Percentile Optimization for Markov Decision Processes with Parameter Uncertainty

Erick Delage, Stanford University, edelage@stanford.edu

Shie Mannor, McGill University, shie.mannor@mcgill.ca

For more information, visit www.stanford.edu/∼edelage

(Research supported by the Fonds Québécois de la recherche sur la nature et les technologies.)

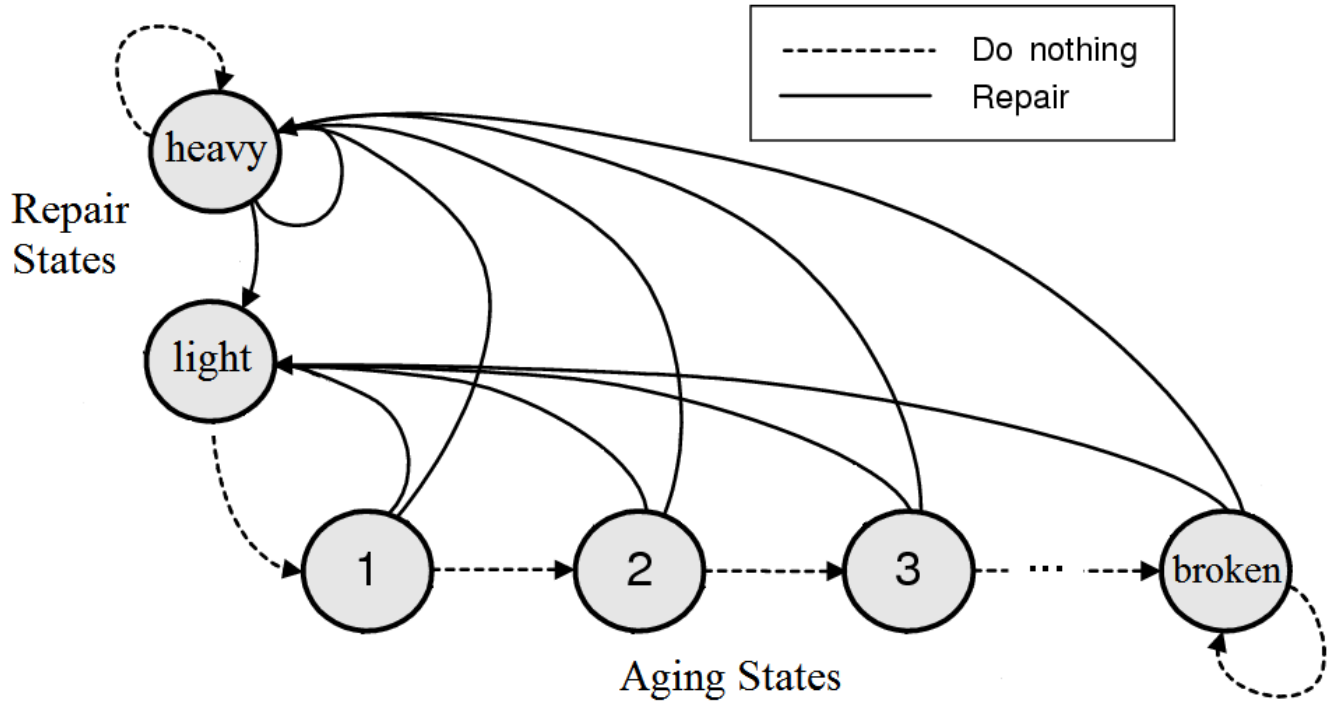
Vanilla Data-Driven Decision Making

- Gather data from your system
- Choose a model for the system
- Point estimate the parameters of your model
- Find a policy that maximizes return

- Hopefully, true system $\approx$ estimated model

A Machine Replacement Problem

- Machine ages and can require “light” or “heavy” repairs
- Two different repair services are available
- Choose repair policy that minimizes long term costs



Example of Model Uncertainty

Given that for the “heavy” repair state historical data says:

- Repair option 1 was successful 90% of 100 trials
- Repair option 2 was successful 100% of 5 trials

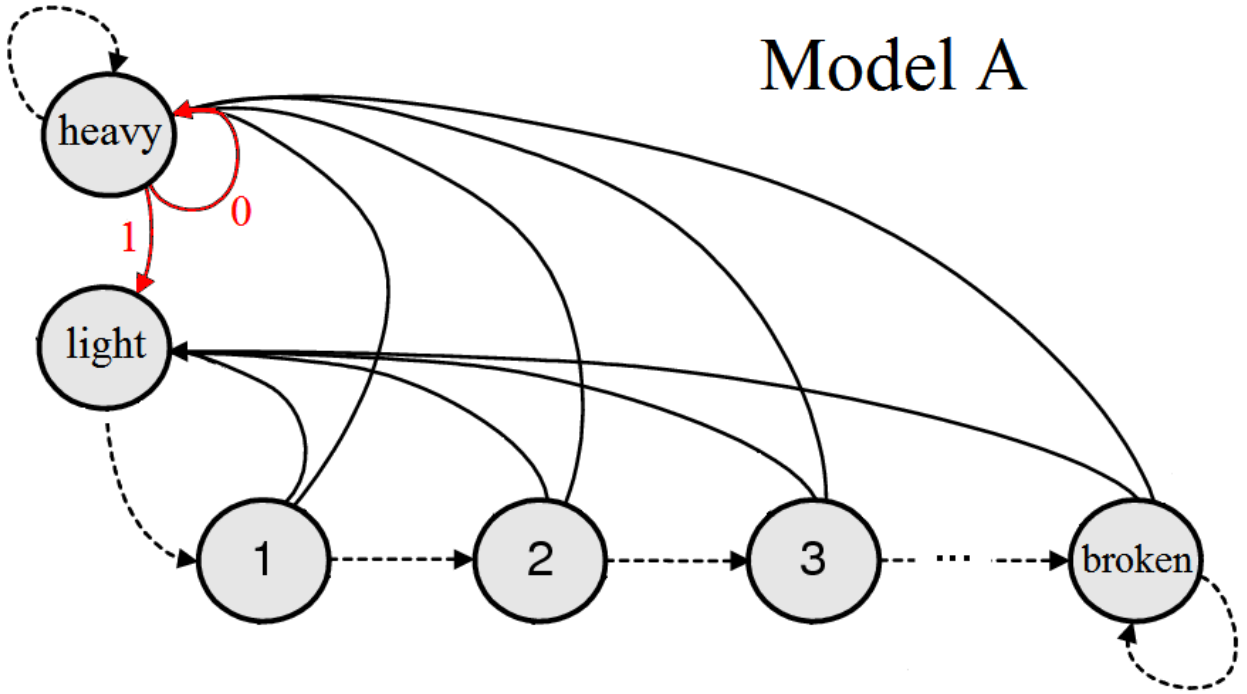
What is the true model for repair option #2?

Example of Model Uncertainty

Given that for the “heavy” repair state historical data says:

- Repair option 1 was successful 90% of 100 trials
- Repair option 2 was successful 100% of 5 trials

What is the true model for repair option #2?

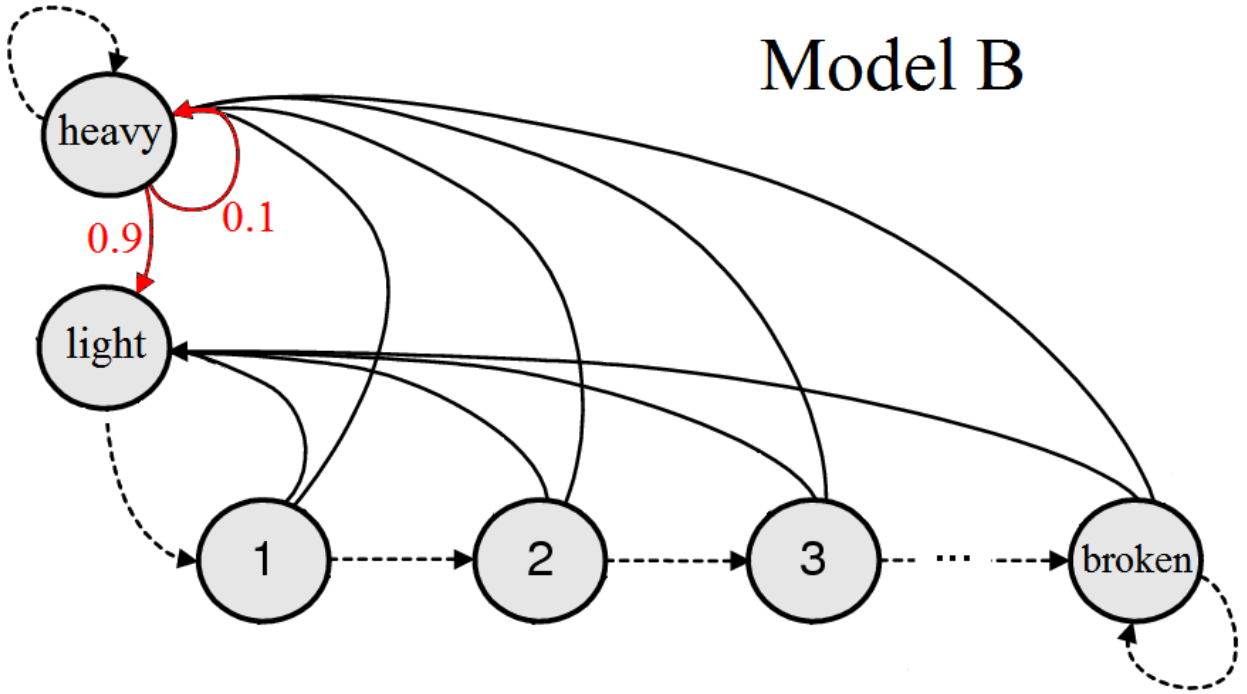


Example of Model Uncertainty

Given that for the “heavy” repair state historical data says:

- Repair option 1 was successful 90% of 100 trials
- Repair option 2 was successful 100% of 5 trials

What is the true model for repair option #2?

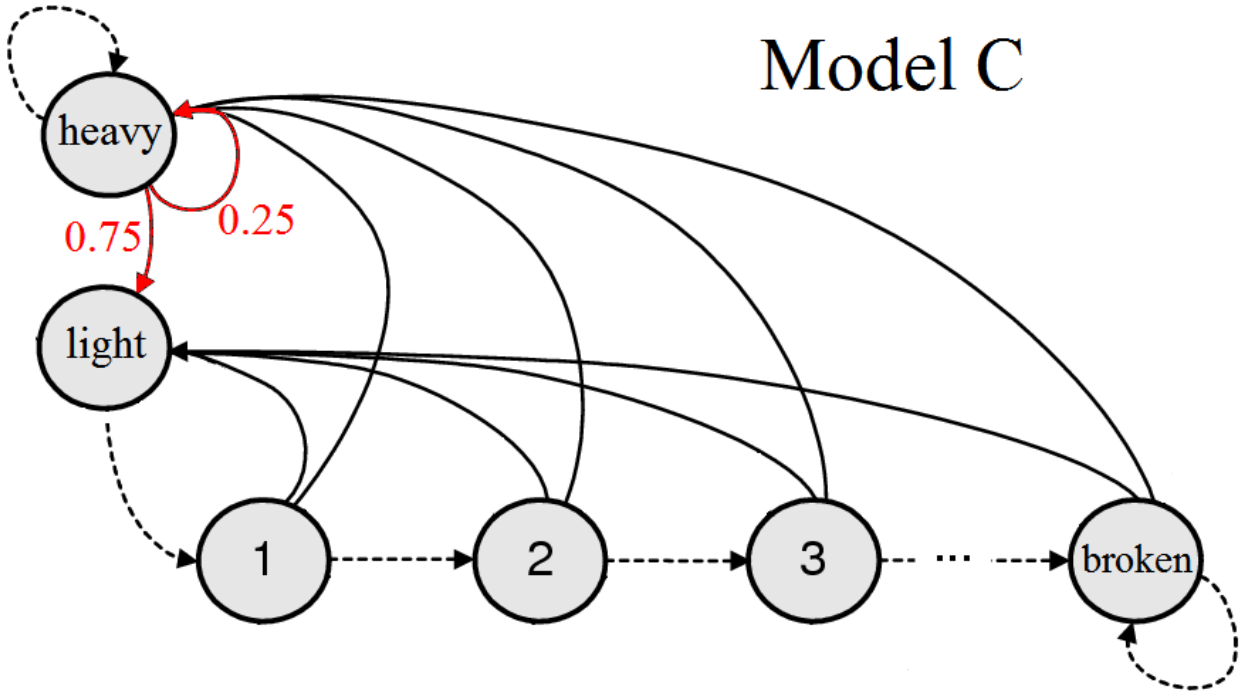


Example of Model Uncertainty

Given that for the “heavy” repair state historical data says:

- Repair option 1 was successful 90% of 100 trials
- Repair option 2 was successful 100% of 5 trials

What is the true model for repair option #2?

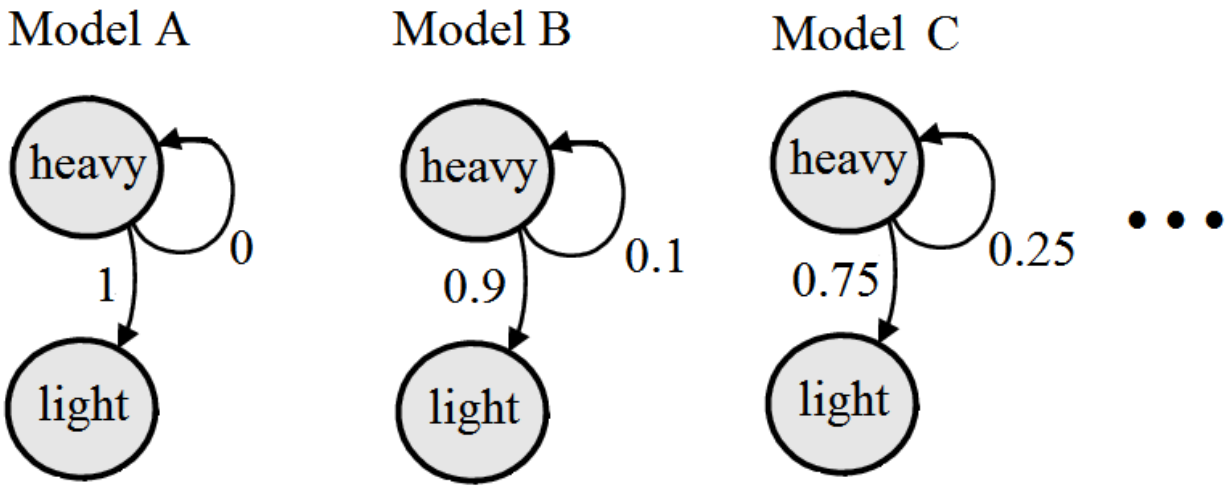


Example of Model Uncertainty

Given that for the “heavy” repair state historical data says:

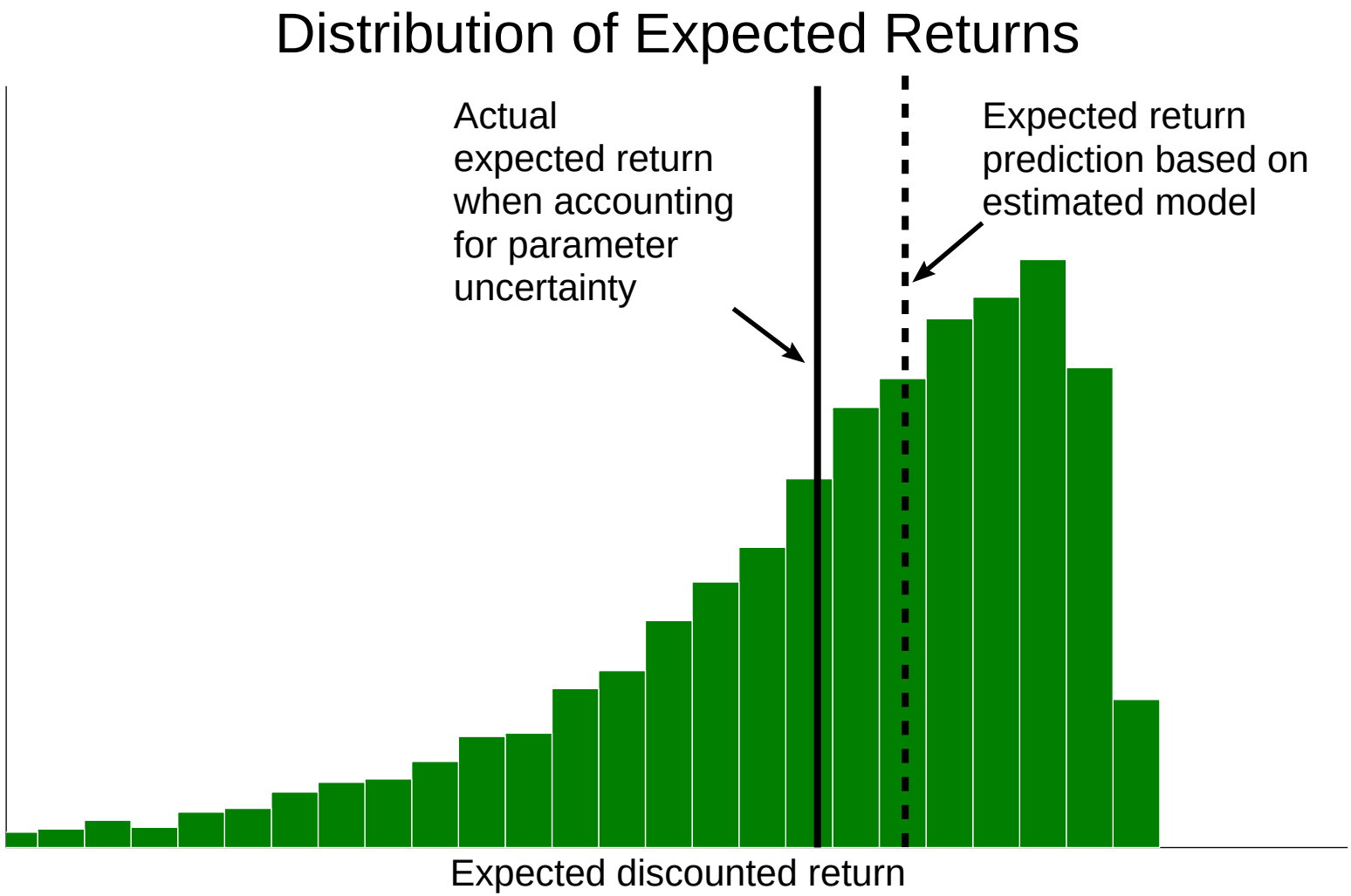
- Repair option 1 was successful 90% of 100 trials
- Repair option 2 was successful 100% of 5 trials

What is the true model for repair option #2?

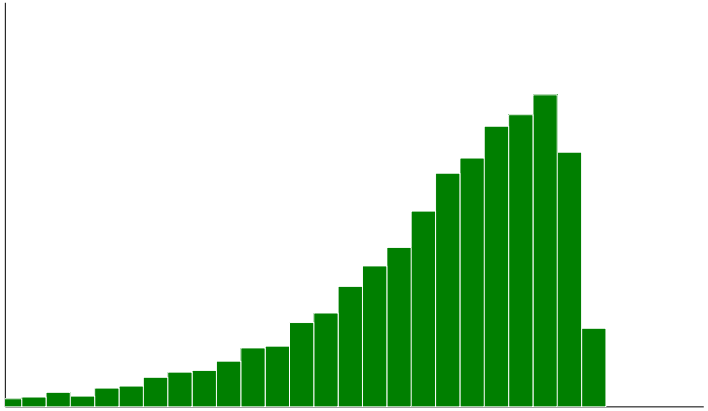


What is the TRUE long term cost of using repair option #2?

The Curse of Model Uncertainty



The Curse of Model Uncertainty (II)



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

Data-Driven MDP Approach

We address the problem of model uncertainty in the context of Markov decision processes

Assumption :

- The system behaves as an MDP with known states and actions

Data is available for :

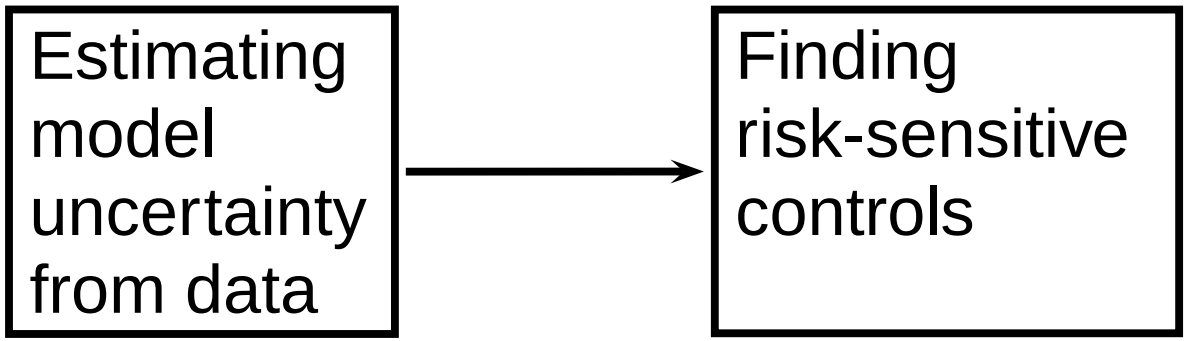
- Transition trajectories & noisy reward measures

Goal :

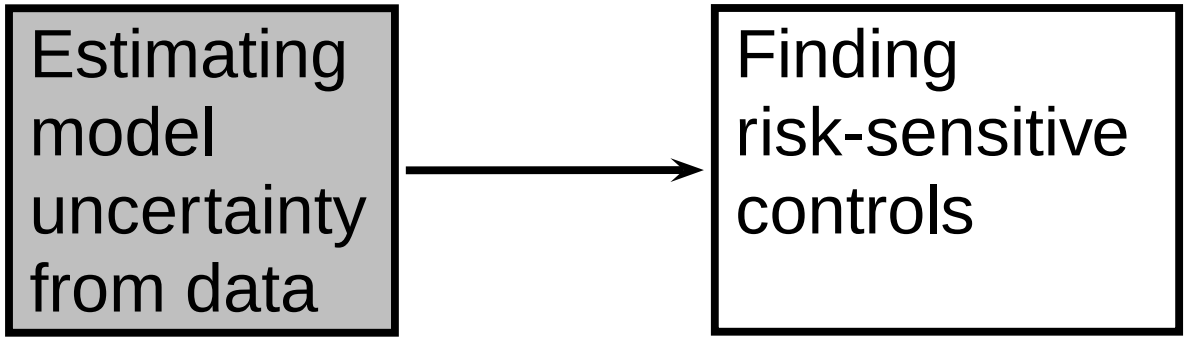
- Find a stationary policy that performs “well” on the true MDP (risk-sensitivity)

Data-Driven MDP Approach

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

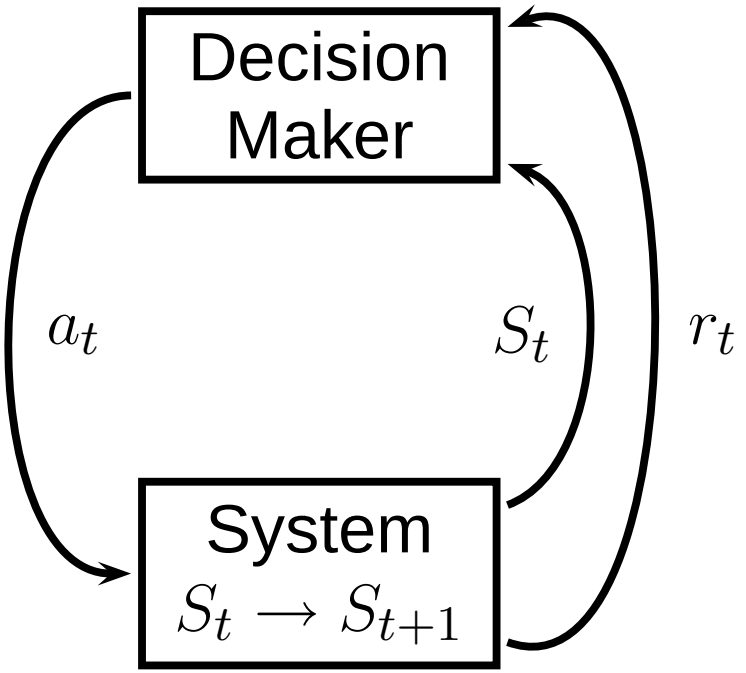


Data-Driven MDP Approach



Markov Decision Processes

A simple and popular model (MDP)



Ingredients:

1. State space $\mathcal{S}$
2. Action space $\mathcal{A}$
3. Reward $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
4. Transition probability $P(s'|s, a)$

Dynamics: $S_t \to A_t \to R_t \to S_{t+1}$

MDPs: The Objective

Objective: maximize (over all policies π)

$$\mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\alpha^{t}R_{t}\middle|S_{0}=s\right] \quad \text{with } \alpha<1\ .$$

There exists an optimal stationary and deterministic policy.

$$\pi:\mathcal{S}\rightarrow\mathcal{A}$$

Algorithmically “easy”: linear programming, policy iteration, value iteration, dynamic programming

Parameter Uncertainty in MDP

We always have uncertainty in the parameters

$$R \text{ and } P(s'|s,a)$$

- I don't have a model - estimate from data
- I know I don't know (part of the model)
- MDP is a model reduction

Bayesian Parameter Uncertainty

We always have uncertainty in the parameters

$$R \text{ and } P(s'|s,a)$$

Bayesian Approach:

- Start with a prior : $\mathbb{P}(R,P)$
- Gather model observations $\mathcal{O}$: $\mathcal{O} \sim \mathbb{P}(\mathcal{O}|R,P)$
- Compute posterior distribution over the model : $\mathbb{P}(R,P|\mathcal{O}) \propto \mathbb{P}(\mathcal{O}|R,P)\mathbb{P}(R,P)$
- Posterior can be evaluated using Gibbs sampling

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:

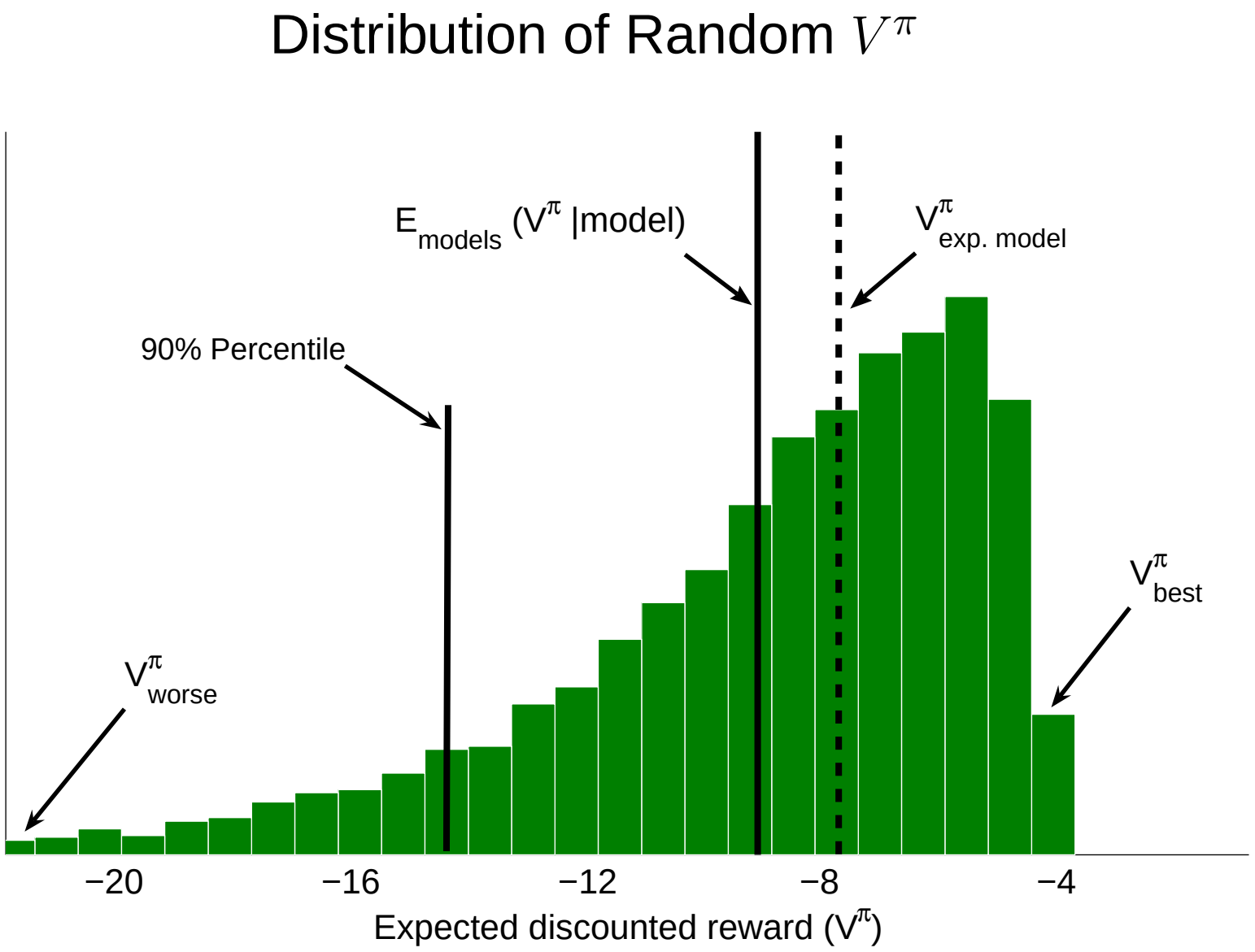
Consider $V^\pi = \mathbb{E}_x^\pi[\sum_{t=0}^{\infty} \alpha^t R(x_t)]$ as a random variable.

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

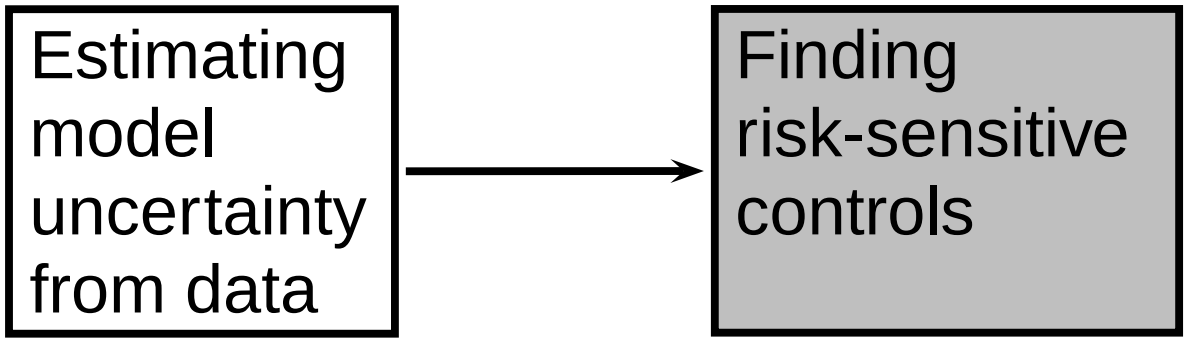
$$\mathbb{E}_{\text{models}}\left[V^\pi\right] = \mathbb{E}_{\text{models}}\left[\mathbb{E}_x^\pi\left[\sum_{t=0}^{\infty} \alpha^t R(x_t)\right]\right]$$

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

The Curse of Parameter Uncertainty



Data-Driven MDP Approach



The Robust Approach

$$\Delta(\mathcal{R}) = \{\text{set of all possible rewards}\}$$
$$\Delta(\mathcal{P}) = \{\text{set of all possible transition probabilities}\}$$

Objective:

$$\max_{\pi} \min_{R \in \Delta(\mathcal{R}), P \in \Delta(P)} \mathbb{E}^{\pi} \left[\sum_{t=0}^{\infty} \alpha^t R(x_t) \right]$$

- Tractable solution via dynamic programming
- Non-probabilistic uncertainty
- Uncertainty may be difficult to calibrate
- Leads to conservative policies

(Iyengar, 2002; Nilim and El-Ghaoui , 2005)

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**
- **Useful approximation for Dirichlet transitions**

Percentile Optimization : Rewards

Suppose $R \approx \mathcal{N}(\mu_R, \Theta_R)$ and q is initial distribution on states, an η-percentile optimal policy can be found using

$$\begin{aligned} \max_{x \in \mathbb{R}^{|S|\times|A|}} \quad & \textstyle\sum_a x_a^\top \mu_R - \Phi^{-1}(\eta) \| \sum_a x_a^\top \Theta_R^{\frac{1}{2}} \|_2 \\ \text{subject to} \quad & \textstyle\sum_a x_a^\top = q^\top + \sum_a \alpha x_a^\top P_a \\ & x_a^\top \geq 0 \,, \quad \forall \;\; a \in A. \end{aligned}$$

- Not much harder than the original problem

Percentile Optimization : Transitions (I)

In the case that rewards are known, but transitions are uncertain, 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

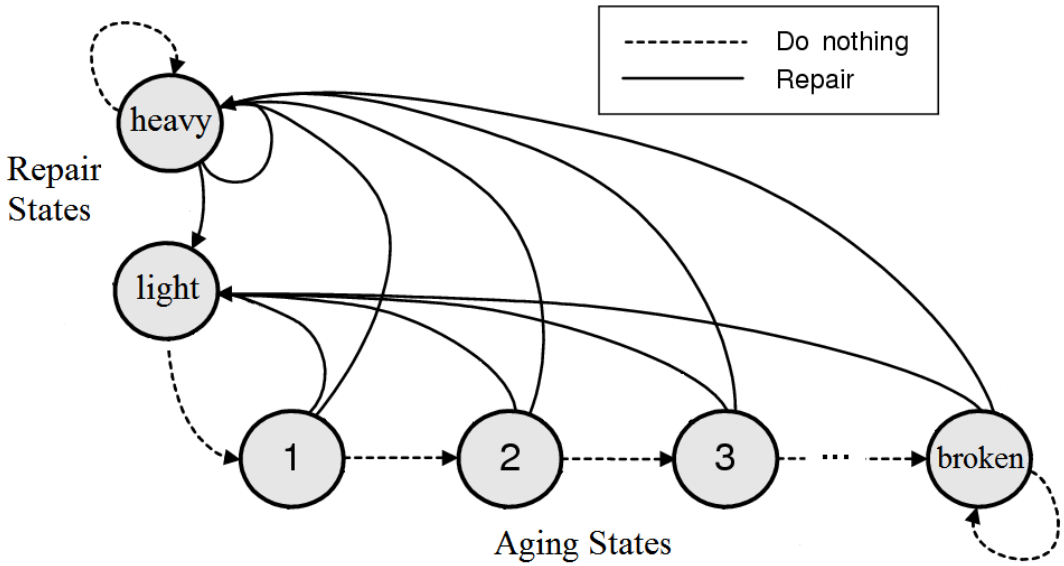
$$\mathbb{F}(\pi) = q^{\mathsf{T}} X^{\pi} R + \alpha^2 q^{\mathsf{T}} X^{\pi} \Pi Q^{\pi} X^{\pi} R$$

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

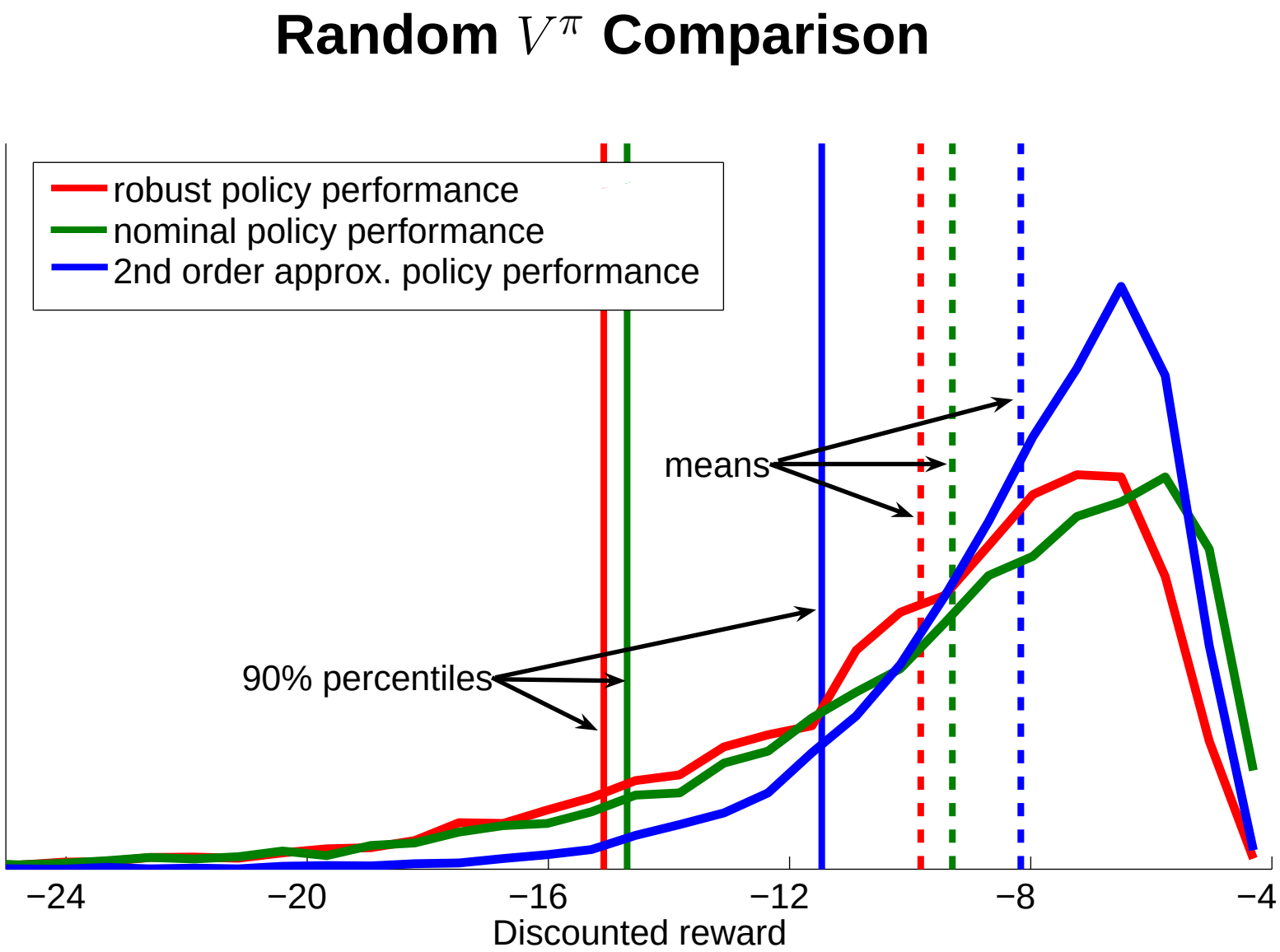
- Given more than M observed transitions from any state-action pair, policy $\hat{\pi} = \arg\max_{\pi} \mathbb{F}(\pi)$ is $o(1/\sqrt{(1-\eta)M})$ optimal according to the percentile problem

Experiment on MDP with Dirichlet Prior

- State related returns R are fully known
- Dirichlet prior for transitions $P(s'|s,a)$
- Observed 5 transitions for each state-action pair
- Choose repair policy that maximizes 90% percentile bound on returns



Experiment on MDP with Dirichlet Prior



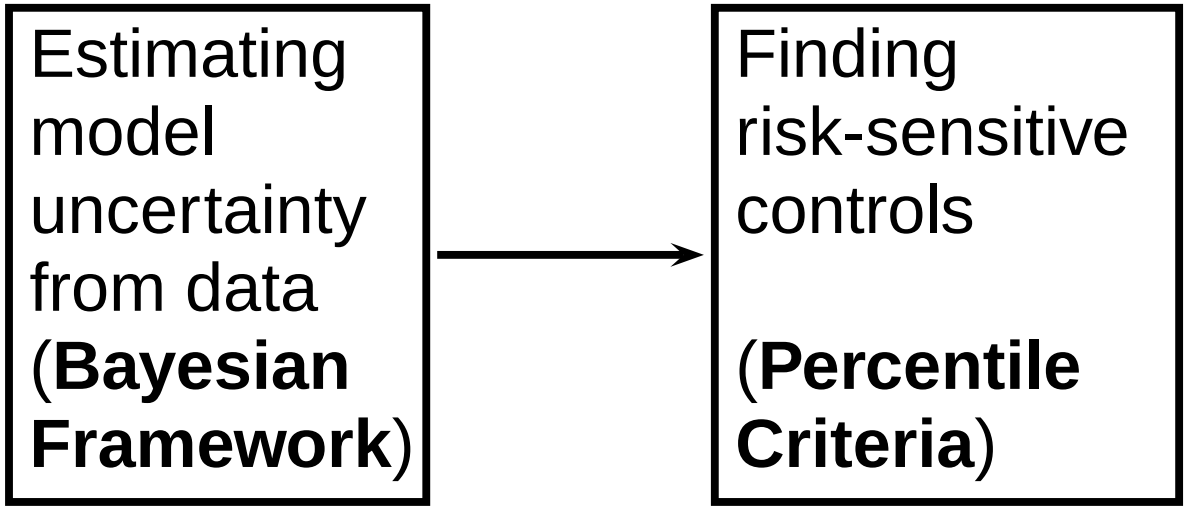
Closing Discussion

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

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

Closing Discussion

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



Closing Discussion

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

Future Work:

- Revisit standard data-driven MDPs with this percentile based method
- Use this framework to provide strategies for parameter exploration
- Address model uncertainty in other forms of decision problems

Thank You !

For more information, visit www.stanford.edu/∼edelage