Safe Reinforcement Learning

Philip S. Thomas

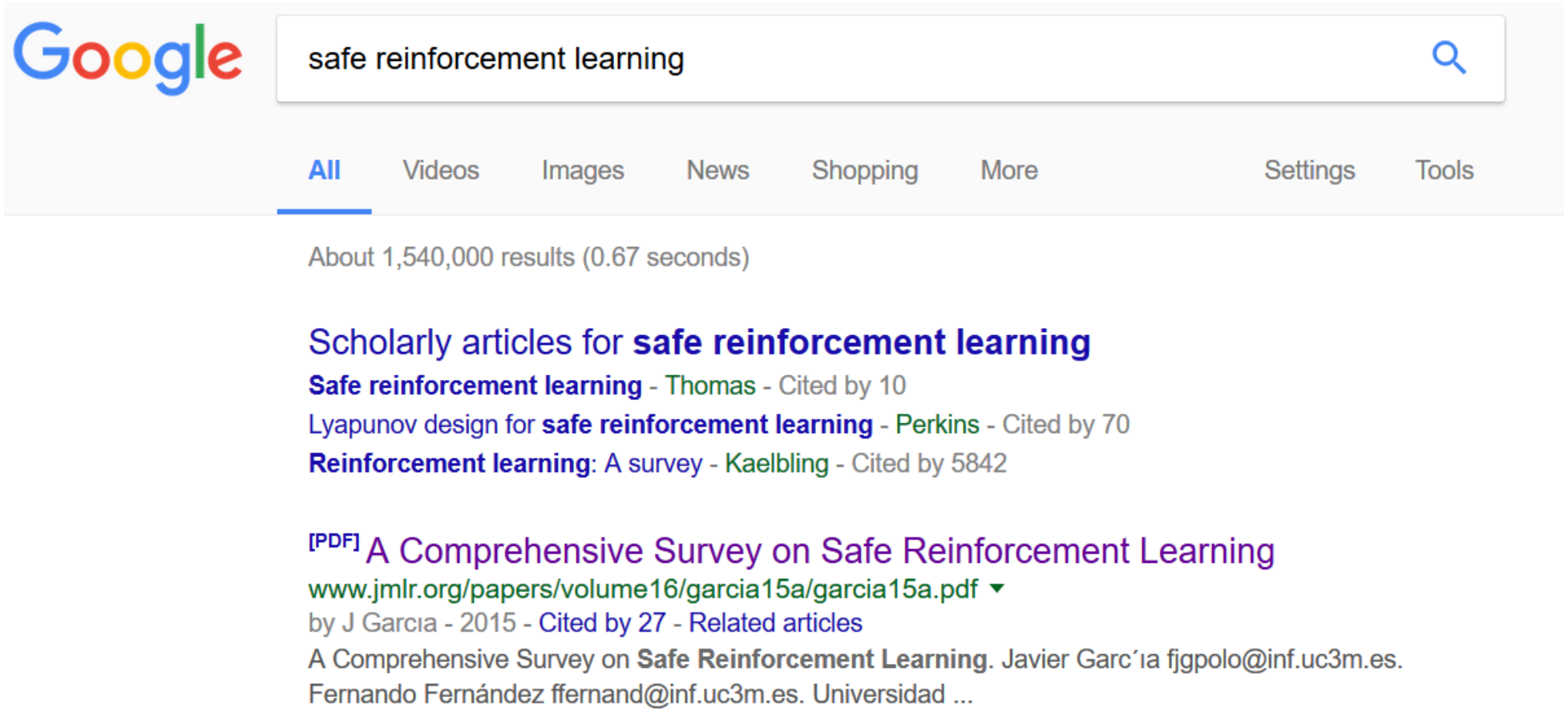
Stanford CS234: Reinforcement Learning, Guest Lecture

May 24, 2017

Lecture overview

- What makes a reinforcement learning algorithm *safe*?
- Notation
- Creating a safe reinforcement learning algorithm
  - Off-policy policy evaluation (OPE)
  - High-confidence off-policy policy evaluation (HCOPE)
  - Safe policy improvement (SPI)
- Empirical results
- Research directions

What does it mean for a reinforcement learning algorithm to be *safe*?



Journal of Machine Learning Research 16 (2015) 1437-1480 Submitted 12/13; Revised 11/14; Published 8/15

A Comprehensive Survey on Safe Reinforcement Learning

Javier García FJGPOLO@INF.UC3M.ES
Fernando Fernández FFERNAND@INF.UC3M.ES
Universidad Carlos III de Madrid,
Avenida de la Universidad 30,
28911 Leganes, Madrid, Spain

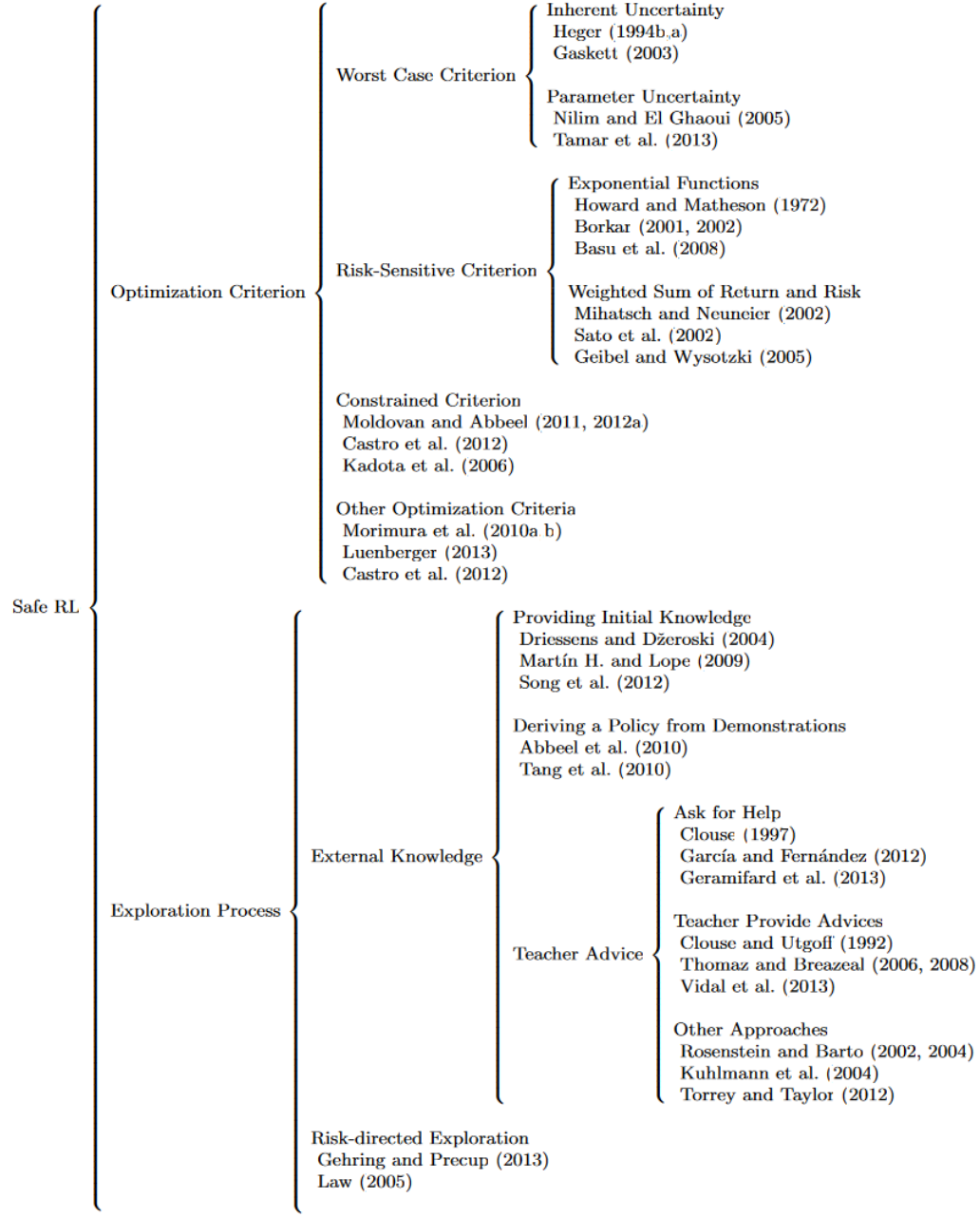
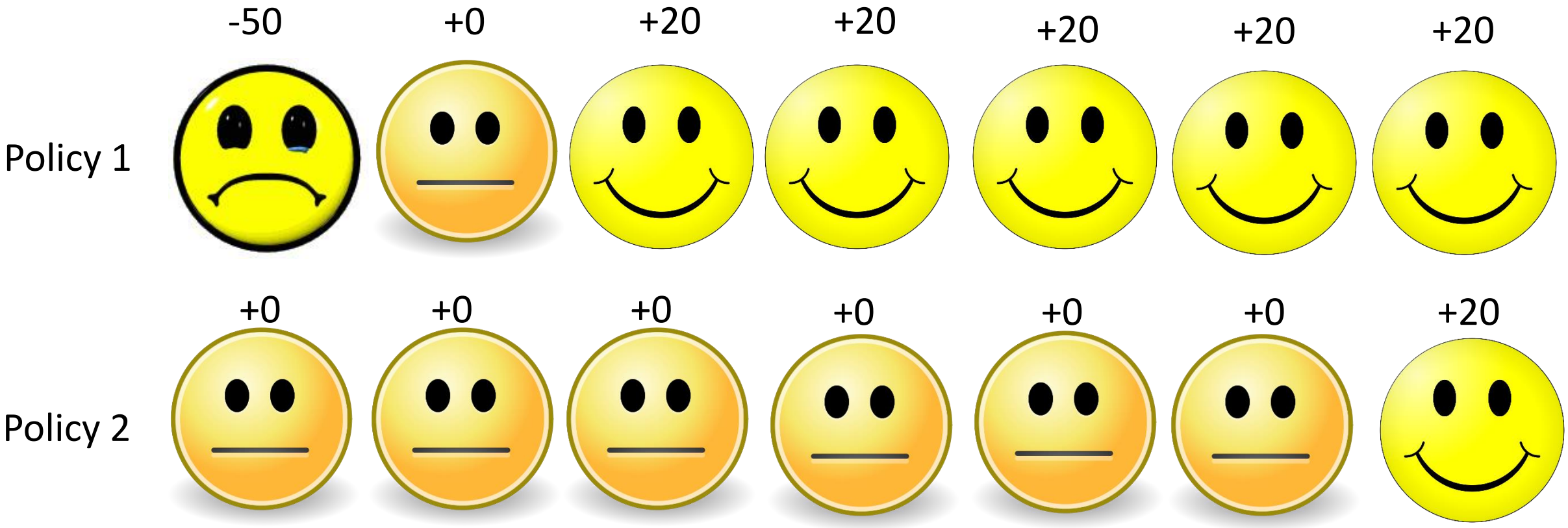


Table 1: Overview of the approaches for Safe Reinforcement Learning considered in this survey.

Changing the objective



Changing the objective

- Policy 1:
  - Reward = 0 with probability 0.999999
  - Reward = 10^9 with probability 1-0.999999
  - Expected reward approximately 1000
- Policy 2:
  - Reward = 999 with probability 0.5
  - Reward = 1000 with probability 0.5
  - Expected reward 999.5

Another notion of safety

Safe and efficient off-policy reinforcement learning

Rémi Munos
munos@google.com
Google DeepMind

Thomas Stepleton
stepleton@google.com
Google DeepMind

Anna Harutyunyan
anna.harutyunyan@vub.ac.be
Vrije Universiteit Brussel

Marc G. Bellemare
bellemare@google.com
Google DeepMind

Another notion of safety (Munos et. al)

We start from the recent work of Harutyunyan et al. (2016), who show that naive off-policy policy evaluation, without correcting for the "off-policyness" of a trajectory, still converges to the desired Q^π value function provided the behavior μ and target π policies are not too far apart (the maximum allowed distance depends on the λ parameter). Their $Q^\pi(\lambda)$ algorithm learns from trajectories generated by μ simply by summing discounted off-policy corrected rewards at each time step. Unfortunately, the assumption that μ and π are close is restrictive, as well as difficult to uphold in the control case, where the target policy is greedy with respect to the current Q-function. In that sense this algorithm is not *safe*: it does not handle the case of arbitrary "off-policyness".

Alternatively, the Tree-backup (TB(λ)) algorithm (Precup et al., 2000) tolerates arbitrary target/behavior discrepancies by scaling information (here called *traces*) from future temporal differences by the product of target policy probabilities. TB(λ) is not *efficient* in the "near on-policy" case (similar μ and π), though, as traces may be cut prematurely, blocking learning from full returns.

Another notion of safety

Reachability-Based Safe Learning with Gaussian Processes

Anayo K. Akametalu* Jaime F. Fisac* Jeremy H. Gillula
Shahab Kaynama Melanie N. Zeilinger Claire J. Tomlin

SAFE REINFORCEMENT LEARNING

A Dissertation Presented

by

PHILIP S. THOMAS

The Problem

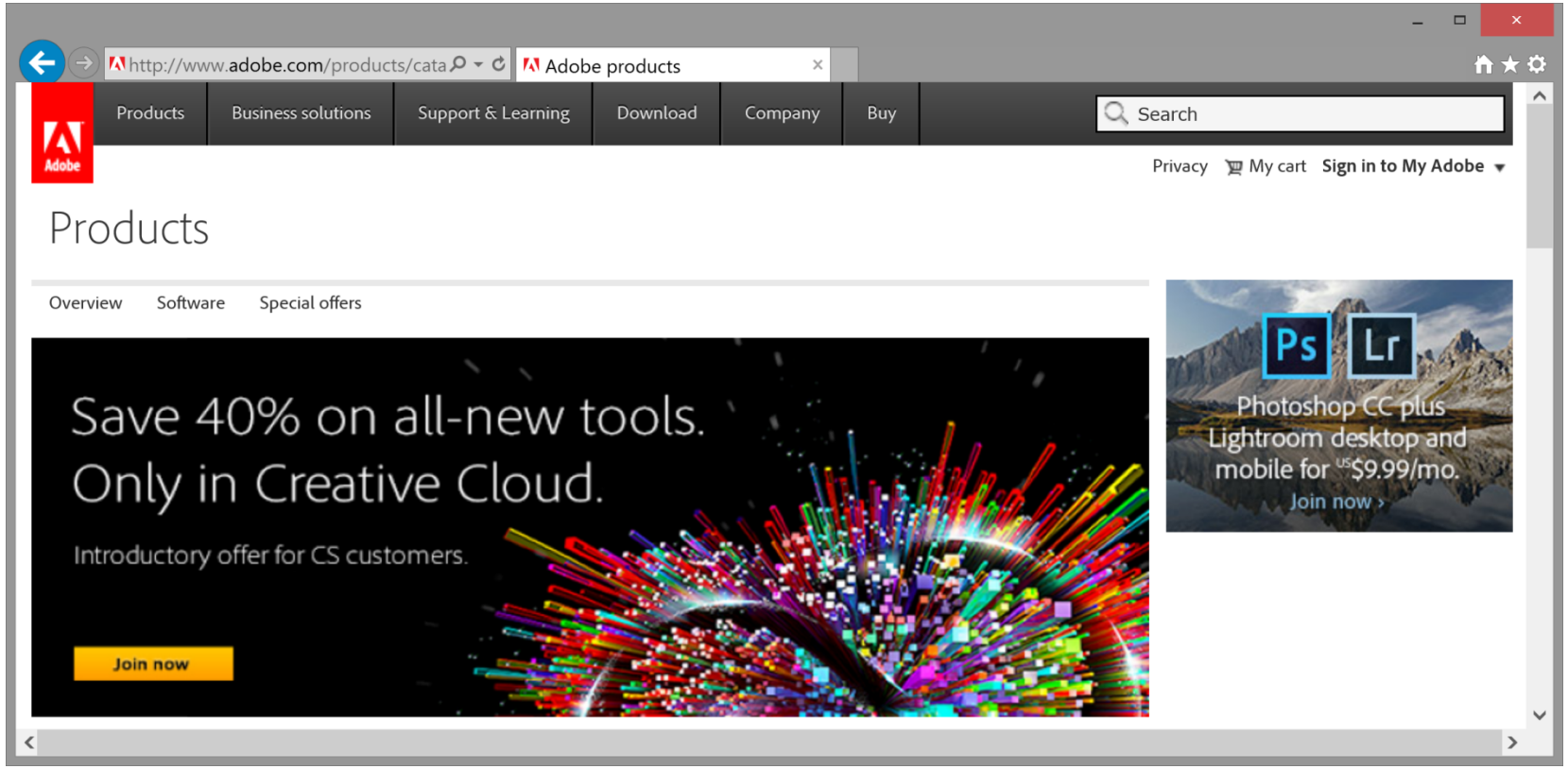
- If you apply an existing method, do you have confidence that it will work?

Reinforcement learning successes

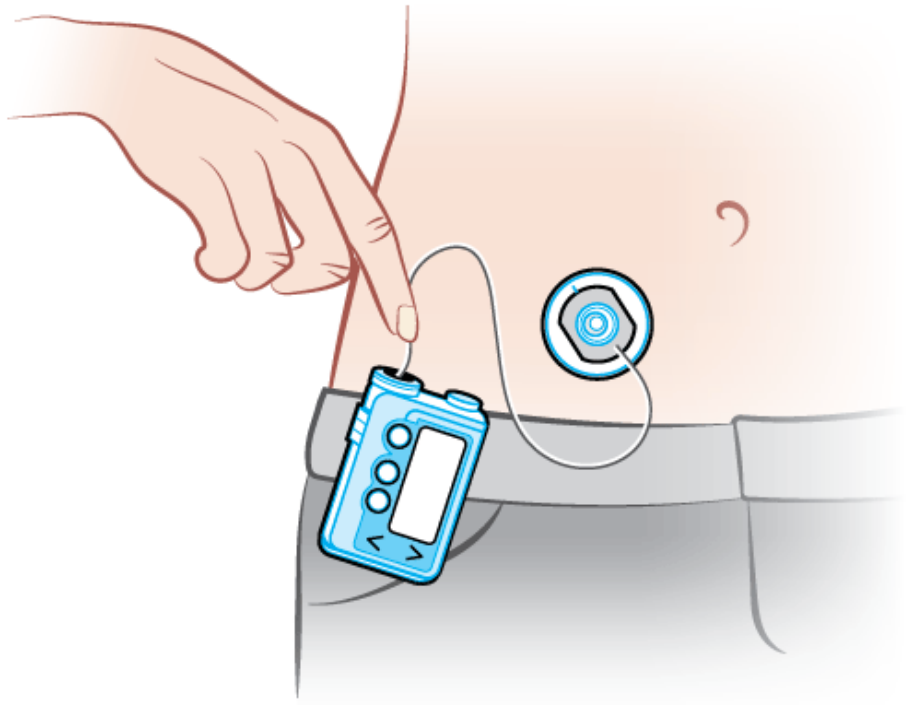
A property of many real applications

- 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* algorithm have?

- Guaranteed to work on the first try
  - "I guarantee that with probability at least $1 - \delta$, I will not change your policy to one that is worse than the current policy."
  - You get to choose δ
  - This guarantee is not contingent on the tuning of any hyperparameters

Lecture overview

- What makes a reinforcement learning algorithm *safe*?
- Notation
- Creating a safe reinforcement learning algorithm
  - Off-policy policy evaluation (OPE)
  - High-confidence off-policy policy evaluation (HCOPE)
  - Safe policy improvement (SPI)
- Empirical results
- Research directions

Notation

- Policy, π

$$\pi(a|s) = \Pr(A_t = a|S_t = s)$$

- History:

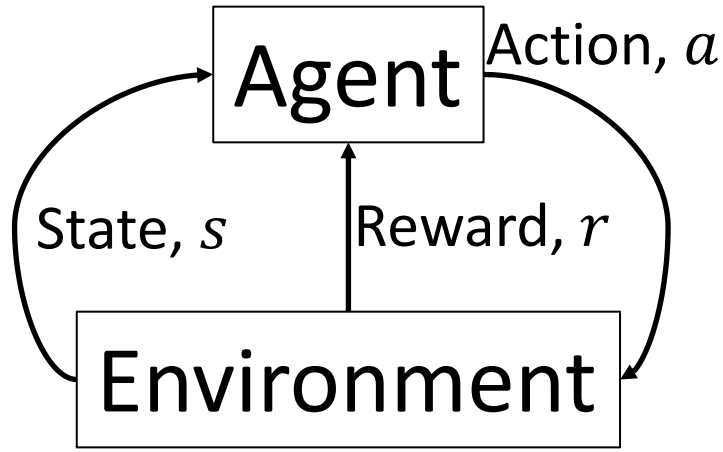
$$H = (s_1, a_1, r_1, s_2, a_2, r_2, \dots, s_L, a_L, r_L)$$

- Historical data:

$$D = \{H_1, H_2, \dots, H_n\}$$

- Historical data from *behavior policy*, π_{b}
- Objective:

$$J(\pi) = \mathbf{E}\left[\sum_{t=1}^{L} \gamma^t R_t \middle| \pi\right]$$



Safe reinforcement learning algorithm

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

$$\Pr\left(J\left(a(D)\right) \geq J(\pi_{\mathrm{b}})\right) \geq 1 - \delta$$

or, in general:

$$\Pr\left(J\left(a(D)\right) \geq J_{\min}\right) \geq 1 - \delta$$

Lecture overview

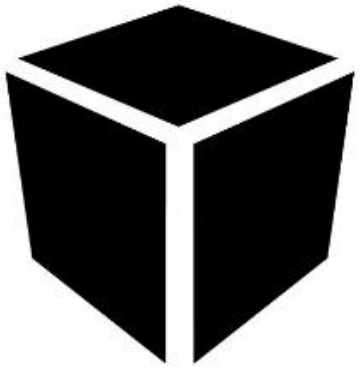
- What makes a reinforcement learning algorithm *safe*?
- Notation
- Creating a safe reinforcement learning algorithm
  - Off-policy policy evaluation (OPE)
  - High-confidence off-policy policy evaluation (HCOPE)
  - Safe policy improvement (SPI)
- Empirical results
- Research directions

Creating a safe 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 $J(\pi_{\mathrm{e}})$
- High-confidence off-policy policy evaluation (HCOPE)
  - Use a concentration inequality to convert the n independent and unbiased estimates of $J(\pi_{\mathrm{e}})$ into a $1-\delta$ confidence lower bound on $J(\pi_{\mathrm{e}})$
- Safe policy improvement (SPI)
  - Use HCOPE method to create a safe reinforcement learning algorithm, a

Off-policy policy evaluation (OPE)

Historical Data, D
Proposed Policy, π_{e}

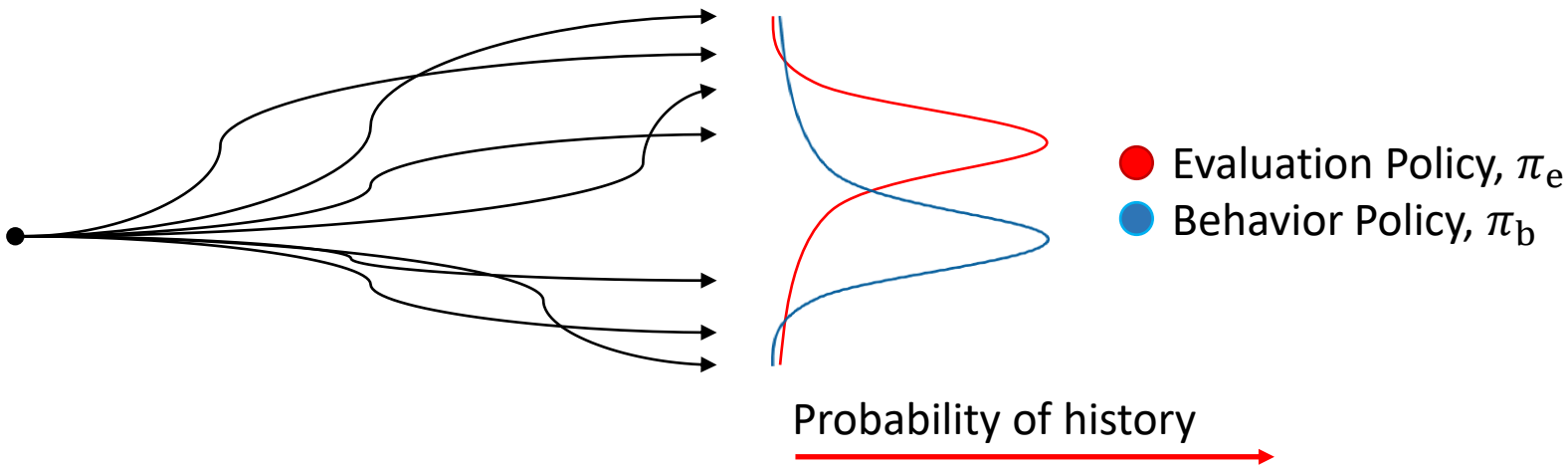


Estimate of $J(\pi_{\mathrm{e}})$

Importance Sampling (Intuition)

- Reminder:
  - History, $H = (s_1, a_1, r_1, s_2, a_2, r_2, \ldots, s_L, a_L, r_L)$
  - Objective, $J(\pi_e) = \mathbf{E}[\sum_{t=1}^{L} \gamma^t R_t \,|\pi_e]$

$$\hat{J}(\pi_e) = \frac{1}{n} \sum_{i=1}^{n} w_i \sum_{t=1}^{L} \gamma^t R_t^i$$



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

Importance sampling (History)

- Kahn, H., Marhshall, A. W. (1953). Methods of reducing sample size in Monte Carlo computations. In *Journal of the Operations Research Society of America*, 1(5):263–278
  - Let $X = 0$ with probability $1 - 10^{-10}$ and $X = 10^{10}$ with probability 10^{-10}
  - $\mathbf{E}[X] = 1$
  - Monte Carlo estimate from $n \ll 10^{10}$ samples of X is almost always zero
  - Idea: Sample X from some other distribution and use importance sampling to "correct" the estimate
  - Can produce lower variance estimates.
- Josiah Hannah et. al, ICML 2017 (to appear).

Importance sampling (History, continued)

- Precup, D., Sutton, R. S., Singh, S. (2000). Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, pp. 759–766. Morgan Kaufmann

Importance sampling (Proof)

- Estimate $\mathbf{E}_p[f(X)]$ given a sample of $X \sim q$
- Let $P = \text{supp}(p)$, $Q = \text{supp}(q)$, and $F = \text{supp}(f)$
- Importance sampling estimate: $\frac{p(X)}{q(X)} f(X)$

$$\mathbf{E}_q\left[\frac{p(X)}{q(X)} f(X)\right] = \sum_{x \in Q} q(X) \frac{p(X)}{q(X)} f(X)$$

$$= \sum_{x \in P} p(X)\, f(X) + \sum_{x \in \bar{P} \cap Q} p(X)\, f(X) - \sum_{x \in P \cap \bar{Q}} p(X) f(X)$$

$$= \sum_{x \in P} p(X)\, f(X) - \sum_{x \in P \cap \bar{Q}} p(X) f(X)$$

Importance sampling (Proof)

- Assume $P \subseteq Q$ (can relax assumption to $P \subseteq Q \cup \bar{F}$)

$$\mathbf{E}_q\left[\frac{p(X)}{q(X)} f(X)\right] = \sum_{x \in P} p(X)\, f(X) - \sum_{x \in P \cap \bar{Q}} p(X) f(X)$$

$$= \sum_{x \in P} p(X)\, f(X)$$

$$= \mathbf{E}_p[f(X)]$$

- Importance sampling is an unbiased estimator of $\mathbf{E}_p[f(X)]$

Importance sampling (proof)

- Assume $f(x) \geq 0$ for all x

$$\begin{aligned}\mathbf{E}_q\left[\frac{p(X)}{q(X)} f(X)\right] &= \sum_{x \in P} p(X)\, f(X) - \sum_{x \in P \cap \bar{Q}} p(X) f(X) \\ &\leq \sum_{x \in P} p(X)\, f(X) \\ &= \mathbf{E}_p[f(X)]\end{aligned}$$

- Importance sampling is a negative-bias estimator of $\mathbf{E}_p[f(X)]$

Importance sampling (reminder)

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

$$\mathbf{E}[\mathrm{IS}(D)] = J(\pi_{\mathrm{e}})$$

Creating a safe 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 $J(\pi_e)$
- High-confidence off-policy policy evaluation (HCOPE)
  - Use a concentration inequality to convert the n independent and unbiased estimates of $J(\pi_e)$ into a $1-\delta$ confidence lower bound on $J(\pi_e)$
- Safe policy improvement (SPI)
  - Use HCOPE method to create a safe reinforcement learning algorithm, a

High confidence off-policy policy evaluation (HCOPE)

Historical Data, D

Proposed Policy, π_e

Probability, $1-\delta$

$1-\delta$ confidence lower bound on $J(\pi_e)$

Hoeffding's inequality

- Let $X_1, \ldots, X_n$ be n independent identically distributed random variables such that $X_\mathrm{i} \in [0, b]$
- Then with probability at least $1 - \delta$:

$$\mathbf{E}[X_i] \geq \overbrace{\frac{1}{n}\sum_{i=1}^{n} X_i}^{\frac{1}{n}\sum_{i=1}^{n}\left(w_i \sum_{t=1}^{L} \gamma^t R_t^i\right)} - b\sqrt{\frac{\ln\left(1/\delta\right)}{2n}}$$

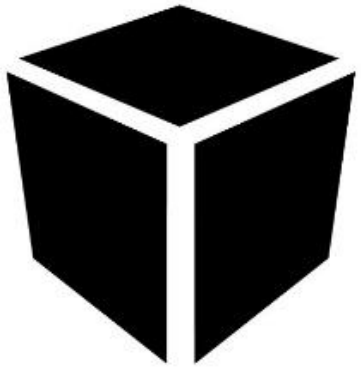
Creating a safe reinforcement learning algorithm

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

Safe policy improvement (SPI)

Historical Data, D

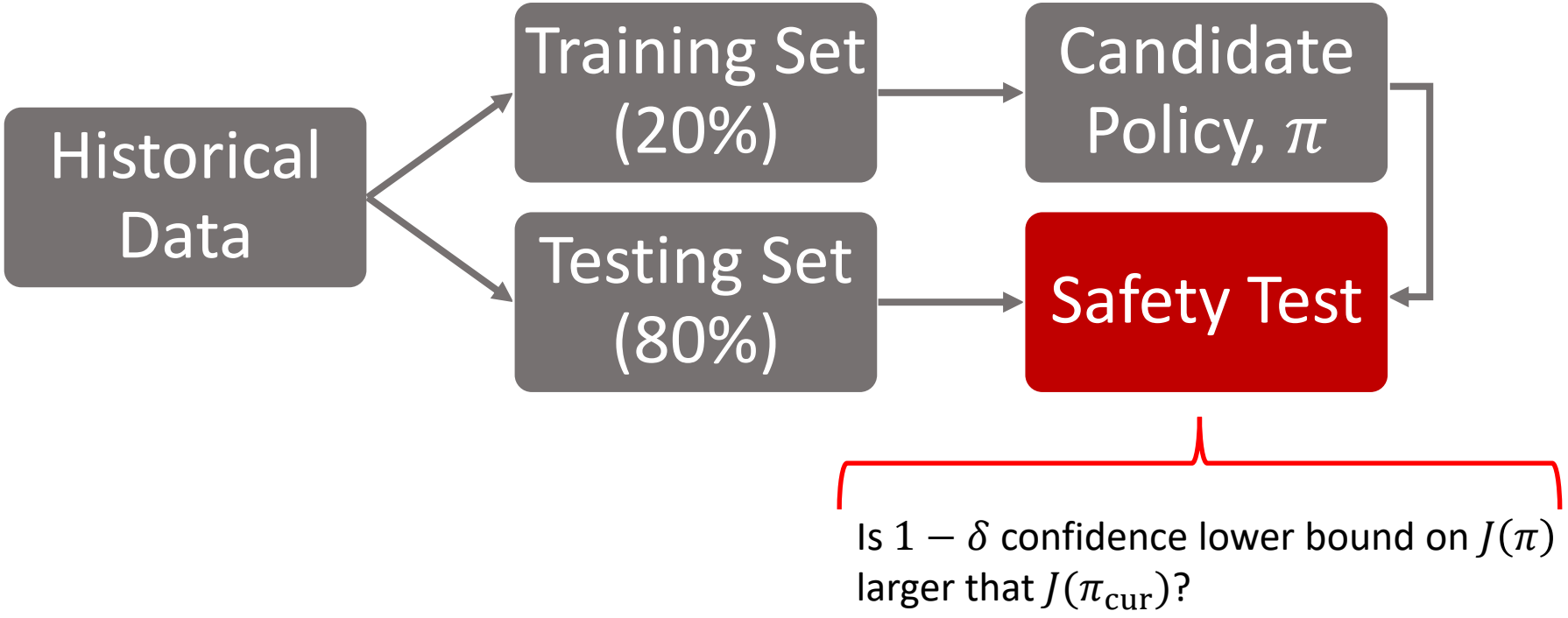
Probability, $1 - \delta$



New policy π, or

`No Solution Found`

Safe policy improvement (SPI)



Creating a safe 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 $J(\pi_e)$
- High-confidence off-policy policy evaluation (HCOPE)
  - Use a concentration inequality to convert the n independent and unbiased estimates of $J(\pi_e)$ into a $1-\delta$ confidence lower bound on $J(\pi_e)$
- Safe policy improvement (SPI)
  - Use HCOPE method to create a safe reinforcement learning algorithm, a

WON'T WORK

Off-policy policy evaluation (revisited)

- Importance sampling (IS):

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

- Per-decision importance sampling (PDIS)

$$\mathrm{PDIS}(D) = \sum_{t=1}^{L}\gamma^t\frac{1}{n}\sum_{i=1}^{n}\left(\prod_{\tau=1}^{t}\frac{\pi_{\mathrm{e}}(a_\tau|s_\tau)}{\pi_{\mathrm{b}}(a_\tau|s_\tau)}\right)R_t^i$$

Off-policy policy evaluation (revisited)

- Importance sampling (IS):

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

- Weighted importance sampling (WIS)

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

Off-policy policy evaluation (revisited)

- Weighted importance sampling (WIS)

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

- **Not** unbiased. When $n = 1$, $\mathbf{E}[\mathrm{WIS}] = J(\pi_\mathrm{b})$
- Strongly consistent estimator of $J(\pi_\mathrm{e})$
  - i.e., $\Pr\left(\lim_{n \to \infty} \mathrm{WIS}(D) = \mathrm{J}(\pi_\mathrm{e}) \right) = 1$
  - 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: $\mu = \mathbf{E}[X]$
- $\hat{\mu} = X$
- Unbiased: $\mathbf{E}[\hat{\mu}] = \mathbf{E}[X] = \mu$
- Variance: $\mathrm{Var}(\hat{\mu}) = \mathrm{Var}(X)$

Control variates

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

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

- Variance:

$$\mathrm{Var}(\hat{\mu}\,) = \mathrm{Var}(X - Y + \mathbf{E}[Y]) = \mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X, Y)$$

- Lower variance if $2\mathrm{Cov}(X, Y) > \mathrm{Var}(Y)$
- We call Y a *control variate.*

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
    - $\mathbf{E}[Y] = 0$ in this case
  - $\mathrm{PDIS}_{\mathrm{CV}}(D) = \mathrm{PDIS}(D) - \mathrm{CV}(D)$
- 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
  - $\mathrm{DR}(D) = \mathrm{PDIS}_{\mathrm{CV}}(D)$
- Non-recursive and weighted forms, as well as control variate view provided by Thomas and Brunskill (2016)

Off-policy policy evaluation (revisited)

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

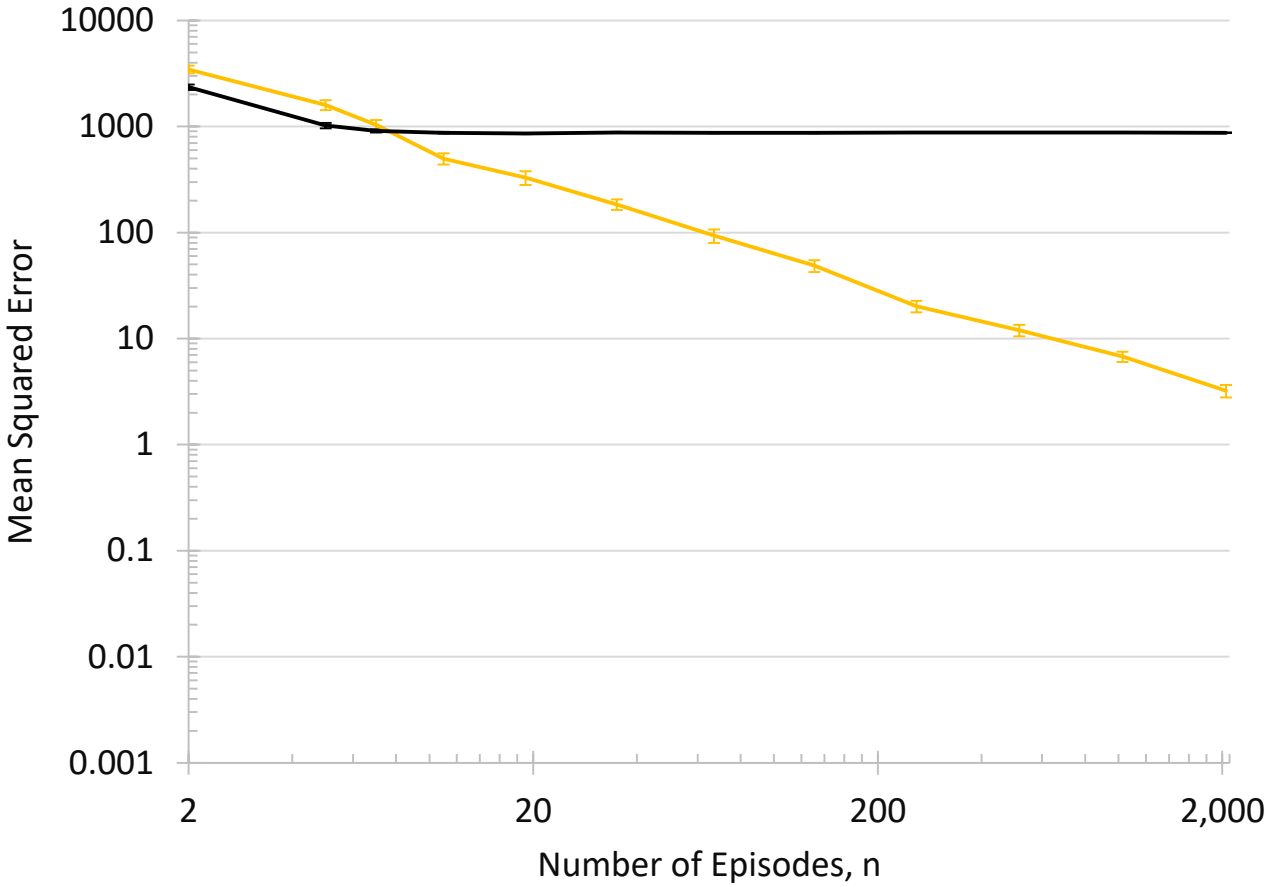
Per-decision importance sampling (PDIS)

$$w_t^i = \prod_{\tau=1}^{t}\frac{\pi_\mathrm{e}(a_\tau|s_\tau)}{\pi_\mathrm{b}(a_\tau|s_\tau)}$$

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

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

Empirical Results (Gridworld)

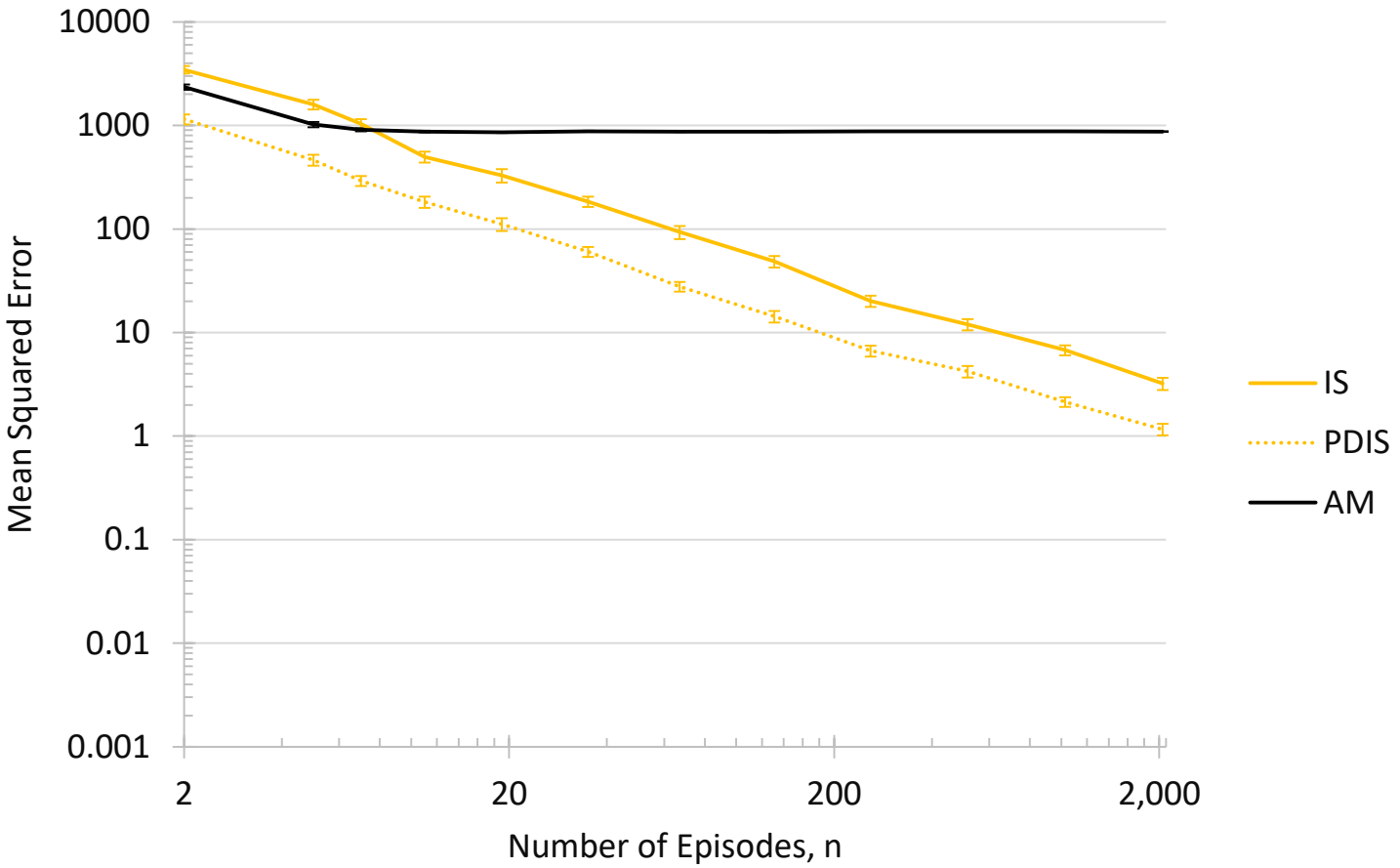


Approximate model

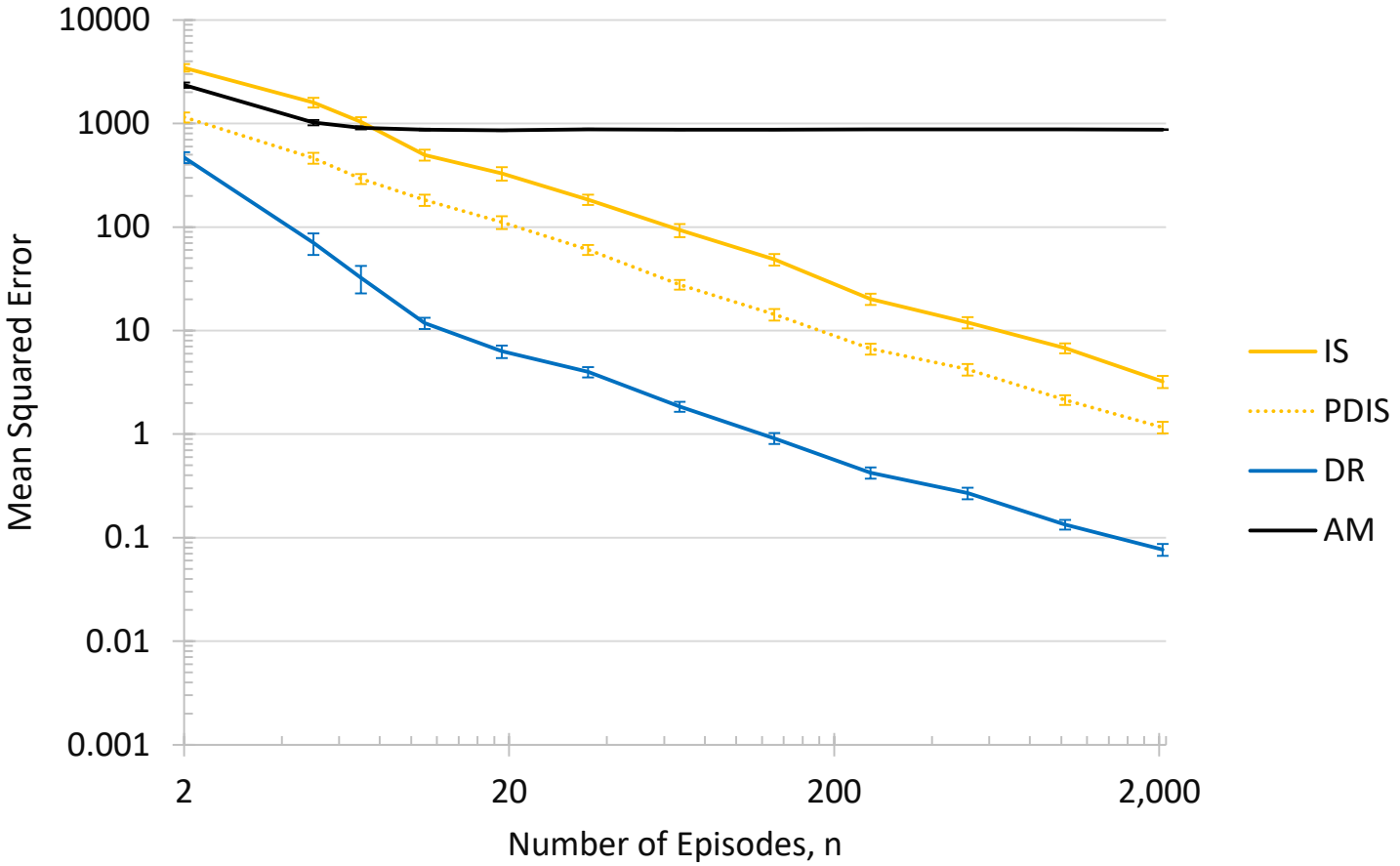
Direct method (Dudik, 2011)

Indirect method (Sutton and Barto, 1998)

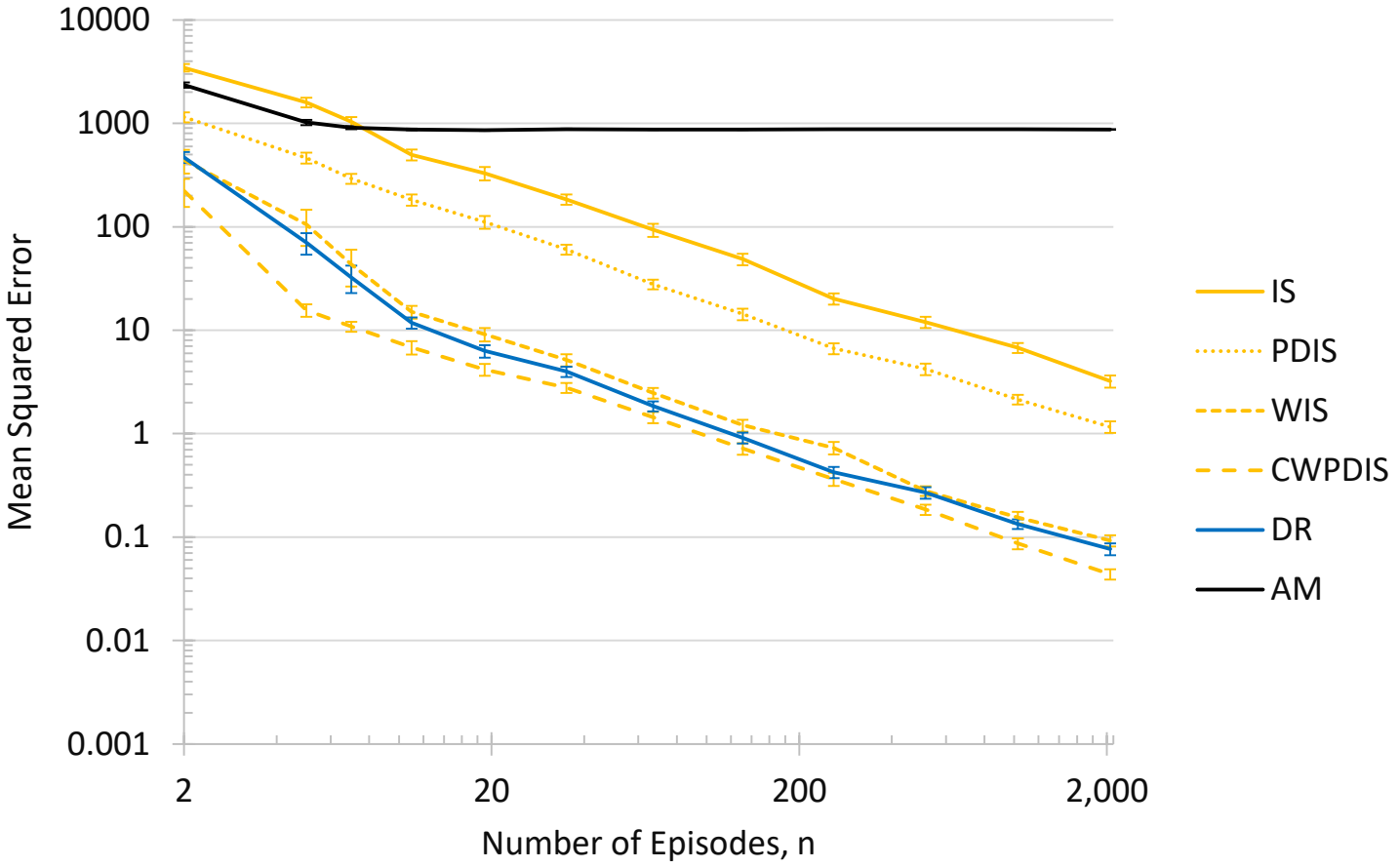
Empirical Results (Gridworld)



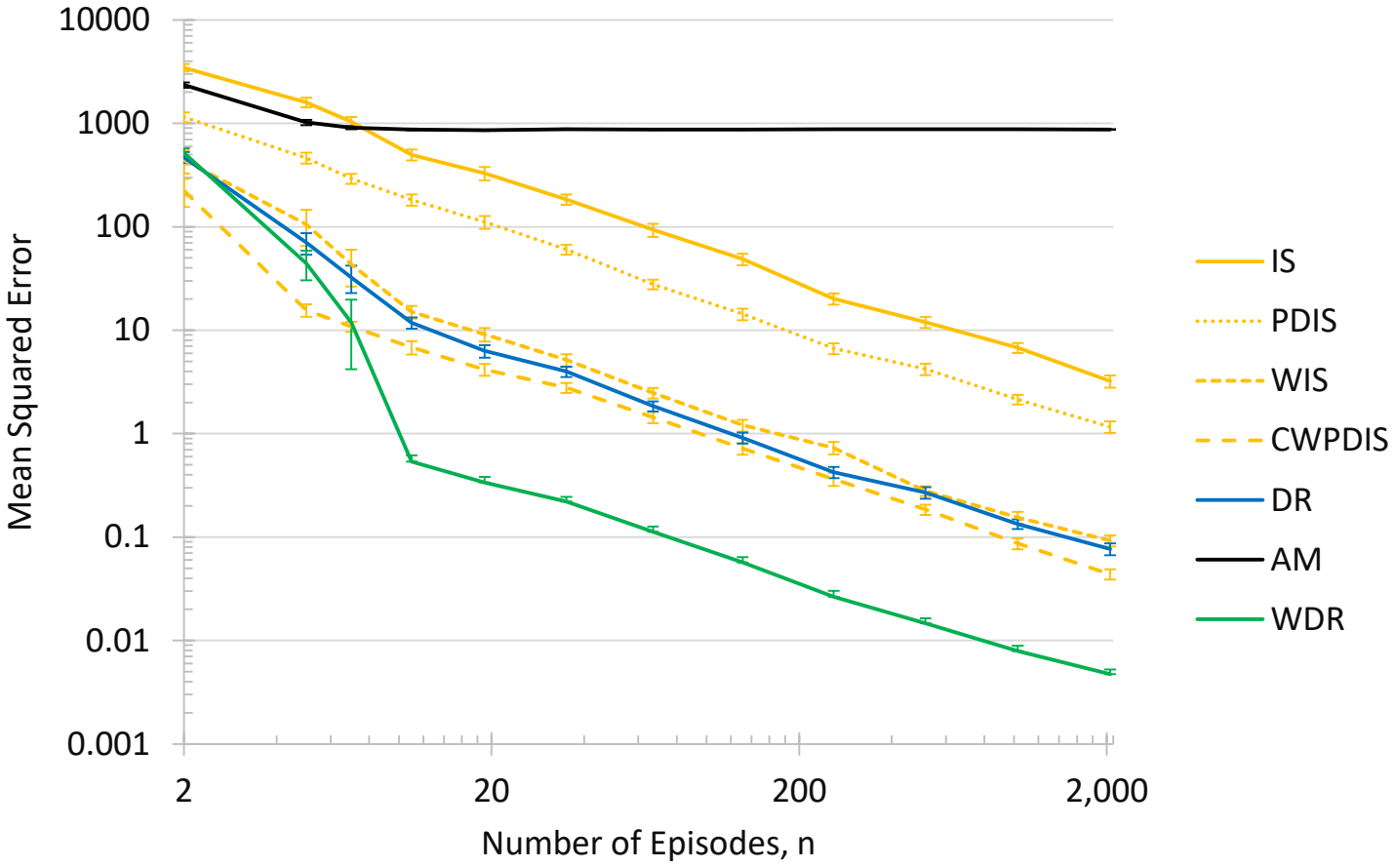
Empirical Results (Gridworld)



Empirical Results (Gridworld)



Empirical Results (Gridworld)

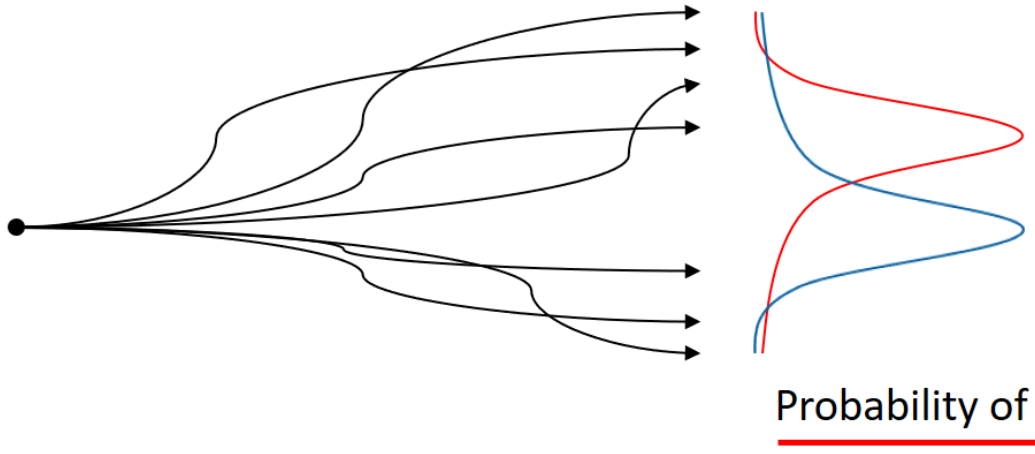


Off-policy policy evaluation (revisited)

- What if $\text{supp}(\pi_e) \subset \text{supp}(\pi_b)$?
- There is a state-action pair, (s, a), such that

$$\pi_e(a|s) = 0, \text{ but } \pi_b(a|s) \neq 0$$

- If we see a history where (s, a) occurs, what weight should we give it?

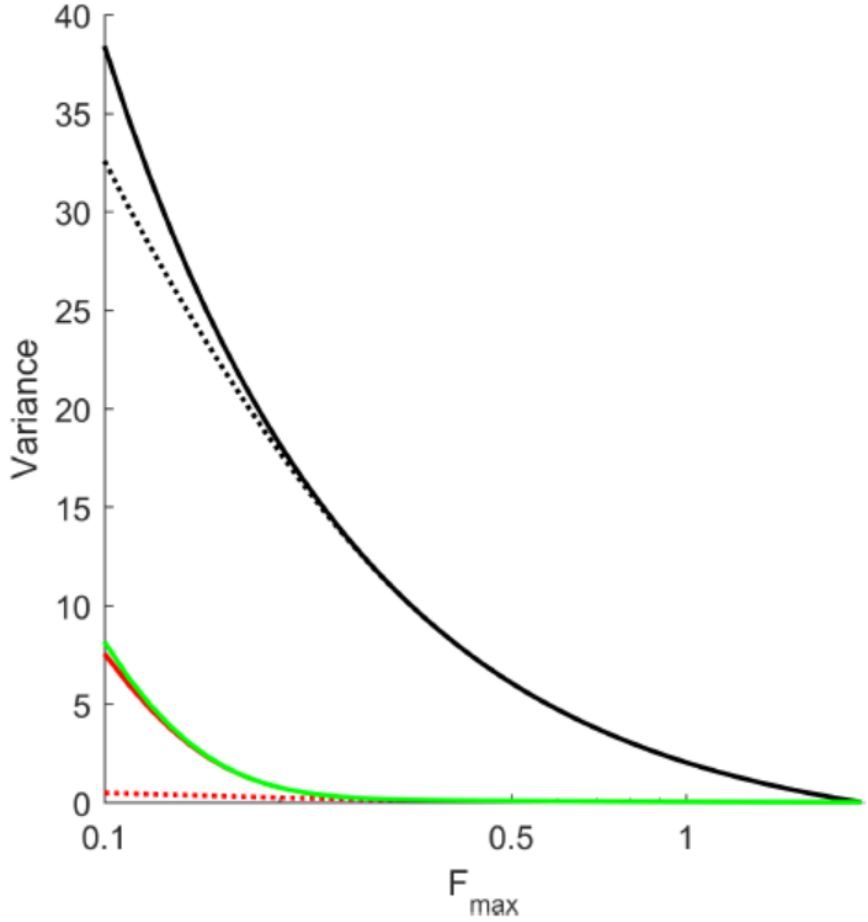


$$\text{IS}(D) = \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{t=1}^{L}\frac{\pi_e(a_t|s_t)}{\pi_b(a_t|s_t)}\right)\left(\sum_{t=1}^{L}\gamma^t R_t^i\right)$$

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)

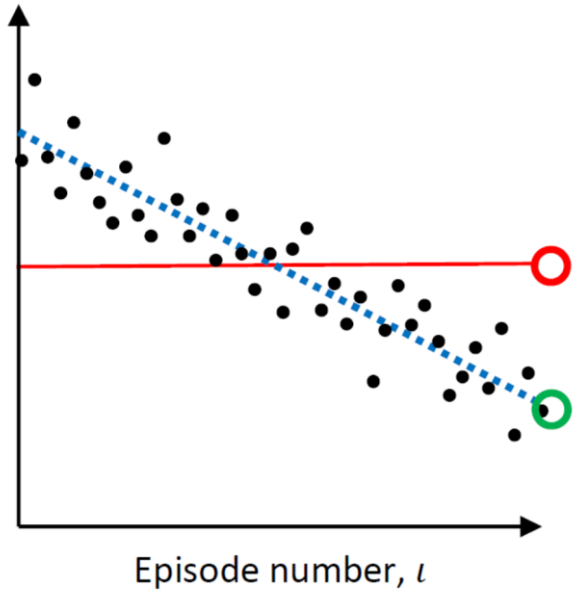
Off-policy policy evaluation (revisited)

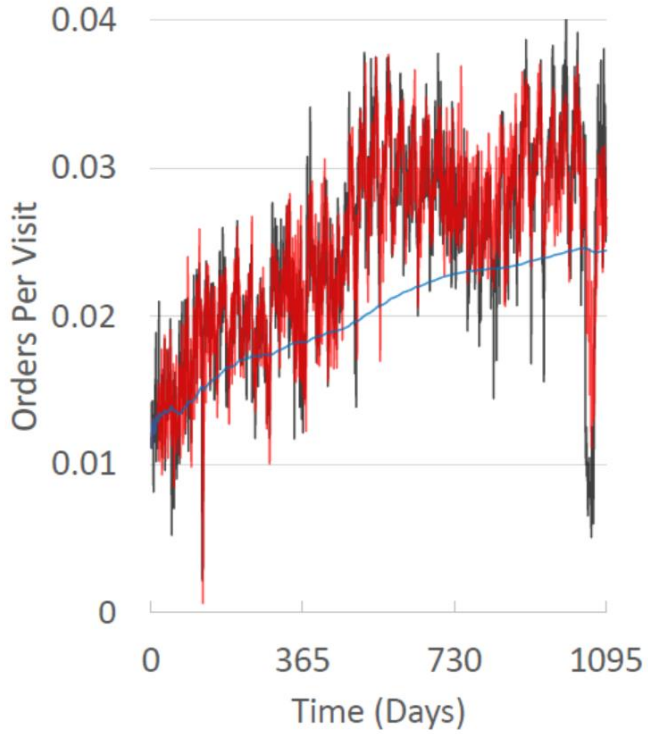


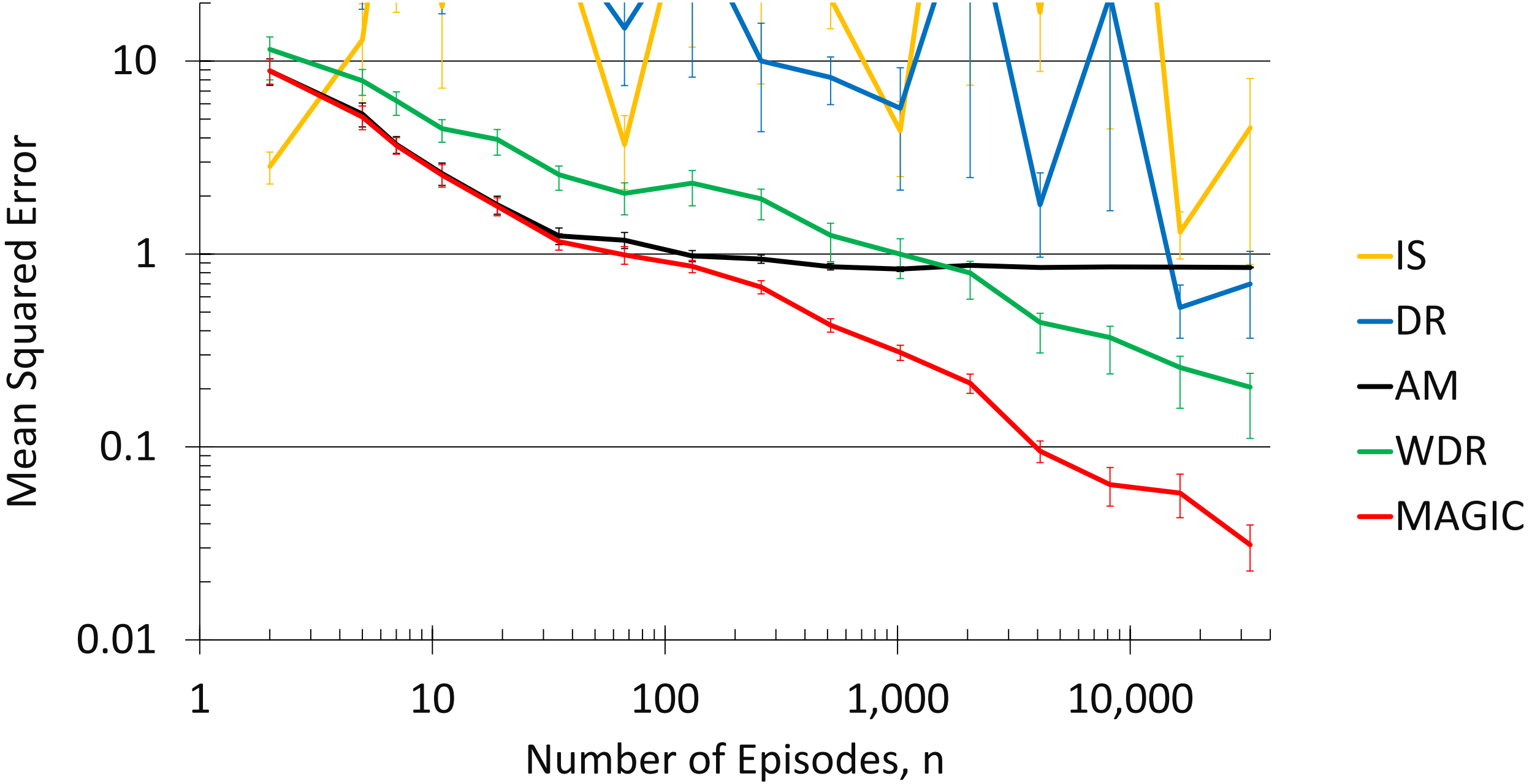
IS, IS (conditional), US, US (conditional) MSE(US)

Off-policy policy evaluation (revisited)

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





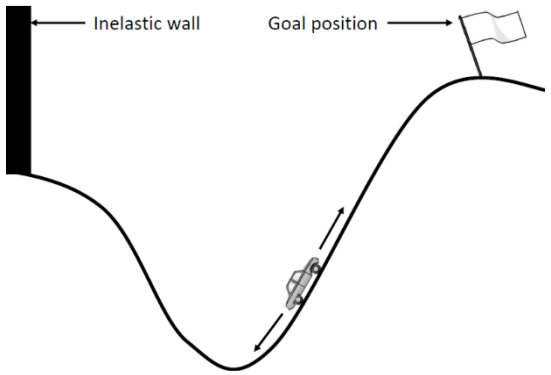

Mean Squared Error
10
1
0.1
0.01
1
10
100
1,000
10,000
Number of Episodes, n
IS
DR
AM
WDR
MAGIC

Creating a safe 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 $J(\pi_e)$
- High-confidence off-policy policy evaluation (HCOPE)
  - Use a concentration inequality to convert the n independent and unbiased estimates of $J(\pi_e)$ into a $1-\delta$ confidence lower bound on $J(\pi_e)$
- Safe policy improvement (SPI)
  - Use HCOPE method to create a safe reinforcement learning algorithm, a

High-confidence off-policy policy evaluation (revisited)

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



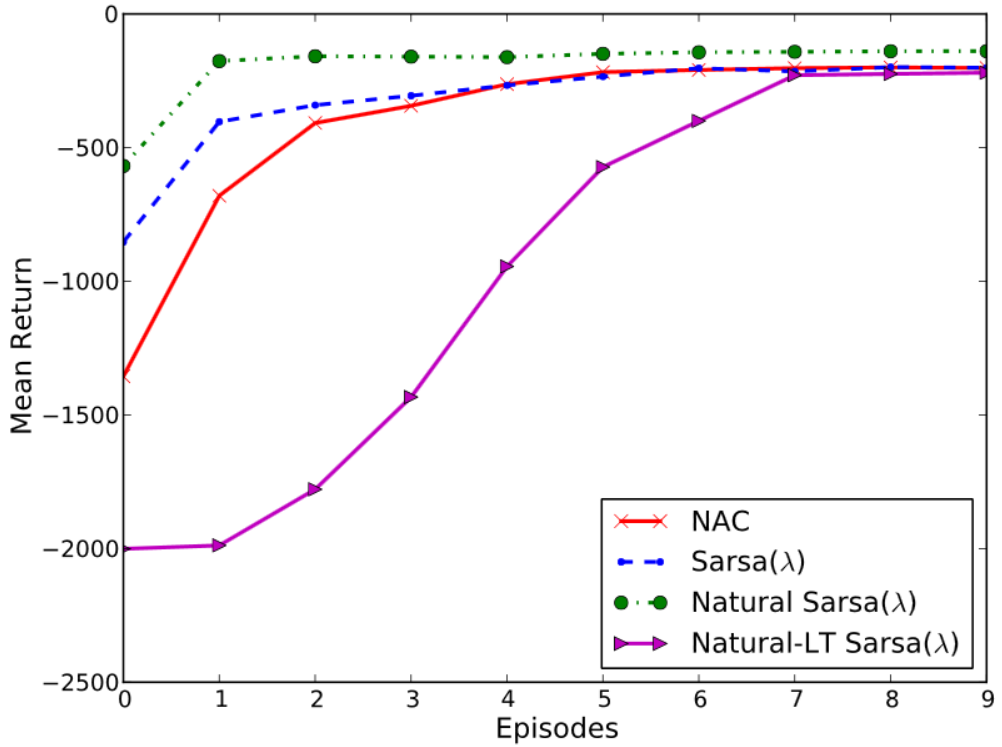


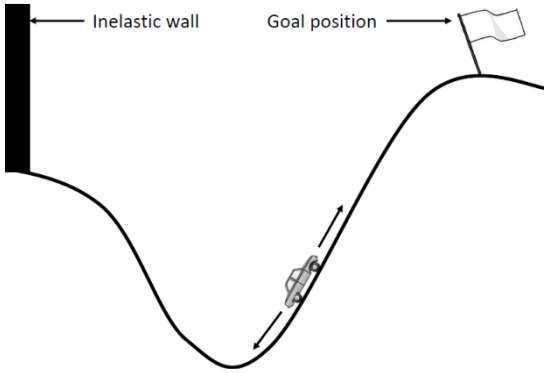
Figure 3: Mountain Car (Sarsa(λ))

Natural Temporal Difference Learning, Dabney and Thomas, 2014

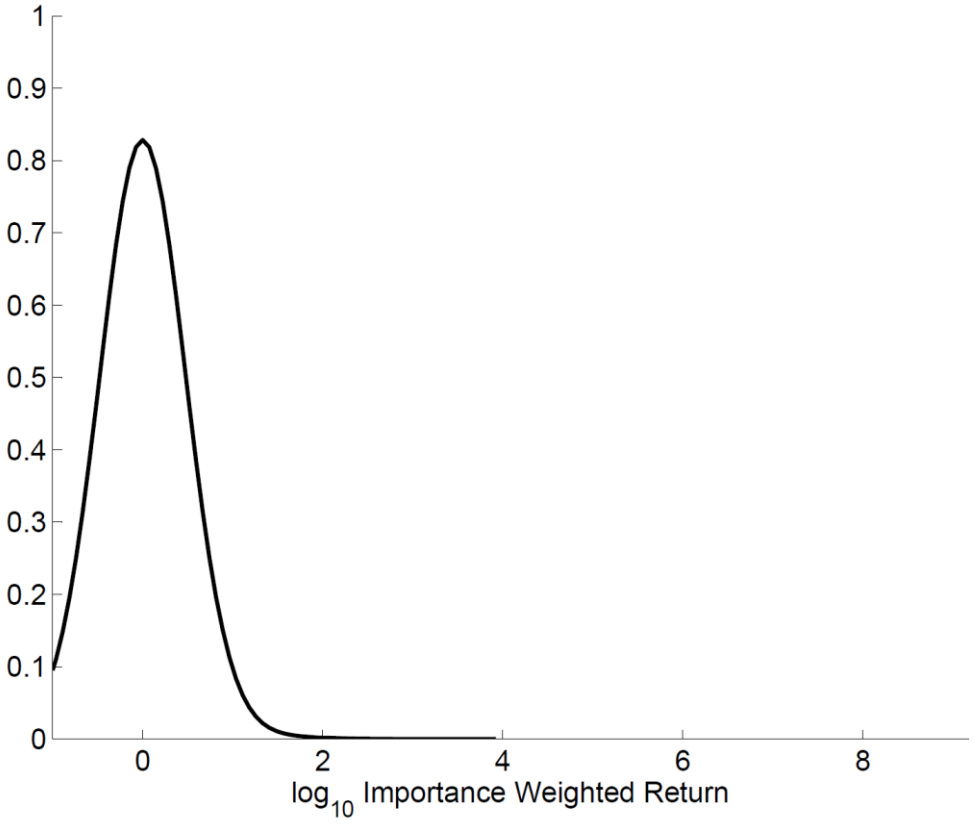
High-confidence off-policy policy evaluation (revisited)

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

$$-5{,}831{,}000$$



What went wrong?



$$w_i = \prod_{t=1}^{L} \frac{\pi_{\mathrm{e}}(a_t|s_t)}{\pi_{\mathrm{b}}(a_t|s_t)}$$

What went wrong?

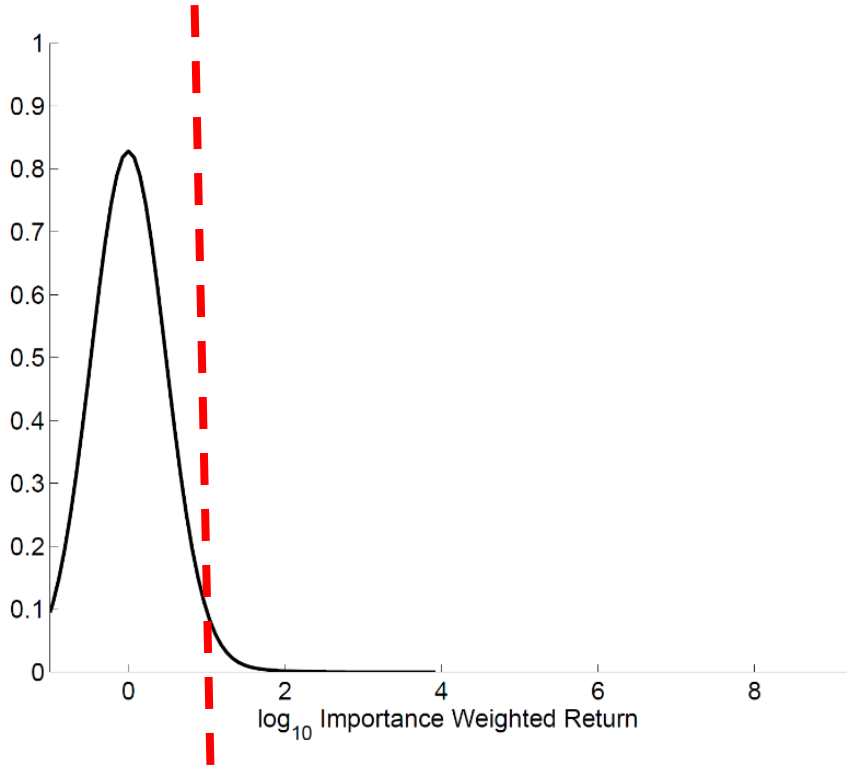
$$\mathbf{E}[X_i] \geq \frac{1}{n}\sum_{i=1}^{n} X_i - b\sqrt{\frac{\ln\left(1/\delta\right)}{2n}}$$

$$b \approx 10^{9.4}$$

Largest observed importance weighted return: 316.

High-confidence off-policy policy evaluation (revisited)

- Removing the upper tail only decreases the expected value.



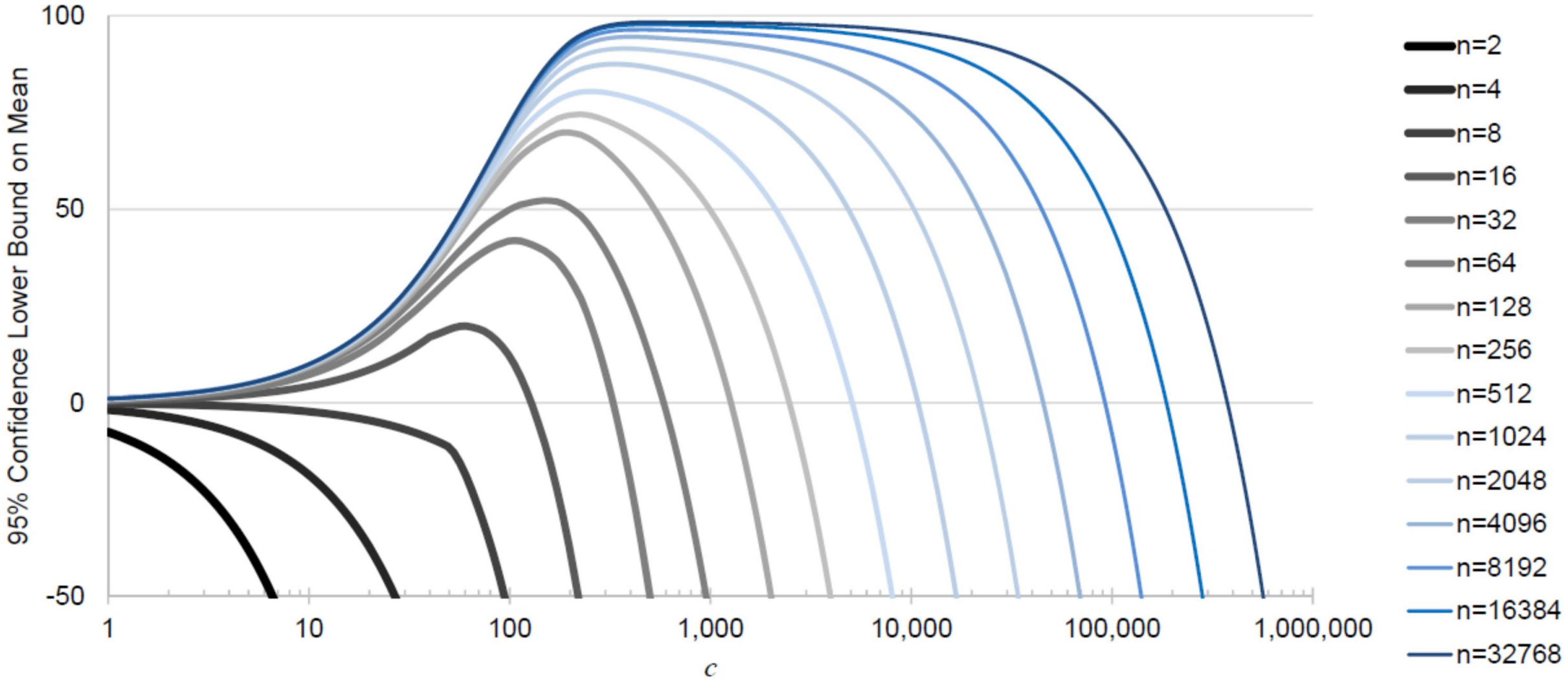
High-confidence off-policy policy evaluation (revisited)

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*

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

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

High-confidence off-policy policy evaluation (revisited)



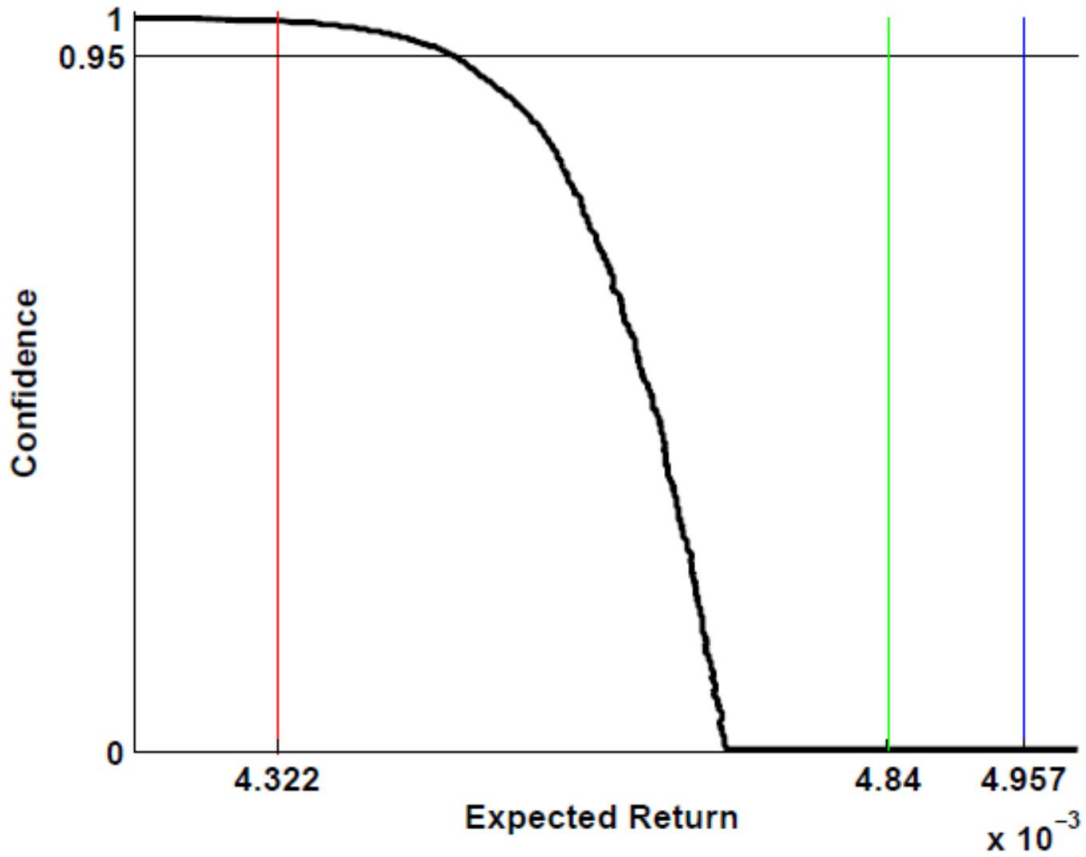
High-confidence off-policy policy evaluation (revisited)

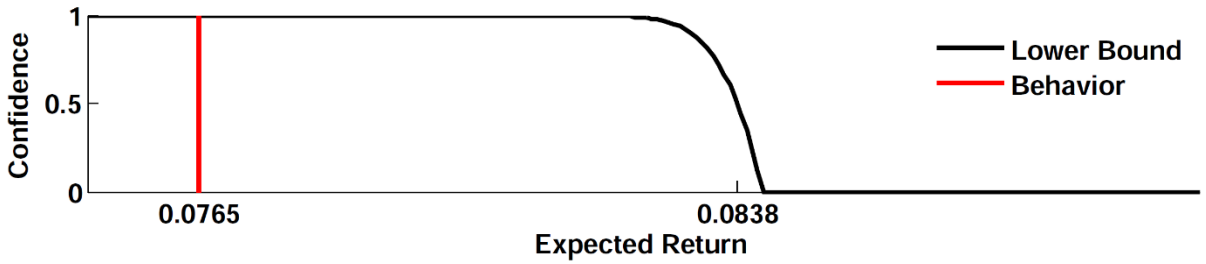
- Use 20% of the data to optimize c.
- Use 80% to compute lower bound with optimized c.
- Mountain car results:

	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)

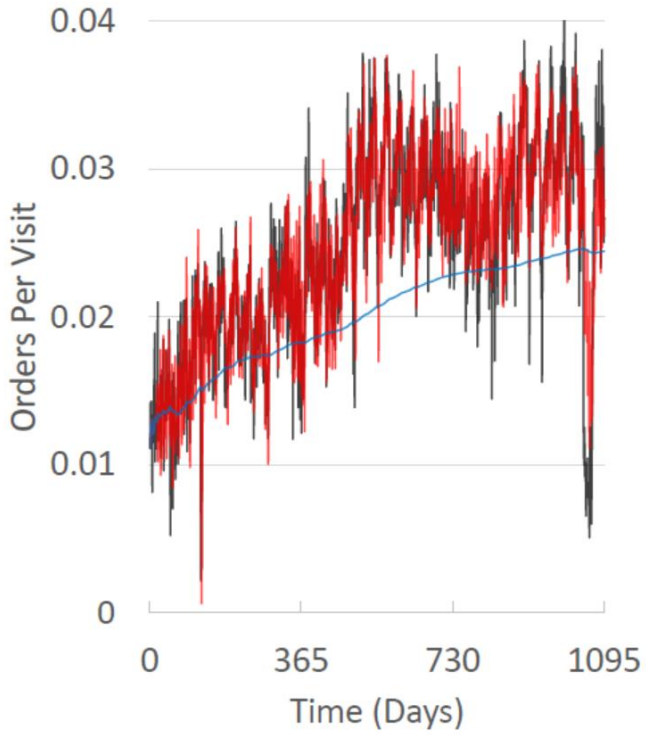
- Digital Marketing:





High-confidence off-policy policy evaluation (revisited)

- Cognitive dissonance



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

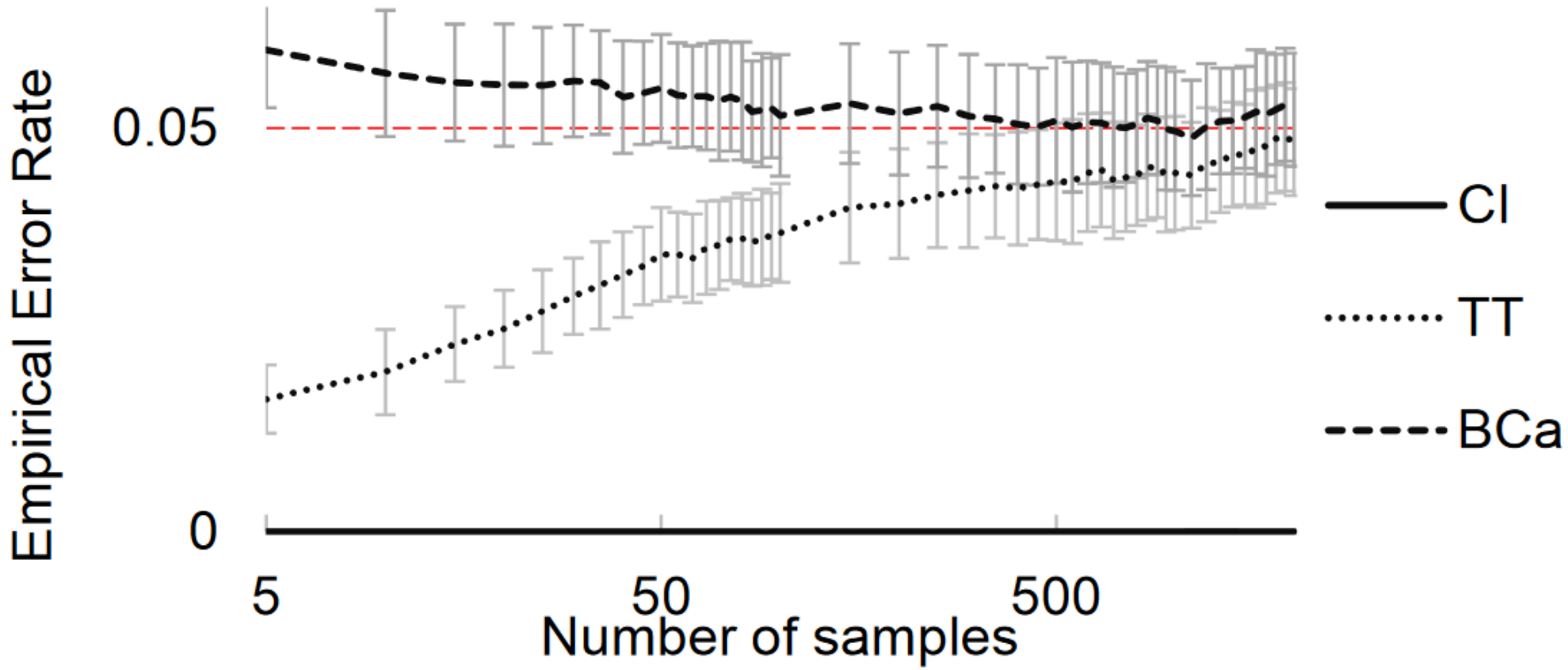
High-confidence off-policy policy evaluation (revisited)

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

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

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

High-confidence off-policy policy evaluation (revisited)



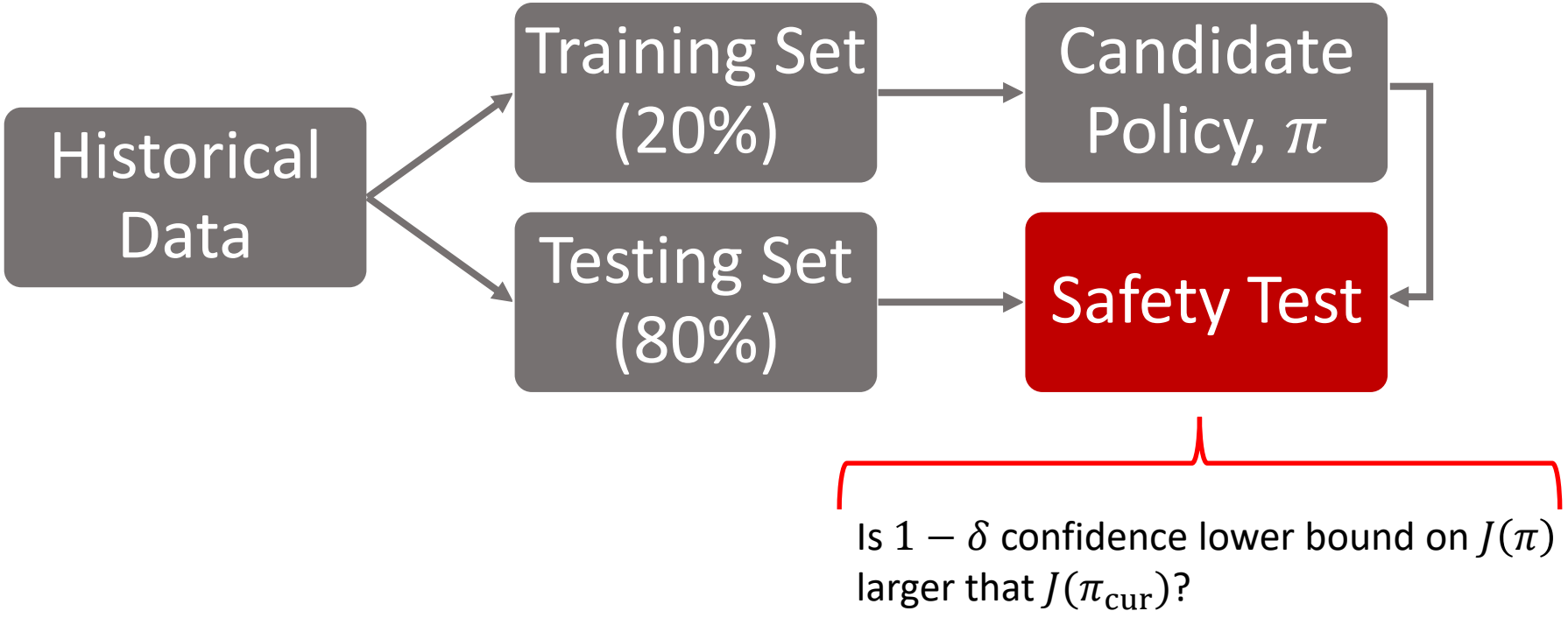
P. S. Thomas. Safe reinforcement learning (PhD Thesis, 2015)

Creating a safe 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 $J(\pi_e)$
- High-confidence off-policy policy evaluation (HCOPE)
  - Use a concentration inequality to convert the n independent and unbiased estimates of $J(\pi_e)$ into a $1-\delta$ confidence lower bound on $J(\pi_e)$
- Safe policy improvement (SPI)
  - Use HCOPE method to create a safe reinforcement learning algorithm, a

Safe policy improvement (revisited)

- Thomas et. al, ICML 2015



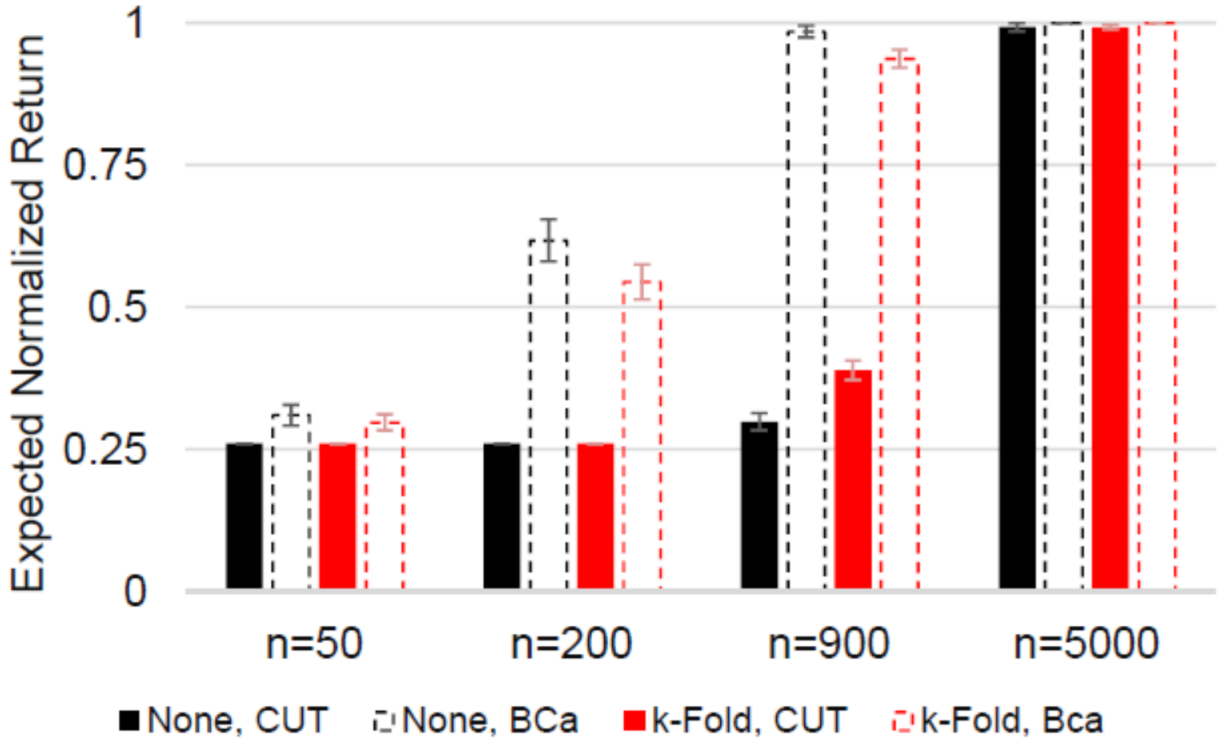
Lecture overview

- What makes a reinforcement learning algorithm *safe*?
- Notation
- Creating a safe reinforcement learning algorithm
  - Off-policy policy evaluation (OPE)
  - High-confidence off-policy policy evaluation (HCOPE)
  - Safe policy improvement (SPI)
- Empirical results
- Research directions

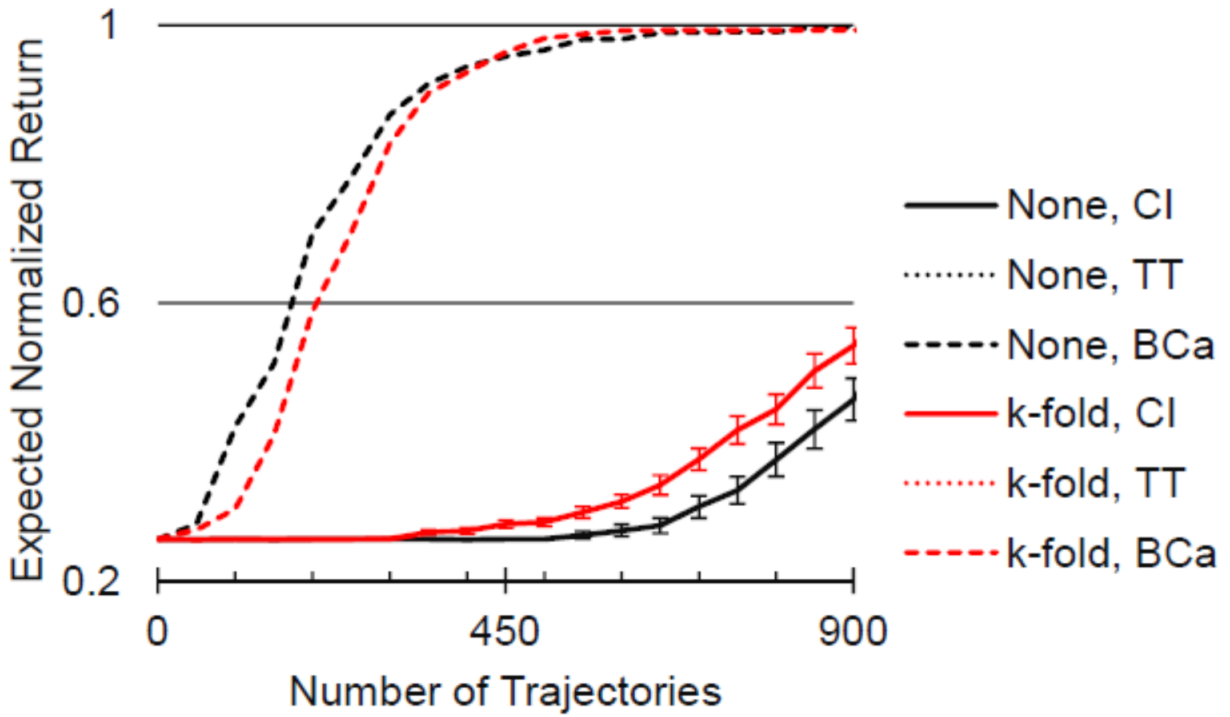
Empirical Results

- What to look for:
  - Data efficiency
  - Error rates

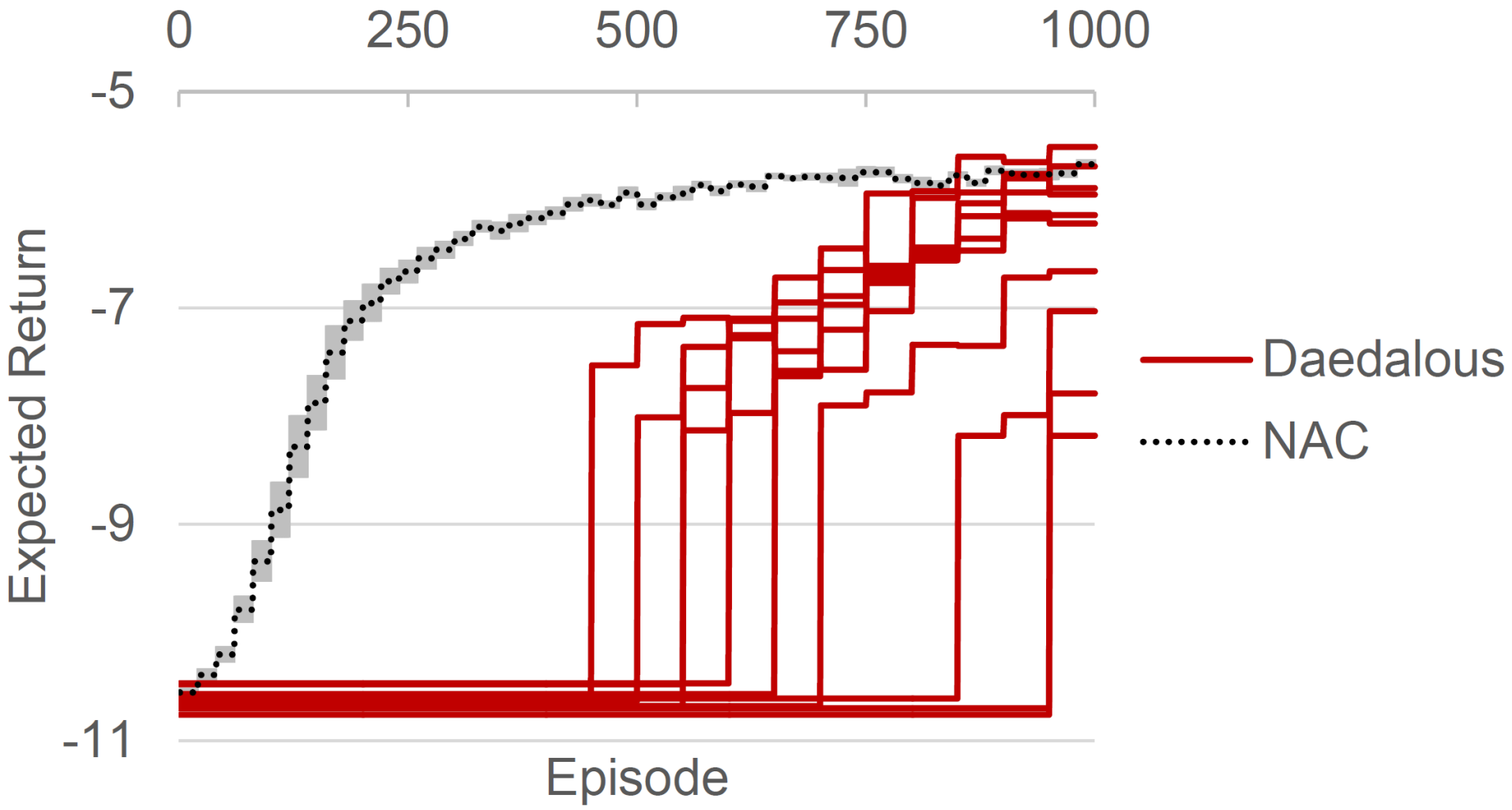
Empirical Results: Mountain Car



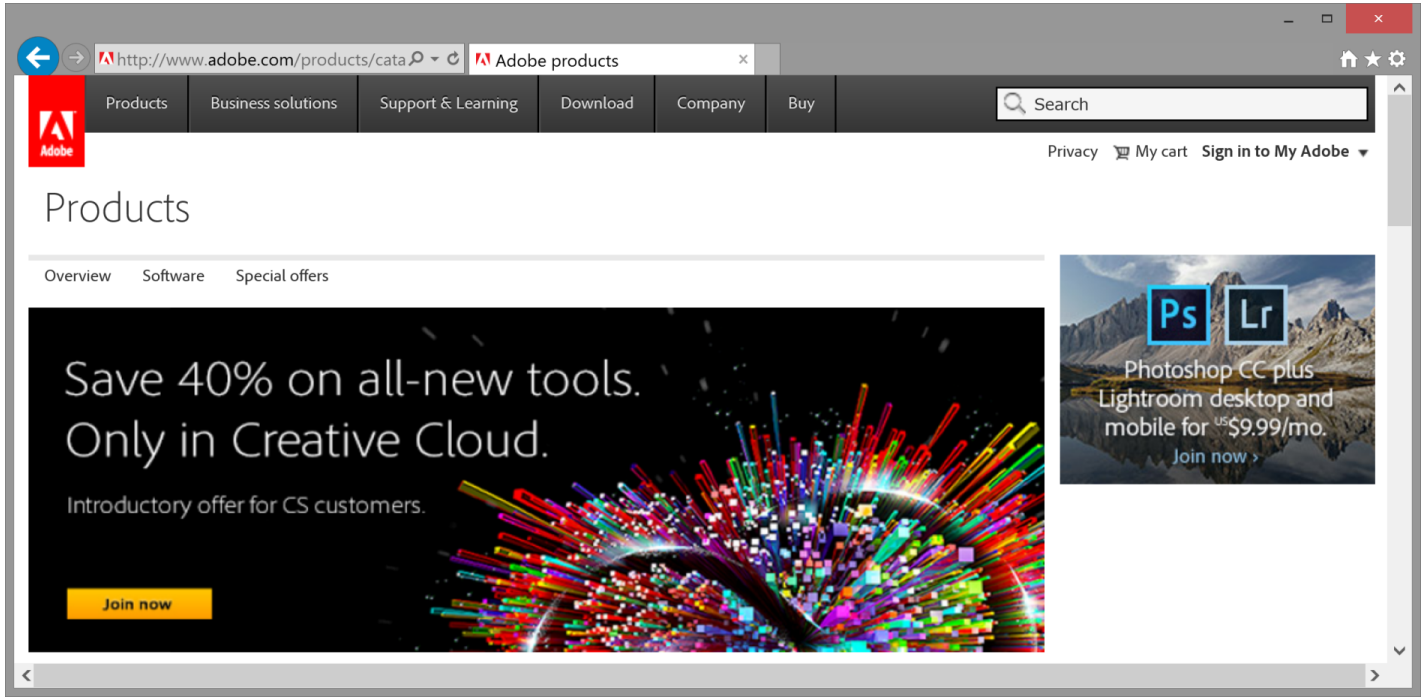
Empirical Results: Mountain Car

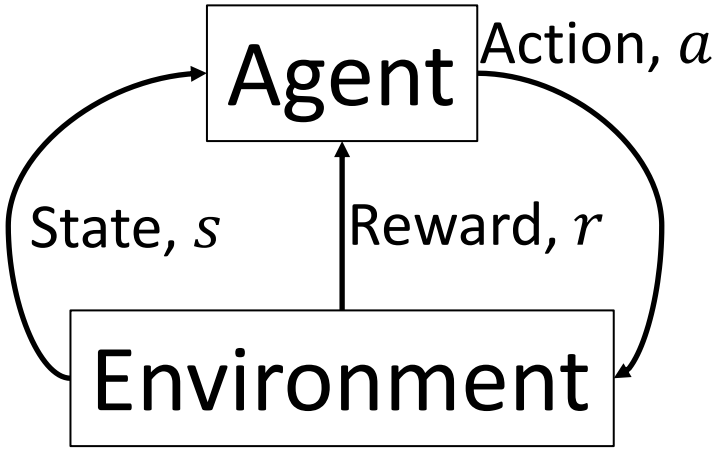


Empirical Results: Mountain Car

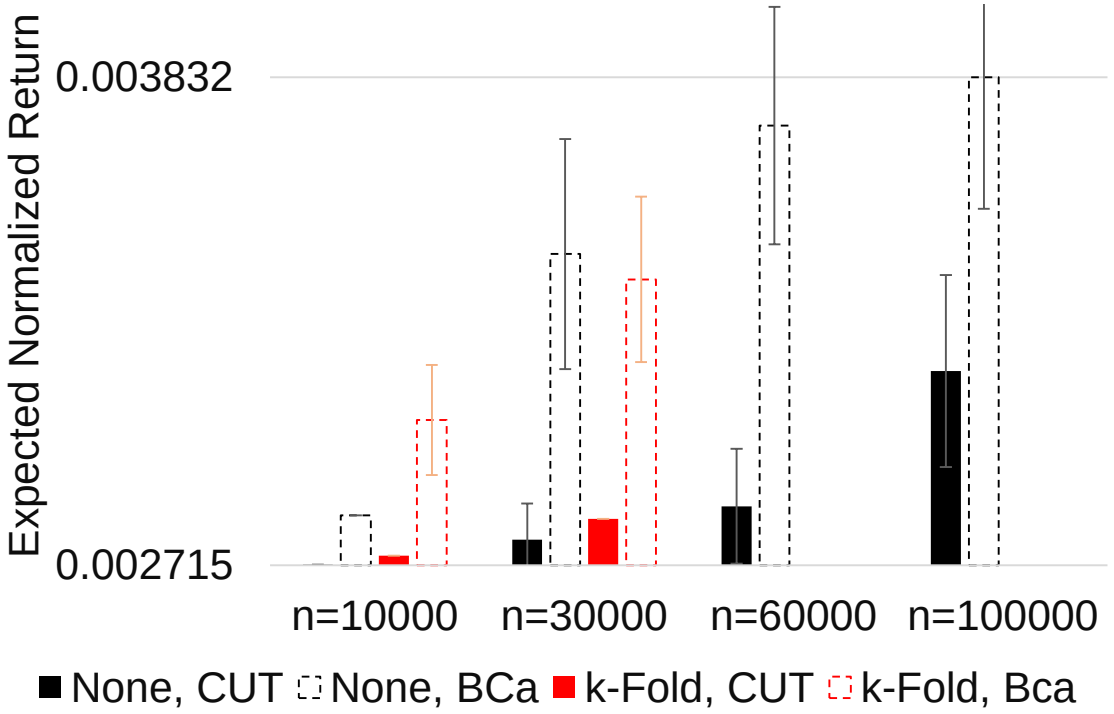


Empirical Results: Digital Marketing

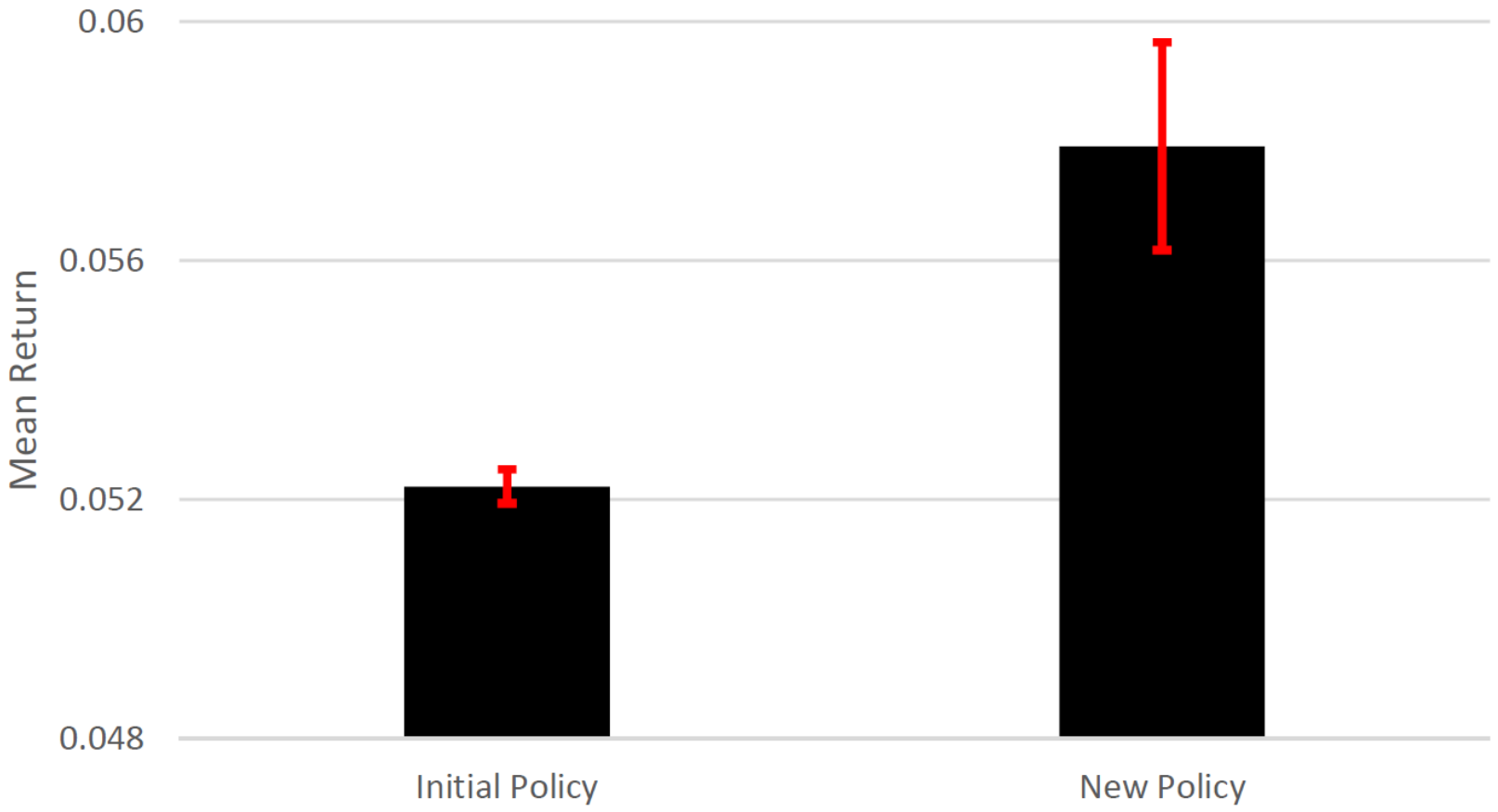




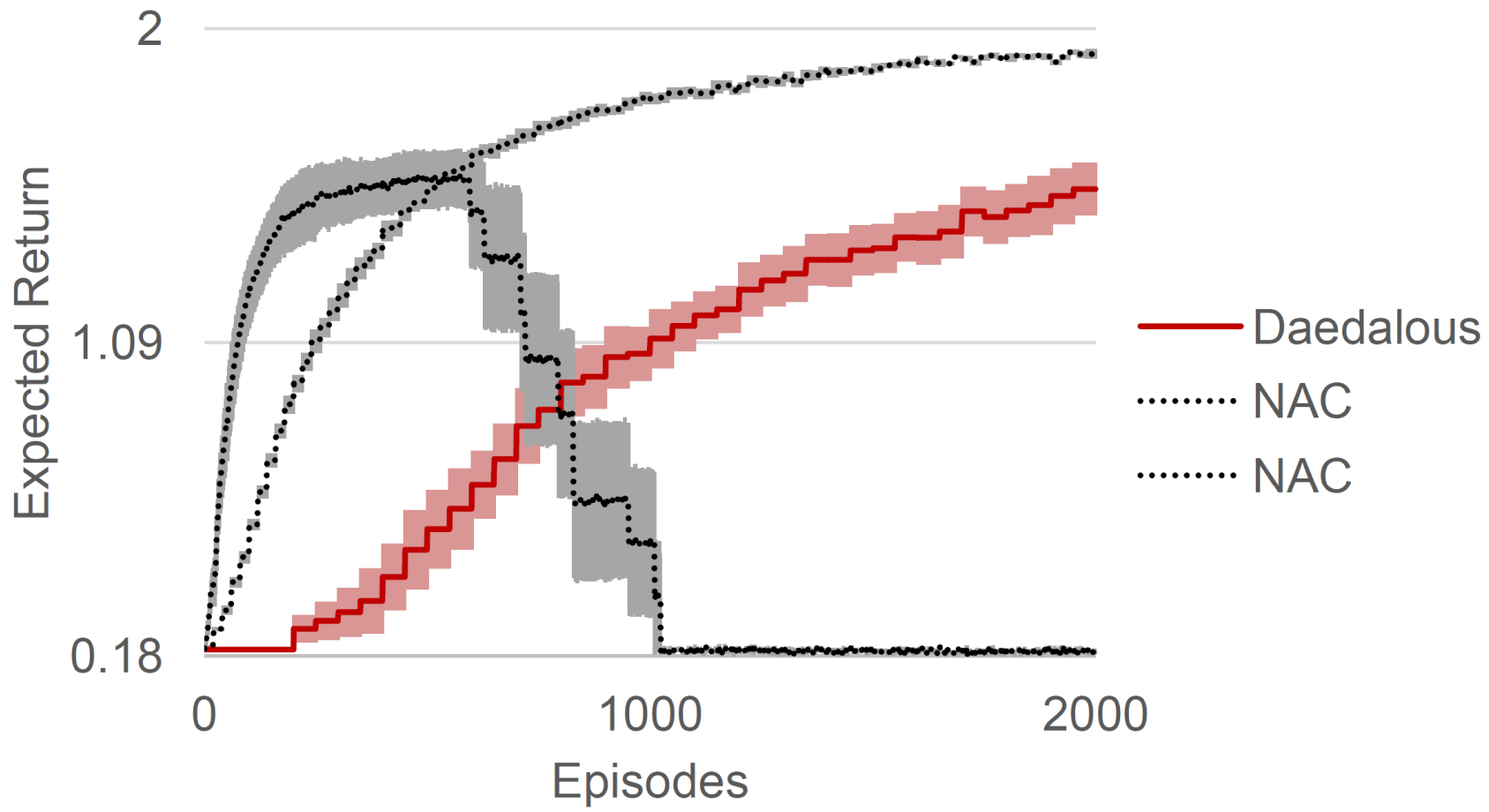
Empirical Results: Digital Marketing



Empirical Results: Digital Marketing

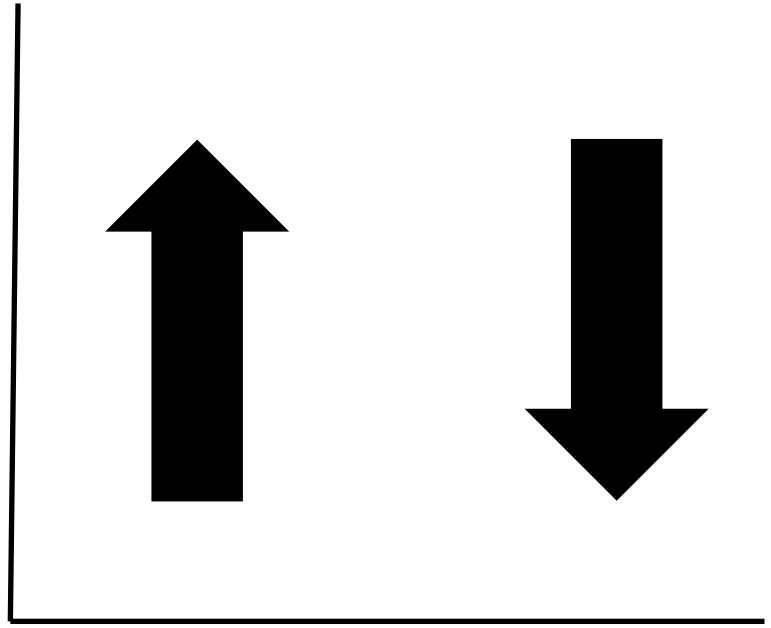


Empirical Results: Digital Marketing

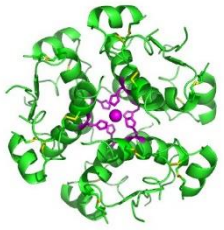


Example Results : Diabetes Treatment

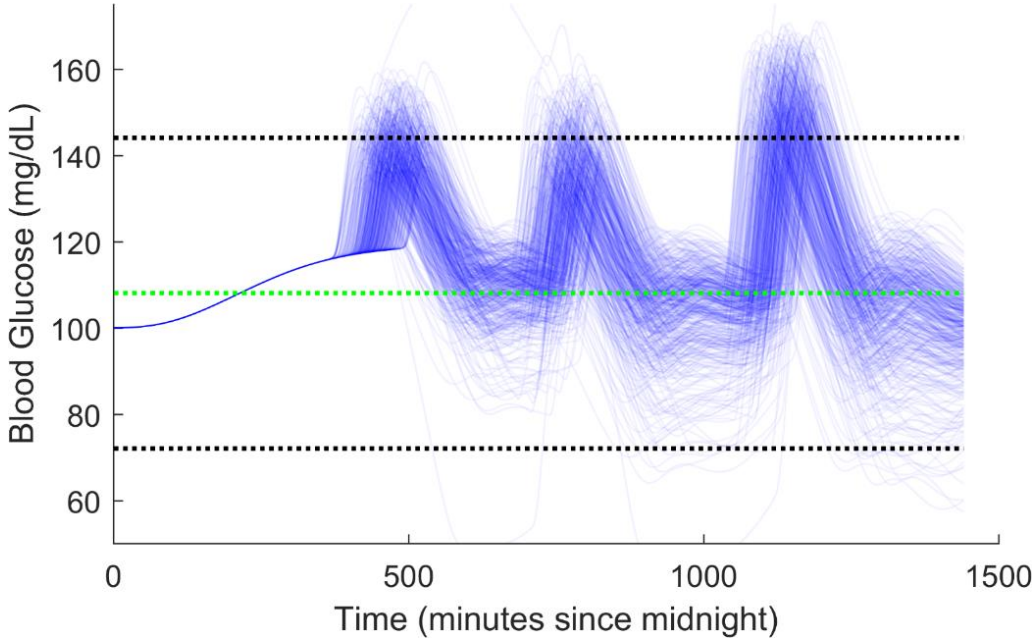
Blood Glucose
(sugar)



Eat Carbohydrates

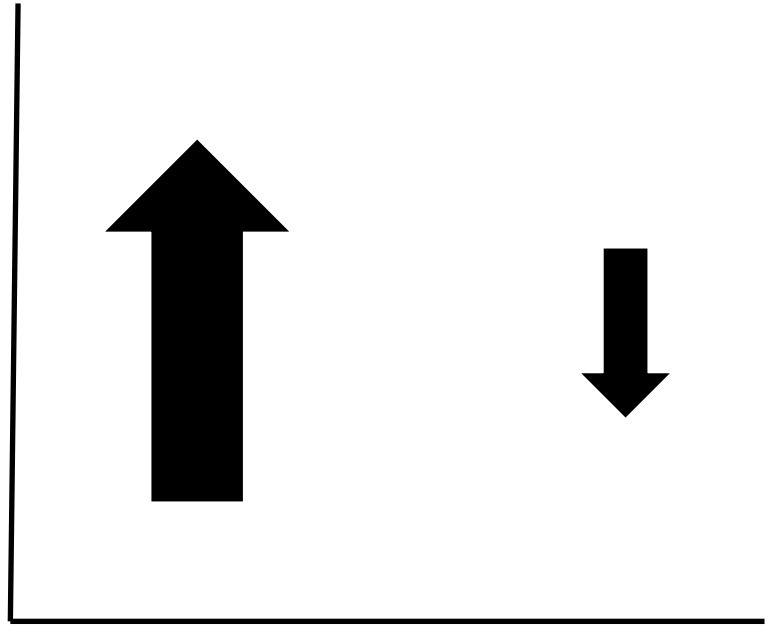


Release Insulin

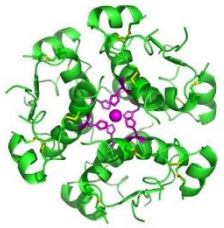


Example Results : Diabetes Treatment

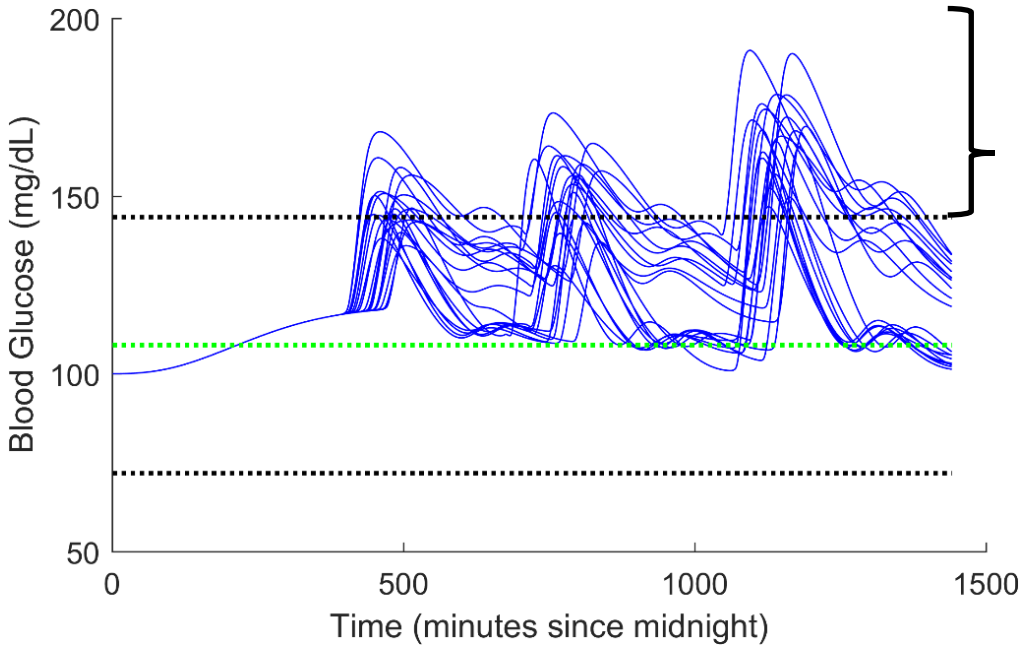
Blood Glucose (sugar)



Eat Carbohydrates



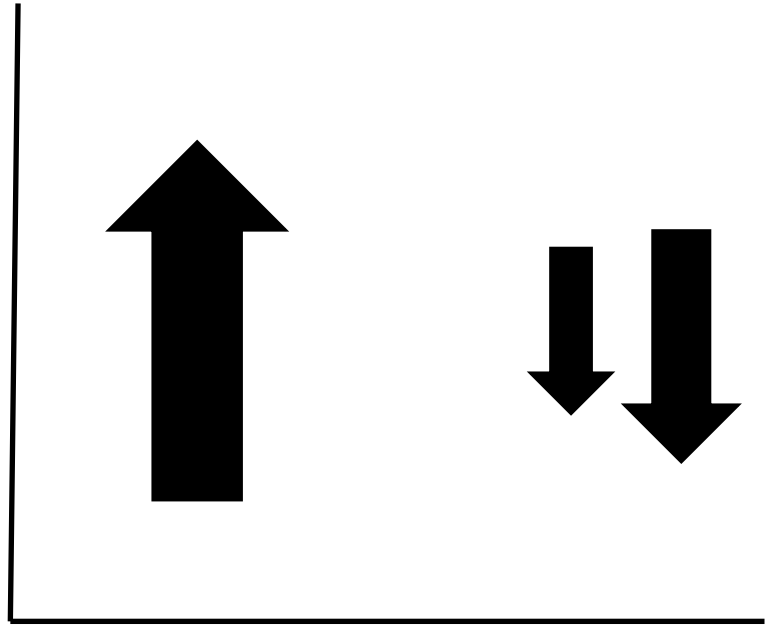
Release Insulin



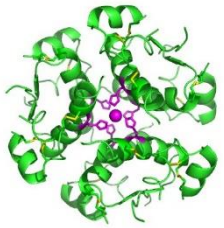
Hyperglycemia

Example Results : Diabetes Treatment

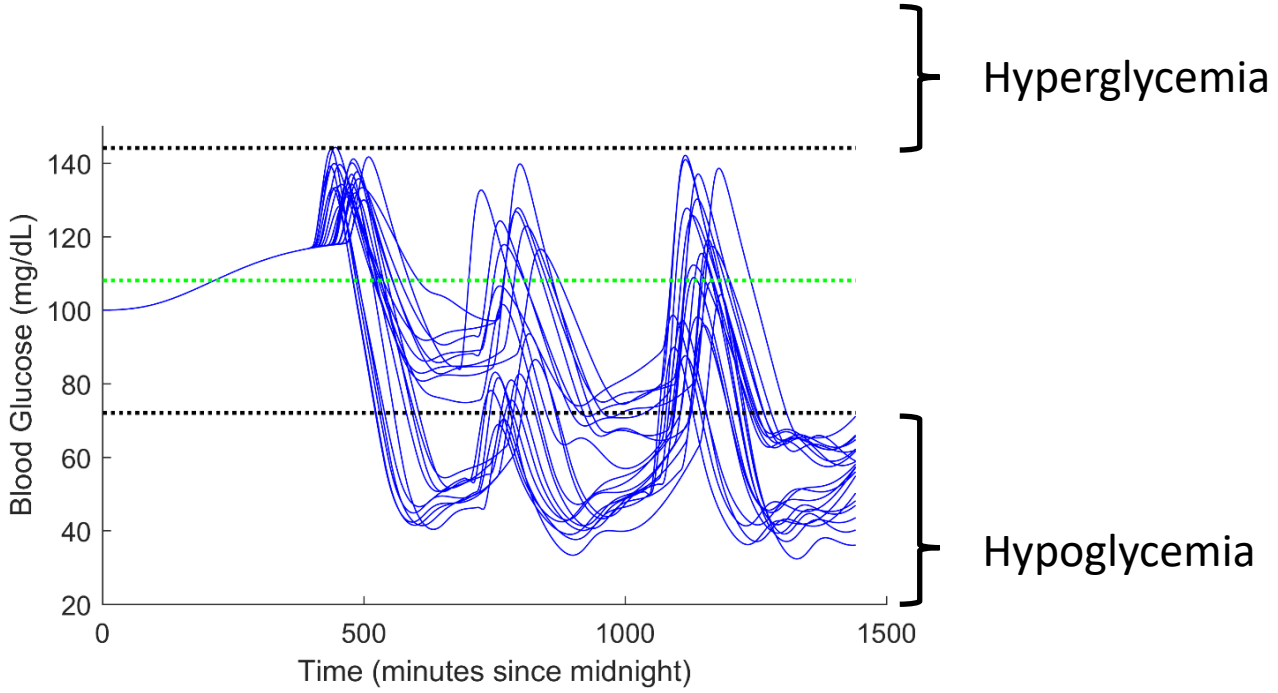
Blood Glucose (sugar)



Eat Carbohydrates



Release Insulin



Example Results : Diabetes Treatment

$$\text{injection} = \frac{\text{blood glucose} - \text{target blood glucose}}{CF} + \frac{\text{meal size}}{CR}$$

January

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

February

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28				

March

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

April

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

May

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

June

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

Example Results : Diabetes Treatment

Alberta Diabetes INSTITUTE

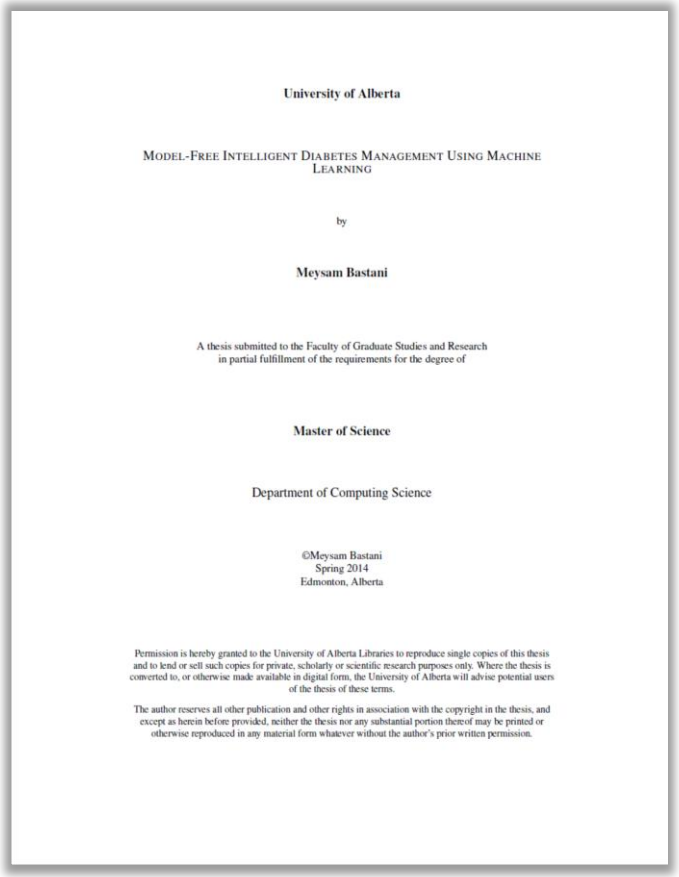
ALBERTA INNOVATES CENTRE FOR MACHINE LEARNING

UNIVERSITY OF ALBERTA

Intelligent Diabetes Management

T1DMS

Type 1 Diabetes Metabolic Simulator



University of Alberta

MODEL-FREE INTELLIGENT DIABETES MANAGEMENT USING MACHINE LEARNING

by

Meysam Bastani

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

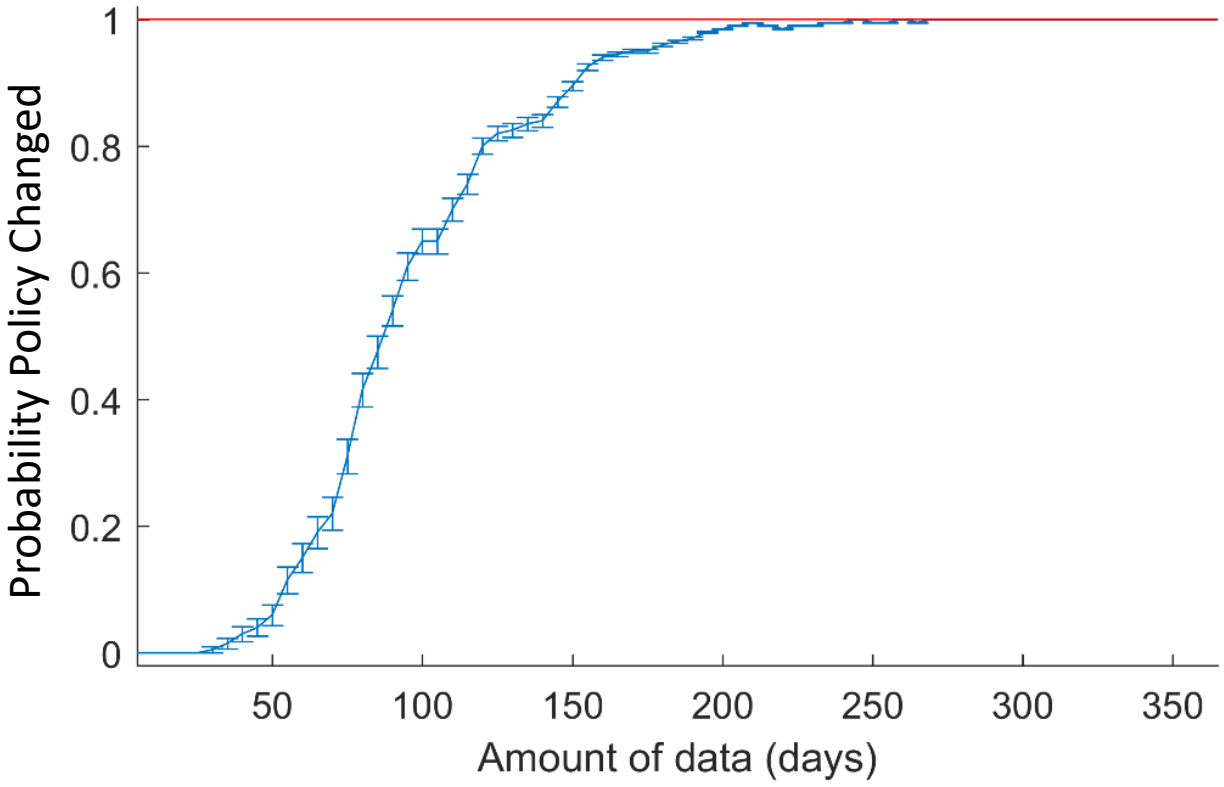
Department of Computing Science

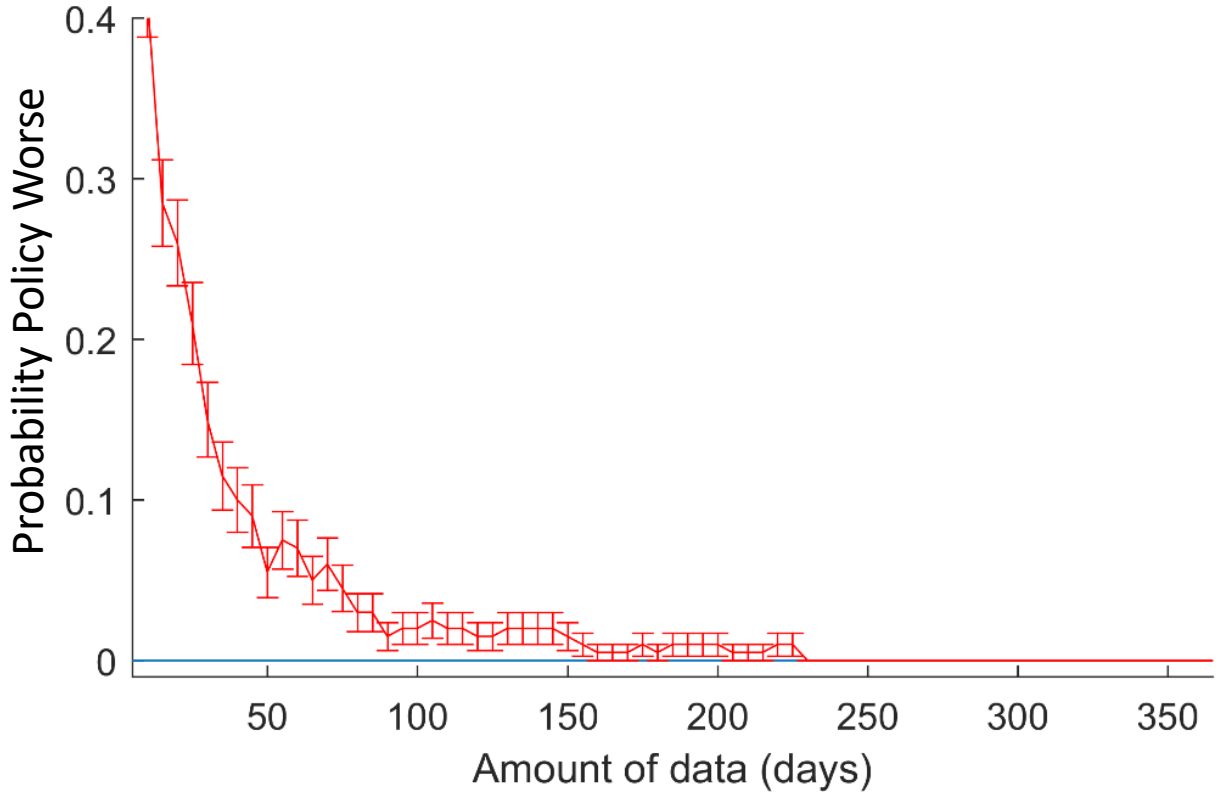
©Meysam Bastani
Spring 2014
Edmonton, Alberta

Permission is hereby granted to the University of Alberta Libraries to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only. Where the thesis is converted to, or otherwise made available in digital form, the University of Alberta will advise potential users of the thesis of these terms.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

Example Results : Diabetes Treatment





Future Directions

- How to deal with long horizons?
- How to deal with importance sampling being “unfair”?
- What to do when the behavior policy is not known?
- What to do when the behavior policy is deterministic?

Summary

- Safe reinforcement learning
  - Risk-sensitive
  - Learning from demonstration
  - Asymptotic convergence even if data is off-policy
  - **Guaranteed (with probability $1 - \delta$) not to make the policy worse**
- Designing a safe reinforcement learning algorithm:
  - Off-policy policy evaluation (OPE)
    - IS, PDIS, WIS, WPDIS, DR, WDR, US, TSP, MAGIC
  - High confidence off-policy policy evaluation (HCOPE)
    - Hoeffding, CUT inequality, Student's t-test, BCa
  - Safe policy improvement (SPI)
    - Selecting the candidate policy

Takeaway

- Safe reinforcement learning is tractable!
  - Not just polynomial amounts of data – practical amounts of data