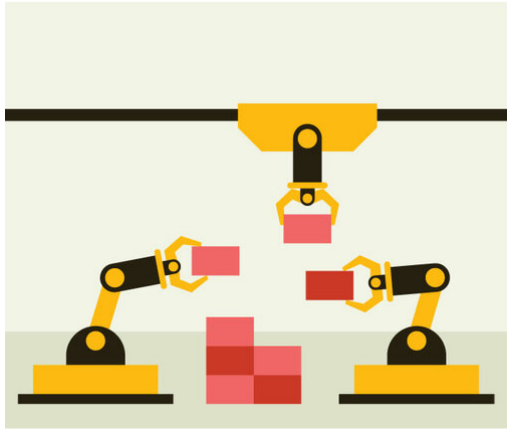
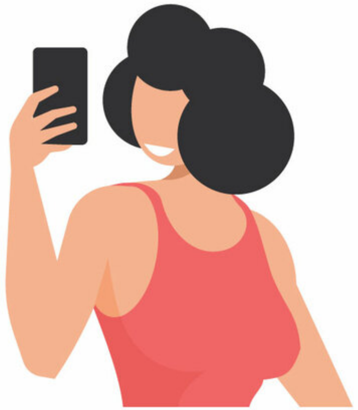
Learning and Decision-Making Under Observer Effects

Sarah Dean, Cornell University

Simons Workshop, April 2025


Large-scale automated systems enabled by machine learning



Large-scale automated systems enabled by machine learning

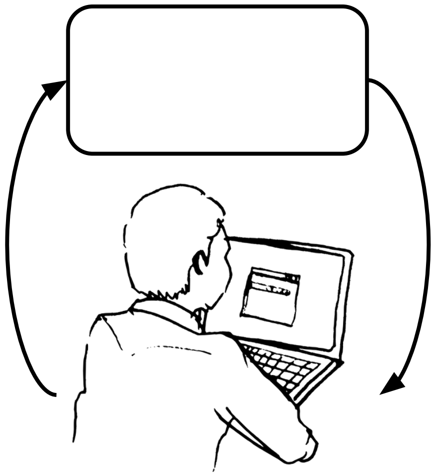
learning & decision-making algorithm

control theory + learning theory

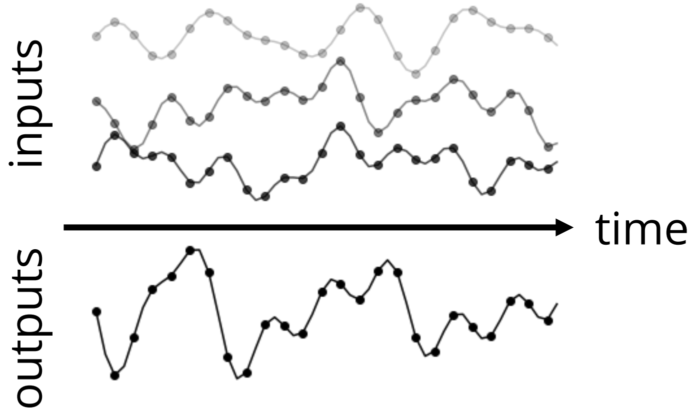
learning & decision-making algorithm

Outline

1. Motivation and Background



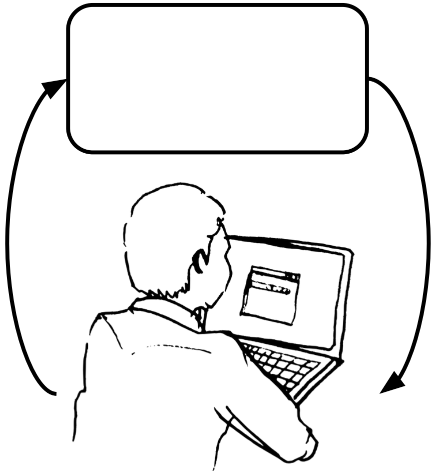
2. Learning Dynamics



3. Optimal Control

Outline

1. Motivation and Background

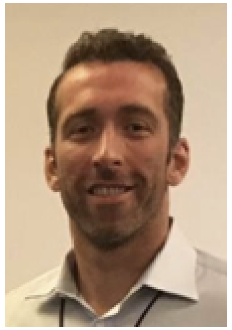
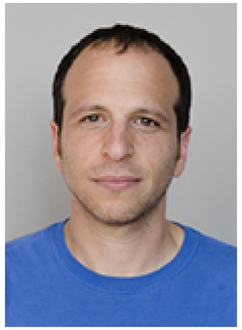


i) LQ Control

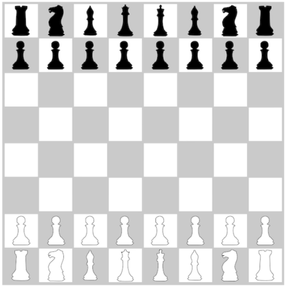
ii) Preference Dynamics

Sample Complexity of Control

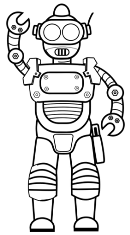
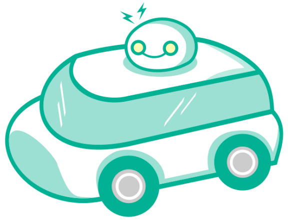
Work with Horia Mania, Nikolai Matni, Ben Recht, and Stephen Tu in 2017


Motivation: foundation for understanding RL & ML-enabled control



Classic RL setting: discrete problems and inspired by games

RL techniques applied to continuous systems interacting with the physical world

Sample Complexity: How much data is necessary to control a system?

- Given error ϵ and failure probability δ, how many samples are necessary to ensure $\mathbb{P}(\text{sub-opt.} \geq \epsilon) \leq \delta$?

Linear Quadratic Control

Simplest problem: linear dynamics, quadratic cost, zero mean noise

$$\text{minimize } \mathbb{E}\left[\sum_{t=0}^{T-1} x_t^\top Q x_t + u_t^\top R u_t\right]$$

$$\text{s.t. } x_{t+1} = Ax_t + Bu_t + w_t$$

Linear policy is optimal and can be computed in closed-form:

$$u_t = K_t^\star x_t$$

- where $K^\star = \{K_0^\star, ..., K_T^\star\}$ defined recursively depending on A, B, Q, R

Partial Observation LQ Control

Simplest problem: linear dynamics, quadratic cost, zero mean noise

$$\text{minimize } \mathbb{E}\left[\textstyle\sum_{t=0}^{T-1} x_t^\top Q x_t + u_t^\top R u_t\right]$$

$$\text{s.t. } x_{t+1} = A x_t + B u_t + w_t$$

$$y_t = C x_t + v_t$$

Linear policy is optimal and can be computed in closed-form:

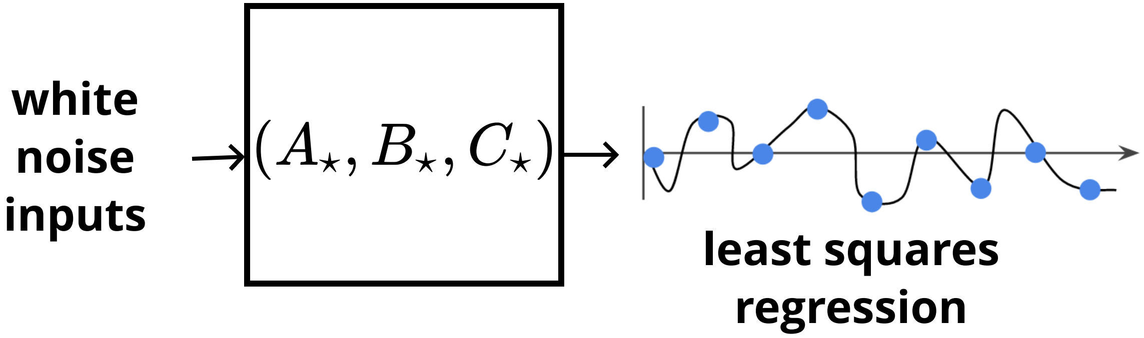
$$u_t = K_t^\star \hat{x}_t, \quad \hat{x}_t = \mathbb{E}[x_t | u_0, ..., u_t, y_0, ..., y_t]$$

(separation principle)

- where $K^\star = \{K_0^\star, ..., K_T^\star\}$ defined recursively depending on A, B, Q, R
- and $\hat{x}_t$ depends on $A, B, C, \Sigma_w, \Sigma_v, \Sigma_0$
- when noise is Gaussian, $\hat{x}_t$ computed efficiently with Kalman filter

Approach: ID then Control

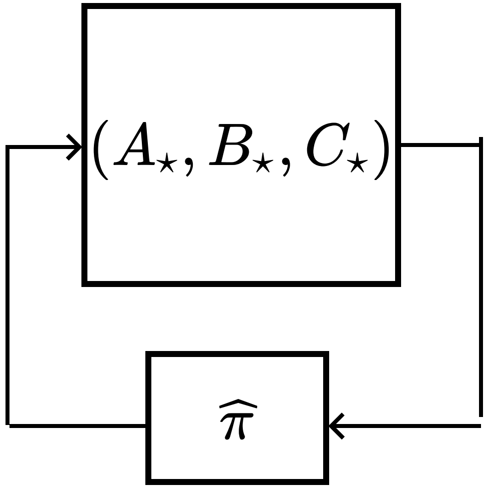
1. Collect N observations and estimate $\widehat{A}, \widehat{B}, \widehat{C}$



Learning Result (Informal):

parameter error $\lesssim \frac{1}{\sqrt{N}}$

2. Design policy as if estimate is true ("certainty equivalent")

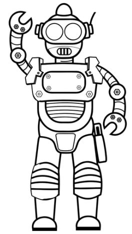


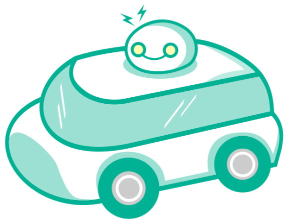
Control Result (Informal):

sub-opt. of $\widehat{\pi} \lesssim$ (param. err.)$^2 \lesssim \frac{1}{N}$

Naive exploration is essentially optimal!

Lessons from LQ Control





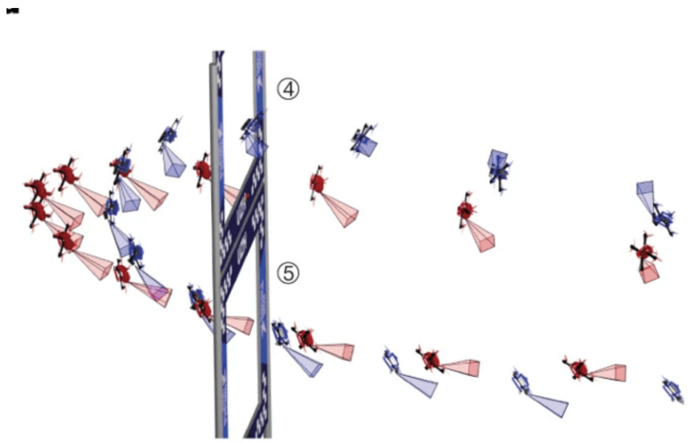
What lessons did we learn about RL & ML-enabled control?

1. Simple model-based approaches work (no need for model-free)
2. Naive exploration is sufficient, or even no exploration
3. No need to account for finite sample uncertainty*

$\Longrightarrow$ Problem does not capture all issues of interest!

*Exceptions: low data regime, safety/actuation limits

New Setting: Observer Effects

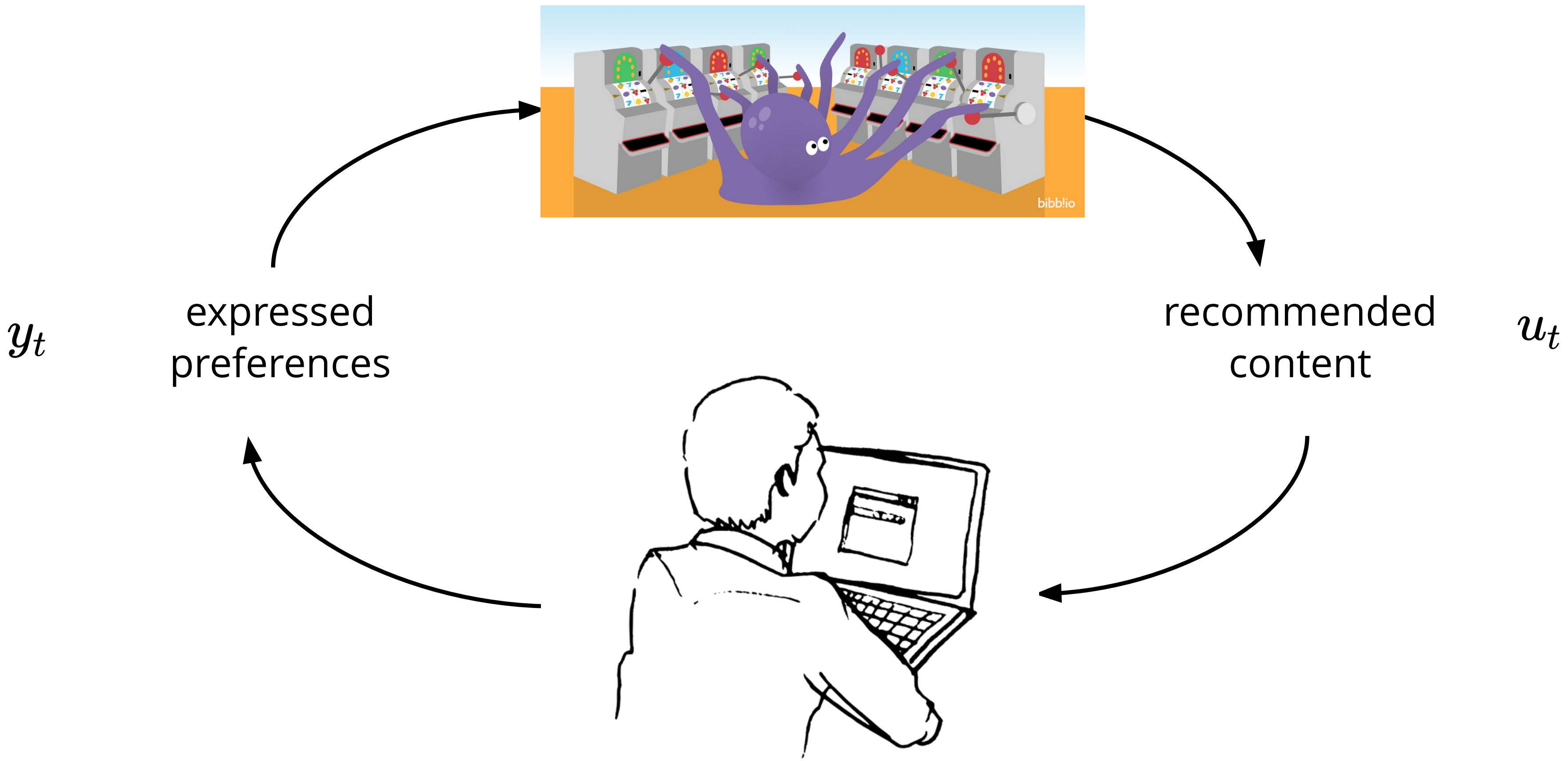


Observer effect: coupling between actuation and observation

- examples: electronic circuits, quantum wave collapse, human psychology, robotics ....

Example: Personalization

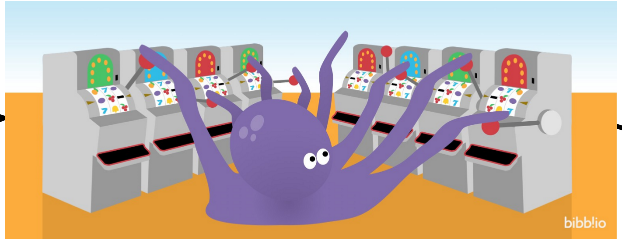
Classically studied as an online decision problem (e.g. *multi-armed bandits*)



unknown preference

Example: Personalization

Classically studied as an online decision problem (e.g. *multi-armed bandits*)



$\mathbb{E}[y_t] = \theta^\top u_t$

expressed
preferences

recommended
content

u_t

approach: identify θ
sufficiently well to make
good recommendations

unknown preference parameters θ

Example: Preference Dynamics

However, interests may be impacted by recommended content

recommender policy

$\mathbb{E}[y_t] = x_t^\top C u_t$ expressed preferences

recommended content u_t

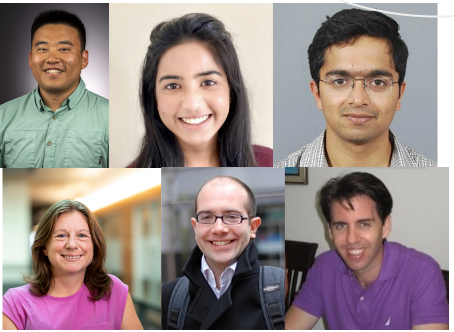
preference state x_t updates to x_{t+1}

Example: Preference Dynamics

- Simple dynamics that capture *assimilation* (adapted from opinion dynamics)

$$x_{t+1} \propto x_t + \eta_t u_t, \qquad y_t = x_t^\top u_t + v_t$$

- If η_t constant, tends to homogenization globally
- If $\eta_t \propto x_t^\top u_t$ (i.e. *biased* assimilation), tends to polarization (Hązła et al., 2019)

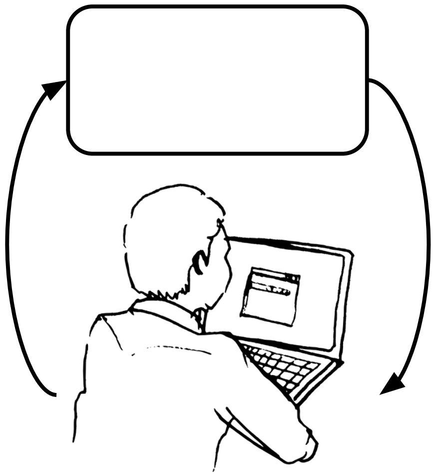


Implications for personalization [DM22]

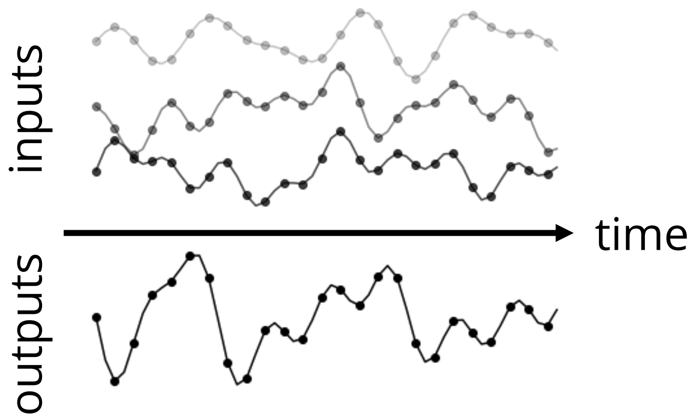
1. It is not necessary to estimate preferences to make "good" recommendations
2. Instead of polarization, preferences "collapse" towards whatever users are often recommended
3. Randomization can prevent such outcomes
4. Even if harmful content is never recommended, can cause harm through preference shifts [CKEWDI24]

Outline

1. Motivation and Background



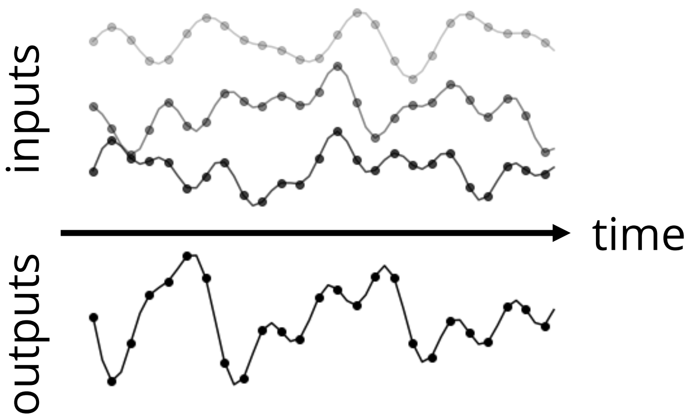
2. Learning Dynamics



3. Optimal Control

Outline

2. Learning Dynamics from Bilinear Observations



i) Setting

ii) Algorithm

iii) Results

Problem Setting: Identification

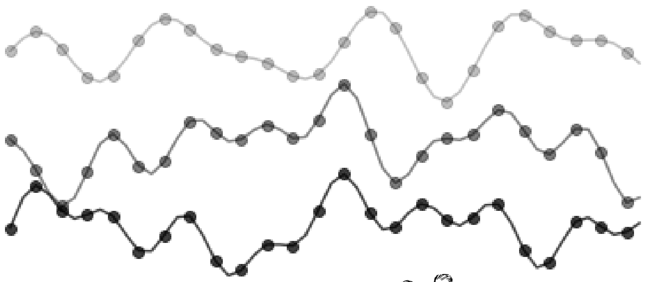
- Unknown dynamics and measurement functions
- Observed trajectory of inputs $u \in \mathbb{R}^p$ and outputs $y \in \mathbb{R}$

$$u_0, y_0, u_1, y_1, \ldots, u_T, y_T$$

- Goal: identify dynamics and measurement models from data
- Setting: linear/bilinear with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times p}$, $C \in \mathbb{R}^{p\times n}$

$$x_{t+1} = Ax_t + Bu_t + w_t$$
$$y_t = u_t^\top C x_t + v_t$$

inputs u_t



e.g. playlist attributes

outputs y_t

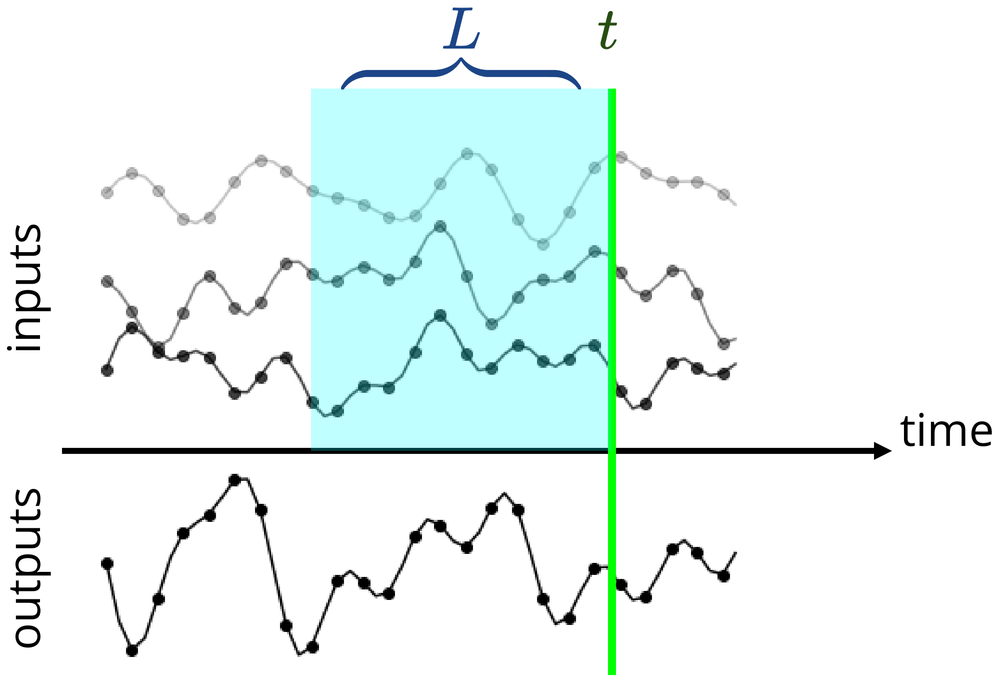
e.g. listen time

Identification Algorithm

Yahya Sattar

Yassir Jedra

Input: data $(u_0, y_0, \ldots, u_T, y_T)$, history length L, state dim n



Step 1: Regression

$$\hat{G} = \arg\min_{G \in \mathbb{R}^{p \times pL}} \sum_{t=L}^{T} \left(y_t - u_t^\top \textstyle\sum_{k=1}^{L} G[k] u_{t-k}\right)^2$$

Step 2: Decomposition $\hat{A}, \hat{B}, \hat{C} = \mathrm{HoKalman}(\hat{G}, n)$

(Omyak & Ozay, 2019)

Estimation Errors

$$\hat{G} = \arg\min_{G \in \mathbb{R}^{p \times pL}} \sum_{t=L}^{T} \left(y_t - \bar{u}_{t-1}^\top \otimes u_t^\top \mathrm{vec}(G) \right)^2$$

- (Biased) estimate of Markov parameters

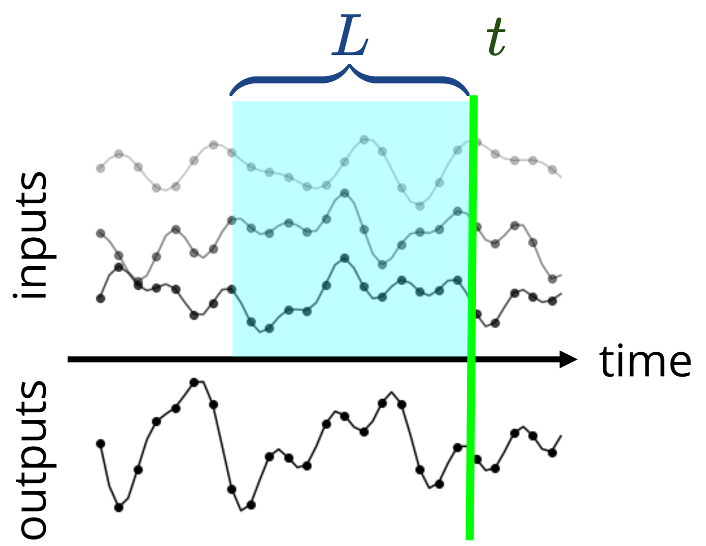
$$G = \begin{bmatrix} CB & CAB & \dots & CA^{L-1}B \end{bmatrix}$$

- Regress y_t against

$$\underbrace{\begin{bmatrix} u_{t-1}^\top & \dots & u_{t-L}^\top \end{bmatrix}}_{\bar{u}_{t-1}^\top} \otimes u_t^\top$$

- Data matrix: circulant-like structure

$$Z = \begin{bmatrix} \bar{u}_{L-1}^\top \otimes u_L^\top \\ \vdots \\ \bar{u}_{T-1}^\top \otimes u_T^\top \end{bmatrix}$$



Main Results

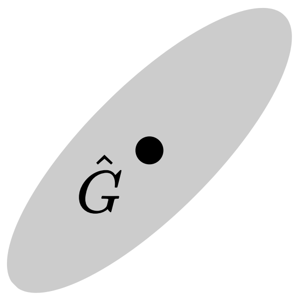
Assumptions:

1. Process and measurement noise w_t, v_t are i.i.d., zero mean, and have bounded second moments
2. Inputs u_t are bounded
3. The dynamics are strictly stable, i.e. $\rho(A) < 1$

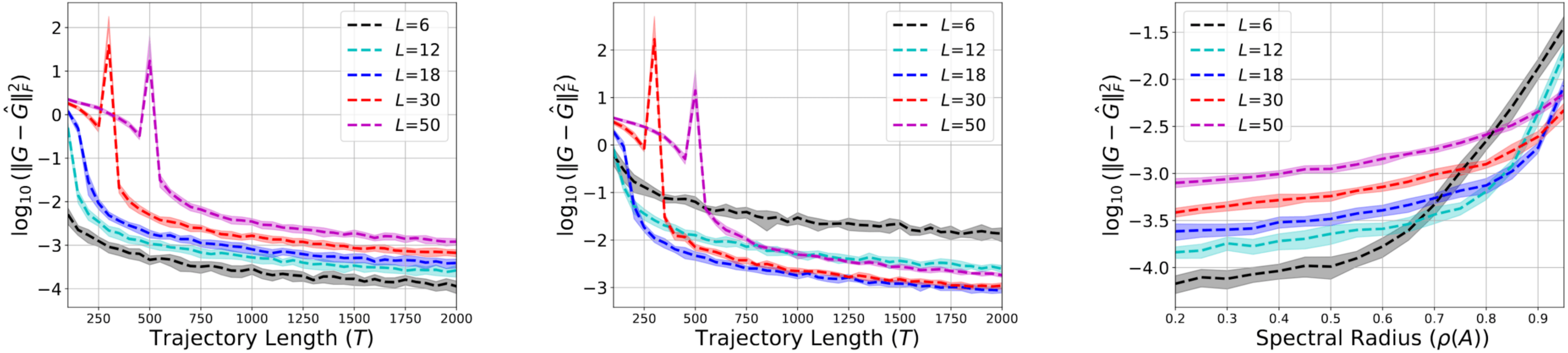
Informal Theorem (Markov parameter estimation)

With probability at least $1 - \delta$,

$$\|G - \hat{G}\|_{Z^\top Z} \lesssim \sqrt{\frac{p^2 L}{\delta}} \cdot c_{\text{stability,noise}} + \rho(A)^L \sqrt{T} c_{\text{stability}}$$



Main Results



(a) $\rho(\boldsymbol{A}) \leq 0.5$ (b) $\rho(\boldsymbol{A}) \leq 0.99$ (c) varying $\rho(\boldsymbol{A})$

Informal Summary Theorem

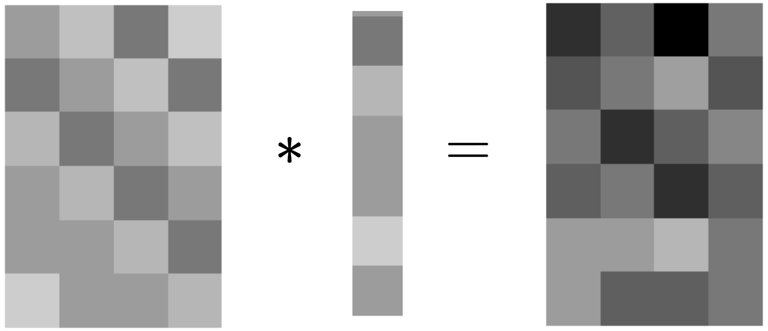
With high probabilty,

$$\text{est. errors} \lesssim \sqrt{\frac{\mathsf{poly}(\text{dimension})}{\sigma_{\min}(Z^\top Z)}} \lesssim \sqrt{\frac{\mathsf{poly}(\text{dim.})}{T}}$$

Proof Sketch

$$\hat{G} = \arg\min_{G \in \mathbb{R}^{p \times pL}} \sum_{t=L}^{T} \left(y_t - \bar{u}_{t-1}^\top \otimes u_t^\top \mathrm{vec}(G) \right)^2$$

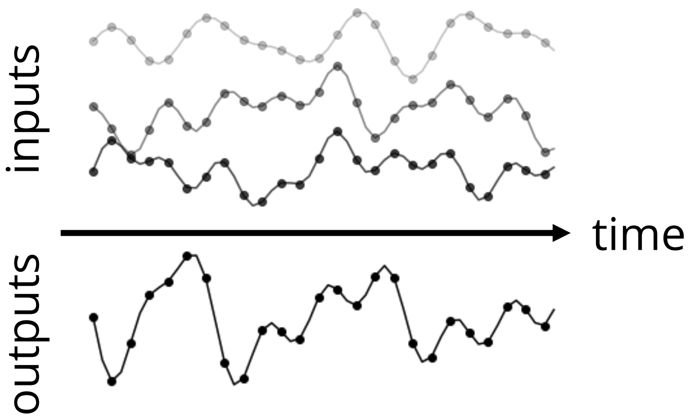
- Claim: this is a biased estimate of Markov parameters
$$G_\star = \begin{bmatrix} CB & CAB & \dots & CA^{L-1}B \end{bmatrix}$$
  - Observe that $x_t = \sum_{k=1}^{L} CA^{k-1}B(u_{t-k} + w_{t-k}) + A^L x_{t-L}$
  - Hence, $y_t = \bar{u}_{t-1}^\top \otimes u_t^\top \mathrm{vec}(G_\star) + u_t^\top \sum_{k=1}^{L} CA^{k-1} w_{t-k} + u_t CA^L x_{t-L} + v_t$
- Least squares: for $y_t = z_t^\top \theta + n_t$, the estimate $\hat{\theta} = \arg\min \sum_t (z_t^\top \theta - y_t)^2$
$$= \arg\min \ \|Z\theta - Y\|_2^2 = (Z^\top Z)^\dagger Z^\top Y = \theta_\star + (Z^\top Z)^\dagger Z^\top N$$
- Estimation errors are therefore $\|G_\star - \hat{G}\|_{Z^\top Z} = \|Z^\top N\|$
- Blocking technique to bound minimum singular value of Z



$$= \begin{bmatrix} \bar{u}_{L-1}^\top \otimes u_L^\top \\ \vdots \\ \bar{u}_{T-1}^\top \otimes u_T^\top \end{bmatrix} = Z$$

Summary

2. Learning Dynamics from Bilinear Observations



Algorithm: nonlinear features

Analysis: blocking technique

Exploration: similar to linear setting

Outline

3. Optimal Control with Bilinear Observations

i) Setting

ii) Separation Principle

iii) Results

Problem Setting: Optimal Control

- Linear state update with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times p}$

$$x_{t+1} = Ax_t + Bu_t + w_t$$

- Bilinear measurements with $C_i \in \mathbb{R}^{m\times n}$

$$y_t = \underbrace{\left(C_0 + \sum_{i=1}^{p} u_t[i]C_i\right)}_{C(u_t)} x_t + v_t$$

- Quadratic costs with $Q, R \succ 0$

$$c(x,u) = x^\top Qx + u^\top Ru$$

- Gaussian process noise, measurement noise, and initial state

$$\{w_t\} \sim \mathcal{N}(0,\Sigma_w), \quad \{v_t\} \sim \mathcal{N}(0,\Sigma_v), \quad x_0 \sim \mathcal{N}(0,\Sigma_0)$$

- Information set for decision-making

$$\mathcal{I}_t = \{u_0, ..., u_{t-1}, y_0, ..., y_{t-1}\}$$

Problem Setting: Optimal Control

$$
\begin{aligned}
\min_{u_t = \pi_t(\mathcal{I}_t)} \quad & \mathbb{E}\left[x_T^\top Q x_T + \sum_{t=1}^{T-1} x_t^\top Q x_t + u_t^\top R u_t \right] \\
\text{s.t.} \quad & x_{t+1} = A x_t + B u_t + w_t \\
& y_t = \left(C_0 + \sum_{i=1}^{p} u_t[i] C_i \right) x_t + v_t
\end{aligned}
$$

Small departure from classic LQG control

Separation Principle

- Separation principle (SP): for partially observed optimal control, it suffices to independently design estimation & control
  - Optimal for partially observed LQ control
- The SP policy has two components:
  1. State estimation $\hat{x}_t = \mathbb{E}[x_t | \mathcal{I}_t]$
  2. State dependent policy $u_t = K_t^\star \hat{x}_t$

State Estimation

- As in LQG, we use the Kalman Filter
  - $\hat{x}_{t+1} = A\hat{x}_t + Bu_t - L_t\big(y_t - C(u_t)\hat{x}_t\big)$
  - $\Sigma_{t+1} = (A + L_tC(u_t))\Sigma_t A^\top + \Sigma_w$
  - $L_t = -A\Sigma_t C(u_t)^\top (C(u_t)\Sigma_t C(u_t)^\top + \Sigma_v)^{-1}$
- **Lemma:** the posterior distribution is given by the Kalman filter

$$x_t | \mathcal{I}_t \sim \mathcal{N}(\hat{x}_t, \Sigma_t)$$

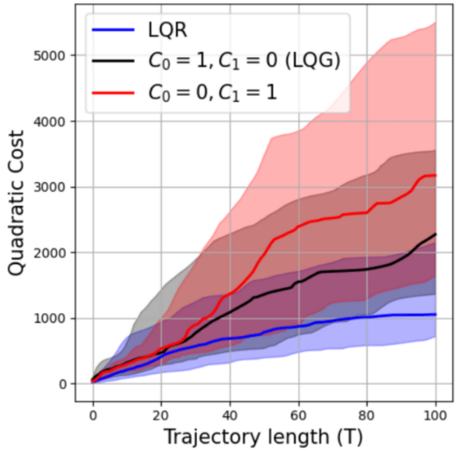
- Unlike the standard linear setting, there is a nonlinear dependence on the inputs

Example

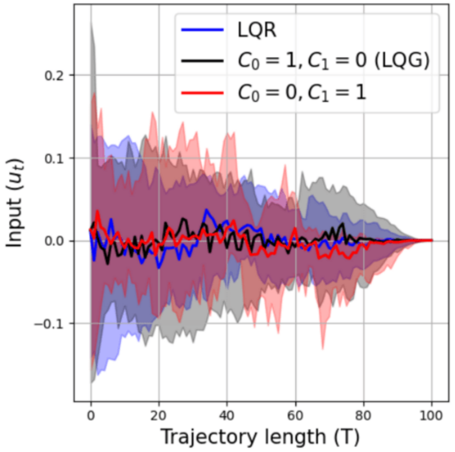
$$x_{t+1} = \begin{bmatrix} 1 & 0.3 \\ 0 & 1 \end{bmatrix} x_t + \begin{bmatrix} 0.3 \\ 0 \end{bmatrix} u_t + w_t$$

$$y_t = (C_0 + C_1 u_t) \begin{bmatrix} 1 & 0 \end{bmatrix} x_t + v_t$$

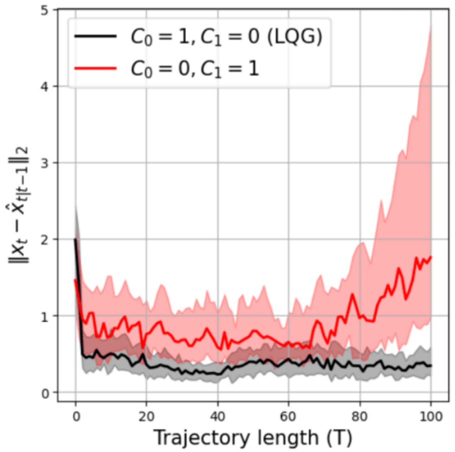
with $Q = I$ and $R = 1000$



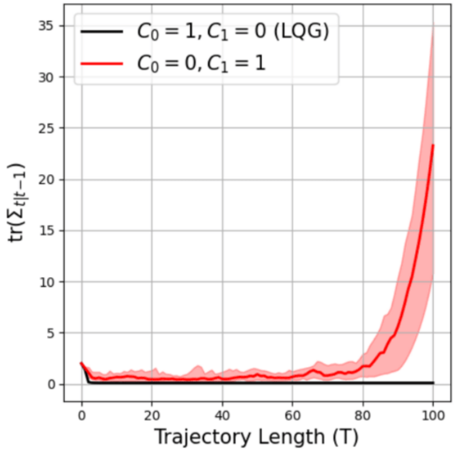
(a) Quadratic cost



(b) Inputs



(c) KF state estimation error



(d) KF covariance trace

Main Results

- **Theorem:** For $T \geq 2$, the optimal policy is not affine in the estimated state
  - as a consequence, the SP policy is not optimal
- **Theorem:** There exist instances in which the SP policy locally maximizes the cost
  - in these instances, the optimal controller is nonlinear and not unique, i.e. for scalar system at $t = T - 2$,

$$u_t^\star = -\alpha \hat{x}_t \left(1 \pm \frac{1}{K_t \hat{x}_t} \sqrt{-\frac{\Sigma_z}{\Sigma_t} + \beta K_t} \right)$$

Proof Sketch

- Strategy: analyze solution to dynamic programming
- At $t = T$, the value function is $V_T(x) = x^\top Q x$
- At $t = T-1$,

$$V_{T-1}(x_t) = \min_u \underbrace{\mathbb{E}[c(x_t, u) + V_T(x_{t+1}) | \mathcal{I}_{T-1}]}_{f_{T-1}(u) = f_{T-1}^{\mathrm{LQ}}(u)}$$

  - The solution coincides with LQG

$$u_{T-1}^\star = K_{T-1}^\star \mathbb{E}[x_{T-1} | \mathcal{I}_{T-1}]$$

- At $t = T-2$, due to dependence of state estimation on input

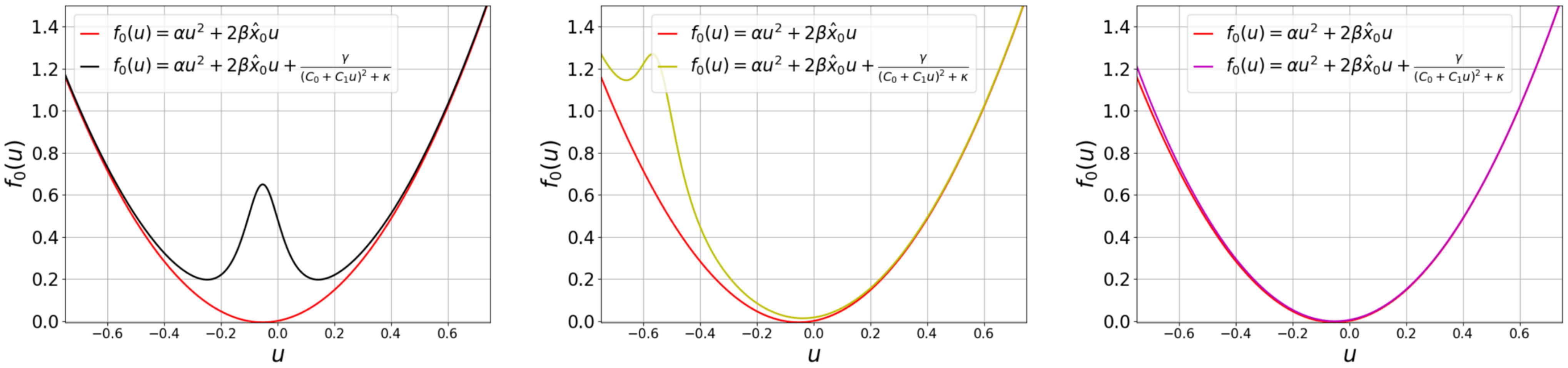
$$\min_u \underbrace{\mathbb{E}[c(x_t, u) + V_{T-1} x_{t+1}) | \mathcal{I}_{T-2}]}_{f_{T-2}(u) = f_{T-2}^{\mathrm{LQ}}(u) + f_{T-2}^{\mathrm{obs}}(u)}$$

Proof Sketch

- Strategy: analyze solution to dynamic programming
- At $t = T$, the value function is $V_T(x) = x^\top Q x$
- At $t = T-1$, $u^\star_{T-1} = K^\star_{T-1}\mathbb{E}[x_{T-1}|\mathcal{I}_{T-1}]$

- At $t = T-2$, due to dependence of state estimation on input

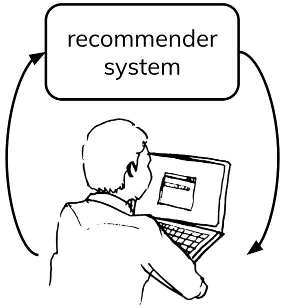
$$\min_u \underbrace{\mathbb{E}[c(x_t, u) + V_{T-1}x_{t+1})|\mathcal{I}_{T-2}]}_{f_{T-2}(u) = f^{\mathrm{LQ}}_{T-2}(u) + f^{\mathrm{obs}}_{T-2}(u)}$$



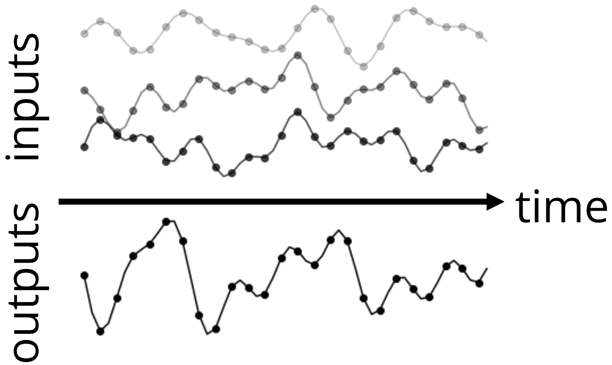
(b) $-\frac{C_0}{C_1} = -\frac{\beta\hat{x}_0}{\alpha}$ (c) $-\frac{C_0}{C_1} = -\frac{\beta\hat{x}_0}{\alpha} - 0.5$ (d) $-\frac{C_0}{C_1} = -\frac{\beta\hat{x}_0}{\alpha} - 1.0$

Summary

1. Motivation and Background



2. Learning Dynamics



3. Optimal Control

Open Directions:

- Learning without stability assumption on A
  - Idea: regress y_t against $u_{0:t}, y_{0:t-1}$
- Control with guarantees of stability or sub-optimality
  - Idea: observability-aware inputs
- Full sample complexity analysis
  - State estimation & control with imperfect models

Thanks! Questions?

- **Learning Linear Dynamics from Bilinear Observations** at ACC25 (arxiv:2409.16499) with Yahya Sattar and Yassir Jedra
- **Sub-optimality of the Separation Principle for Quadratic Control from Bilinear Observations** (arxiv:2504.11555) with Yahya Sattar, Sunmook Choi, Yassir Jedra, Maryam Fazel




- **On the Sample Complexity of the Linear Quadratic Regulator** in FoCM (arxiv.org:1710.01688) with Horia Mania, Nikolai Matni, Benjamin Recht, Stephen Tu
- **Preference Dynamics Under Personalized Recommendations** at EC22 (arxiv:2205.13026) with Jamie Morgenstern
- **Harm Mitigation in Recommender Systems under User Preference Dynamics** at KDD24 (arxiv:2406.09882) with Chee, Kalyanaraman, Ernala, Weinsberg, Ioannidis

$$H^{(K)} := \begin{bmatrix} CB & CAB & \cdots & CA^{K-1}B \\ CAB & CA^2B & \cdots & CA^KB \\ \vdots & \vdots & \ddots & \vdots \\ CA^{K-1}B & CA^KB & \cdots & CA^{2(K-1)}B \end{bmatrix} \in \mathbb{R}^{pK \times pK}$$

$$\hat{H}^{(K)} := \begin{bmatrix} \hat{G}_0 & \hat{G}_1 & \cdots & \hat{G}_{L-2} & \hat{G}_{L-1} & 0 & \cdots & 0 \\ \hat{G}_1 & \hat{G}_2 & \cdots & \hat{G}_{L-1} & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{G}_{L-1} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{pK \times pK}$$

(Oymak & Ozay, 2019)