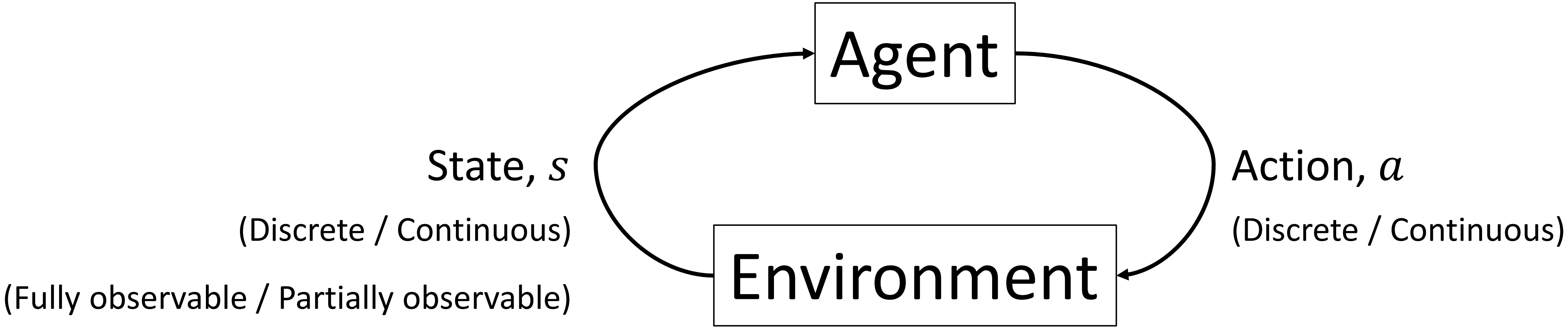
High Confidence Off-Policy Evaluation (HCOPE)

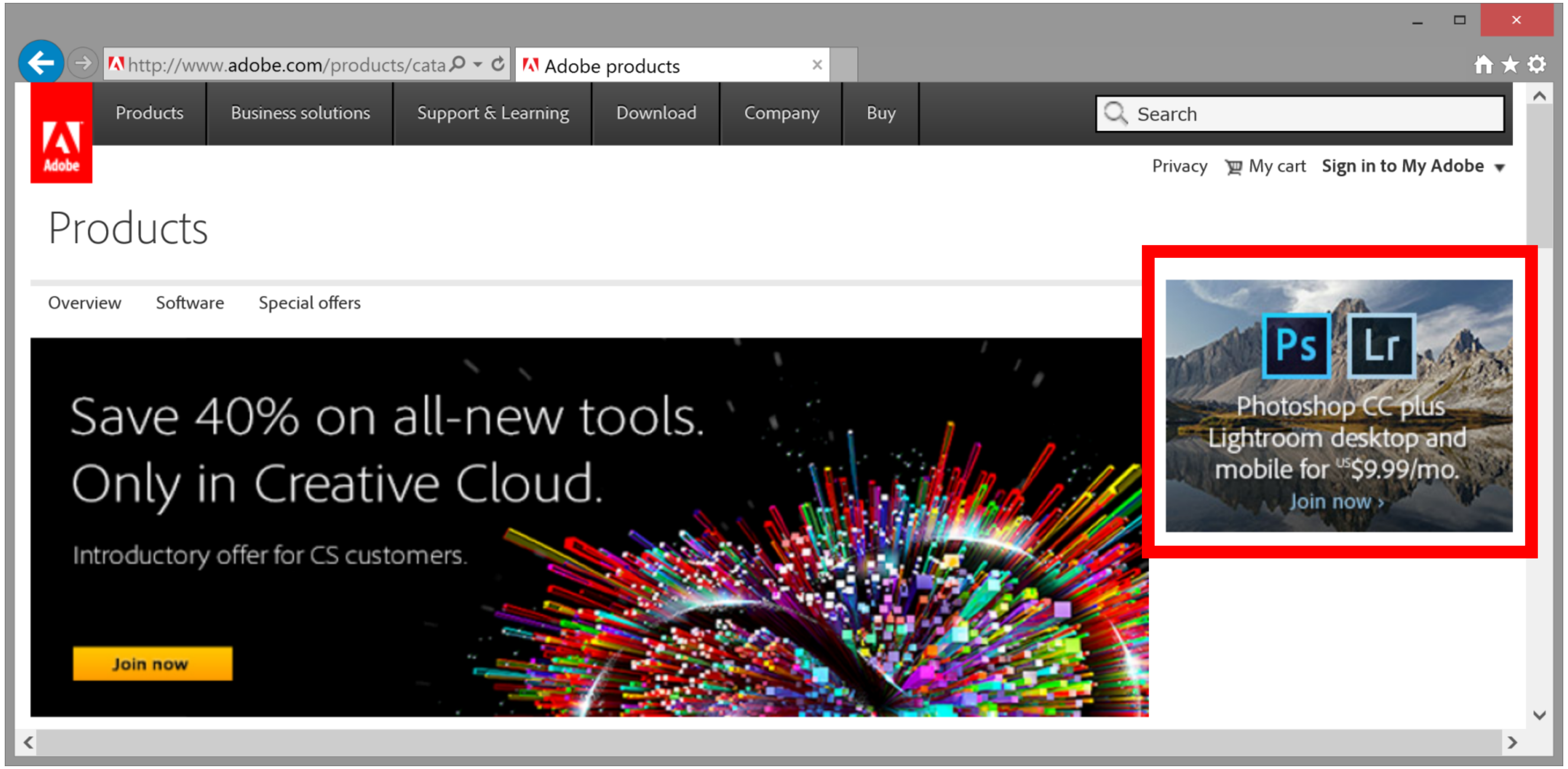
Philip Thomas

CMU 15-899E, Real Life Reinforcement Learning, Fall 2015

Sequential Decision Problems

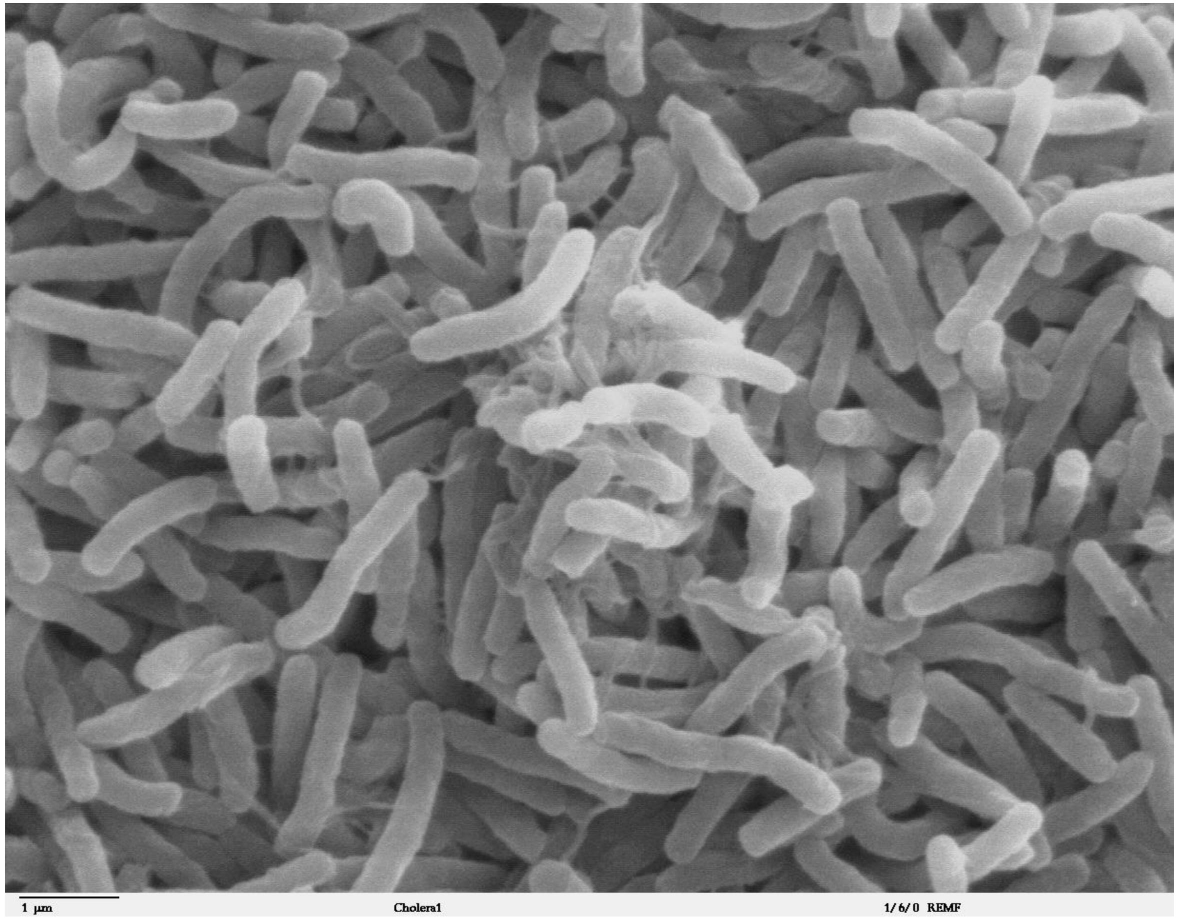


Example: Digital Marketing

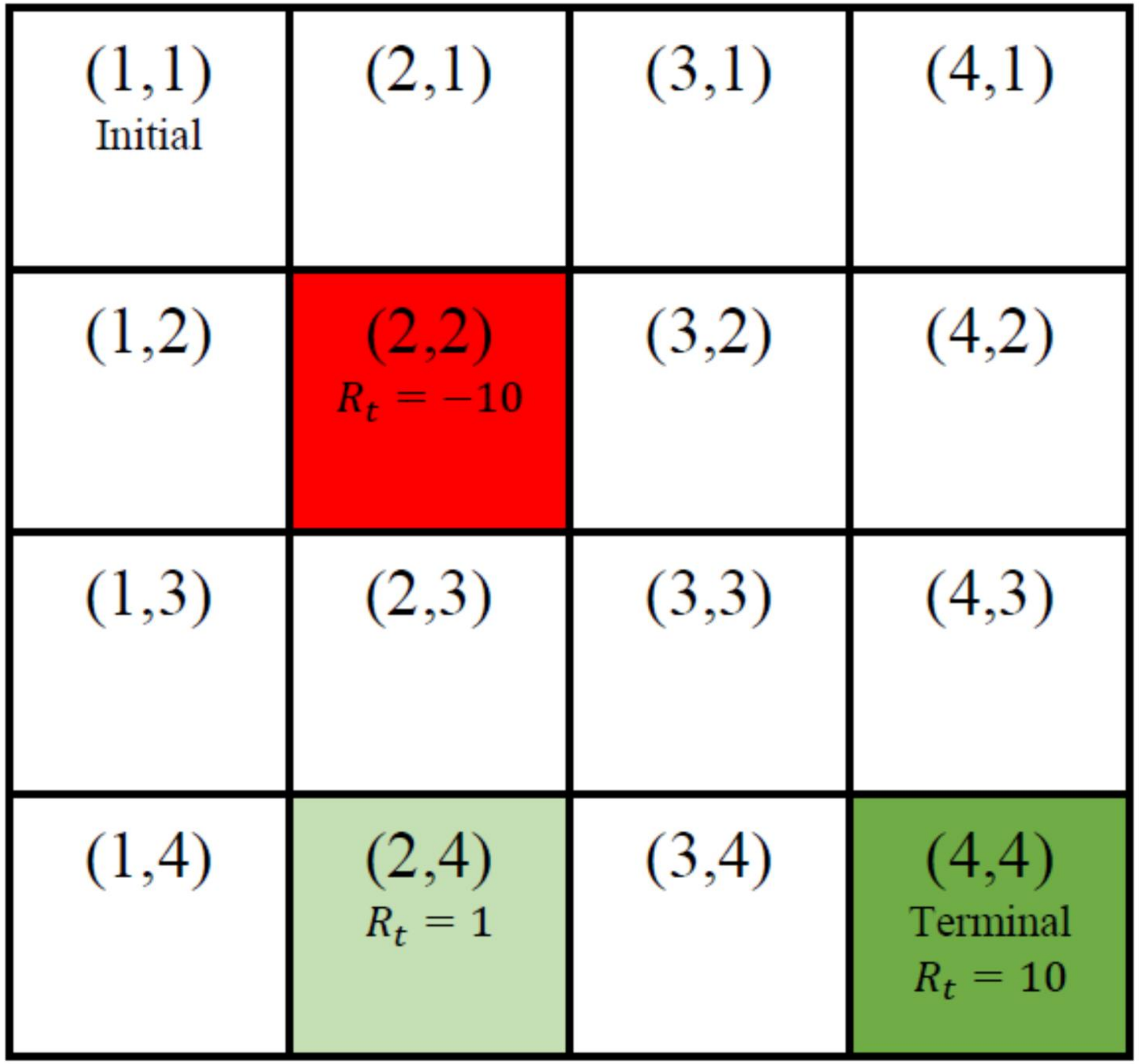


Example: Educational Games

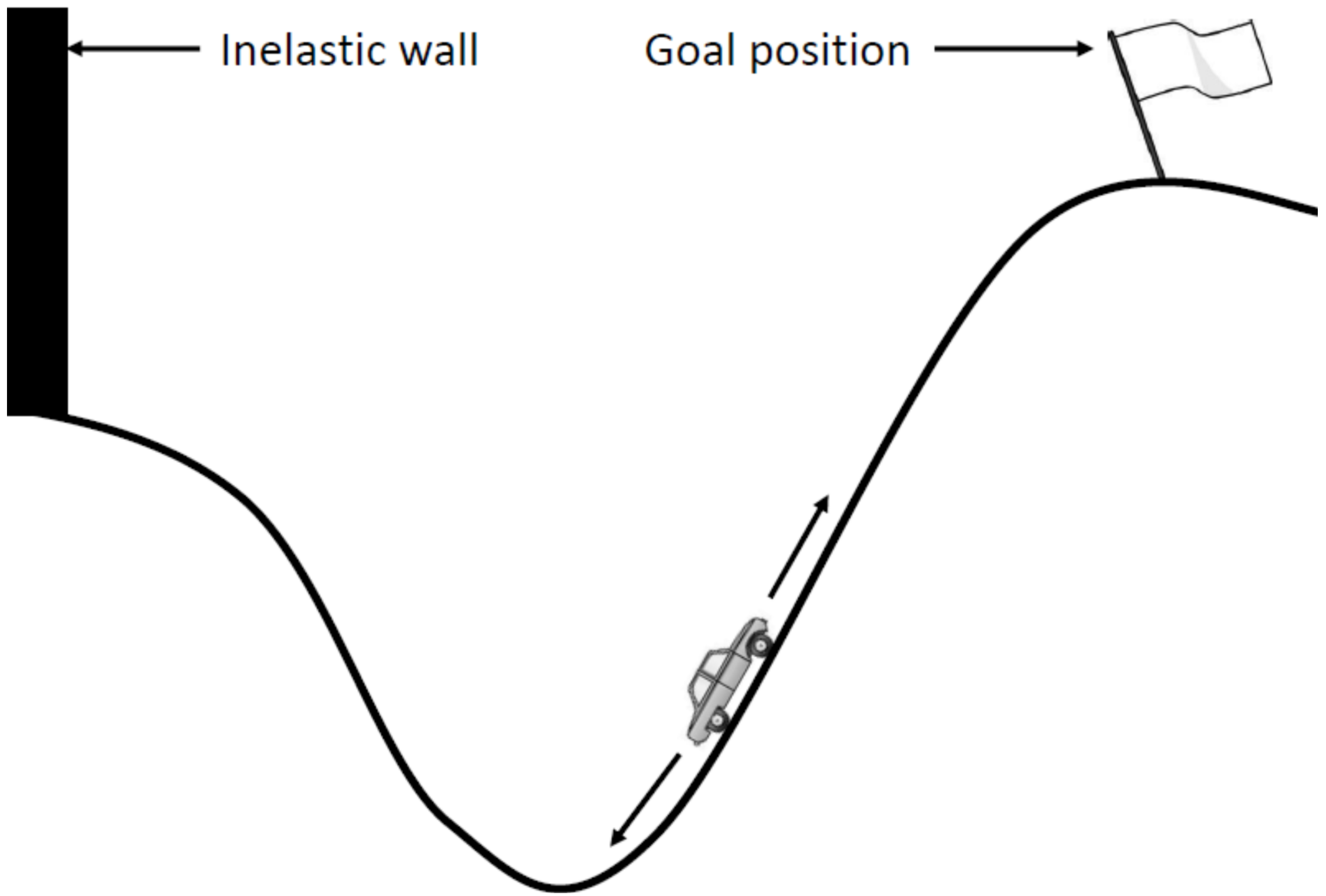
Example: Decision Support Systems



Example: Gridworld



Example: Mountain Car



Reinforcement Learning Algorithms

- Sarsa
- Q-learning
- LSPI
- Fitted Q Iteration
- REINFORCE
- Residual Gradient
- Continuous-Time Actor-Critic
- Value Gradient
- POWER
- PILCO
- LSPI
- PIPI
- Policy Gradient
- DQN
- Double Q-Learning
- Deterministic Policy Gradient
- NAC-LSTD
- INAC
- Average-Reward INAC
- Unbiased NAC
- Projected NAC
- Risk-sensitive policy gradient
- Natural Sarsa
- PGPE / PGPE-SyS
- True Online
- GTD/TDC
- ARP
- GPTD
- Auto-Actor Auto-Critic
- Approximate Value Iteration

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

Notation

- s: State
- a: Action
- S_t, A_t: State, and action at time t
- $\pi(a|s) = \Pr(A_t = a|S_t = s)$
- $\tau = (S_0, A_0, S_1, \dots, S_L, A_L)$
- $G(\tau) \in [0,1]$
- $\rho(\pi) = \mathbf{E}[G(\tau)|\tau \sim \pi]$

Two Goals:

- High confidence off-policy evaluation (HCOPE)

Historical Data, $\mathcal{D}$
Proposed Policy, π_e
Confidence Level, δ

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

- Safe Policy Improvement (SPI)

Historical Data, $\mathcal{D}$
Performance baseline, ρ_-
Confidence Level, δ

$\longrightarrow$ An improved* policy, π

*The probability that π's performance is below ρ_- is at most δ

High Confidence Off-Policy Evaluation

Historical Data, $\mathcal{D}$

Proposed Policy, π_e

Confidence Level, δ

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

- Historical data: $\mathcal{D} = \{(\tau_i, \pi_i) \colon \tau_i \sim \pi_i\}_{i=1}^{n}$
- Evaluation policy, π_e
- Confidence level, δ
- Compute $\mathrm{HCOPE}(\pi_e | \mathcal{D}, \delta)$ such that

$$\Pr(\rho(\pi_e) \geq \mathrm{HCOPE}(\pi_e | \mathcal{D}, \delta)) \geq 1 - \delta$$

Importance Sampling

- We would like to estimate
$$\theta := \mathbf{E}[f(x) | x \sim p]$$
- Monte Carlo estimator:
  - Sample $X_1, \ldots X_n$ from p and set:
$$\hat{\theta}_n := \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$
  - Nice properties
    - The Monte Carlo estimator is strongly consistent:
$$\hat{\theta}_n \xrightarrow{a.s.} \theta$$
    - The Monte Carlo estimator is unbiased for all $n \geq 1$:
$$\mathbf{E}\left[\hat{\theta}_n\right] = \theta$$

Importance Sampling

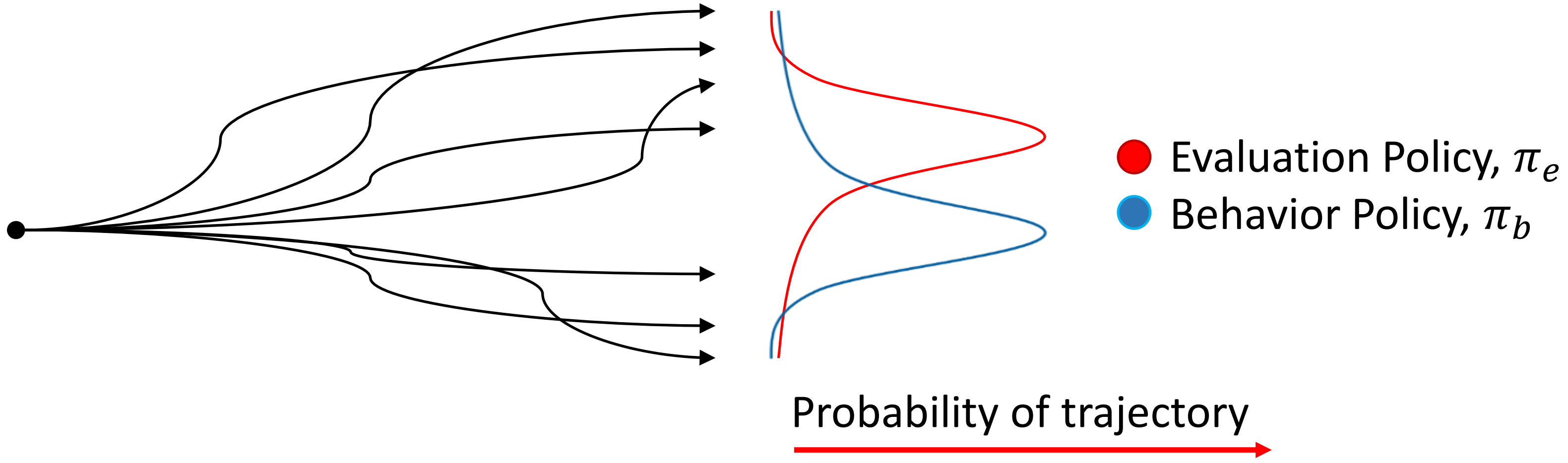
- We would like to estimate
$$\theta := \mathbf{E}[f(x)|x \sim p]$$
- … but we can only sample from a distribution, q, not p.
- Assume: if $q(x) = 0$ then $f(x)p(x) = 0.$ Then:

Importance Sampling

- We would like to estimate
$$\theta := \mathbf{E}[f(x)|x \sim p]$$
- Importance sampling estimator:
  - Sample $X_1, \ldots, X_n$ from q and set:
$$\hat{\theta}_n := \frac{1}{n}\sum_{i=1}^{n}\frac{p(X_i)}{q(X_i)}f(X_i)$$
  - Nice properties (under mild assumptions)
    - The importance sampling estimator is strongly consistent:
$$\hat{\theta}_n \xrightarrow{a.s.} \theta$$
    - The importance sampling estimator is unbiased for all $n \geq 1$:
$$\mathbf{E}[\hat{\theta}_n] = \theta$$

Importance Sampling

$$\rho(\pi_e) = \mathbf{E}_{\tau\sim\pi_e}[G(\tau)] = \mathbf{E}_{\tau\sim\pi_b}\left[\frac{\Pr(\tau|\pi_e)}{\Pr(\tau|\pi_b)}G(\tau)\right]$$



Importance Sampling for Reinforcement Learning (D. Precup, R. S. Sutton, and S. Dasgupta, 2001)

- $\rho(\pi_e) = \mathbf{E}_{\tau\sim\pi_e}[G(\tau)] = \mathbf{E}_{\tau\sim\pi_b}\left[\frac{\Pr(\tau|\pi_e)}{\Pr(\tau|\pi_b)}G(\tau)\right]$
- $\frac{\Pr(\tau|\pi_e)}{\Pr(\tau|\pi_b)}G(\tau) = \frac{\prod_{t=0}^{L}\Pr(S_t|\text{past})\Pr(A_t|\text{past},\pi_e)}{\prod_{t=0}^{L}\Pr(S_t|\text{past})\Pr(A_t|\text{past},\pi_b)}G(\tau)$

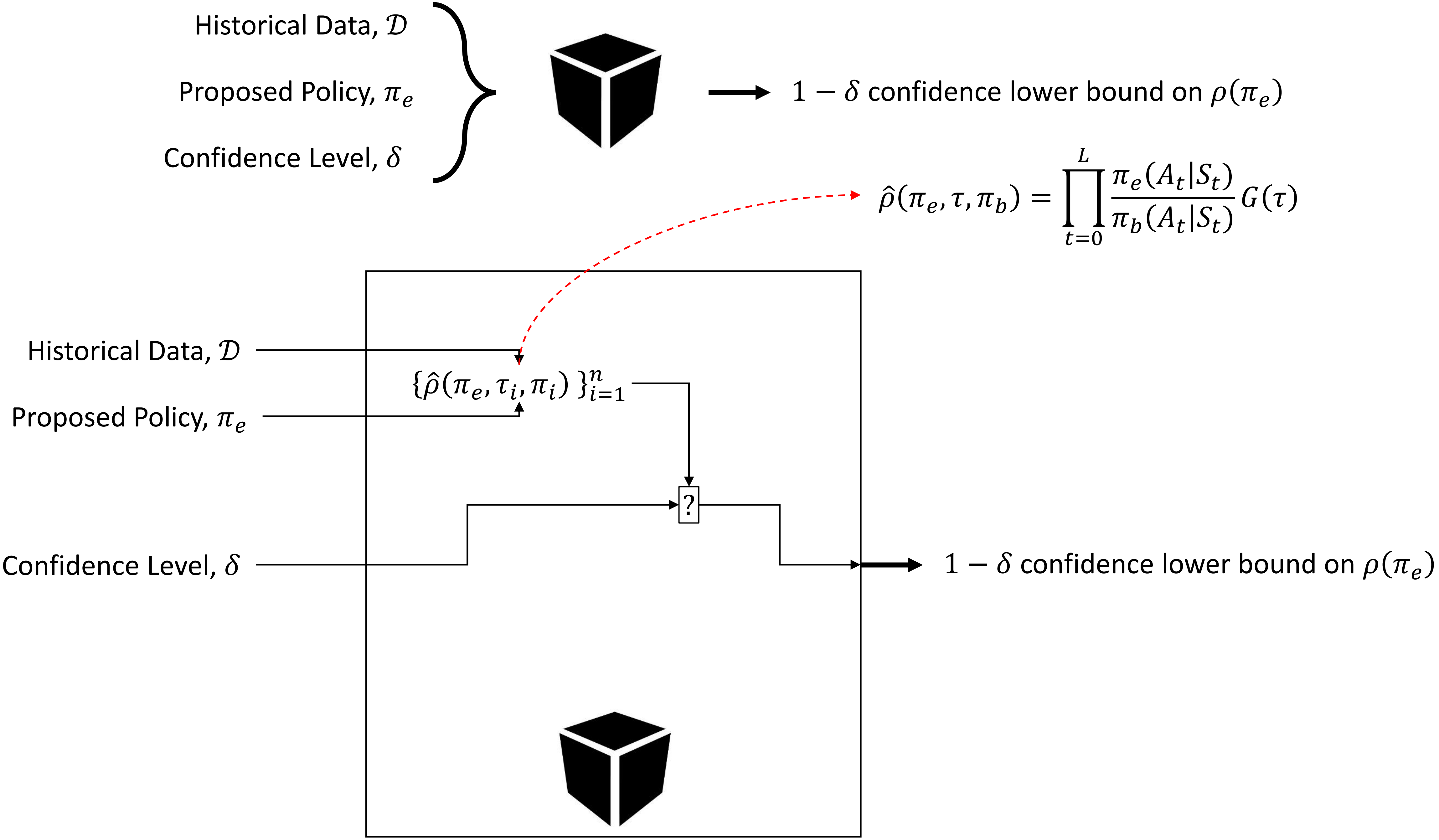
$$=\frac{\prod_{t=0}^{L}\Pr(A_t|\text{past},\pi_e)}{\prod_{t=0}^{L}\Pr(A_t|\text{past},\pi_b)}G(\tau)$$

$$=\frac{\prod_{t=0}^{L}\pi_e(A_t|S_t)}{\prod_{t=0}^{L}\pi_b(A_t|S_t)}G(\tau)$$

- $\hat{\rho}(\pi_e,\tau,\pi_b) = \prod_{t=0}^{L}\frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)}G(\tau)$

Per-Decision Importance Sampling

- Use importance sampling to estimate each R_t.
  - Still and unbiased and strongly consistent estimator of $\rho(\pi_e)$.
  - Often has lower variance than ordinary importance sampling.



Historical Data, $\mathcal{D}$
Proposed Policy, π_e
Confidence Level, δ
$1-\delta$ confidence lower bound on $\rho(\pi_e)$
$\hat{\rho}(\pi_e, \tau, \pi_b) = \prod_{t=0}^{L} \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)} G(\tau)$
Historical Data, $\mathcal{D}$
Proposed Policy, π_e
$\{\hat{\rho}(\pi_e, \tau_i, \pi_i)\}_{i=1}^{n}$
?
Confidence Level, δ
$1-\delta$ confidence lower bound on $\rho(\pi_e)$

Chernoff-Hoeffding Inequality

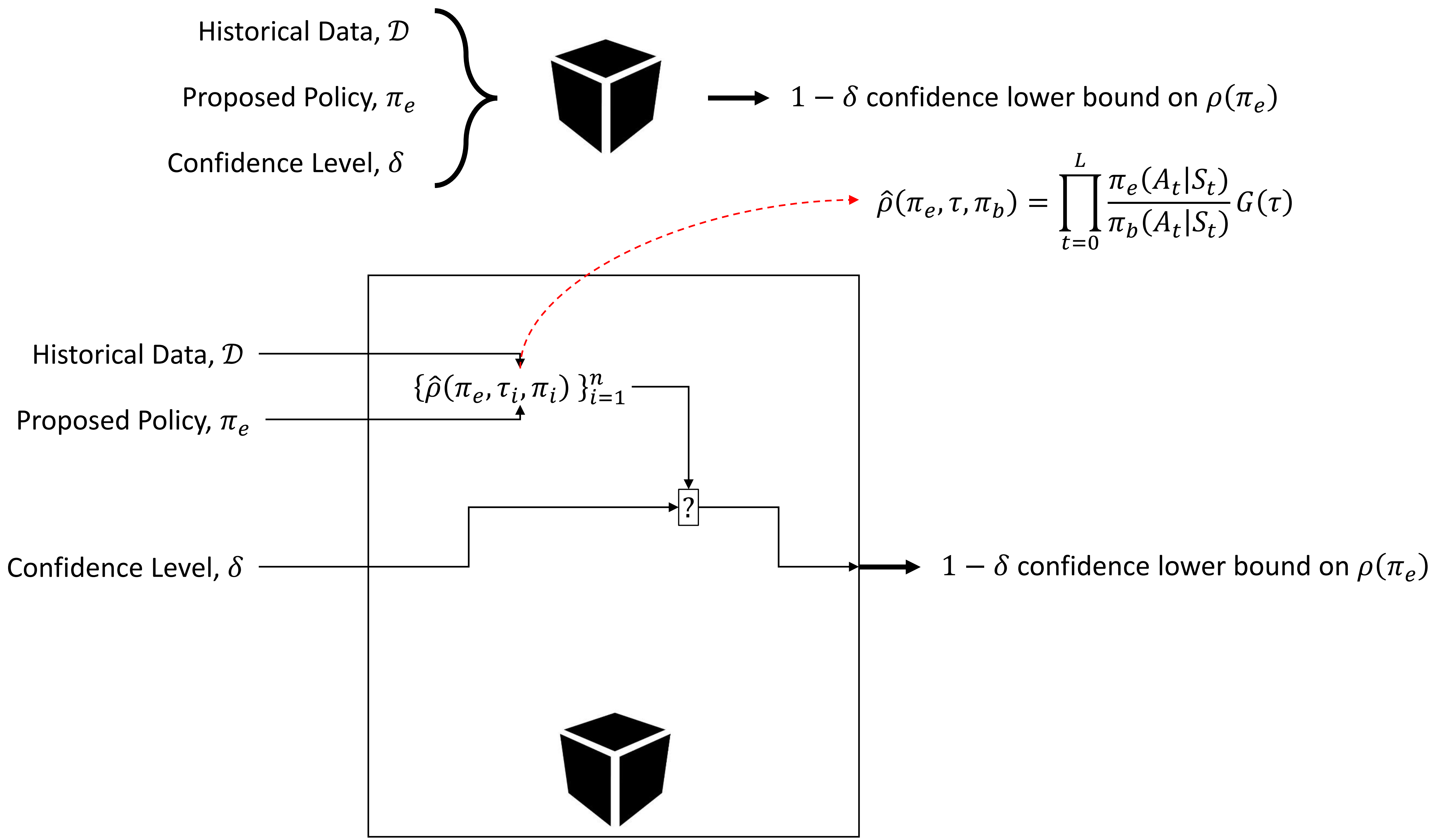
- Let $X_1, \dots, X_n$ be n independent identically distributed random variables such that:
  - $\mathrm{X_i} \in [0, b]$
- Then with probability at least $1 - \delta$:

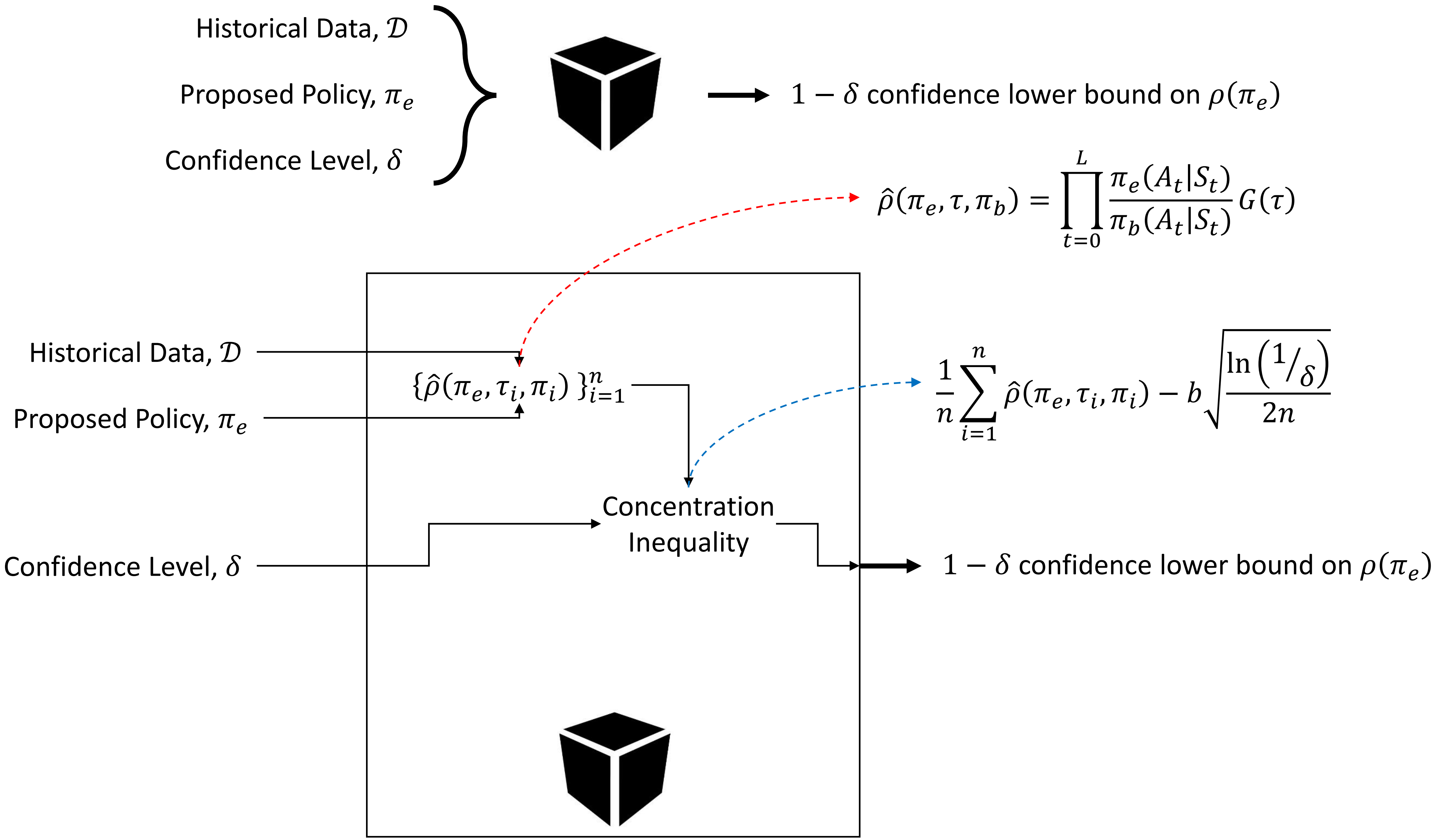
$$E[X_i] \geq \frac{1}{n}\sum_{i=1}^{n} X_i - b\sqrt{\frac{\ln\left({}^{1}/_{\delta}\right)}{2n}}$$

$$\rho(\pi_e) = E[\hat{\rho}(\pi_e, \tau_i, \pi_i)] \geq \frac{1}{n}\sum_{i=1}^{n} \hat{\rho}(\pi_e, \tau_i, \pi_i) - b\sqrt{\frac{\ln\left({}^{1}/_{\delta}\right)}{2n}}$$

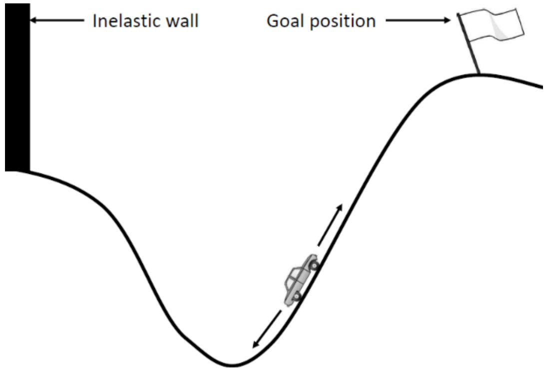
With probability at least $1 - \delta$:

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


Historical Data, $\mathcal{D}$
Proposed Policy, π_e
Confidence Level, δ
$1-\delta$ confidence lower bound on $\rho(\pi_e)$
$\hat{\rho}(\pi_e,\tau,\pi_b) = \prod_{t=0}^{L} \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)} G(\tau)$
Historical Data, $\mathcal{D}$
Proposed Policy, π_e
$\{\hat{\rho}(\pi_e,\tau_i,\pi_i)\}_{i=1}^{n}$
?
Confidence Level, δ
$1-\delta$ confidence lower bound on $\rho(\pi_e)$


Historical Data, $\mathcal{D}$
Proposed Policy, π_e
Confidence Level, δ
$1-\delta$ confidence lower bound on $\rho(\pi_e)$
$\hat{\rho}(\pi_e,\tau,\pi_b) = \prod_{t=0}^{L} \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)} G(\tau)$
Historical Data, $\mathcal{D}$
Proposed Policy, π_e
$\{\hat{\rho}(\pi_e,\tau_i,\pi_i)\}_{i=1}^{n}$
$\frac{1}{n}\sum_{i=1}^{n}\hat{\rho}(\pi_e,\tau_i,\pi_i) - b\sqrt{\frac{\ln(1/\delta)}{2n}}$
Concentration Inequality
Confidence Level, δ
$1-\delta$ confidence lower bound on $\rho(\pi_e)$

Example: Mountain Car



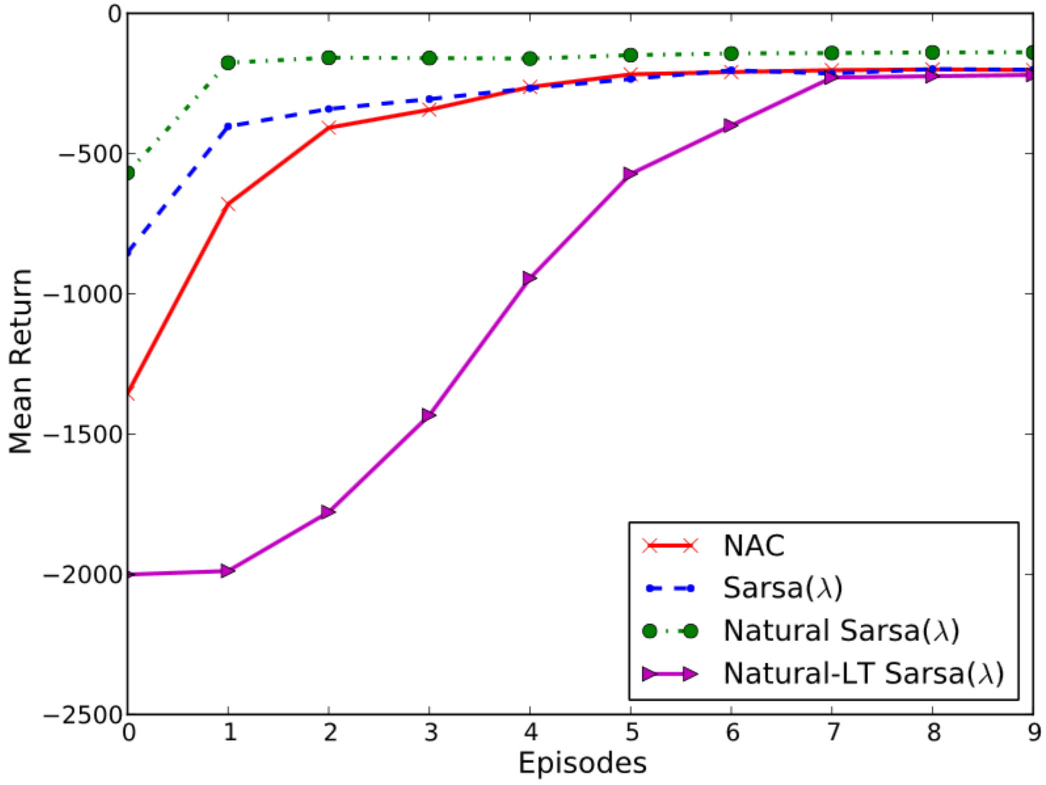
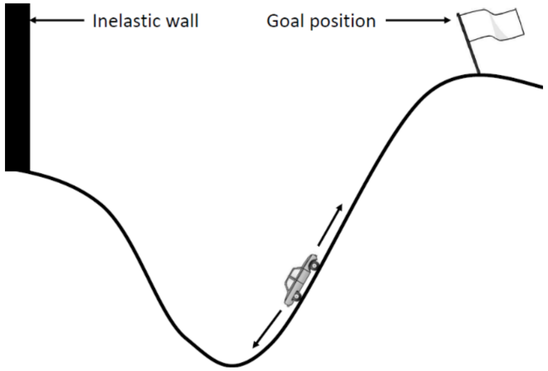


Figure 3: Mountain Car (Sarsa(λ))

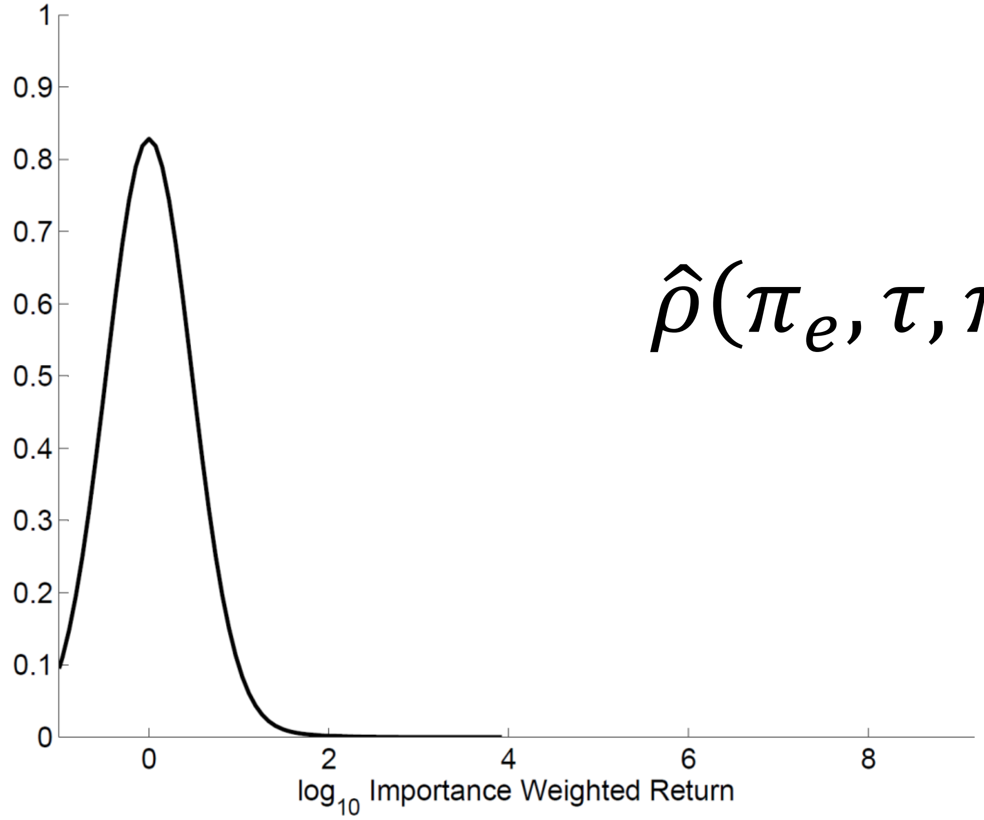
Example: Mountain Car



- 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?



$$\hat{\rho}(\pi_e, \tau, \pi_b) = \prod_{t=0}^{L} \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)} G(\tau)$$

What went wrong?

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

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

Largest observed importance weighted return: 316.

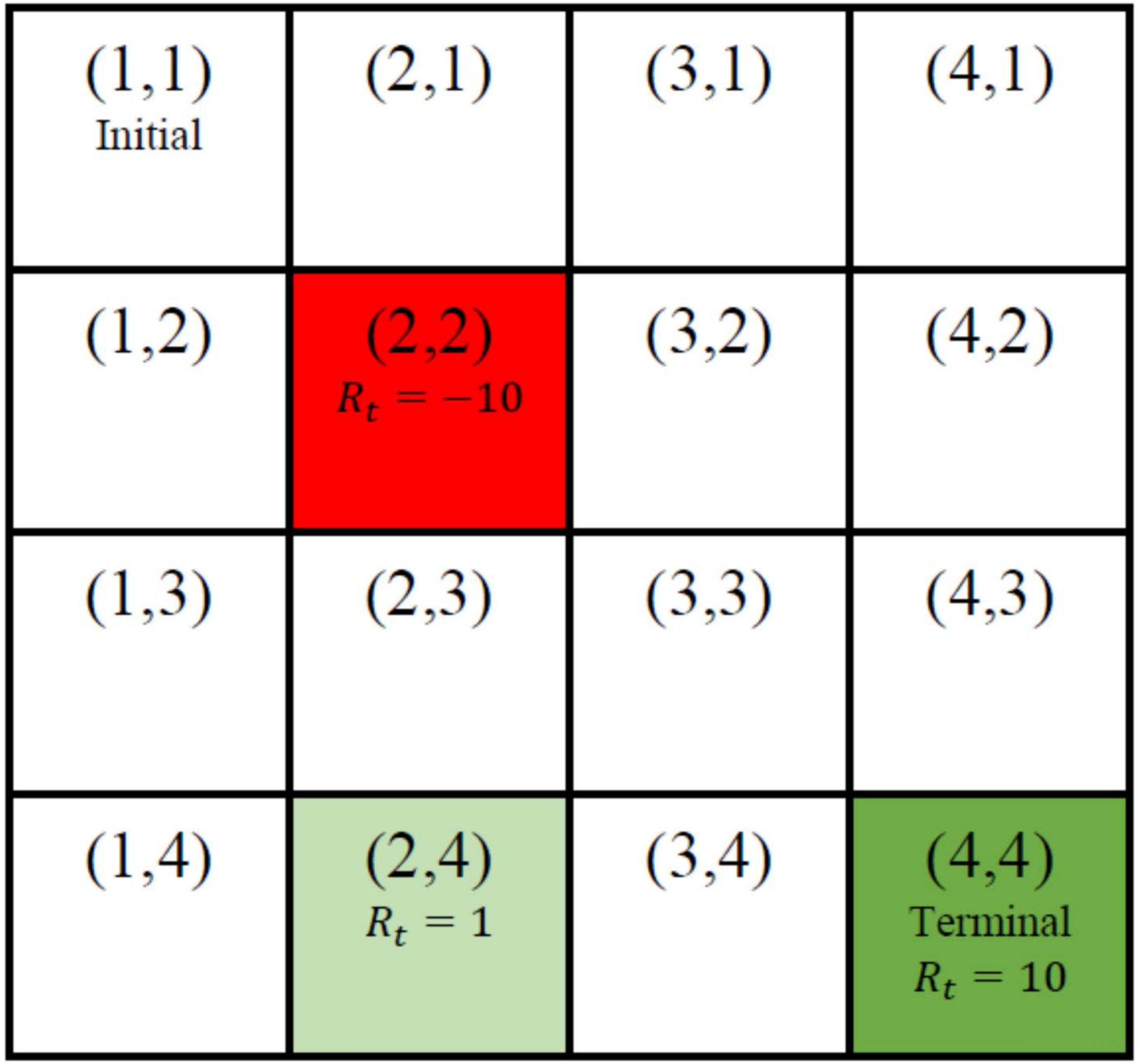
Another problem:

$$\hat{\rho}(\pi_e, \tau, \pi_b) = \prod_{t=0}^{L} \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|\,S_t)} G(\tau)$$

- One behavior policy
  - Independent and identically distributed
- More than one behavior policy
  - Independent

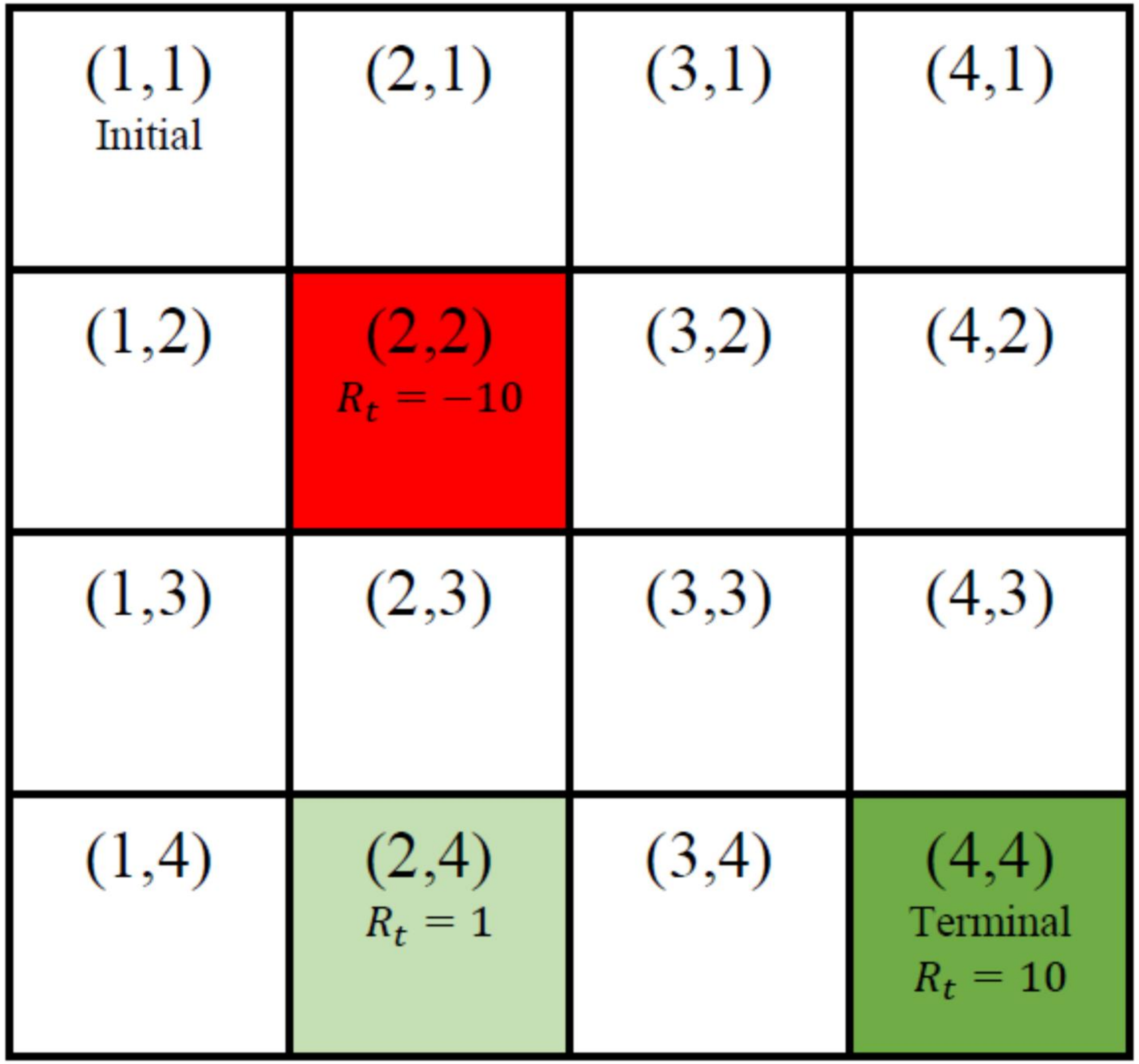
Conservative Policy Iteration (S. Kakade and J. Langford, 2002)

- $\approx 1{,}000{,}000$ trajectories for a single policy improvement.



PAC-RL (T. Lattimore and M. Hutter, 2012)

- $\approx 10^{17}$ time steps to guarantee convergence to a near-optimal policy.



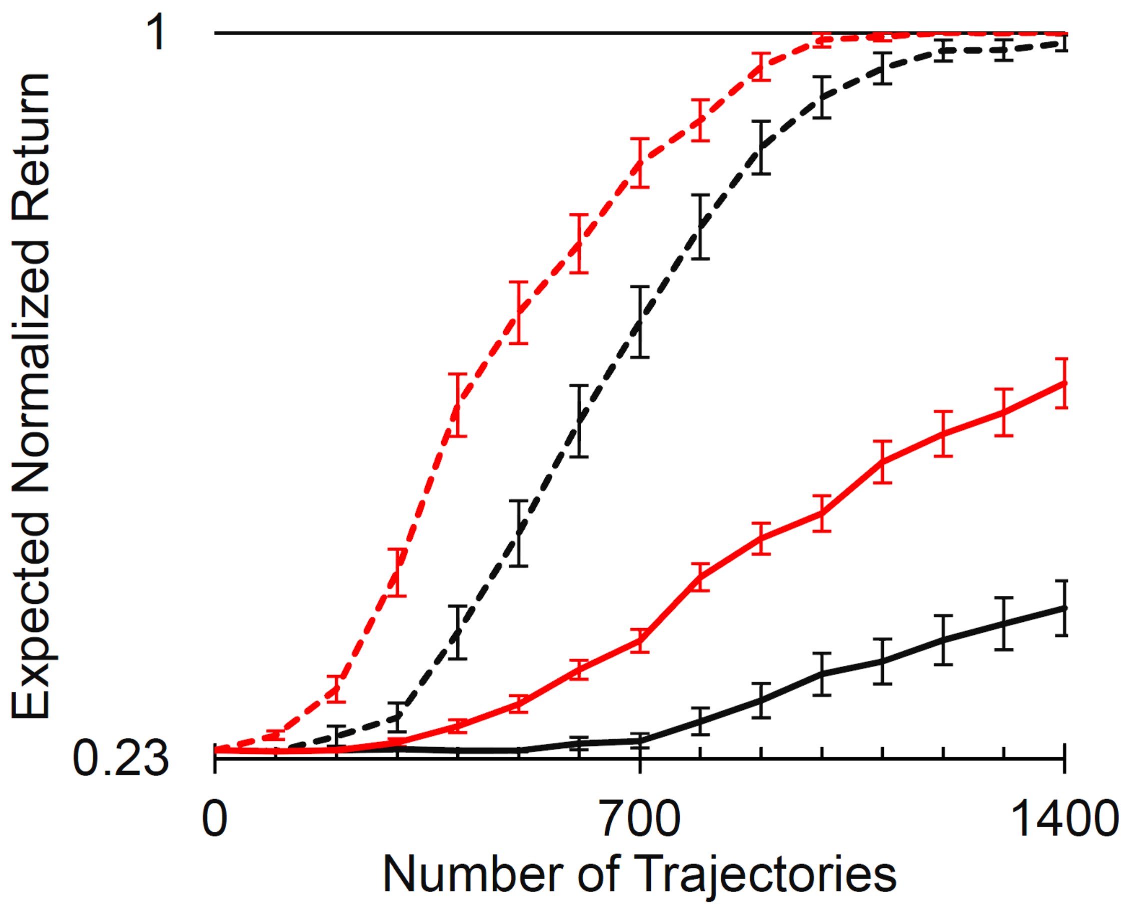
Thesis

High Confidence Off-Policy Evaluation (HCOPE)

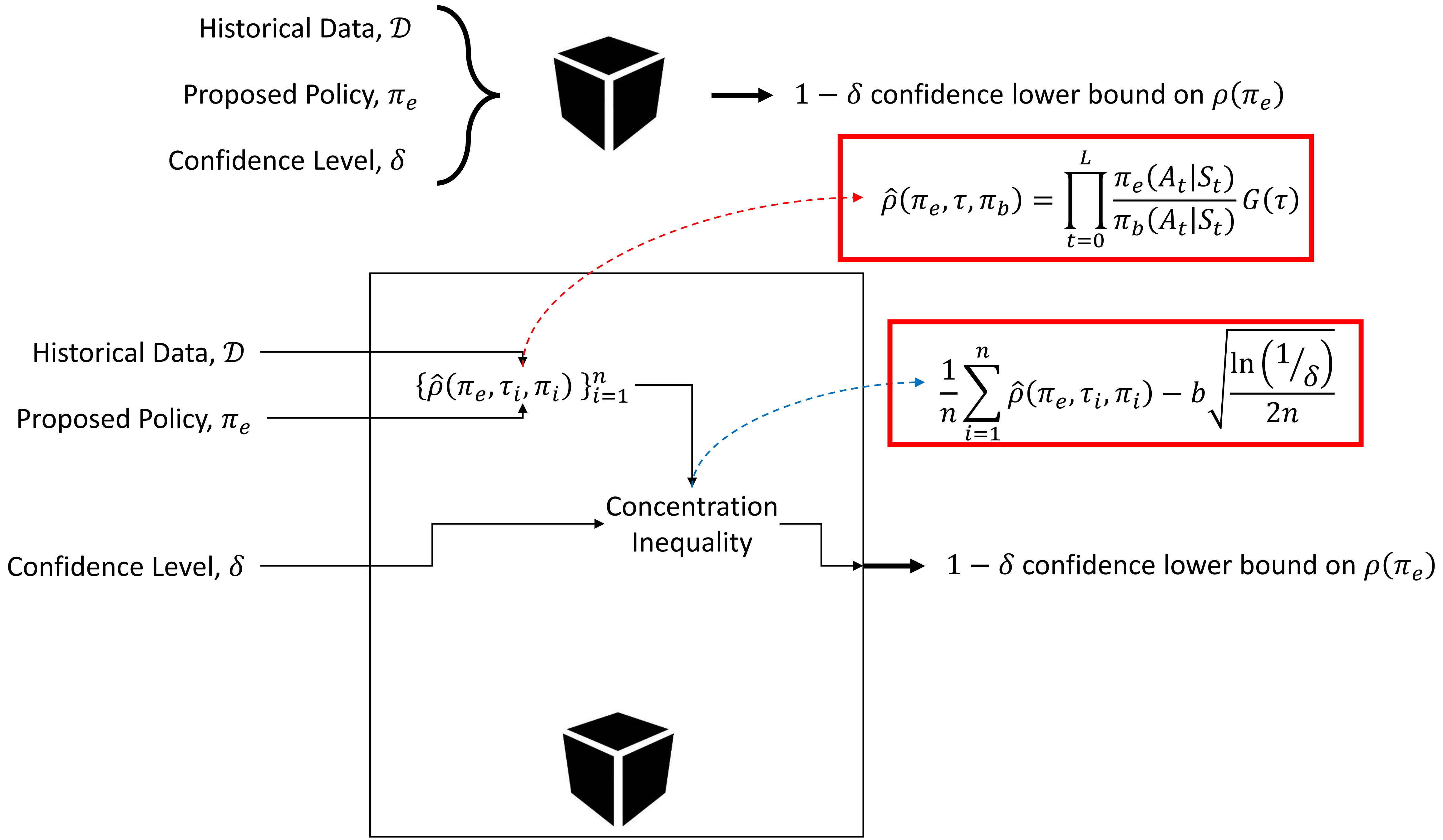
and

Safe Policy Improvement (SPI)

are tractable using a practical amount of data.



1
0.23
Expected Normalized Return
0
700
1400
Number of Trajectories


Historical Data, $\mathcal{D}$
Proposed Policy, π_e
Confidence Level, δ
$1-\delta$ confidence lower bound on $\rho(\pi_e)$
$\hat{\rho}(\pi_e,\tau,\pi_b) = \prod_{t=0}^{L} \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)} G(\tau)$
Historical Data, $\mathcal{D}$
$\{\hat{\rho}(\pi_e,\tau_i,\pi_i)\}_{i=1}^{n}$
Proposed Policy, π_e
$\frac{1}{n}\sum_{i=1}^{n} \hat{\rho}(\pi_e,\tau_i,\pi_i) - b\sqrt{\frac{\ln\left({}^{1}/_{\delta}\right)}{2n}}$
Concentration Inequality
Confidence Level, δ
$1-\delta$ confidence lower bound on $\rho(\pi_e)$

Name	Direct Dependence on b	Identically Distributed Only	Exact or Approximate	Reference	Notes
CH	$\Theta\left(\frac{b}{\sqrt{n}}\right)$	No	Exact	(Massart, 2007)	None
MPeB	$\Theta\left(\frac{b}{n}\right)$	No	Exact	(Maurer and Pontil, 2009, Theorem 11)	Requires all random variables to have the same range.
AM	None	Yes	Exact	(Anderson, 1969, Massart, 1990)	Depends on the largest observed sample. Loose for distributions without heavy tails.
BM	$\Theta\left(\frac{b}{\sqrt{n}}\right)$	Yes	Exact	(Bubeck et al., 2012)	None.
CUT	None	No	Exact	Theorem 23	None.

Theorem 17 (Chernoff-Hoeffding (CH) Inequality)**.** *Let* $\{X_i\}_{i=1}^n$ *be* n *independent random variables such that* $\Pr(X_i \in [a_i, b_i]) = 1$*, for all* $i \in \{1, \ldots, n\}$*, where all* $a_i \in \mathbb{R}$ *and* $b_i \in \mathbb{R}$*. Then*

$$\Pr\left(\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \geq \frac{1}{n}\sum_{i=1}^{n} X_i - \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)\sum_{i=1}^{n}(b_i - a_i)^2}{2n^2}}\right) \geq 1 - \delta. \tag{4.6}$$

Theorem 18 (Maurer and Pontil's Empirical Bernstein (MPeB) Inequality). *Let $\{X_i\}_{i=1}^{n}$ be n independent random variables such that $\Pr\left(X_i \in [a,b]\right) = 1$, for all $i \in \{1,\ldots,n\}$, where $a \in \mathbb{R}$ and $b \in \mathbb{R}$. Then*

$$\Pr\left(\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \geq \underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i}_{\textit{sample mean}} - \frac{7(b-a)\ln\left(\frac{2}{\delta}\right)}{3(n-1)} - \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{n}\underbrace{\frac{1}{n(n-1)}\sum_{i,j=1}^{n}\frac{\left(X_i - X_j\right)^2}{2}}_{\textit{sample variance}}}\right) \geq 1-\delta.$$

Theorem 19 (Anderson and Massart's (AM) Inequality). *Let $\{X_i\}_{i=1}^n$ be n independent and identically distributed random variables such that $X_i \geq a$, for all $i \in \{1, \ldots, n\}$, where $a \in \mathbb{R}$. Then*

$$\Pr\left(\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \geq Z_n - \sum_{i=0}^{n-1}(Z_{i+1} - Z_i)\min\left\{1, \frac{i}{n} + \sqrt{\frac{\ln(2/\delta)}{2n}}\right\}\right) \geq 1 - \delta,$$

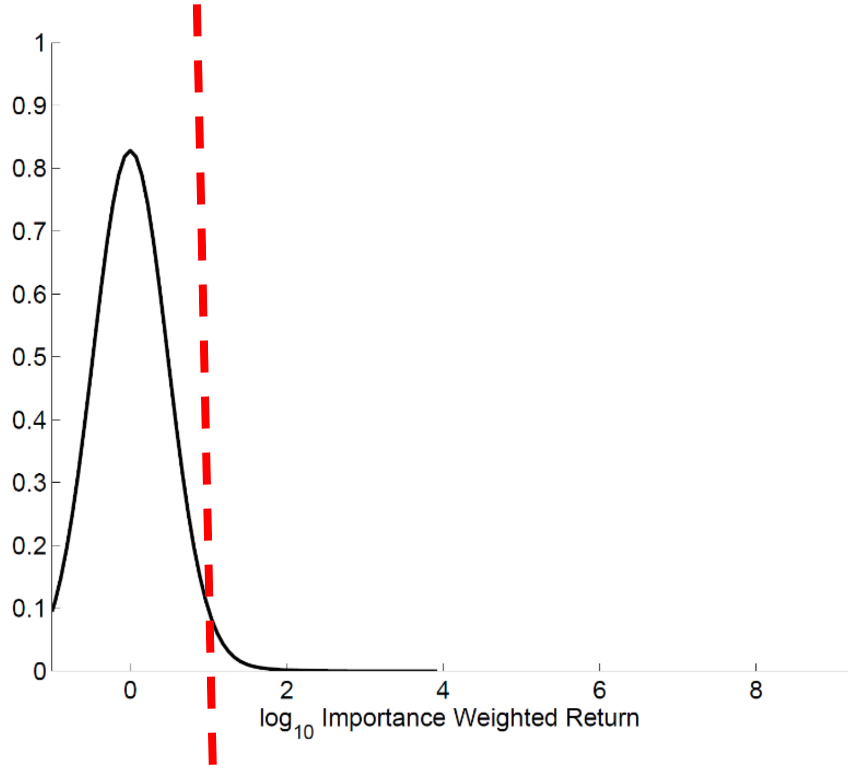
where $Z_0 = a$ and $\{Z_i\}_{i=1}^n$ are $\{X_i\}_{i=1}^n$, sorted such that $Z_1 \leq Z_2 \leq \ldots \leq Z_n$.

Extending Maurer’s Inequality

- First Key Idea:
  - Generalize: random variables with different ranges.
  - Specialize: random variables with the same mean.

Extending Maurer’s Inequality

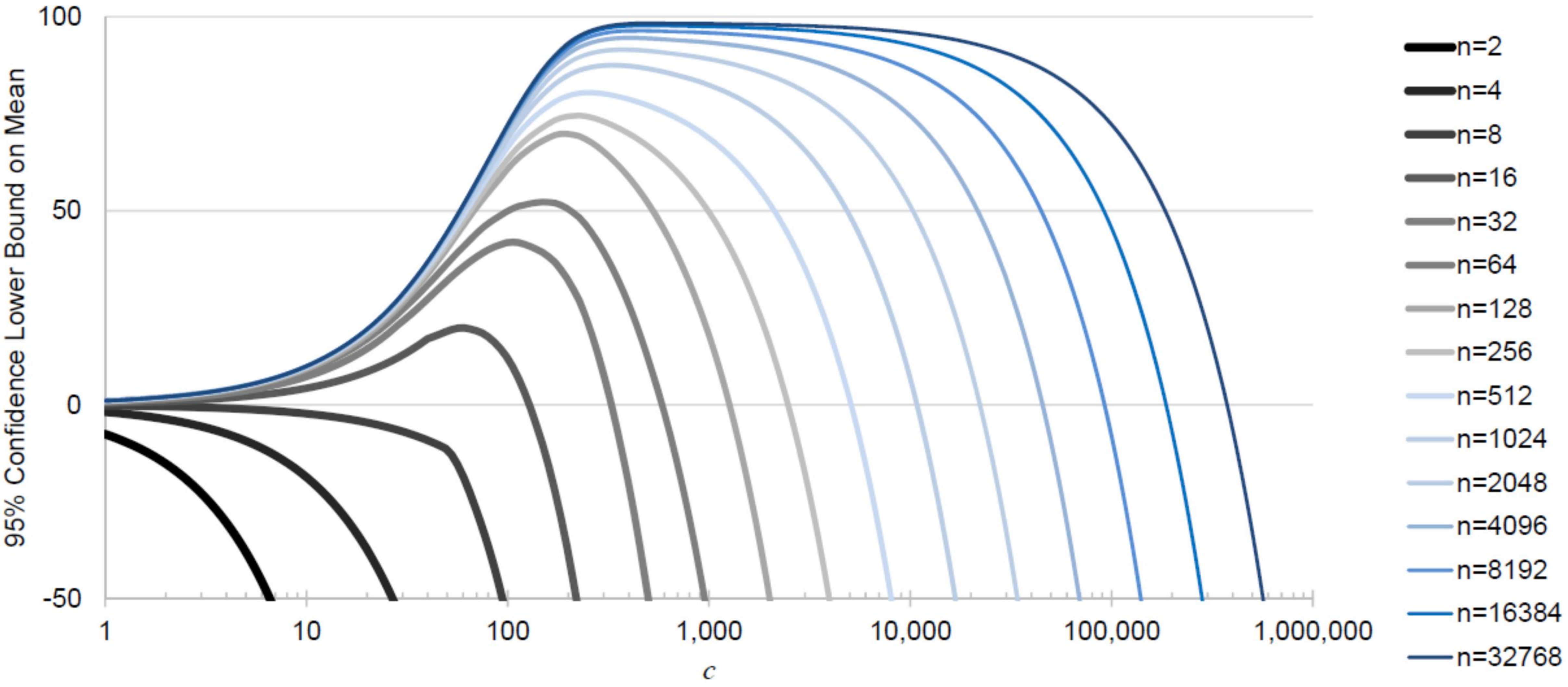
- Second Key Idea:
  - Removing the upper tail only decreases the expected value.



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

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

Tradeoff



95% Confidence Lower Bound on Mean
100
50
0
-50
1
10
100
1,000
10,000
100,000
1,000,000
c
n=2
n=4
n=8
n=16
n=32
n=64
n=128
n=256
n=512
n=1024
n=2048
n=4096
n=8192
n=16384
n=32768

Threshold Optimization

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

Given n samples, $\mathcal{X} \coloneqq \{X_i\}_{i=1}^{n}$, predict what the lower bound would be if computed from m samples, rather than n.

$$\widehat{\mathrm{CUT}}(\mathcal{X}, \delta, c, m) \coloneqq \underbrace{\frac{1}{n}\sum_{i=1}^{n} \min\{X_i, c\}}_{\text{sample mean of } \mathcal{X} \text{ (after being collapsed)}} - \frac{7c\ln(2/\delta)}{3(m-1)} \tag{4.15}$$
$$- \sqrt{\frac{\ln(2/\delta)}{m} \underbrace{\frac{2}{n(n-1)} \left(n\sum_{i=1}^{n} (\min\{X_i, c\})^2 - \left(\sum_{i=1}^{n} \min\{X_i, c\}\right)^2 \right)}_{\text{sample variance of } \mathcal{X} \text{ (after being collapsed)}}},$$

Algorithm 4.11: $\text{CUT}(X_1, \ldots, X_n, \delta)$: Uses the CUT inequality to return a $1-\delta$ confidence lower bound on $\mathbf{E}[\frac{1}{n}\sum_{i=1}^{n} X_i]$.

Constants: This algorithm has a real-valued hyperparameter, $c_{\min} \geq 0$, which is the smallest allowed threshold. It should be chosen based on the application. For HCOPE we use $c_{\min} = 1$.

Assumes: The X_i are independent random variables such that $\Pr(X_i \geq 0) = 1$ for all $i \in \{1, \ldots, n\}$.

Randomly select 1/5 of the X_i and place them in a set $\mathcal{X}_{\text{pre}}$ and the remainder in $\mathcal{X}_{\text{post}}$;

`// Optimize threshold using` $\mathcal{X}_{\text{pre}}$

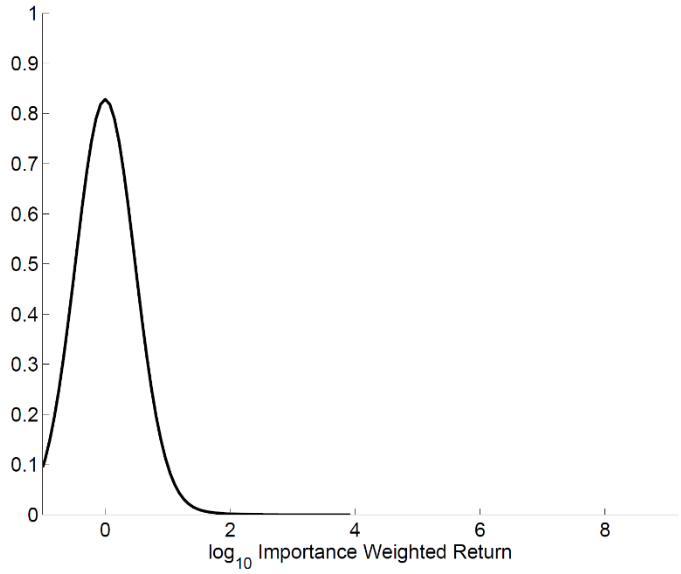
$c^\star \in \arg\max_{c \in [1,\infty]} \widehat{\text{CUT}}(\mathcal{X}_{\text{pre}}, \delta, c, |\mathcal{X}_{\text{post}}|)$; `//` $\widehat{\texttt{CUT}}$ `is defined in` (4.15)

$c^* = \max\{c_{\min}, c^*\}$; `// Do not let` c^* `become too small`

`// Compute lower bound using optimized threshold,` c^* `and` $\mathcal{X}_{\text{post}}$

return $\widehat{\text{CUT}}(\mathcal{X}_{\text{post}}, \delta, c^\star, |\mathcal{X}_{\text{post}}|)$;

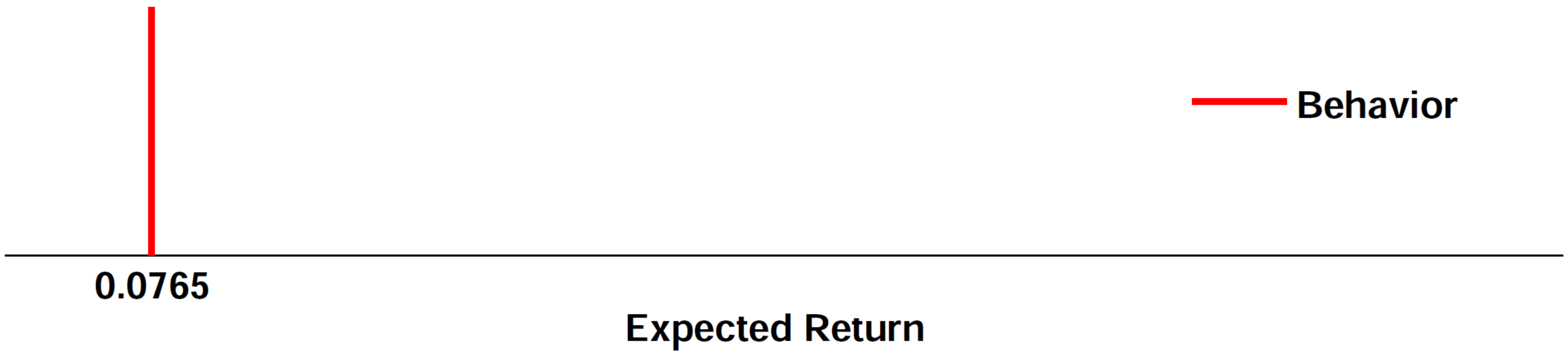
	CUT	Chernoff-Hoeffding	Maurer	Anderson	Bubeck et al.
95% Confidence lower bound on the mean	0.145	$-5{,}831{,}000$	$-129{,}703$	0.055	$-.046$



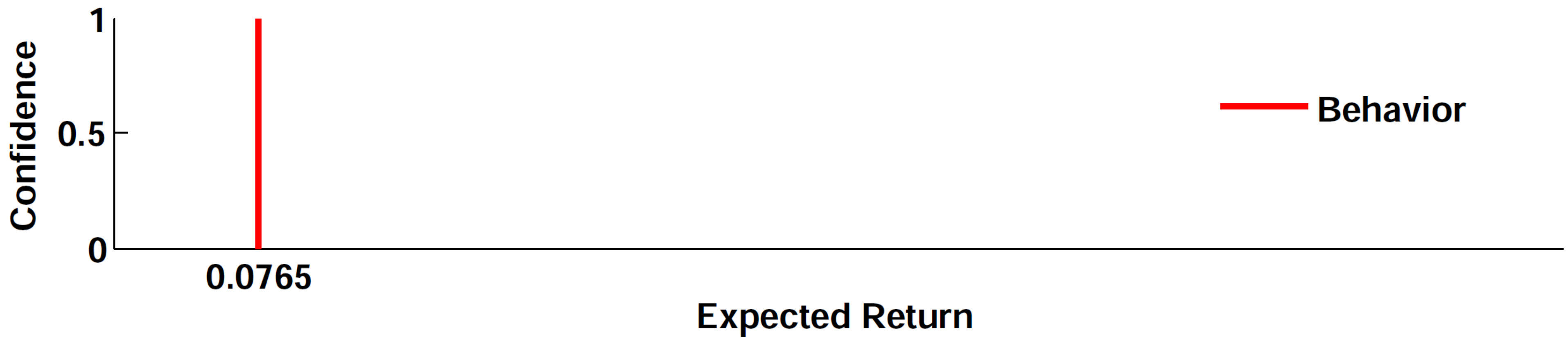
Digital Marketing Example

- 10 real-valued features
- Two groups of campaigns to choose between
- User interactions limited to $L = 10$
- Data collected from a Fortune 20 company
- Data was not used directly.

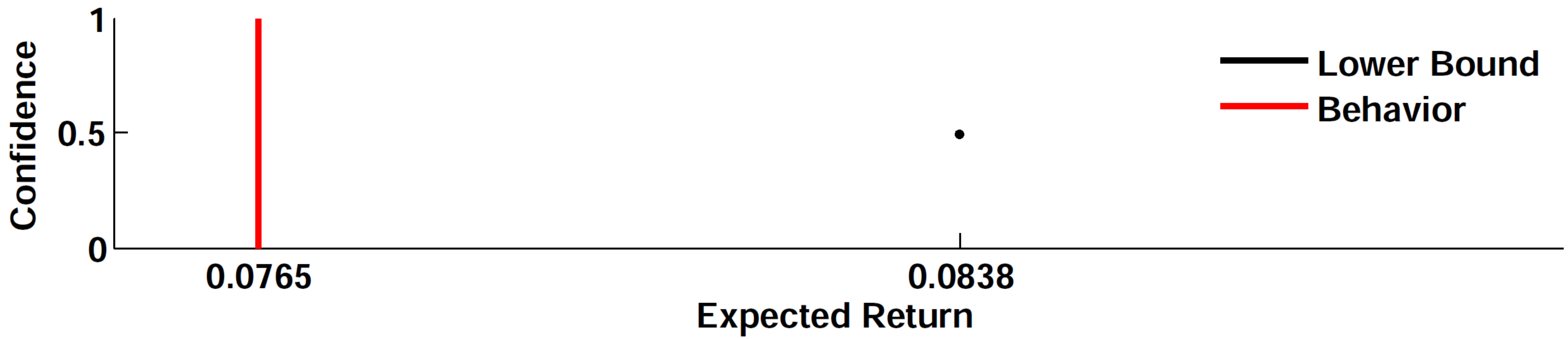
Example: Digital Marketing



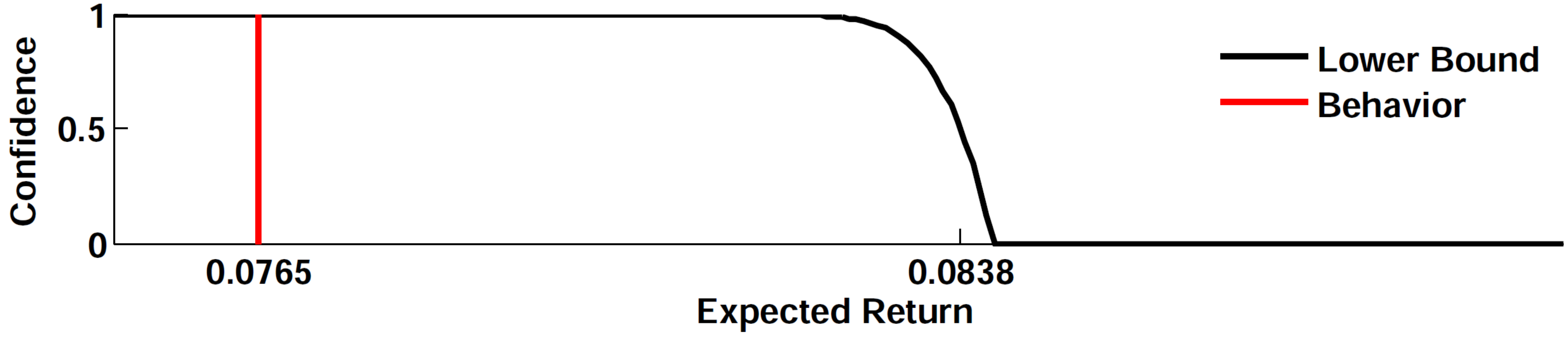
Example: Digital Marketing



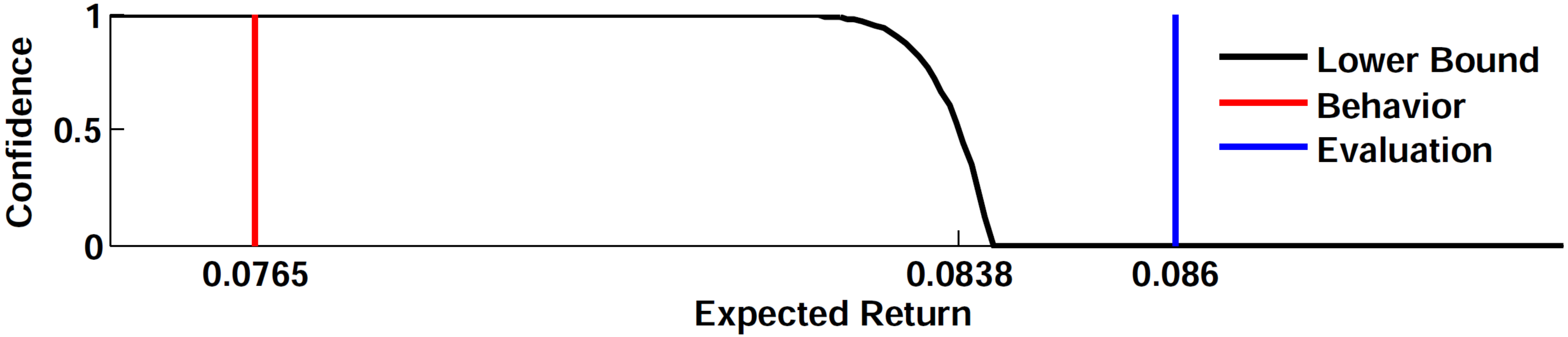
Example: Digital Marketing



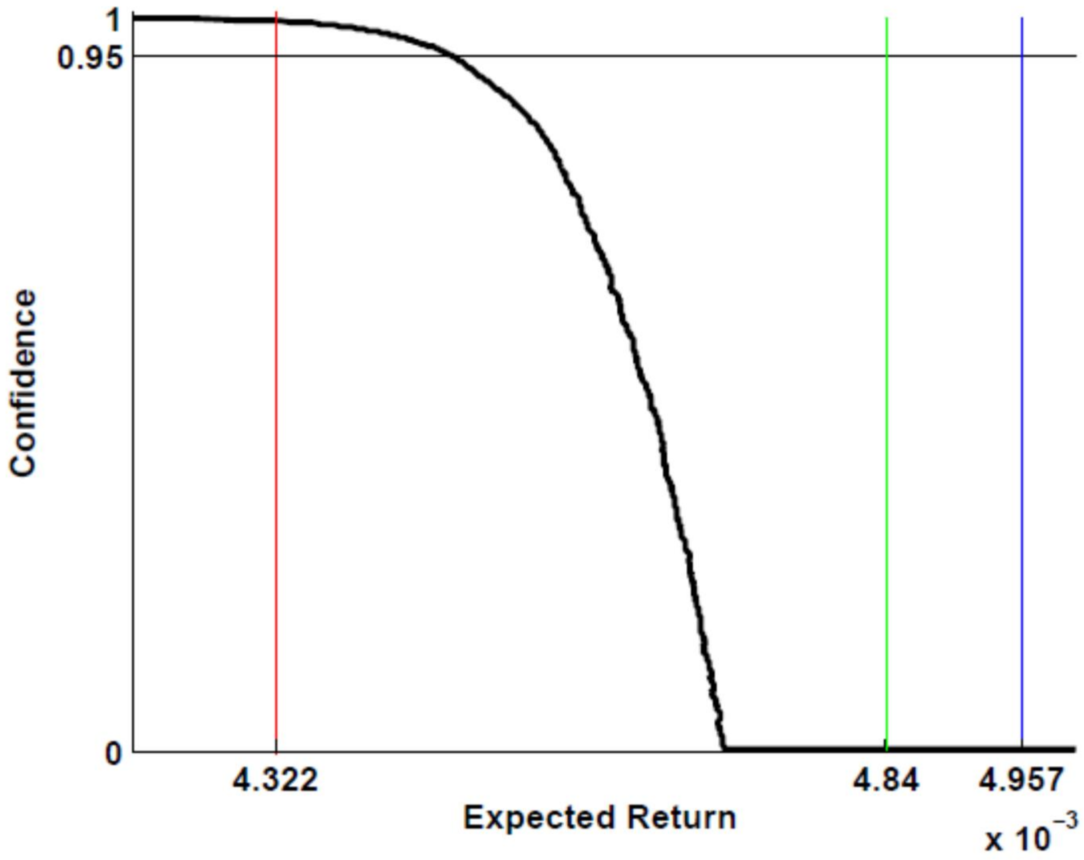
Example: Digital Marketing



Example: Digital Marketing



Example: Digital Marketing



We can now evaluate policies proposed by RL algorithms

Two Goals:

- High confidence off-policy evaluation (HCOPE)

Historical Data, $\mathcal{D}$

Proposed Policy, π_e

Confidence Level, δ

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

- Safe Policy Improvement (SPI)

Historical Data, $\mathcal{D}$

Performance baseline, ρ_-

Confidence Level, δ

$\longrightarrow$ An improved* policy, π

*The probability that π's performance is below ρ_- is at most δ

Safe Policy Improvement (SPI)

$$\mathrm{SPI}(\mathcal{D}, \rho_-, \delta) \in \{\textsc{No Solution Found}\} \cup \Pi$$

$$\Pr\left(\mathrm{SPI}(\mathcal{D}, \rho_-, \delta) \in \{\pi : \rho(\pi) < \rho_-\}\right) < \delta$$

$\boxed{\text{Exact}}$ $\boxed{\text{Approximate}}$

$< \delta$ $\approx \delta$

$\mathrm{SPI}(\mathcal{D}, \rho_{-}, \delta)$

1. Return NO SOLUTION FOUND.

$\mathrm{SPI}(\mathcal{D}, \rho_{-}, \delta)$

1. Return NO SOLUTION FOUND.

$$\Pr\left(\mathrm{SPI}(\mathcal{D}, \rho_{-}, \delta) \in \{\pi : \rho(\pi) < \rho_{-}\}\right) < \delta$$

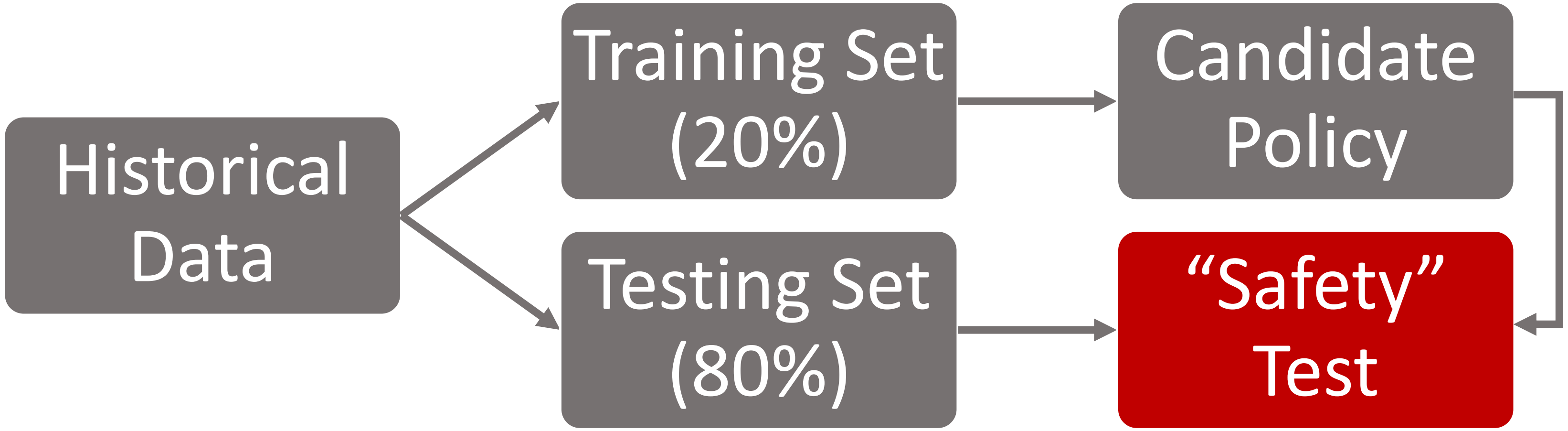
$\mathrm{SPI}(\mathcal{D}, \rho_{-}, \delta)$

1. Return No Solution Found.

$$\Pr\left(\mathrm{SPI}(\mathcal{D}, \rho_{-}, \delta) \in \{\pi : \rho(\pi) < \rho_{-}\}\right) < \delta$$

- Want a batch RL algorithm that
  - Satisfies this inequality
  - Often returns a policy

Thoughts?


Historical Data
Training Set (20%)
Testing Set (80%)
Candidate Policy
“Safety” Test

Algorithm 5.3: $\text{SPI}_{\ddagger}^{\dagger,\star}(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}, \delta, \rho_-)$: Use the historical data, partitioned into $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, to search for a safe policy (with $1-\delta$ confidence lower bound at least ρ_-). If none is found, then return No Solution Found. Although other $\dagger$ and $\ddagger$ could be used, we have only provided complete pseudocode for $(\dagger, \ddagger) \in \{(\text{NPDIS,CUT}), (\text{CWPDIS,BCa})\}$. We allow for $\star \in \{\text{None}, k\text{-fold}\}$. Assumption 1 is not required.

```
π_c ← GetCandidatePolicy‡^{†,⋆}(D_train, δ, ρ_−, |D_test|);
if HCOPE‡^†(π_c, D_test, δ) ≥ ρ_− then
    return π_c;
return No Solution Found
```

$\pi_c \leftarrow \textsc{GetCandidatePolicy}_{\ddagger}^{\dagger,\star}(\mathcal{D}_{\text{train}}, \delta, \rho_-, |\mathcal{D}_{\text{test}}|)$;

if $HCOPE_{\ddagger}^{\dagger}(\pi_c, \mathcal{D}_{test}, \delta) \geq \rho_-$ **then**

  return π_c;

return No Solution Found

$$f^{\dagger}_{\ddagger}(\pi, \mathcal{D}, \delta, \rho_-, m) := \begin{cases} \hat{\rho}(\pi|\mathcal{D}) & \text{if } \text{HCOPE}^{\dagger}_{\ddagger}(\pi, \mathcal{D}, \delta, m) \geq \rho_-, \\ \text{HCOPE}^{\dagger}_{\ddagger}(\pi, \mathcal{D}, \delta, m) & \text{otherwise.} \end{cases}$$

Algorithm 5.4: $\text{GetCandidatePolicy}^{\dagger,\text{None}}_{\ddagger}(\mathcal{D}_{\text{train}}, \delta, \rho_-, m)$: Use the historical data, partitioned into $\mathcal{D}_{\text{train}}$ to search for the candidate policy that is predicted to be safe and perform the best (or to be closest to safe if none are predicted to be safe). Although other † and ‡ could be used, we have only provided complete pseudocode for $(\dagger, \ddagger) \in \{(\text{NPDIS,CUT}), (\text{CWPDIS,BCa})\}$. Assumption 1 is not required.

1 **return** $\arg\max_{\pi} f^{\dagger}_{\ddagger}(\pi, \mathcal{D}_{\text{train}}, \delta, \rho_-, m)$;

Algorithm 5.5: $\textsc{GetCandidatePolicy}_{\ddagger}^{\dagger,k\text{-fold}}(\mathcal{D}_{\text{train}}, \delta, \rho_-, m)$: Use the historical data, $\mathcal{D}_{\text{train}}$, to search for the candidate policy that is predicted to be safe and perform the best (or to be closest to safe if none are predicted to be safe). Although other $\dagger$ and $\ddagger$ could be used, we have only provided complete pseudocode for $(\dagger, \ddagger) \in \{(\text{NPDIS},\text{CUT}), (\text{CWPDIS},\text{BCa})\}$. Assumption 1 is not required.

$\lambda^* \leftarrow \arg\max_{\lambda \in [0,1]} \textsc{CrossValidate}_{\ddagger}^{\dagger}(\lambda, \mathcal{D}_{\text{train}}, \delta, \rho_-, m)$;

$\pi^* \leftarrow \arg\max_{\pi} f_{\ddagger}^{\dagger}(\mu_{\lambda^*,\pi_0,\pi}, \mathcal{D}_{\text{train}}, \delta, \rho_-, m)$;

return $\mu_{\lambda^*,\pi_0,\pi^*}$;

Approximate Confidence Intervals

- What if we knew that the importance weighted returns were normally distributed?
  - One-sided Student's t-test.

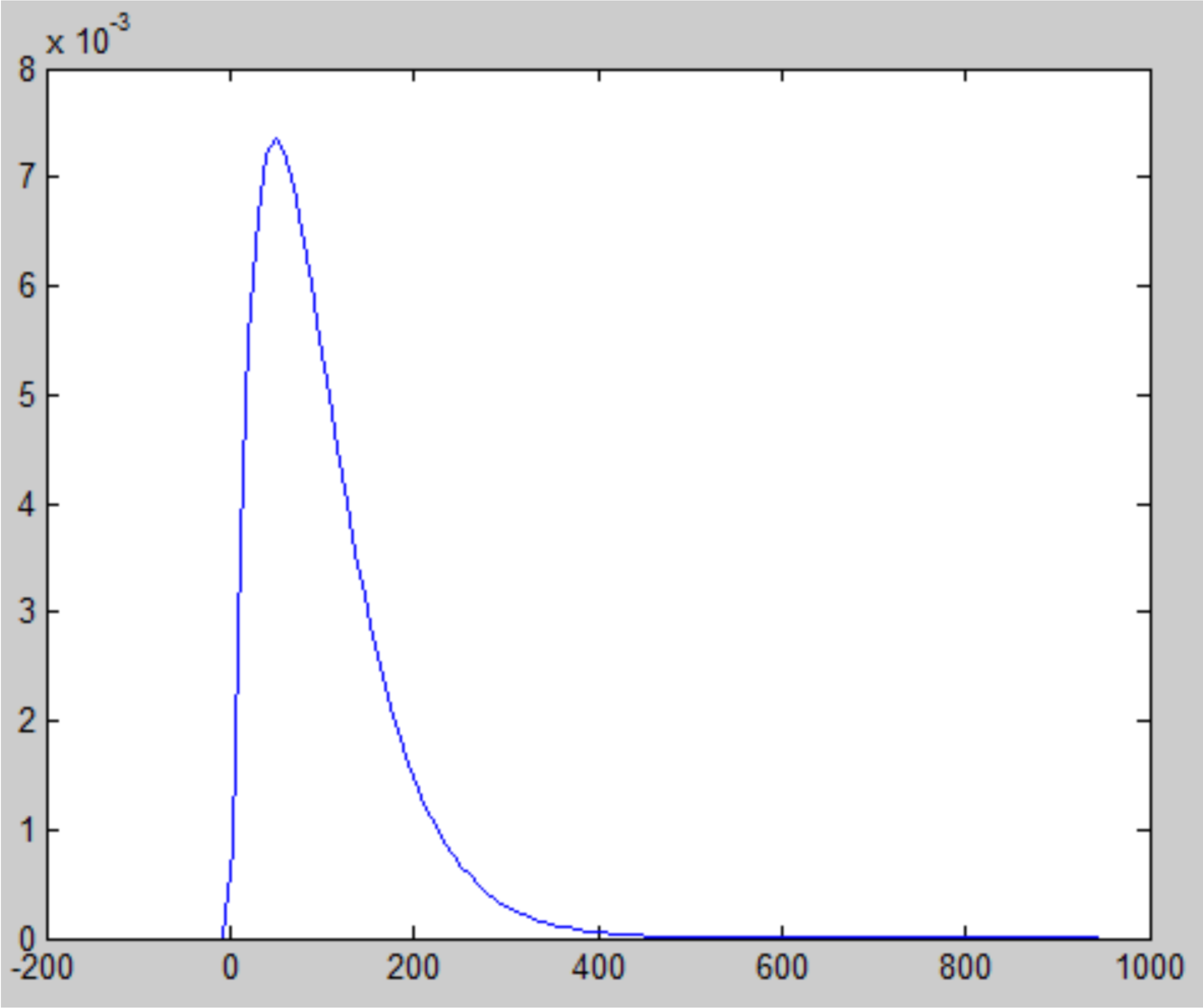
Algorithm 4.5: $\mathrm{TT}(X_1, \ldots, X_n, \delta)$: Uses the TT to return an approximate $1-\delta$ confidence lower bound on $\mathbf{E}[\frac{1}{n}\sum_{i=1}^{n} X_i]$.
Assumes: The X_i are independent random variables with finite variance..

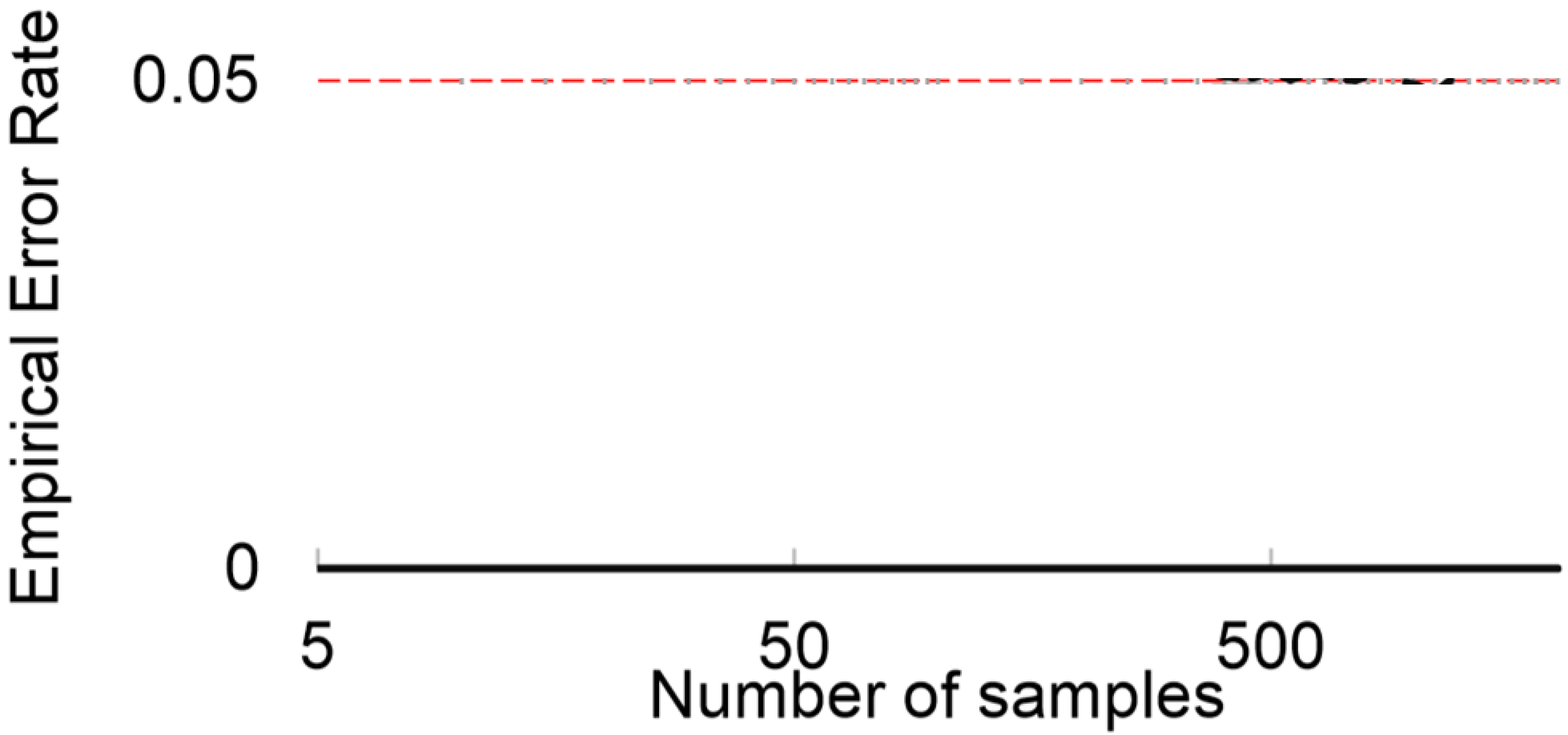
1 **return** $\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2}}{\sqrt{n}} t_{1-\delta,n-1}$;

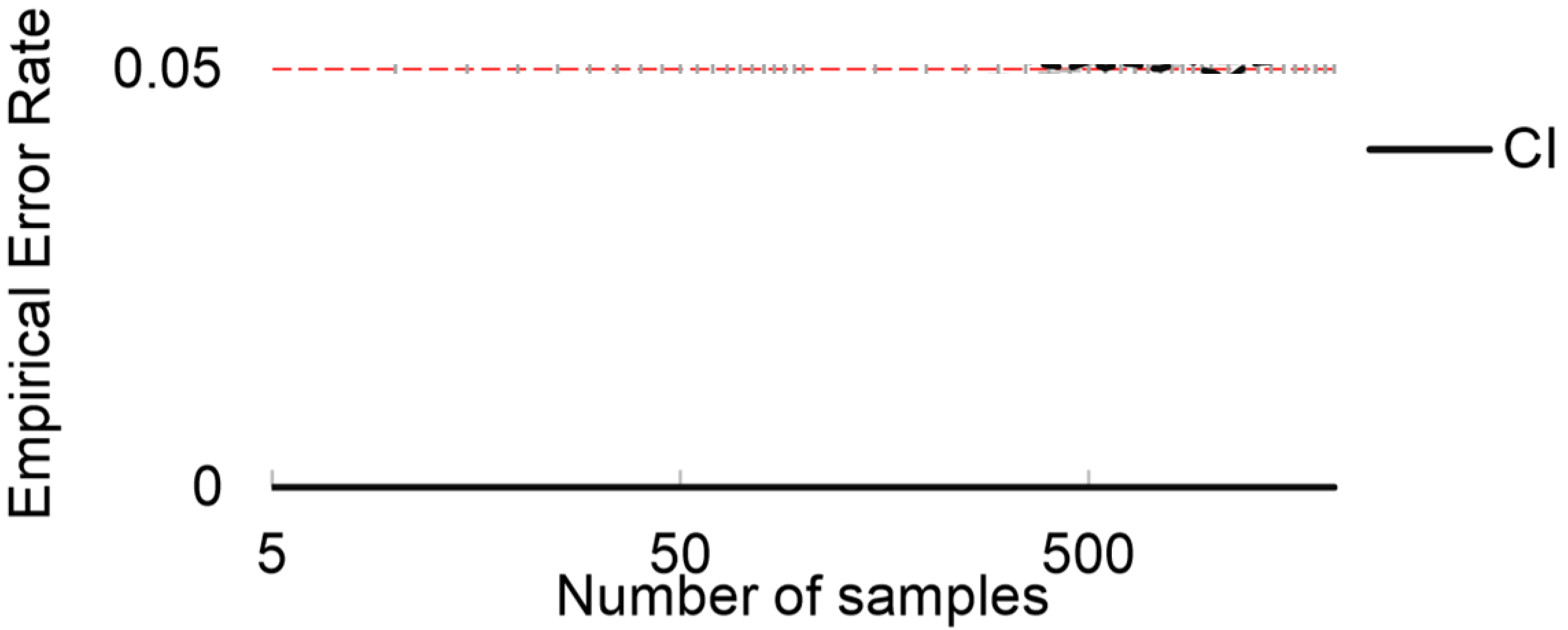
Approximate Confidence Intervals

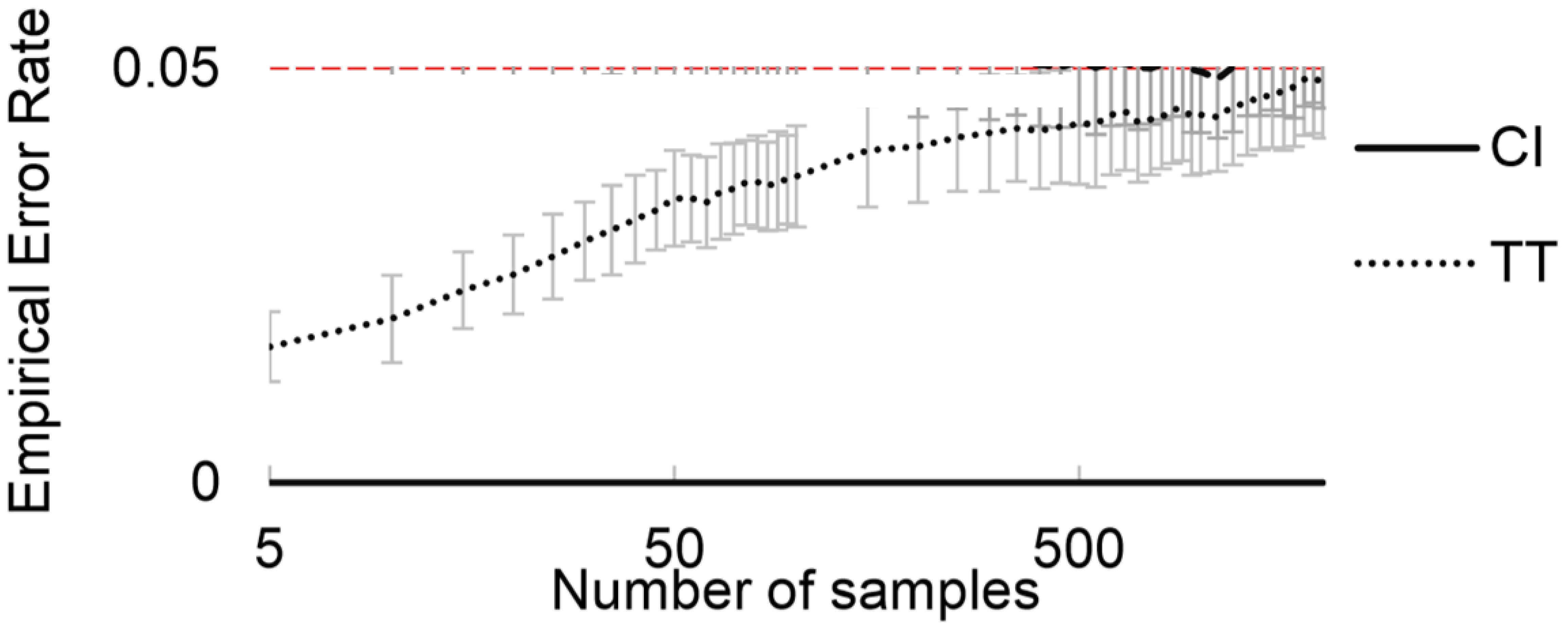
- Bootstrap (BCa)
  - Estimate CDF with sample CDF:
$$F_n(x) \coloneqq \frac{1}{n}\sum_{i=1}^{n} 1_{X_i \leq x}$$
  - What if we assume that the samples come from a distribution like F_n?
    - Sort $X_1, \ldots, X_n$ and return $X_{\delta n}$

ksdensity(gamrnd(2, 50, 10000000, 1))

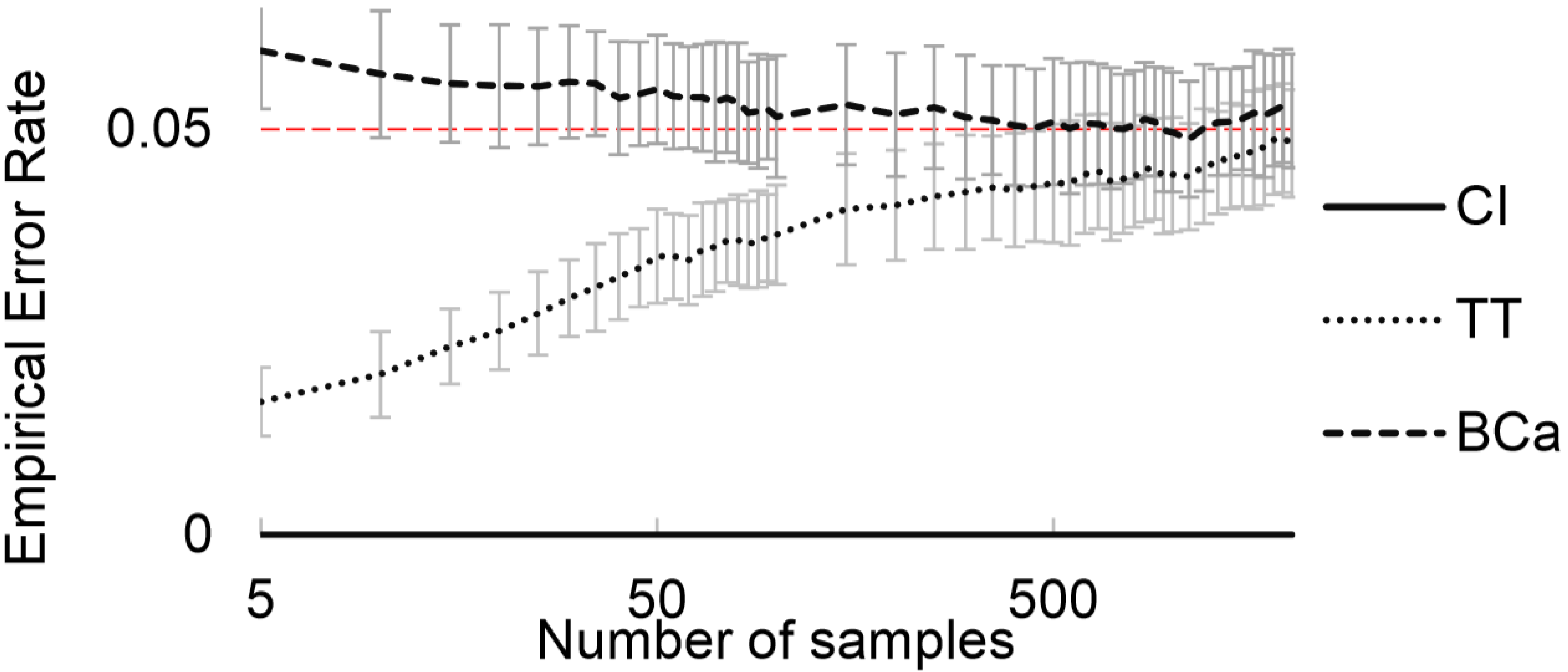



Empirical Error Rate
0.05
0
5
50
500
Number of samples

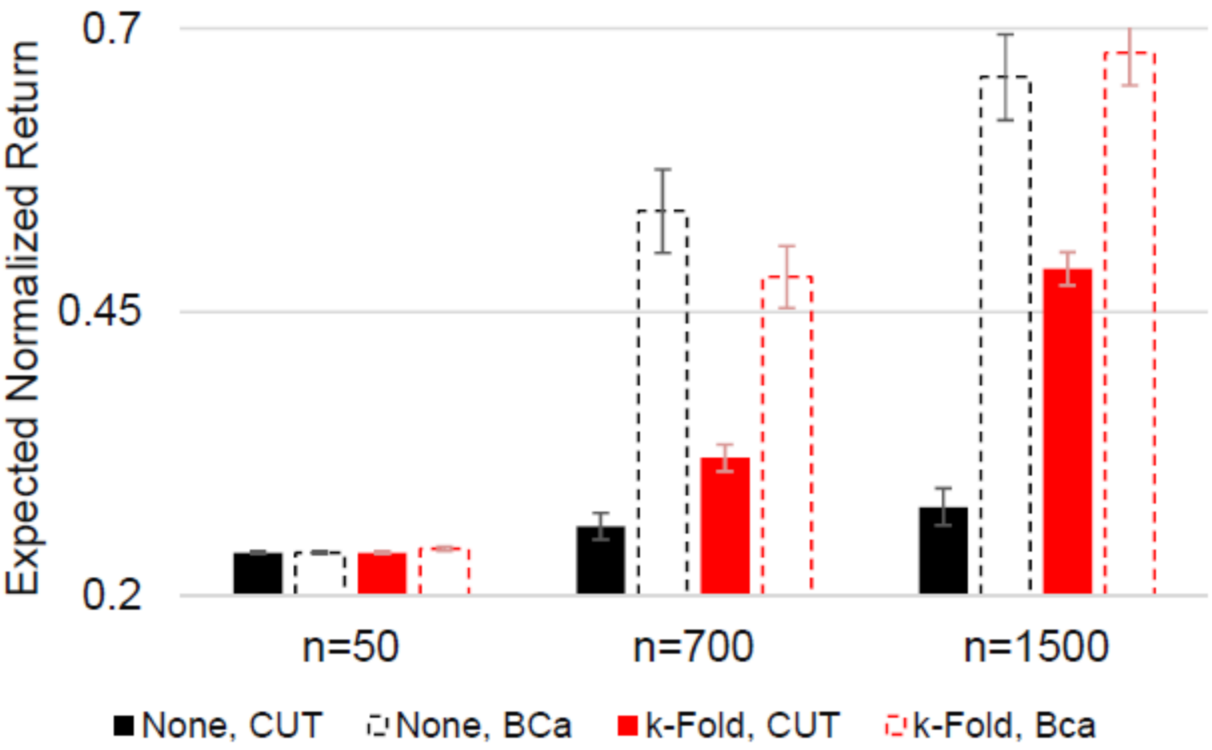

Empirical Error Rate
0.05
0
5
50
500
Number of samples
CI



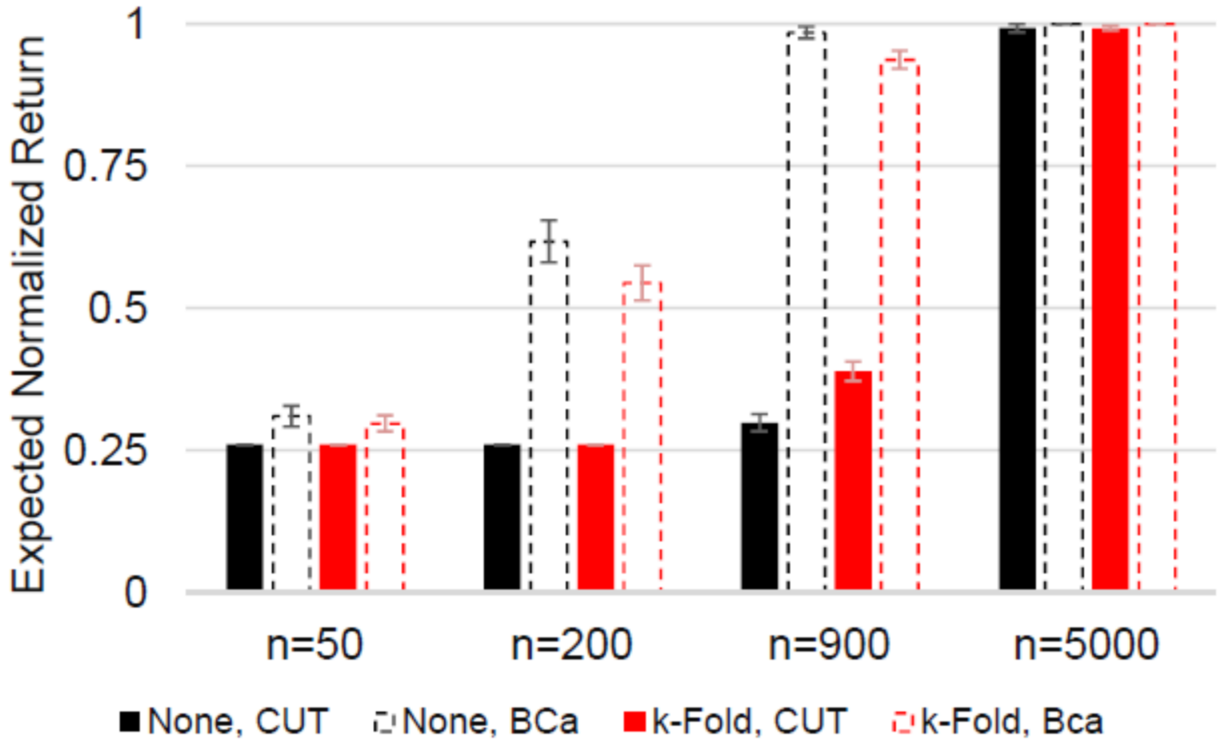
Empirical Error Rate
0.05
0
5
50
500
Number of samples
CI
TT


Empirical Error Rate
0.05
0
5
50
500
Number of samples
CI
TT
BCa

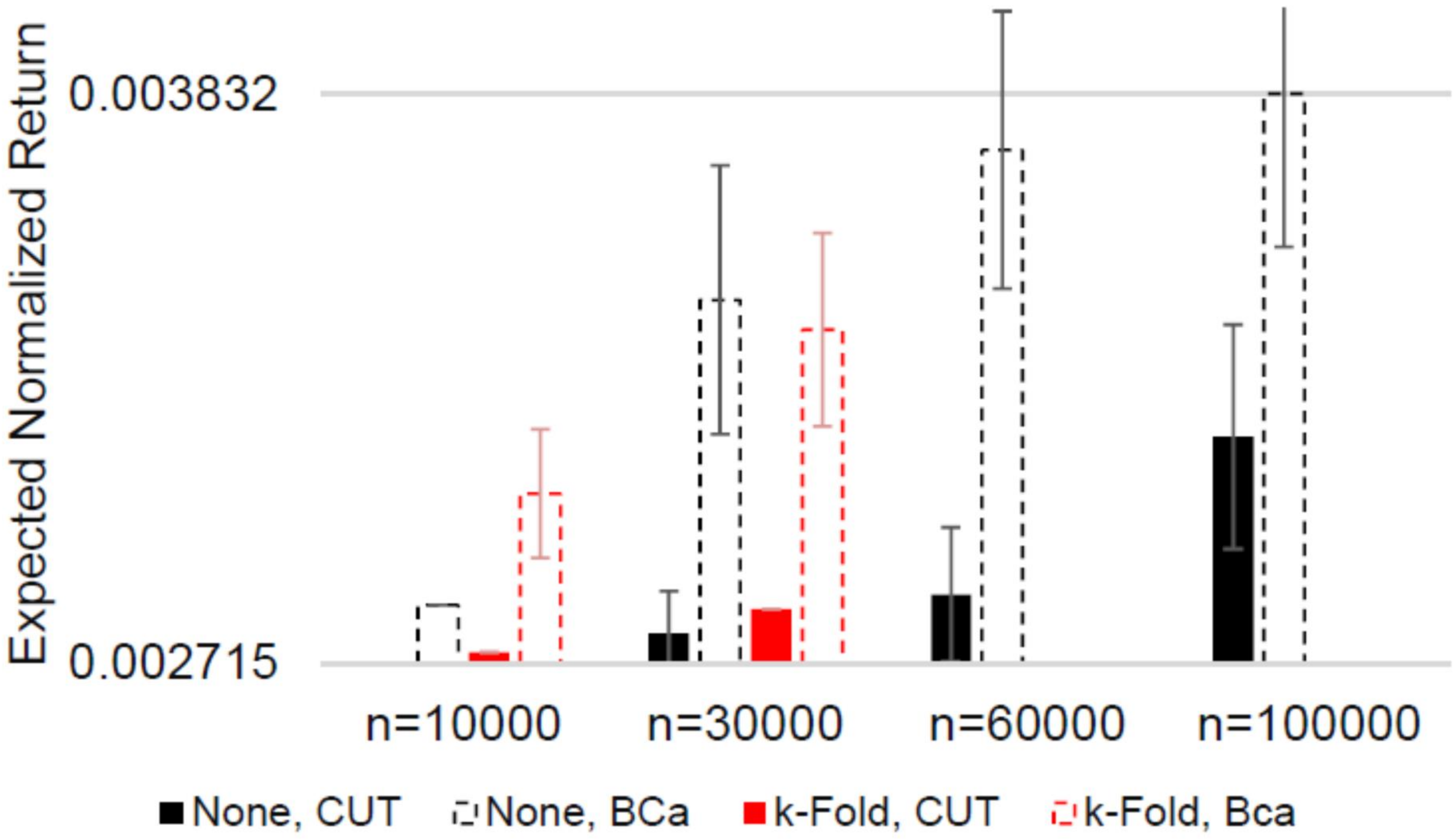
Example: Gridworld



Example: Mountain Car



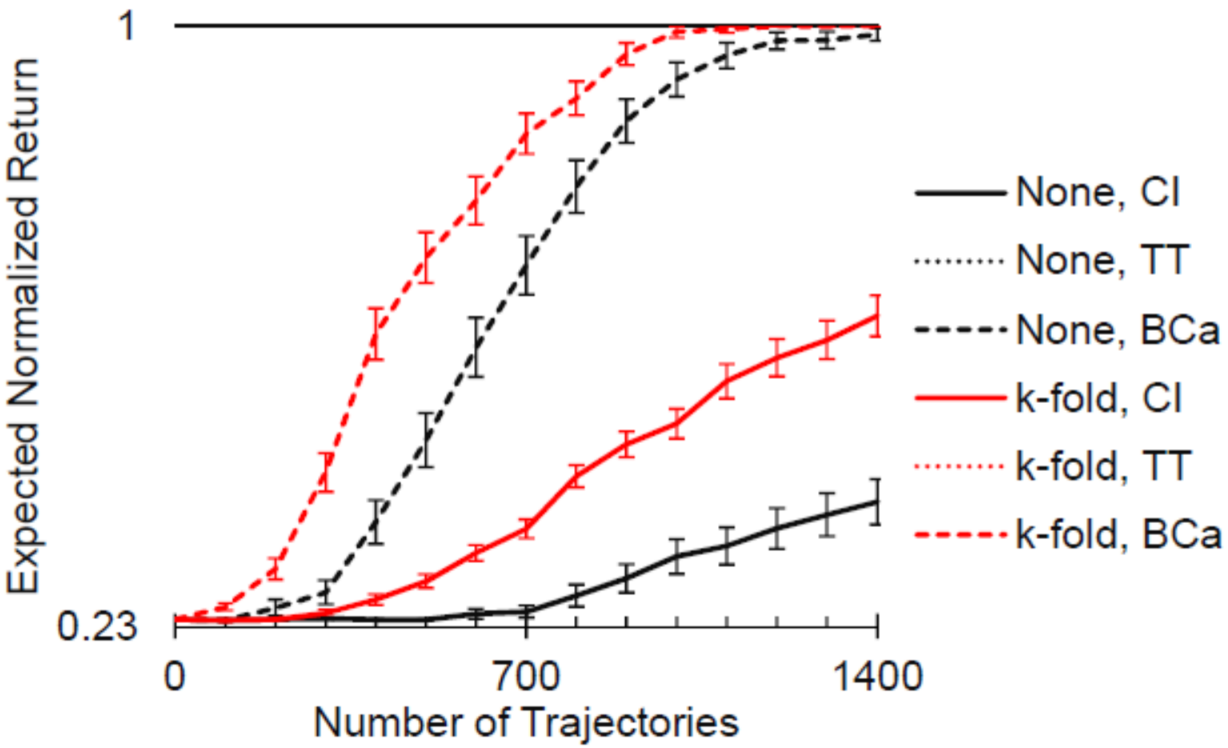
Example: Digital Marketing



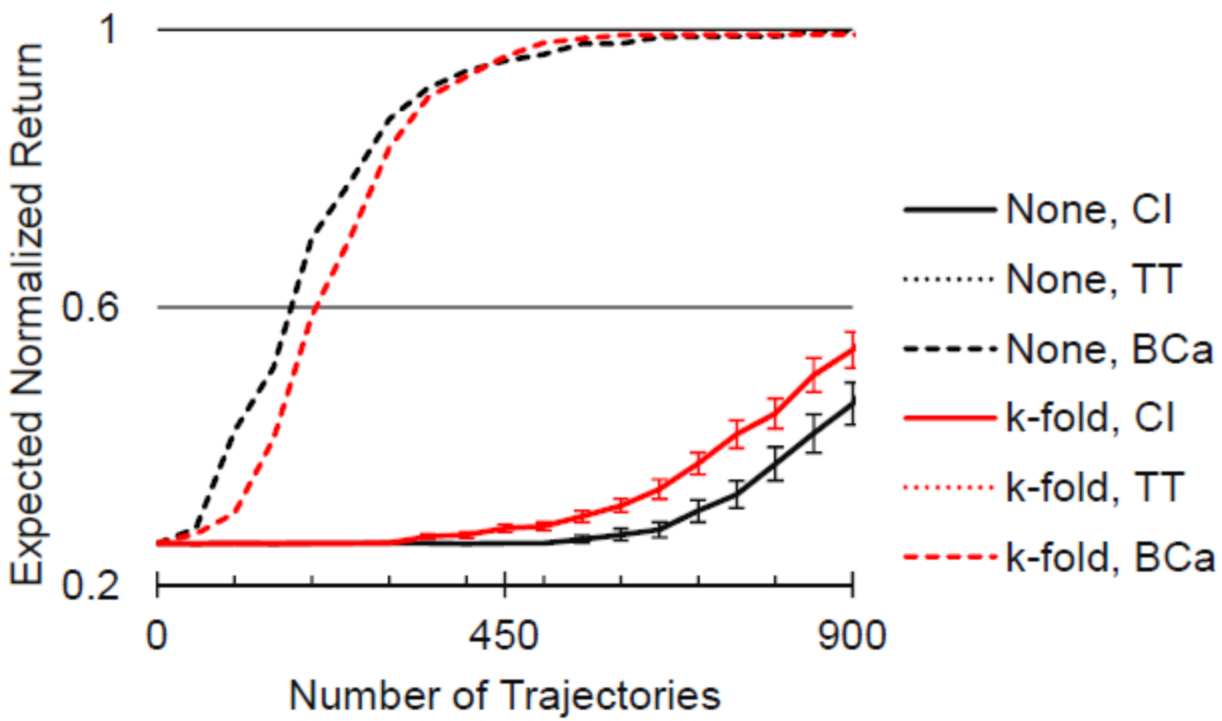
DAEDALUS

- Apply SPI algorithm repeatedly.
- Per-iteration guarantee.
- DAEDALUS2: Exact HCPI until the first change to the policy, then approximate.

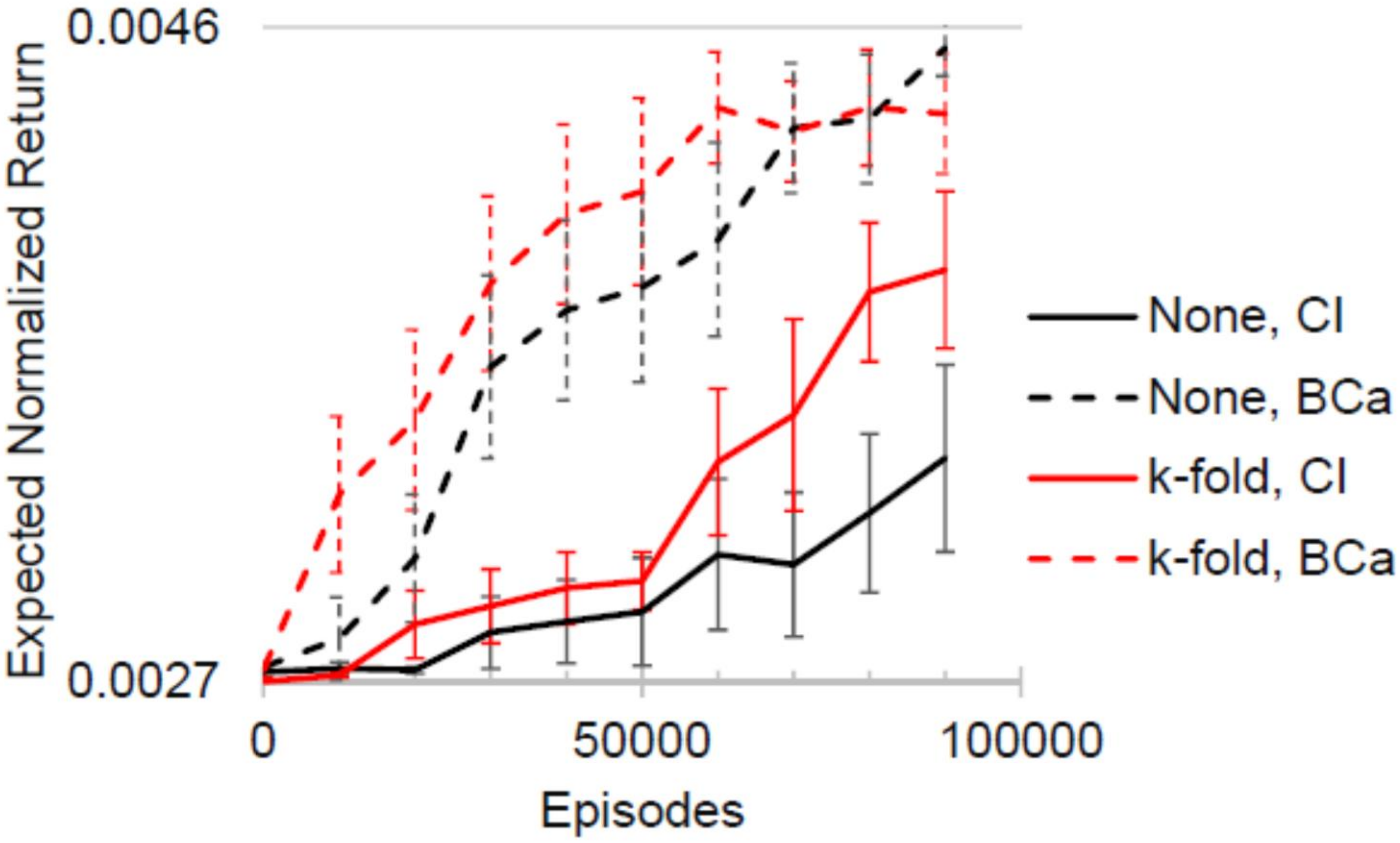
Example: Gridworld



Example: Mountain Car



Example: Digital Marketing



Safe policy improvement is tractable

A policy improvement algorithm that has a low probability of returning a bad policy.