Sample complexity of ARX identification and its implications for learning with parameter redundancy

Zexiang Liu*, Zhe Du*, Jack Weitze, Necmiye Ozay

EECS Department, University of Michigan

Oct. 25th, 2022 @ ELLIIT Focus Period

Autoregressive with External Input (ARX) Model

An ARX model is in form of

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

with output y_t,

input u_t,

noise $\eta_t \sim \mathcal{N}\left(0, \sigma_\eta^2\right)$.

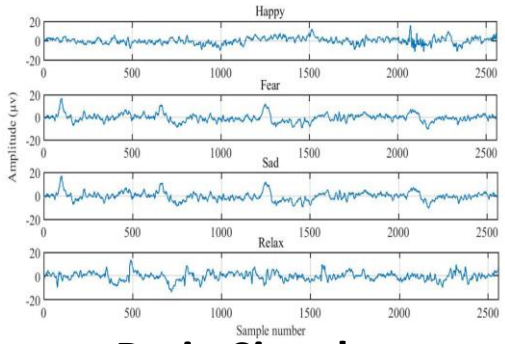
Orders: $n_\alpha,\ n_\beta$

Data: Trajectories $\{u_t, y_t\}_{t=0}^{T}$

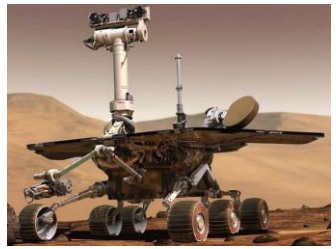
Goal: estimate the model parameters α_i, β_i

A simple model used many places

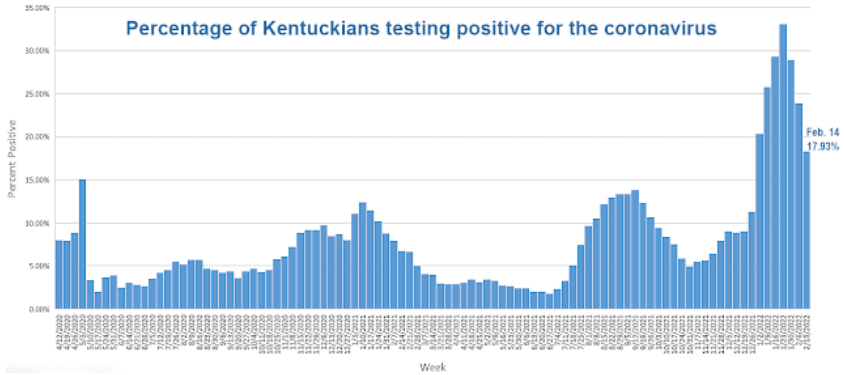
Price of Bitcoin



Brain Signals



Man-made Control Systems



Infection Rates

Is a linear model a strong assumption?

A nonlinear model

$$y_{t+1} = f(y_t, u_t)$$

Collect data $y_0, u_0, y_1, u_1, \dots, u_k, y_k, \dots$

Fit an ARX model

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

Recent progress in Koopman theory shows that many nonlinear systems can be approximated by an ARX-type model very well as long as n_α and n_β are large enough.

-> delay coordinates are “universal” class of observables [1]

[1] Brunton, Steven L., et al. "Modern Koopman theory for dynamical systems." *arXiv preprint arXiv:2102.12086* (2021).

Is a linear model a strong assumption?

A nonlinear model

$$y_{t+1} = \sin(y_t)$$

Collect data $y_0, y_1, \ldots, y_T$

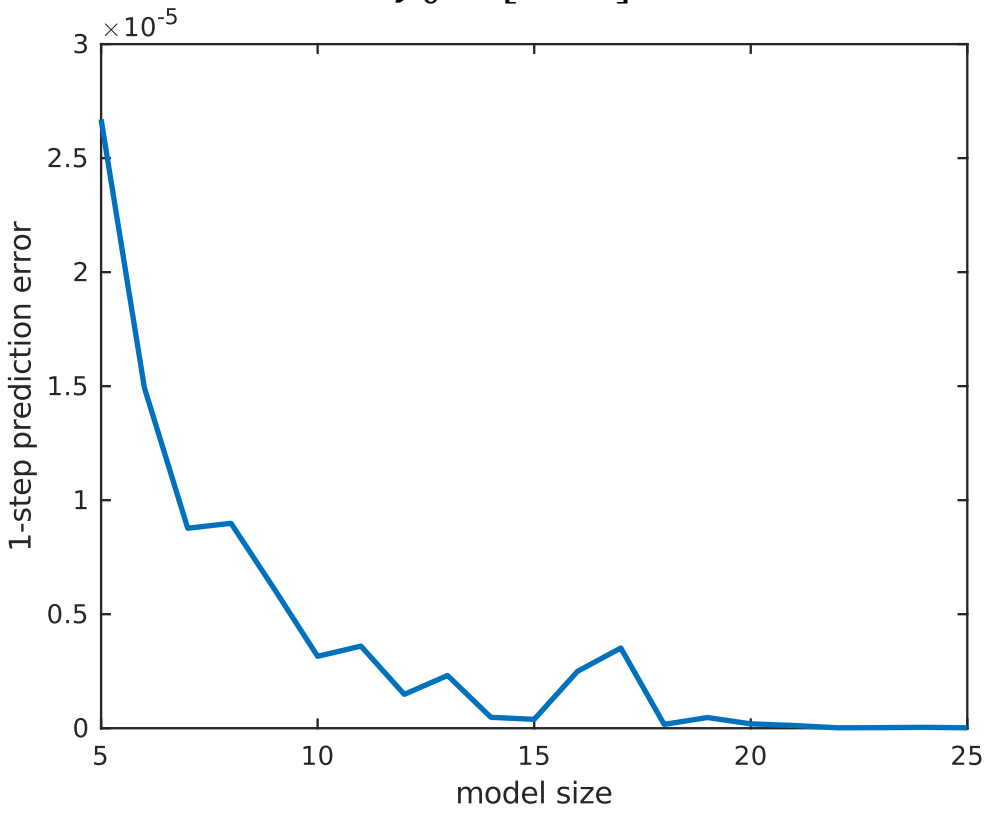
Fit an ARX model with order n

$$y_t = \sum_{i=1}^{n} \alpha_i y_{t-i} \,.$$

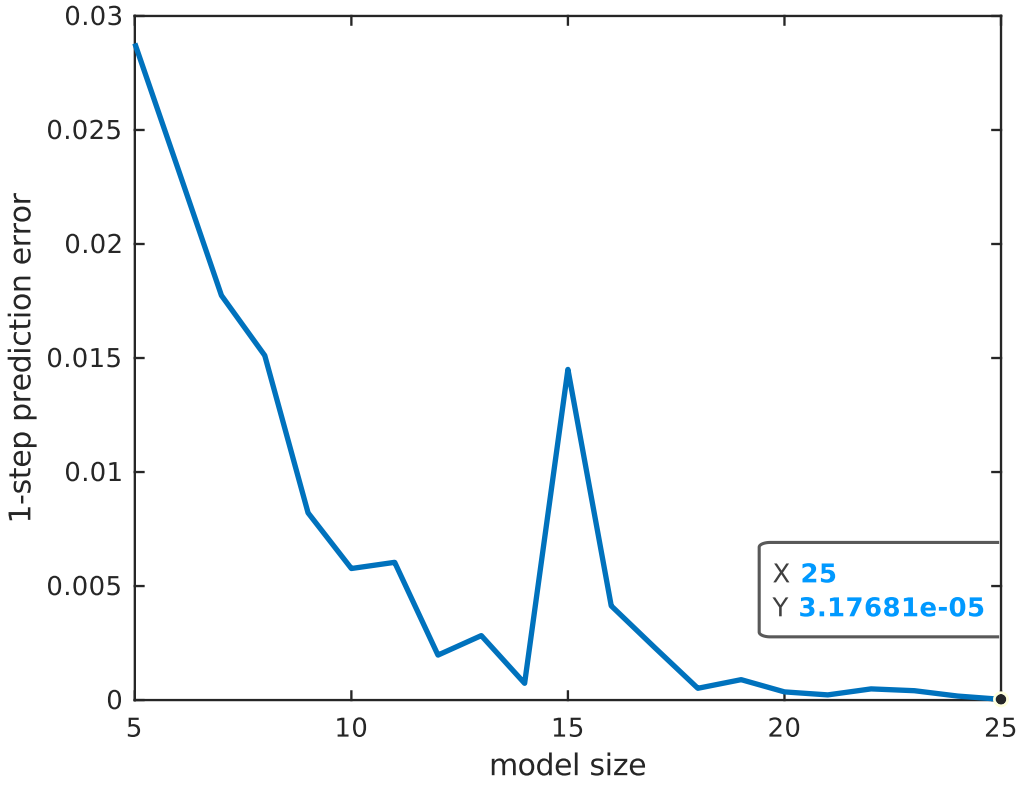
Test with one-step prediction averaged over 5000 samples

Is a linear model a strong assumption?

$y_0 \in [-1,1]$

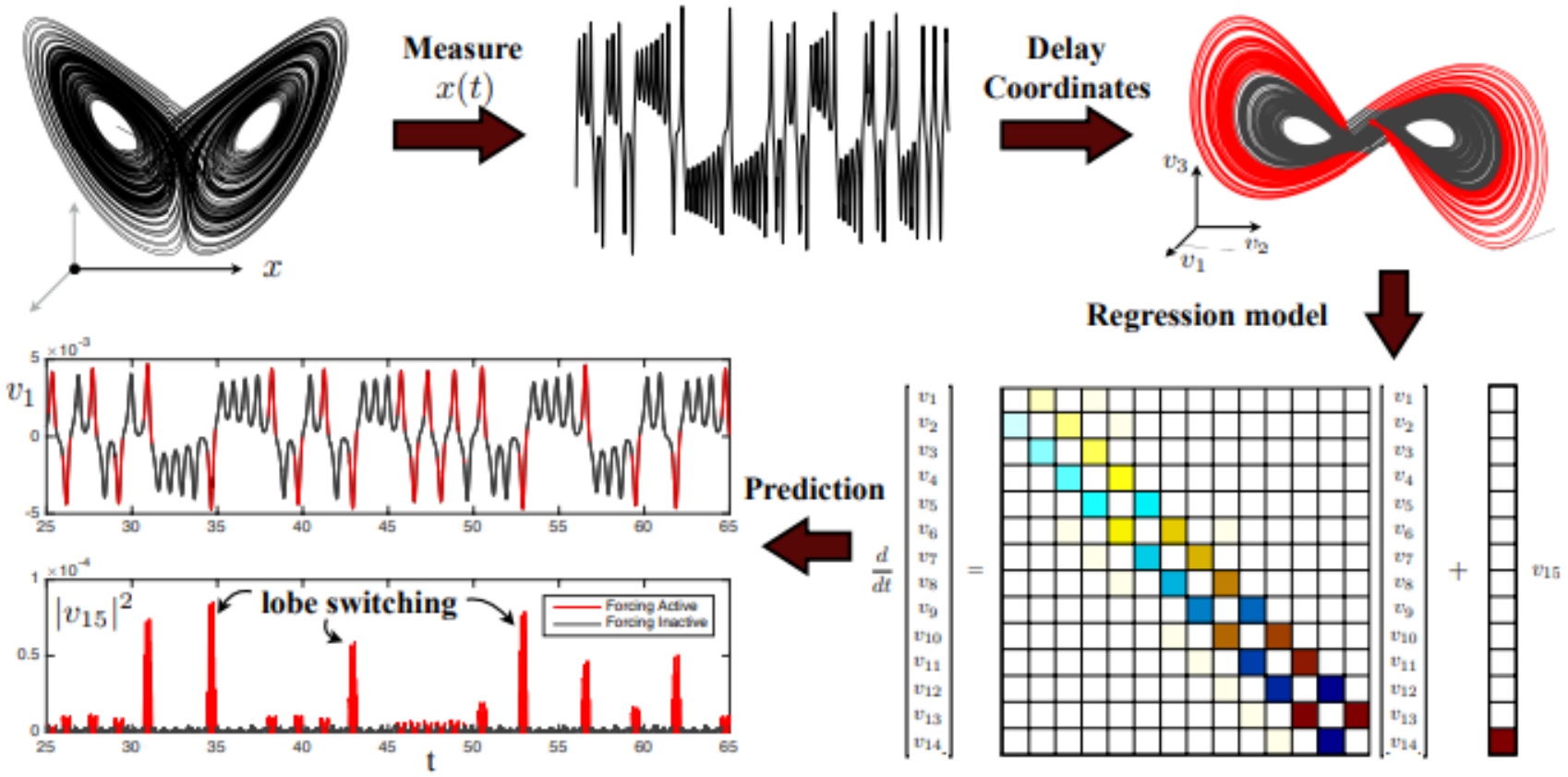


$y_0 \in [-5,5]$



Is a linear model a strong assumption?

Brunton, Steven L., et al. "Modern Koopman theory for dynamical systems." *arXiv preprint arXiv:2102.12086* (2021).



A linear model can approximately the 3D curve very well except at the “switching point”!

ARX Identification Setup

- True ARX Model:

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

Assume that the orders n_α and n_β are **unknown**.

- Hypothesis class:

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

for some $\bar{n}_\alpha$ and $\bar{n}_\beta$ greater than the true orders n_α and n_β

-> Overparameterization (learn more parameters than required).

- Goal: Learn $\theta = (\alpha_1, \dots, \alpha_{\bar{n}_\alpha}, \beta_1, \dots, \beta_{\bar{n}_\beta})$ from data.

What are the ground truth parameters?

Goal: Learn $\theta = (\alpha_1, \dots, \alpha_{\bar{n}_\alpha}, \beta_1, \dots, \beta_{\bar{n}_\beta})$ from data.

Intuitively speaking, the ground truth is the parameters in the true model + zeros. That is,

$$\theta_0 = \left(\alpha_1, \dots, \alpha_{n_\alpha}, 0, \dots, 0, \beta_1, \dots, \beta_{n_\beta}, 0, \dots, 0\right).$$

What are the ground truth parameters?

Say the ground-truth ARX model is

$$y_t = 0.5y_{t-1} + u_{t-1} + \eta_{t-1} \qquad (1)$$

Its order is $\mathrm{n}_\alpha = n_\beta = 1$.

Note that $y_{t-1} = 0.5y_{t-2} + u_{t-2} + \eta_{t-2}$

$$0 = -y_{t-1} + 0.5y_{t-2} + u_{t-2} + \eta_{t-2} \qquad (2)$$

Then, (1) +c(2) for any constant c gives another ground-truth model!

$$y_t = (0.5 - c)y_{t-1} + 0.5cy_{t-2} + u_{t-1} + cu_{t-2} + \eta_{t-1} + c\eta_{t-2} \qquad (3)$$

(3) is equivalent to (1) but with orders $\mathrm{n}_\alpha = n_\beta = 2$.

What are the ground truth parameters?

The ground-truth ARX model is

$$y_t = 0.5y_{t-1} + u_{t-1} + \eta_{t-1} \qquad (1)$$

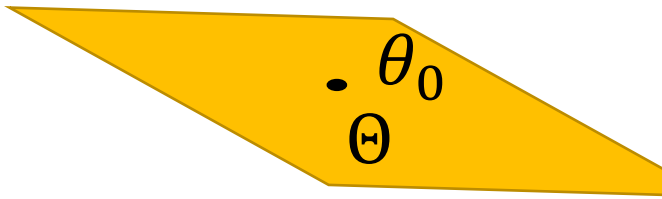
If we select the hypothesis class with $\bar{n}_\alpha = \bar{n}_\beta = 2$:

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

The unknown parameters are $\theta = (\alpha_1, \alpha_2, \beta_1, \beta_2)$.

The set of ground truth parameters is

$$\Theta = \{(0.5 - c, 0.5c, 1, c) | c \in R\}.$$



$$y_t = (0.5 - c)y_{t-1} + 0.5cy_{t-2} + u_{t-1} + cu_{t-2} + \eta_{t-1} + c\eta_{t-2}$$

Question: are those ground truth parameters equally good?

True model is equivalent to:

$$y_t = (0.5 - c)y_{t-1} + 0.5cy_{t-2} + u_{t-1} + cu_{t-2} + \eta_{t-1} + c\eta_{t-2}$$

Suppose an algorithm that learns one of the ground truth for some c:

$$\theta_c = (0.5 - c, 0.5c, 1, c).$$

Then, the expected prediction error of the learned model is $\eta_t \sim \mathcal{N}(0, \sigma_\eta^2)$

$$\mathbb{E}[(y_t - \hat{y}_t)^2] = \mathbb{E}[(y_t - \hat{y}_t)^2] = \mathbb{E}[(\eta_{t-1} + c\eta_{t-2})^2] = (1 + c^2)\sigma_\eta^2$$

So among all the ground truth, $\theta_0 = (0.5, 0, 1, 0)$ minimizes the prediction error!

Properties of θ_0

- True model: $y_t = 0.5y_{t-1} + u_{t-1} + \eta_{t-1}$
- $\theta_0 = (0.5, 0, 1, 0)$ is
  - the parameters of the minimal-order true model;
  - the parameters that minimizes the prediction error;
  - the most sparse parameter in $\Theta = \{(0.5 - c, 0.5c, 1, c) | c \in R\}$
- Ideally, we want the system identification algorithm to learn θ_0 in Θ.

Properties of θ_0

- True model: $y_t = 0.5y_{t-1} + u_{t-1} + \eta_{t-1}$
- $\theta_0 = (0.5, 0, 1, 0)$ is
  - the parameters of the minimal-order true model;
  - the parameters that minimizes the prediction error;
  - the most sparse parameter in $\Theta = \{(0.5 - c, 0.5c, 1, c) | c \in R\}$ →need a l_1 regularization?
- Ideally, we want the system identification algorithm to learn θ_0 in Θ.
- Next, we show that regular least-squares does the job!

Least-squares setup (multiple-trajectory case)

- Suppose that multiple independent trajectories can be sampled

Experiment 1: $y_0^{(1)}, u_0^{(1)}, \dots, u_{T-1}^{(1)}, y_T^{(1)}$

Experiment 2: $y_0^{(2)}, u_0^{(2)}, \dots, u_{T-1}^{(2)}, y_T^{(2)}$

…

Experiment N: $y_0^{(N)}, u_0^{(N)}, \dots, u_{T-1}^{(N)}, y_T^{(N)}$

$$\hat{\theta} = \min_{\alpha_i, \beta_i} \frac{1}{N} \sum_{k=1}^{N} \left(y_T^{(k)} - \hat{y}_T^{(k)} \right)^2$$

$$s.t\ \hat{y}_T^{(k)} = \sum_{i=1}^{\bar{n}_\alpha} \alpha_i y_{T-i}^{(k)} + \sum_{i=1}^{\bar{n}_\beta} \beta_i u_{T-i}^{(k)}$$

- Since we have closed-solution of $\hat{\theta}$, it is possible to show that $\mathbb{E}\left[\hat{\theta}\right] = \theta_0$.
- For the RLS estimate $\hat{\theta}$, we study its sample complexity (how many samples are needed to guarantee $\left|\left|\hat{\theta} - \theta_0\right|\right| < \epsilon$ with high probability?)

Sample Complexity Analysis (Multi-Trajectory)

$\bar{n} = \bar{n}_\alpha + \bar{n}_\beta$

Σ_z is covariance matrix of the regressor

$Z\hat{\theta} = \mathrm{Y}$

Theorem 1:

Suppose that

- u_t and e_t are i.i.d zero-mean Gaussian variables with $\sigma_u > 0$, $\sigma_\eta > 0$.
- $T \geq \max(\bar{n}_\alpha, \bar{n}_\beta) + 1$

Then, if the sample size N is large enough, with probability $1 - \delta$,

$$\|\hat{\theta} - \theta_0\|_2 \leq \frac{16\sigma_\eta}{\underline{\sigma}(\Sigma_z)} \sqrt{\frac{(1+\bar{n})||\Sigma_z|| \log\left(\frac{18}{\delta}\right)}{N}} = \mathcal{O}\left(\sqrt{\frac{\log\left(\frac{18}{\delta}\right)}{N}}\right).$$

Least-squares setup (single-trajectory case)

- Suppose that only one trajectory is sampled

Experiment: $y_0\ , u_0\ , \dots , u_{T-1}, y_T$ $\longrightarrow$

$$\hat{\theta} = \min_{\alpha_i,\beta_i} \frac{1}{T} \sum_{k=\max(\bar{n}_\alpha,\bar{n}_\beta)}^{T} (y_k - \hat{y}_k)^2$$

$$s.t\ \ \hat{y}_k = \sum_{i=1}^{\bar{n}_\alpha} \alpha_i y_{k-i} + \sum_{i=1}^{\bar{n}_\beta} \beta_i u_{k-i}$$

- Different from the multiple trajectory case, analyze the RLS estimate $\hat{\theta}$ in the single-trajectory case is difficult due to correlations between regressors.

Correlations in LS

$$\hat{\theta} = \min_{\theta} \left|\left|Z\theta - \mathrm{Y}\right|\right|_2^2 = (Z^T Z)^{-1} Z^T Y$$

Proof idea: We want to show how fast the least singular value of the matrix $Z^T Z$ goes to infinity as N increases.

Multiple trajectory setup

$$Z = \begin{bmatrix} & & & & \cdots & & & & \\ & & & & \cdots & & & & \\ y_{T-1}^{(k)} & \cdots & y_{T-\bar{n}_\alpha}^{(k)} & u_{T-1}^{(k)} & \cdots & u_{T-\bar{n}_\beta}^{(k)} \\ y_{T-1}^{(k+1)} & \cdots & y_{T-\bar{n}_\alpha}^{(k+1)} & u_{T-1}^{(k+1)} & \cdots & u_{T-\bar{n}_\beta}^{(k+1)} \\ & & & & \cdots & & & & \\ & & & & \cdots & & & & \end{bmatrix}$$

The rows of Z are independent to each other;
The entries in each row are correlated.

- $\mathbb{E}\left[\hat{\theta}\right]$ is easy to compute
- Standard tools in high-dimensional probability can be applied if you know the trick.

Single trajectory setup

$$Z = \begin{bmatrix} & & & & \cdots & & & & \\ & & & & \cdots & & & & \\ y_{k-1} & y_{k-2} & \cdots & y_{k-\bar{n}_\alpha} & u_{k-1} & u_{k-2} & \cdots & u_{k-\bar{n}_\beta} \\ y_k & y_{k-1} & \cdots & y_{k-\bar{n}_\alpha+1} & u_k & u_{k-1} & \cdots & u_{k+1-\bar{n}_\beta} \\ & & & & \cdots & & & & \\ & & & & \cdots & & & & \end{bmatrix}$$

The rows of Z are correlated;
The entries in each row are correlated too.

- $\mathbb{E}\left[\hat{\theta}\right]$ is difficult to compute
- Mixing-time type arguments are needed to bound the correlation between rows.

Sample Complexity Analysis (Single-Trajectory)

- **Theorem 2**: Suppose that
- the magnitude ρ^* of the largest pole is less than 1;
- u_t and e_t are i.i.d zero-mean Gaussian variables with $\sigma_u > 0$, $\sigma_\eta > 0$.

If $\bar{n}_\alpha \geq n_\alpha$ and $\bar{n}_\beta \geq n_\beta$, then for T large enough, with probability $1-\delta$, the OLS estimation error satisfies

- $$\|\hat{\theta} - \theta_0\| \leq O\left(\frac{\tau}{1-\rho^*}\frac{\sigma_\eta}{\sigma_u}\sqrt{\frac{(\bar{n}_\alpha+\bar{n}_\beta)\log T}{T}}\log\left(\frac{T}{\delta}\right)\right).$$

Remark:

- $\tau/(1-\rho^*)$ is an upper bound of the H_∞ norm of the ARX model, and σ_η/σ_u is the noise-to-input ratio.
- The OLS estimator $\hat{\theta}$ converges to the most sparse solution θ_0 in the single-trajectory setup.

More remarks:

- Since AR and FIR models are special cases of ARX model, the theorem above can be applied to those models.
- For multiple-trajectory setup, the ARX model does not need to have all poles inside the unit circle.
- Theorem 1 can be extended to vector ARX models.

A brief literature review

- The overparameterized LS estimate converges to the most sparse solution θ_0 in both single-trajectory and multi-trajectory setup

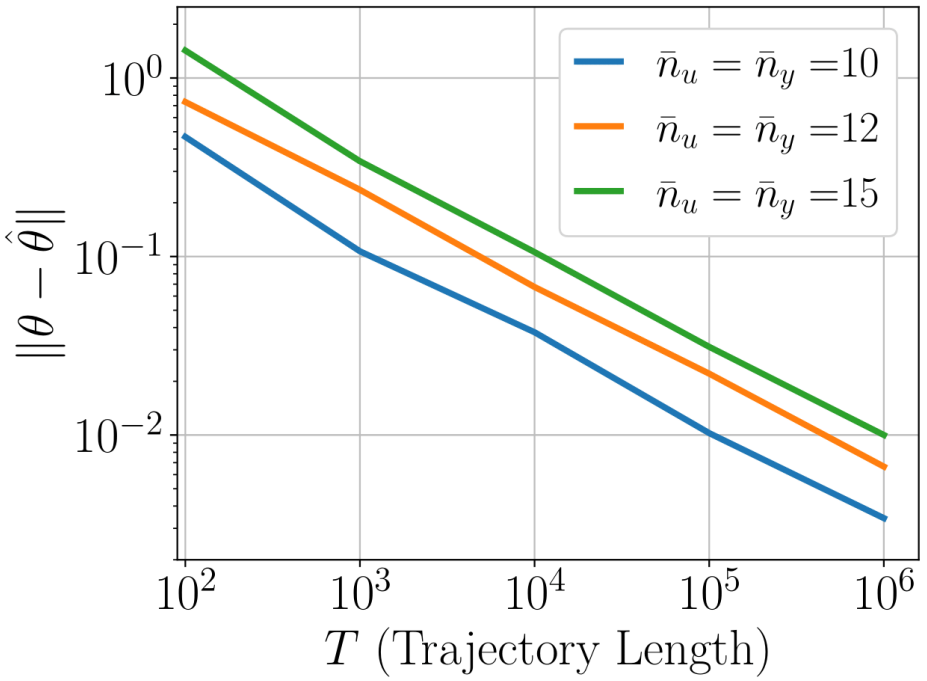
→self regularization or implicit regularization!

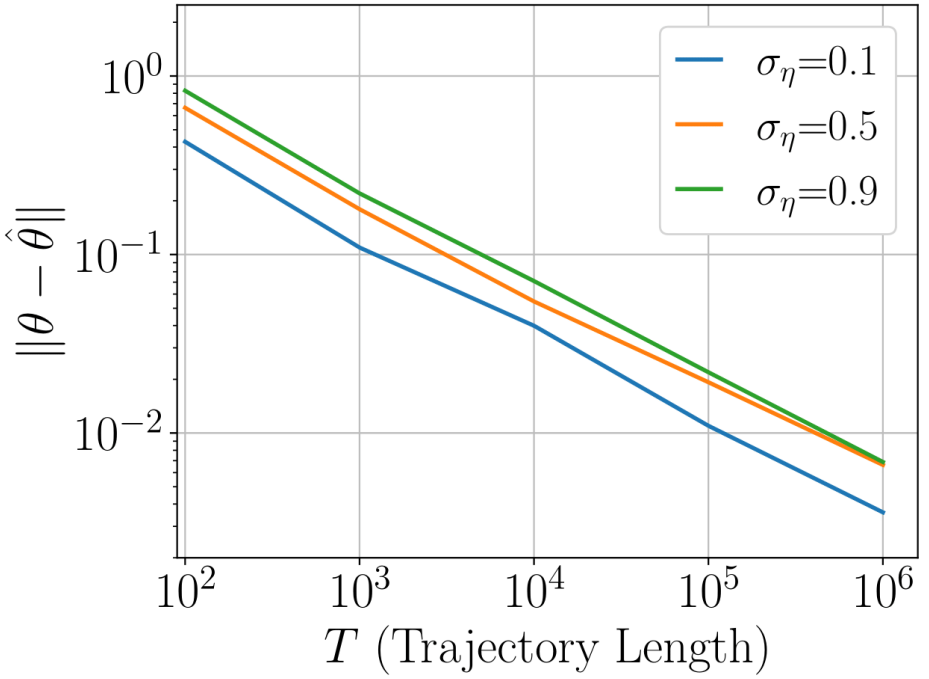
	Previous works [Jones and Dahleh, 2022] [Ljung and Wahlberg, 1992]	Our work
Method	RLS: $\hat{\theta} \leftarrow \underset{\hat{\theta}}{\operatorname{argmax}} \sum_t (y_t - \hat{\theta} z_t)^2 + \lambda \|\hat{\theta}\|^2$	OLS: $\hat{\theta} \leftarrow \underset{\hat{\theta}}{\operatorname{argmax}} \sum_t (y_t - \hat{\theta} z_t)^2$
Guarantees	$\|\hat{\theta} - \theta\| \leq O(\frac{\sqrt{\log(T/\lambda) + \lambda}}{\sqrt{T}})$	$\|\hat{\theta} - \theta\| \leq O(\frac{\sqrt{\log(T)}}{\sqrt{T}})$
Pros/ Cons/ Comments	Tuning λ is needed. Manifests the "oracle property" [Candes 2006]	"self-regularization" The first finite sample result for OLS on unknown (and known)-order ARX models

- Jones, D., Dahleh, M., Non-Parametric Finite Time Identification of Closed Loop Systems, ACC 2022, forthcoming
- Ljung, L., & Wahlberg, B. (1992). Asymptotic properties of the least-squares method for estimating transfer functions and disturbance spectra. *Advances in Applied Probability*, *24*(2), 412-440.
- Candes, E. J. (2006). Modern statistical estimation via oracle inequalities. *Acta numerica*, *15*, 257-325.

Experiments

$n_\alpha = n_\beta = 10, \sigma_u = \sigma_\eta = 1, \rho^* = 0.85$





Summary

- A self-regularization property in RLS-based ARX identification
- Sample complexity bounds for RLS estimates in different setups.

“Sample Complexity Analysis and Self-regularization in Identification of Over-parameterized ARX Models”

Conference version will be presented at CDC 2022.