Safe Reinforcement Learning

Philip S. Thomas

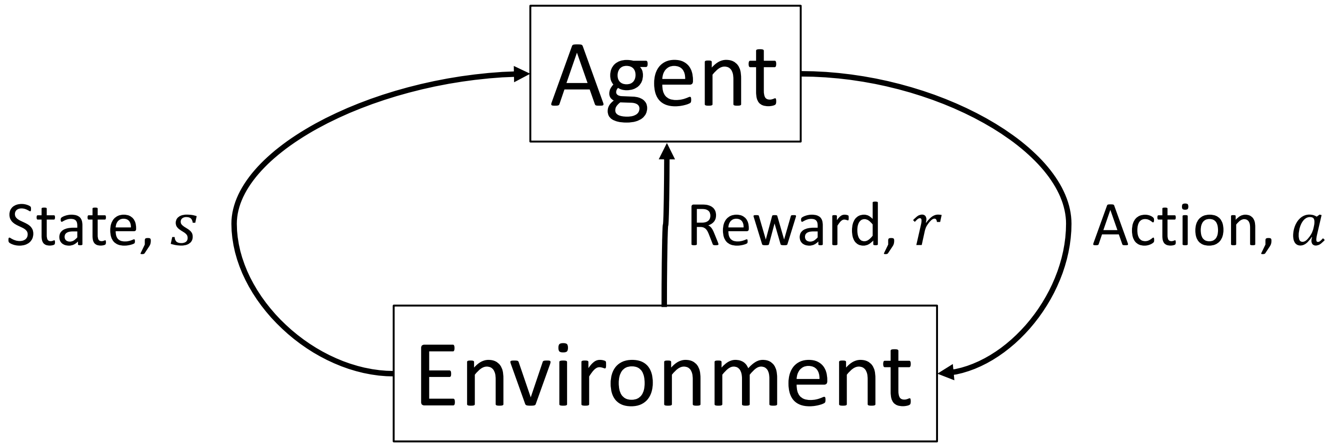
Carnegie Mellon University

Reinforcement Learning Summer School 2017

Overview

- Background and motivation
- Definition of “safe”
- Three steps towards a safe algorithm
  - Off-policy policy evaluation
  - High-confidence off-policy policy evaluation
  - Safe policy improvement
- Experimental results
- Conclusion

Background



Policy: Decision rule $s \rightarrow a$

Notation

- Policy, π

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

- History:

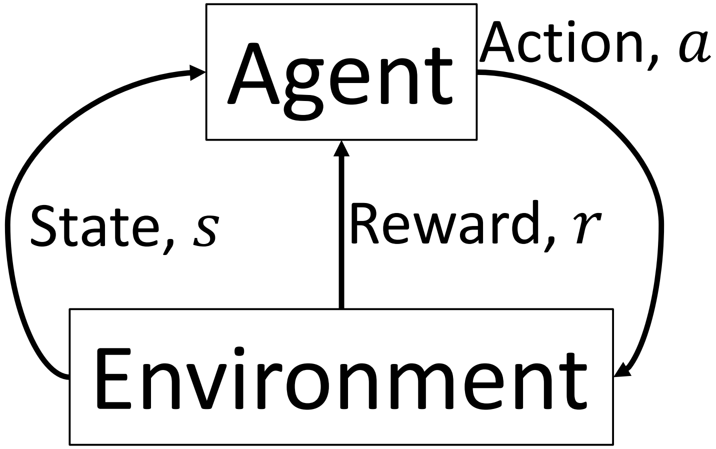
$$H = (S_1, A_1, R_1, S_2, A_2, R_2, \dots, S_L, A_L, R_L)$$

- Historical data:

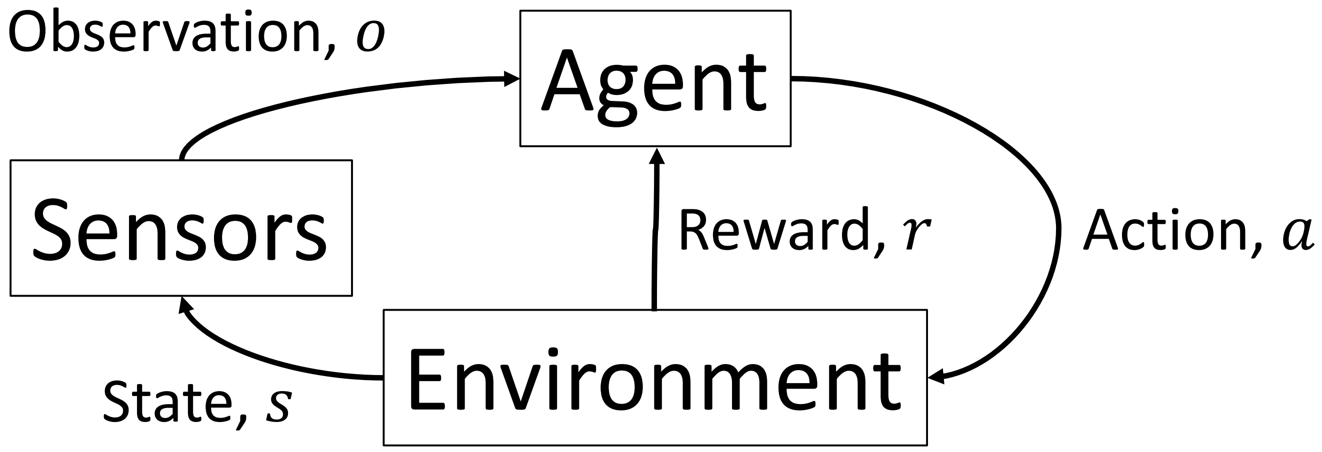
$$D = \{H_1, H_2, \dots, H_n\}$$

- Historical data from *behavior policy*, π_b
- Objective:

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

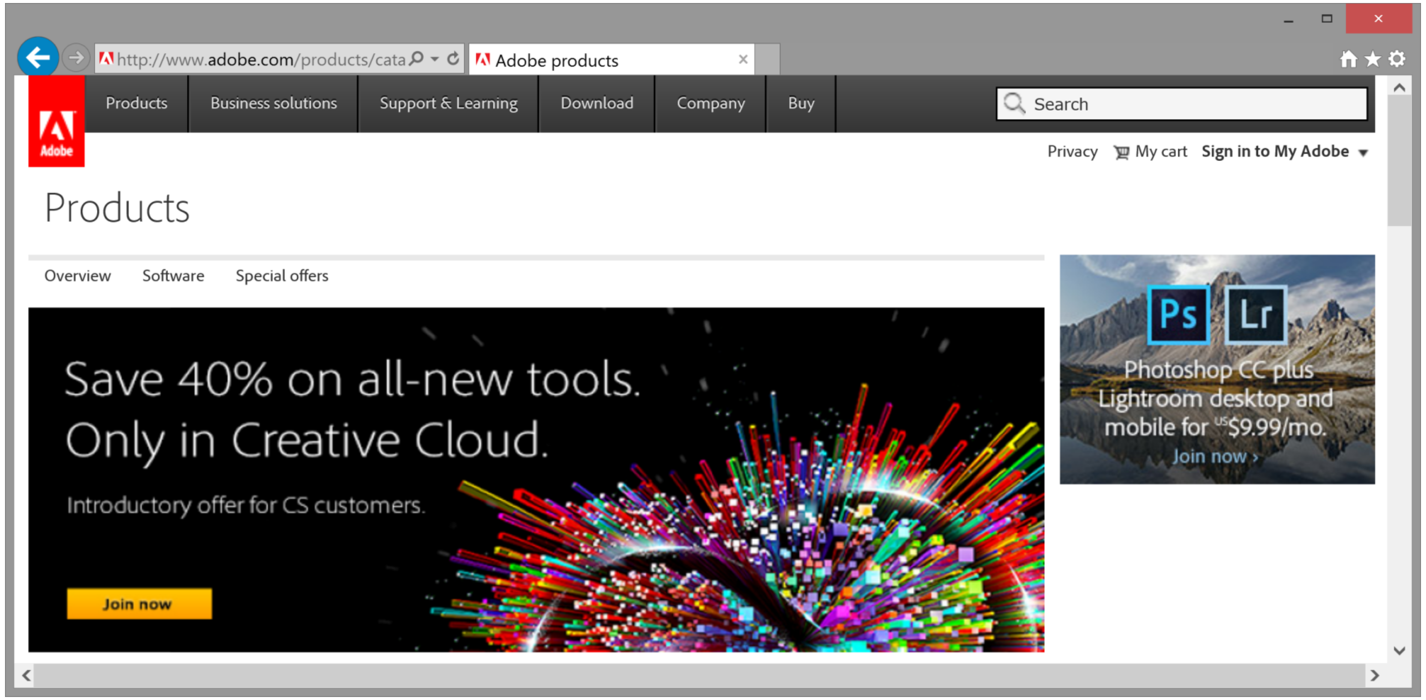


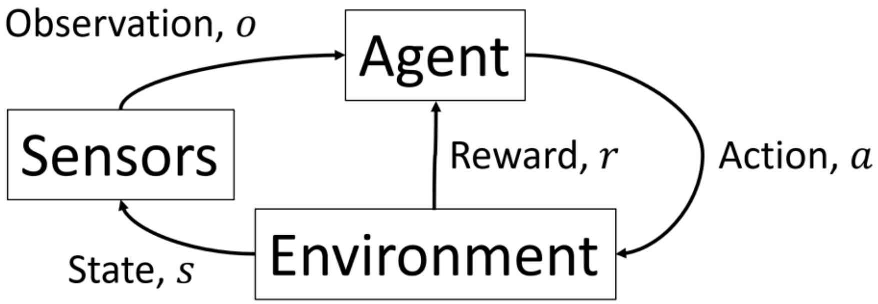
Background



Policy: Decision rule $s \rightarrow a$

Potential Application: Digital Marketing





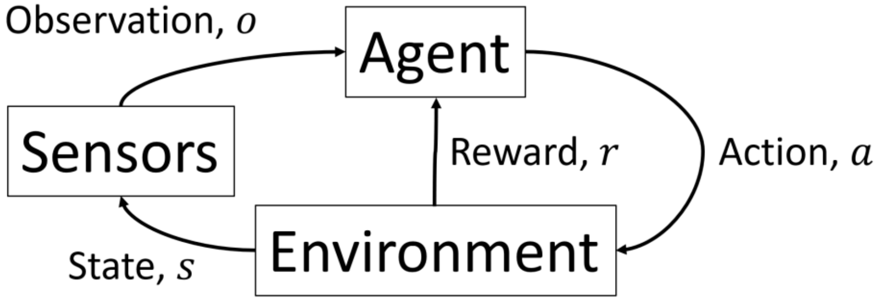
- History:

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

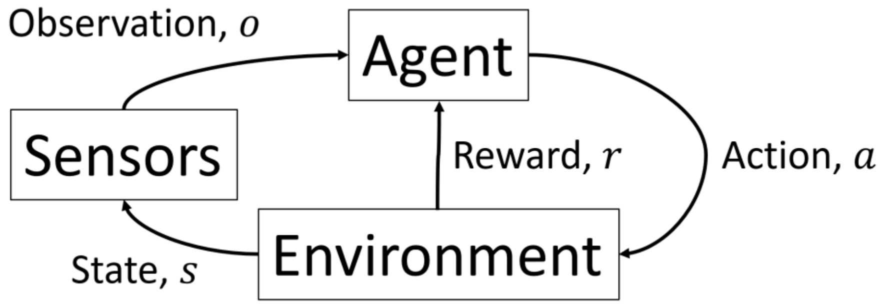
- Historical data:

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

Potential Application: Intelligent Tutoring Systems

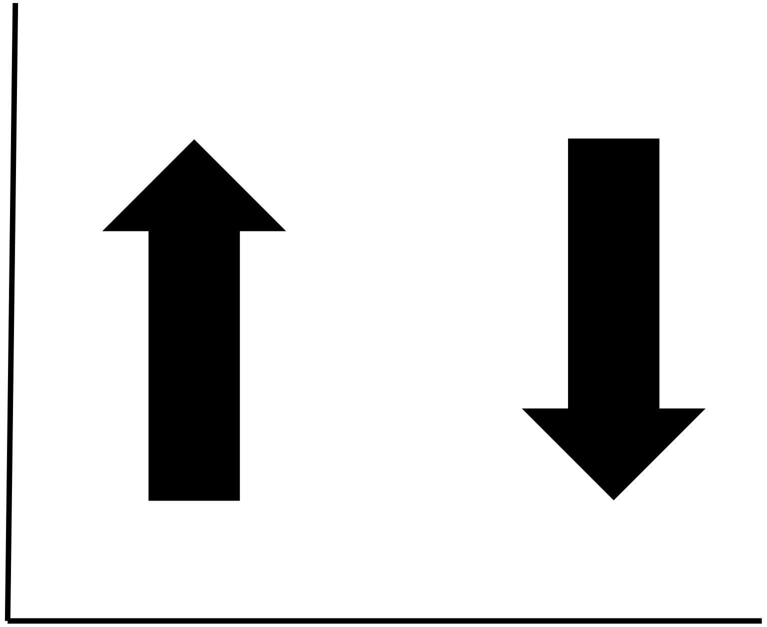


Potential Application: Functional Electrical Stimulation

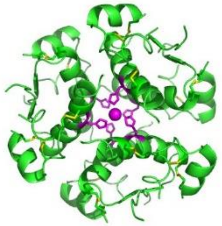


Potential Application: Diabetes Treatment

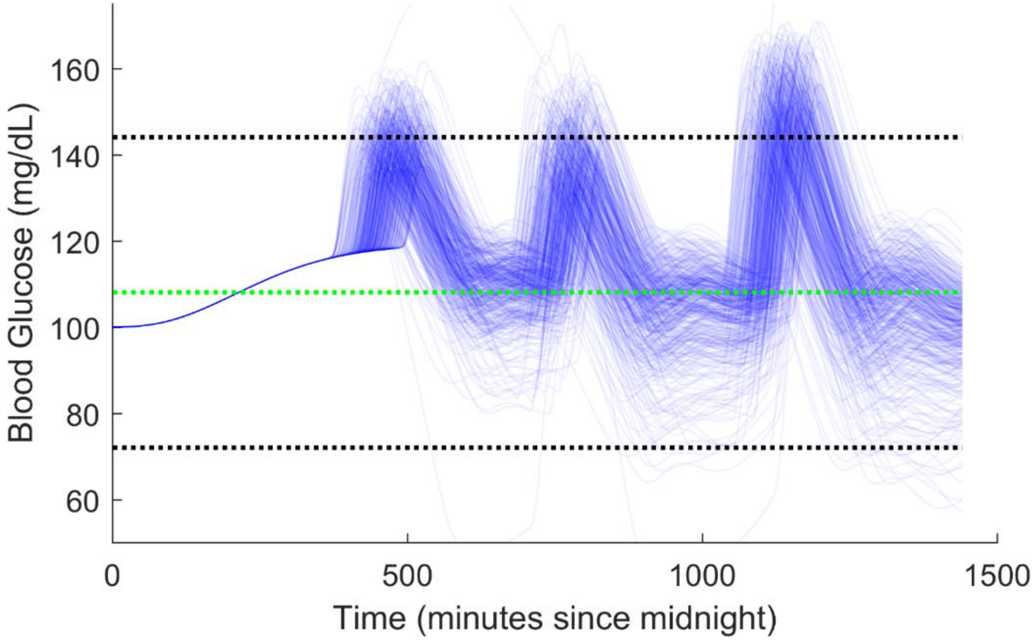
Blood Glucose (sugar)



Eat Carbohydrates

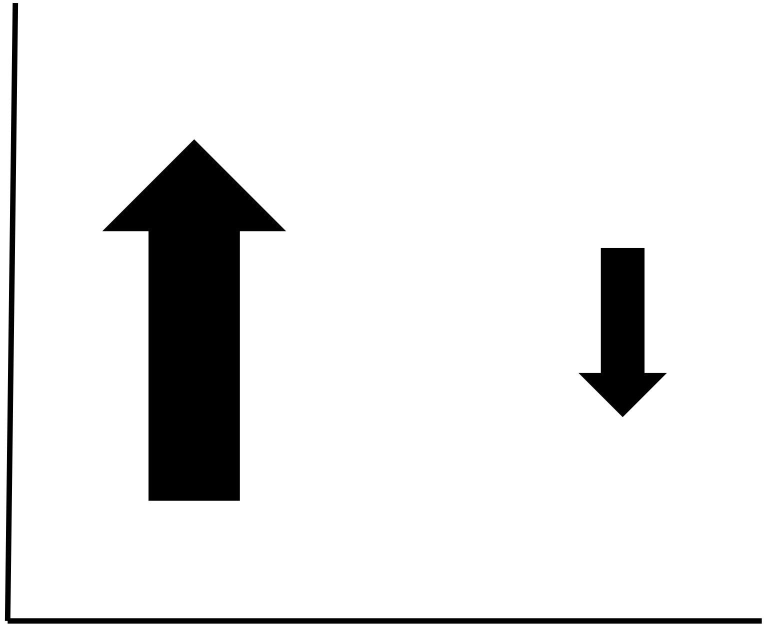


Release Insulin

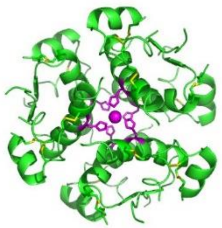


Potential Application: Diabetes Treatment

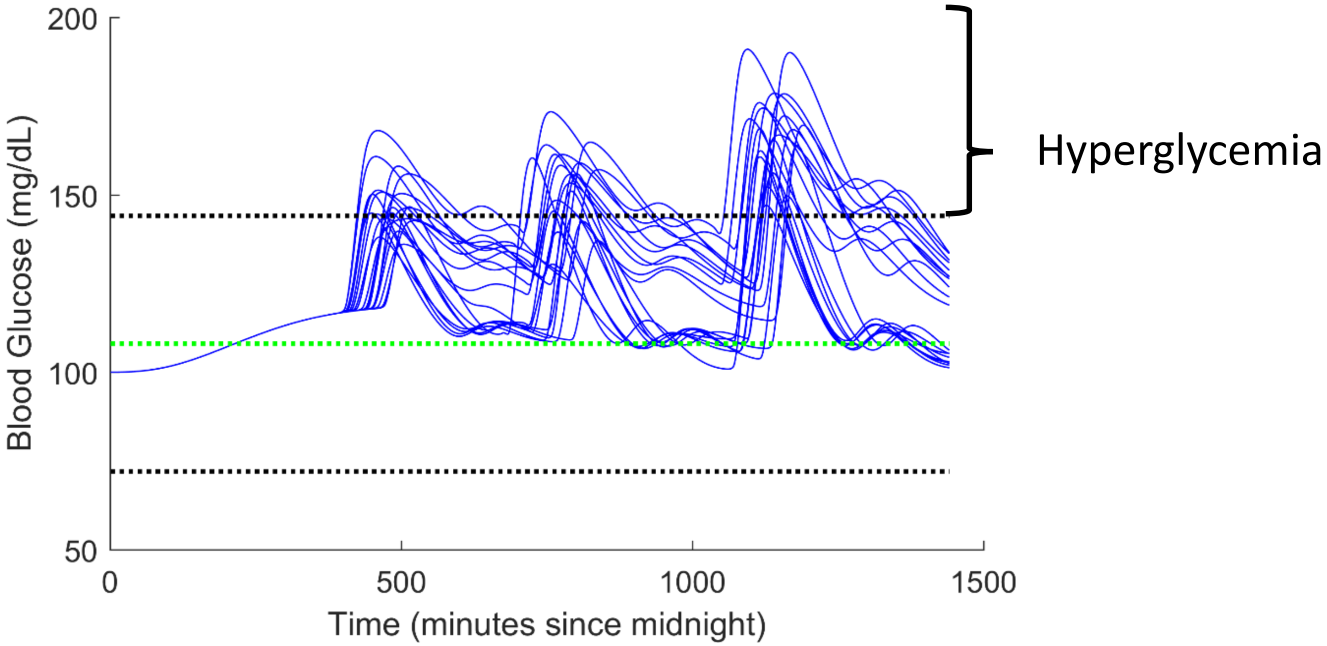
Blood Glucose
(sugar)



Eat Carbohydrates

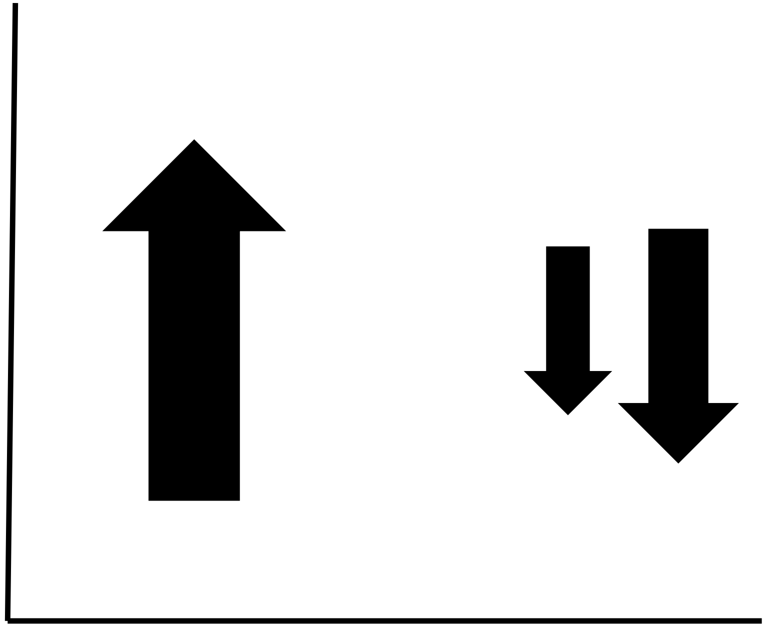


Release Insulin

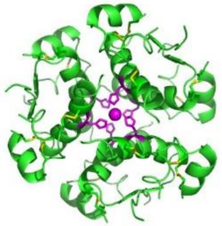


Potential Application: Diabetes Treatment

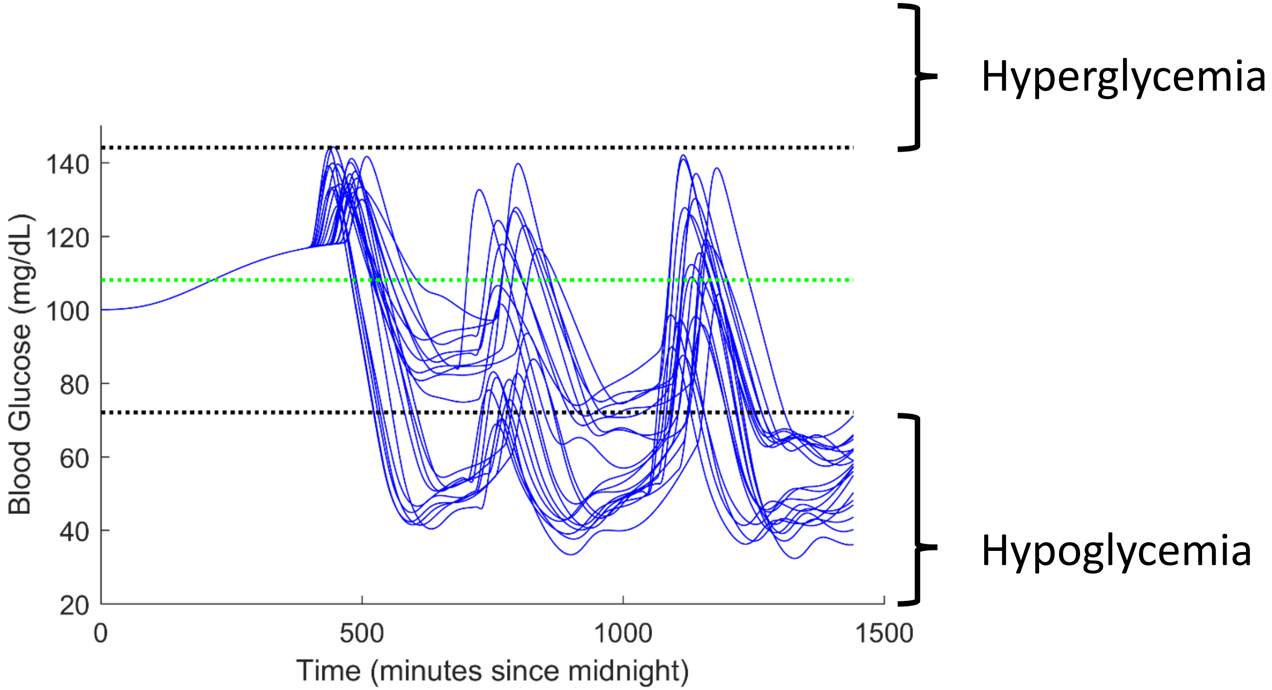
Blood Glucose (sugar)



Eat Carbohydrates

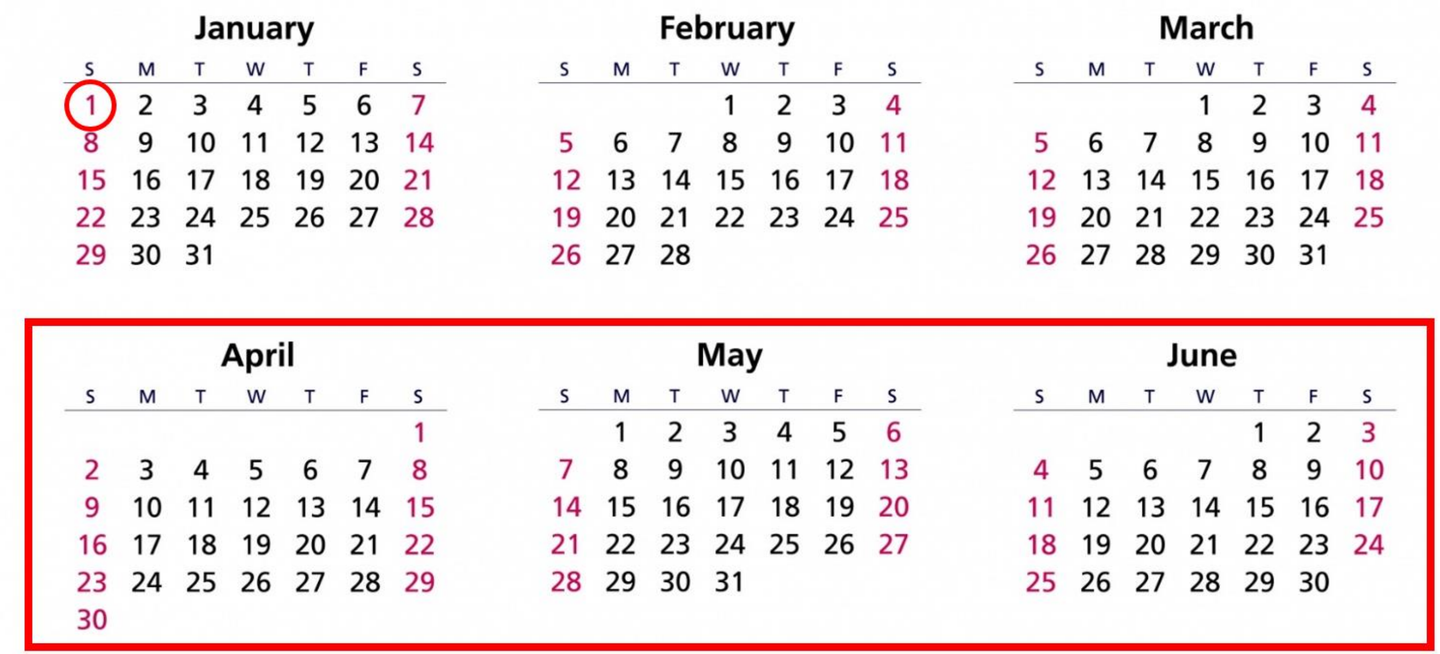


Release Insulin



Potential Application: Diabetes Treatment

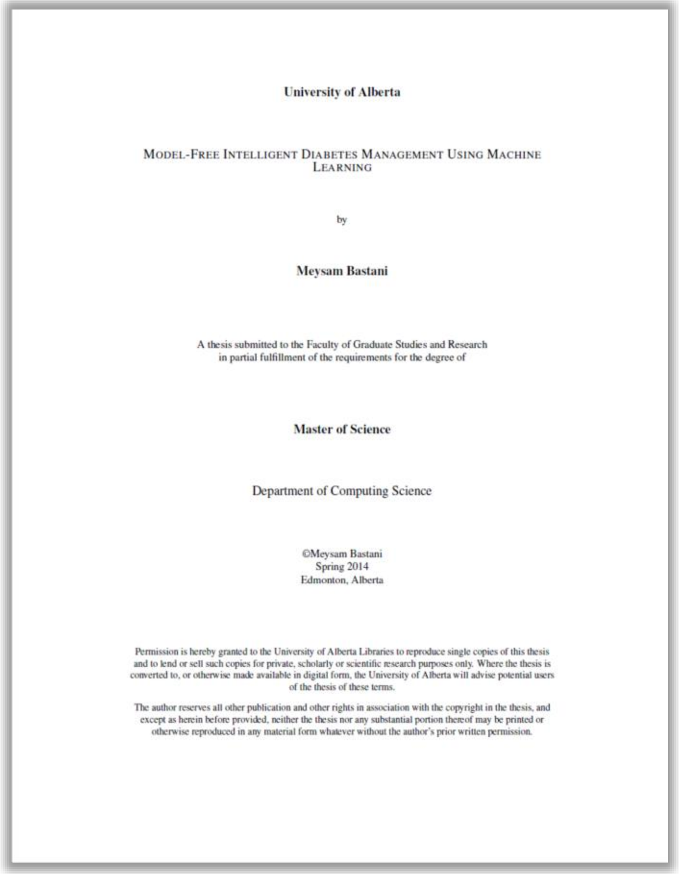
$$\text{injection} = \frac{\text{blood glucose} - \text{target blood glucose}}{CF} + \frac{\text{meal size}}{CR}$$



Potential Application: Diabetes Treatment

Intelligent Diabetes Management

Alberta Diabetes Institute
AICML Alberta Innovates Centre for Machine Learning
University of Alberta



University of Alberta

Model-Free Intelligent Diabetes Management Using Machine Learning

by

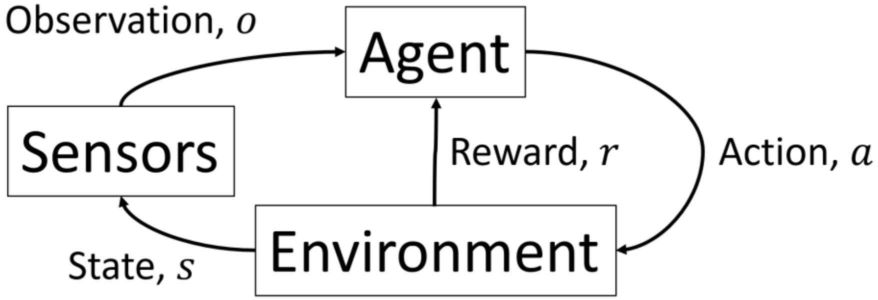
Meysam Bastani

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

Master of Science

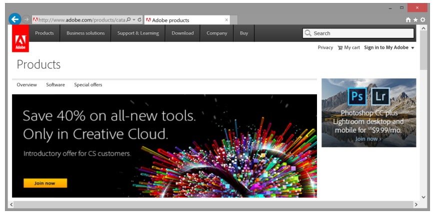
Department of Computing Science

©Meysam Bastani
Spring 2014
Edmonton, Alberta



Motivation for Safe Reinforcement Learning

- If you deploy an existing reinforcement learning algorithm to one of these problems, do you have confidence that the policy that it produces will be better than the current policy?



vs.

Learning Curves are Deceptive

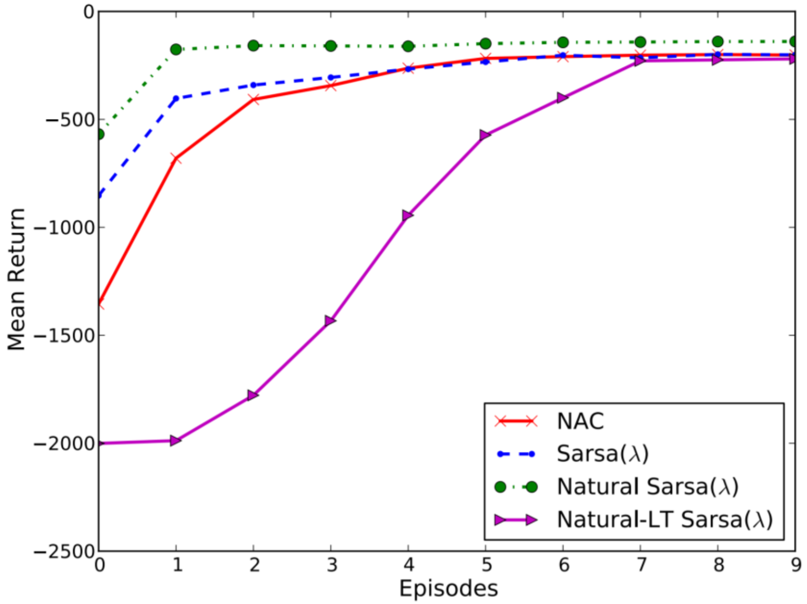


Figure 3: Mountain Car (Sarsa(λ))

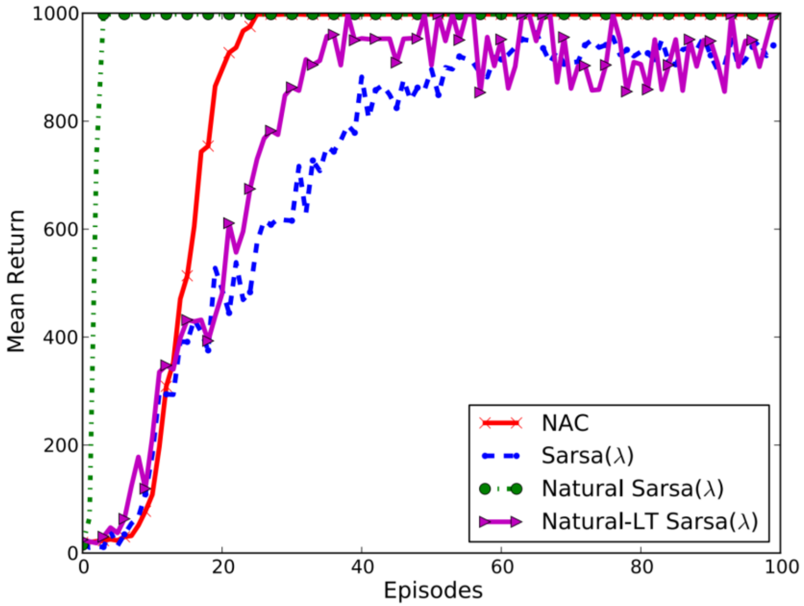


Figure 5: Cart Pole (Sarsa(λ))

- … after *billions* of episodes
  - Millions (billions?) of episodes of parameter optimization
  - Human intuition from past experience with these domains
  - Billions of episodes of experimental design

Overview

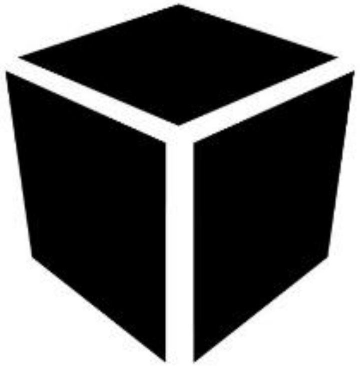
- Background and motivation
- Definition of “safe”
- Three steps towards a safe algorithm
  - Off-policy policy evaluation
  - High-confidence off-policy policy evaluation
  - Safe policy improvement
- Experimental results
- Conclusion

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

Historical Data, D

Probability, $1-\delta$



New policy π, or

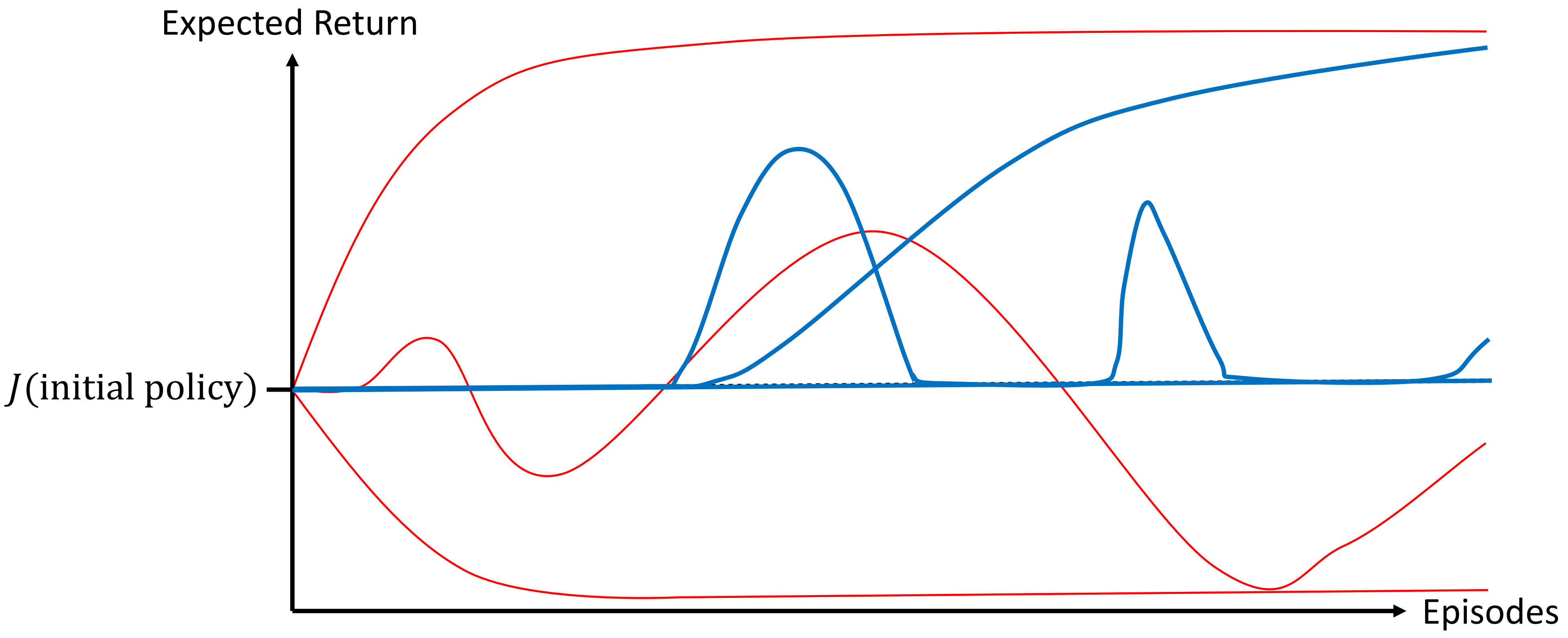
`No Solution Found`

$$\Pr\big(J(\pi) \geq J(\pi_b)\big) \geq 1-\delta$$

Limitations of the Safe RL Setting

- Assumes that an initial policy is available
- Often assumes that the initial policy is known
- Often assumes that the initial policy is stochastic
- Batch setting

Standard RL vs Safe RL



Standard

Safe: $\Pr\big(J(\pi) \geq J(\pi_b)\big) \geq 1 - \delta$

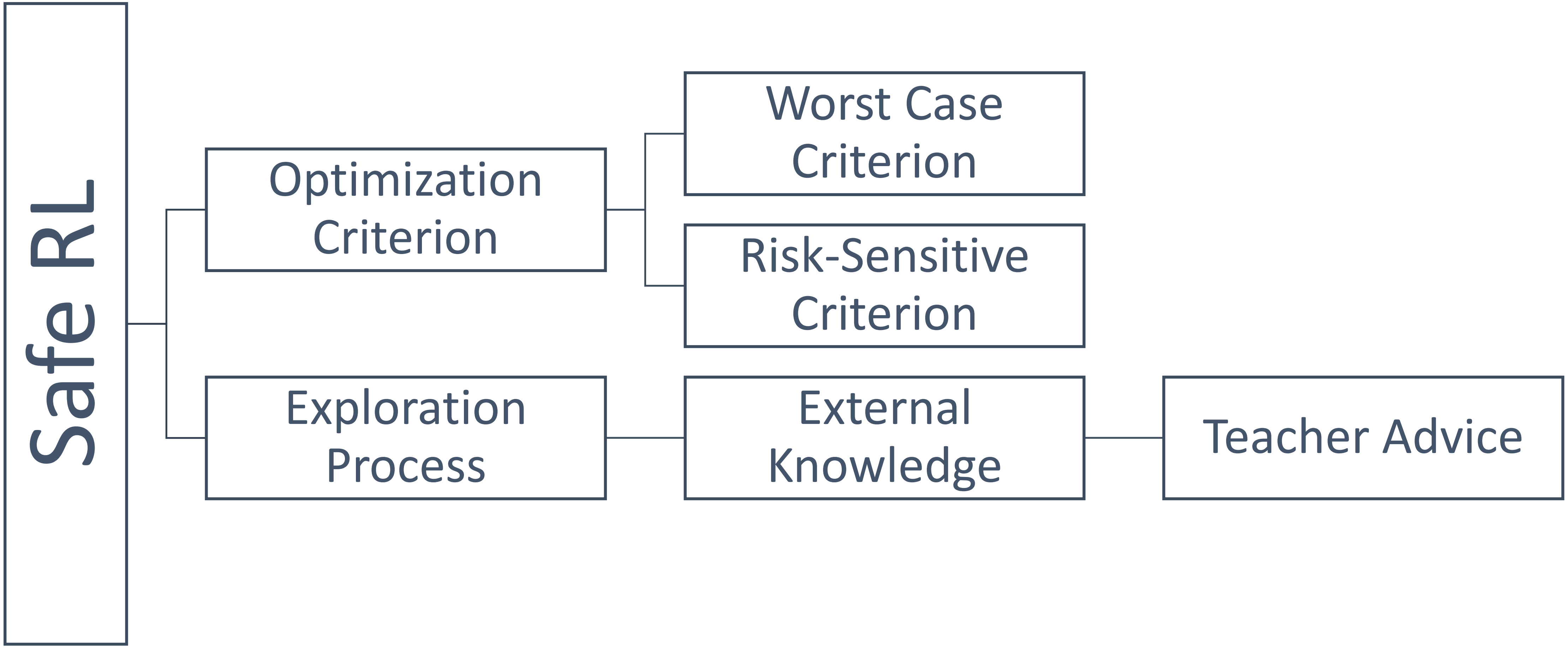
Other Definitions of “Safe”

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

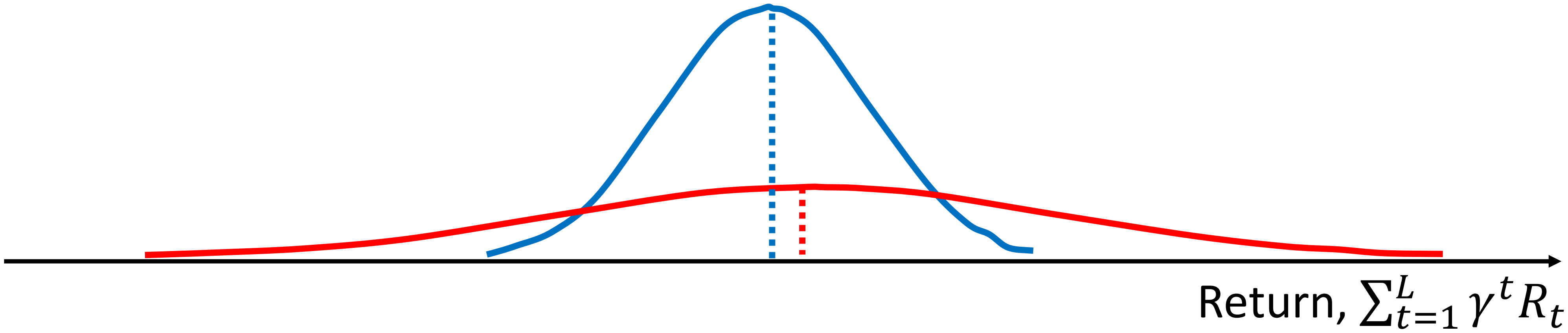
Other Definitions of “Safe”



Risk-Sensitive Criterion

- Expected return:

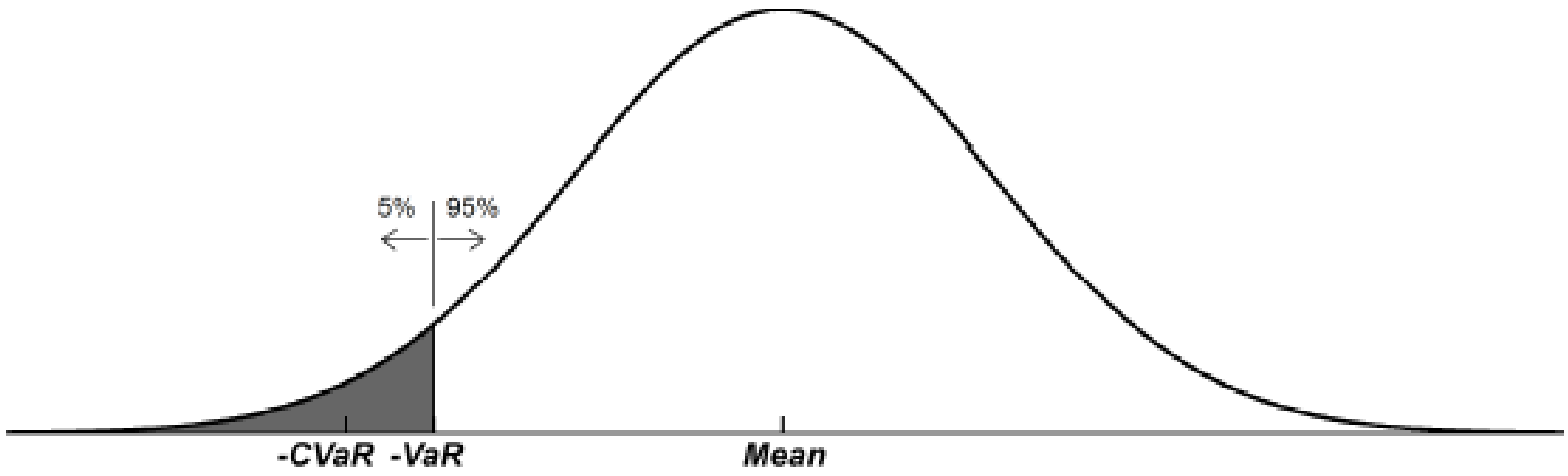
$$J(\pi) = \mathbf{E}\left[\sum_{t=1}^{L} \gamma^t R_t \,\middle|\, \pi\right]$$



- Which policy is better if I am a casino?
- Which policy is better if I am a doctor?

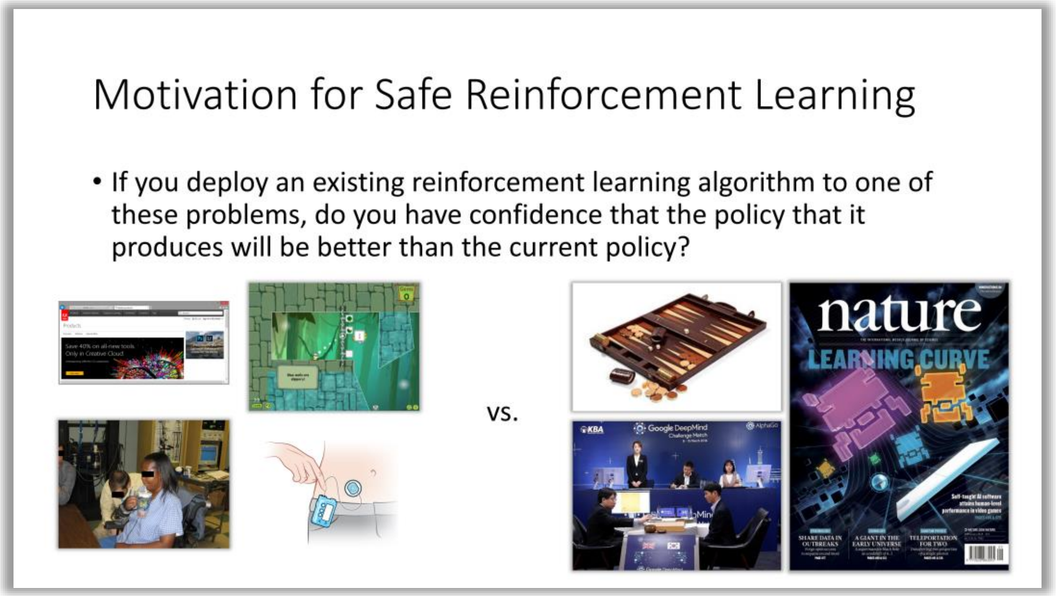
Risk-Sensitive Criterion

- Idea: Change our objective to minimize a notion of risk
  - Penalize variance: $J(\pi) = \mathbf{E}\left[\sum_{t=1}^{L} \gamma^t R_t \,\middle|\, \pi\right] - \lambda \mathrm{Var}\left(\sum_{t=1}^{L} \gamma^t R_t \,\middle|\, \pi\right)$
  - Maximize *Value at Risk* (VaR), *Conditional Value at Risk* (CVaR), or another *robust* objective



Benefits and Limitations of Changing Objectives

- For some applications a risk-sensitive objective is more appropriate
- Changing the objective does not address our motivation



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 Definition of Safety

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 Definition of Safety

Reachability-Based Safe Learning with Gaussian Processes

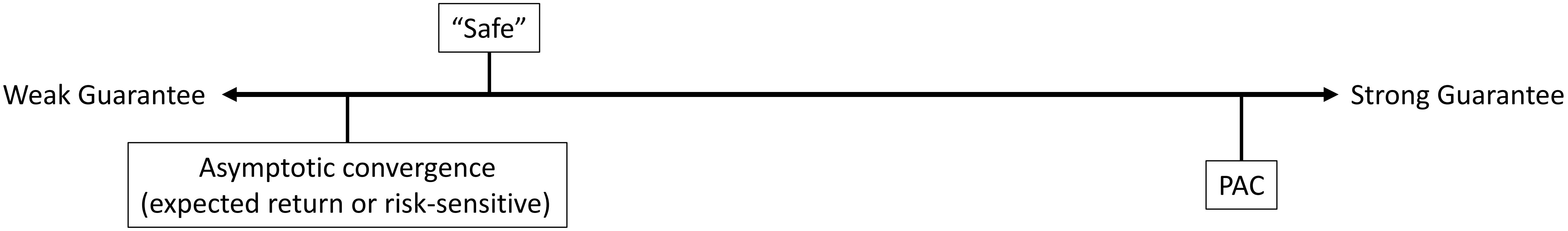
Anayo K. Akametalu*
Shahab Kaynama

Jaime F. Fisac*
Melanie N. Zeilinger

Jeremy H. Gillula
Claire J. Tomlin

Another Definition of Safety

- Probably Approximately Correct (PAC) RL
  - Guarantee that with probability at least $1 - \delta$ the policy (or q-function) will be within ϵ of optimal after n episodes
    - Typically an equation is given for n in terms of the number of states and actions, the horizon, L, and both ϵ and δ



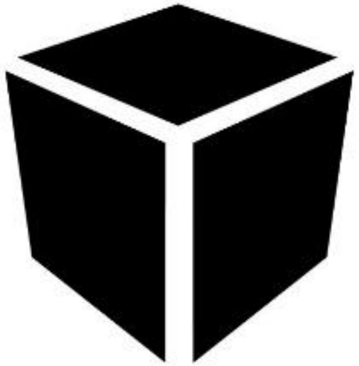
Overview

- Background and motivation
- Definition of “safe”
- Three steps towards a safe algorithm
  - Off-policy policy evaluation
  - High-confidence off-policy policy evaluation
  - Safe policy improvement
- Experimental results
- Conclusion

Off-Policy Policy Evaluation (OPE)

- Given the historical data, D, produced by a *behavior policy*, π_b
- Given a new policy, which we call the *evaluation policy*, π_e
- Predict the performance, $J(\pi_e)$, of the evaluation policy
- Do not deploy π_e since doing so could be costly or dangerous

Historical Data, D
Proposed Policy, π_e



Estimate of $J(\pi_e)$

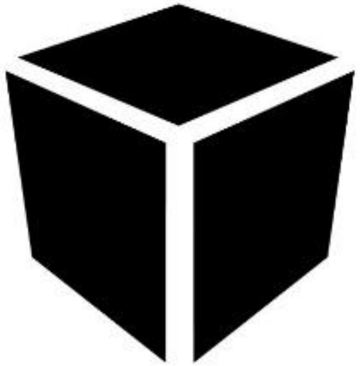
High Confidence Off-Policy Policy Evaluation (HCOPE)

- Given the historical data, D, produced by the behavior policy, π_b
- Given a new policy, which we call the *evaluation policy*, π_e
- Given a probability, $1 - \delta$
- Lower bound the performance, $J(\pi_e)$, of the evaluation policy with probability $1 - \delta$
- Do not deploy π_e since doing so could be costly or dangerous

Historical Data, D

Proposed Policy, π_e

Probability, $1 - \delta$



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

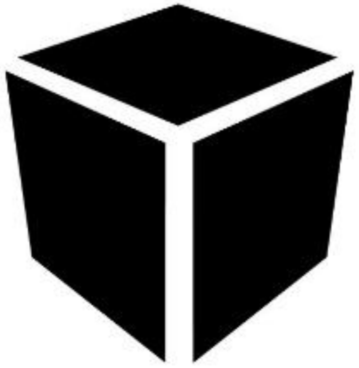
Safe Policy Improvement (SPI)

- Given the historical data, D, produced by the behavior policy, π_b
- Given a probability, $1 - \delta$
- Produce a policy, π, that we predict maximizes $J(\pi)$ and which satisfies:

$$\Pr\big(J(\pi) \geq J(\pi_b)\big) \geq 1 - \delta$$

Historical Data, D

Probability, $1 - \delta$



New policy π, or
`No Solution Found`

Overview

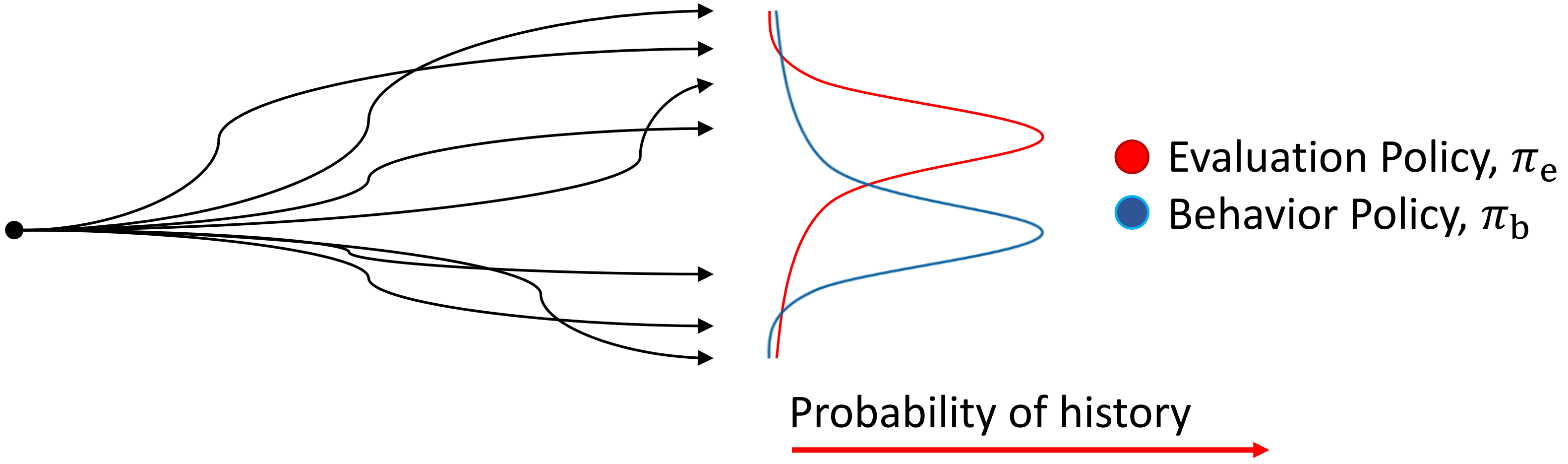
- Background and motivation
- Definition of “safe”
- Three steps towards a safe algorithm
    - Off-policy policy evaluation
    - High-confidence off-policy policy evaluation
    - Safe policy improvement
- Experimental results
- Conclusion

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

Importance weighted return

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



Importance Sampling (Derivation)

- Let X be a random variable with *probability mass function* (PMF) p
  - X is a history generated by the evaluation policy
- Let Y be a random variable with PMF q and the same range as X
  - Y is a history generated by the behavior policy
- Let f be a function
  - $f(X)$ is the return of the history X
- We want to estimate $\mathbf{E}[f(X)]$ given samples of Y
  - Estimate the expected return if trajectories are generated by the evaluation policy given trajectories generated by the behavior policy
- Let $P = \text{supp}(p)$, $Q = \text{supp}(q)$, and $F = \text{supp}(f)$

Importance Sampling (Derivation)

- Given one sample, Y, the importance sampling estimate of $\mathbf{E}_p[f(X)]$ is:

$$\mathrm{IS}(Y) = \frac{p(Y)}{q(Y)} f(Y)$$

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

Importance Sampling (Derivation)

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

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

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

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

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

Importance Sampling (Derivation)

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

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

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

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

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

Importance Sampling for Reinforcement Learning

- $X \leftarrow H$ produced by π_e
- $Y \leftarrow H$ produced by π_b
- $p \leftarrow \Pr(\cdot\,|\pi_e)$
- $q \leftarrow \Pr(\cdot\,|\pi_b)$
- $f(H) = \sum_{t=1}^{L} \gamma^t R_t$
- $\mathbf{E}[f(X)] \leftarrow J(\pi_e)$
- $\text{IS}(Y) = \frac{p(Y)}{q(Y)} f(Y)$
- Assume either:
  - Support of π_e is a subset of the support of π_b
  - Returns are non-negative

- Importance sampling estimator from one history, $H \sim \pi_b$:

$$\text{IS}(H) = \frac{\Pr(H|\pi_e)}{\Pr(H|\pi_b)} \sum_{t=1}^{L} \gamma^t R_t$$

- $\text{IS}(H)$ is an unbiased estimate of $J(\pi_e)$
- Estimate from D:

$$\begin{aligned}\text{IS}(D) &= \frac{1}{n}\sum_{i=1}^{n} \text{IS}(H_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \boxed{\frac{\Pr(H|\pi_e)}{\Pr(H|\pi_b)}} \sum_{t=1}^{L} \gamma^t R_t\end{aligned}$$

Computing the Importance Weight

$$\frac{\Pr(H|\pi_e)}{\Pr(H|\pi_b)}$$

$$= \frac{\Pr(S_1)\pi_e(A_1|S_1)\Pr(R_1, S_2|S_1, A_1)\pi_e(A_2|S_2)\Pr(R\_2, S\_3|S_2, A_2)\ldots}{\Pr(S_1)\pi_b(A_1|S_1)\Pr(R_1, S_2|S_1, A_1)\pi_b(A_2|S_2)\Pr(R\_2, S\_3|S_2, A_2)\ldots}$$

$$= \frac{\pi_e(A_1|S_1)\pi_e(A_2|S_2)\ldots}{\pi_b(A_1|S_1)\pi_b(A_2|S_2)\ldots}$$

$$= \prod_{t=1}^{L} \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)}$$

Importance Sampling for Reinforcement Learning

$$\text{IS}(D) = \frac{1}{n}\sum_{i=1}^{n}\frac{\Pr(H|\pi_e)}{\Pr(H|\pi_b)}\sum_{t=1}^{L}\gamma^t R_t$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{t=1}^{L}\frac{\pi_e\left(A_t^i|S_t^i\right)}{\pi_b\left(A_t^i|S_t^i\right)}\right)\sum_{t=1}^{L}\gamma^t R_t$$

Per-Decision Importance Sampling

- Use importance sampling to estimate $\mathbf{E}[R_t|\pi_e]$ independently for each t

$$\mathrm{IS}_t(D) = \frac{1}{n}\sum_{i=1}^{n}\frac{\Pr\left(H_t^i|\pi_e\right)}{\Pr\left(H_t^i|\pi_b\right)}R_t^i$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{j=1}^{t}\frac{\pi_{\mathrm{e}}\left(A_j^i|S_j^i\right)}{\pi_{\mathrm{b}}\left(A_j^i|S_j^i\right)}\right)R_t^i$$

$$\mathrm{PDIS}(D) = \sum_{t=1}^{L}\gamma^t \mathrm{IS}_t(D) = \sum_{t=1}^{L}\gamma^t\frac{1}{n}\sum_{i=1}^{n}\left(\prod_{j=1}^{t}\frac{\pi_{\mathrm{e}}\left(A_j^i|S_j^i\right)}{\pi_{\mathrm{b}}\left(A_j^i|S_j^i\right)}\right)R_t^i$$

Importance Sampling Range / Variance

- What is the range of the importance sampling estimator?

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

  - Mountain car with mediocre behavior policy, $L \approx 1000$
  - $\frac{\pi_e(a|s)}{\pi_b(a|s)} \in [0, 2.0]$, $\sum_{t=1}^{L}\gamma^t r_t \in [0,1]$
  - $\mathrm{IS}(D) \in [0, 2^{1000}]$
- The importance sampling estimator may be unbiased, but it has **high variance**.
  - Particularly when π_e and π_b are quite different
  - MSE = Bias^2 + Var, $\mathbf{E}\left[\left(\mathrm{IS}(D) - J(\pi_e)\right)^2\right] = \left(\mathbf{E}[\mathrm{IS}(D)] - J(\pi_e)\right)^2 + \mathrm{Var}(\mathrm{IS}(D))$

Importance Sampling (More Intuition)

- What value does the IS estimator take in practice if π_e and π_b are very different?

$$\text{IS}(D) = \frac{1}{n}\sum_{i=1}^{n}\frac{\Pr(H_i|\pi_e)}{\Pr(H_i|\pi_b)}\text{Return}(H_i)$$

- $\text{IS}(D) \approx 0$
- As n (the number of histories in D) increases, $\text{IS}(D)$ tends towards $J(\pi_e)$
  - Formally, $\text{IS}(D)$ is a *strongly consistent* estimator of $J(\pi_e)$
    - $\text{IS}(D)$ converges almost surely to $J(\pi_e)$ as $n \to \infty$
    - $\Pr\left(\lim_{n\to\infty} IS(D) = J(\pi_e)\right) = 1$

An Idea

- Recall that MSE = Bias^2 + Var
- Bias(IS) = 0
- Var(IS) = Huge
- Can we make a new importance sampling estimator that has some bias, but drastically lower variance?
  - Perhaps make $\mathbf{E}[\text{new estimator}] = J(\pi_b)$ when there is little data
  - As we gather more data, have the expected value converge to $J(\pi_e)$
    - The new estimator should remain strongly consistent

Weighted Importance Sampling

$$w_i = \prod_{t=1}^{L} \frac{\pi_e\left(A_t^i \middle| S_t^i\right)}{\pi_b\left(A_t^i \middle| S_t^i\right)}$$

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

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

Weighted Importance Sampling

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

- What if $n = 1$?

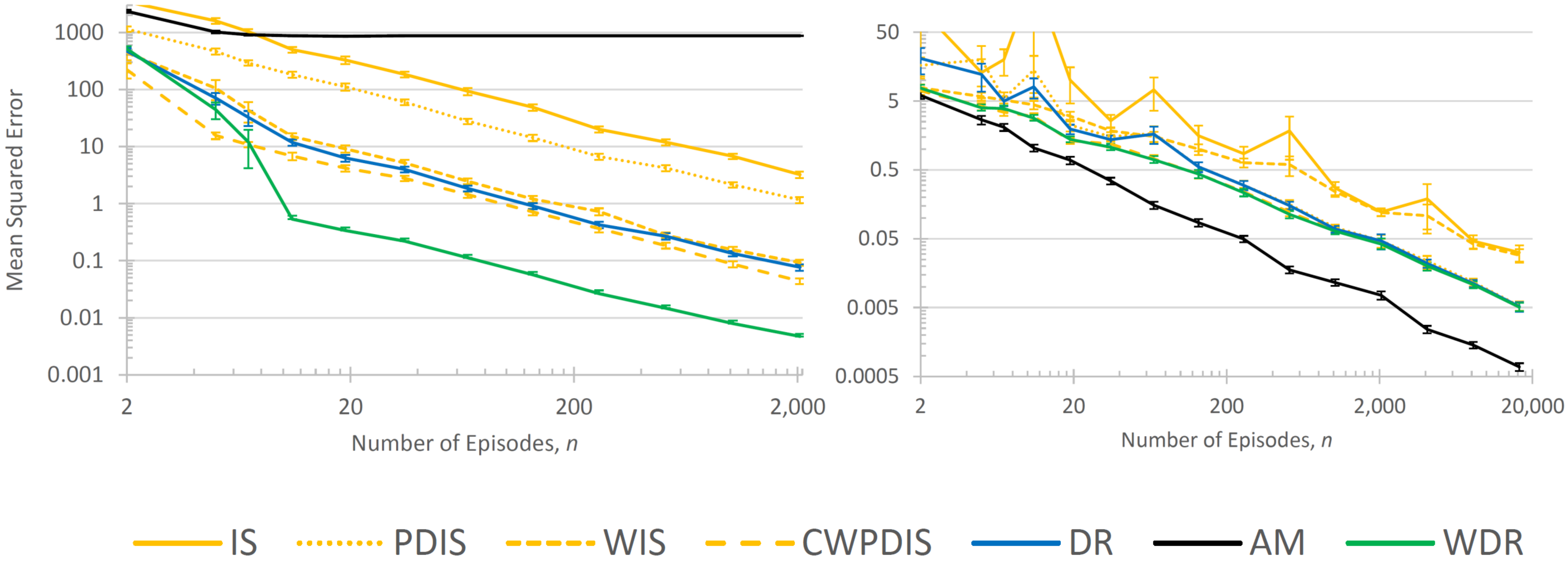
$$\text{WIS}(H) = \sum_{t=1}^{L} \gamma^t R_t^i$$

- $\mathbf{E}[w_i] = \mathbf{E}\left[\frac{p(Y)}{q(Y)}\right] = \sum_y q(y) \frac{p(y)}{q(y)} = \sum_y p(y) = 1$
  - $\sum_{j=1}^{n} w_j \to n$ almost surely
  - WIS acts like the Monte Carlo estimator of $J(\pi_b)$ with little data and $\text{IS}(D)$ with lots of data

Off-Policy Policy Evaluation (OPE) Overview

- Importance Sampling (IS)
- Per-Decision Importance Sampling (PDIS)
- Weighted Importance Sampling (WIS)
- Others
  - Weighted Per-Decision Importance Sampling (WPDIS or CWPDIS)
  - Importance sampling with unequal support (US)
  - Model-based estimators (Direct Method / Indirect Method / Approximate Model)
  - Doubly robust importance sampling
  - Weighted doubly robust importance sampling
  - Importance Sampling (IS) + Time Series Prediction (TSP)
  - MAGIC (Model And Guided Importance sampling Combined)

Off-Policy Policy Evaluation (OPE) Examples



Overview

- Background and motivation
- Definition of "safe"
- Three steps towards a safe algorithm
  - Off-policy policy evaluation
  - High-confidence off-policy policy evaluation
  - Safe policy improvement
- Experimental results
- Conclusion

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

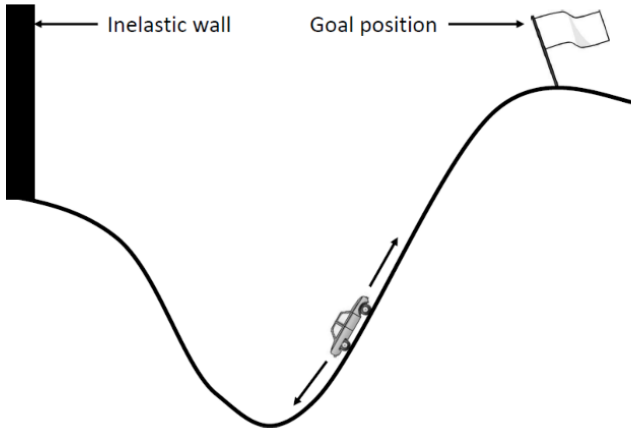
- Let $X_1, \ldots, X_n$ be n independent identically distributed random variables such that $X_\mathrm{i} \in [0, b]$
- Then with probability at least $1 - \delta$:

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

Applying Hoeffding's Inequality

- Example: Mountain Car
  - $J(\pi_e) = 0.19 \in [0,1]$
  - $n = 100{,}000$
  - Lower bound from Hoeffding's inequality:

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



What went wrong?

- Recall: $\mathrm{IS}(D) \in [0, 2^{1000}]$
  - $b = 2^{1000}$

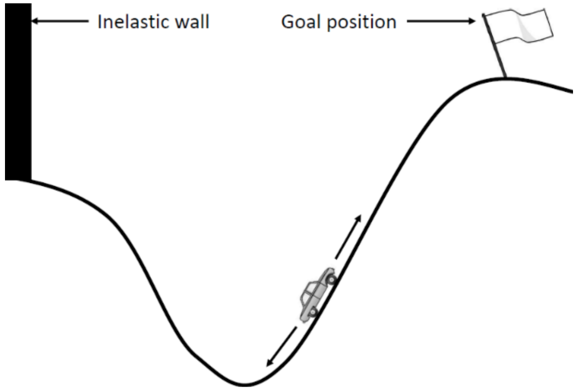
$$\mathbf{E}[X_i] \geq \frac{1}{n}\sum_{i=1}^{n} X_i - b\sqrt{\frac{\ln\left({}^{1}/{}_{\delta}\right)}{2n}}$$

Applying Other Concentration Inequalities

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 $\mathbb{P}[0 \le X_i] = 1$, $\mathbb{E}[X_i] \le \mu$, 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*

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

See “High Confidence Off-Policy Policy Evaluation”, AAAI 2015 for how to select c_i



Actual	Hoeffding	Maurer & Pontil	Anderson & Massart	CUT Inequality
0.19	-5,831,000	-129,703	0.055	0.154

Approximate Confidence Intervals: t-Test

- If $\frac{1}{n}\sum_{i=1}^{n} X_i$ is normally distributed, then by Student's t-test, with probability at least $1-\delta$:

$$\mathbf{E}[X_i] \geq \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{\sigma}{\sqrt{n}} t_{1-\delta,n-1}$$

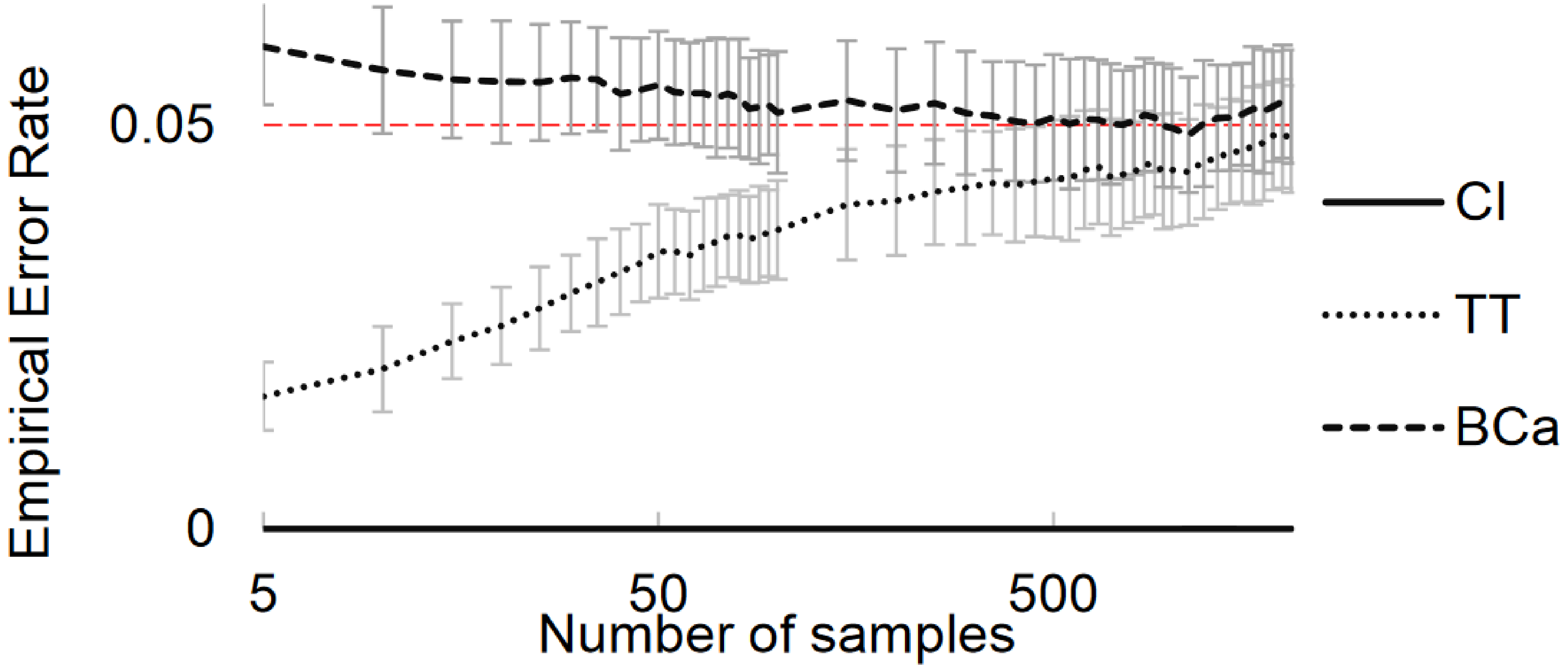
where σ is the sample standard deviation of $X_1, \ldots, X_n$ with Bessel's correction.

- By the central limit theorem, $\frac{1}{n}\sum_{i=1}^{n} X_i$ is approximately normally distributed
- If rewards non-negative then the t-test tends to be *conservative.*

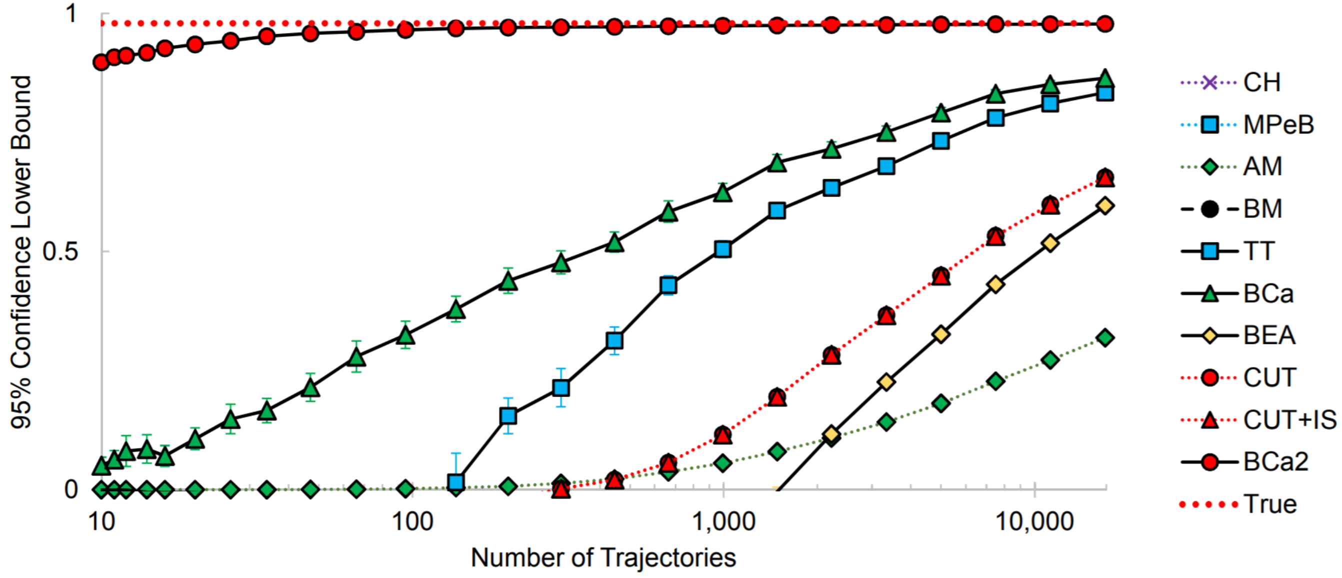
Approximate Confidence Intervals: Bootstrap

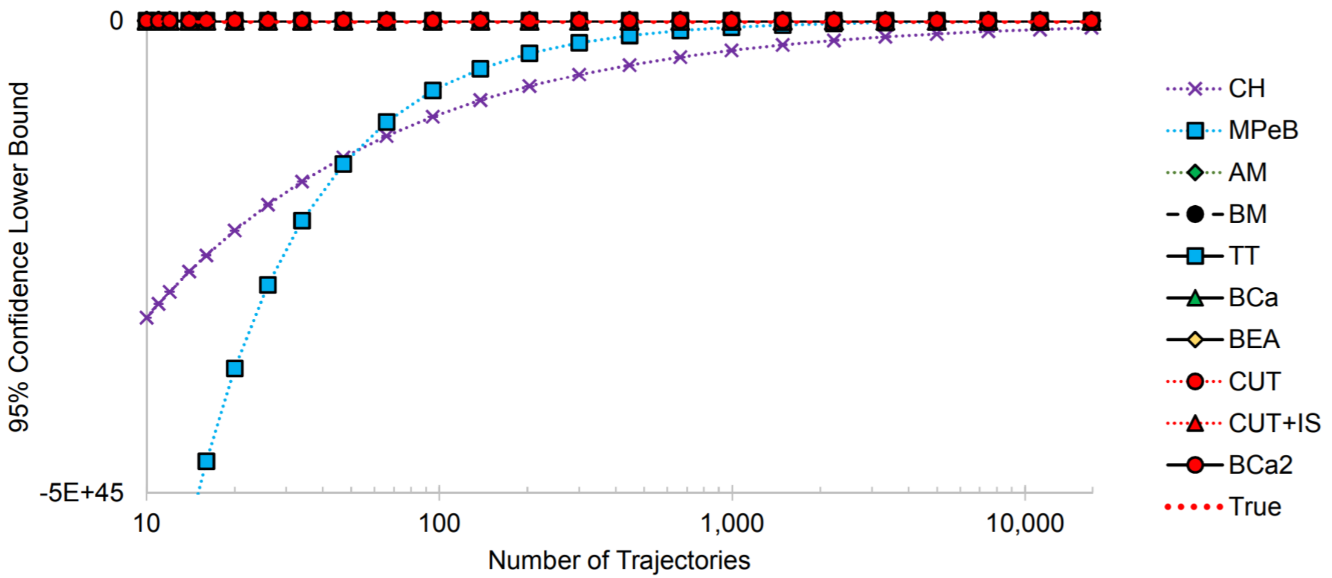
- Efron's bootstrap, not TD's bootstrap
- Resample n samples from $X_1, \dots, X_n$ with replacement to create a new data set, D_1
- Repeat this process $\beta \approx 2{,}000$ times to create β data sets, $D_1, \dots, D_\beta$
- Pretend that these β data sets represent new independent runs
- Run importance sampling (or any OPE method) on each data set:
$$\mathrm{IS}(D_1), \dots, \mathrm{IS}\left(D_\beta\right)$$
- Sort these estimates and return the $\delta\beta$'th smallest

CI vs t-Test vs Bootstrap (non-negative rewards)

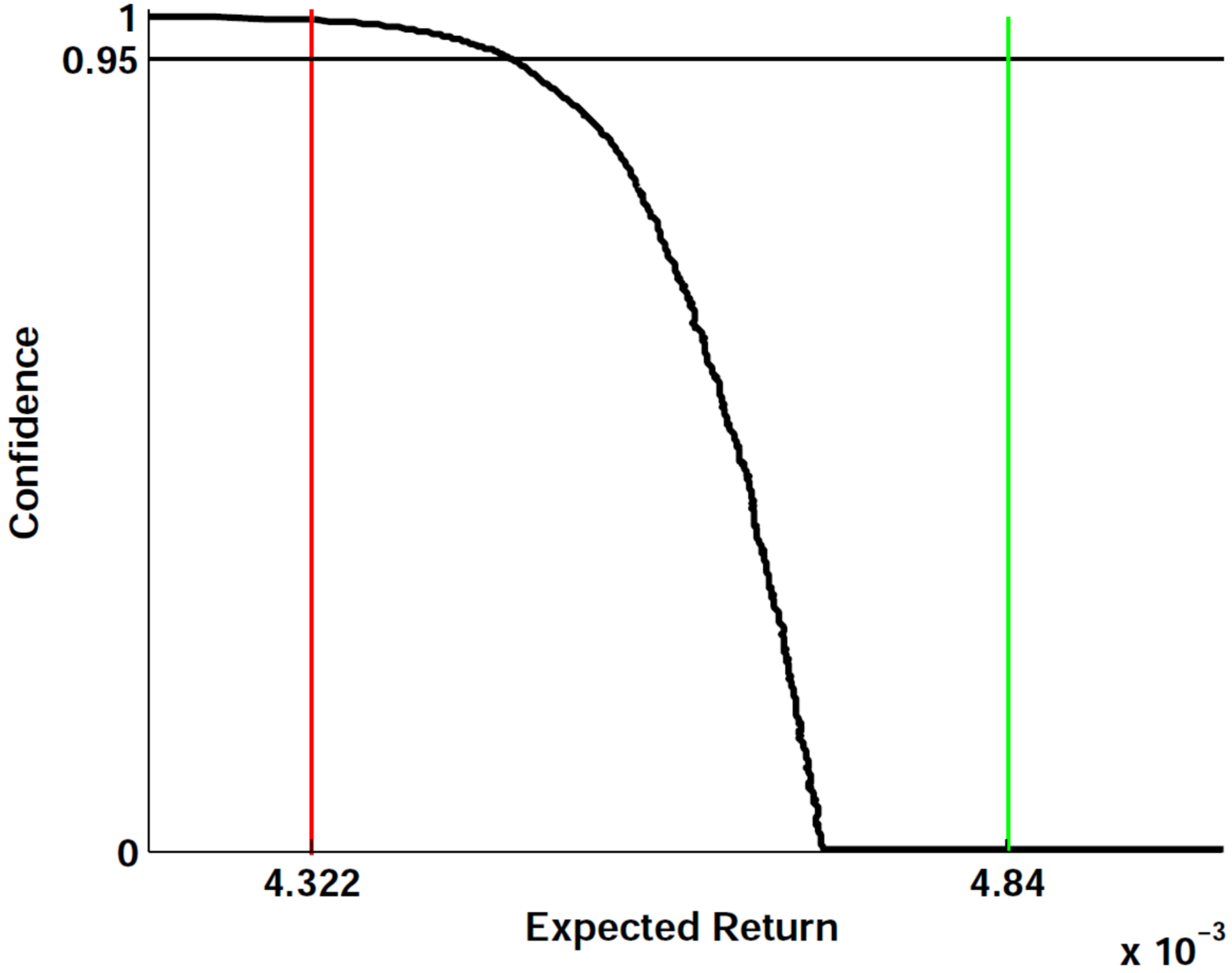


HCOPE: Mountain Car





HCOPE: Digital Marketing



HCOPE Summary

- Use OPE method (e.g., importance sampling) to produce an estimate of $J(\pi_e)$ from each history
- Use a concentration inequality to bound $J(\pi_e)$ given these n estimates
- Suggested method:
  - Weighted doubly robust + Student's t-Test
- Suggested simple method:
  - Weighted per-decision importance sampling + Student's t-Test
- Suggested method if computation is not an issue:
  - Weighted doubly robust + Bias-Corrected and Accelerated Bootstrap (BCa)

HCOPE Using Weighted Per-Decision Importance Sampling and Student's t-Test

- **Input**: **1)** n histories, $H_1, \ldots, H_n$ produced by a known policy, π_b. **2)** An evaluation policy, π_e. **3)** A probability, $1-\delta$.
- Allocate 2-dimensional array, $\rho[L][n]$, and 1-dimensional arrays $\xi[L]$ and $\hat{J}[n]$. Initialize $\hat{J}$ array to zero.
- For $t = 1$ to L
  - For $i = 1$ to n
    - $\rho[t][i] = \prod_{j=1}^{t} \frac{\pi_\mathrm{e}\left(A_j^i \middle| S_j^i\right)}{\pi_\mathrm{b}\left(A_j^i \middle| S_j^i\right)}$
  - $\xi[t] = \sum_{i=1}^{n} \rho[t][i]$
- For $i = 1$ to n
  - For $t = 1$ to L
    - $\hat{J}[i] = \hat{J}[i] + \frac{\rho[t][i]}{\xi[t]} \gamma^t R_t^i$
- $\bar{J} = \text{average}(\hat{J}[1], \hat{J}[2], \ldots, \hat{J}[n])$
- $\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(\hat{J}[i] - \bar{J}\right)^2}$
- Return $\bar{J} - \frac{\sigma}{\sqrt{n}} \mathrm{tinv}(1-\delta, n-1)$ // See MATLAB documentation for tinv

Note: More efficient implementations exist.
E.g., $\rho[t][i]$ can be computed starting from $\rho[t-1][i]$

Overview

- Background and motivation
- Definition of “safe”
- Three steps towards a safe algorithm
  - Off-policy policy evaluation
  - High-confidence off-policy policy evaluation
  - Safe policy improvement
- Experimental results
- Conclusion

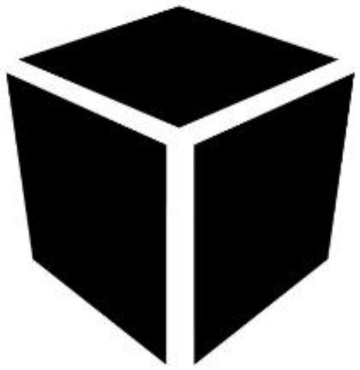
Safe Policy Improvement (SPI)

- Given the historical data, D, produced by the behavior policy, π_b
- Given a probability, $1 - \delta$
- Produce a policy, π, that we predict maximizes $J(\pi)$ and which satisfies:

$$\Pr\big(J(\pi) \geq J(\pi_b)\big) \geq 1 - \delta$$

Historical Data, D

Probability, $1 - \delta$

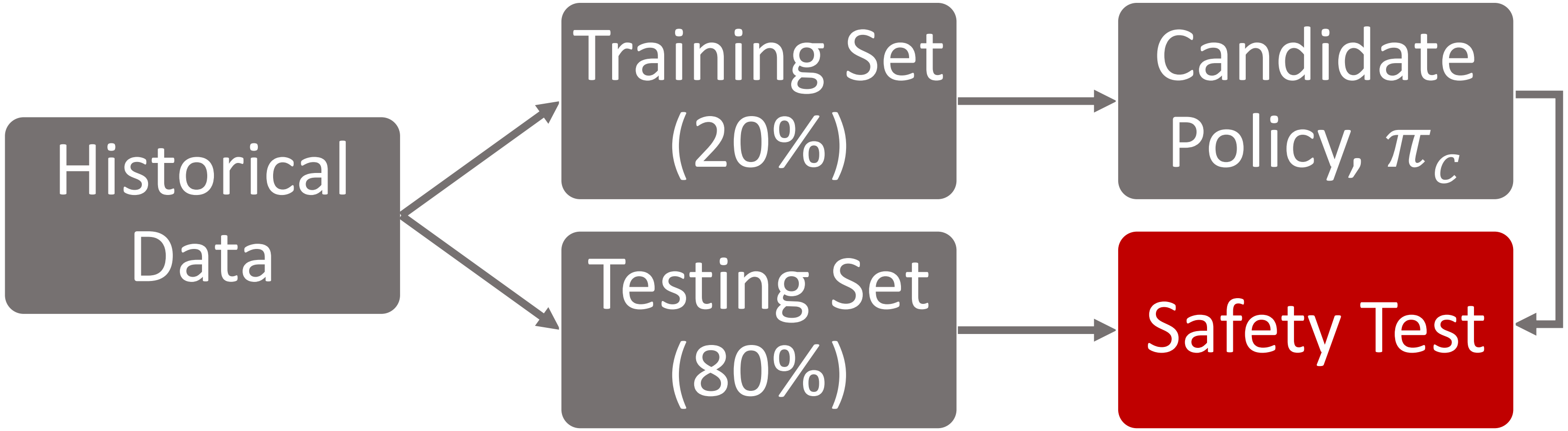


New policy π, or
`No Solution Found`

Safe Policy Improvement

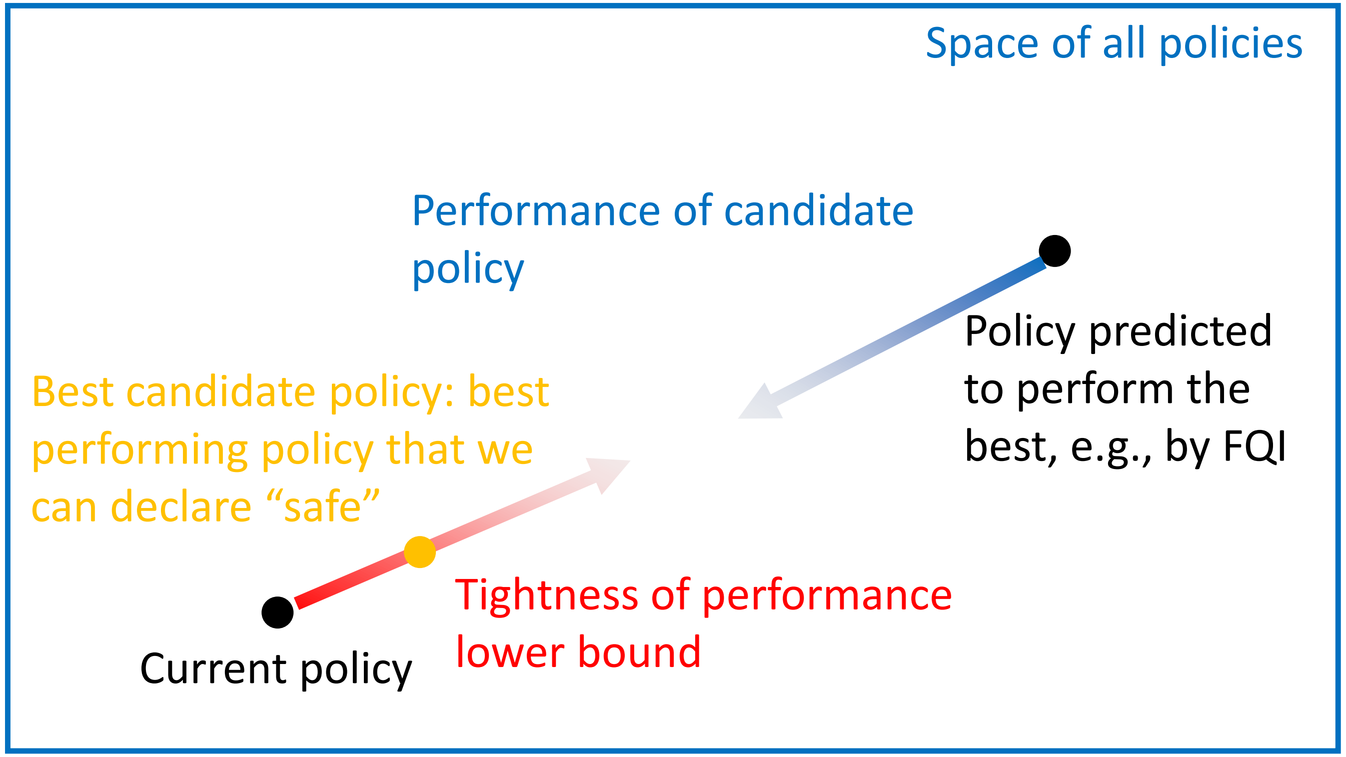
- Split data, D, into two sets, D_{train} and D_{test}
- Use batch RL algorithm on D_{train}
  - Call output policy, π_c, the *candidate policy*
- Use HCOPE algorithm and D_{test} to lower bound $J(\pi_c)$ with probability $1-\delta$. Store this value in `lower_bound`.
- If `lower_bound` $\geq J(\pi_b)$, return π_c
- Else, return No Solution Found, i.e., π_b

Safe Policy Improvement



Is $1-\delta$ confidence lower bound on $J(\pi_c)$ larger that $J(\pi_b)$?

Selecting the Candidate Policy

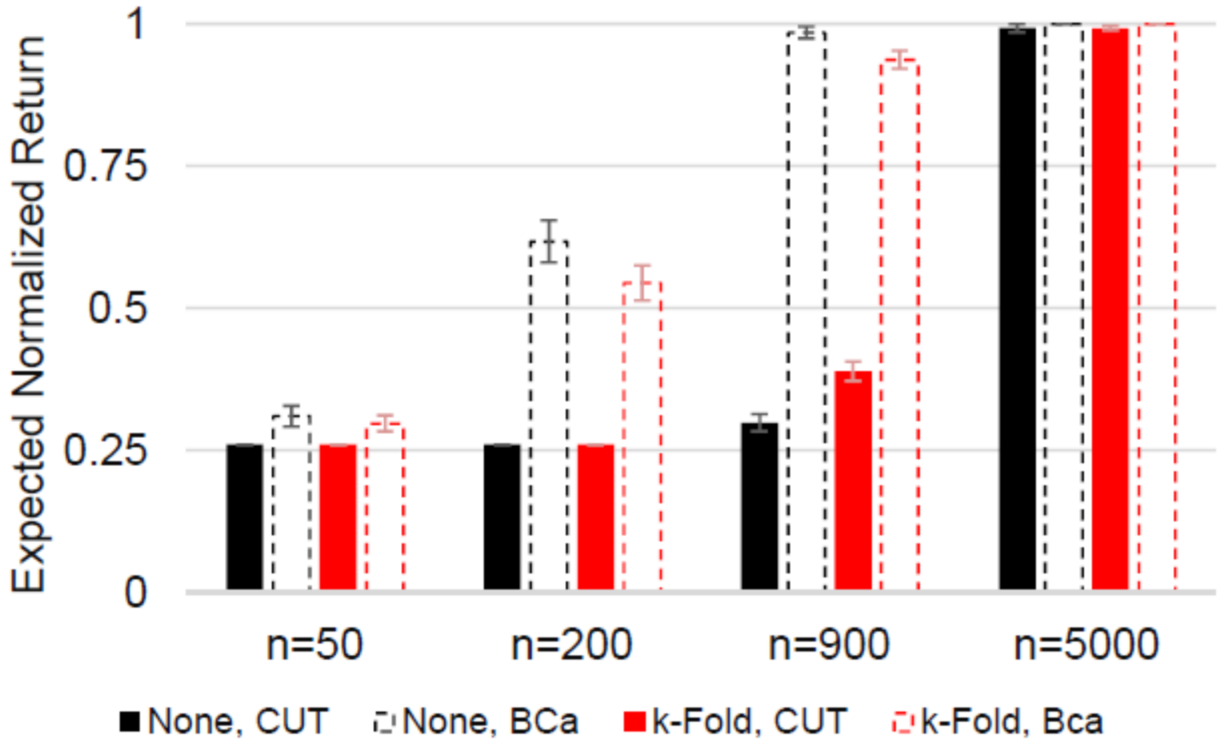


- Use regularization when selection candidate policy to stay "close" to the current policy.

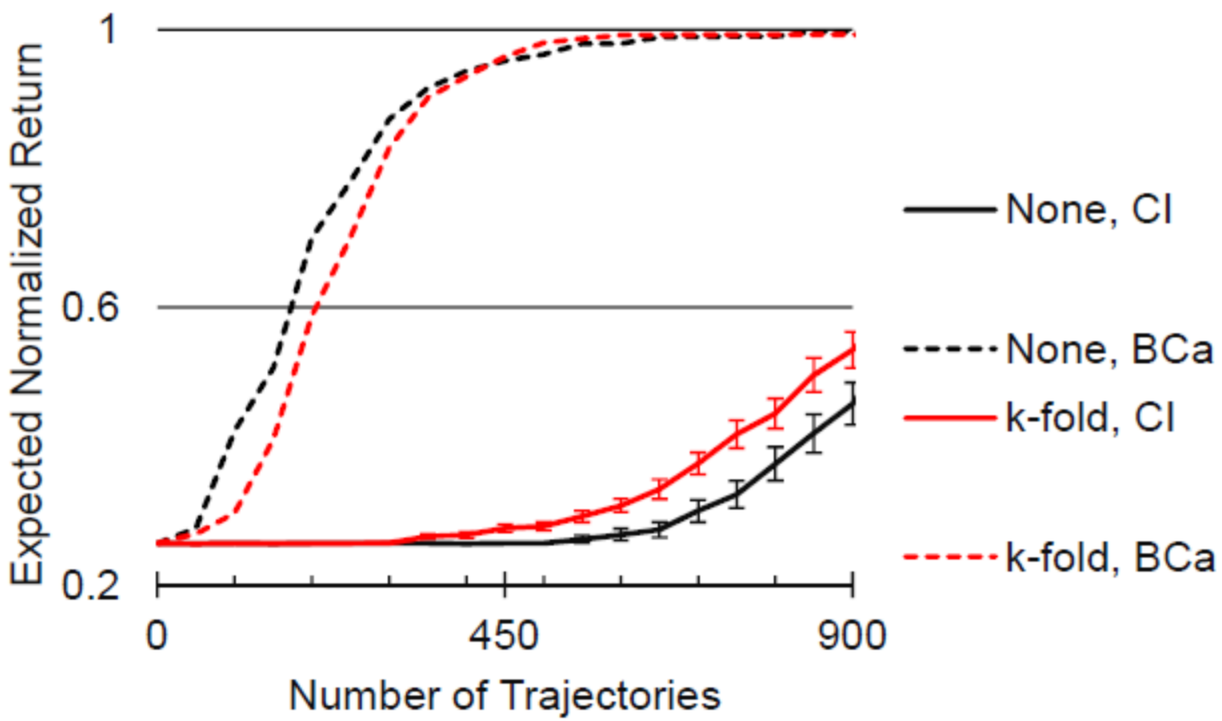
Overview

- Background and motivation
- Definition of “safe”
- Three steps towards a safe algorithm
  - Off-policy policy evaluation
  - High-confidence off-policy policy evaluation
  - Safe policy improvement
- Experimental results
- Conclusion

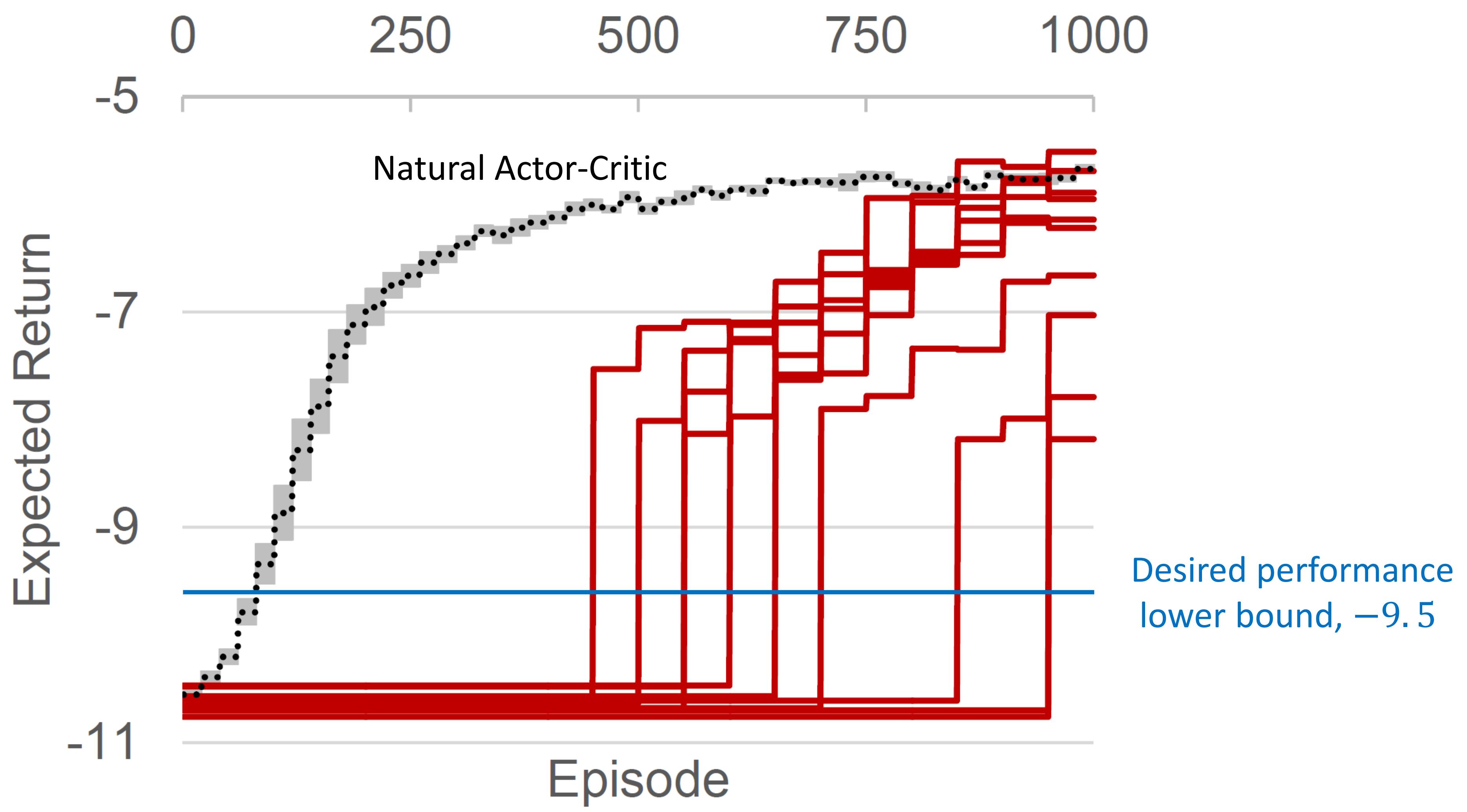
Experimental Results: Mountain Car



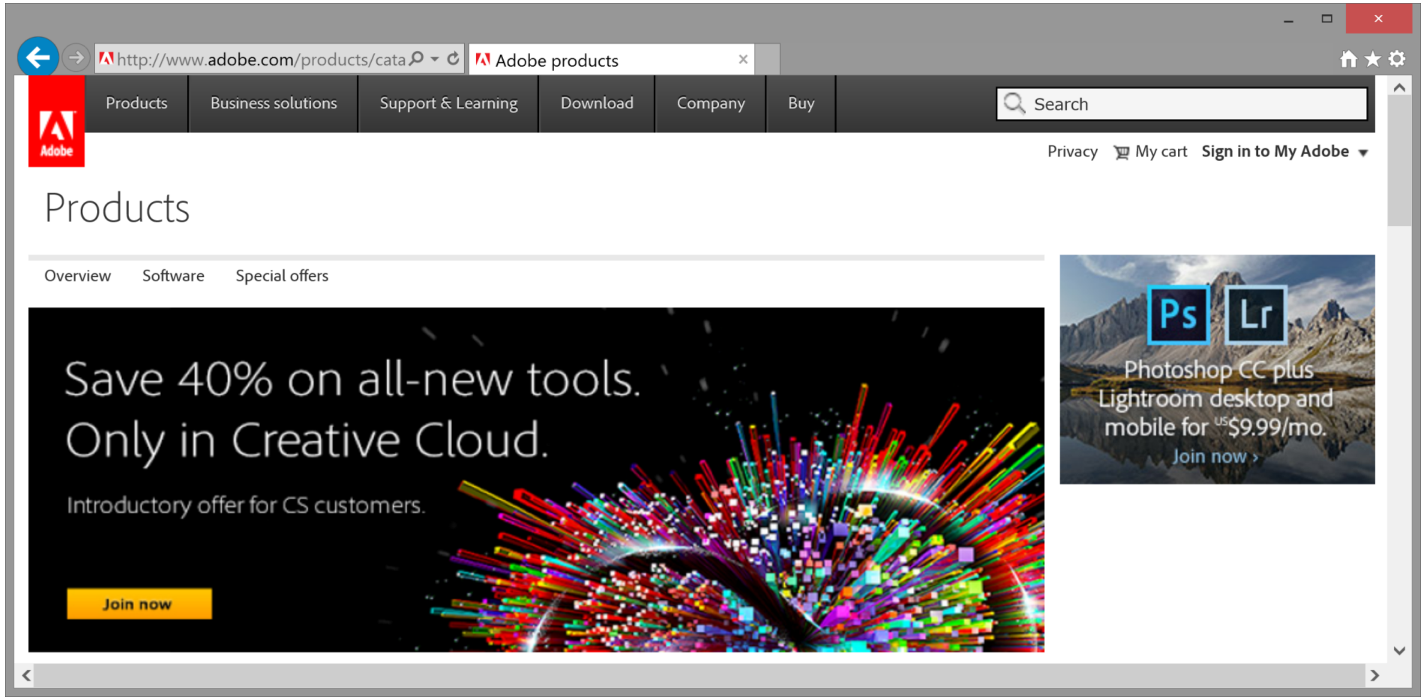
Experimental Results: Mountain Car

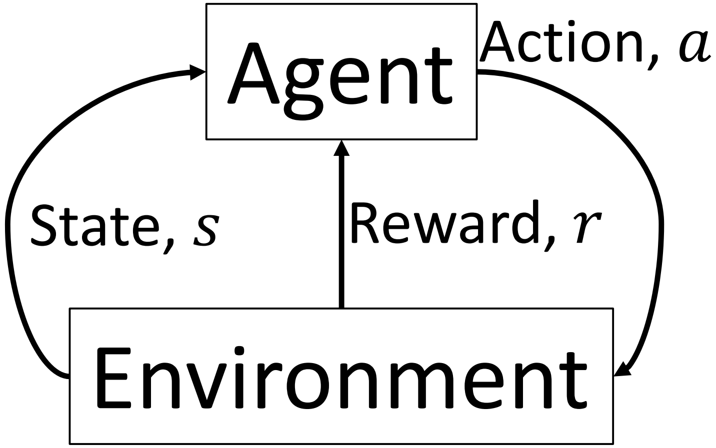


Experimental Results: Mountain Car

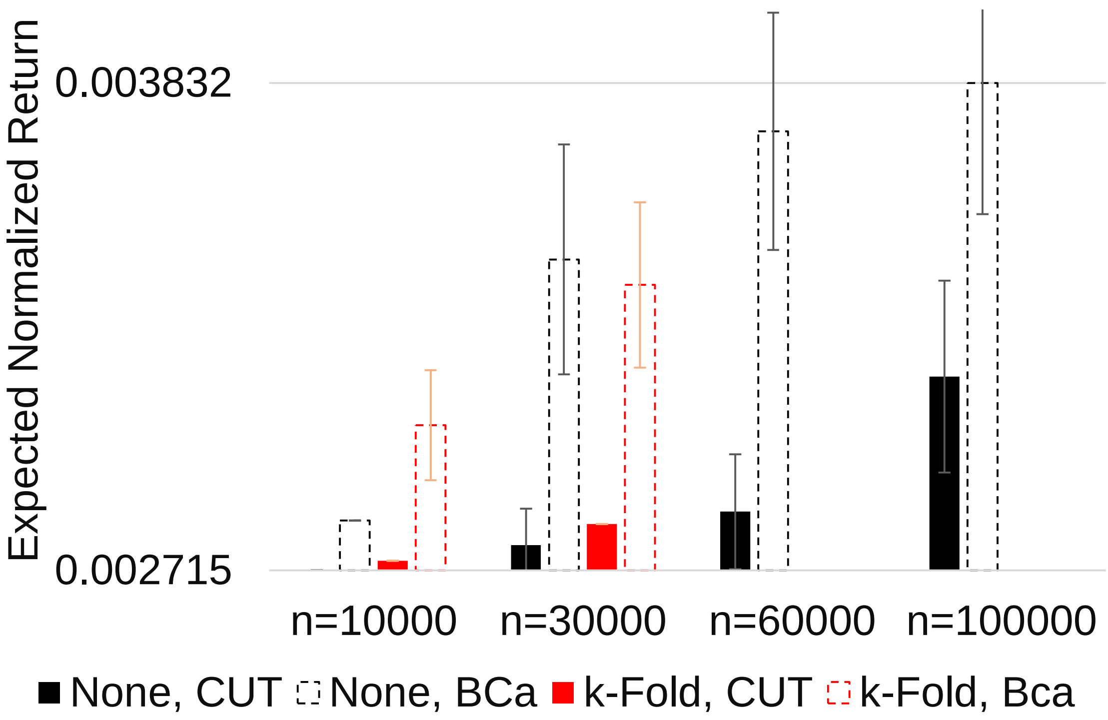


Experimental Results: Digital Marketing

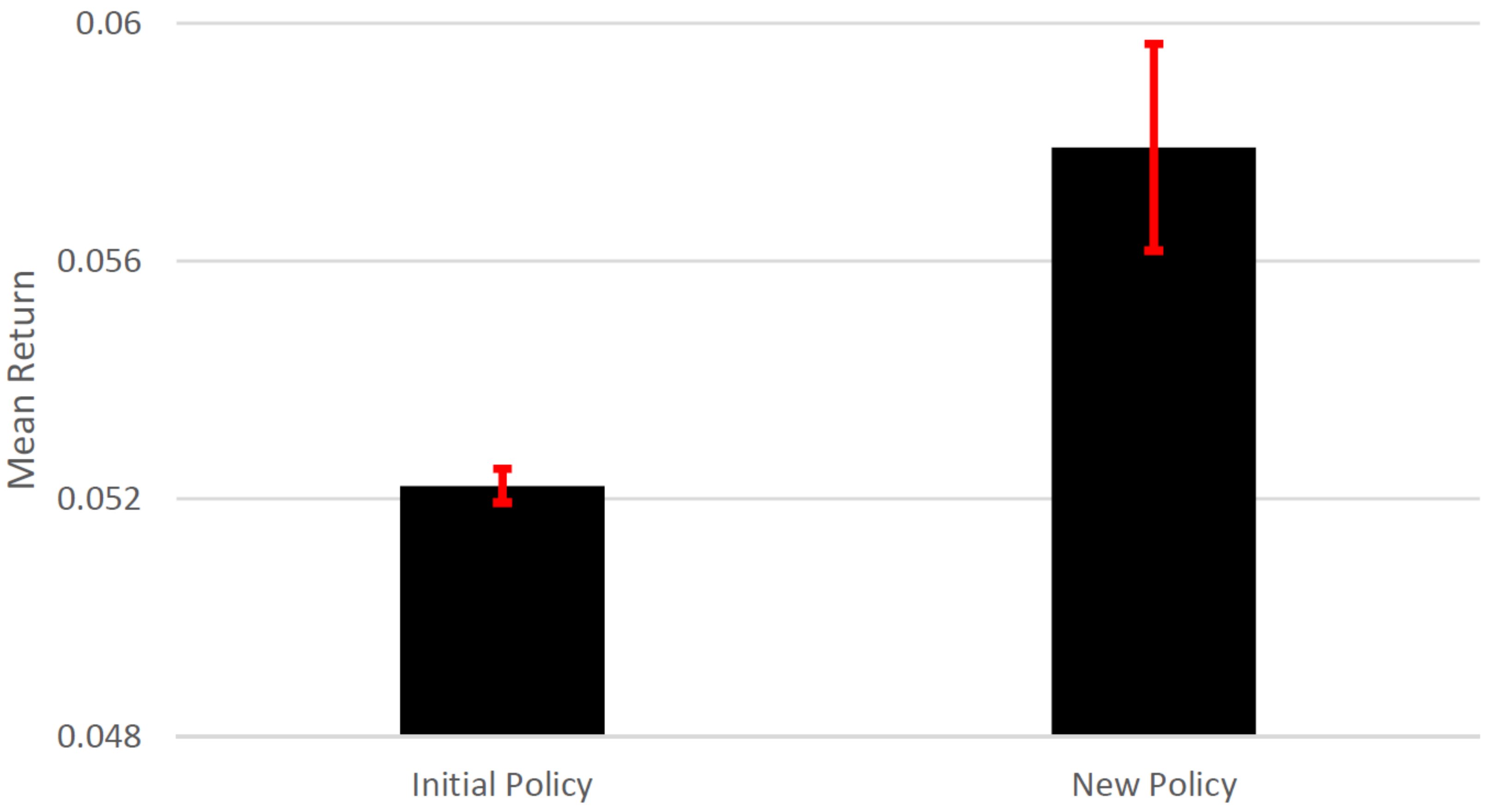




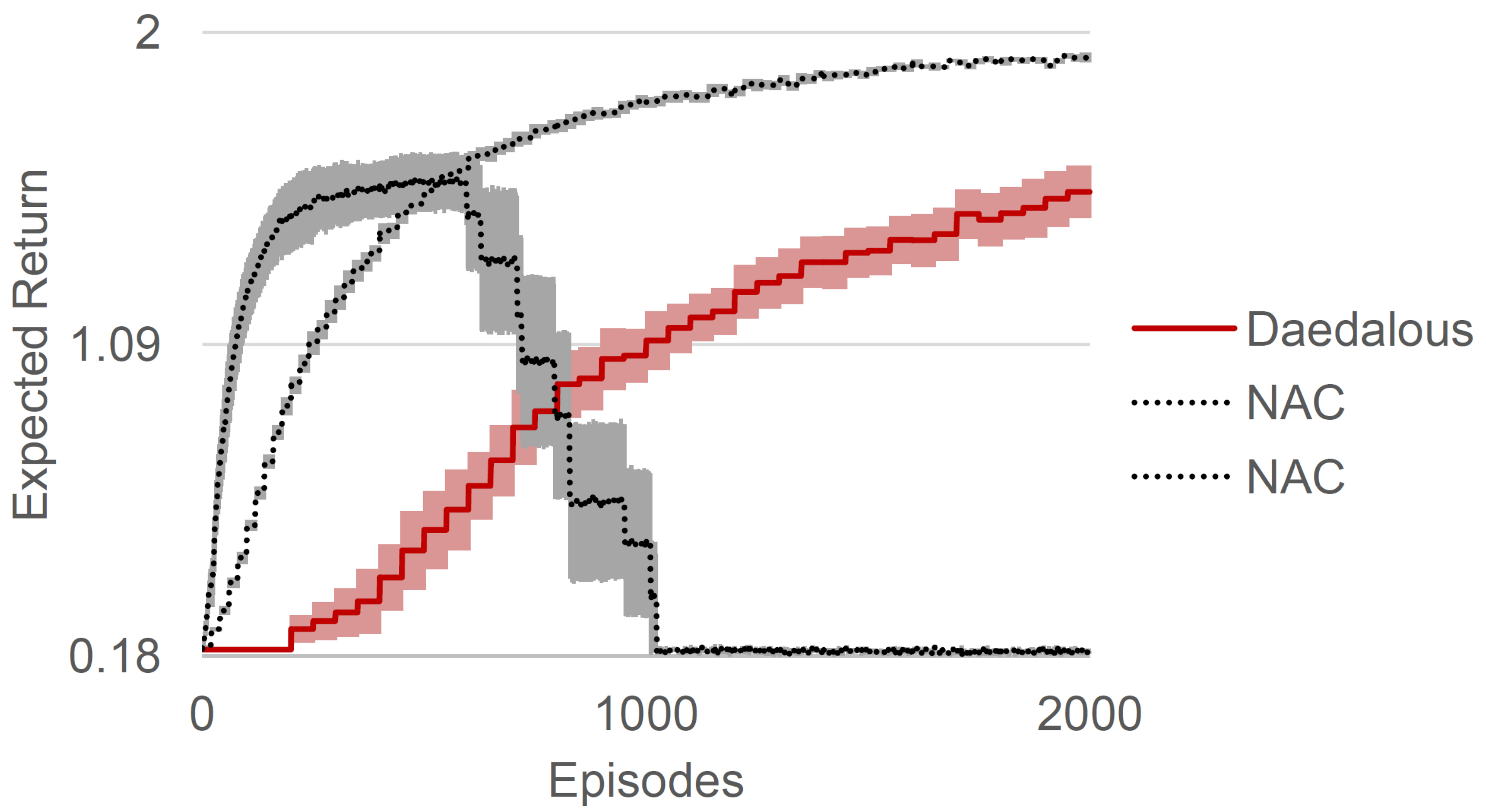
Experimental Results: Digital Marketing



Experimental Results: Digital Marketing



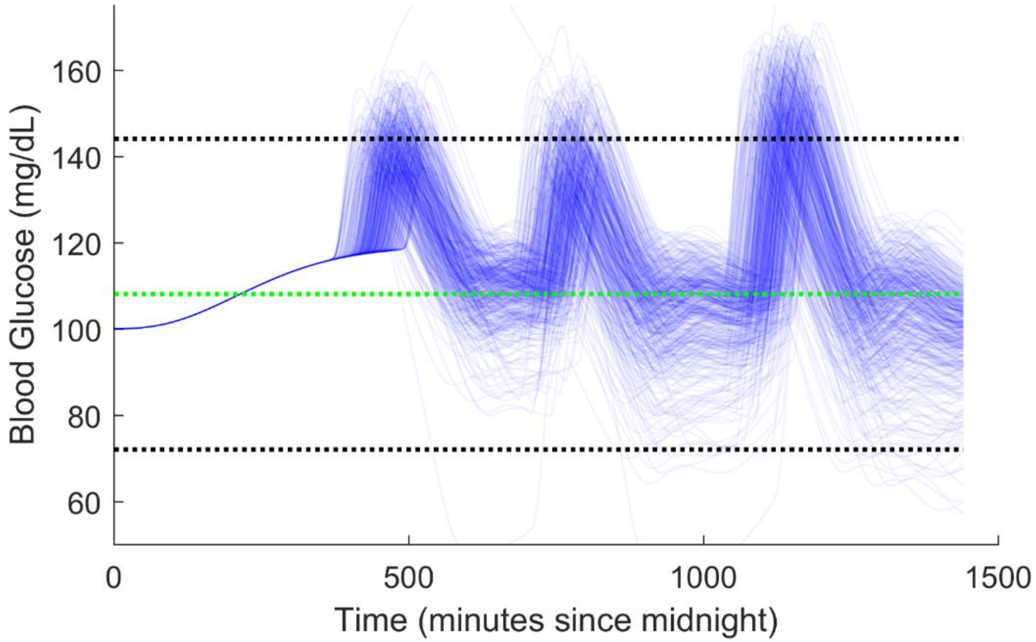
Experimental Results: Digital Marketing



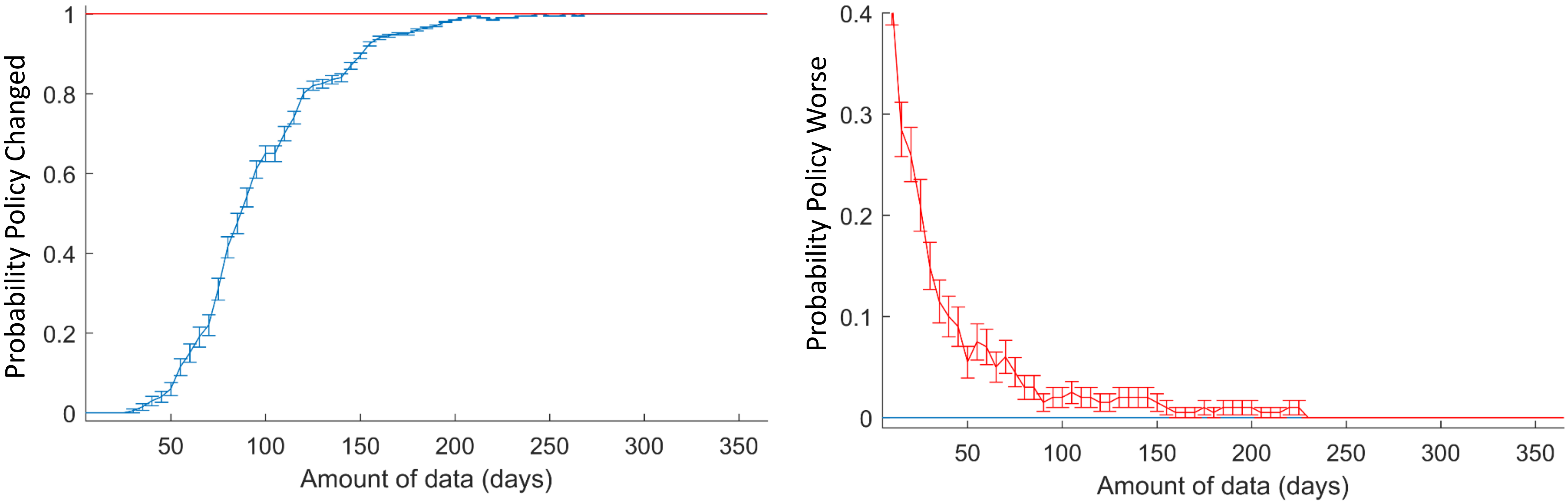
Experimental Results: Diabetes Treatment

T1DMS

Type 1 Diabetes Metabolic Simulator



Experimental Results : Diabetes Treatment



Overview

- Background and motivation
- Definition of “safe”
- Three steps towards a safe algorithm
  - Off-policy policy evaluation
  - High-confidence off-policy policy evaluation
  - Safe policy improvement
- Experimental results
- Conclusion

Conclusion: Summary

- Many definitions of “safe reinforcement learning”.
  - With probability at least $1-\delta$ the algorithm will not return a worse policy
- Three steps to making a safe reinforcement algorithm
  - Off-policy Policy Evaluation (OPE)
    - Importance sampling variants
  - High Confidence Off-policy Policy Evaluation (HCOPE)
    - Concentration inequalities / Student’s t-Test / Bootstrap
  - Safe Policy Improvement
    - Select candidate policy using some data and bound its performance using the rest
- Empirical Results
  - Safe RL is tractable!

Conclusion: Future Directions

- Improvements have been by orders of magnitude. Several orders left to go.
- OPE
  - Can we handle long horizon problems?
  - Can we handle non-episodic problems?
  - What if the behavior policy is not known?
  - What if the environment is non-stationary?
  - How best to leverage prior knowledge like an estimate of the transition function?
- HCOPE
  - Better concentration inequalities for importance sampling?
- Safe Policy Improvement
  - Better techniques for selecting the candidate policy?
  - Automate decision of how much data to use in D_{train}?

Conclusion: References and Additional Reading

- Importance sampling for RL (IS, PDIS, WIS, CWPDIS)
  - D. Precup, R. S. Sutton, and S. Singh. Eligibility traces for off-policy policy evaluation. In Proceedings of the 17th International Conference on Machine Learning, pages 759–766, 2000. [NOTE: WPDIS estimator has a typo]
  - P. S. Thomas. Safe reinforcement learning. PhD Thesis, UMass Amherst, 2015.
- Doubly robust importance sampling and MAGIC for RL
  - N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. ICML 2016
  - P. S. Thomas and E. Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. ICML 2016.
- Other importance sampling estimators for RL (more for bandits)
  - P. S. Thomas and E. Brunskill. Importance Sampling with Unequal Support. AAAI 2017
  - P. S. Thomas., G. Theocharous, M. Ghavamzadeh, I. Durugkar, and E. Brunskill. Predictive Off-Policy Policy Evaluation for Nonstationary Decision Problems, with Applications to Digital Marketing. IAAI 2017.
  - S. Daroudi, P. S. Thomas, and E. Brunskill. Importance Sampling for Fair Policy Selection. UAI 2017.
  - Z. Guo, P. S. Thomas, and E. Brunskill. Using Options for Long-Horizon Off-Policy Evaluation. RLDM 2017.
  - Y. Liu, P. S. Thomas, and E. Brunskill. Model Selection for Off-Policy Policy Evaluation. RLDM 2017.
  - P. S. Thomas, S. Niekum, G. Theocharous, and G.D. Konidaris. Policy Evaluation Using the Omega-Return. NIPS 2015.
- HCOPE
  - L. Bottou, J. Peters, J. Quinonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. JMLR 2013.
  - J.P. Hanna, P. Stone, and S. Niekum. Bootstrapping with Models: Confidence Intervals for Off-Policy Evaluation. AAMAS 2017.
  - P. S. Thomas, G. Theocharous, and M. Ghavamzadeh. High Confidence Off-Policy Evaluation. AAAI 2015.
  - P. S. Thomas . Safe reinforcement learning. PhD Thesis, UMass Amherst, 2015.
- Safe Policy Improvement
  - P. S. Thomas, G. Theocharous, and M. Ghavamzadeh. High Confidence Policy Improvement. ICML 2015
  - P. S. Thomas. Safe reinforcement learning. PhD Thesis, UMass Amherst, 2015.