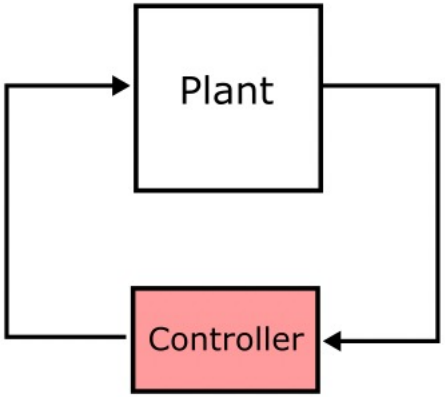
Sample complexity analysis of input/output model identification

representations, over-parameterization, and self-regularization

Necmiye Ozay
Electrical Engineering and Computer Science
Robotics Institute
University of Michigan

ELLIIT Hybrid AI Workshop
November 2, 2022

Models + Task → Control

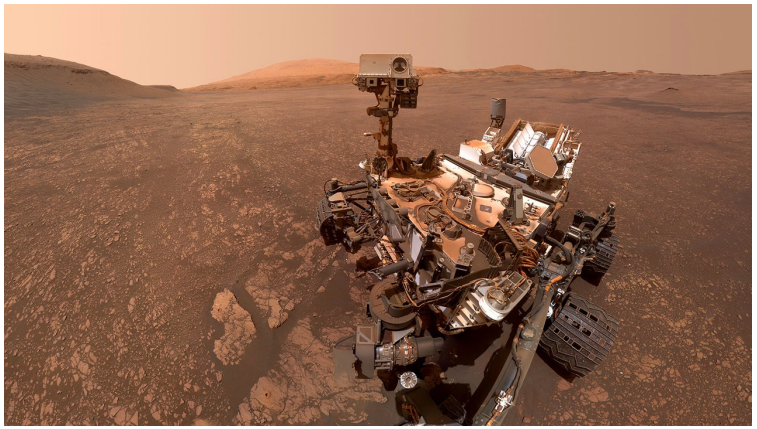


Data → Models + Task → Control

Data → Control

Data → **Models** + Task → Control

Models useful for (i) control design, (ii) fast simulations (iii) system monitoring, (iv) anomaly detection, etc.



From: NASA

Exploring unknown environments

From: nbcnews.com

Handling unexpected failures

"Small" data

Online system identification ➔ learning models at run-time

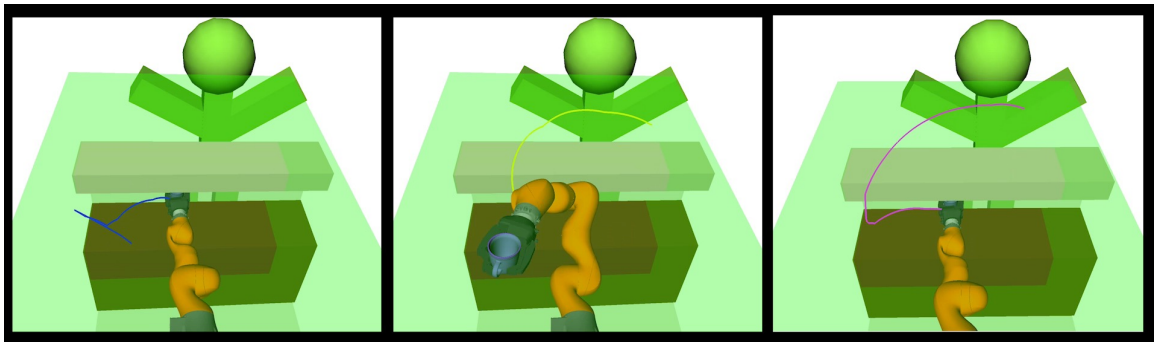
↓

Adaptation (change the objective)

(repurposing, changing mission objectives)

Data → Models + **Task** → Control

Task specifications useful for (i) control design
(ii) system monitoring, (iii) anomaly detection, etc.



Teaching a robot to do a task

"Small" data

Learning from demonstrations → constraint/task learning

↓

Generalization
(performing the tasks in new environments)

LTI system identification

Given input/output data ($\{u_t, y_t\}_{t=0}^{N}$) coming from a linear time invariant (LTI) system subject to noise (w_t, z_t), find a system model.

LTI system identification

Given input/output data ($\{u_t, y_t\}_{t=0}^{N}$) coming from a linear time invariant (LTI) system subject to noise (w_t, z_t), find a system model.

Some questions:

- What model/parametrization to choose?
  - LTI systems can be modeled in different ways (state-space models, auto-regressive models, impulse response models)
  - How to pick a system order?
- What algorithm to use?
  - Typically posed as an optimization: What is the computational complexity? Do we need regularization and if so, how?
- What type of guarantees can we give for these algorithms?

LTI system identification

Given input/output data ($\{u_t, y_t\}_{t=0}^{N}$) coming from a linear time invariant (LTI) system subject to noise (w_t, z_t), find a system model.

- **Asymptotic analysis:**
  - As the data size *N* goes to infinity and/or noise (w_t, z_t) level goes to zero, can we learn the system model?
- **Non-asymptotic analysis:**
  - Given finite amount of noisy data, how does the identification accuracy depend on the data size *N* and noise?
  - What can the best identification algorithm achieve in this case?

Existing results (incomplete list)

- **Asymptotic analysis:**
  - As the data size *N* goes to infinity and/or noise (w_t, z_t) level goes to zero, can we learn the system model?

    Textbook on sys id: [Ljung 99], *standard methods*: Ho-Kalman (Eigen Realization Algorithm-ERA), N4SID, etc.

- **Non-asymptotic analysis:**
  - Given finite amount of noisy data, how does the identification accuracy depend on the data size *N* and noise?
  - What can the best identification algorithm achieve in this case?

    Control theoretic methods: [Weyer et al. 99], [Vidyasagar & Karandikar 01], [Campi & Weyer 02], [Akcay 04], [Carè et al. 18], etc.

    Statistical machine learning methods: [Hardt et al. 16], [Dean et al. 17], [Hazan et al. 17], [Tu et al. 17], [Sarkar & Rakhlin 18], [Simchowitz et al. 18], etc.

Detour: random design linear regression

Assume we want to estimate a static quantity α from *N* noisy measurements y:

$$y = U\alpha + \varepsilon$$

where U has iid Gaussian entries and ε is zero-mean iid.

- **Asymptotic analysis:** least squares estimator is consistent, i.e., as *N* goes to infinity, the estimate converges to the true value in probability
- **Non-asymptotic analysis:** least squares estimate satisfies, with high probability,

$$\|\hat{\alpha} - \alpha\| \leq O(\frac{1}{\sqrt{N}})$$

Hsu, Kakade, Zhang, COLT'12

Plan

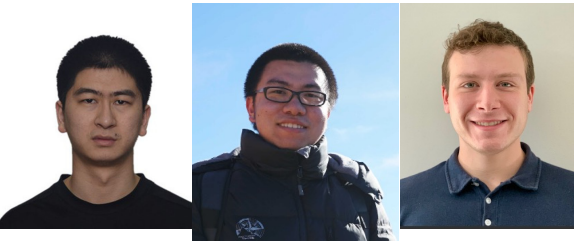
Focus on a well-known system identification algorithms:

- **Can we achieve similar sample complexity results for system identification algorithms where the data is highly correlated?**

- **Part 1: State-space models – Ho-Kalman Algorithm**

Joint work with Samet Oymak, UC Riverside
ACC'19, TAC'22

- **Part 2: Autoregressive models – Ordinary least squares**



Joint work with Zhe Du, Zexiang Liu, Jack Weitze, Michigan
CDC'22

LTI system identification

Given input/output data ($\{u_t, y_t\}_{t=0}^{N}$) coming from a linear time invariant (LTI) system subject to noise, find a system model.

State-space model:

$$x_{t+1} = Ax_t + Bu_t + w_t$$
$$y_t = Cx_t + Du_t + z_t$$

ARX model:

$$y_t = \sum_{i=1}^{n_\alpha} \alpha_i y_{t-i} + \sum_{i=1}^{n_\beta} \beta_i u_{t-i} + \eta_{t-1}$$

Plan

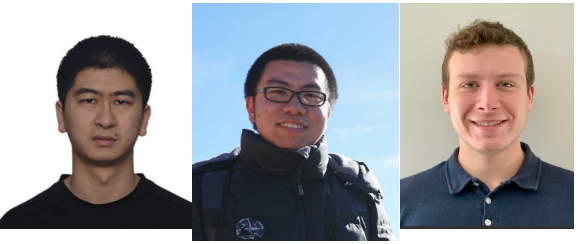
Focus on a well-known system identification algorithms:

- **Can we achieve similar sample complexity results for system identification algorithms where the data is highly correlated?**

- **Part 1: State-space models – Ho-Kalman Algorithm**

Joint work with Samet Oymak, UC Riverside
ACC'19, TAC'22

- **Part 2: Autoregressive models – Ordinary least squares**



Joint work with Zhe Du, Zexiang Liu, Jack Weitze, Michigan
CDC'22

Ho-Kalman Algorithm

Given input/output data ($\{u_t, y_t\}_{t=0}^{N}$), find a model of the form:

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

$$y_t = Cx_t + Du_t$$

- Identification problem is **ill-posed**:
  - we can only learn up to a similarity transformation (change of basis)

Ho-Kalman Algorithm

Given input/output data ($\{u_t, y_t\}_{t=0}^{N}$), find a model of the form:

$$x_{t+1} = Ax_t + Bu_t$$
$$y_t = Cx_t + Du_t$$

- Identification problem is **ill-posed**:
  - we can only learn up to a similarity transformation (change of basis)

$$\tilde{x} = Px \Rightarrow \begin{aligned} \tilde{x}_{t+1} &= PAP^{-1}\tilde{x}_t + PBu_t \\ y_t &= CP^{-1}\tilde{x}_t + Du_t \end{aligned}$$

Ho-Kalman Algorithm

Given input/output data ($\{u_t, y_t\}_{t=0}^{N}$), find a model of the form:

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

$$y_t = Cx_t + Du_t$$

- Identification problem is **ill-posed**:
  - we can only learn up to a similarity transformation (change of basis).
  - we can only learn the controllable and observable part
- *Assume:* The system is controllable and observable

Ho-Kalman Algorithm

Given input/output data ($\{u_t, y_t\}_{t=0}^{N}$), find a model of the form:

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

$$y_t = Cx_t + Du_t$$

- Identification problem is **ill-posed**:
  - we can only learn up to a similarity transformation (change of basis).
  - we can only learn the controllable and observable part
- *Assume:* The system is controllable and observable
- Two step procedure:
  1. Estimate the Markov parameters of the system:

     $$D, CB, CAB, CA^2B, \ldots, CA^tB, \ldots$$

     Markov parameters are **invariant** to the choice of basis
  2. Estimate the "system matrices" from Markov parameters

Ho-Kalman Algorithm

Given input/output data ($\{u_t, y_t\}_{t=0}^{N}$), find a model of the form:

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

$$y_t = Cx_t + Du_t$$

- Identification problem is **ill-posed**:
  - we can only learn up to a similarity transformation (change of basis).
  - we can only learn the controllable and observable part
- *Assume:* The system is controllable and observable
- Two step procedure:
  1. Estimate the Markov parameters of the system:

     $$D, CB, CAB, CA^2B, \ldots, CA^tB, \ldots$$

     Markov parameters are **invariant** to the choice of basis
  2. **Estimate the "system matrices" from Markov parameters**

Ho-Kalman Algorithm

- Assume Markov parameters of the system are given:

$$D, CB, CAB, CA^2B, \ldots, CA^tB, \ldots$$

- Form the Hankel matrix of Markov parameters

$$\overbrace{\begin{bmatrix} CB & CAB & CA^2B & \ldots & CA^{T_2}B \\ CAB & CA^2B & \iddots & \ldots & CA^{T_2+1}B \\ CA^2B & \iddots & \iddots & \ldots & \vdots \\ \vdots & \iddots & \iddots & \ldots & \vdots \\ CA^{T_1}B & \iddots & \iddots & \ldots & CA^{T_1+T_2}B \end{bmatrix}}^{H:\ \text{hankel matrix}} = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{T_1} \end{bmatrix} \begin{bmatrix} B & AB & A^2B & \ldots & A^{T_2}B \end{bmatrix}$$

$$rank(H) = n$$

n: state-space dim

Ho-Kalman Algorithm

- Assume Markov parameters of the system are given:

$$D, CB, CAB, CA^2B, \ldots, CA^tB, \ldots$$

- Form the Hankel matrix of Markov parameters

$$\overbrace{\begin{bmatrix} CB & CAB & CA^2B & \ldots & CA^{T_2}B \\ CAB & CA^2B & \iddots & \ldots & CA^{T_2+1}B \\ CA^2B & \iddots & \iddots & \ldots & \vdots \\ \vdots & \iddots & \iddots & \ldots & \vdots \\ CA^{T_1}B & \iddots & \iddots & \ldots & CA^{T_1+T_2}B \end{bmatrix}}^{H:\ \text{hankel matrix}}$$

$$H^+ \qquad\qquad H^-$$

$$H^+ = OQ \Longrightarrow H^- = OAQ$$

Ho-Kalman Algorithm

- Assume Markov parameters of the system are given:

$$D, CB, CAB, CA^2B, \ldots, CA^tB, \ldots$$

- Form the Hankel matrix of Markov parameters

$$\overbrace{\begin{bmatrix} CB & CAB & CA^2B & \ldots & CA^{T_2}B \\ CAB & CA^2B & \iddots & \ldots & CA^{T_2+1}B \\ CA^2B & \iddots & \iddots & \ldots & \vdots \\ \vdots & \iddots & \iddots & \ldots & \vdots \\ CA^{T_1}B & \iddots & \iddots & \ldots & CA^{T_1+T_2}B \end{bmatrix}}^{H:\ \text{hankel matrix}}$$

H^+ $\qquad$ H^-

$$H^+ = OQ \Longrightarrow H^- = OAQ$$

$$L \doteq \mathrm{SVD}_n(H^+) = U\Sigma V^\intercal$$

$$O \doteq U\Sigma^{1/2},\ Q \doteq \Sigma^{1/2}V^\intercal$$

$$\bar{C} \doteq \text{first } m \text{ rows of } O$$

$$\bar{B} \doteq \text{first } p \text{ columns of } Q$$

$$\bar{A} \doteq O^\dagger H^- Q^\dagger$$

n: states
m: outputs
p: inputs

Ho-Kalman Algorithm

- Assume Markov parameters of the system are given:

$$D, CB, CAB, CA^2B, \ldots, CA^tB, \ldots$$

- Form the Hankel matrix of Markov parameters

$$\overbrace{\begin{bmatrix} CB & CAB & CA^2B & \ldots & CA^{T_2}B \\ CAB & CA^2B & \iddots & \ldots & CA^{T_2+1}B \\ CA^2B & \iddots & \iddots & \ldots & \vdots \\ \vdots & \iddots & \iddots & \ldots & \vdots \\ CA^{T_1}B & \iddots & \iddots & \ldots & CA^{T_1+T_2}B \end{bmatrix}}^{H:\ \text{hankel matrix}}$$

H^+ $\qquad$ H^-

$$H^+ = OQ \Longrightarrow H^- = OAQ$$

Hankel singular values

$$L \doteq \mathrm{SVD}_n(H^+) = U\boxed{\Sigma}V^\intercal$$

$$O \doteq U\Sigma^{1/2},\ Q \doteq \Sigma^{1/2}V^\intercal$$

$$\bar{C} \doteq \text{first } m \text{ rows of } O$$

$$\bar{B} \doteq \text{first } p \text{ columns of } Q$$

$$\bar{A} \doteq O^\dagger H^- Q^\dagger$$

Balanced realization

Ho-Kalman Algorithm

- **Estimated** Markov parameters:

$$G = [D, CB, CAB, \ldots, CA^T B]$$

$$\hat{G} = [\hat{G}_{-1}, \hat{G}_0, \hat{G}_1, \ldots, \hat{G}_T]$$

- Form the Hankel matrix:

$$\overbrace{\begin{bmatrix} \hat{G}_0 & \hat{G}_1 & \hat{G}_2 & \cdots & \hat{G}_{T_2} \\ \hat{G}_1 & \hat{G}_2 & \iddots & \cdots & \hat{G}_{T_2+1} \\ \hat{G}_2 & \iddots & \iddots & \cdots & \vdots \\ \vdots & \iddots & \iddots & \cdots & \vdots \\ \hat{G}_{T_1} & \iddots & \iddots & \cdots & \hat{G}_{T_1+T_2} \end{bmatrix}}^{\hat{H}}$$

$\hat{H}^+$ $\hat{H}^-$

$$\hat{L} \doteq \mathrm{SVD}_n(\hat{H}^+) = \hat{U}\hat{\Sigma}\hat{V}^\intercal$$

$$\hat{O} \doteq \hat{U}\hat{\Sigma}^{1/2}, \ \hat{Q} \doteq \hat{\Sigma}^{1/2}\hat{V}^\intercal$$

$$\hat{C} \doteq \text{first } m \text{ rows of } \hat{O}$$

$$\hat{B} \doteq \text{first } p \text{ columns of } \hat{Q}$$

$$\hat{A} \doteq \hat{O}^\dagger \hat{H}^- \hat{Q}^\dagger$$

Noise sensitivity of Ho-Kalman

- **Estimated** Markov parameters:

$$G = [D, CB, CAB, \ldots, CA^T B]$$

$$\hat{G} = [\hat{G}_{-1}, \hat{G}_0, \hat{G}_1, \ldots, \hat{G}_T]$$

- Given a bound on $||G - \hat{G}||$

how good are the other estimates?

Algorithm

$\hat{L} \doteq \mathrm{SVD}_n(\hat{H}^+) = \hat{U}\hat{\Sigma}\hat{V}^\intercal$

$\hat{O} \doteq \hat{U}\hat{\Sigma}^{1/2},\ \hat{Q} \doteq \hat{\Sigma}^{1/2}\hat{V}^\intercal$

$\hat{C} \doteq$ first m rows of $\hat{O}$

$\hat{B} \doteq$ first p columns of $\hat{Q}$

$\hat{A} \doteq \hat{O}^\dagger \hat{H}^- \hat{Q}^\dagger$

Lemma:

$$\|H - \hat{H}\| \leq \sqrt{\min\{T_1, T_2\}}\|G - \hat{G}\|$$

$$\|L - \hat{L}\| \leq 2\sqrt{\min\{T_1, T_2\}}\|G - \hat{G}\|$$

Theorem: Assume, $\|L - \hat{L}\| \leq \sigma_{min}(L)/2$. Then, there exists a unitary matrix *P* s.t.

$$\|\bar{C} - \hat{C}P\|_F \leq \sqrt{5n\|L - \hat{L}\|}$$

$$\|\bar{B} - P^\intercal \hat{B}\|_F \leq \sqrt{5n\|L - \hat{L}\|}$$

$$\|\bar{A} - P^\intercal \hat{A} P\|_F \leq \frac{14\sqrt{n}}{\sigma_{min}(L)}\left(2\|H^- - \hat{H}^-\| + \sqrt{\frac{\|L - \hat{L}\|}{\sigma_{min}(L)}}\|H^-\|\right)$$

Estimation of Markov parameters

Consider

$$x_{t+1} = Ax_t + Bu_t + w_t$$
$$y_t = Cx_t + Du_t + z_t$$

assume,

$$x_0 = 0,\ u_t \sim \mathcal{N}(0, \sigma_u^2 I_p),\ w_t \sim \mathcal{N}(0, \sigma_w^2 I_n),\ \text{and}\ z_t \sim \mathcal{N}(0, \sigma_z^2 I_m)$$

Then, cross-correlations of input and output give Markov parameters:

$$\mathbb{E}\left[\frac{y_t u_{t-k}^*}{\sigma_u^2}\right] = \begin{cases} D & \text{if} \quad k = 0, \\ CA^{k-1}B & \text{if} \quad k \geq 1. \end{cases}$$

Estimation of Markov parameters

Given input/output data ($\{u_t, y_t\}_{t=0}^{\bar{N}}$), from a process of the form:

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

$$y_t = Cx_t + Du_t + z_t$$

where

$$x_0 = 0,\ u_t \sim \mathcal{N}(0, \sigma_u^2 I_p),\ w_t \sim \mathcal{N}(0, \sigma_w^2 I_n),\ \text{and}\ z_t \sim \mathcal{N}(0, \sigma_z^2 I_m)$$

consider N subsequences of data of length T+1:

$$\begin{array}{cccccccccc}
x_0, & x_1, & \dots & x_T, & x_{T+1}, & \dots & x_{\bar{N}-T} & \dots & x_{\bar{N}} \\
u_0, & u_1, & \dots & u_T, & u_{T+1}, & \dots & u_{\bar{N}-T} & \dots & u_{\bar{N}} \\
 & & & & & & & & \\
y_0, & y_1, & \dots & y_T, & y_{T+1}, & \dots & y_{\bar{N}-T} & \dots & y_{\bar{N}}
\end{array}$$

Estimation of Markov parameters

Given input/output data ($\{u_t, y_t\}_{t=0}^{\bar{N}}$), from a process of the form:

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

$$y_t = Cx_t + Du_t + z_t$$

where

$$x_0 = 0,\ u_t \sim \mathcal{N}(0, \sigma_u^2 I_p),\ w_t \sim \mathcal{N}(0, \sigma_w^2 I_n),\ \text{and}\ z_t \sim \mathcal{N}(0, \sigma_z^2 I_m)$$

consider N subsequences of data of length T+1:

$$\begin{array}{cccccccccc}
x_0, & x_1, & \dots & x_T, & x_{T+1}, & \dots & x_{\bar{N}-T} & \dots & x_{\bar{N}} \\
u_0, & u_1, & \dots & u_T, & u_{T+1}, & \dots & u_{\bar{N}-T} & \dots & u_{\bar{N}} \\
 & & & & & & & & \\
y_0, & y_1, & \dots & y_T, & y_{T+1}, & \dots & y_{\bar{N}-T} & \dots & y_{\bar{N}}
\end{array}$$

Estimation of Markov parameters

Given input/output data ($\{u_t, y_t\}_{t=0}^{\bar{N}}$), from a process of the form:

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

$$y_t = Cx_t + Du_t + z_t$$

where

$$x_0 = 0,\ u_t \sim \mathcal{N}(0, \sigma_u^2 I_p),\ w_t \sim \mathcal{N}(0, \sigma_w^2 I_n),\ \text{and}\ z_t \sim \mathcal{N}(0, \sigma_z^2 I_m)$$

consider N subsequences of data of length T+1:

$$\begin{array}{cccccccccc}
x_0, & x_1, & \dots & x_T, & x_{T+1}, & \dots & x_{\bar{N}-T} & \dots & x_{\bar{N}} \\
u_0, & u_1, & \dots & u_T, & u_{T+1}, & \dots & u_{\bar{N}-T} & \dots & u_{\bar{N}} \\
 & & & & & & & & \\
y_0, & y_1, & \dots & y_T, & y_{T+1}, & \dots & y_{\bar{N}-T} & \dots & y_{\bar{N}}
\end{array}$$

Estimation of Markov parameters

$$x_0, \quad x_1, \quad \dots \quad x_T, \quad x_{T+1}, \quad \dots \quad x_{\bar{N}-T} \quad \dots \quad x_{\bar{N}}$$

$$u_0, \quad u_1, \quad \dots \quad u_T, \quad u_{T+1}, \quad \dots \quad u_{\bar{N}-T} \quad \dots \quad u_{\bar{N}}$$

$$y_0, \quad y_1, \quad \dots \quad y_T, \quad y_{T+1}, \quad \dots \quad y_{\bar{N}-T} \quad \dots \quad y_{\bar{N}}$$

for all $t \in \{T+1, \dots, \bar{N}\}$

$$y_t = CA^T x_{t-T} + Du_t + \sum_{i=1}^{T} CA^{i-1} B u_{t-i} + \sum_{i=1}^{T} CA^{i-1} w_{t-i} + z_t$$

$$= G\bar{\mathbf{u}}_t + F\bar{\mathbf{w}}_t + z_t + e_t$$

Recall:

$G = [D, CB, CAB, \dots, CA^T B]$

Estimation of Markov parameters

$$x_0, \quad x_1, \quad \dots \quad x_T, \quad x_{T+1}, \quad \dots \quad x_{\bar{N}-T} \quad \dots \quad x_{\bar{N}}$$

$$u_0, \quad u_1, \quad \dots \quad u_T, \quad u_{T+1}, \quad \dots \quad u_{\bar{N}-T} \quad \dots \quad u_{\bar{N}}$$

$$y_0, \quad y_1, \quad \dots \quad y_T, \quad y_{T+1}, \quad \dots \quad y_{\bar{N}-T} \quad \dots \quad y_{\bar{N}}$$

for all $t \in \{T+1, \dots, \bar{N}\}$

$$y_t = CA^T x_{t-T} + Du_t + \sum_{i=1}^{T} CA^{i-1} B u_{t-i} + \sum_{i=1}^{T} CA^{i-1} w_{t-i} + z_t$$

$$= G\bar{\mathbf{u}}_t + \underbrace{F\bar{\mathbf{w}}_t + z_t}_{\text{Noise terms}} + e_t$$

Effect of x_{t-T} (characterize statistics and treat as noise)

Recall:

$$G = [D, CB, CAB, \dots, CA^T B]$$

Estimation of Markov parameters

$$x_0, \quad x_1, \quad \dots \quad x_T, \quad x_{T+1}, \quad \dots \quad x_{\bar{N}-T} \quad \dots \quad x_{\bar{N}}$$
$$u_0, \quad u_1, \quad \dots \quad u_T, \quad u_{T+1}, \quad \dots \quad u_{\bar{N}-T} \quad \dots \quad u_{\bar{N}}$$
$$y_0, \quad y_1, \quad \dots \quad y_T, \quad y_{T+1}, \quad \dots \quad y_{\bar{N}-T} \quad \dots \quad y_{\bar{N}}$$

for all $t \in \{T+1, \dots, \bar{N}\}$

$$y_t = CA^T x_{t-T} + Du_t + \sum_{i=1}^{T} CA^{i-1} B u_{t-i} + \sum_{i=1}^{T} CA^{i-1} w_{t-i} + z_t$$
$$= G\bar{\mathbf{u}}_t + \underbrace{F\bar{\mathbf{w}}_t + z_t}_{\text{Noise terms}} + e_t$$

Effect of x_{t-T} (characterize statistics and treat as noise)

Use least squares to estimate *G:*

$$\hat{G} = \arg\min_X \sum_{t=T}^{\bar{N}} ||y_t - X\bar{\mathbf{u}}_t||_2^2$$

How good is this estimate?

Estimation of Markov parameters

$$y_t = CA^T x_{t-T} + Du_t + \sum_{i=1}^{T} CA^{i-1}Bu_{t-i} + \sum_{i=1}^{T} CA^{i-1}w_{t-i} + z_t$$

$$= G\bar{\mathbf{u}}_t + \underbrace{F\bar{\mathbf{w}}_t + z_t}_{\text{Noise terms}} + e_t$$

Effect of x_{t-T} (characterize statistics and treat as noise)

Use least squares to estimate *G:*

$$\hat{G} = \arg\min_X \sum_{t=T}^{\bar{N}} ||y_t - X\bar{\mathbf{u}}_t||_2^2$$

How good is this estimate?

Concatenating further through $\bar{N}$ to obtain matrix form*:*

$$Y = GU + FW + Z + E$$

$$\hat{G} = YU^*(UU^*)^{-1} = G + (FW + Z + E)U^*(UU^*)^{-1}$$

$$||\hat{G} - G|| \leq (||F||\underbrace{||WU^*||}_{} + \underbrace{||ZU^*||}_{} + \underbrace{||EU^*||}_{})||\underbrace{(UU^*)^{-1}||}_{}$$

concentration inequalities for circulant matrices (Krahmer et al. 14)

martingale argument

Sample complexity

Theorem:

Given input/output data ($\{u_t, y_t\}_{t=0}^{\bar{N}}$), from a process of the form:

$$x_{t+1} = Ax_t + Bu_t + w_t$$
$$y_t = Cx_t + Du_t + z_t$$

where A is stable and

$$x_0 = 0,\ u_t \sim \mathcal{N}(0, \sigma_u^2 I_p),\ w_t \sim \mathcal{N}(0, \sigma_w^2 I_n),\ \text{and}\ z_t \sim \mathcal{N}(0, \sigma_z^2 I_m)$$

let $N \geq N_0 = cTq\log^2(2Tq)\log^2(2Nq)$ where $q = n + m + p$

and take $\bar{N} = N + T$

then with **very high** probability, we have

$$\|G - \hat{G}\| \leq \frac{\sigma_z + \sigma_e + \sigma_w\|F\|}{\sigma_u}\sqrt{\frac{N_0}{N}}$$

*Recall the error of least squares: $\hat{G} - G = (FW + Z + E)U^*(UU^*)^{-1}$

Combining it all…

Given any δ, ε, we can find a "tight" $\bar{N}$ such that if we have input/output data of length $\bar{N}$, with probability (1-δ), we can estimate the system matrices (of balanced realization) by accuracy at most ε.

Similarly, given input/output data of length $\bar{N}$, and any δ, we can give a bound ε on the accuracy of the system matrix estimates that is valid with probability (1- δ).

Several applications/extensions: estimates of H-infinity norm error, system order, etc.

Numerical examples

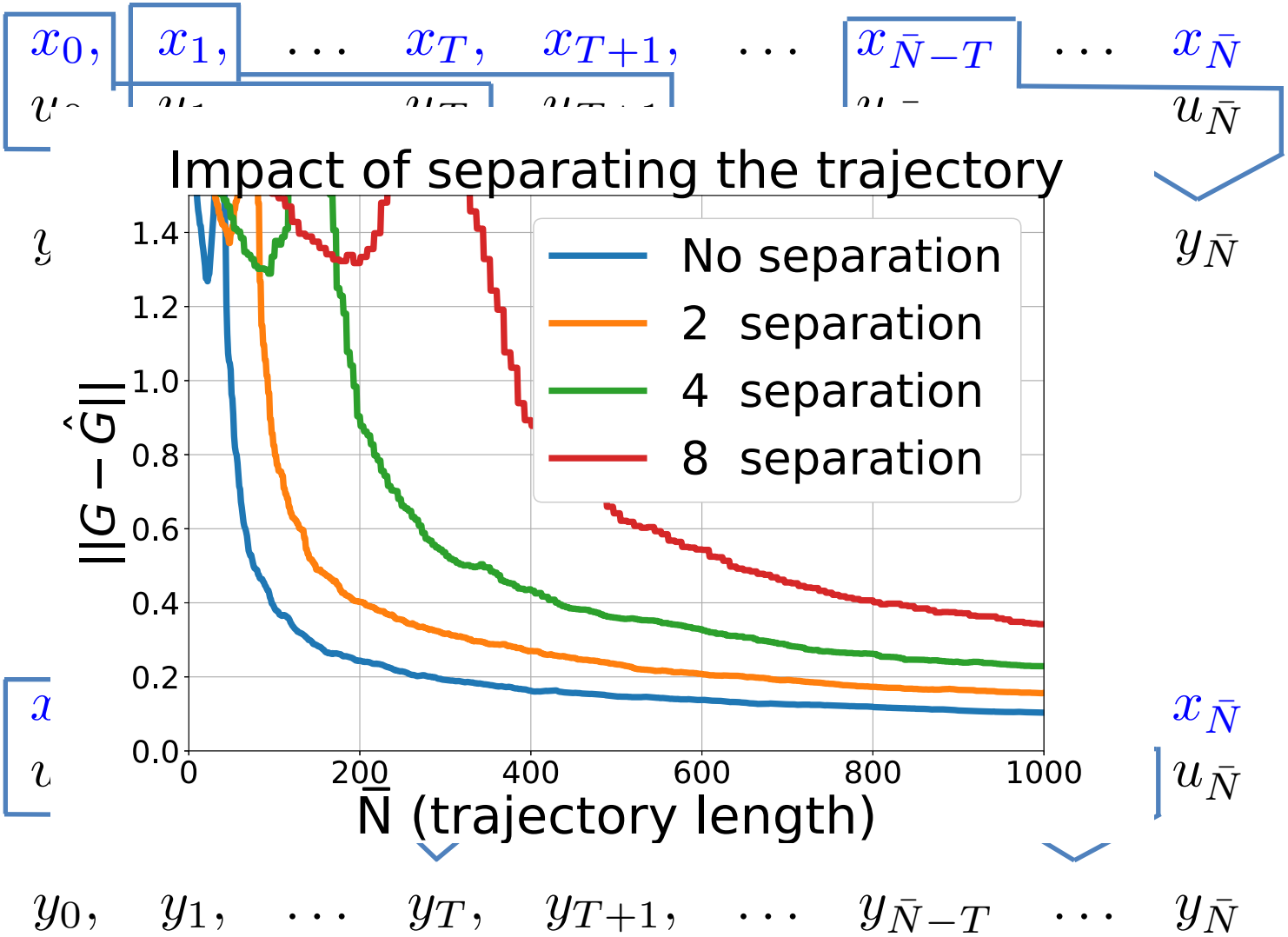
$$
\begin{array}{cccccccccc}
x_0, & x_1, & \dots & x_T, & x_{T+1}, & \dots & x_{\bar{N}-T} & \dots & x_{\bar{N}} \\
u_0, & u_1, & \dots & u_T, & u_{T+1}, & \dots & u_{\bar{N}-T} & \dots & u_{\bar{N}} \\
y_0, & y_1, & \dots & y_T, & y_{T+1}, & \dots & y_{\bar{N}-T} & \dots & y_{\bar{N}}
\end{array}
$$

Statistically easier to analyze (but less "efficient") alternative:
a variant of the i.i.d. trajectory view point in [Oymak 19]

Separation K

$$
\begin{array}{cccccccccc}
x_0, & x_1, & \dots & x_T, & x_{T+1}, & \dots & x_{\bar{N}-T} & \dots & x_{\bar{N}} \\
u_0, & u_1, & \dots & u_T, & u_{T+1}, & \dots & u_{\bar{N}-T} & \dots & u_{\bar{N}} \\
y_0, & y_1, & \dots & y_T, & y_{T+1}, & \dots & y_{\bar{N}-T} & \dots & y_{\bar{N}}
\end{array}
$$

Numerical examples

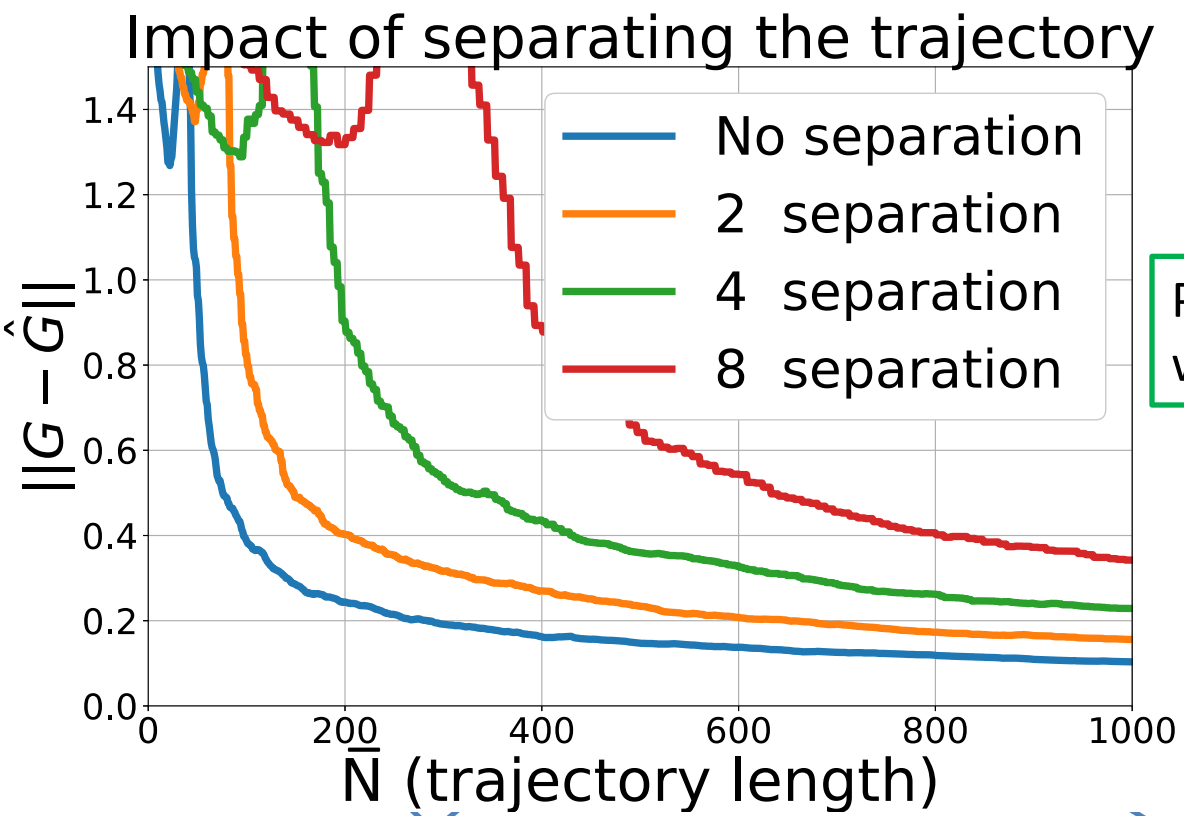


Numerical examples

$x_0, \quad x_1, \quad \dots \quad x_T, \quad x_{T+1}, \quad \dots \quad x_{\bar{N}-T} \quad \dots \quad x_{\bar{N}}$

$u_{\bar{N}}$

$y_{\bar{N}}$



Practically better use of data with tight statistical bounds!

$x_{\bar{N}}$

$u_{\bar{N}}$

$y_0, \quad y_1, \quad \dots \quad y_T, \quad y_{T+1}, \quad \dots \quad y_{\bar{N}-T} \quad \dots \quad y_{\bar{N}}$

Plan

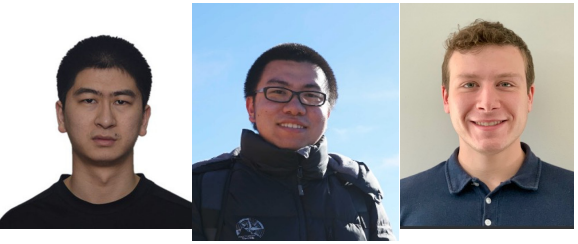
Focus on a well-known system identification algorithms:

- **Can we achieve similar sample complexity results for system identification algorithms where the data is highly correlated?**

- **Part 1: State-space models – Ho-Kalman Algorithm**

Joint work with Samet Oymak, UC Riverside
ACC'19, TAC'22

- **Part 2: Autoregressive models – Ordinary least squares**



Joint work with Zhe Du, Zexiang Liu, Jack Weitze, Michigan
CDC'22

A different representation

- **AutoRegressive eXogenous (ARX) models**

$$y_t = \sum_{i=1}^{n_\alpha} \alpha_i y_{t-i} + \sum_{i=1}^{n_\beta} \beta_i u_{t-i} + \eta_{t-1}$$

- output y_t, input u_t, noise $\eta_t \sim \mathcal{N}(0, \sigma_\eta^2)$ at time t
- n_α and n_β are model orders

A different representation

- **AutoRegressive eXogenous (ARX) models**

$$y_t = \sum_{i=1}^{n_\alpha} \alpha_i y_{t-i} + \sum_{i=1}^{n_\beta} \beta_i u_{t-i} + \eta_{t-1}$$

- **Problem Step**
  - **Data:** a single trajectory $\{u_t, y_t\}_{t=0}^{N}$ where $u_t \sim \mathcal{N}(0, \sigma_u^2)$.
  - **Goal:** estimate n_α, n_β and $\{\alpha_i\}_{i=1}^{n_\alpha}, \{\beta_i\}_{i=1}^{n_\beta}$.

- **Over-parameterization:** pick $\bar{n}_\alpha \geqslant n_\alpha$, $\bar{n}_\beta \geqslant n_\beta$ and fit the data $\{u_t, y_t\}_{t=0}^{N}$ with

$$y_t = \sum_{i=1}^{\bar{n}_\alpha} \hat{\alpha}_i y_{t-i} + \sum_{i=1}^{n_\beta} \hat{\beta}_i u_{t-i}$$

A different rep

- **AutoRegressive eXogenous (ARX) models**

$$y_t = \sum_{i=1}^{n_\alpha} \alpha_i y_{t-i} + \sum_{i=1}^{n_\beta}$$

- **Problem Step**
  - **Data:** a single trajectory $\{u_t, y_t\}_{t=0}^{N}$ w
  - **Goal:** estimate n_α, n_β and $\{\alpha_i\}_{i=1}^{n_\alpha}$

Challenges

- Statistical dependency among $\{u_t, y_t\}_{t=0}^{N}$
- Over-parameterization enlarges the model capacity but also introduces model ambiguity!
  - Consider $y_t = 0.5y_{t-1} + u_{t-1}$ with $n_\alpha = n_\beta = 1$.
  - Then,

$$\begin{gathered}(y_t = 0.5y_{t-1} + u_{t-1}) \\ + \\ c \times (y_{t-1} = 0.5y_{t-2} + u_{t-2}) \\ \Downarrow \\ y_t = (0.5 - c)y_{t-1} + 0.5cy_{t-2} + u_{t-1} + cu_{t-2}\end{gathered}$$

  - Hence, when $\bar{n}_\alpha = \bar{n}_\beta = 2$, we have <u>infinitely</u> many equivalent models!

- **Over-parameterization:** pick $\bar{n}_\alpha \geqslant n_\alpha$, $\bar{n}_\beta \geqslant n_\beta$ and fit the data $\{u_t, y_t\}_{t=0}^{N}$ with

$$y_t = \sum_{i=1}^{\bar{n}_\alpha} \hat{\alpha}_i y_{t-i} + \sum_{i=1}^{n_\beta} \hat{\beta}_i u_{t-i}$$

Earlier work

Recall …

- **Over-parameterization:** pick orders $\bar{n}_\alpha \geqslant n_\alpha,\ \bar{n}_\beta \geqslant n_\beta$ and fit the data $\{u_t, y_t\}_{t=0}^{N}$ with

$$y_t = \sum_{i=1}^{\bar{n}_\alpha} \hat{\alpha}_i y_{t-i} + \sum_{i=1}^{\bar{n}_\beta} \hat{\beta}_i u_{t-i}$$

vectorize

- **Over-parameterization:** pick orders $\bar{n}_\alpha \geqslant n_\alpha,\ \bar{n}_\beta \geqslant n_\beta$ and fit the data $\{\mathbf{z}_t, y_t\}_{t=0}^{T}$ with

$$y_t = \hat{\boldsymbol{\theta}}^\intercal \mathbf{z}_t$$

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &:= [\hat{\alpha}_1, \ldots, \hat{\alpha}_{n_\alpha}, \hat{\alpha}_{n_\alpha+1}, \ldots, \hat{\alpha}_{\bar{n}_\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_{n_\beta}, \hat{\beta}_{n_\beta+1}, \ldots, \hat{\beta}_{\bar{n}_\beta}]^\intercal \\
\mathbf{z}_t &:= [y_{t-1}, \ldots, y_{t-n_\alpha}, y_{t-n_\alpha+1}, \ldots, y_{t-\bar{n}_\alpha}, u_{t-1}, \ldots, u_{t-n_\beta}, u_{t-n_\beta+1}, \ldots, u_{t-\bar{n}_\beta}]^\intercal \\
\boldsymbol{\theta} &:= [\alpha_1, \ldots, \alpha_{n_\alpha}, 0, \ldots, 0, \beta_1, \ldots, \beta_{n_\beta}, 0, \ldots, 0]^\intercal
\end{aligned}$$

- Estimation error $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|$

exact-parameterization

over-parameterization

If estimation error → 0, then

- exact-param → true parameters
- over-param → 0.

Earlier work

Recall ...

- **Over-parameterization:** pick orders $\bar{n}_\alpha \geqslant n_\alpha,\ \bar{n}_\beta \geqslant n_\beta$ and fit the data $\{u_t, y_t\}_{t=0}^{N}$ with

$$y_t = \sum_{i=1}^{\bar{n}_\alpha} \hat{\alpha}_i y_{t-i} + \sum_{i=1}^{\bar{n}_\beta} \hat{\beta}_i u_{t-i}$$

vectorize

- **Over-parameterization:** pick orders $\bar{n}_\alpha \geqslant n_\alpha,\ \bar{n}_\beta \geqslant n_\beta$ and fit the data $\{\mathbf{z}_t, y_t\}_{t=0}^{T}$ with

$$y_t = \hat{\theta}^\intercal \mathbf{z}_t$$

$$\begin{aligned}
\hat{\theta} &:= [\hat{\alpha}_1, \ \ldots, \ \hat{\alpha}_{n_\alpha}, & \hat{\alpha}_{n_\alpha+1}, \ \ldots, \ \hat{\alpha}_{\bar{n}_\alpha}, & \hat{\beta}_1, \ \ldots, \ \hat{\beta}_{n_\beta}, & \hat{\beta}_{n_\beta+1}, \ \ldots, \ \hat{\beta}_{\bar{n}_\beta}]^\intercal \\
\mathbf{z}_t &:= [y_{t-1}, \ldots, y_{t-n_\alpha}, & y_{t-n_\alpha+1}, \ldots, y_{t-\bar{n}_\alpha}, & u_{t-1}, \ldots, u_{t-n_\beta}, & u_{t-n_\beta+1}, \ldots, u_{t-\bar{n}_\beta}]^\intercal \\
\theta &:= [\alpha_1, \ \ldots, \ \alpha_{n_\alpha}, & 0, \ \ldots, \ 0, & \beta_1, \ \ldots, \ \beta_{n_\beta}, & 0, \ \ldots, \ 0]^\intercal
\end{aligned}$$

- Estimation error $\|\hat{\theta} - \theta\|$

exact-parameterization

over-parameterization

	Previous work [Jones and Dahleh, 2022], [Ljung and Wahlberg, 1992]	Our work
Method	**RLS**: $\hat{\theta} \leftarrow \arg\min_{\hat{\theta}} \sum_t (y_t - \hat{\theta}^\intercal \mathbf{z}_t)^2 + \lambda \|\hat{\theta}\|^2$	**OLS**: $\hat{\theta} \leftarrow \arg\min_{\hat{\theta}} \sum_t (y_t - \hat{\theta}^\intercal \mathbf{z}_t)^2$
Guarantees	$\|\hat{\theta} - \theta\| \leqslant \mathcal{O}(\frac{\sqrt{\log(N/\lambda)+\lambda}}{\sqrt{N}})$	$\|\hat{\theta} - \theta\| \leqslant \mathcal{O}(\frac{\sqrt{\log(N)}}{\sqrt{N}})$
Pros/Cons/Comments	Hyper-parameter λ can be neither too large nor small, tuning is needed. Manifests the "oracle property" [Candès, 2006]	Regularizer (hyper-parameter) free — "self-regularization" The first finite sample result for OLS on unknown(and known)-order ARX models

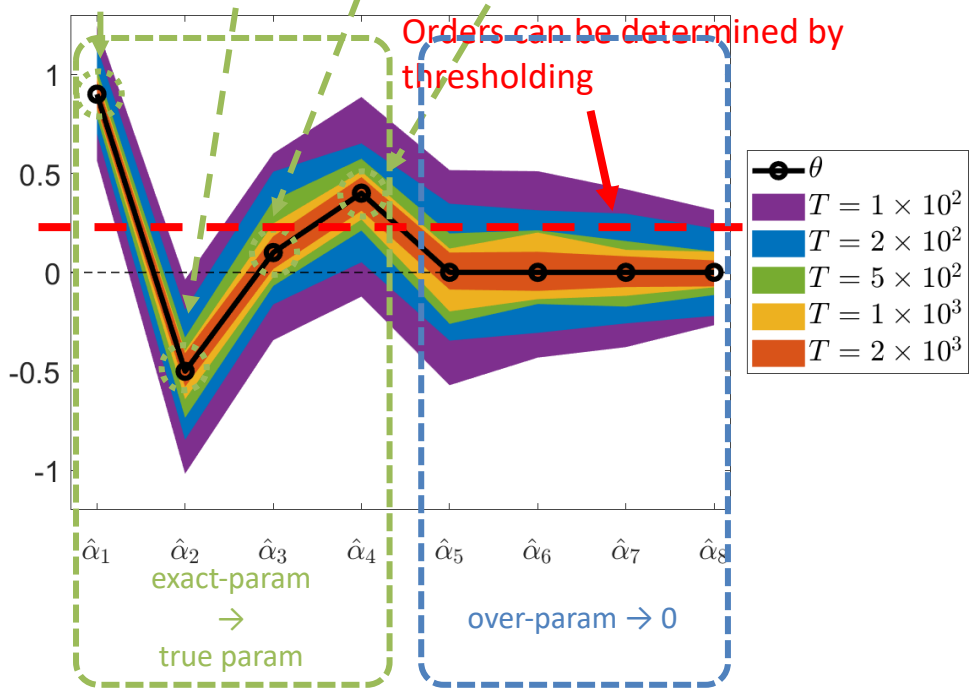
Earlier work

A Teaser Example

- Consider learning the following ARX $(n_\alpha = 4)$

$$y_t = 0.9y_{t-1} - 0.5y_{t-2} + 0.1y_{t-3} + 0.4y_{t-4} + u_{t-1} + \eta_{t-1}$$

- $u_t \sim \mathcal{N}(0,1)$, $\eta_t \sim \mathcal{N}(0,1)$
- Consider OLS with $\bar{n}_\alpha = 8$. Plot gives the "shape" of $\hat{\theta}$ (only the α part)



rders $\bar{n}_\alpha \geqslant n_\alpha$, $\bar{n}_\beta \geqslant n_\beta$ and fit the data

$y_t = \hat{\theta}^\intercal \mathbf{z}_t$

$\ldots, \hat{\alpha}_{\bar{n}_\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_{n_\beta}, \hat{\beta}_{n_\beta+1}, \ldots, \hat{\beta}_{\bar{n}_\beta}]^\intercal$

$1, \ldots, y_{t-\bar{n}_\alpha}, u_{t-1}, \ldots, u_{t-n_\beta}, u_{t-n_\beta+1}, \ldots, u_{t-\bar{n}_\beta}]^\intercal$

$., 0, \beta_1, \ldots, \beta_{n_\beta}, 0, \ldots, 0]^\intercal$

zation

over-parameterization

Our work
OLS: $\hat{\theta} \leftarrow \arg\min_{\hat{\theta}} \sum_t (y_t - \hat{\theta}^\intercal \mathbf{z}_t)^2$
$\lVert \hat{\theta} - \theta \rVert \leqslant \mathcal{O}(\frac{\sqrt{\log(N)}}{\sqrt{N}})$
Regularizer (hyper-parameter) free — "self-regularization" The first finite sample result for OLS on unknown(and known)-order ARX models

Sample complexity

Theorem: *Let* $q(z) := z^{n_\alpha} - \sum_{i=1}^{n_\alpha} \alpha_i z^{n_\alpha - i}$ *and* $\rho := \max_{z:q(z)=0} |z|$. *Suppose* $\rho < 1$, $\bar{n}_\alpha \geqslant n_\alpha$, $\bar{n}_\beta \geqslant n_\beta$ *, then w.p.* $1-\delta$ *, the OLS estimator* $\hat{\theta}$ *satisfies*

$$\|\hat{\theta} - \theta\| \leqslant \mathcal{O}\left(\frac{\sqrt{\bar{n}_\alpha + \bar{n}_\beta}}{1-\rho} \cdot \frac{\sigma_\eta}{\sigma_u} \cdot \sqrt{\frac{\log(N)}{N}} \log\left(\frac{N}{\delta}\right) \right).$$

Sample complexity

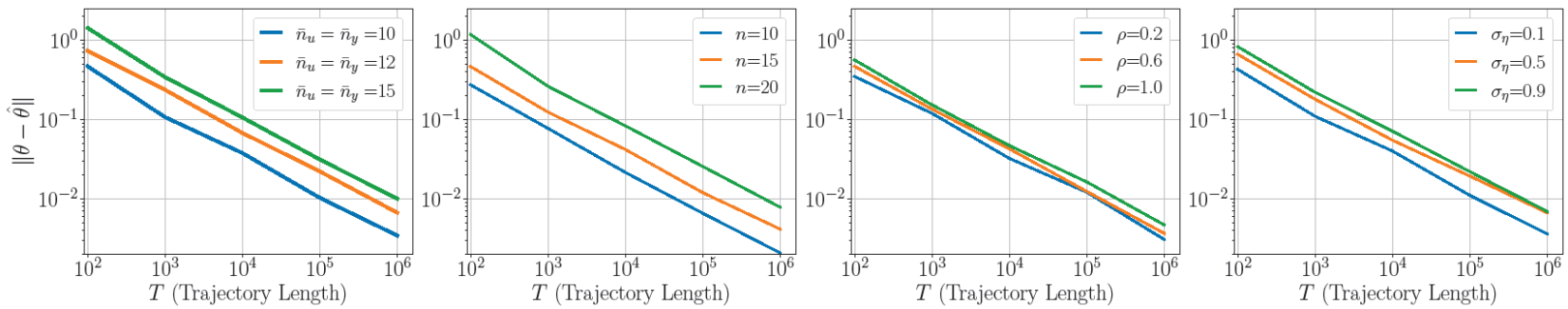
Theorem: *Let* $q(z) := z^{n_\alpha} - \sum_{i=1}^{n_\alpha} \alpha_i z^{n_\alpha - i}$ *and* $\rho := \max_{z:q(z)=0} |z|$. *Suppose* $\rho < 1$, $\bar{n}_\alpha \geqslant n_\alpha$, $\bar{n}_\beta \geqslant n_\beta$ *, then w.p.* $1-\delta$ *, the OLS estimator* $\hat{\theta}$ *satisfies*

$$\|\hat{\theta} - \theta\| \leqslant \mathcal{O}\left(\frac{\sqrt{\bar{n}_\alpha + \bar{n}_\beta}}{1-\rho} \cdot \frac{\sigma_\eta}{\sigma_u} \cdot \sqrt{\frac{\log(N)}{N}} \log\left(\frac{N}{\delta}\right) \right).$$

- ρ — stability of the model; if $\rho > 1$, then $y_t \to \infty$ as $t \to \infty$.

Numerical results

- True model (unless otherwise specified): $n_\alpha = n_\beta =: n = 10$, $\rho = 0.85$, $\sigma_u = 1$, $\sigma_\eta = 1$.



$$\|\hat{\theta} - \theta\| \leqslant \mathcal{O}\left(\frac{\sqrt{\bar{n}_\alpha + \bar{n}_\beta}}{1-\rho} \cdot \frac{\sigma_\eta}{\sigma_u} \cdot \sqrt{\frac{\log(N)}{N}} \log\left(\frac{N}{\delta}\right) \right).$$

- ρ — stability of the model; if $\rho > 1$, then $y_t \to \infty$ as $t \to \infty$.
- $\frac{\sigma_\eta}{\sigma_u}$ — noise-to-signal ratio.
- The bound is almost optimal (in terms of the order dependency on $\bar{n}_\alpha, \bar{n}_\beta, N$) compared with min-max lower bound [Simchowitz et al., 2018].

More recent results (incomplete list)

	Type	Rollouts	Rates	Burn-in time
Oymak & Ozay (part 1)	MIMO SS stable	Single	O(N^-1/2)	pT
Sun, Oymak, Fazel 2020	MIMO SS (un)stable	Multiple	O(N^-1/2)	pn
Zheng & Li 2020	MIMO SS (un)stable	Multiple	O(N^-1/2)	mT+q
Fattahi 2021	MIMO SS stable	Single	O(N^-1/4)	polylog(pT)
Du, Liu, Weitze, Ozay (part 2)	SISO ARX (un)stable	Single/Multiple	O(N^-1/2)	n^2

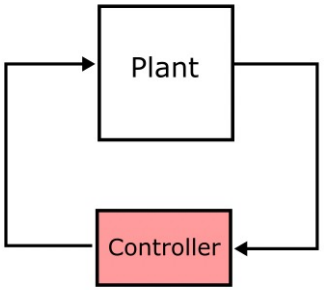
N: samples, n: system order, T: FIR order
m: # of inputs, p: # of outputs, q: #of inputs+outputs+states

Summary & Conclusions

System identification:

- We can learn from dynamical data as efficiently as we can learn from static data
- Downstream tasks:
  - Certainty equivalent control:

  regret guarantees (ACC'22)



- Extensions:
  - Bilinear system identification (CDC'22)
  - Markov jump linear system identification (ACC'22)
- Current directions:
  - Nonlinear system identification via liftings