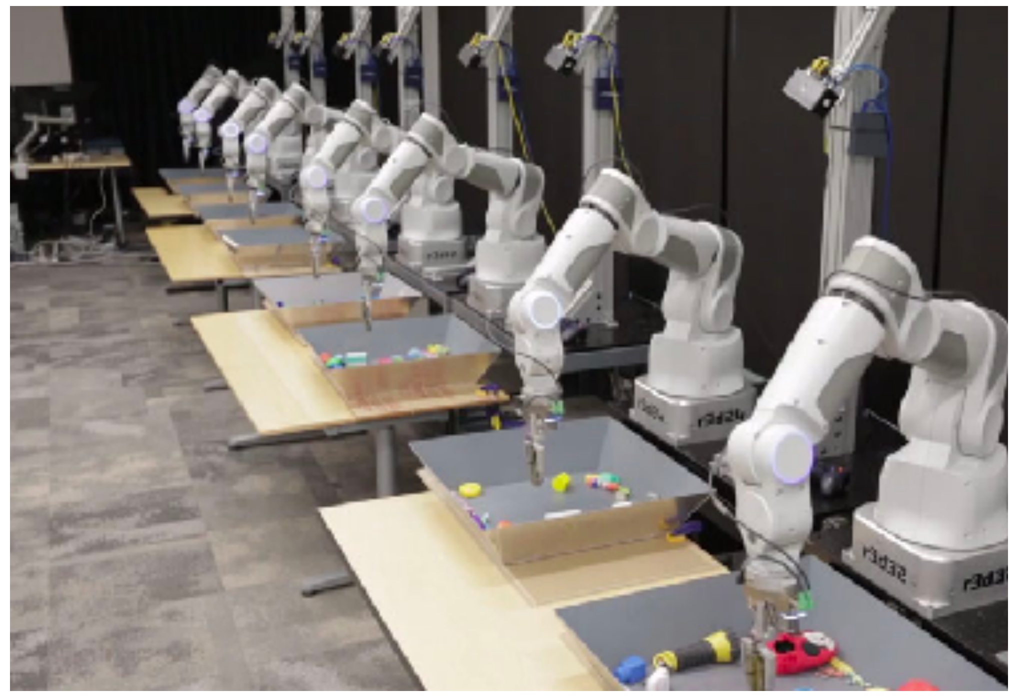
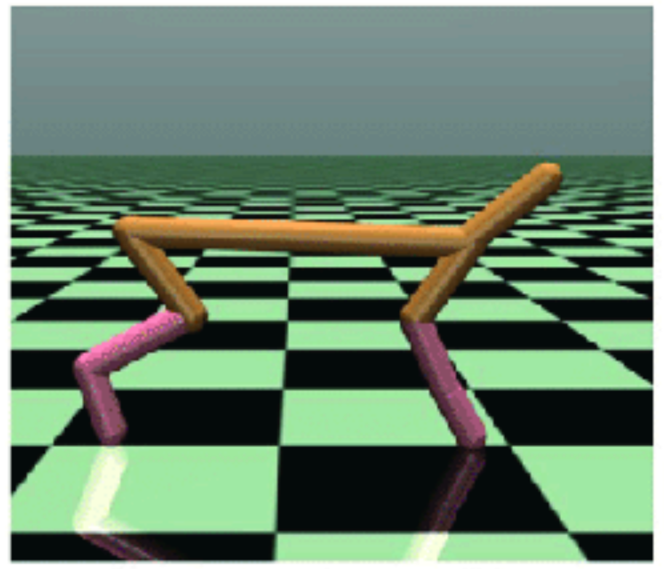
# Robust application-oriented exploration in LQ control

Jack Umenberger
Thomas B. Schön

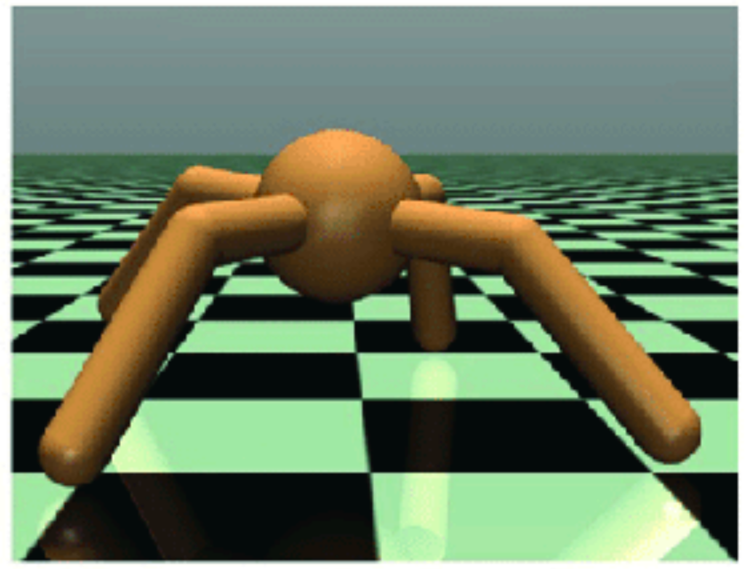
Mina Ferizbegovic
Håkan Hjalmarsson

# Goals

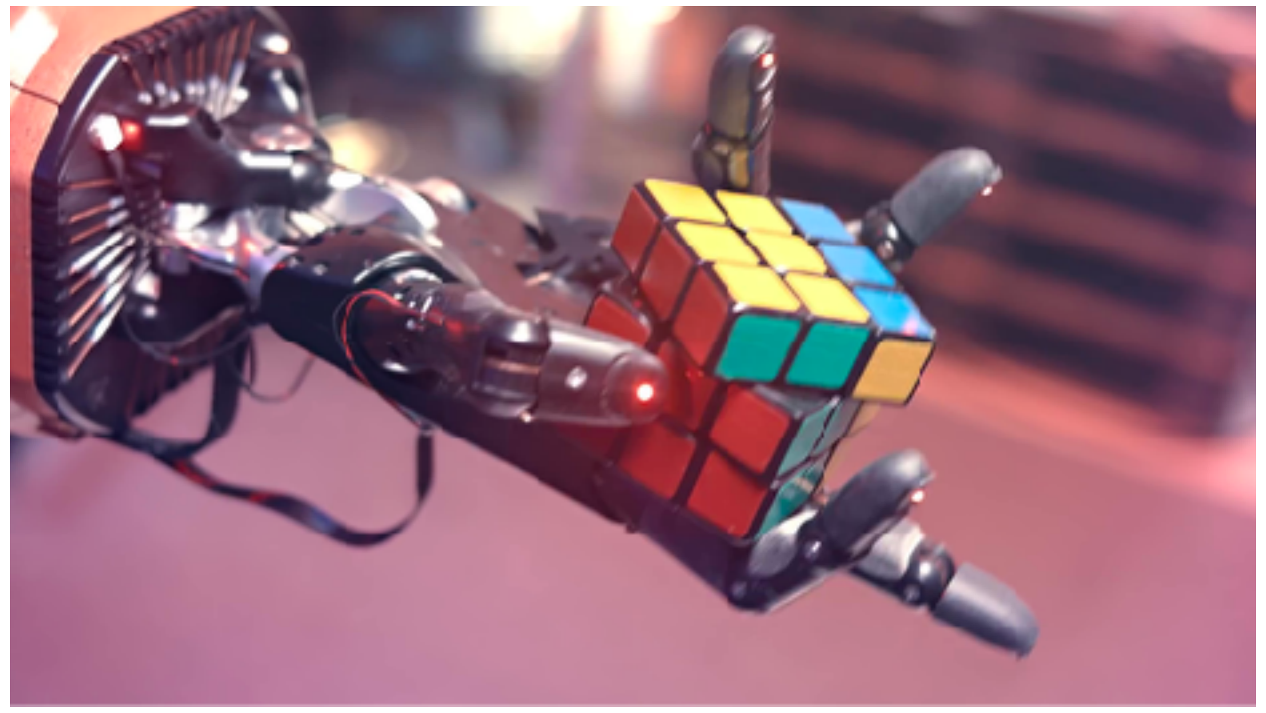


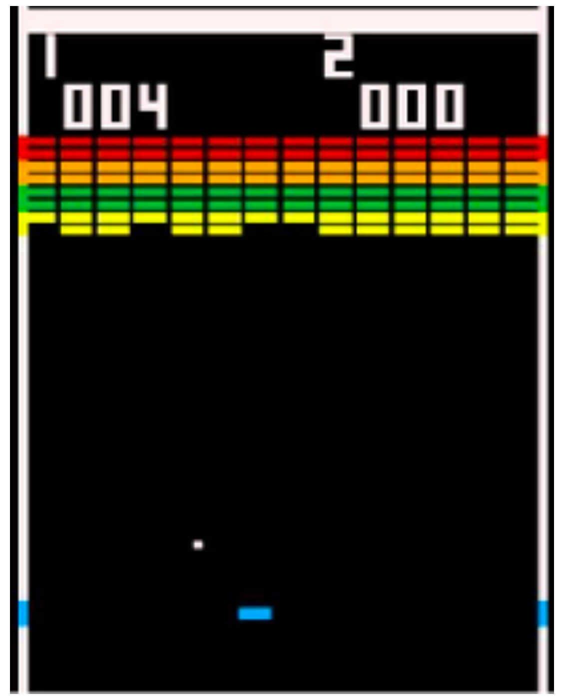


Half-Cheetah



Ant






- In reinforcement learning, agents need to explore their environment to learn which actions maximize rewards

- exploration is *often* random trial and error

- we want to do exploration that is **robust** and **targeted**

  - **robust:** does not destabilize the system or cause failure

  - **targeted:** provides knowledge that helps complete the task

# Balancing the tradeoff…

- There typically exists a **tradeoff** between exploration and exploitation
  - actions that provide the most information about the environment may incur high short term costs

# Balancing the tradeoff…

- There typically exists a **tradeoff** between exploration and exploitation
  - actions that provide the most information about the environment may incur high short term costs
- We achieve the **optimal tradeoff** between exploration and exploitation by formulating the search for a policy as a **convex optimization** problem, that can be solved to **global optimality**

# Balancing the tradeoff…

- There typically exists a **tradeoff** between exploration and exploitation
  - actions that provide the most information about the environment may incur high short term costs
- We achieve the **optimal tradeoff** between exploration and exploitation by formulating the search for a policy as a **convex optimization** problem, that can be solved to **global optimality**
- Relies only strong assumptions about the environment…

# Linear quadratic control

$$u_t \longrightarrow \boxed{\begin{aligned} x_{t+1} &= Ax_t + Bu_t + w_t \\ w_t &\sim \mathcal{N}(0, \Pi) \end{aligned}} \longrightarrow x_t$$

**Task:** find a state-feedback controller to minimize the quadratic cost

$$\sum_{t=0}^{T} x_t^\top Q x_t + u_t^\top R u_t$$

# Linear quadratic control

$$u_t \longrightarrow \boxed{\begin{array}{l} x_{t+1} = Ax_t + Bu_t + w_t \\ w_t \sim \mathcal{N}(0, \Pi), \end{array}} \longrightarrow x_t$$

**Task:** find a state-feedback controller to minimize the quadratic cost

$$\sum_{t=0}^{T} x_t^\top Q x_t + u_t^\top R u_t$$

**Challenge:** we don't know the system parameters $A, B, \Pi$

# Key quantity: empirical covariance

- The key quantity in our formulation is the empirical covariance $D = \sum_{t=0}^{T} \begin{bmatrix} x_t \\ u_t \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top$

- Shows up in both

# Key quantity: empirical covariance

- The key quantity in our formulation is the empirical covariance $D = \sum_{t=0}^{T} \begin{bmatrix} x_t \\ u_t \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top$

- Shows up in both

  - the cost, trace $\begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} D$, where "**smaller" is better** for lower cost

# Key quantity: empirical covariance

- The key quantity in our formulation is the empirical covariance $D = \sum_{t=0}^{T} \begin{bmatrix} x_t \\ u_t \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top$
- Shows up in both
  - the cost, trace $\begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} D$ , where “**smaller” is better** for lower cost
  - the system uncertainty (i.e. inverse variance of the posterior), $D \otimes I$, where “**bigger**” **is better** for reduced uncertainty
- This clearly illustrates the exploration (“big”) and exploitation (“small”) tradeoff

# Convex formulation

(Simplified) optimization problem:

$$\min_{D} \ \text{trace} \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} D \quad \leftarrow \quad \text{quadratic cost function} \ \sum_{t=0}^{T} x_t^\top Q x_t + u_t^\top R u_t$$

$$\text{s.t.} \ \begin{bmatrix} \star & \star \\ \star & D \end{bmatrix} \succeq 0 \quad \leftarrow$$

**robust stability constraint:**
**ensures stability of the true system with high probability**

# Convex formulation

(Simplified) optimization problem:

$$\min_{D} \ \text{trace} \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} D \quad \leftarrow \quad \textbf{quadratic cost function} \ \sum_{t=0}^{T} x_t^\top Q x_t + u_t^\top R u_t$$

$$\text{s.t.} \ \begin{bmatrix} \star & \star \\ \star & D \end{bmatrix} \succeq 0 \quad \leftarrow$$

**robust stability constraint:**
**ensures stability of the true system with high probability**

- As this is a convex program (linear objective + linear matrix inequality constraint) we can solve efficiently to **global optimality**
- This provides the **optimal balance** between exploration and exploitation

# Preserving structure in uncertainty

- recent works consider **absolute** measures of uncertainty

$$\|A_{\text{est}} - A_{\text{true}}\| \leq \epsilon_A \,, \quad \|B_{\text{est}} - B_{\text{true}}\| \leq \epsilon_B$$

- all uncertainty information is collapsed into a single **scalar**; structure is lost

# Preserving structure in uncertainty

- recent works consider **absolute** measures of uncertainty

$$\|A_{\text{est}} - A_{\text{true}}\| \leq \epsilon_A \,, \quad \|B_{\text{est}} - B_{\text{true}}\| \leq \epsilon_B$$

- all uncertainty information is collapsed into a single **scalar**; structure is lost

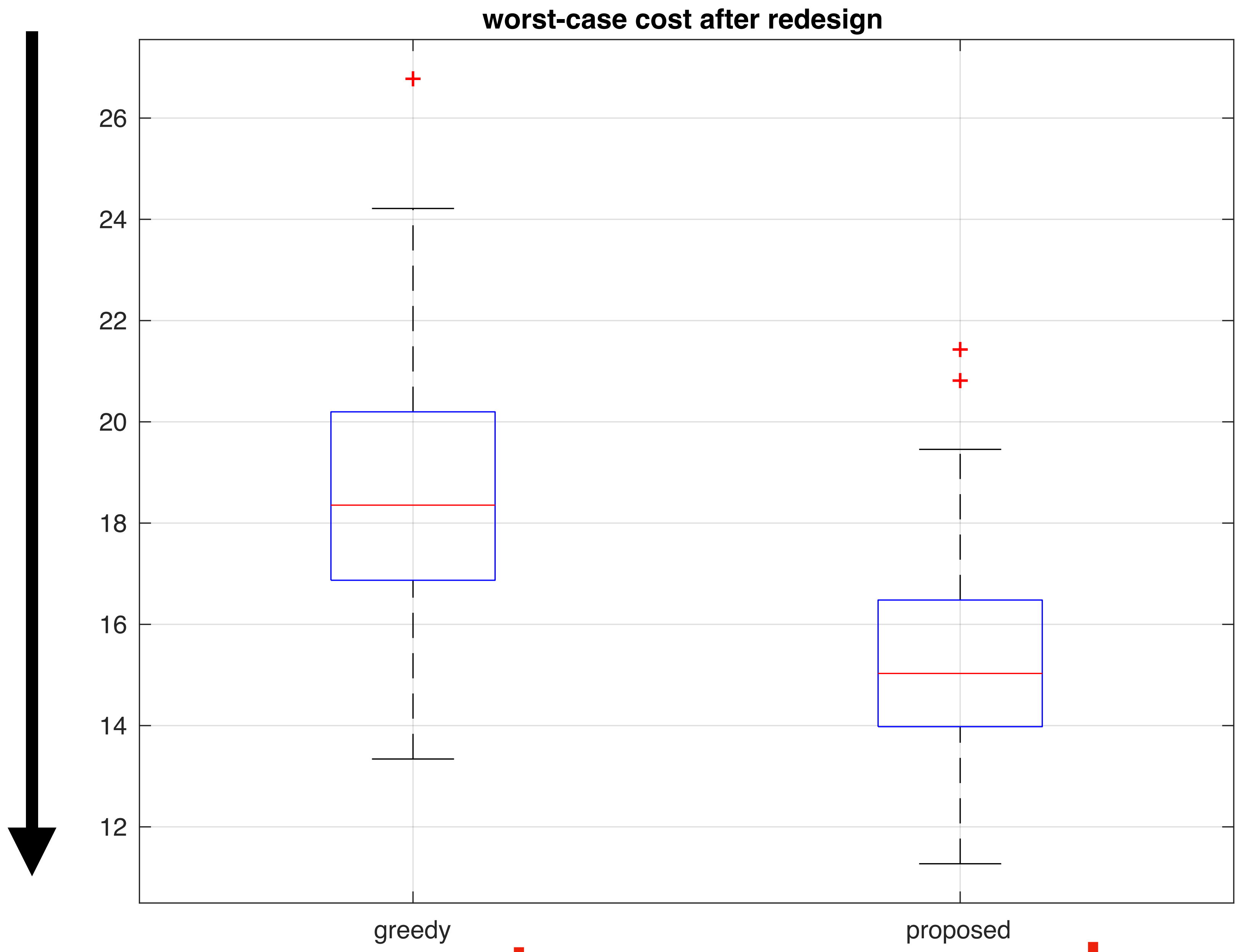
- we preserve structure by working with the **matrix** $D = \sum_{t=0}^{T} \begin{bmatrix} x_t \\ u_t \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top$

- allows to optimize for reduction of uncertainty in the parameters that **matter for the task**

# less uncertainty ≠ lower cost

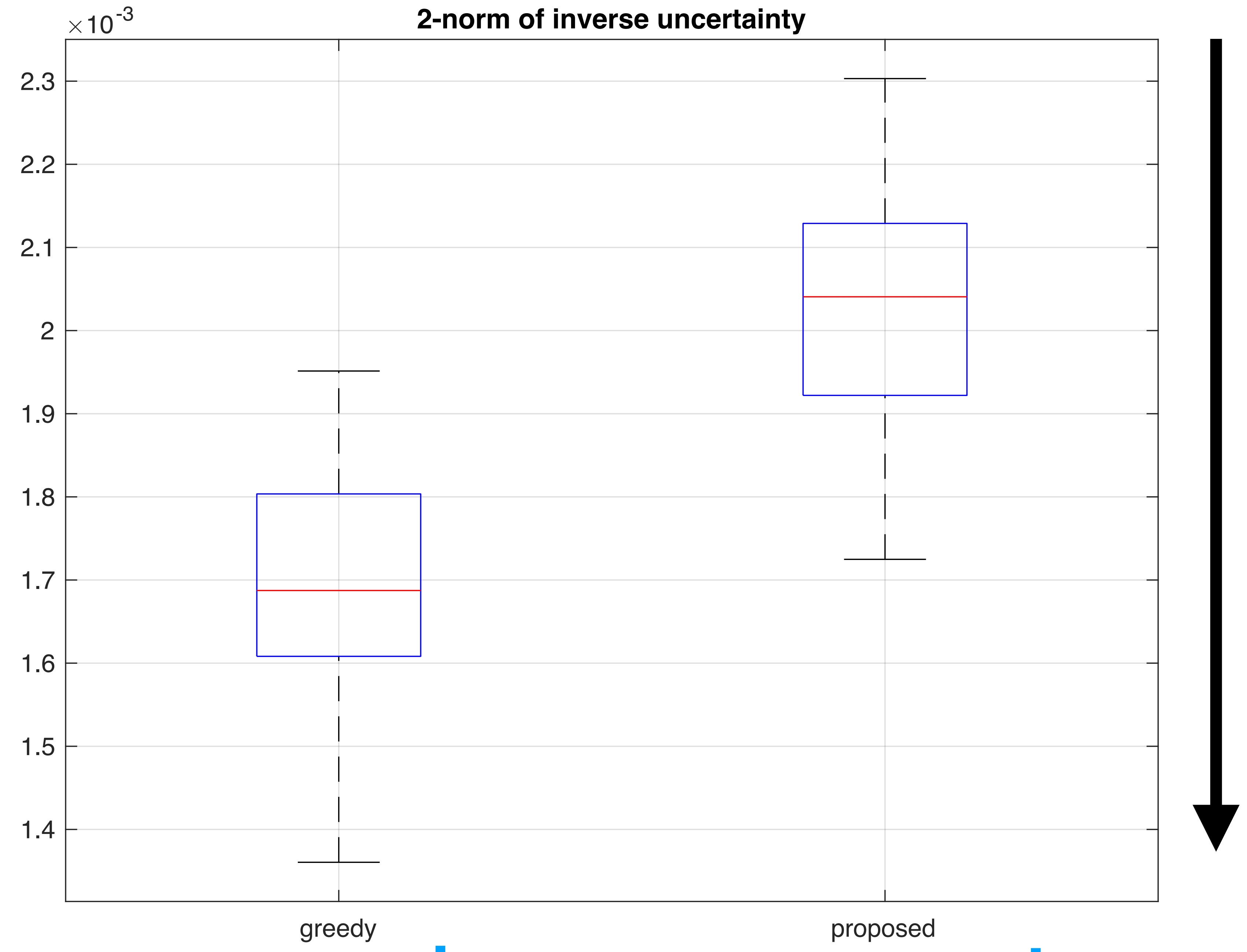
**total cost of policy**

lower cost (better performance)




$\epsilon$-greedy proposed

**absolute uncertainty**




lower absolute uncertainty

$\epsilon$-greedy proposed

- **performance** of proposed method is better, even though absolute **uncertainty** is larger
- the **structure** of the uncertainty matters
- uncertainty has been reduced in the parameters that matter for the control task

# Poster presentation

## Poster #177

### Today 05:00 -- 07:30 PM @ East Hall B+C

Swedish Foundation for Strategic Research