

# Decision Making under Parameter Uncertainty

Shie Mannor  
McGill University



McGill

Joint work with: Erick Delage (Stanford), Xu Huan (McGill),  
Duncan Simester (MIT), Peng Sun (Duke), John Tsitsiklis (MIT)

# DECISION MAKING

Classical decision making:

I know where I am

I know what I can do

I know what will happen (or at least the distribution of future events)

## DECISION MAKING

Parameter Uncertainty:

I know where I am

I know what I can do

I am not sure what what is the distribution of future events

## WHY HAVE UNCERTAINTY?

- I don't have a model - sample from data
- I know I don't know
- Model includes uncertainty
- Things change with time
- I would like to have a full posterior, not just point estimates

# BEING BAYESIAN



- $Z$  – parameter generating **hidden process**,  $Y$  – observable
- We want to infer  $Z$  from measurements of  $Y$
- Statistical dependence between  $Z$  and  $Y$  known:  $P(Y|Z)$
- We have a prior over  $Z$ , reflecting our uncertainty:  $P(Z)$
- Observe  $Y = y$
- Compute posterior:  $P(Z|Y = y) = \frac{P(y|Z)P(Z)}{\int dZ' P(y|Z')P(Z')}$

# BEING FREQUENTIST



- $Z$  – unknown **parameter**,  $Y$  – observable
- Distribution of  $Y$  is parametric
- Statistical dependence between  $Z$  and  $Y$  known:  $P_Z(Y)$
- Observe  $Y = y$
- Compute maximum likelihood parameter:  
 $Z_{ML} = \arg \max_Z P(Y|Z)$
- Other parameters  $Z_{\text{other}}$  may true

# MARKOV DECISION PROCESSES

A simple and popular model (MDP)

Ingredients:

1. State space  $S$
2. Action space  $A$
3. Reward  $R$  (a random variable)
4. Transition probability  $P(s'|s, a)$ .

Dynamics:

$$s_t \rightarrow a_t \rightarrow r_t \rightarrow s_{t+1}$$

## MDP: OBJECTIVE

Objective: maximize (over all policies)

$$\mathbb{E}_{\pi} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \right]$$

where  $\gamma < 1$

Reminder: There exists an **optimal stationary** and **deterministic** policy.

Other objectives are possible - average cost, finite horizon.

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

## SOME APPLICATIONS OF MDPs

OR: Supply chain, machine replacement, scheduling, inventory control

Robotics: Inverted pendulum, path finding, etc.

Marketing: Mail-order catalogs

IT: Power management

Health-care: Medical treatment

Problems of dynamic nature with both **inherent** uncertainty and **parametric** uncertainty.

We focus on **parametric** uncertainty

## ANOTHER SOURCE OF UNCERTAINTY

Very high dimensional observation spaces.

Examples:

- Power management problem
- Mail-order catalog problem

Manageable MDPs are small ( $\approx 1000$  states)

Actual MDP represents a simplification

## MDPs WITH PARAMETRIC UNCERTAINTY

We assume we know  $S$  and  $A$

But  $R$  and  $P$  are not known (exactly)

If  $S$  is not known  $\Rightarrow$  a different talk

What are we going to do?

But first - should we care?

## VARIANCE: ILLUSTRATION

Catalog Circulation Problem

Womens clothing retailer

1.7 million customers

Entire purchase history

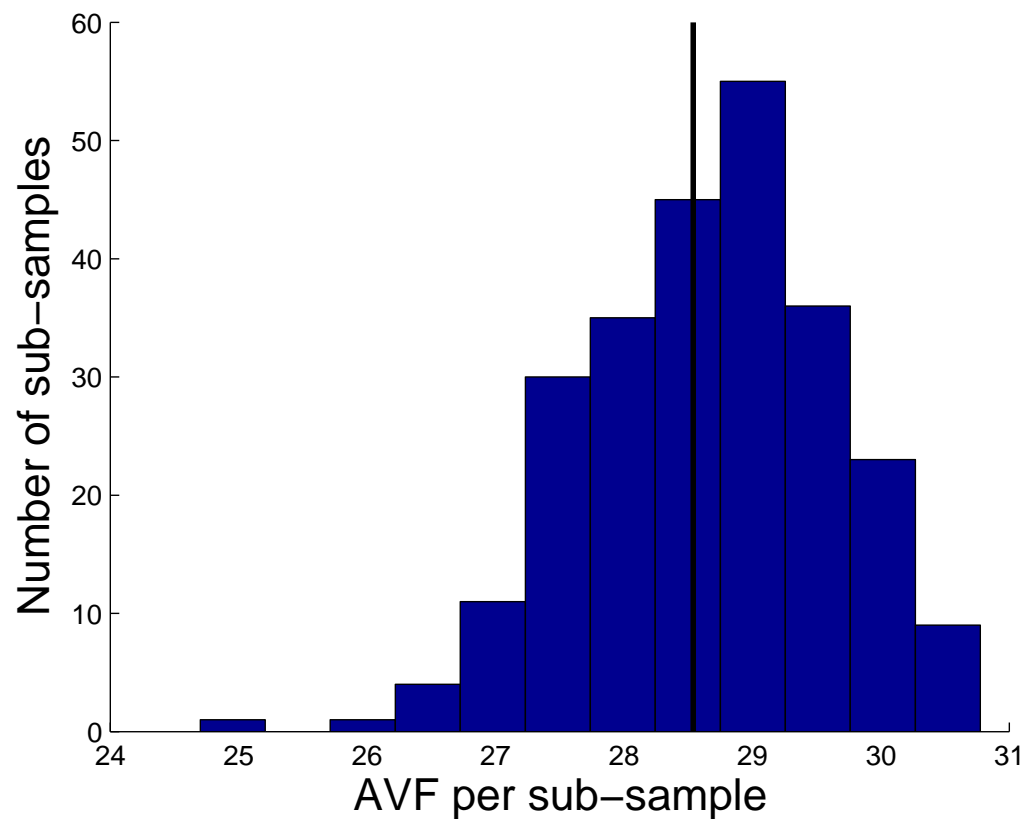
6 years of mailing history

64 states: Recency, Frequency and Monetary Value Quartile ( $4^3$ )

250 Sub-samples: 657,000 observations in each

”True” model: All 1.7 million customers

# VALUE FUNCTION: TRUE VS. ESTIMATED



STD = \$2

## THE CONTROL PROBLEM

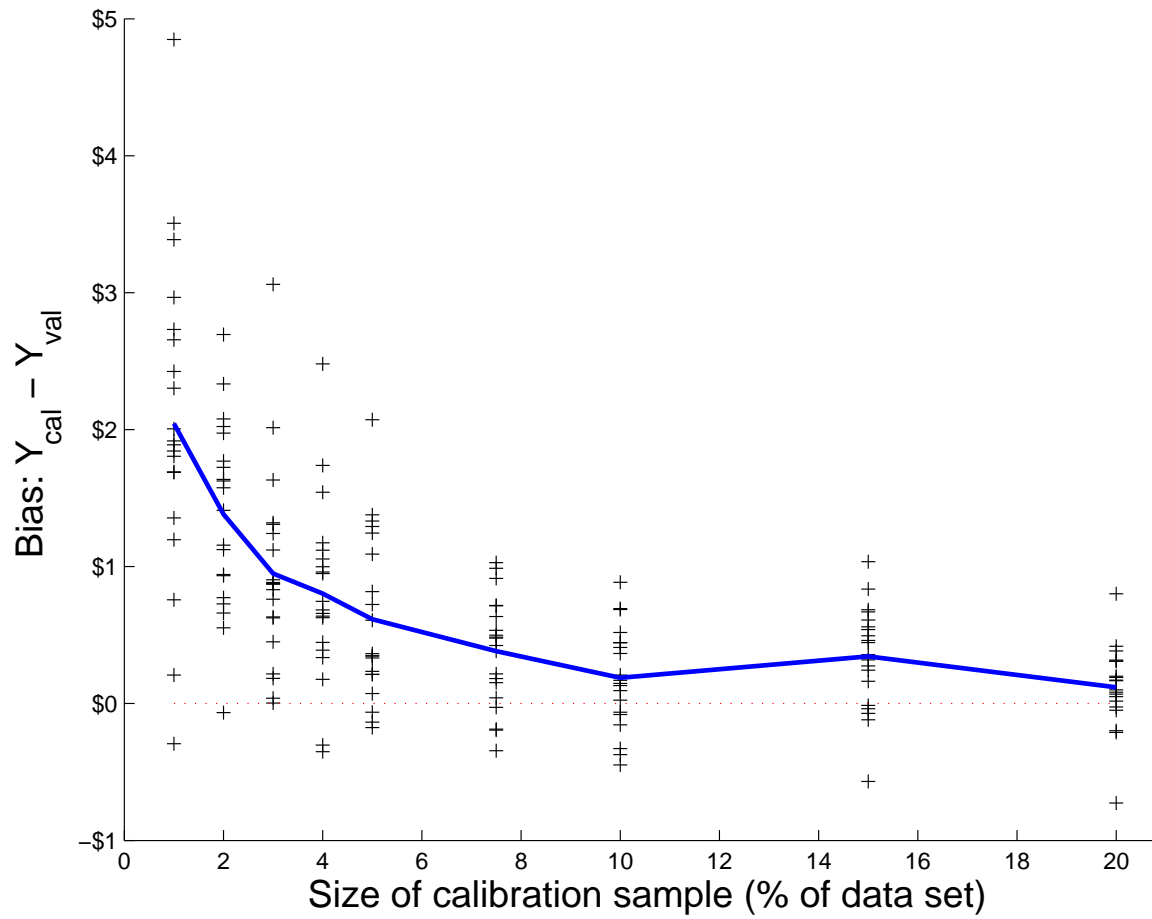
Optimization may induce additional bias (!)

How big is this bias?

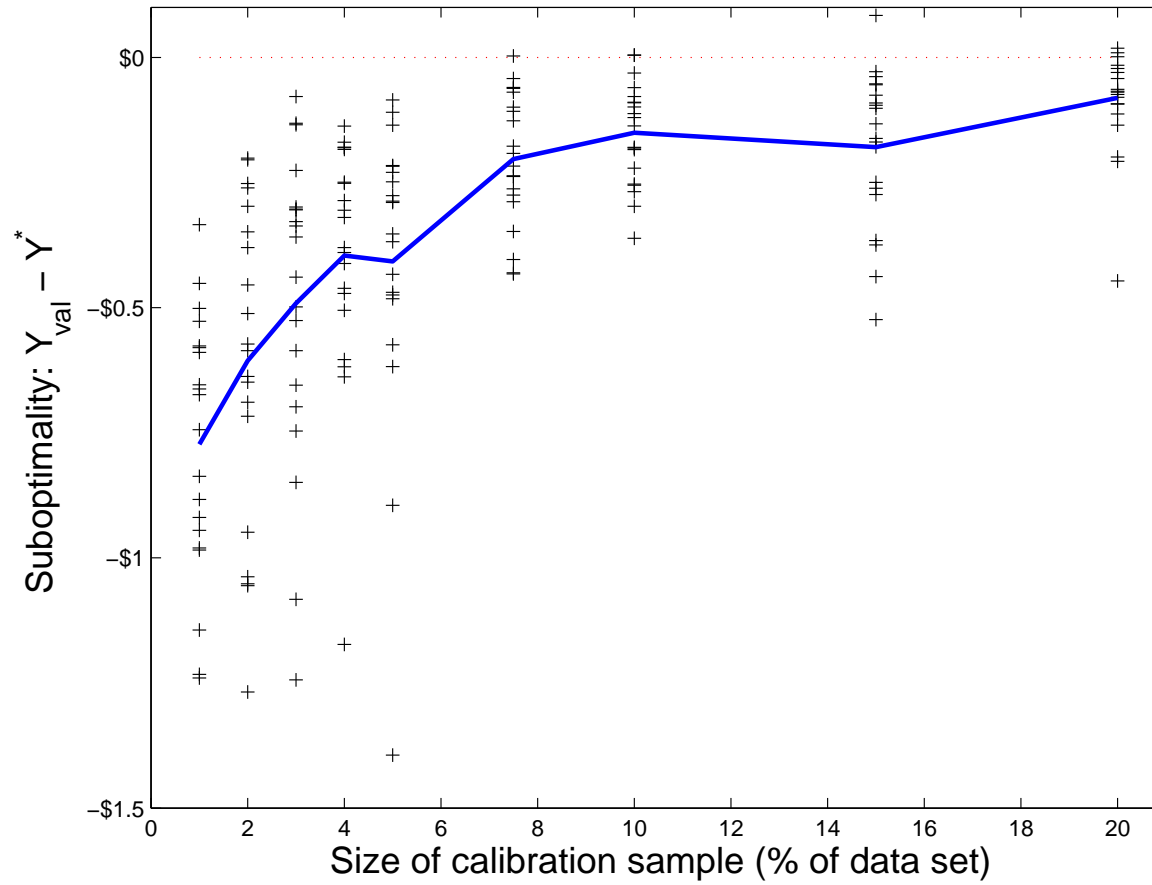
Recipe:

1. Divide data to calibration and validation set
2. Solve on calibration
3. Evaluate on validation
4. Estimate the magnitude of bias

# THE CONTROL PROBLEM: BIAS



# THE CONTROL PROBLEM: SUB-OPTIMALITY



## SOLUTIONS NEEDED

A partial list:

- Ignore uncertainty: hope for the best.
- Robustify: expect the worst.
- Risk aware approach: Be Bayesian and consider a small probability of “disaster”.
- Tradeoff robustness with ignorance: balance pessimism and optimism

## IGNORE UNCERTAINTY

Assume all parameters take nominal value:

If  $R(s) \in [0, 1]$ : take 0.5

If sampled 0, 1, 1: take mean =  $2/3$ .

A frequentist approach

Simple: Ignorance is bliss

But, can be disastrous if uncertainty not small.

Standard approach in ML/OR

## THE ROBUST APPROACH

Pessimism in face of uncertainty.

Let:

$\Delta(R) = \{\text{set of all possible rewards}\}.$

$\Delta(P) = \{\text{set of all possible probabilities}\}.$

Objective: maximize (over all policies)

$$\mathbb{E}_\pi \left[ \min_{R \in \Delta(R), P \in \Delta(P)} \sum_{t=0}^{\infty} \gamma^t r_t \right]$$

A max-min problem.

Game against Nature.

Non-probabilistic uncertainty.

## THE ROBUST APPROACH

Suppose:

1.  $\Delta(R)$  is a polytope:  $R(s, a) \in [\underline{r}(s, a), \bar{r}(s, a)]$
2.  $\Delta(P)$  is a polytope:  $p(\cdot|s, a) \in \text{polytope}$

Then:

1. There exists an optimal stationary policy
2. This policy can be computed via dynamic programming
3. We solve the 0-sum game against Nature

Best when uncertainty is small and mistakes are disasters (e.g., air traffic control problem)

(Nilim and El-Ghaoui:, OR, 2005)

## THE BAYESIAN APPROACH

$R(s, a) \sim \mathcal{N}$ : what if uncertainty is large?

Suppose we have a prior on  $R$  and  $P$ . That is we **believe** that

$$R(s, a) \sim \mathcal{N}: P(x; \alpha) = C(\alpha) e^{-(x - \alpha_{mean})^2 / \alpha_{var}}$$

$$P(\cdot | s, a) \sim \text{Dirichlet}: \Pr(x | \alpha) = C(\alpha) \prod_{i=1}^n x_i^{\alpha_i - 1}$$

After observing data we can update our **belief**.

Magic: If we start from  $R(s, a) \sim \mathcal{N}$  and  $P(\cdot | s, a) \sim \text{Dirichlet}$  we maintain the form after the update.

## THE BAYESIAN APPROACH II

For a given  $\pi$  we can ask what is:

$$E_{\text{models}} \left[ \mathbb{E}_{\pi} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \right] \right]$$

We can also ask (percentile optimization):

$$\begin{aligned} & \max_{\text{strategies}} && g \\ & \text{s.t.} && \Pr_{\text{models}} \left( \left[ \mathbb{E}_{\pi} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \right] \right] > g \right) \geq \rho \end{aligned}$$

$\rho$  is the **risk** parameter. Maximization is over the value gained at risk level  $\rho$ .

## THE BAYESIAN APPROACH III

It turns out that solving the percentile optimization is:

1. NP-hard in general.
2. NP-hard even if transitions are known.

But: For Gaussian reward parameters, problem is polytime.

Problem is solvable by 2nd order cone programming. (slightly more complex than LP).

## THE NON BAYESIAN APPROACH

Suppose we know the reward's mean and variance but cannot make Gaussian assumptions.

It turns out we can also solve:

$$\begin{aligned} & \max_{\text{strategies}} && g \\ & \text{s.t.} && \inf_{\text{models with given variance and mean}} \Pr_{\text{model}} \left( \left[ \mathbb{E}_{\pi} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \right] \right] > g \right) \geq \rho \end{aligned}$$

This is not Bayesian - we compute the statistics of the model and find the best solution.

## A HEURISTIC

Still, we want to have a handle on the complete problem (uncertainty in both transitions and rewards).

So we can look at the maximization problem.

$$\text{Maximize }_{\pi} E_{\text{models}} \left[ \mathbb{E}_{\pi} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \right] \right]$$

equivalent to :

$$\text{Maximize }_{\pi} E_{\text{models}} \left[ (I - \gamma P_{\pi}^{\text{model}})^{-1} R_{\pi}^{\text{model}} \right]$$

where  $P_{\pi}^{\text{model}}$  and  $R_{\pi}^{\text{model}}$  are transition probabilities and rewards when using  $\pi$  and following the model.

Non-linear expression inside the expectation  $\Rightarrow$  problem is tough.

## VARIANCE ESTIMATES

Suppose you:

1. construct model from data,
2. have a policy  $\pi$  (source - not important),
3. feel Bayesian

You estimate:  $V_{\pi}^{\text{emp model}} = (I - \gamma P_{\pi}^{\text{emp model}})^{-1} R_{\pi}^{\text{emp model}}$

Question 1: Can we compute/approximate the bias:

$$V_{\pi}^{\text{emp model}} - E_{\text{models}} \left[ (I - \gamma P_{\pi}^{\text{model}})^{-1} R_{\pi}^{\text{model}} \right]$$

Question 2: Can we compute/approximate the variance

$$E_{\text{models}} \left[ \left( V_{\pi}^{\text{emp model}} - (I - \gamma P_{\pi}^{\text{model}})^{-1} R_{\pi}^{\text{model}} \right)^2 \right]$$

## VARIANCE ESTIMATES

We can estimate both bias and variance when uncertainty is moderate.

Technology used: second order approximation of  $(I - \gamma P_{\pi}^{\text{model}})^{-1}$ , some analysis (but not too much).

Some observations:

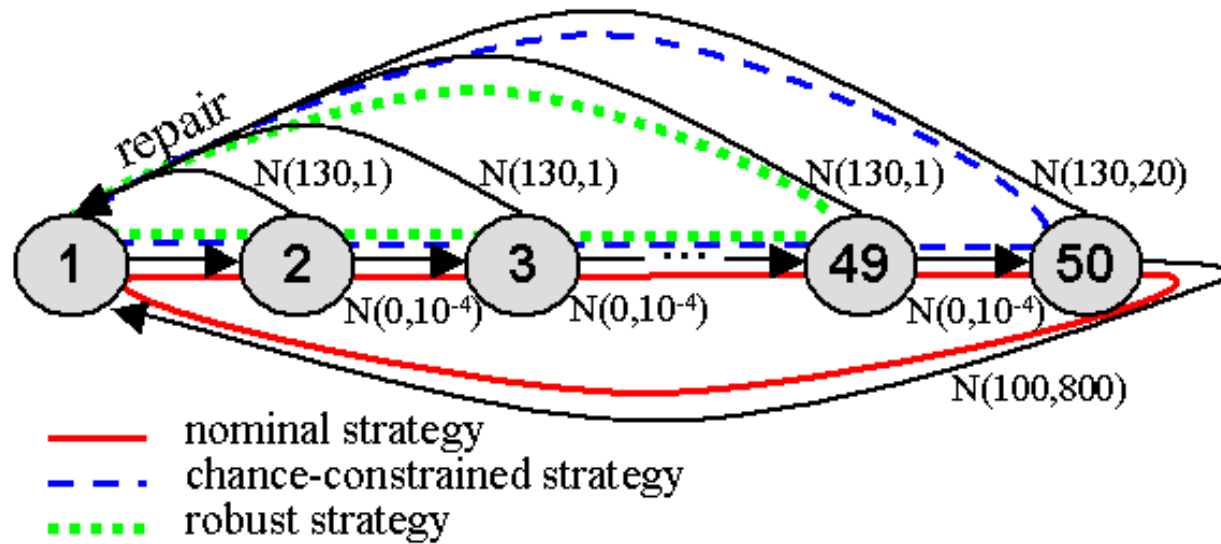
1. Bias is smaller than  $\sqrt{\text{variance}}$
2. Asymptotic normality

One can optimize the second order approximation of the reward and obtain:

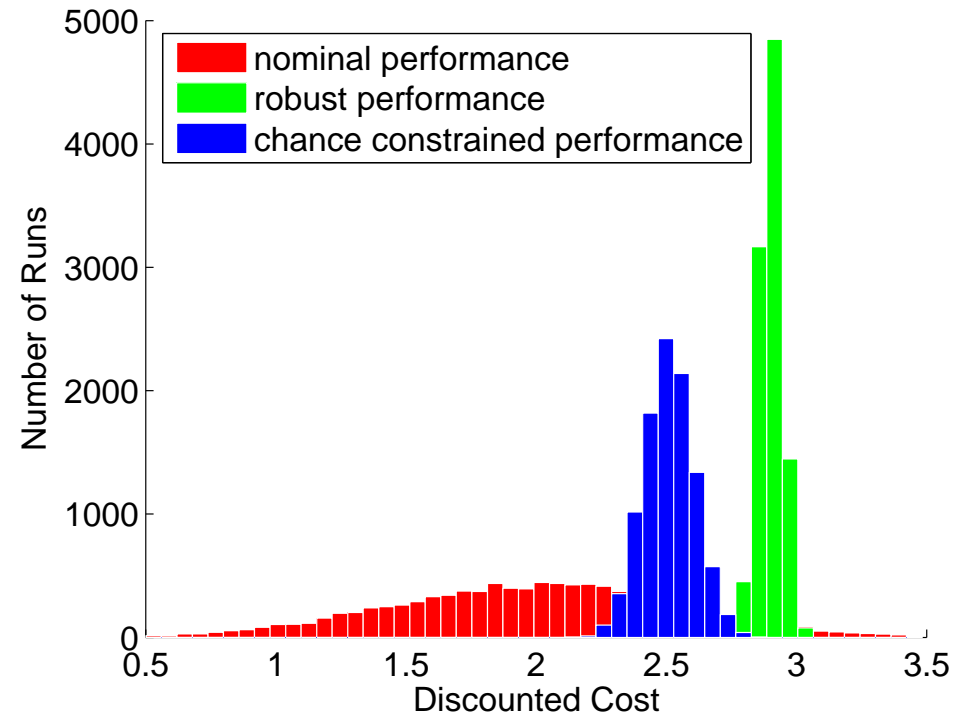
Optimizing 2nd order approximation is  $o(1/\sqrt{\rho M_{\text{minimal count}}})$  away from the chance constrained MDP with risk  $\rho$ .

# RESULTS I

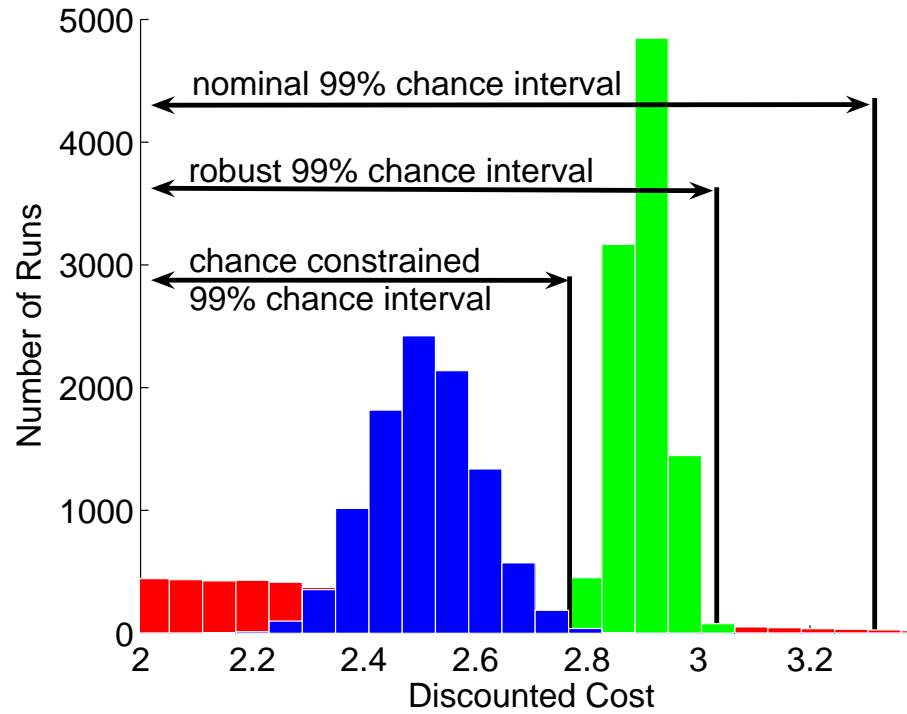
Machine replacement problem.



# RESULTS II



# RESULTS III



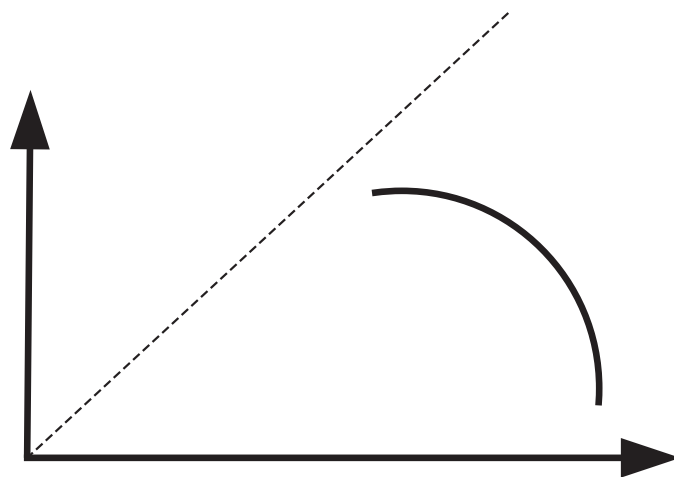
# TRADEOFF ROBUSTNESS WITH IGNORANCE

Want to balance pessimism with optimism

$$\text{Nominal value: } N(\pi) = \mathbb{E}_{\pi}^{\text{nominal model}} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \right]$$

$$\text{Worst-case value: } W(\pi) = \mathbb{E}_{\pi} \left[ \min_{R \in \Delta(R), P \in \Delta(P)} \sum_{t=0}^{\infty} \gamma^t r_t \right]$$

Performance  
under  
worst-case



Nominal performance

The curve is indeed convex.

## FINDING THE CURVE

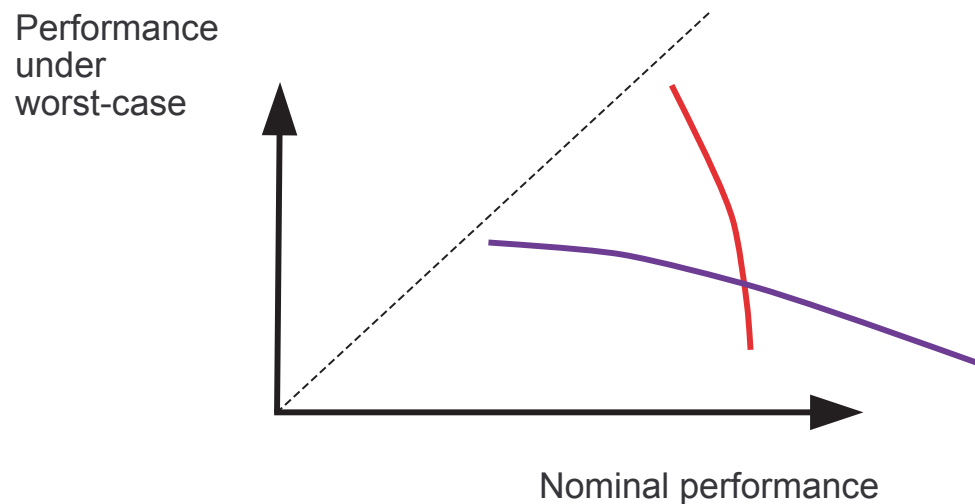
Maximize (over all policies)

$$\lambda N(\pi) + (1 - \lambda)W(\pi)$$

$\lambda = 1$ : nominal.  $\lambda = 0$ : robust.

Every value of  $\lambda$  gives a point on the Pareto front.

Why would you care?



## FINDING THE CURVE

Theorem: (finite horizon)

1. For a fixed  $\lambda$  principle of optimality works.
2. The value function is piecewise linear in  $\lambda$ .

Theorem: (infinite horizon with discounting)

Can compute (in one shot!) performance for all  $\lambda$  if there is uncertainty only in rewards.

Technique: parametric linear programming.

Theorem:

Optimal policy in general is non-Markovian.

## WRAP-UP

- Data-driven analysis
- Parameter uncertainty is significant
- Robust approach - pessimistic
- Percentile optimization
- Tradeoff robustness and nominal performance

The learning connection:

1. Can reduce uncertainty actively
2. Must take uncertainty into account when building agents