Safety in Sequential Decision-making

Mohammad Ghavamzadeh

August 3, 2018

Outline

Safety: Problem Formulation

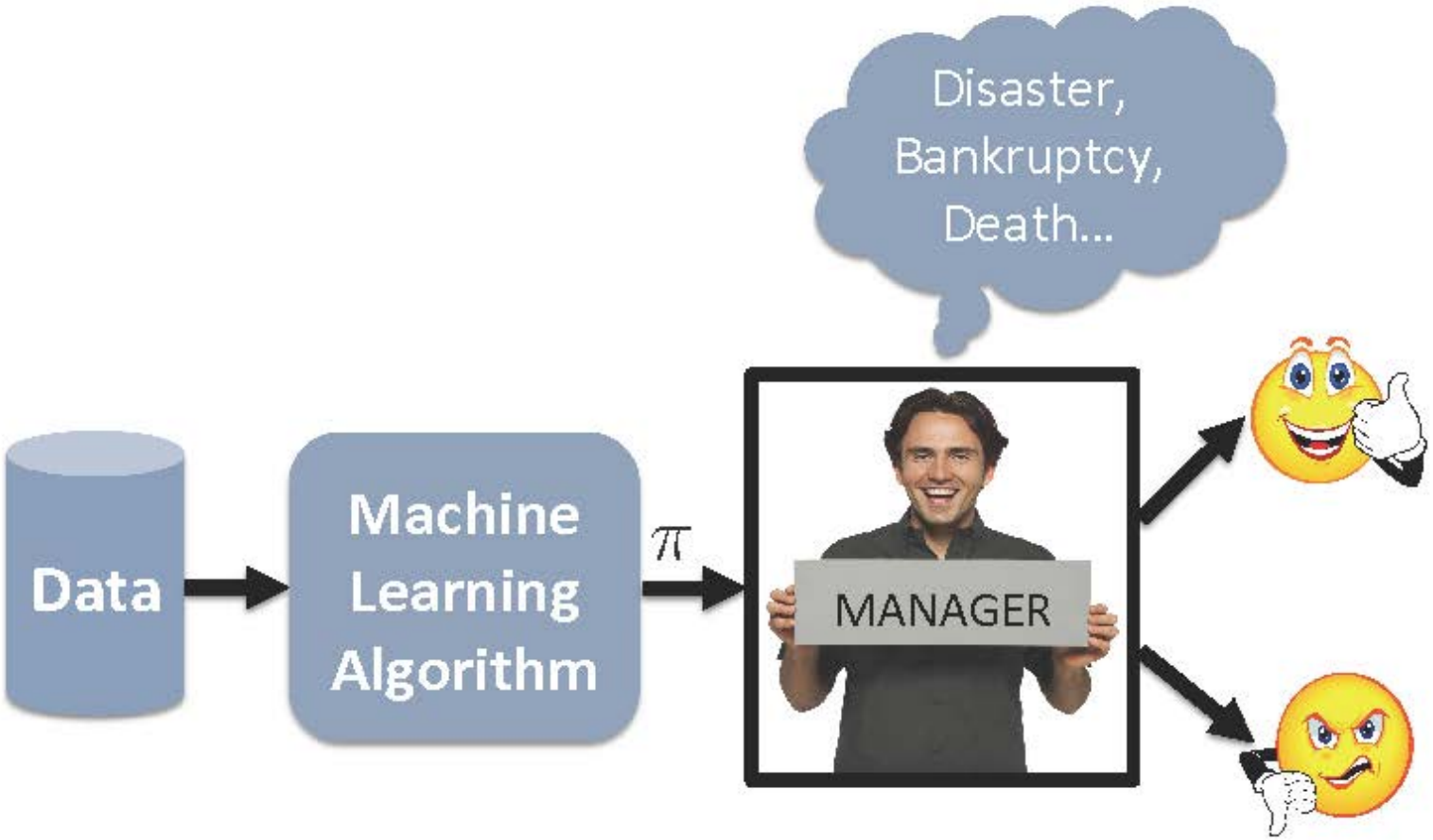
Different Approaches to Safety

Risk-sensitive Decision-making *(optional)*

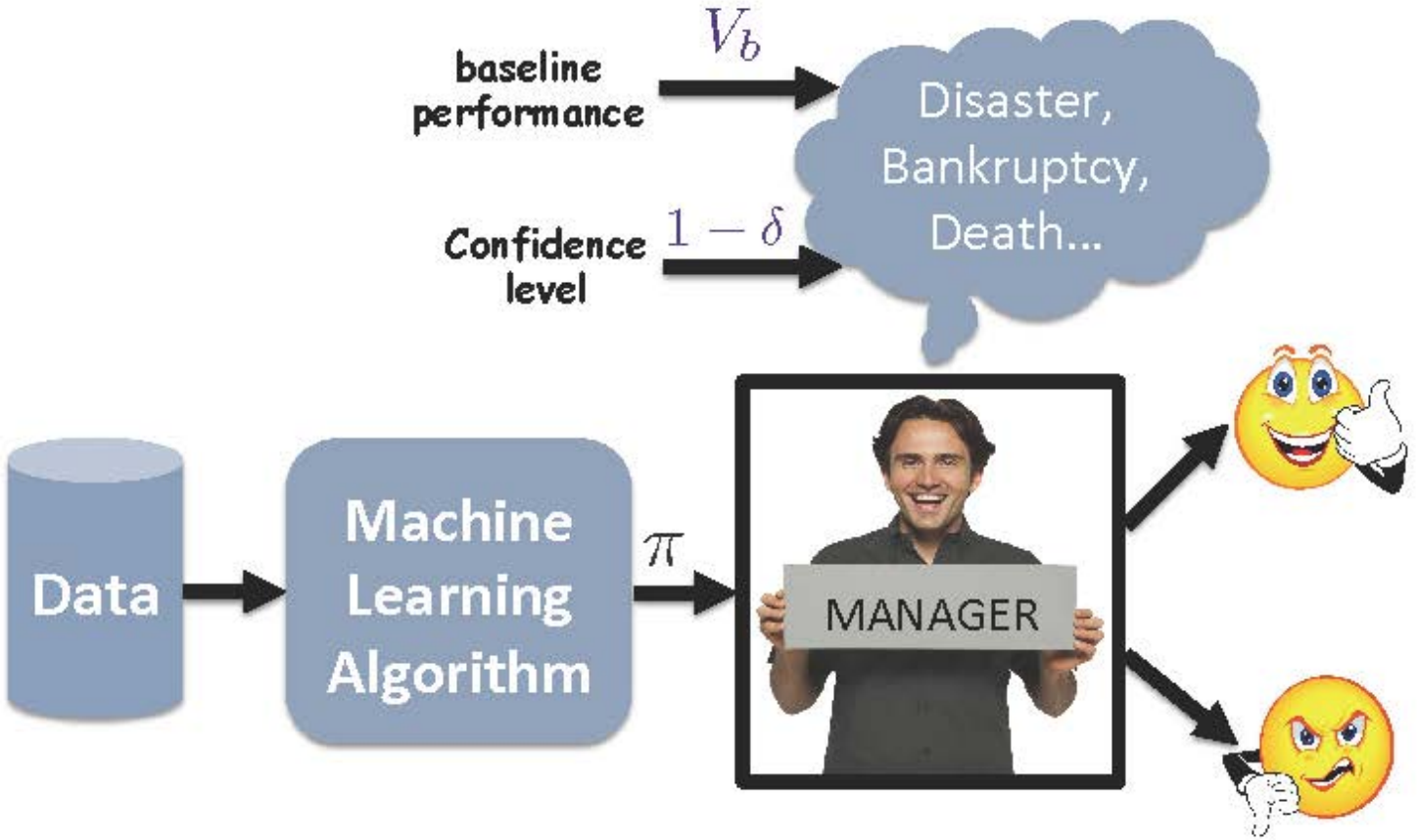
Safety w.r.t. Undesirable Situations *(optional)*

Outline

Safety: Problem Formulation



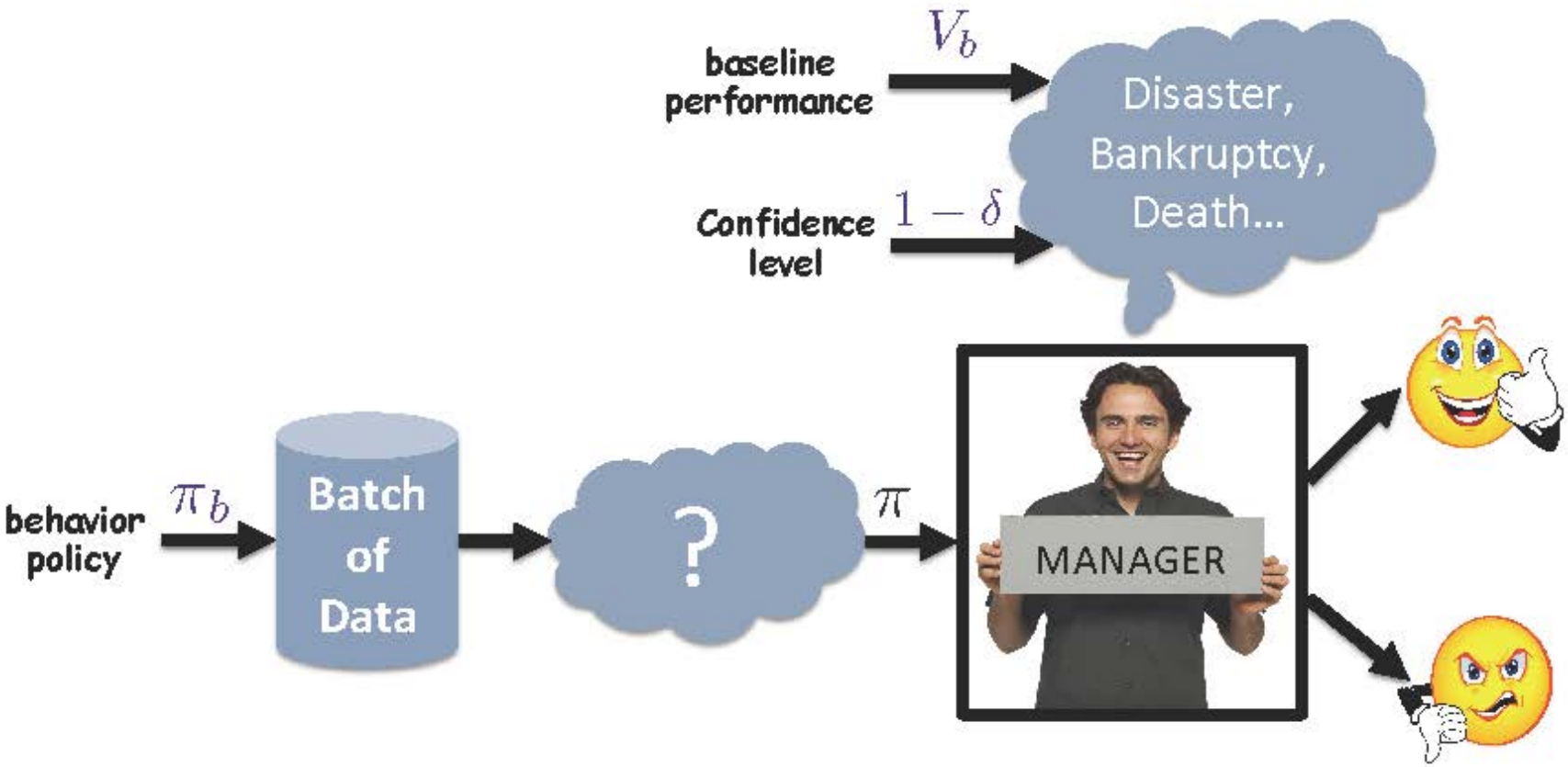
Safety: Problem Formulation



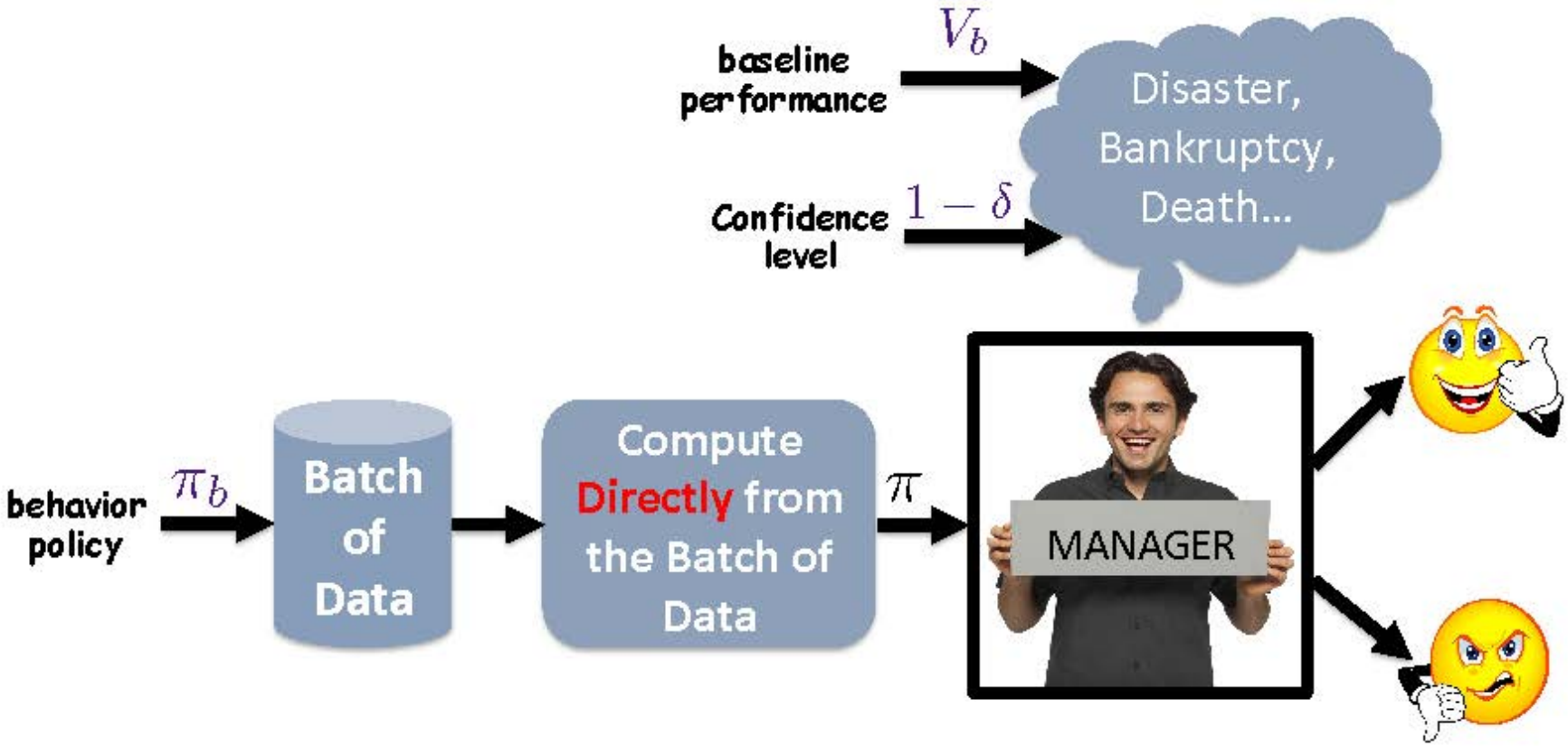
Outline

1. Model-free Approach

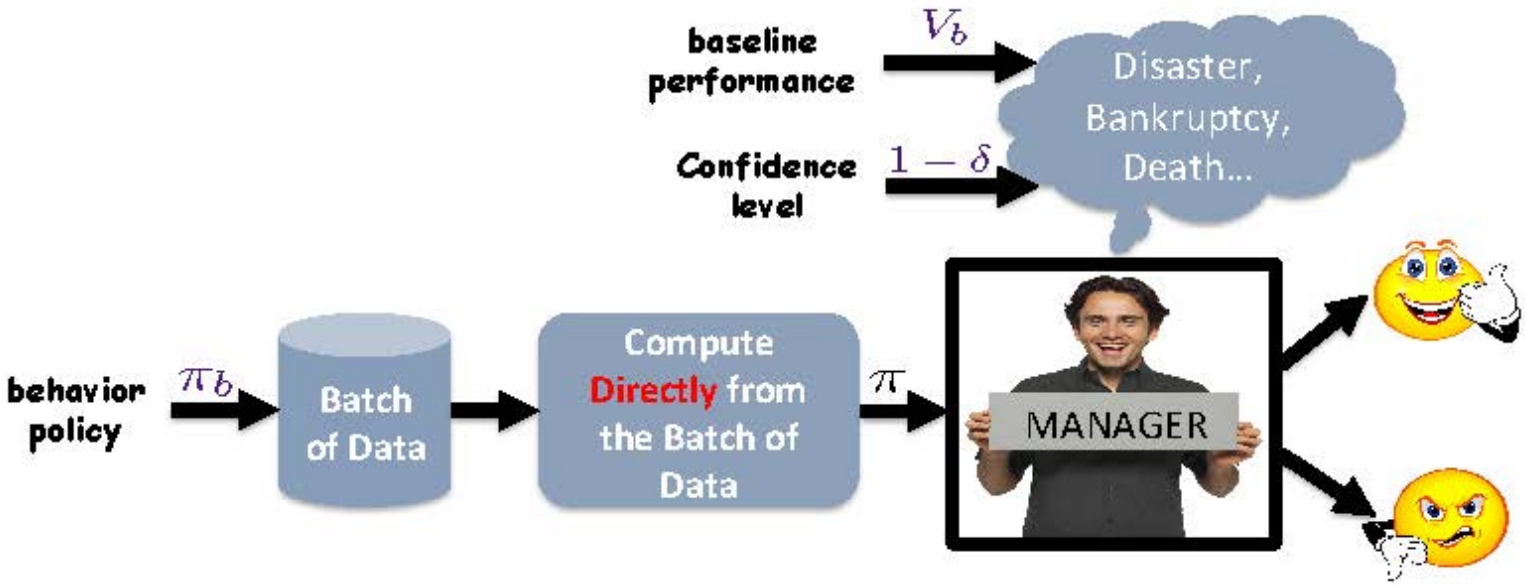
Offline Setting

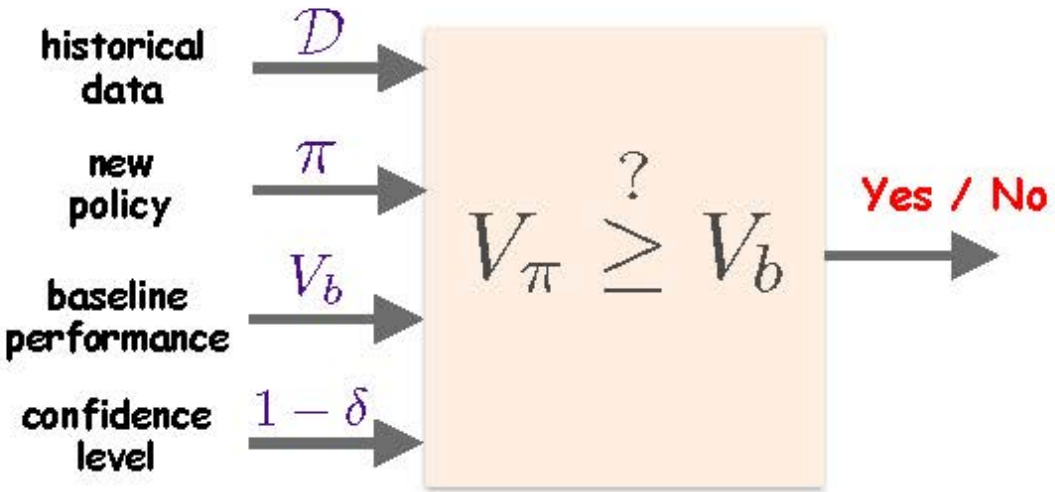


Model-free Approach

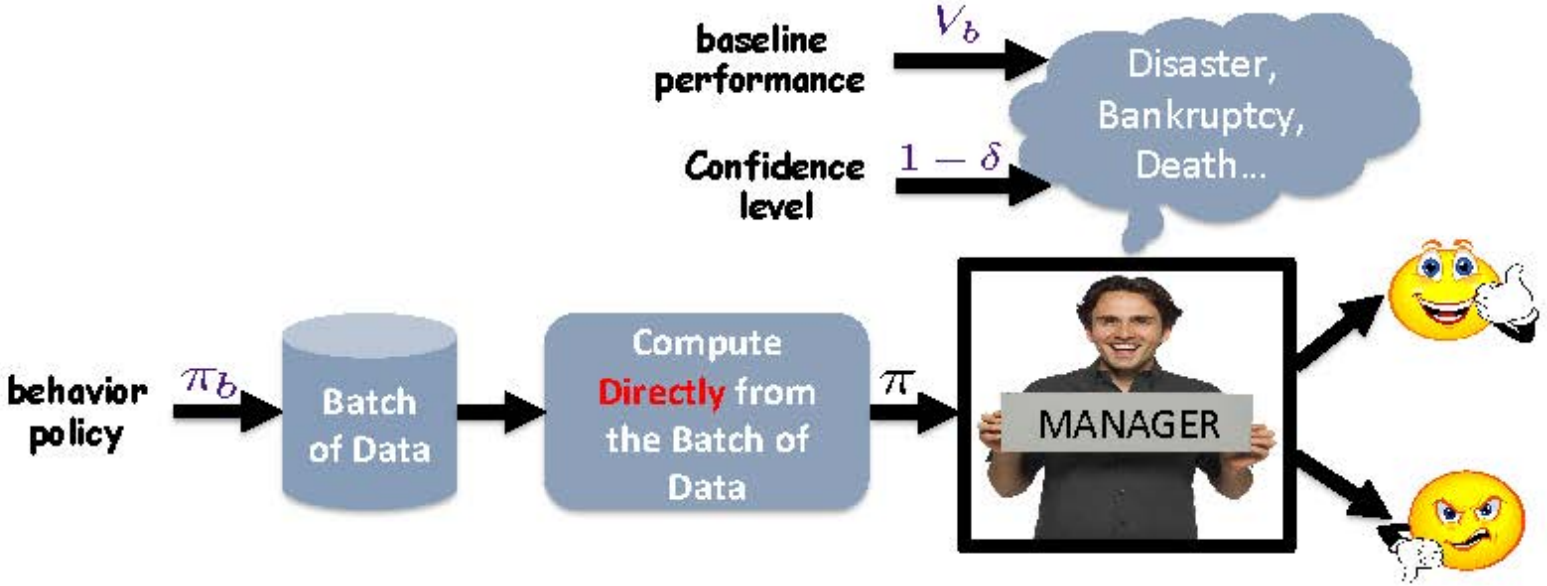


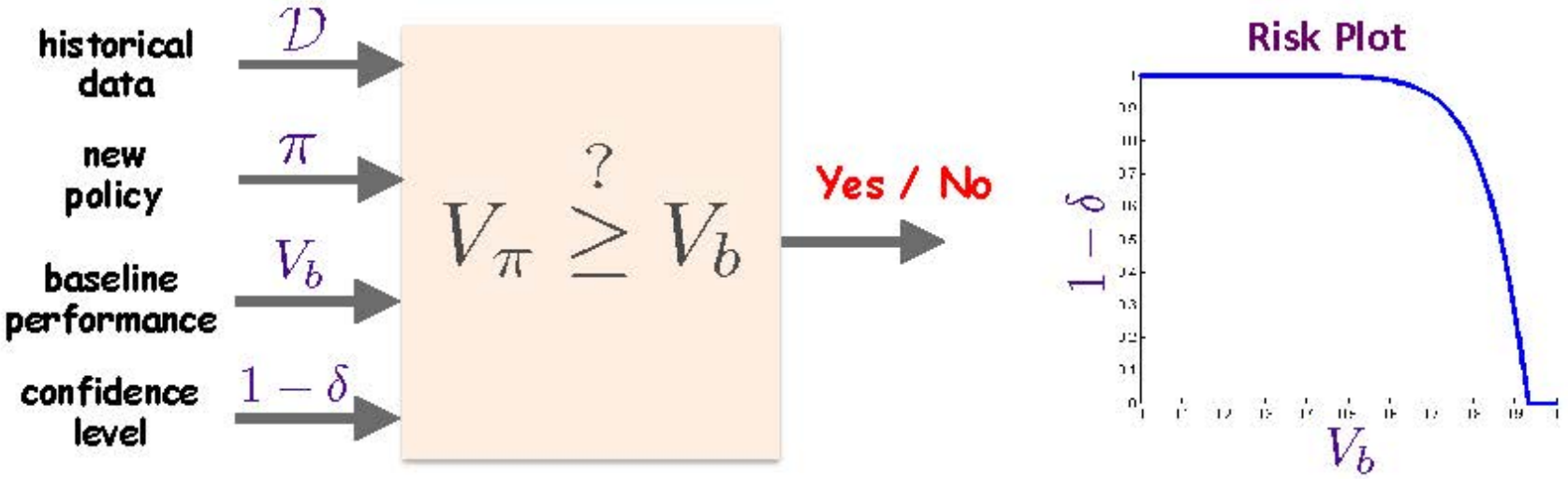
Model-free Approach





Model-free Approach





Publications

1. P. Thomas, G. Theocharous, and **MGH**. *"High Confidence Off-Policy Evaluation"*. ***AAAI-2015***.

2. P. Thomas, G. Theocharous, and **MGH**. *"High Confidence Policy Improvement"*. ***ICML-2015***.

3. G. Theocharous, P. Thomas, and **MGH**. *"Building Personal Ad Recommendation Systems for Life-Time Value Optimization with Guarantees"*. ***IJCAI-2015***.

Other Publications

- L. Bottou, J. Peters, J. Quinonero-Candela, D. Charles, D. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. *"Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising"*. **JMLR-2013**.

- A. Swaminathan and T. Joachims. *"Counterfactual Risk Minimization: Learning from Logged Bandit Feedback"*. **ICML-2015**.

- N. Jiang and L. Li. *"Doubly Robust Off-Policy Value Evaluation for Reinforcement Learning"*. **ICML-2016**.

- P. Thomas and E. Brunskill. *"Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning"*. **ICML-2016**.

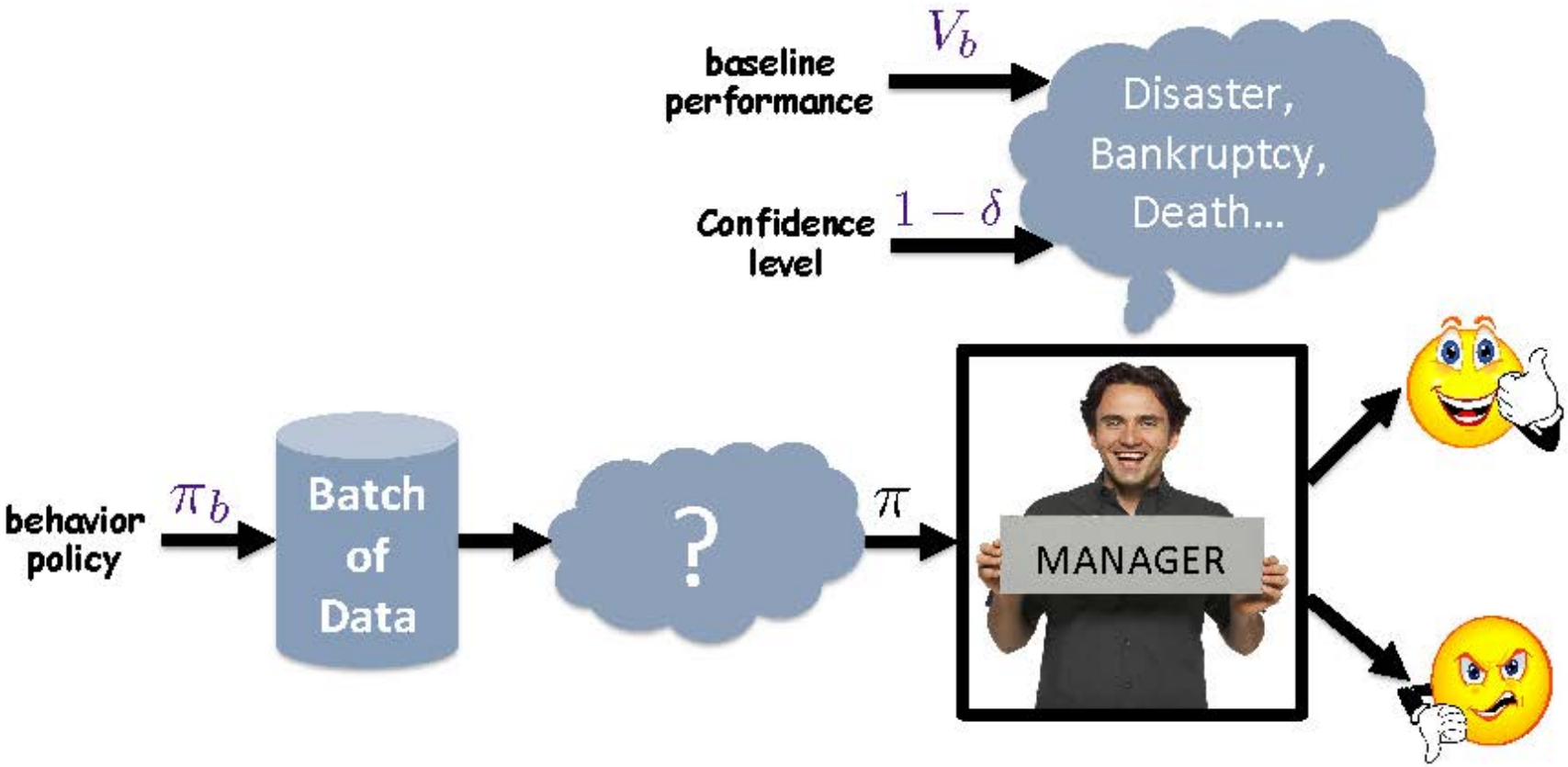
- A. Swaminathan, A. Krishnamurthy, A. Agarwal, M. Dudik, J. Langford, D. Jose, and I. Zitouni. *"Off-policy Evaluation for Slate Recommendation"*. **NIPS-2017**.

- 

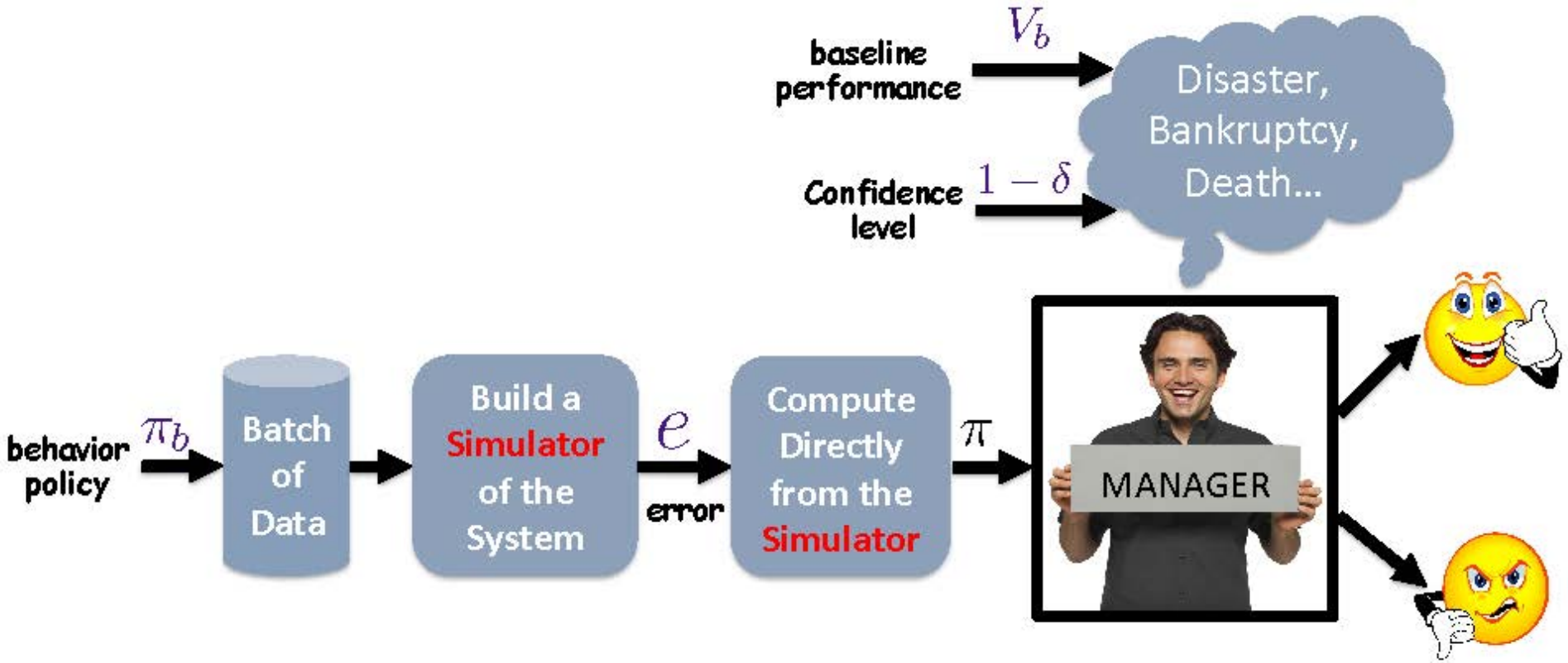
M. Farajtabar, Y. Chow, and **MGH**. *"More Robust Doubly Robust Off-policy Evaluation"*. **ICML-2018**.

2. Model-based Approach

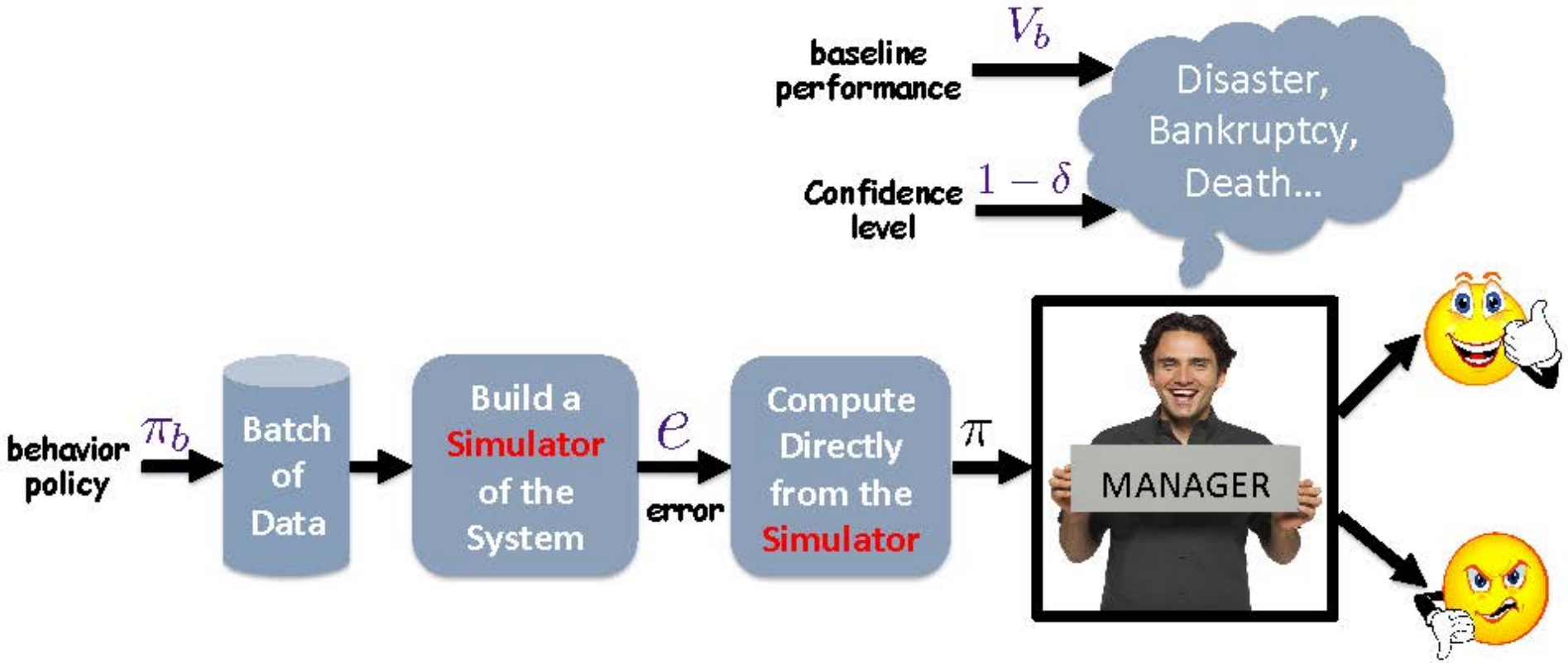
Offline Setting



Model-based Approach



Model-based Approach



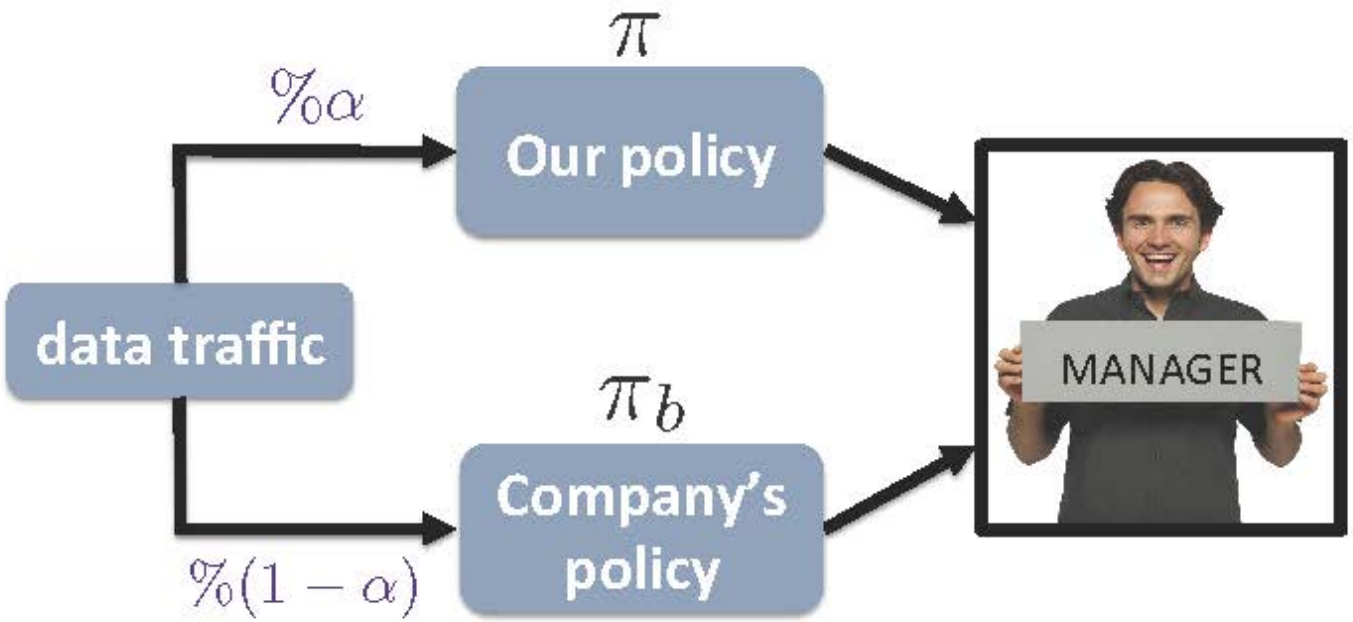
Main Question: Given the simulator and the error in building it, how to compute a policy that is guaranteed ***(with a given confidence level)*** to perform at least as well as a baseline???

Publications

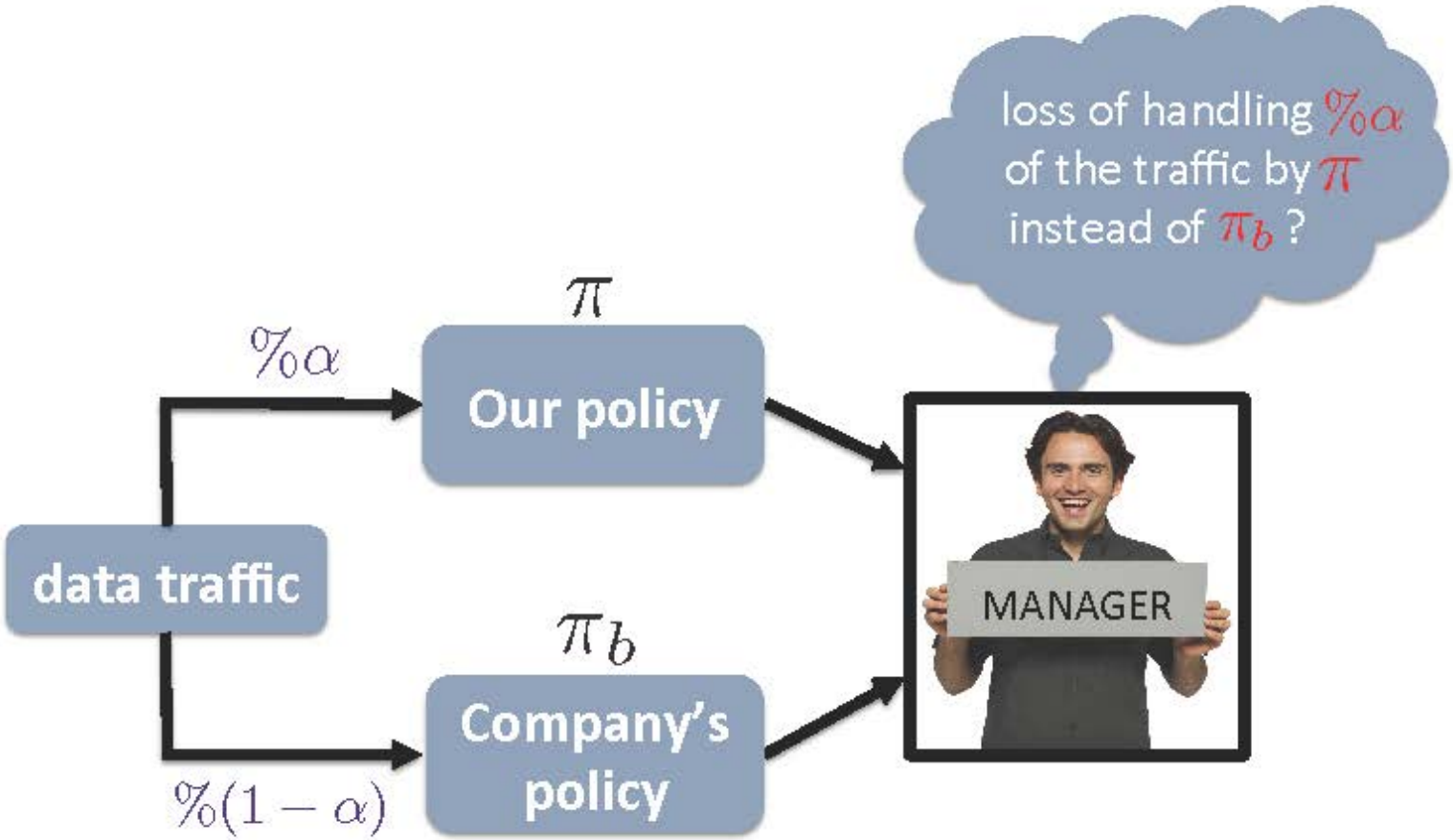
1. M. Petrik, **MGH**, and Y. Chow. *"Safe Policy Improvement by Minimizing Robust Baseline Regret"*. ***NIPS-2016***.

3. Online Approach

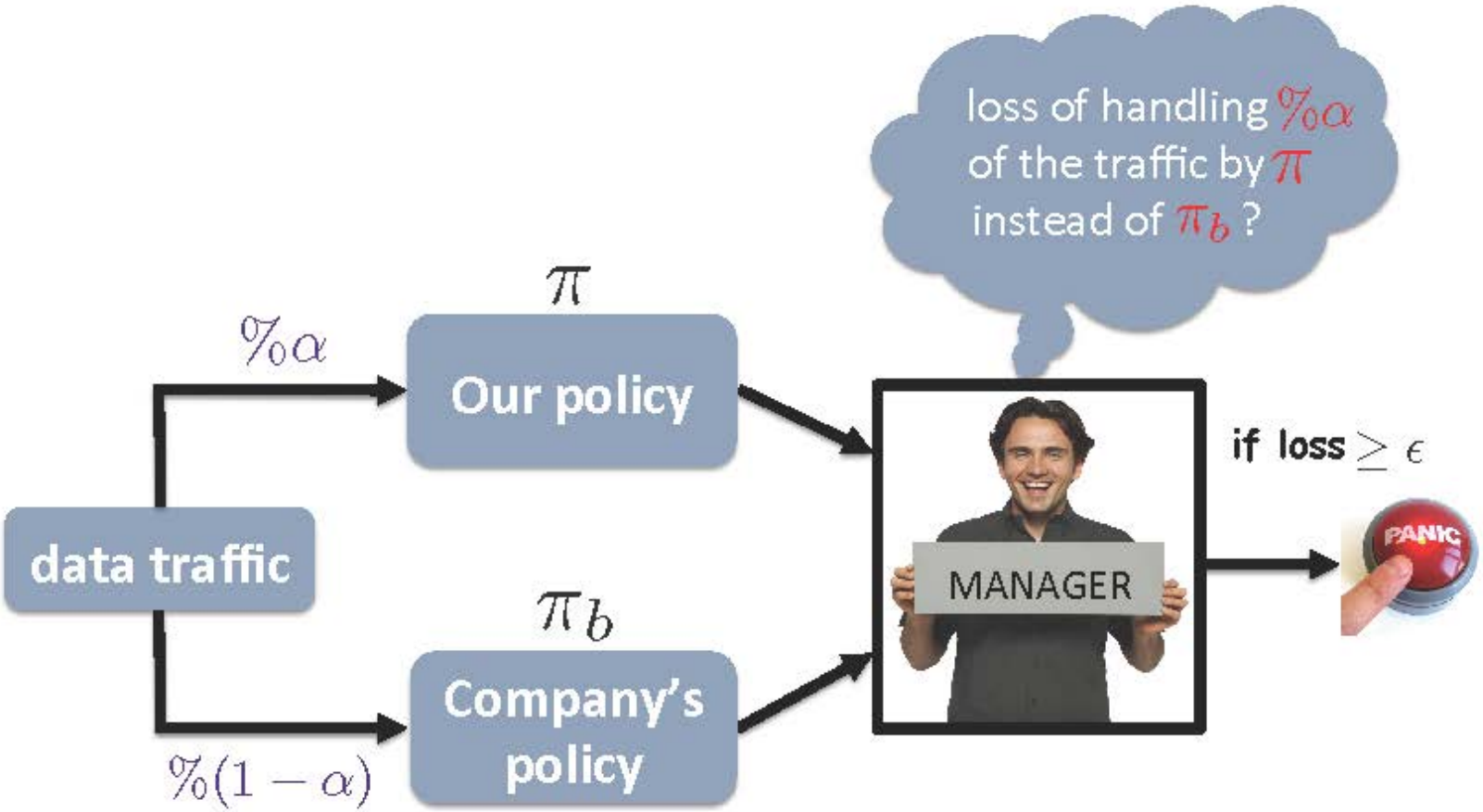
Online Approach



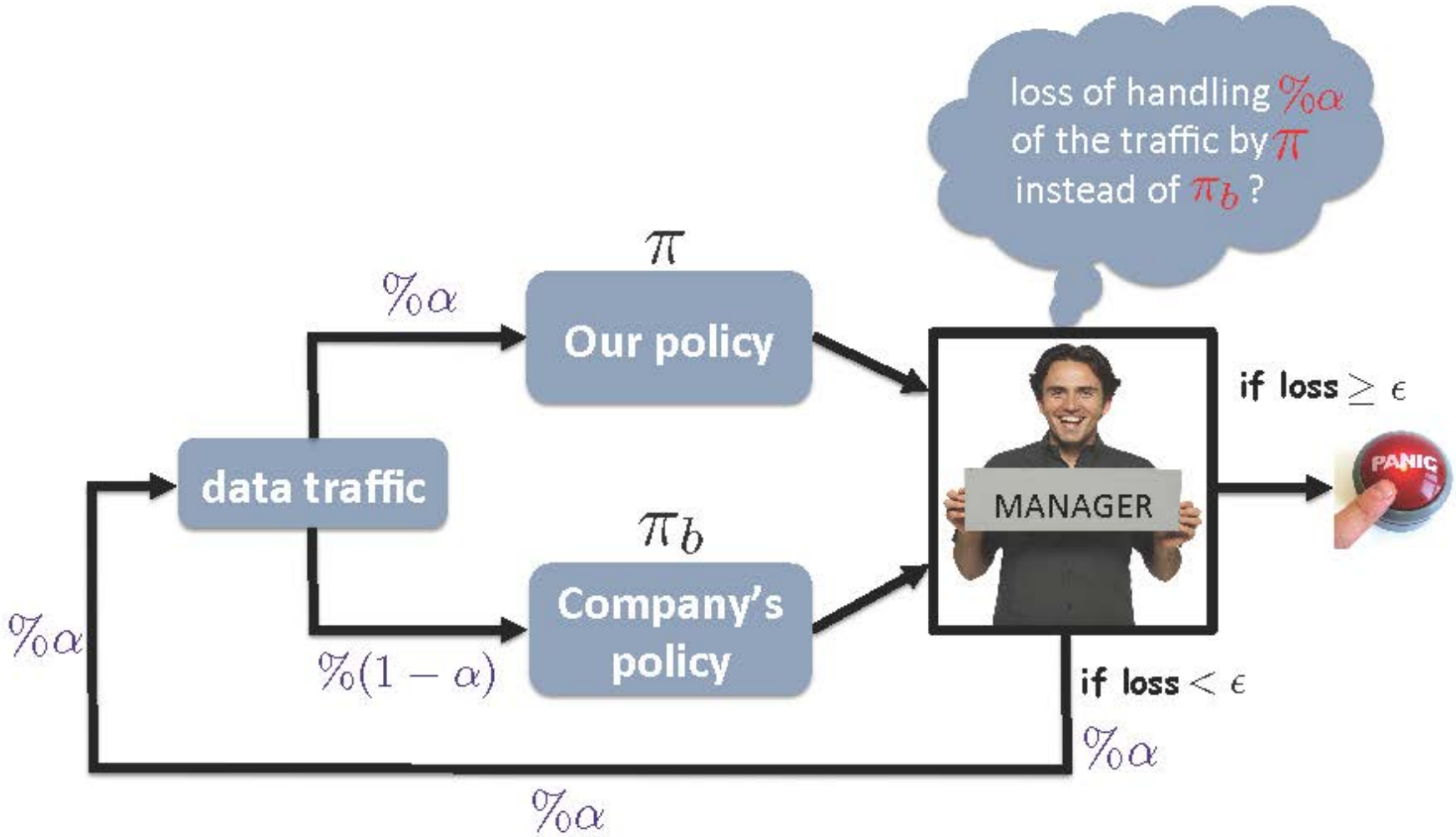
Online Approach



Online Approach



Online Approach



Publications

1. A. Kazerouni, **MGH**, Y. Abbasi-Yadkori, and B. Van Roy. *"Conservative Contextual Linear Bandits"*. ***NIPS-2017***.

2. Y. Wu, R. Shariff, T. Lattimore, and C. Szepesvári. *"Conservative bandits"*. ***ICML-2016***.

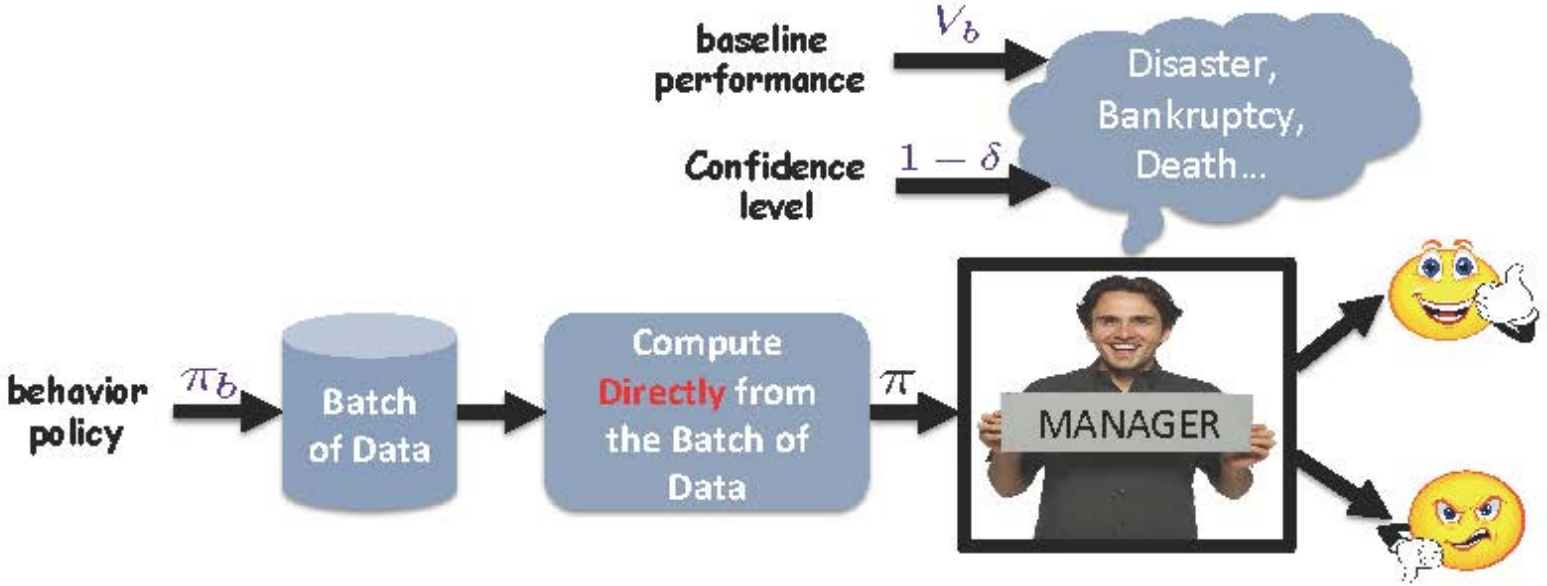
Other Publications

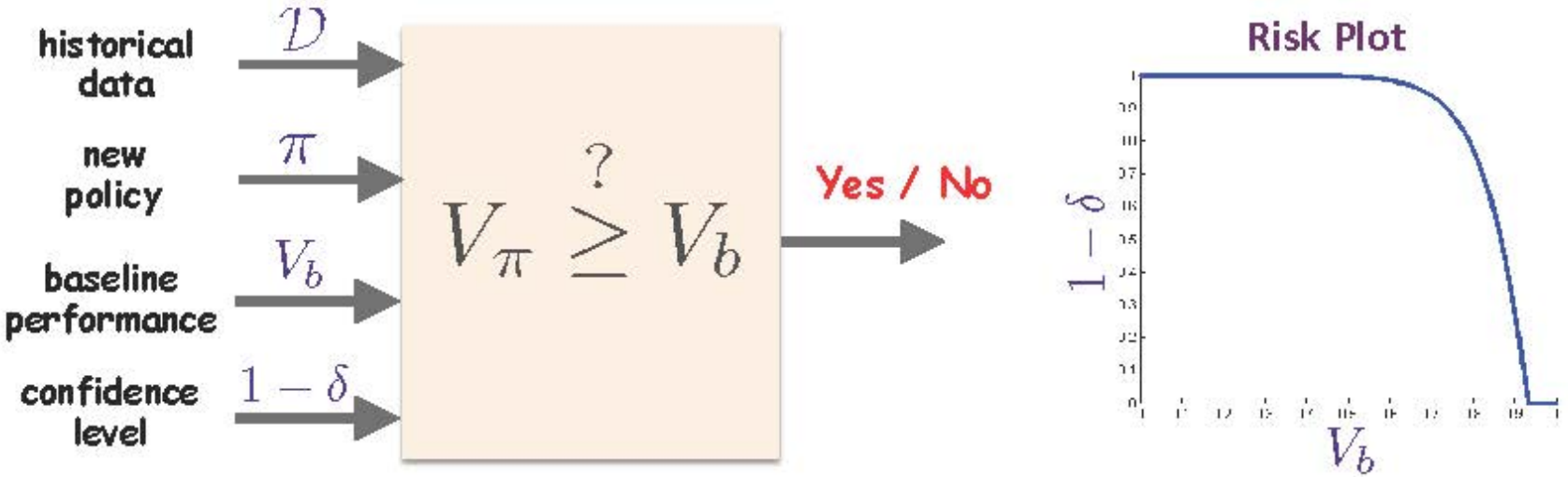
1. Y. Mansour, A. Slivkins, and V. Syrgkanis. *"Bayesian Incentive-Compatible Bandit Exploration"*. ***EC-2015***.

Outline

Model-free Approach

Model-free Approach





Problem Definition

new policy π has been computed using a finite number of samples

- Is it ***safe*** to deploy policy π?
- Is π guaranteed to perform at least as well as π_b?
- Is the performance of π guaranteed to be at least V_b?

Problem Definition

new policy π has been computed using a finite number of samples

- Is it ***safe*** to deploy policy π?
- Is π guaranteed to perform at least as well as π_b?
- Is the performance of π guaranteed to be at least V_b?

an important problem in many different fields including marketing, health, and finance

High-Confidence Off-Policy Evaluation

Problem Formulation

- **System Trajectory**

$$\tau = \{x_1, a_1, r_1, x_2, a_2, r_2, \ldots, s_T, a_T, r_T\} \qquad r_t = r(x_t, a_t)$$

- **Return of a Trajectory** *(assume rewards are in $[0,1]$)*

$$D(\tau) = \sum_{t=1}^{T} \gamma^{t-1} r_t \ \in \ [0, 1/1-\gamma]$$

- **Policy Performance**

$$V^{\pi} = \mathbb{E}\big[D(\tau)\big] \ \in \ [0, 1/1-\gamma] \qquad \text{if} \quad a_t \sim \pi(\cdot|x_t)$$

Problem Formulation

- **Historical Data**

$$\mathcal{D} = \left\{ (\tau_i, \pi_i) \right\}_{i=1}^{n}$$

τ_i ***has been generated by*** π_i

Problem Formulation

- **Historical Data**

$$\mathcal{D} = \left\{ (\tau_i, \pi_i) \right\}_{i=1}^{n} \qquad \tau_i \textbf{ has been generated by } \pi_i$$

- **Behavior Policies** $\pi_1, \ldots, \pi_n$ **Target Policy** π

Problem Formulation

- **Historical Data**

$$\mathcal{D} = \left\{ (\tau_i, \pi_i) \right\}_{i=1}^{n} \qquad \tau_i \textbf{ has been generated by } \pi_i$$

- **Behavior Policies** $\pi_1, \ldots, \pi_n$ **Target Policy** π

- **Baseline Performance** V_b

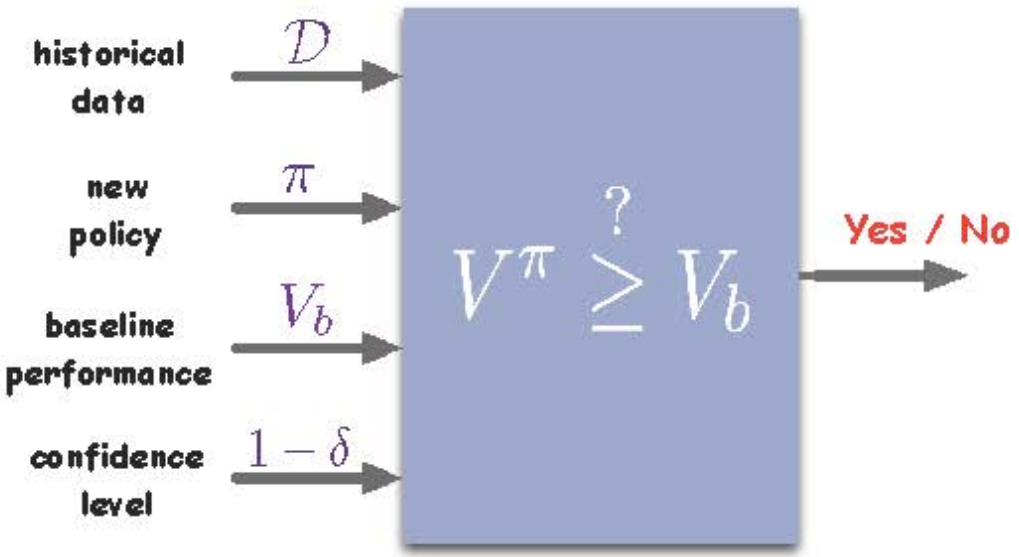
Problem Formulation

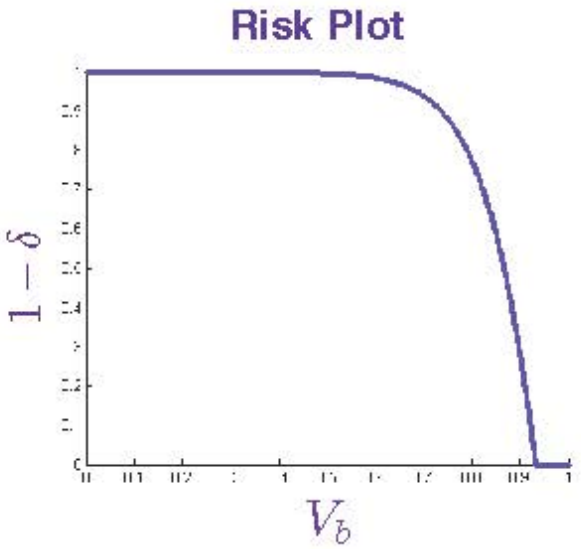
- **Historical Data**

$$\mathcal{D} = \left\{ (\tau_i, \pi_i) \right\}_{i=1}^{n} \qquad \tau_i \textit{ has been generated by } \pi_i$$

- **Behavior Policies** $\pi_1, \ldots, \pi_n$ **Target Policy** π

- **Baseline Performance** V_b





Weighted Importance Return

For any $(\tau_i, \pi_i) \in \mathcal{D}$

$$\widehat{D}(\tau_i, \pi, \pi_i) = D(\tau_i) \frac{\Pr(\tau_i|\pi)}{\Pr(\tau_i|\pi_i)} = \underbrace{D(\tau_i)}_{\text{return}} \underbrace{\prod_{t=1}^{T} \frac{\pi(a_t|x_t)}{\pi_i(a_t|x_t)}}_{\text{importance weight}}$$

Weighted Importance Return

For any $(\tau_i, \pi_i) \in \mathcal{D}$

$$\widehat{D}(\tau_i, \pi, \pi_i) = D(\tau_i)\ \frac{\Pr(\tau_i|\pi)}{\Pr(\tau_i|\pi_i)} = \underbrace{D(\tau_i)}_{\textbf{return}}\ \underbrace{\prod_{t=1}^{T} \frac{\pi(a_t|x_t)}{\pi_i(a_t|x_t)}}_{\textbf{importance weight}}$$

For each π_i, $\widehat{D}(\tau_i, \pi, \pi_i)$ is a random variable

(by generating a trajectory τ_i)

Weighted Importance Return

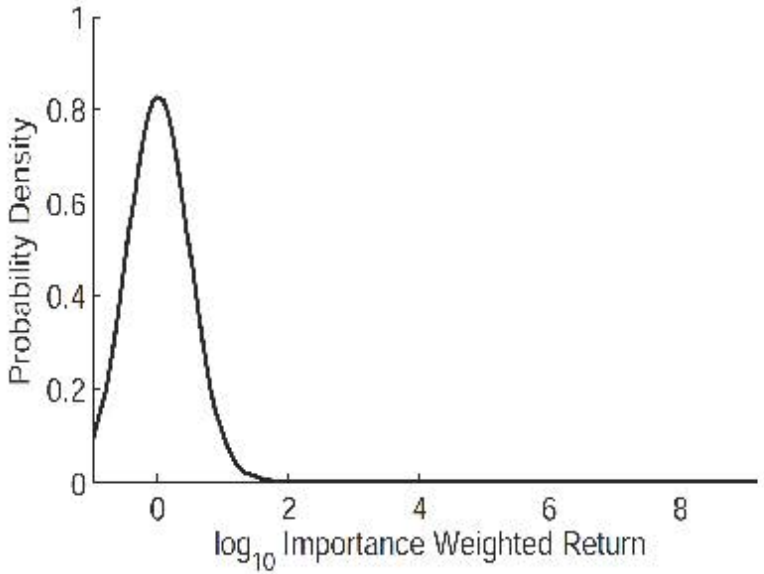
$\widehat{D}(\tau_i, \pi, \pi_i)$ is a random variable such that

- $\mathbb{E}\big[\widehat{D}(\tau_i, \pi, \pi_i)\big] \leq V^{\pi}$
  - if $\forall x, a \ni \pi_i(a|x) = 0 \Longrightarrow \pi(a|x) = 0$ then $\mathbb{E}\big[\widehat{D}(\tau_i, \pi, \pi_i)\big] = V^{\pi}$
- $\widehat{D}(\tau_i, \pi, \pi_i) \geq 0$ and $\mathbb{E}\big[\widehat{D}(\tau_i, \pi, \pi_i)\big] \in [0, 1/1-\gamma]$
- $\widehat{D}(\tau_i, \pi, \pi_i)$ may have a very large upper-bound ***(very long tail)***

Weighted Importance Return

$\widehat{D}(\tau_i, \pi, \pi_i)$ is a random variable such that

- $\mathbb{E}\big[\widehat{D}(\tau_i, \pi, \pi_i)\big] \leq V^{\pi}$
  - if $\forall x, a \ni \pi_i(a|x) = 0 \Longrightarrow \pi(a|x) = 0$ then $\mathbb{E}\big[\widehat{D}(\tau_i, \pi, \pi_i)\big] = V^{\pi}$
- $\widehat{D}(\tau_i, \pi, \pi_i) \geq 0$ and $\mathbb{E}\big[\widehat{D}(\tau_i, \pi, \pi_i)\big] \in [0, 1/1-\gamma]$
- $\widehat{D}(\tau_i, \pi, \pi_i)$ may have a very large upper-bound *(very long tail)*



- two policies in the Mountain Car problem
- T = 20
- PDF is estimated from 100,000 trajectories
- sample mean = 0.191
- maximum observed WIR = 316
- upper-bound on WIR = $10^{9.4}$

Problem Formulation

Given the data set $\mathcal{D} = \left\{(\tau_i, \pi_i)\right\}_{i=1}^{n}$, we have

- n independent bounded random variables

$$X_1 = \widehat{D}(\tau_1, \pi, \pi_1) \quad X_2 = \widehat{D}(\tau_2, \pi, \pi_2) \quad \dots \quad X_n = \widehat{D}(\tau_n, \pi, \pi_n)$$

- for all $i \in \{1, \dots, n\}$,

$$\Pr\left(X_i \in [0, b_i]\right) = 1 \qquad \textbf{and} \qquad \mathbb{E}[X_i] = \mu \leq V^{\pi}$$

Problem Formulation

Given the data set $\mathcal{D} = \left\{(\tau_i, \pi_i)\right\}_{i=1}^{n}$, we have

- n independent bounded random variables

$$X_1 = \widehat{D}(\tau_1, \pi, \pi_1) \quad X_2 = \widehat{D}(\tau_2, \pi, \pi_2) \quad \dots \quad X_n = \widehat{D}(\tau_n, \pi, \pi_n)$$

- for all $i \in \{1, \dots, n\}$,

$$\Pr\left(X_i \in [0, b_i]\right) = 1 \qquad \textbf{and} \qquad \mathbb{E}[X_i] = \mu \leq V^{\pi}$$

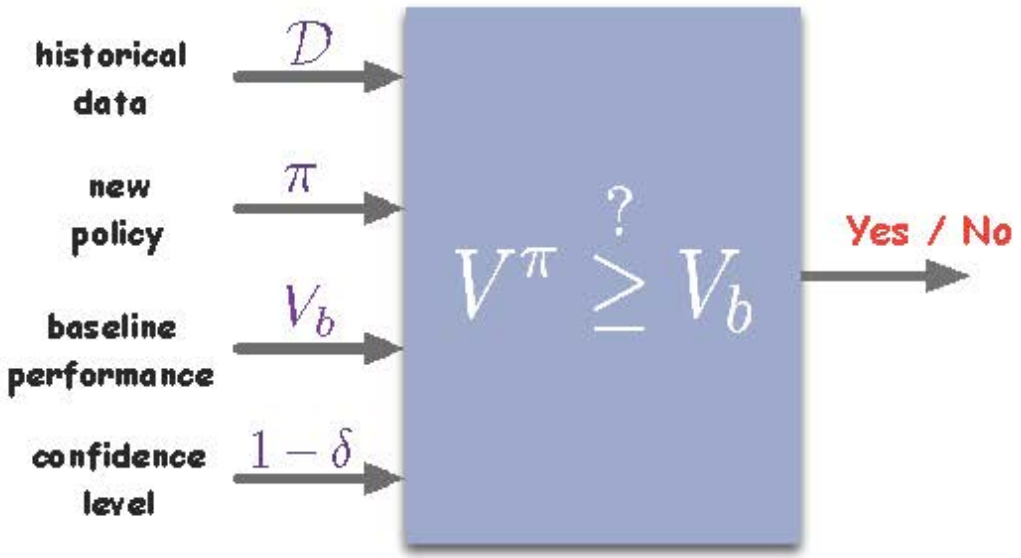
Main Objective: to derive a ***tight*** high probability (w.p. $\geq 1 - \delta$) lower-bound μ_{-} on μ

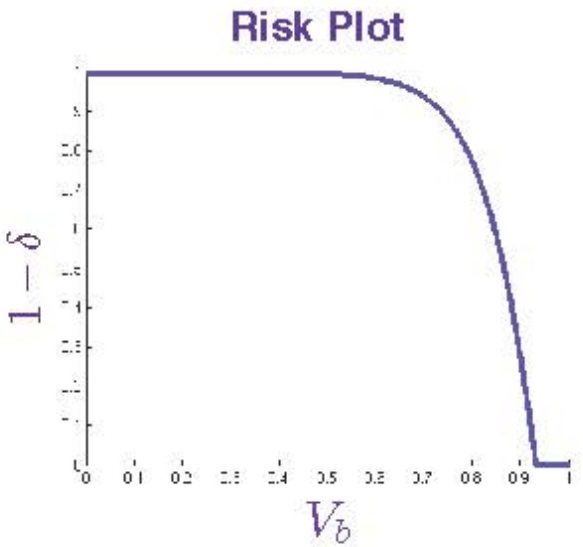
(note that b_i's can be very large)

Problem Formulation

if $\mu_- \geq V_b$, ***then*** so does V^π ***(w.p.*** $\geq 1-\delta$***)*** ***Yes***

otherwise ***no answer***

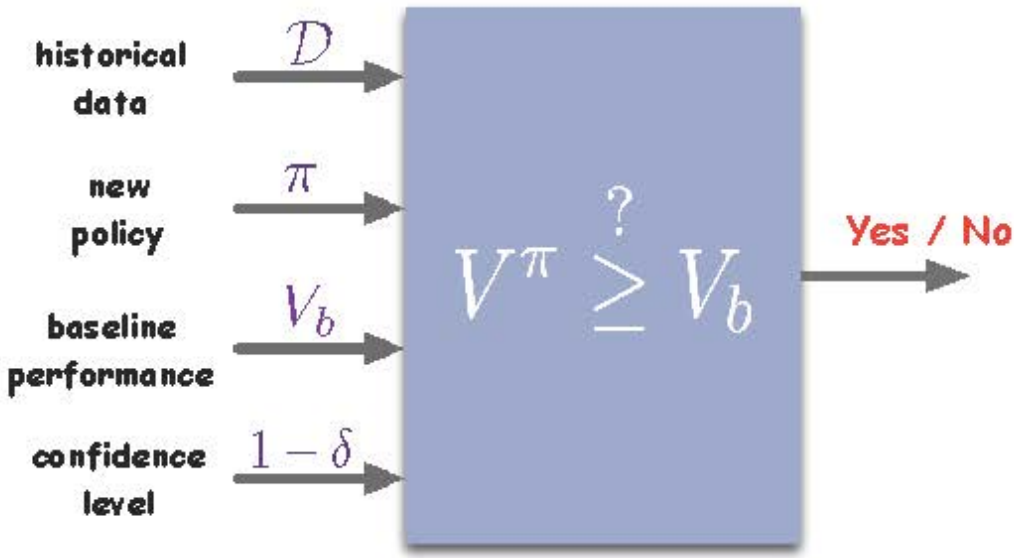


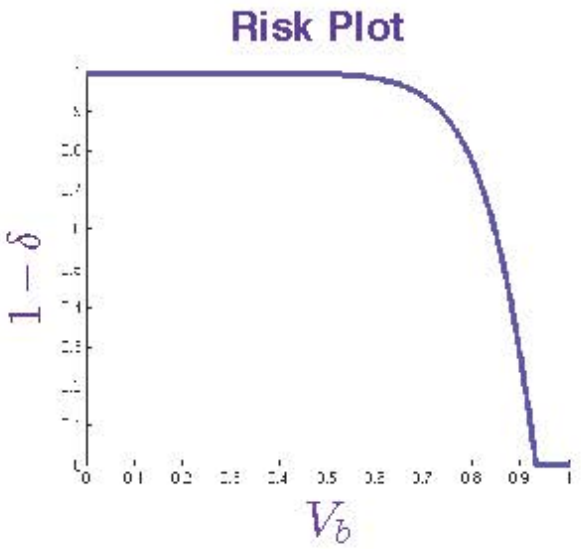


Problem Formulation

if $\mu_- \geq V_b$, ***then*** so does V^π ***(w.p.*** $\geq 1-\delta$***)*** ***Yes***

otherwise ***no answer***





how to derive a tight lower-bound on μ***???***

(note that b_i***'s can be very large)***

Concentration Inequalities

Chernoff-Hoeffding (CH)

$$\mu \geq \frac{1}{n}\sum_{i=1}^{n} X_i - b\sqrt{\frac{\log(1/\delta)}{2n}} \qquad \text{w.p.} \geq 1-\delta$$

Maurer-Pontil empirical Bernstein (MPeB) *(Maurer & Pontil, 2009)*

$$\mu \geq \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{7b\log(2/\delta)}{3(n-1)} - \frac{1}{n}\sqrt{\frac{\log(2/\delta)}{n-1}\sum_{i,j=1}^{n}(X_i - X_j)^2} \qquad \text{w.p.} \geq 1-\delta$$

Anderson-Massart (AM) *(Anderson, 1969; Massart, 1990)*

$$\mu \geq z_n - \sum_{i=0}^{n-1}(z_{i+1} - z_i)\ \min\left\{1, \frac{i}{n} + \sqrt{\frac{\log(2/\delta)}{2n}}\right\} \qquad \text{w.p.} \geq 1-\delta$$

$z_1 \leq z_2 \leq \ldots \leq z_n$ are sorted samples of the random variables $X_1, \ldots, X_n$

Concentration Inequalities

Chernoff-Hoeffding (CH)

$$\mu \geq \frac{1}{n}\sum_{i=1}^{n} X_i - b\sqrt{\frac{\log(1/\delta)}{2n}} \qquad \text{w.p.} \geq 1-\delta$$

Maurer-Pontil empirical Bernstein (MPeB) *(Maurer & Pontil, 2009)*

$$\mu \geq \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{7b\log(2/\delta)}{3(n-1)} - \frac{1}{n}\sqrt{\frac{\log(2/\delta)}{n-1}\sum_{i,j=1}^{n}(X_i - X_j)^2} \qquad \text{w.p.} \geq 1-\delta$$

Anderson-Massart (AM) *(Anderson, 1969; Massart, 1990)*

$$\mu \geq z_n - \sum_{i=0}^{n-1}(z_{i+1} - z_i)\,\min\left\{1, \frac{i}{n} + \sqrt{\frac{\log(2/\delta)}{2n}}\right\} \qquad \text{w.p.} \geq 1-\delta$$

$z_1 \leq z_2 \leq \ldots \leq z_n$ are sorted samples of the random variables $X_1, \ldots, X_n$

Concentration Inequalities

Chernoff-Hoeffding (CH)

(for i.d. and i.i.d.)

$$\mu \geq \frac{1}{n}\sum_{i=1}^{n} X_i - b\sqrt{\frac{\log(1/\delta)}{2n}} \qquad \text{w.p.} \geq 1-\delta$$

Maurer-Pontil empirical Bernstein (MPeB)

(for i.d. and i.i.d.)

$$\mu \geq \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{7b\log(2/\delta)}{3(n-1)} - \frac{1}{n}\sqrt{\frac{\log(2/\delta)}{n-1}\sum_{i,j=1}^{n}(X_i - X_j)^2} \qquad \text{w.p.} \geq 1-\delta$$

Anderson-Massart (AM)

(only for i.i.d.)

$$\mu \geq z_n - \sum_{i=0}^{n-1}(z_{i+1} - z_i)\ \min\left\{1, \frac{i}{n} + \sqrt{\frac{\log(2/\delta)}{2n}}\right\} \qquad \text{w.p.} \geq 1-\delta$$

$z_1 \leq z_2 \leq \ldots \leq z_n$ are sorted samples of the random variables $X_1, \ldots, X_n$

Concentration Inequalities

Maurer-Pontil empirical Bernstein (MPeB) ***(Maurer & Pontil, 2009)***

$$\mu \geq \underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{7b\log(2/\delta)}{3(n-1)} - \frac{1}{n}\sqrt{\frac{\log(2/\delta)}{n-1}\sum_{i,j=1}^{n}(X_i - X_j)^2}}_{\mu_-} \quad \text{w.p.} \geq 1-\delta$$

Concentration Inequalities

Maurer-Pontil empirical Bernstein (MPeB) ***(Maurer & Pontil, 2009)***

$$\mu \geq \underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{7b\log(2/\delta)}{3(n-1)} - \frac{1}{n}\sqrt{\frac{\log(2/\delta)}{n-1}\sum_{i,j=1}^{n}(X_i - X_j)^2}}_{\mu_-} \quad \text{w.p.} \geq 1-\delta$$

New MPeB bound: by collapsing the tails of the distributions of the random variables and then bounding the means of the new distributions ***a from of Winsorization*** ***(Wilcox & Keselman, 2003)***

New MPeB Bound

Theorem

Let $X_1, \ldots, X_n$ be n independent bounded ***(in*** $[0, b_i]$***)*** random variables with the ***same*** mean $\mathbb{E}[X_i] = \mu$. Let $Y_1, \ldots, Y_n$ such that $Y_i = \min\{X_i, c_i\}$ be bounded versions of $X_1, \ldots, X_n$ with ***fixed*** thresholds $c_i > 0$. Then ***w.p.*** $\geq 1 - \delta$, we have

$$\mu \geq \underbrace{\left(\sum_{i=1}^{n} \frac{1}{c_i}\right)^{-1} \sum_{i=1}^{n} \frac{Y_i}{c_i}}_{\text{empirical mean}} - \underbrace{\left(\sum_{i=1}^{n} \frac{1}{c_i}\right)^{-1} \frac{7n \log(2/\delta)}{3(n-1)}}_{\text{term that goes to zero as } 1/n}$$

$$- \underbrace{\left(\sum_{i=1}^{n} \frac{1}{c_i}\right)^{-1} \sqrt{\frac{\log(2/\delta)}{n-1} \sum_{i,j=1}^{n} \left(\frac{Y_i}{c_i} - \frac{Y_j}{c_j}\right)^2}}_{\text{term that goes to zero as } 1/\sqrt{n}}.$$

New MPeB Bound

Theorem

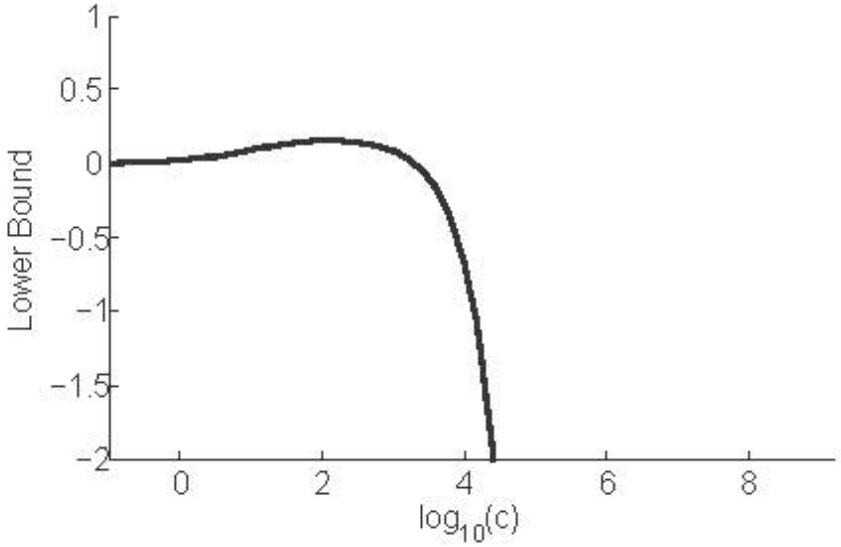
Let $X_1, \ldots, X_n$ be n independent bounded ***(in*** $[0, b_i]$***)*** random variables with the ***same*** mean $\mathbb{E}[X_i] = \mu$. Let $Y_1, \ldots, Y_n$ such that $Y_i = \min\{X_i, c_i\}$ be bounded versions of $X_1, \ldots, X_n$ with ***fixed*** thresholds $c_i > 0$. Then ***w.p.*** $\geq 1 - \delta$, we have

$$\mu \geq \underbrace{\left(\sum_{i=1}^{n} \frac{1}{c_i}\right)^{-1} \sum_{i=1}^{n} \frac{Y_i}{c_i}}_{\text{empirical mean}} - \underbrace{\left(\sum_{i=1}^{n} \frac{1}{c_i}\right)^{-1} \frac{7n \log(2/\delta)}{3(n-1)}}_{\text{term that goes to zero as } 1/n}$$

$$- \underbrace{\left(\sum_{i=1}^{n} \frac{1}{c_i}\right)^{-1} \sqrt{\frac{\log(2/\delta)}{n-1} \sum_{i,j=1}^{n} \left(\frac{Y_i}{c_i} - \frac{Y_j}{c_j}\right)^2}}_{\text{term that goes to zero as } 1/\sqrt{n}}.$$

how to select c_i***'s independent of the realization of*** X_i***'s???***

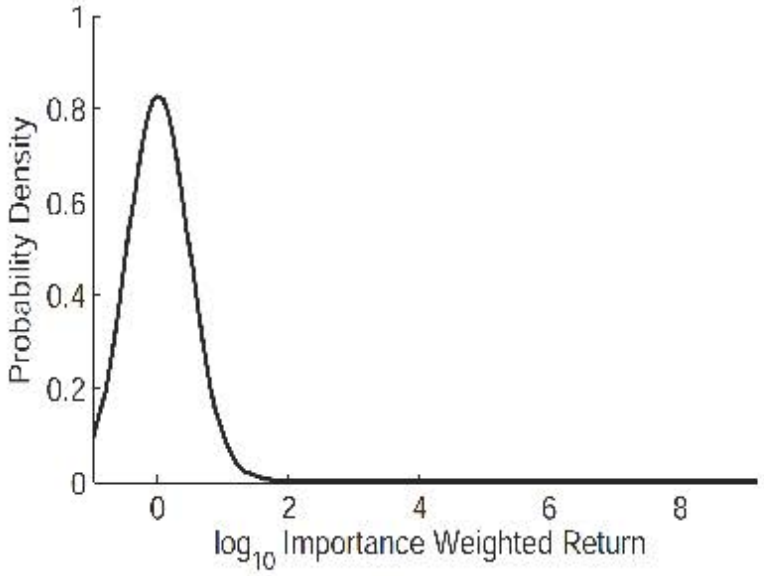
Selecting the Thresholds c_i's



- dividing the data $\mathcal{D}$ into two parts $\mathcal{D}_{\text{pre}}$ and $\mathcal{D}_{\text{post}}$
  ($1/20$ in $\mathcal{D}_{pre}$, $19/20$ in $\mathcal{D}_{post}$)
- calculating the optimal values of c_i by maximizing μ_- using $\mathcal{D}_{\text{pre}}$
  (computing c_i^*'s)
- calculating μ_- with c_i^*'s using $\mathcal{D}_{\text{post}}$
- ***if*** $\mu_- \geq V_b$, *return* ***Yes***, ***else*** *return* ***no answer***

Experimental Results

Results - Mountain Car Problem

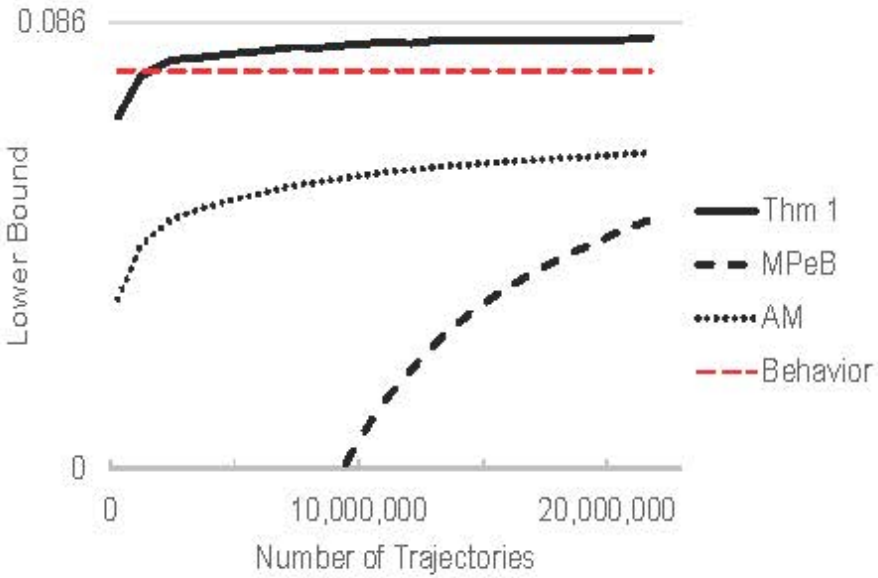


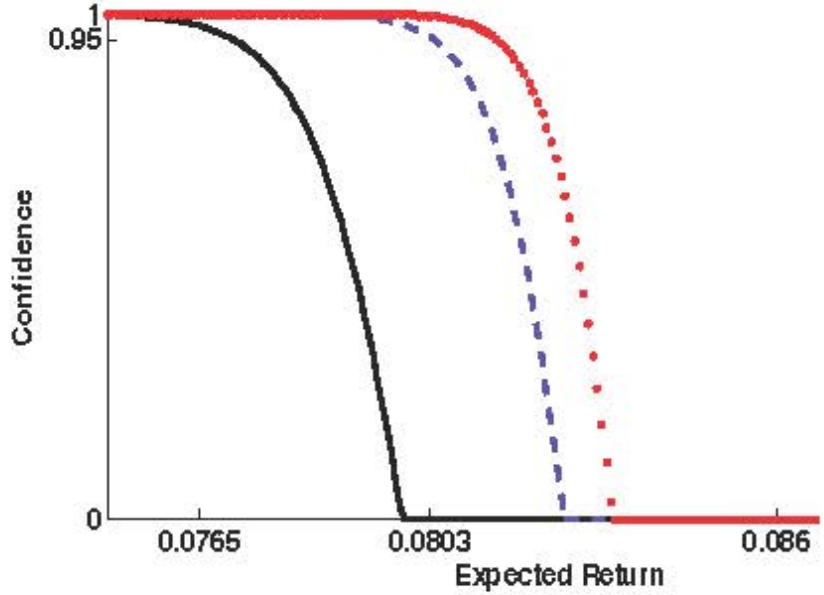
- two policies in the Mountain Car problem
- T = 20
- PDF is estimated from 100,000 trajectories
- sample mean = 0.191
- maximum observed WIR = 316
- upper-bound on WIR = $10^{9.4}$

95% *confidence lower-bounds*

	New MPeB	CH	MPeB	AM
μ_-	0.154	$-5{,}831{,}000$	$-129{,}703$	0.055

Results - Personalized Ad Recommendation *(simulated data)*





- target policy = a RL policy
- performance of behavior policy = 0.0765
- performance of target policy = 0.086
- trajectories of size T=20
- 95% confidence lower-bound

2M traj (behavior)
95% lower-bound = 0.077
5M traj (behavior)
95% lower-bound = 0.0803
5M traj (behavior) + 1M traj (target)
95% lower-bound = 0.081

Improving the Lower-Bound

MPeB lower-bound

general - no assumption on the data

One-sided Student's t-Test

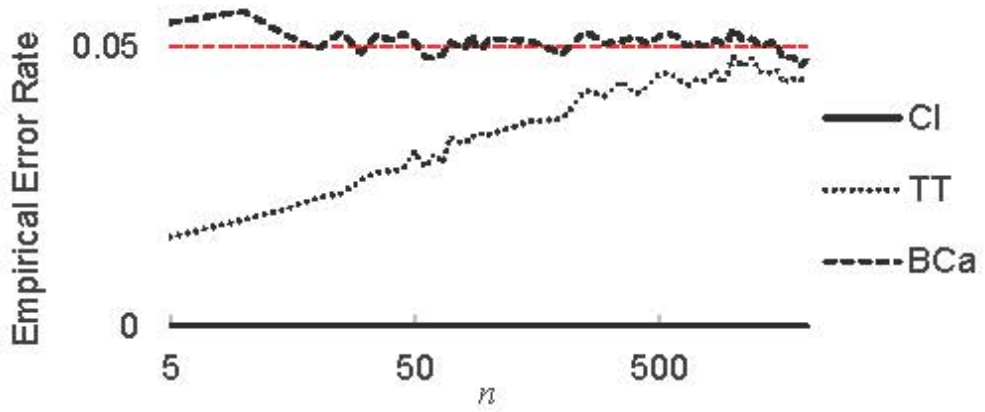
assumes $\widehat{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is normally distributed *(is true as $n \to \infty$ - CLT)*

Bias Corrected and accelerated (BCa) Bootstrap: *(Efron, 1987)*

use bootstrapping to estimate the true distribution of $\widehat{X}$ and then use it to produce a lower-bound

Improving the Lower-Bound

heavy upper-tail Gamma distribution

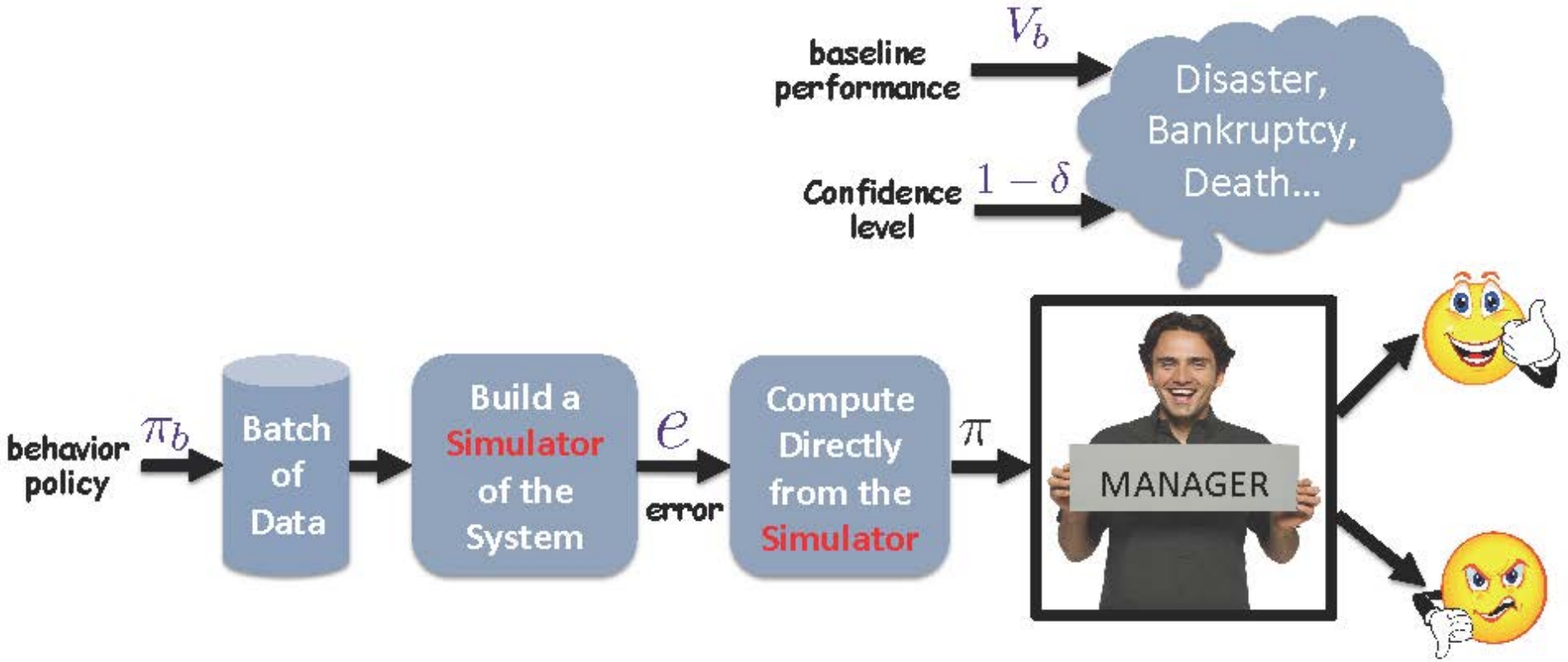


95% confidence lower-bound on the mean - 100,000 trials

Outline

Model-based Approach

Model-based Approach



Main Question: Given the simulator and the error in building it, how to compute a policy that is guaranteed ***(with a given confidence level)*** to perform at least as well as a baseline???

Safe Policy Improvement by Minimizing Robust Baseline Regret

M. Petrik, **MGH**, and Y. Chow. *"Safe Policy Improvement by Minimizing Robust Baseline Regret"*. ***NIPS-2016***.

Problem Definition *(Model Uncertainty)*

- **True Dynamics** $P^\star$
- **Simulator** $\widehat{P}$
- **Error Function**

$$\forall (x,a) \in \mathcal{X} \times \mathcal{A} \quad : \quad ||P^\star(\cdot|x,a) - \widehat{P}(\cdot|x,a)||_1 \leq e(x,a)$$

Problem Definition *(Model Uncertainty)*

- **True Dynamics** $P^\star$
- **Simulator** $\widehat{P}$
- **Error Function**

$$\forall (x,a) \in \mathcal{X} \times \mathcal{A} \quad : \quad ||P^\star(\cdot|x,a) - \widehat{P}(\cdot|x,a)||_1 \leq e(x,a)$$

- **Uncertainty Set**

$$\Xi(\widehat{P}, e) = \Big\{ \xi : \mathcal{X} \times \mathcal{A} \to \Delta^{\mathcal{X}} \ : \\ ||\xi(\cdot|x,a) - \widehat{P}(\cdot|x,a)||_1 \leq e(x,a), \ \forall (x,a) \in \mathcal{X} \times \mathcal{A} \Big\}$$

Problem Definition *(Safety)*

- **Baseline Policy** *(deterministic)* π_b
- **Return of a Policy π in world ξ** $V(\pi, \xi)$
- **True Return of a Policy π** $V(\pi, P^\star)$
- **True Optimal Policy** $\pi^\star \in \arg\max_\pi V(\pi, P^\star)$
- **Policy π is safe if** $V(\pi, P^\star) \geq V(\pi_b, P^\star)$

Problem Formulation I

Solving the Simulator

$$\pi_{\text{sim}} \in \arg\max_{\pi} \; V(\pi, \widehat{P})$$

Problem Formulation I

Solving the Simulator

$$\pi_{\text{sim}} \in \arg\max_{\pi} \; V(\pi, \widehat{P})$$

- no guarantee that π_{sim} is ***safe***

Theorem: Bound on Performance Loss of π_{sim}

$$\Phi(\pi_{\text{sim}}) \triangleq V(\pi^\star, P^\star) - V(\pi_{\text{sim}}, P^\star) \leq \frac{2\gamma R_{\max}}{(1-\gamma)^2} ||e||_\infty$$

Problem Formulation II

Solving the Robust MDP

$$\pi_0 \in \arg\max_{\pi} \min_{\xi \in \Xi} V(\pi, \xi) \tag{1}$$

Problem Formulation II

Solving the Robust MDP

$$\pi_0 \in \arg\max_{\pi} \min_{\xi \in \Xi} V(\pi, \xi) \tag{1}$$

$$\pi_R = \begin{cases} \pi_0 & \text{if} \quad \min_{\xi \in \Xi} V(\pi_0, \xi) > \max_{\xi \in \Xi} V(\pi_b, \xi), \\ \pi_b & \text{otherwise.} \end{cases}$$

Problem Formulation II

Solving the Robust MDP

$$\pi_0 \in \arg\max_{\pi} \min_{\xi \in \Xi} V(\pi, \xi) \tag{1}$$

$$\pi_R = \begin{cases} \pi_0 & \text{if} \quad \min_{\xi \in \Xi} V(\pi_0, \xi) > \max_{\xi \in \Xi} V(\pi_b, \xi), \\ \pi_b & \text{otherwise.} \end{cases}$$

- π_R is guaranteed to be ***safe***

Theorem: Bound on Performance Loss of π_R

$$\Phi(\pi_R) \leq \min \left\{ \frac{2\gamma R_{\max}}{(1-\gamma)^2} \left(||e_{\pi^*}||_{1,u^*_{\pi^*}} + ||e_{\pi_b}||_{1,u^*_{\pi_b}} \right), \Phi(\pi_b) \right\}$$

Our Proposed Formulation

Robust Policy Improvement

$$\pi_S \in \arg\max_{\pi} \min_{\xi \in \Xi} \Big(V(\pi, \xi) - V(\pi_b, \xi) \Big) \tag{2}$$

Our Proposed Formulation

Robust Policy Improvement

$$\pi_S \in \arg\max_{\pi} \min_{\xi \in \Xi} \Big(V(\pi, \xi) - V(\pi_b, \xi) \Big) \tag{2}$$

- π_S is guaranteed to be ***safe***
- π_S can outperform π_R by an arbitrarily large margin

Comparison with the Robust Solution

Solving the Robust MDP

$$\pi_0 \in \arg\max_{\pi} \min_{\xi \in \Xi} V(\pi, \xi)$$

$$\pi_R = \begin{cases} \pi_0 & \text{if} \quad \min_{\xi \in \Xi} V(\pi_0, \xi) > \max_{\xi \in \Xi} V(\pi_b, \xi), \\ \pi_b & \text{otherwise.} \end{cases}$$

Robust Policy Improvement

$$\pi_S \in \arg\max_{\pi} \min_{\xi \in \Xi} \Big(V(\pi, \xi) - V(\pi_b, \xi) \Big)$$

Comparison with the Robust Solution

Solving the Robust MDP

$$\pi_0 \in \arg\max_{\pi} \min_{\xi \in \Xi} V(\pi, \xi)$$

$$\pi_R = \begin{cases} \pi_0 & \text{if} \quad \min_{\xi \in \Xi} V(\pi_0, \xi) > \max_{\xi \in \Xi} V(\pi_b, \xi), \\ \pi_b & \text{otherwise.} \end{cases}$$

Robust Policy Improvement

$$\pi_S \in \arg\max_{\pi} \min_{\xi \in \Xi} \Big(V(\pi, \xi) - V(\pi_b, \xi) \Big)$$

$$\overbrace{\max_{\pi} \min_{\xi} \Big(V(\pi, \xi) - V(\pi_b, \xi) \Big)}^{\text{Robust Policy Improvement}} \geq \underbrace{\max_{\pi} \min_{\xi} V(\pi, \xi) - \max_{\xi} V(\pi_b, \xi)}_{\text{Robust MDP}}$$

Properties of Robust Policy Improvement Formulation

1. Policy Class

Theorem

Optimization problem (2) may not have solution in the space of deterministic policies and ignoring this may cause huge loss.

Properties of Robust Policy Improvement Formulation

1. Policy Class

Theorem

Optimization problem (2) may not have solution in the space of deterministic policies and ignoring this may cause huge loss.

2. Performance Bound

Theorem

$$\Phi(\pi_S) \leq \min \left\{ \frac{2\gamma R_{\max}}{(1-\gamma)^2} \left(||e_{\pi^*}||_{1,u^*_{\pi^*}} + ||e_{\pi_b}||_{1,u^*_{\pi_b}} \right), \Phi(\pi_b) \right\}$$

Properties of Robust Policy Improvement Formulation

1. Policy Class

Theorem

Optimization problem (2) may not have solution in the space of deterministic policies and ignoring this may cause huge loss.

2. Performance Bound

Theorem

$$\Phi(\pi_S) \leq \min \left\{ \frac{2\gamma R_{\max}}{(1-\gamma)^2} \left(||e_{\pi^*}||_{1,u^*_{\pi^*}} + ||e_{\pi_b}||_{1,u^*_{\pi_b}} \right), \Phi(\pi_b) \right\}$$

3. Computational Complexity

Theorem

Optimization problem (2) is NP-hard.

Solutions to Robust Policy Improvement Problem

1. Exact Solution with Extra Information

Assumption: Markov chain induced by π_b is known

Solutions to Robust Policy Improvement Problem

1. Exact Solution with Extra Information

Assumption: Markov chain induced by π_b is known

$$\forall x \in \mathcal{X}, \qquad \widehat{P}\big(\cdot\,|x, \pi_b(x)\big) = P^\star\big(\cdot\,|x, \pi_b(x)\big)$$

Solutions to Robust Policy Improvement Problem

1. Exact Solution with Extra Information

Assumption: Markov chain induced by π_b is known

$$\forall x \in \mathcal{X}, \qquad \widehat{P}\big(\cdot\,|x, \pi_b(x)\big) = P^{\star}\big(\cdot\,|x, \pi_b(x)\big)$$

$$\forall x \in \mathcal{X}, \qquad e\big(x, \pi_b(x)\big) = 0$$

Solutions to Robust Policy Improvement Problem

1. Exact Solution with Extra Information

Assumption: Markov chain induced by π_b is known

$$\forall x \in \mathcal{X}, \qquad \widehat{P}\big(\cdot\,|x, \pi_b(x)\big) = P^\star\big(\cdot\,|x, \pi_b(x)\big)$$

$$\forall x \in \mathcal{X}, \qquad e\big(x, \pi_b(x)\big) = 0$$

$$\forall \xi \in \Xi(\widehat{P}, e), \ \ \forall x \in \mathcal{X}, \qquad \xi\big(\cdot\,|x, \pi_b(x)\big) = \widehat{P}\big(\cdot\,|x, \pi_b(x)\big)$$

Solutions to Robust Policy Improvement Problem

1. Exact Solution with Extra Information

Assumption: Markov chain induced by π_b is known

$$\forall x \in \mathcal{X}, \qquad \widehat{P}\big(\cdot\,|x, \pi_b(x)\big) = P^\star\big(\cdot\,|x, \pi_b(x)\big)$$

$$\forall x \in \mathcal{X}, \qquad e\big(x, \pi_b(x)\big) = 0$$

$$\forall \xi \in \Xi(\widehat{P}, e), \ \ \forall x \in \mathcal{X}, \qquad \xi\big(\cdot\,|x, \pi_b(x)\big) = \widehat{P}\big(\cdot\,|x, \pi_b(x)\big)$$

$$\forall \xi \in \Xi(\widehat{P}, e), \qquad V(\pi_b, \xi) \text{ is fixed}$$

Solutions to Robust Policy Improvement Problem

1. Exact Solution with Extra Information

Assumption: Markov chain induced by π_b is known

$$\forall x \in \mathcal{X}, \qquad \widehat{P}\big(\cdot\,|x, \pi_b(x)\big) = P^\star\big(\cdot\,|x, \pi_b(x)\big)$$

$$\forall x \in \mathcal{X}, \qquad e\big(x, \pi_b(x)\big) = 0$$

$$\forall \xi \in \Xi(\widehat{P}, e), \ \ \forall x \in \mathcal{X}, \qquad \xi\big(\cdot\,|x, \pi_b(x)\big) = \widehat{P}\big(\cdot\,|x, \pi_b(x)\big)$$

$$\forall \xi \in \Xi(\widehat{P}, e), \qquad V(\pi_b, \xi) \text{ is fixed}$$

$$\underset{\pi \in \Pi_R}{\arg\max} \ \min_{\xi \in \Xi} \Big(V(\pi, \xi) - V(\pi_b, \xi)\Big) \longrightarrow \underbrace{\underset{\pi \in \Pi_R}{\arg\max} \ \min_{\xi \in \Xi} V(\pi, \xi)}_{\text{Robust MDP}}$$

Solutions to Robust Policy Improvement Problem

1. Exact Solution with Extra Information

Assumption: Markov chain induced by π_b is known

$$\forall x \in \mathcal{X}, \qquad \widehat{P}\big(\cdot\,|x, \pi_b(x)\big) = P^\star\big(\cdot\,|x, \pi_b(x)\big)$$

$$\forall x \in \mathcal{X}, \qquad e\big(x, \pi_b(x)\big) = 0$$

$$\forall \xi \in \Xi(\widehat{P}, e), \;\; \forall x \in \mathcal{X}, \qquad \xi\big(\cdot\,|x, \pi_b(x)\big) = \widehat{P}\big(\cdot\,|x, \pi_b(x)\big)$$

$$\forall \xi \in \Xi(\widehat{P}, e), \qquad V(\pi_b, \xi) \text{ is fixed}$$

$$\arg\max_{\pi \in \Pi_R} \min_{\xi \in \Xi} \Big(V(\pi, \xi) - V(\pi_b, \xi)\Big) \longrightarrow \overbrace{\underbrace{\arg\max_{\pi \in \Pi_R} \min_{\xi \in \Xi} V(\pi, \xi)}_{\text{Robust MDP}}}^{\text{strongly polynomial time}}$$

Solutions to Robust Policy Improvement Problem

2. A Heuristic Solution

Assumption: Simulator is ***accurate*** for all the actions suggested by π_b

$$\forall x \in \mathcal{X}, \qquad \widehat{P}\big(\cdot\,|x, \pi_b(x)\big) \approx P^{\star}\big(\cdot\,|x, \pi_b(x)\big)$$

Solutions to Robust Policy Improvement Problem

2. A Heuristic Solution

Assumption: Simulator is ***accurate*** for all the actions suggested by π_b

$$\forall x \in \mathcal{X}, \qquad \widehat{P}\big(\cdot\,|x, \pi_b(x)\big) \approx P^{\star}\big(\cdot\,|x, \pi_b(x)\big)$$

solve the robust MDP

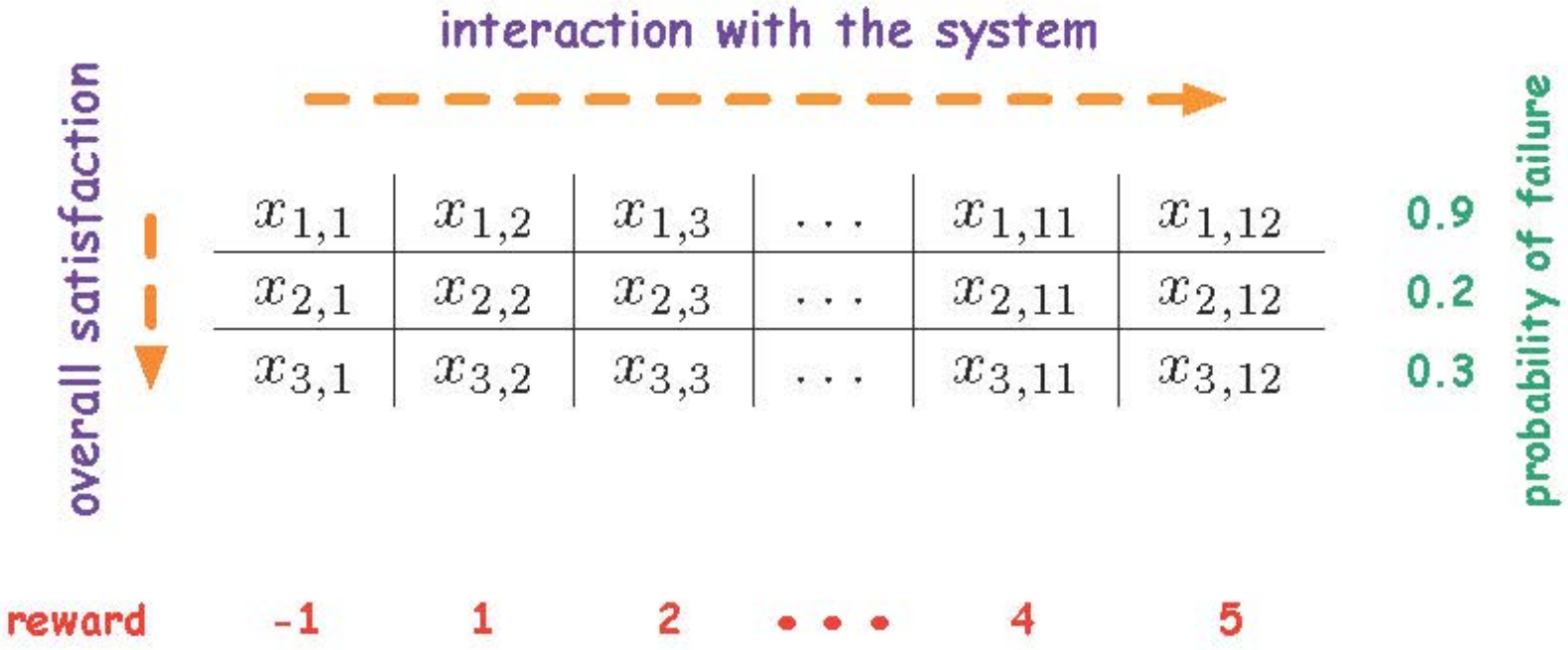
$$\underset{\pi \in \Pi_R}{\arg\max} \quad \underset{\xi \in \Xi(\widehat{P}, e)}{\min} \quad V(\pi, \xi)$$

where in the uncertainty set $\Xi(\widehat{P}, e)$, we force

$$\forall x \in \mathcal{X}, \qquad \widehat{P}\big(\cdot\,|x, \pi_b(x)\big) = P^{\star}\big(\cdot\,|x, \pi_b(x)\big)$$

Experimental Results

Grid Problem

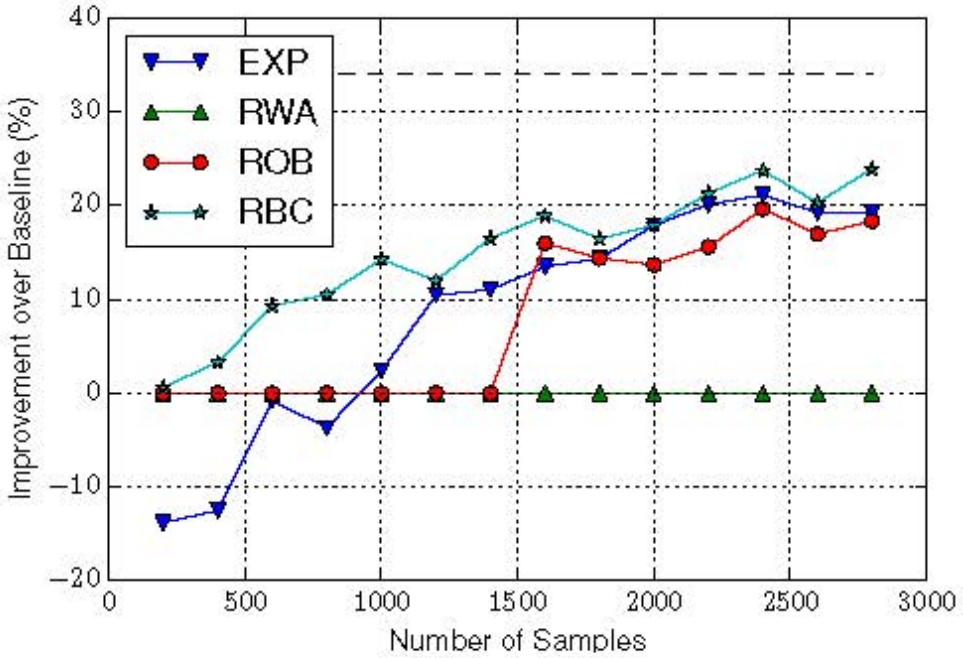


Actions: *left* *right* *up* *down*

Results - Grid Problem

baseline policy: optimal ignoring the row part of the state

simulator: built by samples from random policy



EXP: solving simulator

RWA: reward-adjusted MDP

ROB: robust MDP

RBC: robust policy improvement

Energy Arbitrage Problem

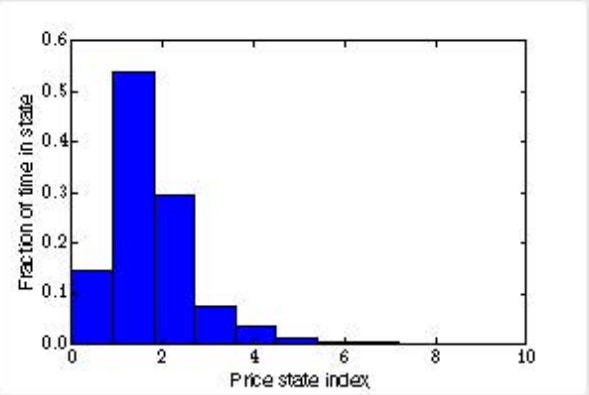
Problem Description

State: *charge level* *capacity* *price*

Action: amount of energy to buy or sell

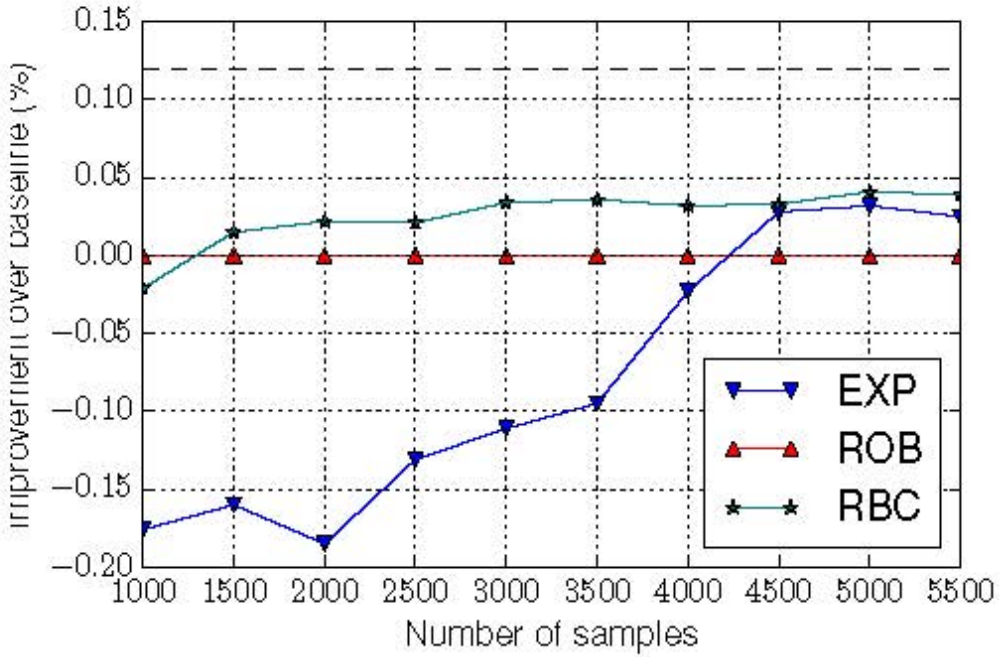
Dynamics:

- capacity degrades when energy is stored or retrieved
- model for charge level and capacity are ***known***
- model for price is ***unknown***
- price is discretized to 10 separate levels



Results - Energy Arbitrage Problem

baseline policy: optimal for the model in which price is discretized to 3 levels



EXP: solving simulator **ROB:** robust MDP

RBC: robust policy improvement

Summary

Safe Policy Improvement

Error Function $\forall(x,a) \in \mathcal{X} \times \mathcal{A} \quad : \quad ||P^*(\cdot|x,a) - \widehat{P}(\cdot|x,a)||_1 \leq e(x,a)$

Uncertainty Set

$$\Xi(\widehat{P}, e) = \left\{\xi : \mathcal{X} \times \mathcal{A} \to \Delta^{\mathcal{X}} \ : \ ||\xi(\cdot|x,a) - \widehat{P}(\cdot|x,a)||_1 \leq e(x,a),\ \forall(x,a) \in \mathcal{X} \times \mathcal{A}\right\}$$

Robust Policy Improvement

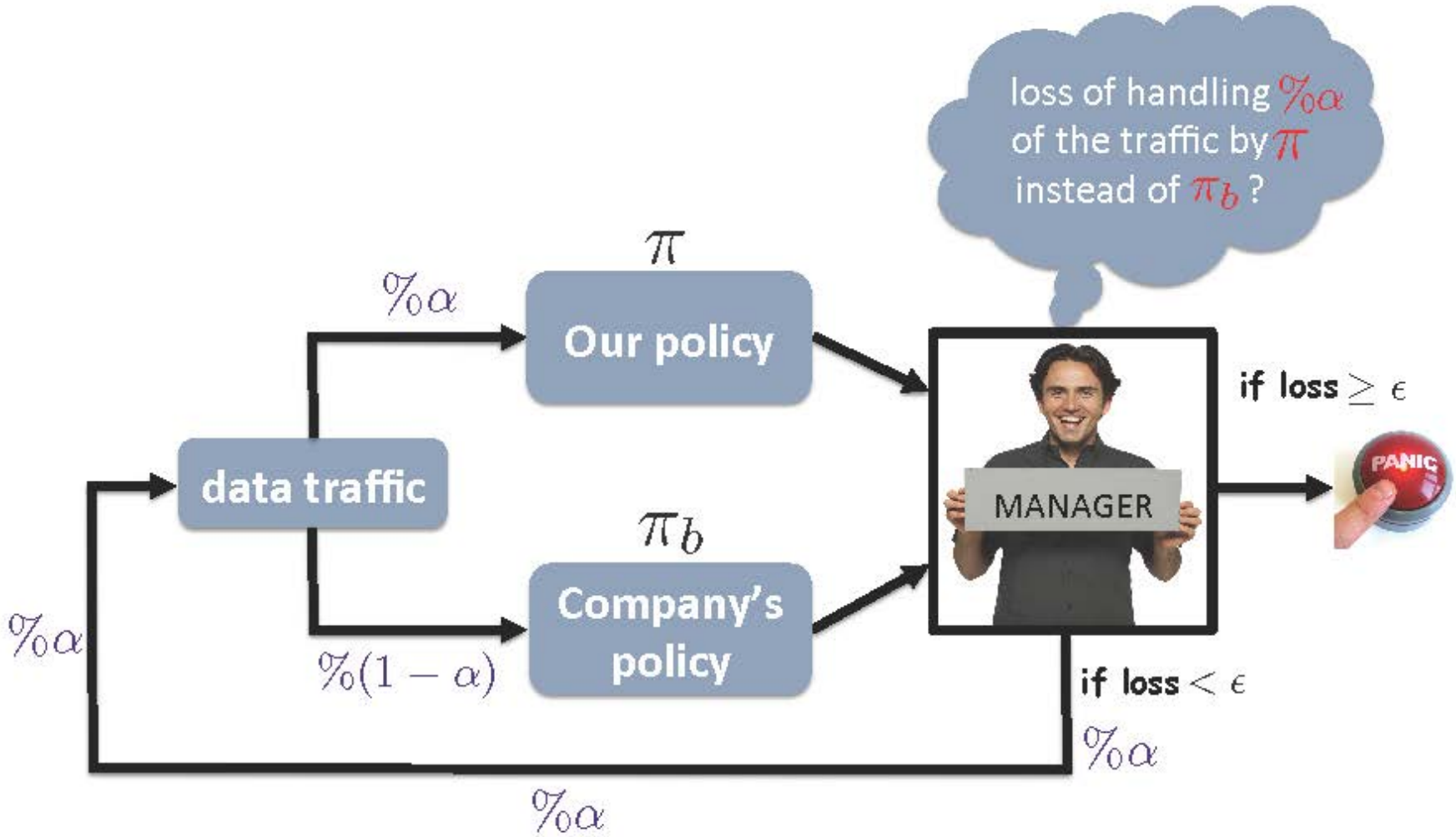
$$\pi_S \in \arg\max_{\pi} \min_{\xi \in \Xi} \Big(V(\pi, \xi) - V(\pi_b, \xi)\Big) \tag{3}$$

- π_S is guaranteed to be ***safe***
- π_S is less conservative than the other solutions
- (3) may not have solution in the space of deterministic policies
- Optimization problem (3) is NP-hard
- Under an assumption, (3) can be solved with a polynomial algorithm

Outline

Online Approach

Online Approach



Conservative Contextual Linear Bandits

A. Kazerouni, **MGH**, Y. Abbasi-Yadkori, and B. Van Roy. *"Conservative Contextual Linear Bandits"*. ***NIPS-2017***.

Contextual Linear Bandit

At each round t,

- Learner selects an action $a_t \in \mathcal{A}_t$ *($\mathcal{A}_t$ possibly infinite set)*
  - a_t is associated with the feature vector $\phi^t_{a_t} \in \mathbb{R}^d$

Contextual Linear Bandit

At each round t,

- Learner selects an action $a_t \in \mathcal{A}_t$ *($\mathcal{A}_t$ possibly infinite set)*
  - a_t is associated with the feature vector $\phi^t_{a_t} \in \mathbb{R}^d$
- Learner observes a random reward

$$Y_t = \langle \theta^*, \phi^t_{a_t} \rangle + \eta_t$$

  - $\langle \theta^*, \phi^t_{a_t} \rangle = \mathbb{E}[Y_t] = r^t_{a_t}$ *(expected reward)*
  - η_t is conditionally σ-sub-Gaussian noise
  - $\theta^* \in \mathbb{R}^d$ is the ***unknown*** parameter

Contextual Linear Bandit

At each round t,

- Learner selects an action $a_t \in \mathcal{A}_t$ *($\mathcal{A}_t$ possibly infinite set)*
  - a_t is associated with the feature vector $\phi^t_{a_t} \in \mathbb{R}^d$
- Learner observes a random reward

$$Y_t = \langle \theta^*, \phi^t_{a_t} \rangle + \eta_t$$

  - $\langle \theta^*, \phi^t_{a_t} \rangle = \mathbb{E}[Y_t] = r^t_{a_t}$ *(expected reward)*
  - η_t is conditionally σ-sub-Gaussian noise
  - $\theta^* \in \mathbb{R}^d$ is the ***unknown*** parameter

Assumption: $||\theta^*||_2 \leq B$ and $||\phi^t_a||_2 \leq D$ and $r^t_a \in [0, 1]$

Contextual Linear Bandit *(Main Objective)*

Optimal Action

$$a_t^* = \arg\max_{a \in \mathcal{A}_t} \langle \theta^*, \phi_a^t \rangle$$

Minimizing (pseudo)-Regret

$$R_T = \sum_{t=1}^{T} \langle \theta^*, \phi_{a_t^*}^t \rangle - \sum_{t=1}^{T} \langle \theta^*, \phi_{a_t}^t \rangle$$

Conservative Contextual Linear Bandit

same as contextual linear bandit ***except***

Conservative Contextual Linear Bandit

same as contextual linear bandit ***except***

- there exists a ***baseline*** policy π_b that at each round t
  - selects action $b_t \in \mathcal{A}_t$
  - observes expected reward $r^t_{b_t} = \langle \theta^*, \phi^t_{b_t} \rangle$ ($r^t_{b_t}$ ***is known***)

Conservative Contextual Linear Bandit

same as contextual linear bandit ***except***

- there exists a ***baseline*** policy π_b that at each round t
  - selects action $b_t \in \mathcal{A}_t$
  - observes expected reward $r^t_{b_t} = \langle \theta^*, \phi^t_{b_t} \rangle$ ($r^t_{b_t}$ ***is known***)
- **Performance Constraint:** At each round t, $\alpha \in (0, 1)$

$$\underbrace{\sum_{i=1}^{t} r^i_{b_i} - \sum_{i=1}^{t} r^i_{a_i}}_{\textit{cumulative loss}} \leq \alpha \sum_{i=1}^{t} r^i_{b_i}$$

Conservative Contextual Linear Bandit

same as contextual linear bandit ***except***

- there exists a ***baseline*** policy π_b that at each round t
  - selects action $b_t \in \mathcal{A}_t$
  - observes expected reward $r^t_{b_t} = \langle \theta^*, \phi^t_{b_t} \rangle$ ($r^t_{b_t}$ ***is known***)
- **Performance Constraint:** At each round t, $\alpha \in (0, 1)$

$$\underbrace{\sum_{i=1}^{t} r^i_{b_i} - \sum_{i=1}^{t} r^i_{a_i}}_{\textit{cumulative loss}} \leq \alpha \sum_{i=1}^{t} r^i_{b_i} \quad , \quad \sum_{i=1}^{t} r^i_{a_i} \geq (1-\alpha) \sum_{i=1}^{t} r^i_{b_i}$$

small α more conservative, large α less conservative

A Conservative Contextual Linear Bandit Algorithm

Given α and $r^t_{b_t}$, it should

minimize (pseudo)-regret

$$R_T = \sum_{t=1}^{T} \langle \theta^*, \phi^t_{a^*_t} \rangle - \sum_{t=1}^{T} \langle \theta^*, \phi^t_{a_t} \rangle$$

satisfy performance constraint

$$\sum_{i=1}^{t} r^i_{a_i} \geq (1-\alpha) \sum_{i=1}^{t} r^i_{b_i}$$

A Conservative Contextual Linear Bandit Algorithm

At each round t, the CLUCB algorthm

A Conservative Contextual Linear Bandit Algorithm

At each round t, the CLUCB algorthm

- uses the previous observations and builds a ***confidence set*** $\mathcal{C}_t$ ***(w.h.p. contains*** θ^* ***)***

A Conservative Contextual Linear Bandit Algorithm

At each round t, the CLUCB algorthm

- uses the previous observations and builds a ***confidence set*** $\mathcal{C}_t$ ***(w.h.p. contains*** θ^* ***)***
- computes the ***optimistic action***

$$a'_t = \arg\max_{a \in \mathcal{A}_t} \max_{\theta \in \mathcal{C}_t} \langle \theta, \phi_a^t \rangle$$

A Conservative Contextual Linear Bandit Algorithm

At each round t, the CLUCB algorithm

- uses the previous observations and builds a ***confidence set*** $\mathcal{C}_t$ ***(w.h.p. contains*** θ^* ***)***
- computes the ***optimistic action***

$$a'_t = \arg\max_{a \in \mathcal{A}_t} \max_{\theta \in \mathcal{C}_t} \langle \theta, \phi^t_a \rangle$$

- plays the ***optimistic action*** $(a_t = a'_t)$, only if

$$\overbrace{\sum_{i \in S^o_{t-1}} r^i_{b_i}}^{\text{baseline rounds}} + \min_{\theta \in \mathcal{C}_t} \left(\underbrace{\langle \theta, \phi^t_{a'_t} \rangle}_{\text{round } t} + \overbrace{\sum_{i \in S_{t-1}} \langle \theta, \phi^i_{a'_i} \rangle}^{\text{optimistic rounds}} \right) \geq (1-\alpha) \sum_{i=1}^{t} r^i_{b_i}$$

plays the ***baseline action*** $(a_t = b_t)$, otherwise.

Construction of Confidence Sets *(Abbasi-Yadkori et al., 2011)*

At each round t,

given the observed data $\left\{(\phi^i_{a_i}, Y_i)\right\}_{i=1}^{|S_t|}$, CLUCB updates the ***confidence set*** as *(S_t = set of rounds we play optimistic)*

$$C_{t+1} = \left\{\theta \in \mathbb{R}^d \ : \ ||\theta - \widehat{\theta}||_{V_t} \leq \beta_{t+1}\right\}$$

$$\widehat{\theta}_t = (\Phi_t \Phi_t^\top + \lambda I)^{-1} \Phi_t Y_t \quad , \quad V_t = \lambda I + \Phi_t \Phi_t^\top$$

$$\beta_{t+1} = \sigma \sqrt{d \log\left(\frac{1 + (|S_t| + 1)D^2/\lambda}{\delta}\right)} + \sqrt{\lambda} B$$

Construction of Confidence Sets *(Abbasi-Yadkori et al., 2011)*

At each round t,

given the observed data $\left\{(\phi^i_{a_i}, Y_i)\right\}_{i=1}^{|S_t|}$, CLUCB updates the ***confidence set*** as *(S_t = set of rounds we play optimistic)*

$$C_{t+1} = \left\{\theta \in \mathbb{R}^d \ : \ ||\theta - \widehat{\theta}||_{V_t} \leq \beta_{t+1}\right\}$$

$$\widehat{\theta}_t = (\Phi_t \Phi_t^\top + \lambda I)^{-1} \Phi_t Y_t \quad , \quad V_t = \lambda I + \Phi_t \Phi_t^\top$$

$$\beta_{t+1} = \sigma \sqrt{d \log\left(\frac{1 + (|S_t| + 1)D^2/\lambda}{\delta}\right)} + \sqrt{\lambda} B$$

Proposition: For any C_t and $\delta > 0$, we have $\mathbb{P}\left[\theta^* \in C_t, \ \forall t \in \mathbb{N}\right] \geq 1 - \delta$.

Regret Analysis

Assumption

There exists $0 \leq \Delta_l \leq \Delta_h$ and $0 < r_l$ such that at each round t,

$$\Delta_l \leq \Delta_{b_t}^t = r_{a_t^*}^t - r_{b_t}^t \leq \Delta_h \qquad \text{and} \qquad r_l \leq r_{b_t}^t.$$

Proposition

The regret of CLUCB can be decomposed into two terms as follows:

$$R_T(\text{CLUCB}) \leq R_{S_T}(\text{LUCB}) + |S_T^c|\Delta_h$$

Regret Analysis

Theorem

With probability at least $1-\delta$, CLUCB satisfies the performance constraint for all $t \in \mathbb{N}$ and has the regret bound

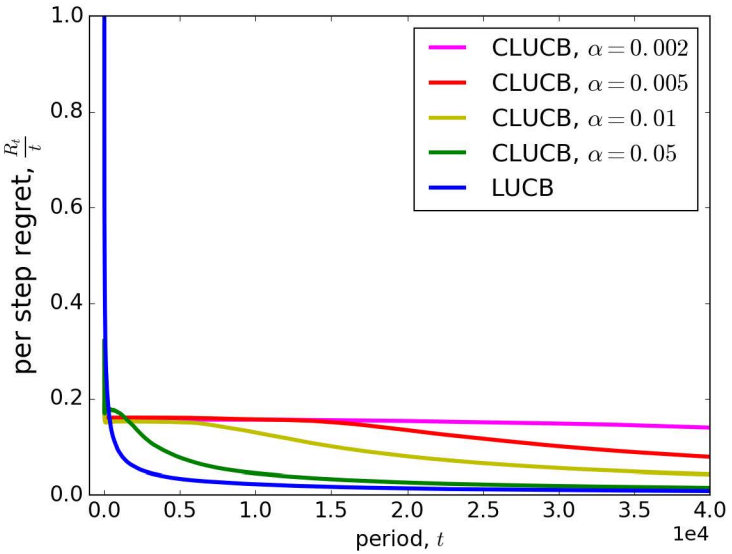
$$R_T(CLUCB) = O\left(\overbrace{\underbrace{d\sqrt{T}\log\left(\frac{DT}{\lambda\delta}\right)}_{R_{S_T}(\text{LUCB})}}^{O\left(\sqrt{T}\log(T)\right)} + \underbrace{\frac{K\Delta_h}{\alpha r_l}}_{\text{constant}} \right)$$

$$K = 1 + 114d^2 \frac{(B\sqrt{\lambda}+\sigma)^2}{\Delta_l + \alpha r_l} \left[\log\left(\frac{62d(B\sqrt{\lambda}+\sigma)}{\sqrt{\delta}(\Delta_l+\alpha r_l)}\right)\right]^2$$

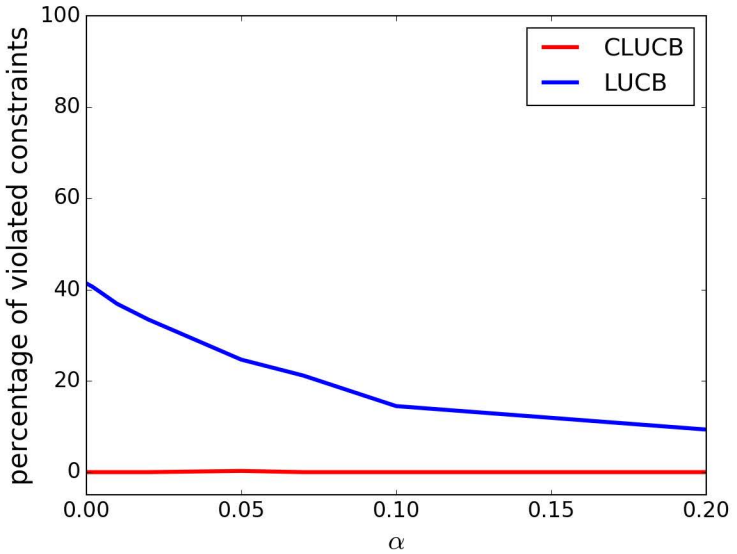
Experimental Results

$$d = 4, \quad \lambda = 1, \quad \delta = 0.001, \quad \theta^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I_4}), \quad \eta_t \sim \mathcal{N}(0, 1)$$

$$\#\text{ of arms} = 100, \quad \text{baseline} = 10\text{'th best arm}, \quad \text{averaged over } 1,000 \text{ runs}$$



per-step regret



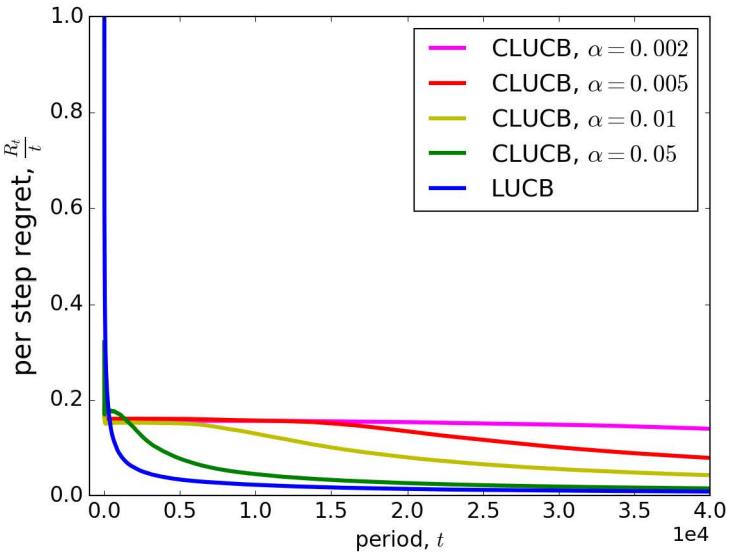
% constraint violation in first 1000 rounds

CLUCB starts to play optimistically more quickly for larger values of α

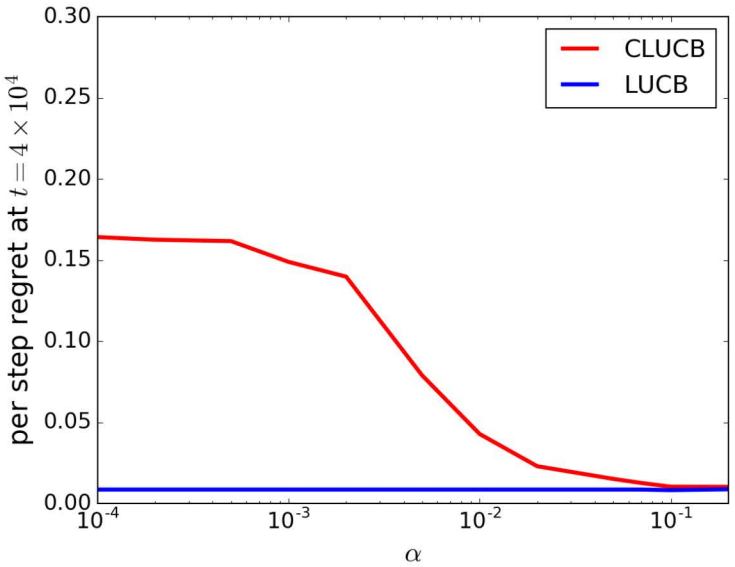
Experimental Results

$$d = 4, \quad \lambda = 1, \quad \delta = 0.001, \quad \theta^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I_4}), \quad \eta_t \sim \mathcal{N}(0, 1)$$

$\#$ of arms $= 100$, baseline $= 10$'th best arm, averaged over $1,000$ runs



per-step regret



per-step regret at $t = 40,000$

performance of CLUCB converges to that of LUCB more quickly for larger values of α

Summary

Conservative Contextual Linear Bandit

Given α and $r^t_{b_t}$,

minimize (pseudo)-regret

$$R_T = \sum_{t=1}^{T} \langle \theta^*, \phi^t_{a^*_t} \rangle - \sum_{t=1}^{T} \langle \theta^*, \phi^t_{a_t} \rangle$$

s.t. the performance constraint

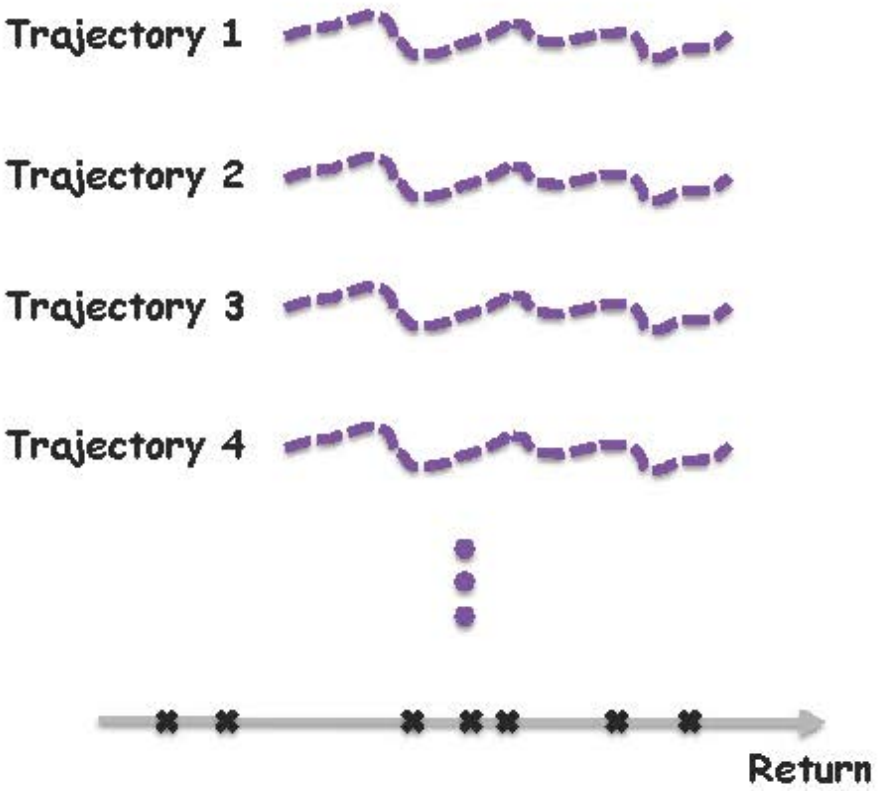
$$\forall t \in \{1, \dots, T\} \qquad \sum_{i=1}^{t} r^i_{a_i} - \sum_{i=1}^{t} r^i_{b_i} \leq \alpha \sum_{i=1}^{t} r^i_{b_i}$$

Outline

Risk-sensitive Decision-making
(brief overview)

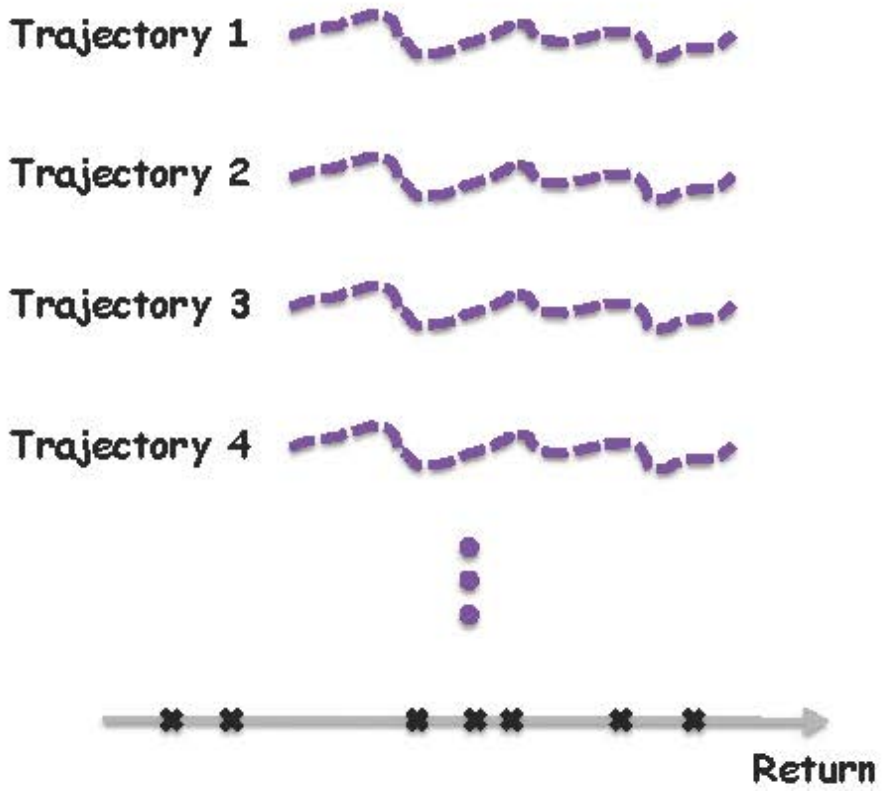
Risk-sensitive Decision-making

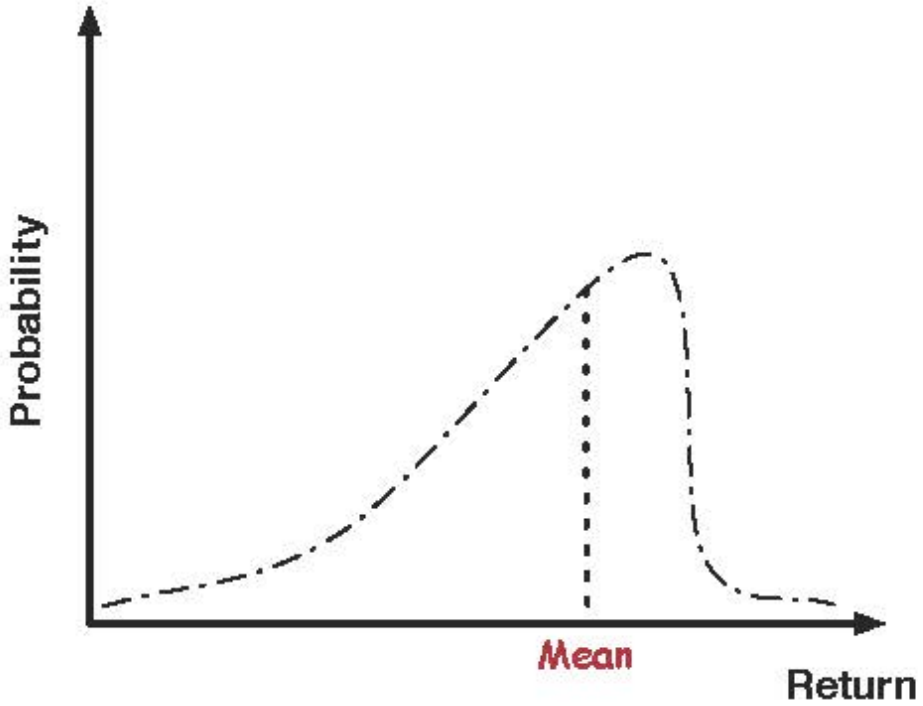
Policy π



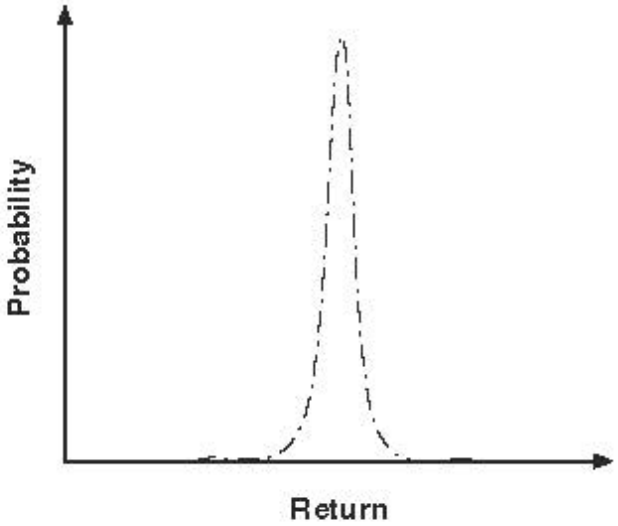
Risk-sensitive Decision-making

Policy π



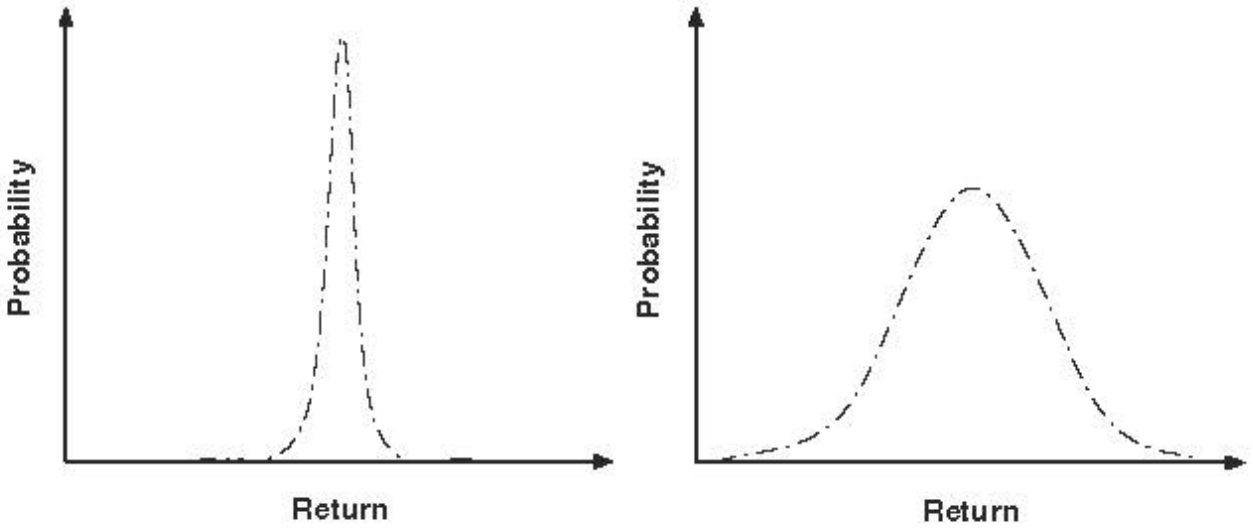


Risk-sensitive Decision-making



$$\max_{\pi} \ \text{Mean}(D^{\pi})$$

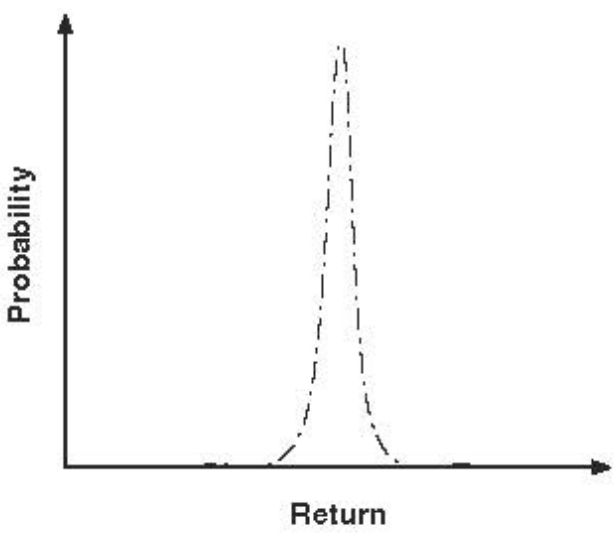
Risk-sensitive Decision-making

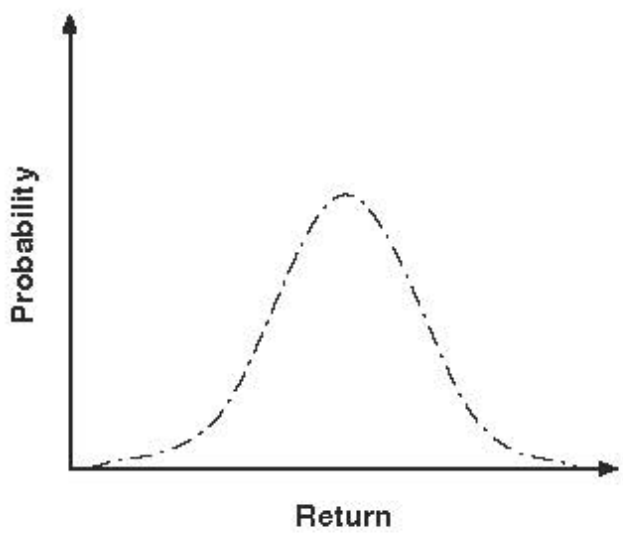


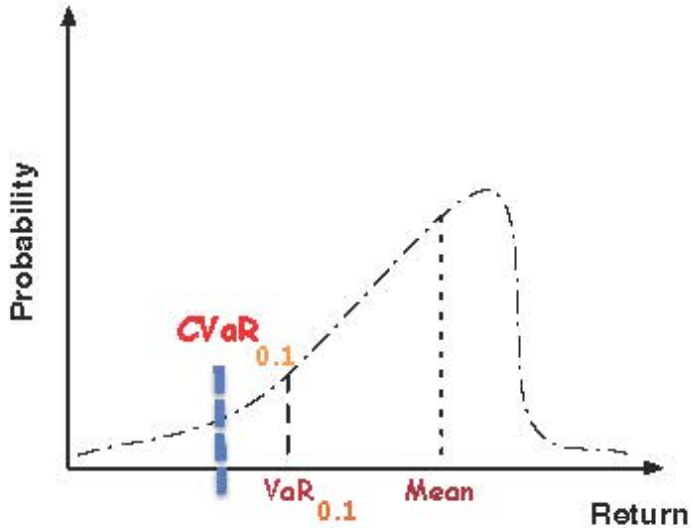
$$\max_{\pi} \quad \text{Mean}(D^{\pi})$$

$$\max_{\pi} \quad \text{Mean}(D^{\pi})$$
$$\text{s.t.} \quad \text{var}(D^{\pi}) \leq \beta$$

Risk-sensitive Decision-making







$$\max_{\pi} \quad \text{Mean}(D^{\pi})$$

$$\max_{\pi} \quad \text{Mean}(D^{\pi})$$
$$\text{s.t.} \quad \text{var}(D^{\pi}) \leq \beta$$

$$\max_{\pi} \quad \text{Mean}(D^{\pi})$$
$$\text{s.t.} \quad \text{CVaR}_{\alpha}(D^{\pi}) \geq \beta$$

Publications

1. Prashanth L. A. and **MGH**. *"Actor-Critic Algorithms for Risk-Sensitive MDPs"*. ***NIPS-2013***.

2. Y. Chow and **MGH**. *"Algorithms for CVaR Optimization in MDPs"*. ***NIPS-2014***.

3. A. Tamar, Y. Chow, **MGH**, and S. Mannor. *"Policy Gradient for Coherent Risk Measures"*. ***NIPS-2015***.

4. Prashanth L. A. and **MGH**. *"Variance-constrained Actor-Critic Algorithms for Discounted and Average Reward MDPs"*. ***MLJ-2016***.

5. A. Tamar, Y. Chow, **MGH**, and S. Mannor. *"Optimization of Coherent Risk Measures"*. ***IEEE TAC 2017***.

6. Y. Chow, **MGH**, L. Janson, and M. Pavone. *"Risk-Constrained Reinforcement Learning with Percentile Risk Criteria"*. ***JMLR-2017***.

7. NEW! J. Lacotte, Y. Chow, **MGH**, and M. Pavone. *"Risk-sensitive Generative Adversarial Imitation Learning"*. ***submitted***.

8. NEW! B. Liu and **MGH**. *"A Block Coordinate Ascent Algorithm for Mean-Variance Optimization"*. ***submitted***.

Tutorial on

Risk-averse Decision-making & Control

Marek Petrik and Mohammad Ghavamzadeh

AAAI-2017

Outline

Safety w.r.t. Undesirable Situations
(brief overview)

Safe RL

CMDP Formulation

$$\min_{\pi} \mathbb{E}\left[\sum_{t=0}^{T-1} c(x_t, a_t) \mid x_0, \pi\right] \quad , \quad \text{s.t.} \quad \mathbb{E}\left[\sum_{t=0}^{T-1} d(x_t) \mid x_0, \pi\right] \leq d_0$$

Safe Policy Iteration (SPI)

1. finding the Lyapunov function

$$\max_{\epsilon: \mathcal{X} \to \mathbb{R}+} ||\epsilon||_1 \quad , \quad \text{s.t.} \quad \mathcal{T}_{d+\epsilon}^{\pi_k}[L_k](x) = L_k(x), \; \forall x \in \mathcal{X} \quad , \quad L_k(x_0) \leq d_0$$

$$L_k(x) = V_{d+\epsilon}^{\pi_k}(x), \; \forall x \in \mathcal{X}$$

2. policy evaluation $\quad V_k = V_c^{\pi_k}$
3. policy improvement $\quad \pi_{k+1} \in \arg\min_{\pi \in \mathcal{F}_{L_k}(x)} \mathcal{T}_c^{\pi}[V_k]$

$$\mathcal{F}_{L_k}(x) = \left\{\pi(\cdot|x) \mid \mathcal{T}_d^{\pi}[L_k](x) \leq L_k(x)\right\}$$

(a) all π_k's are safe, **(b)** π_{k+1} is no worse than π_k, **(c)** SPI converges

Safe RL

SPI / Large Penalty | **Optimal / Small Penalty**

Thank you!!

Mohammad Ghavamzadeh

ghavamza@google.com OR
mohammad.ghavamzadeh@inria.fr