Percentile Optimization in Uncertain Markov Decision Processes with Application to Efficient Exploration

Erick Delage

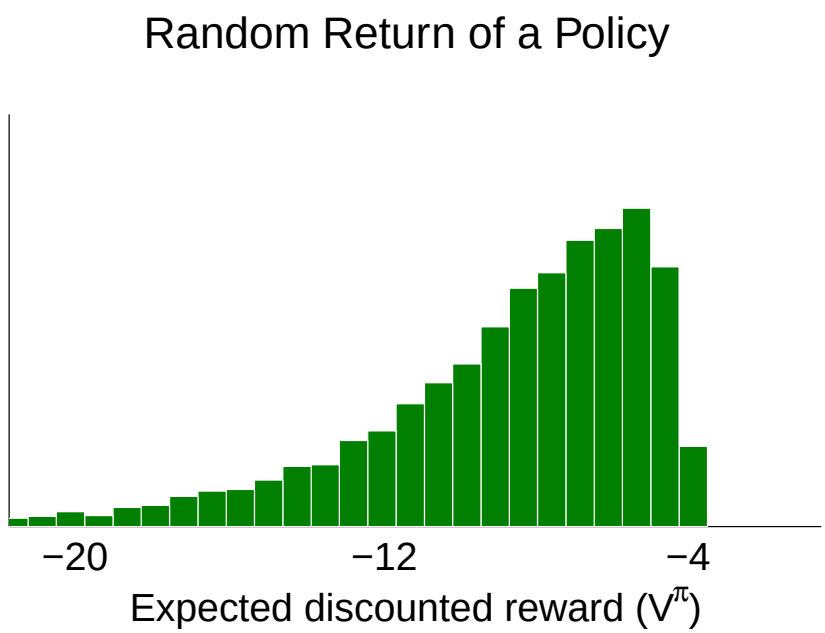
Stanford University

Joint work with: Shie Mannor (McGill University)

Vanilla Data-Driven MDP

- Assume system behaves as an MDP
- Gather transition and reward data
- Estimate R and $P(s'|s,a)$ parameters
- Maximize long term expected discounted reward
- Assume true MDP behaves as expected one

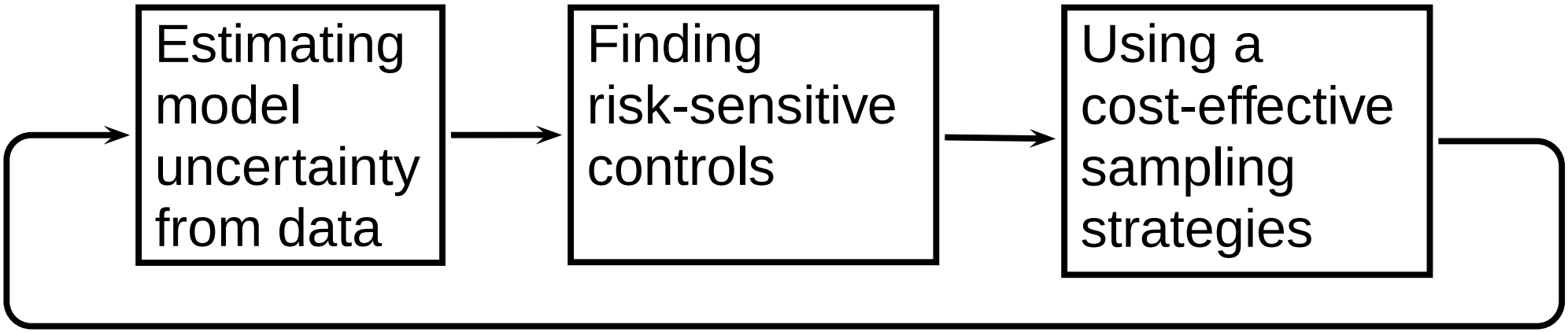
The Curse of Parameter Uncertainty



- Parameter uncertainty is always present
- We cannot always afford to make it negligible
- Robust methods are difficult to apply and deceptively conservative

Our Approach

We propose a complete risk-sensitive method for addressing data-driven Markov decision processes.



Data-Driven MDP Approach

Estimating model uncertainty from data

A Distribution over MDP Models

A Gaussian prior on Rewards:

- Prior belief is $R(i,a) \propto \mathcal{N}(\mu_{(i,a)}, \sigma^2_{(i,a)})$
- Given a new measurement $\hat{R}(i,a) = R(i,a) + \nu$ (Gaussian noise), belief remains Gaussian

A Dirichlet prior on transition parameters $P(.|i,a) = \overrightarrow{p}$:

- Prior belief on $\overrightarrow{p}$ is $f(\overrightarrow{p}) \propto \prod_{j=1}^{|\mathcal{S}|} p_j^{\beta_j - 1}$
- Given a new transition from (i,a), belief remains a Dirichlet distribution.

The Bayesian Approach

We have a probability over models:

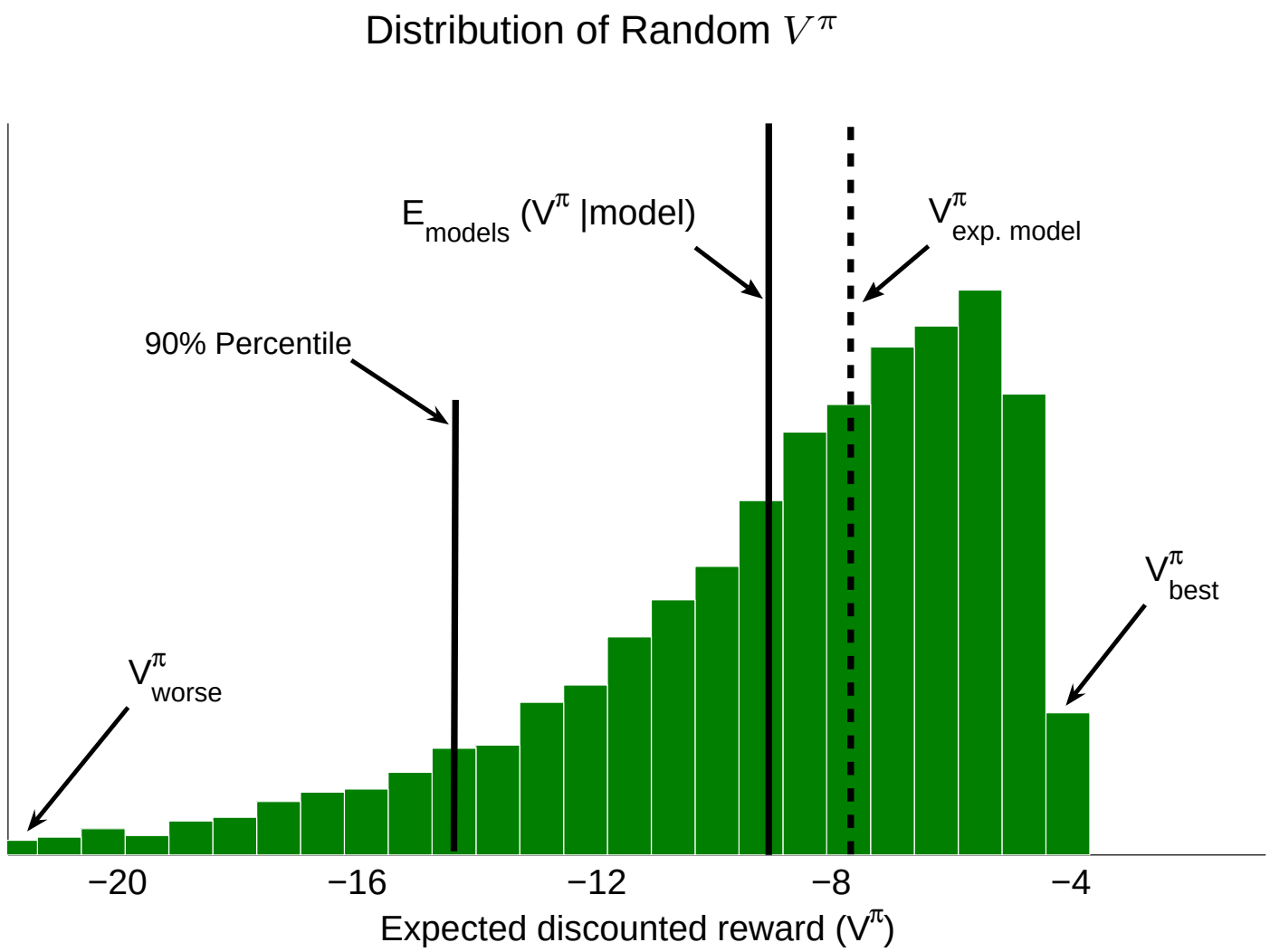
Consider $V^\pi = \mathbb{E}_x^\pi[\sum_{t=0}^\infty \alpha^t R(x_t)]$ as a random variable.

For a given π and a current belief we can ask what is:

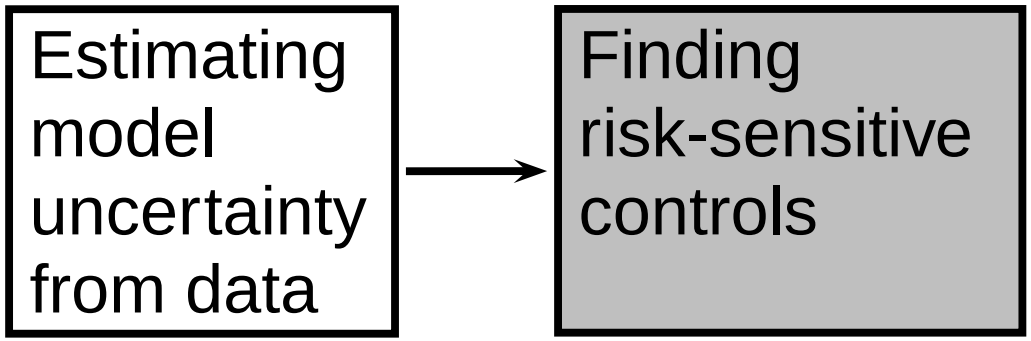
$$\mathbb{E}_{\text{models}}\left[V^\pi\right] = \mathbb{E}_{\text{models}}\left[\mathbb{E}_x^\pi\left[\sum_{t=0}^{\infty}\alpha^t R(x_t)\right]\right]$$

What if true MDP does not behave like $\mathbb{E}_{\text{models}}[V^\pi]$?

The Curse of Parameter Uncertainty



Data-Driven MDP Approach



Percentile Optimization

Find optimal policy according to:

$$\max_{\text{policy } \pi, y \in \mathbb{R}} \quad y$$

$$\text{sub. to} \quad \mathbb{P}_{\text{models}} \left(\mathbb{E}_x^\pi \left(\sum_{t=0}^{\infty} \alpha^t R(x_t) \right) \geq y \right) \geq \eta \ ,$$

Value-at-risk: η is the risk parameter.

Percentile Optimization

Find optimal policy according to:

$$\max_{\text{policy } \pi, y \in \mathbb{R}} \quad y$$

$$\text{sub. to} \quad \mathbb{P}_{\text{models}} \left(\mathbb{E}_x^{\pi} \left(\sum_{t=0}^{\infty} \alpha^t R(x_t) \right) \geq y \right) \geq \eta \ ,$$

Value-at-risk: η is the risk parameter.

It turns out that solving the percentile optimization is:

- NP-hard in general
- NP-hard even if transitions are known
- **Polytime for Gaussian reward parameters**

Percentile Optimization : Transitions (I)

It is already hard to solve:

$$\max_{\text{policy } \pi} \mathbb{E}_{\text{models}} \left[\mathbb{E}_x^\pi \left[\sum_{t=0}^{\infty} \alpha^t R_t \right] \right]$$

equivalent to:

$$\max_{\pi} \mathbb{E}_{\text{models}} \left[(I - \alpha P_\pi^{\text{model}})^{-1} R \right]$$

The objective depends non-linearly on all moments of P.

Percentile Optimization : Transitions (II)

Let $\mathbb{F}(\pi)$ be the second order approximation

$$\mathbb{F}(\pi) = q^{\mathsf{T}} X^{\pi} R + \alpha^2 q^{\mathsf{T}} X^{\pi} \Pi Q^{\pi} X^{\pi} R$$

- $\mathbb{F}(\pi)$ only depends on first and second moments of P
- Optimizing $\mathbb{F}(\pi)$ is tractable for problem $\approx$ 1000 states

Percentile Optimization : Transitions (II)

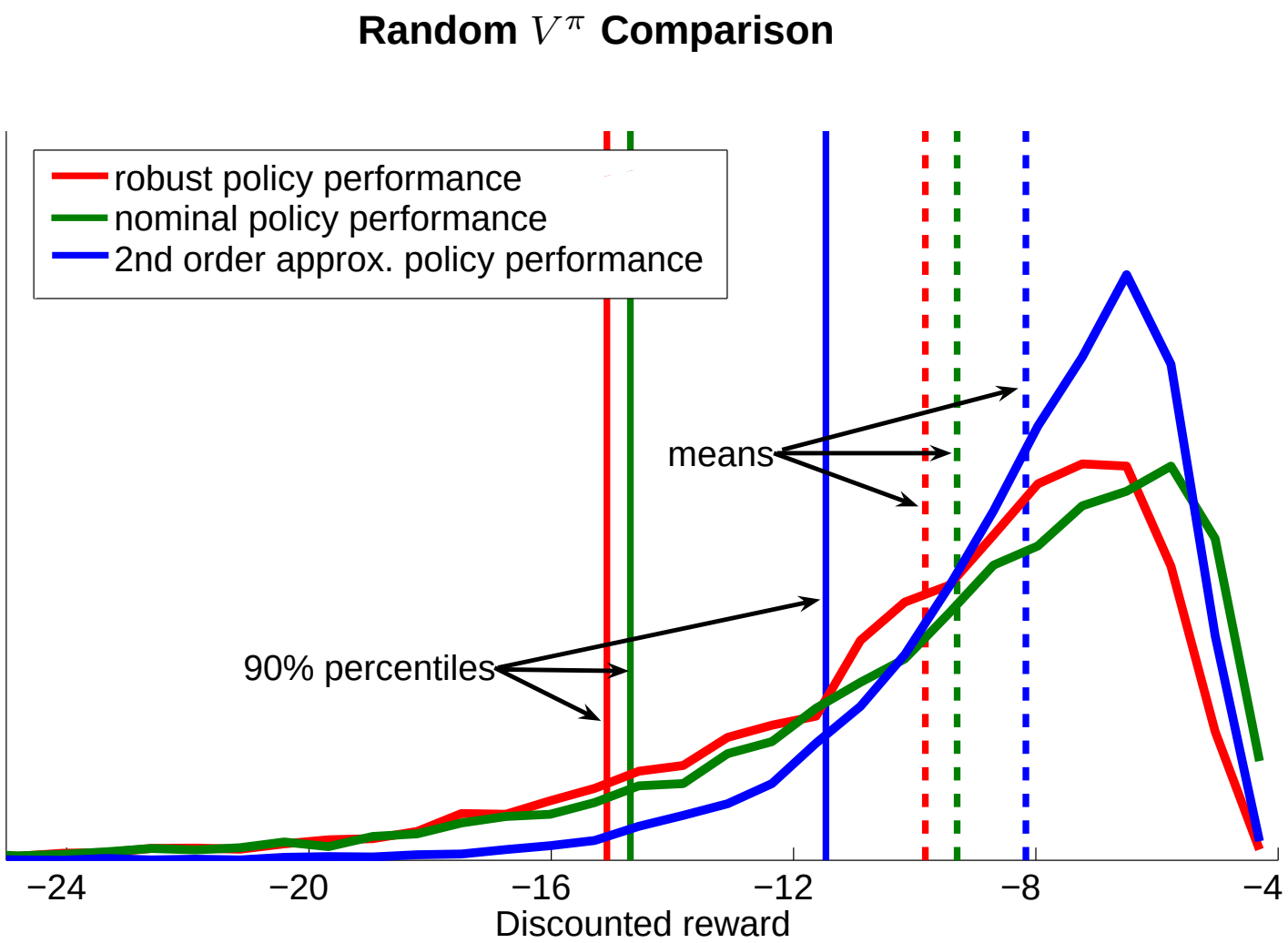
Let $\mathbb{F}(\pi)$ be the second order approximation

$$\mathbb{F}(\pi) = q^{\mathsf{T}} X^{\pi} R + \alpha^2 q^{\mathsf{T}} X^{\pi} \Pi Q^{\pi} X^{\pi} R$$

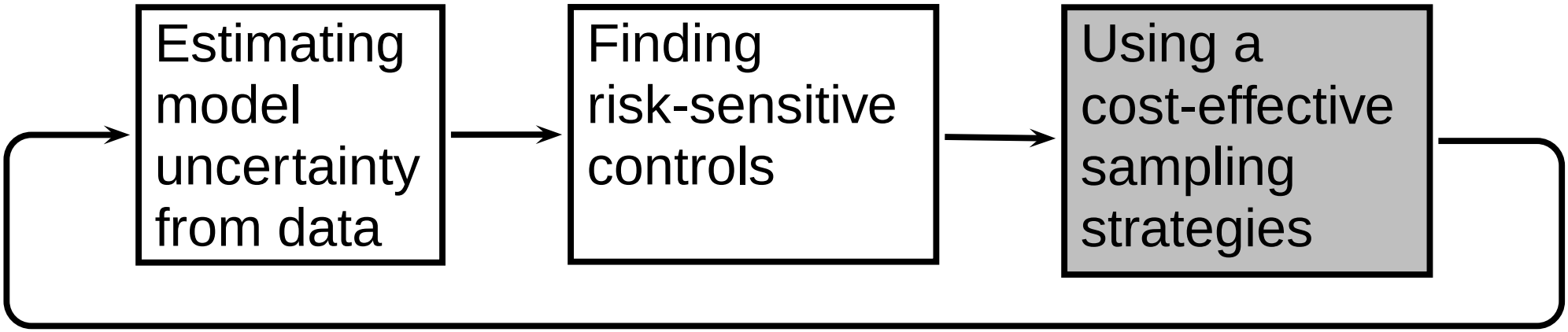
- $\mathbb{F}(\pi)$ only depends on first and second moments of P
- Optimizing $\mathbb{F}(\pi)$ is tractable for problem $\approx$ 1000 states

- Given more than M observed transitions from any state-action pair, policy $\hat{\pi} = \arg\max_{\pi} \mathbb{F}(\pi)$ is $o(1/\sqrt{(1-\eta)M})$ optimal according to the percentile problem

Experiment on MDP with Dirichlet Prior



Data-Driven MDP Approach



Cost-Effective Parameter Exploration

Sampling can be expensive in time or dollars.
Is there value in reducing my uncertainty in some parameters before exploiting the system?

Cost-Effective Parameter Exploration

Sampling can be expensive in time or dollars.
Is there value in reducing my uncertainty in some parameters before exploiting the system?

- Model based interval estimation (Strehl & Littman, 2005)
  Optimal for large horizon problems
- Budgeted learning problem (Guha & Munagala, 2007)
  Disregards the exploration-exploitation dilemma
- Value of Information (Howard, 1966)

Expected Value of Information

Given a prior on reward and transitions f_R and f_P and a risk-sensitive measure of return $\mathcal{G}(\pi, f_R, f_P)$, the value of sampling at (i, a) is

$$\mathcal{V}(i,a) = \mathbb{E}_{f'_R, f'_P}(\max_{\pi'} \mathcal{G}(\pi', f_{R'}, f_{P'})) - \max_{\pi} \mathcal{G}(\pi, f_R, f_P)$$

Expected Value of Information

Given a prior on reward and transitions f_R and f_P and a risk-sensitive measure of return $\mathcal{G}(\pi, f_R, f_P)$, the value of sampling at (i, a) is

$$\mathcal{V}(i,a) = \mathbb{E}_{f'_R, f'_P}(\max_{\pi'} \mathcal{G}(\pi', f_{R'}, f_{P'})) - \max_{\pi} \mathcal{G}(\pi, f_R, f_P)$$

Strategy:

if $\max_{i,a} \mathcal{V}(i,a) \geq$ sampling cost,

then sample at $(i, a) = \arg\max \mathcal{V}(i,a)$,

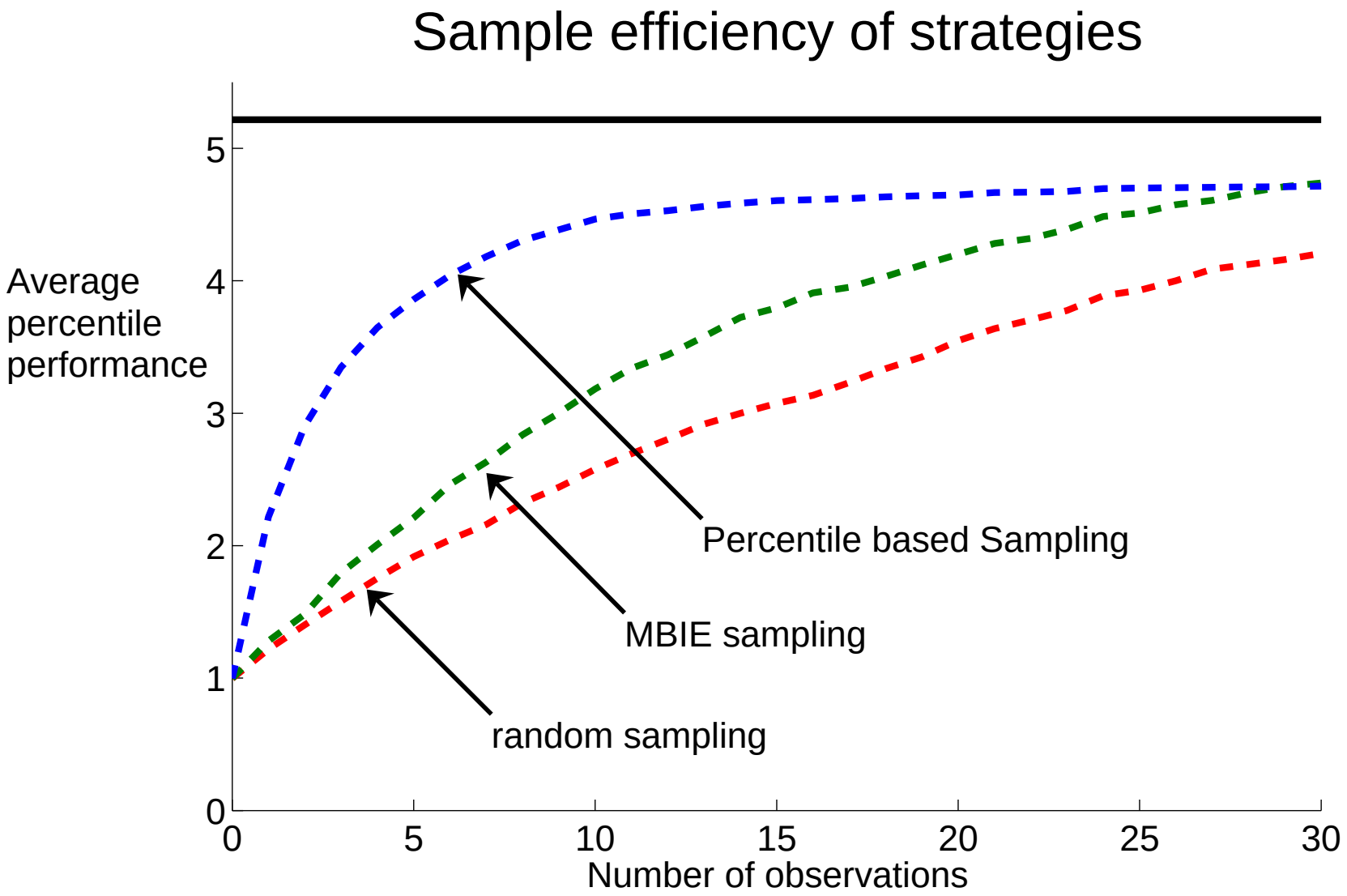
else exploit with $\pi = \arg\max_\pi \mathcal{G}(\pi, f_R, f_P)$.

Experiments on Random MDPs (I)

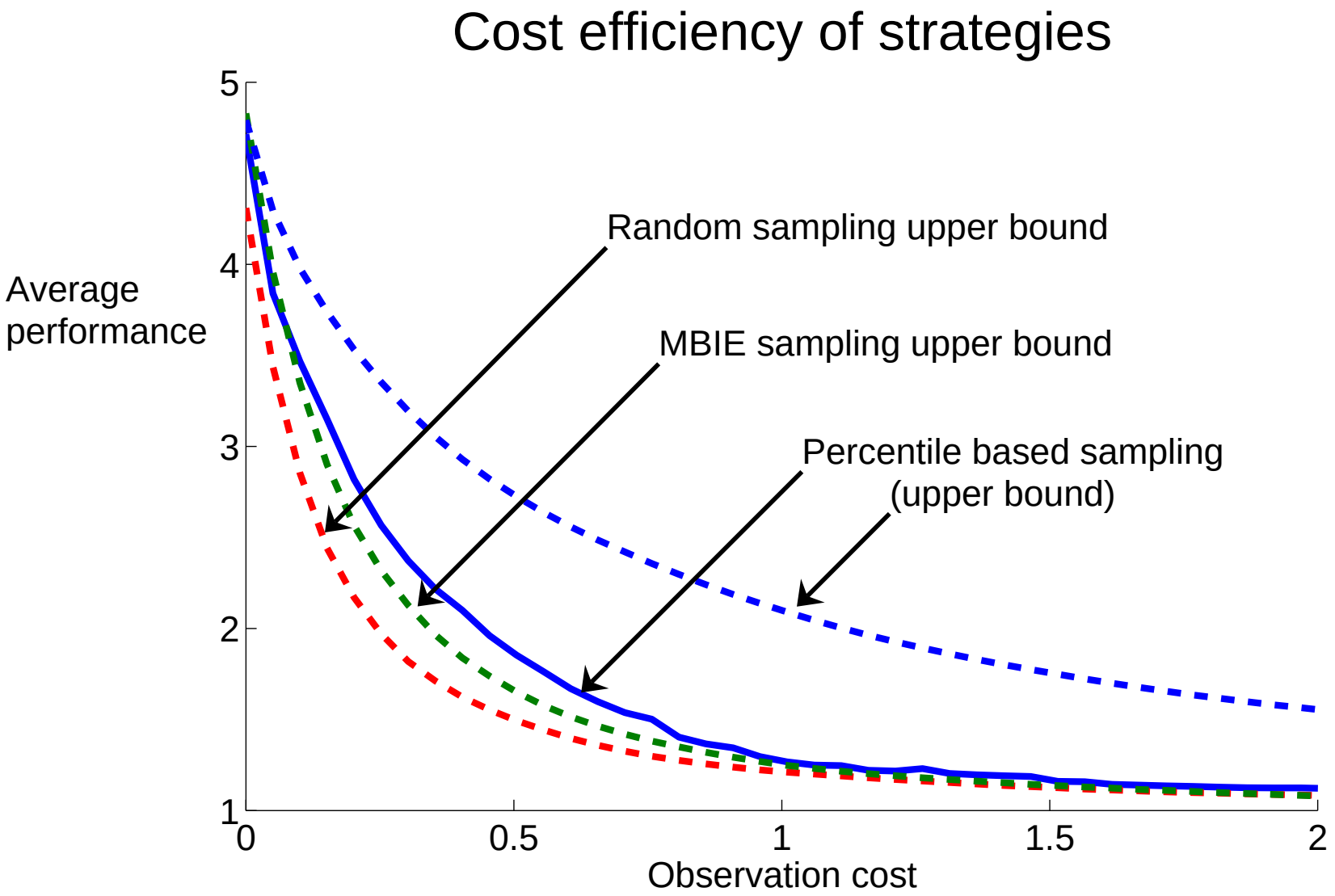
- 1000 random MDPs with reward uncertainty
- True reward drawn from $\mu_R \propto \mathcal{N}(0, 1)$
- For each MDP, prior is $R(i, a) \propto \mathcal{N}(\mu_R(i, a), 1)$
- Agent is allowed to take reward measurements at cost c before committing to optimal 90% percentile strategy

Given initial belief about the MDP, what is a good exploration strategy?

Experiments on Random MDPs (II)



Experiments on Random MDPs (II)



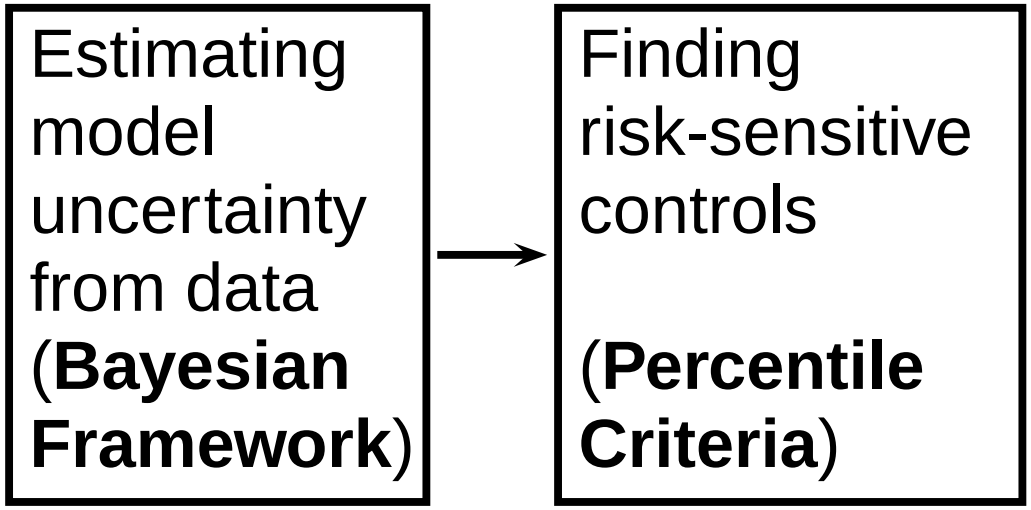
Closing Discussion

We proposed a complete risk-sensitive approach for data-driven MDP optimization:

Estimating model uncertainty from data (**Bayesian Framework**)

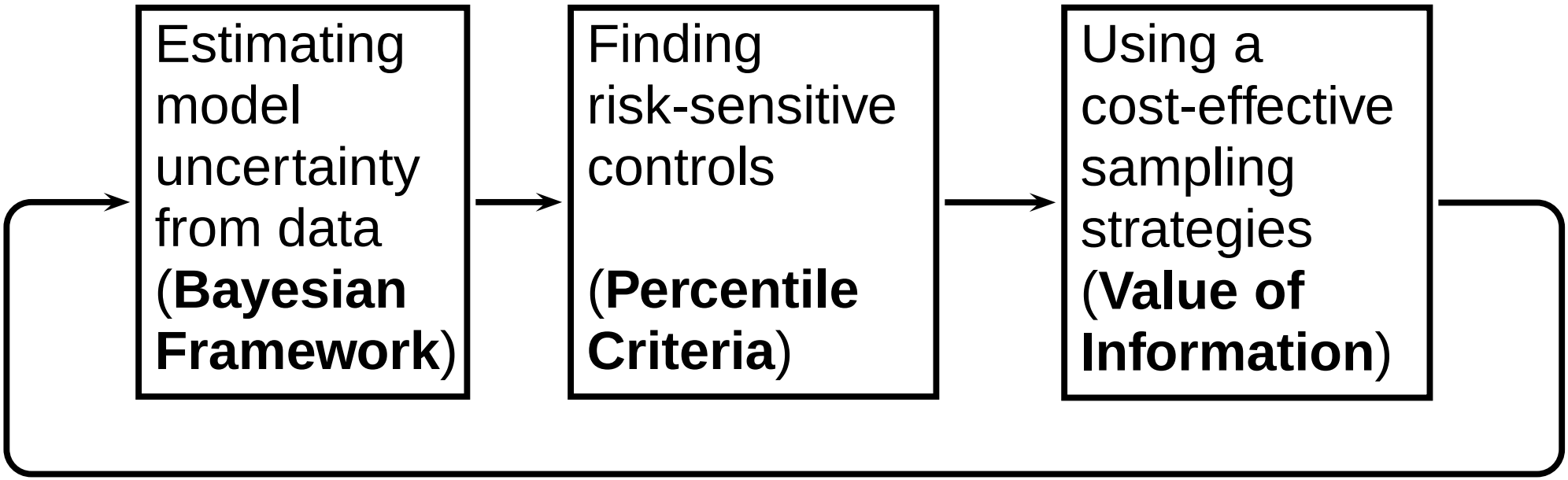
Closing Discussion

We proposed a complete risk-sensitive approach for data-driven MDP optimization:



Closing Discussion

We proposed a complete risk-sensitive approach for data-driven MDP optimization:



Closing Discussion

We proposed a complete risk-sensitive approach for data-driven MDP optimization.

Future Work:

- Revisit standard data-driven MDPs with this percentile based method for exploration and exploitation
- Apply this method in the context of online learning

Thank You !