2021 61st IEEE Conference on Decision and Control (CDC)
December 6-9, 2022. Cancún, Mexico

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

Zhe Du* Zexiang Liu* Jack Weitze Necmiye Ozay

Abstract— **AutoRegressive eXogenous (ARX) models form one of the most important model classes in control theory, econometrics, and statistics, but they are yet to be understood in terms of their finite sample identification analysis. The technical challenges come from the strong statistical dependency not only between data samples at different time instances but also between elements within each individual sample. In this work, for ARX models with potentially unknown orders, we study how ordinary least squares (OLS) estimator performs in terms of identifying model parameters from data collected from either a single length-T trajectory or N i.i.d. trajectories. Our main results show that as long as the orders of the model are chosen optimistically, i.e., we are learning an over-parameterized model compared to the ground truth ARX, the OLS will converge with the optimal rate $\mathcal{O}(1/\sqrt{T})$ (or $\mathcal{O}(1/\sqrt{N})$) to the true (low-order) ARX parameters. This occurs without the aid of any regularization, thus is referred to as *self-regularization*. Our results imply that the oracle knowledge of the true orders and usage of regularizers are not necessary in learning ARX models — over-parameterization is all you need.**

I. Introduction

AutoRegressive eXogenous (ARX) models are versatile in terms of modeling real world signals that have temporal dependency and affected by external excitations, including those emerging in genomics, neuroscience, medicine, macroeconomics, stock markets, etc. [2]–[7]. Especially in such domains where physics-based modeling is hard, the identification of ARX models from input/output data is a fundamental problem [8]. Existing work on ARX identification has generally consisted of new algorithmic techniques, and asymptotic analysis, where one studies the behavior of the estimators as the number of data samples goes to infinity. Asymptotic properties of classical estimators such as OLS and maximum likelihood estimators for ARX are studied, e.g., in [9]–[13]. However, practitioners concerned with safety and risk aversion ask: how many samples are needed to guarantee at least a certain level of performance? In the identification context, the result is a probabilistic bound on error in estimated parameters. This requires non-asymptotic finite sample analysis, which is crucial in tasks such as prediction, filtering, and adaptive control [14], [15].

The major challenge in determining finite-sample error bounds for ARX identification comes from the strong correlations among the data generated by ARX models. Recent advances in the understanding of high-dimensional statistics, random matrix theory, and random processes [16]–[20] have enabled rapid progress in non-asymptotic result for learning dynamical models with a single data trajectory [21]–[26]. However, ARX models are more challenging to analyze: even if we simplify the setting to collecting a single sample from multiple i.i.d. trajectories, correlation still exists within individual sample vectors due to outputs regressing onto their past values. This makes it difficult to lower bound the sample covariance matrix and obtain an estimation error upper bound using only quantities depending on model properties.

Contributions: In this work, we study the non-asymptotic behavior of the ordinary least squares (OLS) estimators for identifying stable ARX models with potentially unknown orders. We show that when the orders of ARX model are chosen optimistically, i.e., we are learning an over-parameterized model compared to the true one, the OLS estimator will converge to the true model parameters with the following rate:

- $\mathcal{O}(\sqrt{\log(T)/T})$ — if the data is collected from a single length-T trajectory,
- $\mathcal{O}(1/\sqrt{N})$ — if the data is collected from N i.i.d. trajectories,

where the latter scheme can also deal with explosive (unstable) ARX models.

The implication of our work is that, while learning ARX models, neither knowledge of the exact model orders nor regularization are mandatory. As long as orders used by the estimator are larger than the true ones, sufficient amount of data guarantees that the OLS automatically converges to the true parameter with correct orders. Particularly, the redundant parameters estimated by the OLS estimator, i.e., elements that should be 0 in a naive lifting of the true parameter, indeed converge to 0. This phenomenon is also known as the *oracle property* [27] of estimators. We refer to it as *self-regularization* in this work as the explicit regularization typically required for estimators to satisfy the oracle property is not needed in our setting.

Furthermore, when the true model orders are used in the OLS, we obtain the estimation accuracy upper bound $\mathcal{O}(\sqrt{n\log(T)/T})$ and $\mathcal{O}(\sqrt{n/N})$, where n is the model order. Comparing to minimax lower bounds for similar problems, these upper bounds achieve the optimal (up to logarithmic factors) rates in terms of the dependency on n and T (or N). Perhaps, most relevant to our work is that in [15], where similar rates for an ℓ_2-regularized least squares

This work was supported in part by ONR grants N00014-18-1-2501 and N00014-21-1-2431 and NSF grants EECS-1553873 and CNS-1931982. Z. Du was also supported in part by AFOSR YIP award FA9550-19-1-0026 and NSF CAREER award CCF-1845076.

* First two authors contributed equally.

The authors are with the Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor zhedu,zexiang,jweitze,necmiye@umich.edu

A full version with extended proofs is available at [1].

978-1-6654-6761-2/22/$31.00 ©2022 IEEE

estimator for ARX identification with known orders from a single trajectory are obtained.

Organization: In Section II, we introduce ARX models and its over-parameterized OLS estimators. Section III uses a simple example to illustrate and give insight about the self-regularization phenomenon. Section IV and V provide the non-asymptotic results for OLS estimators with single and multiple trajectories respectively. Section VI shows numerical results that support our theoretical results.

II. Preliminaries

In this paper, boldface uppercase (lowercase) letters denote matrices (vectors); plain letters mainly denote scalars. For a matrix $\mathbf{E}$, $\mathbf{E}(i,:)$ denotes the ith row of $\mathbf{E}$, and $\mathbf{E}(i,j{:}k)$ denotes the ith row preserving only the jth to kth elements. Let $\bar{\sigma}(\mathbf{E})$ and $\underline{\sigma}(\mathbf{E})$ denote the largest and smallest singular values. $\bar{\lambda}(\mathbf{E})$ and $\underline{\lambda}(\mathbf{E})$ are defined similarly for the eigenvalues of positive semi-definite matrices. Given an arbitrary square matrix $\mathbf{E}$, let $\rho(\mathbf{E})$ denote its spectral radius. For any $s \in \mathbb{N}$, we let $[s] := \{1,2,\dots,s\}$. For a sequence of variables $X_0, X_1, \dots, X_N$, let $X_{0:N} := \{X_i\}_{i=0}^N$. Notation $\mathbf{I}_n$ denotes the n dimensional identity matrix; $0_{m\times n}$ denotes the $m \times n$ dimensional zero matrix; $\mathbf{e}_i$ denotes the one-hot vector with the ith element being 1.

For an arbitrary square matrix $\mathbf{M} \in \mathbb{R}^{m \times m}$ and a free parameter ρ such that $\rho \geq \rho(\mathbf{M})$, we define

$$\tau(\mathbf{M},\rho) := \begin{cases} \sup_{k\in\mathbb{N}} \|\mathbf{M}^k\|/\rho^k & \text{if } \rho(\mathbf{M}) > 0, \\ \sum_{k=0}^{m-1} \|\mathbf{M}^k\| & \text{if } \rho(\mathbf{M}) = 0. \end{cases} \tag{1}$$

When $\rho(\mathbf{M}) > 0$, this gives $\|\mathbf{M}^k\| \leq \tau(\mathbf{M},\rho)\rho^k$ for any $k \in \mathbb{N}$. By definition, (i) when $\rho > \rho(\mathbf{M})$, $\tau(\mathbf{M},\rho)$ is finite by Gelfand's formula; (ii) when $\rho = \rho(\mathbf{M})$, $\tau(\mathbf{M},\rho)$ is finite as long as $\mathbf{M}$ is diagonalizable, a generic property of matrices in $\mathbb{R}^{m\times m}$. In the latter case, $\tau \leq \|\mathbf{V}\|\|\mathbf{V}^{-1}\|$, where $\mathbf{V}$ is a matrix whose columns span the eigenspace of $\mathbf{M}$.

For a zero-mean random vector $\mathbf{x}$, we let $\mathbf{Cov}(\mathbf{x}) := \mathbb{E}[\mathbf{x}\mathbf{x}^\intercal]$ denote its covariance matrix. Claims such as "for all j, with probability $1-\delta$, for all i, event $\mathcal{E}_{i,j}$ occurs" is equivalent to "for all j, $\mathbb{P}(\cap_i \mathcal{E}_{i,j}) \geq 1-\delta$".

A. ARX Basics

In this work, we seek to learn the following ARX models

$$y_t = \sum_{i=1}^{n_\alpha} \alpha_i y_{t-i} + \sum_{i=1}^{n_\beta} \beta_i u_{t-i} + \eta_{t-1}, \tag{2}$$

with unknown model orders n_α and n_β. The output y_t is driven by i.i.d. input $u_t \sim \mathcal{N}(0,\sigma_u^2)$ and process noise $\eta_t \sim \mathcal{N}(0,\sigma_\eta^2)$ for all time t. We assume α_{n_α} and β_{n_β} are non-zero, as the orders can be reduced otherwise. The parameters $\alpha_{1:n_\alpha}$, $\beta_{1:n_\beta}$ are to be learned from the input-output data.

Let $n{:=}\max(n_\alpha, n_\beta)$, and $\alpha_i{=}\beta_j{=}0$ for any $i > n_y$, $j > n_u$. Define polynomials $q(z) := z^n - \sum_{i=1}^n \alpha_i z^{n-i}$ and $p(z) := \sum_{i=1}^n \beta_i z^{n-i}$, and $\rho^\star := \max_{z:q(z)=0} |z|$. Define $\mathbf{A} := \left[\begin{array}{c|c} 0_{1\times(n-1)} & \alpha_n \\ \hline \mathbf{I}_{n-1} & \begin{matrix}\vdots \\ \alpha_1\end{matrix} \end{array}\right]$, $\mathbf{B} := [\beta_n, \dots, \beta_1]^\intercal$, $\mathbf{C} := [0,\dots,0,1]$, and $\mathbf{\Gamma} := \mathbf{C}^\intercal$. The ARX model (2) has an equivalent single-input single-output (SISO) state space representation in the following observable canonical form

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{A}\mathbf{x}_t + \mathbf{B}u_t + \mathbf{\Gamma}\eta_t \\ y_t &= \mathbf{C}\mathbf{x}_t \end{aligned} \tag{3}$$

where $\mathbf{x}_t \in \mathbb{R}^n$ denotes the internal state that bridges the input and output in the ARX model. We can see $\rho^\star$ is equal to $\rho(\mathbf{A})$, the spectral radius of $\mathbf{A}$, as its characteristic polynomial is given by $q(z)$. For $i \in \mathbb{N}$, define $g_i^x := \mathbf{C}\mathbf{A}^i$, $g_i^u := \mathbf{C}\mathbf{A}^{i-1}\mathbf{B}$, $g_0^u := 0$, $g_i^\eta := \mathbf{C}\mathbf{A}^{i-1}\mathbf{\Gamma}$, $g_0^\eta := 0$, which are referred to as Markov parameters. Using the internal state $\mathbf{x}_t$, the ARX model (2) can be further written as:

$$y_t = g_{t_0}^x \mathbf{x}_{t-t_0} + \sum_{i=1}^{t_0} g_i^u u_{t-i} + \sum_{i=1}^{t_0} g_i^\eta \eta_{t-i} \tag{4}$$

for some starting time index $t - t_0$. Compared to the ARX model (2), introducing the internal state $\mathbf{x}$ eliminates the dependency of output y_t on past y's.

B. Over-parameterized OLS Estimator for ARX

To make sure the true ARX model (2) with unknown orders n_α and n_β can be learned, one could fit the data using an ARX model with orders $\bar{n}_\alpha$ and $\bar{n}_\beta$ large enough such that $\bar{n}_\alpha \geq n_\alpha$ and $\bar{n}_\beta \geq n_\beta$. We refer to this as over-parameterization. Let $\bar{n} := \bar{n}_\alpha + \bar{n}_\beta$ and

$$\boldsymbol{\theta} := [\alpha_{1:n_\alpha}, 0_{1\times(\bar{n}_\alpha - n_\alpha)}, \beta_{1:n_\beta}, 0_{1\times(\bar{n}_\beta - n_\beta)}]^\intercal \in \mathbb{R}^{\bar{n}} \tag{5}$$

which embeds the true parameter. Under this over-parameterization setup, we study the OLS estimator with data collected from (i) a single trajectory and (ii) multiple independent trajectories generated by (2) with system starting from rest at time $1 - \bar{n}_\beta$, i.e., y_t, u_t, and η_t are zero for all $t < 1 - \bar{n}_\alpha$. This initial condition implies that the internal state $\mathbf{x}_t = 0$ for $t \leq 1 - \bar{n}_\alpha$. The results in this paper hold similarly for non-zero initial conditions as long as the initial conditions have finite second-order moment.

Single Trajectory: Consider a single trajectory $\{(u_t, y_t, \eta_t)_{t=1-\max(\bar{n}_\beta,\bar{n}_\alpha)}^T\}$ of length T. Define

$$\begin{aligned} \mathbf{z}_t &:= [y_{t-1}, \dots, y_{t-\bar{n}_\alpha}, u_{t-1}, \dots, u_{t-\bar{n}_\beta}]^\intercal \in \mathbb{R}^{\bar{n}}, \\ \mathbf{Z} &:= [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_T]^\intercal \in \mathbb{R}^{T\times\bar{n}}, \\ \mathbf{Y} &:= [y_1, y_2, \dots, y_T]^\intercal \in \mathbb{R}^T, \\ \mathbf{E} &:= [\eta_0, \eta_1, \dots, \eta_{T-1}]^\intercal \in \mathbb{R}^T, \end{aligned} \tag{6}$$

where we allow for potentially negative time indices, e.g. $y_{1-\bar{n}_\alpha}$, to ease the exposition.

Multiple Trajectories: Consider N i.i.d. trajectories $\{(u_t^{(i)}, y_t^{(i)}, \eta_t^{(i)})_{t=1}^T\}_{i=1}^N$ each with length T. Define

$$\begin{aligned} \mathbf{z}^{(i)} &:= [y_{T-1}^{(i)}, \dots, y_{T-\bar{n}_\alpha}^{(i)}, u_{T-1}^{(i)}, \dots, u_{T-\bar{n}_\beta}^{(i)}]^\intercal \in \mathbb{R}^{\bar{n}}, \\ \mathbf{Z} &:= [\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \dots, \mathbf{z}^{(N)}]^\intercal \in \mathbb{R}^{N\times\bar{n}}, \\ \mathbf{Y} &:= [y_T^{(1)}, y_T^{(2)}, \dots, y_T^{(N)}]^\intercal \in \mathbb{R}^N, \\ \mathbf{E} &:= [\eta_{T-1}^{(1)}, \eta_{T-1}^{(2)}, \dots, \eta_{T-1}^{(N)}]^\intercal \in \mathbb{R}^N, \end{aligned} \tag{7}$$

where only the final portion, i.e., $y_{T-\bar{n}_\alpha:T-1}$ and $u_{T-\bar{n}_\beta:T-1}$, of each trajectory is used to ease the analysis, similar to the setup in [28].

For both single- and multiple-trajectory cases, we have $\mathbf{Z}\boldsymbol{\theta} + \mathbf{E} = \mathbf{Y}$, and the OLS estimator is given by

$$\hat{\boldsymbol{\theta}} := (\mathbf{Z}^\mathsf{T}\mathbf{Z})^{-1}\mathbf{Z}^\mathsf{T}\mathbf{Y}. \tag{8}$$

For an explosive ($\rho^\star > 1$) ARX model, the output y_t can grow exponentially over time. Generating a single trajectory for learning may not be safe or practically feasible, and even if an experiment can be run, the collected data can lead to numerical issues. For such models, using multiple trajectories with small trajectory length T could be a better choice.

C. Model Ambiguity in Over-parameterization

Over-parameterization increases model capacity but also induces model ambiguity. For all $i = 1, \dots, \min(\bar{n}_\alpha - n_\alpha, \bar{n}_\beta - n_\beta)$, let

$$\begin{aligned}\mathbf{v}_i := [0_{1\times i-1},\ -1,\ \alpha_{1:n_\alpha},\ 0_{1\times(\bar{n}_\alpha - n_\alpha - i)},\\ 0_{1\times i},\ \beta_{1:n_\beta},\ 0_{1\times(\bar{n}_\beta - n_\beta - i)}]^\mathsf{T} \in \mathbb{R}^{\bar{n}}.\end{aligned} \tag{9}$$

Then, for $\tilde{\boldsymbol{\theta}} := \boldsymbol{\theta} + \sum_i c_i \mathbf{v}_i$ for any $c_i \in \mathbb{R}$, with some algebra, one can see the model

$$y_t = \sum_{i=1}^{\bar{n}_\alpha} \tilde{\boldsymbol{\theta}}_i y_{t-i} + \sum_{i=1}^{\bar{n}_\beta} \tilde{\boldsymbol{\theta}}_{\bar{n}_\alpha + i} u_{t-i} + \eta_{t-1} + \sum_i c_i \eta_{t-1-i} \tag{10}$$

is equivalent to the true ARX model (2). We use the "true parameter" set $\Theta := \{\tilde{\boldsymbol{\theta}} : \tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} + \sum_i c_i \mathbf{v}_i, c_i \in \mathbb{R}\}$ to denote all $\tilde{\boldsymbol{\theta}}$ such that the over-parameterized model is equivalent to the true ARX model (2). Note that the ambiguity carried by $\mathbf{v}_i$'s and Θ essentially translates to multiplying both $q(z)$ and $p(z)$ by a monic polynomial $\prod_i (z_i - \xi_i)$ for some $\{\xi_i\}_i$. Hence, the transfer function, $p(z)/q(z)$, is unchanged due to the cancellation.

In Section III and IV, we numerically and theoretically show that even though the true parameter set Θ has infinitely many elements, $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}$, the one with the smallest order and the original $\alpha_{1:n_\alpha}$ and $\beta_{1:n_\beta}$ in model (2).

III. A Simple Illustrative Example

In this section, we show how self-regularization emerges through a numerical experiment, followed by high-level insight into the cause of such phenomenon.

We consider a simple ARX model

$$y_t = -0.3y_{t-1} + 0.4y_{t-2} + u_{t-1} + \eta_{t-1} \tag{11}$$

with $n_\alpha = 2, n_\beta = 1$ and $\sigma_u = \sigma_\eta = 1$. The OLS over-parameterization orders are chosen as $\bar{n}_\alpha = 4, \bar{n}_\beta = 2$, which gives $\boldsymbol{\theta} = [-0.3, 0.4, 0, 0, 1, 0]^\mathsf{T}$. By definition, the true parameter set is given by $\Theta = \{\boldsymbol{\theta} + c_1\mathbf{v}_1 : c_1 \in \mathbb{R}\}$ for $\mathbf{v}_1 := [-1, -0.3, 0.4, 0, 0, 1]$.

Fig. 1 numerically shows the convergence of the over-parameterized single-trajectory OLS. We can see that as the trajectory length T increases, the range of $\hat{\boldsymbol{\theta}}$ shrinks to $\boldsymbol{\theta}$. Specifically, the exact-parameterized part, i.e., $[\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1]$ converges to the true model parameter $[-0.3, 0.4, 1]$, while the over-parameterized part, i.e., $[\hat{\alpha}_3, \hat{\alpha}_4, \hat{\beta}_2]$, converges to 0.

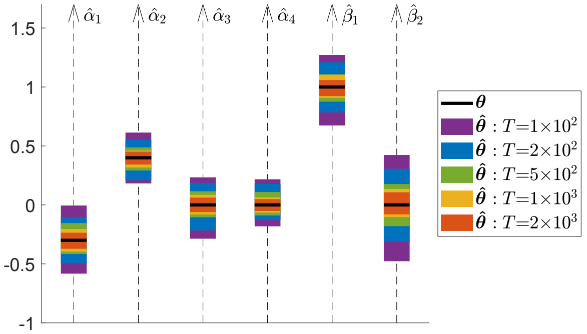


Fig. 1. True $\boldsymbol{\theta}$ vs. OLS estimator $\hat{\boldsymbol{\theta}}$ computed using single-trajectory. For 500 i.i.d. experiments per trajectory length T, the plot shows the range (maximum and minimum values) of each element in $\hat{\boldsymbol{\theta}}$ with varying T, where $\hat{\alpha}_{1:4} := \hat{\boldsymbol{\theta}}_{1:4}$ and $\hat{\beta}_{1:2} := \hat{\boldsymbol{\theta}}_{5:6}$. Observe that as T increases, the redundant parameters (i.e. $\hat{\alpha}_3$, $\hat{\alpha}_4$, and $\hat{\beta}_2$) converge to zero, and the remaining parameters converge to their corresponding true value.

For any $\tilde{\boldsymbol{\theta}} \in \Theta$, the prediction error at time t with respect to the regressor $\mathbf{z}_t := [y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, u_{t-1}, u_{t-2}]^\mathsf{T}$ and output y_t is given by

$$\mathbb{E}[(y_t - \tilde{\boldsymbol{\theta}}^\mathsf{T}\mathbf{z}_t)^2] = \mathbb{E}[(\eta_{t-1} + c\eta_{t-2})^2] = (1 + c^2)\sigma_\eta^2, \tag{12}$$

which achieves the minimum when $c = 0$, i.e. $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}$. This implies that as long as there is noise, $\sigma_\eta \neq 0$, among the infinitely many true parameters $\tilde{\boldsymbol{\theta}}$ the one with the smallest order, $\boldsymbol{\theta}$, uniquely minimizes the prediction error. Since OLS minimizes the empirical prediction error that converges to the expectation in (12), it is reasonable to conjecture that $\hat{\boldsymbol{\theta}}$ would converge to $\boldsymbol{\theta}$. In what follows, we prove this conjecture through finite-sample analysis.

IV. Single-trajectory Case

We first present a few assumptions and notations to be used in our theoretical analysis. As a key assumption in the single-trajectory analysis, we require the ARX models to be stable as stated next.

Assumption 1: The ARX model has $\rho^\star < 1$.

Under this Assumption, the output y_t will be non-explosive. Recall that $\rho^\star$ is also the spectral radius of matrix $\mathbf{A}$ defined in the state space representation (3). Then, for a free parameter ρ such that $\rho \geq \rho^\star$, $\tau(\mathbf{A}, \rho)$ defined in (1) will be used to bound the growth of $\mathbf{A}$'s powers, as in [29]. We drop the arguments of $\tau(\mathbf{A}, \rho)$ and use τ as a shorthand notation. When $\rho^\star > 0$, one can see $\|\mathbf{A}^k\| \leq \tau\rho^k$ for any $k \in \mathbb{N}$.

With the Markov parameters g_i^u and g_i^η in SectionII-A, we define the following two Toeplitz matrices: for $h \geq \bar{n}_\alpha$, let

$$\begin{aligned}\mathbf{G}^u(h) &:= \begin{bmatrix} g_1^u & g_2^u & & \cdots & & g_h^u \\ & g_1^u & & \cdots & & g_{h-1}^u \\ & & \ddots & & & \vdots \\ & & & g_1^u & \cdots & g_{h-\bar{n}_\alpha+1}^u \end{bmatrix} \in \mathbb{R}^{\bar{n}_\alpha \times h} \\ \mathbf{G}^\eta(h) &:= \begin{bmatrix} g_1^\eta & g_2^\eta & & \cdots & & g_h^\eta \\ & g_1^\eta & & \cdots & & g_{h-1}^\eta \\ & & \ddots & & & \vdots \\ & & & g_1^\eta & \cdots & g_{h-\bar{n}_\alpha+1}^\eta \end{bmatrix} \in \mathbb{R}^{\bar{n}_\alpha \times h}\end{aligned} \tag{13}$$

From the representation (4), one can see $\mathbf{G}^u(h)$ and $\mathbf{G}^\eta(h)$ correspondingly map $u_{t-2:t-h-1}$ and $\eta_{t-2:t-h-1}$, to $y_{t-1:t-\bar{n}_\alpha}$. Note $y_{t-1:t-\bar{n}_\alpha}$ is the first part of the regressor vector $\mathbf{z}_t$ defined in (6).

Table I summarizes some notation that is used in the statements of the main results. In this table $\delta \in (0, 1)$. Note that some of the defined variables such as L and $\underline{T}_1$ depend on T, δ, the free parameter ρ, and its corresponding τ.

TABLE I

NOTATIONS — SINGLE TRAJECTORY

L	$\lceil \frac{\log(T)}{\log(1/\rho)} \rceil + 2\bar{n} - 1$
$\sigma_{u,\eta}$	$\sqrt{\sigma_u^2 \lVert \mathbf{B} \rVert^2 + \sigma_\eta^2}$
$\mathbf{G}_0$	$\left[\begin{array}{ccc} 0_{\bar{n}_\alpha \times 1} & \mathbf{G}^u(\max(\bar{n}_\alpha, \bar{n}_\beta - 1)) & \mathbf{G}^\eta(\bar{n}_\alpha) \\ \hline \mathbf{I}_{\bar{n}_\beta} & & \cdots 0 \cdots \end{array}\right]$
$\mathbf{G}_1$	$\left[\begin{array}{cc} 0_{\bar{n}_\alpha \times 1} & \mathbf{G}^u(\bar{n} + n) \\ \hline \mathbf{I}_{\bar{n}_\beta} & \cdots 0 \cdots \end{array}\right]$
$\mathbf{G}_2$	$\left[\begin{array}{c} \mathbf{G}^\eta(\bar{n}_\alpha) \\ \hline 0_{\bar{n}_\beta \times \bar{n}_\alpha} \end{array}\right]$
$\underline{\lambda}_0$	$\max\big(\underline{\sigma}(\mathbf{G}_0)^2 \min(\sigma_u^2, \sigma_\eta^2)\ ,$ $\underline{\sigma}(\mathbf{G}_1)^2 \sigma_u^2 + \underline{\sigma}(\mathbf{G}_2)^2 \sigma_\eta^2 \big)$
C_o	$192 \max(\sigma_{u,\eta}^2, \sigma_u^2)/\underline{\lambda}_0$
$\underline{T}_1$	$4L \log(8L/\delta) + 2L$
$\underline{T}_{\underline{\lambda},1}$	$\left(\frac{640\bar{n}\tau^2 \max(\sigma_{u,\eta}^2, \sigma_{u,\eta}\sigma_u) \log(8T/\delta)}{(1-\rho)^2 \underline{\lambda}_0} \right)^2$
$\underline{T}_{\underline{\lambda},2}$	$16 \log(4L/\delta)L + \frac{32\bar{n}\tau^2 \max(\sigma_{u,\eta}^2, \sigma_u^2)L}{(1-\rho)^2 \underline{\lambda}_0}$

Theorem 1 (OLS Convergence — Single Trajectory): Suppose Assumption 1 holds, and the OLS orders satisfy $\bar{n}_\alpha \geq n_\alpha$, $\bar{n}_\beta \geq n_\beta$. Then, for any $\rho \in [\rho^\star, 1)$ and its corresponding τ, as long as $T \geq \max(\underline{T}_1, \underline{T}_{\underline{\lambda},1}, \underline{T}_{\underline{\lambda},2})$, with probability at least $1 - \delta$,

$$\|\hat{\theta} - \theta\| \leq \frac{C_o \tau}{1-\rho} \cdot \frac{\sigma_\eta}{\sigma_u} \sqrt{\frac{\bar{n} \log(T/\rho^{2\bar{n}})}{T \log(1/\rho)}} \log\left(\frac{36T}{\delta}\right). \quad (14)$$

According to Theorem 1, with rate $\mathcal{O}(\sqrt{\log(T)/T})$, the OLS estimator $\hat{\theta}$ converges to $\theta = [\alpha_{1:n_\alpha}, 0_{1\times(\bar{n}_\alpha - n_\alpha)}, \beta_{1:n_\beta}, 0_{1\times(\bar{n}_\beta - n_\beta)}]^\intercal$, the true parameter with orders n_α and n_β, and the original $\alpha_{1:n_\alpha}$, $\beta_{1:n_\beta}$ in model (2). Thus, if the true orders n_β and n_α are unknown, using over-parameterization orders $\bar{n}_\alpha$ and $\bar{n}_\beta$ guarantees the parameter with the smallest order and the most trailing zeros can be recovered, with sufficient data. Similar observations have been in the literature for ℓ_2-regularized and [30] ℓ_1-regularized [31] least squares where it is shown that even when an over-parameterized model is used in estimation, the estimate converges to the true parameter, i.e., oracle property [27] exists. Since this sparse parameter is obtained without any explicit regularizer in our work, we refer to this phenomenon as self-regularization.

When the over-parameterization order $\bar{n}$, spectral radius $\rho^\star$, and noise-to-signal ratio σ_η/σ_u are larger, the upper bound on $\|\hat{\theta} - \theta\|$ become looser, and the conditions on the trajectory length T for the analysis to hold become more stringent.

Of course, Theorem 1 also shows the convergence for the identification setup where the true orders n_α and n_β are assumed to be known and used in the identification. In this case, Theorem 1 yields upper bounds $\mathcal{O}(\sqrt{(n_\alpha + n_\beta)\log(T)/T})$. Compared with minimax lower bounds for time series parameter estimation error [21] and prediction error [32], our developed upper bound almost (up to $\log(T)$ term) meets those lower bounds in terms of the order dependency on the model order and trajectory length.

Note that in Theorem 1, the coefficient C_o has denominator $\underline{\lambda}_0 := \max\big(\underline{\sigma}(\mathbf{G}_0)^2 \min(\sigma_u^2, \sigma_\eta^2), \underline{\sigma}(\mathbf{G}_1)^2\sigma_u^2 + \underline{\sigma}(\mathbf{G}_2)^2\sigma_\eta^2\big)$. $\underline{\lambda}_0$ helps to lower bound the sample covariance matrix $\mathbf{Z}^\intercal\mathbf{Z}$ from 0, i.e. $\underline{\lambda}(\mathbf{Z}^\intercal\mathbf{Z}) > 0$, so that $\hat{\theta} := (\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal\mathbf{Y}$ exists. The following proposition guarantees that $\underline{\lambda}_0 \neq 0$ so that both C_o and the OLS estimator $\hat{\theta}$ are well-defined.

Proposition 1: In the definition of $\underline{\lambda}_0$, matrices $\mathbf{G}_0, \mathbf{G}_2$, and $\mathbf{G}_1$ satisfy the following:

- $\mathbf{G}_0 \in \mathbb{R}^{\bar{n} \times \max(\bar{n}, 2\bar{n}_\alpha + 1)}$ has full row rank.
- $\mathbf{G}_2 \in \mathbb{R}^{\bar{n} \times \bar{n}_\alpha}$ has full row rank as long as $\bar{n}_\beta = 0$.
- $\mathbf{G}_1 \in \mathbb{R}^{\bar{n} \times (\bar{n} + n + 1)}$ has full row rank as long as (i) $q(z)$ and $p(z)$ have no common roots, and (ii) $\bar{n}_\alpha = n_\alpha$ *or* $\bar{n}_\beta = n_\beta$.

The first two bullets can be proved simply by the structures of block matrices. The third one also utilizes basic properties of controllability and observability of the true order state space model (3). Note that both controllability and observability can be guaranteed when $q(z)$ and $p(z)$ have no common roots. This proposition shows $\underline{\sigma}(\mathbf{G}_0), \underline{\sigma}(\mathbf{G}_1)$, and $\underline{\sigma}(\mathbf{G}_1)$ can be non-zero, even under special cases such as $n_\beta = \bar{n}_\beta = 0$ or $\sigma_\eta = 0$, which, by definition, further guarantees $\underline{\lambda}_0 \neq 0$. More discussions on those special cases are in the next section.

A. Special Cases

Theorem 1 is proved for the general ARX models, but it also accounts for several special cases such as Autoregressive (AR) models and noise-free cases. These cases not only serve as sanity checks for our theoretical results but also enjoy tighter guarantees.

AR Case: When the ARX model is not driven by exogenous input, i.e., the order $n_\beta = 0$, it reduces to an AR model. In this case, only the outputs are used to compute the OLS estimator $\hat{\theta}$. To obtain the theoretical guarantees, one only needs to set $n_\beta = \bar{n}_\beta = 0$, $\bar{n} = \bar{n}_\alpha$, and $\sigma_u = 0$ in Theorem 1. Correspondingly, certain parameters have more concise and informative expressions. To list a few, in Table I, we have $\sigma_{u,\eta} = \sigma_\eta$, $\underline{\lambda}_0 = \underline{\sigma}(\mathbf{G}_2)^2\sigma_\eta^2$, and $C_o = 192/\underline{\sigma}(\mathbf{G}_2)^2$. Note that $\underline{\lambda}_0 \neq 0$ according to Proposition 1. We mention two recent works related to this special case. Finite sample analysis of OLS estimator for AR models with known order appeared in [33], where the parameter error in terms of ℓ_∞-norm is analyzed as opposed to the ℓ_2-norm error in our case. Non-asymptotic results for using lasso estimator to learn AR models appeared in [34], which requires mixing time and a slightly different excitation condition, which can be less interpretable for control audience than system-related properties in our work.

FIR Case: When there is no autoregressive part in the ARX model, i.e., $\alpha_{1:n_\alpha} = 0$, y_t depends only on past $u_{t-1:t-n_\beta}$, which is an FIR model. In this case, the data matrix $\mathbf{Z}$ is constructed with the inputs only. Thus, the dependency between the data (rows) in $\mathbf{Z}$ are greatly reduced compared to general ARX models. For example, according to the definitions in (6), for any t and any $t' > t + \bar{n}_\beta$, the rows $\mathbf{z}_t$ and $\mathbf{z}_{t'}$ in $\mathbf{Z}$ are completely independent. Therefore, the estimation error $\|\hat{\theta} - \theta\|$ can be more easily analyzed, and one may expect the guarantees to be tighter as well. On the other hand, one can directly adapt the result in Theorem 1 to this case by setting $\rho^\star = 0$, $n_\alpha = \bar{n}_\alpha = 0$, $\bar{n} = \bar{n}_\beta$. Matrices $\mathbf{G}_0$, $\mathbf{G}_1$, and $\mathbf{G}_2$ will be left only with the lower blocks. It is worth noticing that in the upper bound (14), the order term $\sqrt{\frac{\log(T/\rho^{2\bar{n}})}{T\log(1/\rho)}}$ reduces to $\sqrt{2\bar{n}_\alpha/T}$ when choosing $\rho = \rho^\star = 0$, which is indeed tighter than the general case.

Noise-free Case: The last special case we consider is the noise-free case, i.e. $\sigma_\eta = 0$, and the ARX model is driven by the input only. In this ideal case, it is tempting to believe in the following seemingly true statement: as long as certain amount of data is collected, the estimation error would become exactly 0. It is easy to see this statement when (i) polynomials $q(z)$ and $p(z)$ have no common roots, and (ii) $\bar{n}_\alpha = n_\alpha$ *and* $\bar{n}_\beta = n_\beta$, i.e. the ground truth orders are used to parameterize the OLS estimator $\hat{\theta}$. Note that condition (i) guarantees that there does not exist an equivalent ARX model with smaller orders.

However, it is not necessarily true when $\bar{n}_\alpha \geq n_\alpha$ and $\bar{n}_\beta \geq n_\beta$, i.e., the over-parameterization scenario. This is because with Gaussian input $\mathbf{u}_t$ and condition (i), one can show the covariance matrix $\mathbf{Z}^\intercal\mathbf{Z} \in \mathbb{R}^{(\bar{n}_\alpha+\bar{n}_\beta)\times(\bar{n}_\alpha+\bar{n}_\beta)}$ has rank $n_\alpha + n_\beta + \max(\bar{n}_\alpha - n_\alpha, \bar{n}_\beta - n_\beta)$ almost surely. If both $\bar{n}_\alpha > n_\alpha$ and $\bar{n}_\beta > n_\beta$, it gives rank-deficient $\mathbf{Z}^\intercal\mathbf{Z}$ thus the OLS estimator $\hat{\theta} := (\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal\mathbf{Y}$ is ill-defined. But the pseudo inverse estimator $\hat{\theta}_{pinv} := \mathbf{Z}^\dagger\mathbf{Y}$ still exists. Since any $\tilde{\theta}$ in the true parameter set Θ fits the data perfectly, $\hat{\theta}_{pinv}$ ends up becoming the one with the minimum norm in Θ, i.e., $\hat{\theta}_{pinv} = \arg\min_{\tilde{\theta}\in\Theta}\|\tilde{\theta}\|$. On the bright side, if either $\bar{n}_\alpha = n_\alpha$ *or* $\bar{n}_\beta = n_\beta$, we see $\mathbf{Z}^\intercal\mathbf{Z}$ has full rank, and $\hat{\theta}$ not only exists but also has $\hat{\theta} = \theta$.

These can be explained by Theorem 1 as well: $\sigma_\eta = 0$ gives $\underline{\lambda}_0 = \underline{\sigma}(\mathbf{G}_1)^2\sigma_u^2$, $C_o = 192\max(1, \|\mathbf{B}\|^2)/\underline{\sigma}(\mathbf{G}_1)^2$, and $\|\hat{\theta} - \theta\| \leq \mathcal{O}(\frac{\sigma_\eta}{\underline{\sigma}(\mathbf{G}_1)^2\sigma_u}\sqrt{\frac{\log(T/\rho^{2\bar{n}})}{T\log(1/\rho)}})$. This upper bound is automatically 0 when $\underline{\sigma}(\mathbf{G}_1) \neq 0$, which can be satisfied, according to Proposition 1, when $q(z)$ and $p(z)$ have no common roots and either $\bar{n}_\alpha = n_\alpha$ *or* $\bar{n}_\beta = n_\beta$ is satisfied.

Recall that under the noise-free case, $q(z)$ and $p(z)$ having common roots implies that there exists an equivalent ARX model with orders strictly smaller than n_α and n_β. This means that even if the true orders $\bar{n}_\alpha = n_\alpha$ and $\bar{n}_\beta = n_\beta$ are used in the identification, one may end up with this reduced order model. This ambiguity led by reduced orders can be otherwise eliminated, as long as there exists noise: by Theorem 2, when $\sigma_\eta \neq 0$, $\|\hat{\theta} - \theta\|$ always has a valid error upper bound with decay rate $\mathcal{O}(\log(T)/\sqrt{T})$. In other words, $\hat{\theta}$ converges to θ that embeds the true orders n_α and n_β, even if $q(z)$ and $p(z)$ have common roots. In this sense, the existence of noise helps the identification.

B. Proof Sketch of Theorem 1

Due to space limitation, we only highlight the key steps in this proof sketch and leave the derivation details in [1]. The main challenge in the analysis is the strong dependency among the rows in the data matrix $\mathbf{Z}$. To reduce the dependency in the analysis, we use the idea of sub-sampling and decoupling [35]–[37]: if we sub-sample the trajectory with large enough sampling period, the sub-sampled data has much weaker dependency; then, by ignoring the dependent factors among the sub-sampled data, we construct decoupled "data" that are completely independent. We refer to this as decoupling. The decoupled data is for analysis purpose only and the overall data in the OLS problem can be written as a sum of these decoupled terms plus some residuals. The upper bound on $\|\hat{\theta} - \theta\|$ will mainly come from analyzing these decoupled data, while the impact of the ignored dependent factors can be guaranteed to be small with large sampling period.

For L defined in Table I, it is easy to see $\rho^{\frac{L}{2}} \leq 1/\sqrt{T}$. We will use L as the sampling period, and to ease the analysis, we assume the trajectory length T is a multiple of L, i.e., $K := T/L \in \mathbb{N}$. Similar results can be established for general T.

Sub-trajectory: For all $l \in [L]$, $k \in [K]$, define notation $(l,k) := l + (k-1)L$. Similar to (6) that defines $\mathbf{z}_t, \mathbf{Z}$, and $\mathbf{E}$, we define

$$\begin{aligned}
\mathbf{z}_{(l,k)} &:= \left[y_{(l,k)-1}, \dots, y_{(l,k)-\bar{n}_\alpha}, u_{(l,k)-1}, \dots, u_{(l,k)-\bar{n}_\beta}\right]^\intercal \\
\mathbf{Z}_{(l)} &:= \left[\mathbf{z}_{(l,1)}, \mathbf{z}_{(l,2)}, \dots, \mathbf{z}_{(l,K)}\right]^\intercal \\
\mathbf{E}_{(l)} &:= \left[\eta_{(l,1)-1}, \eta_{(l,2)-1}, \dots, \eta_{(l,K)-1}\right]^\intercal.
\end{aligned}$$

These can be viewed as the variables associated with the lth sub-trajectory. Hence, we have a total of L sub-trajectories each with length K, and (l,k) indexes the kth data in the lth sub-trajectory. These notations also give the relations $\mathbf{Z}^\intercal\mathbf{Z} = \sum_{l=1}^{L}\mathbf{Z}_{(l)}^\intercal\mathbf{Z}_{(l)}$ and $\mathbf{Z}^\intercal\mathbf{E} = \sum_{l=1}^{L}\mathbf{Z}_{(l)}^\intercal\mathbf{E}_{(l)}$.

Decoupling: First note that in $\mathbf{z}_{(l,k)}$, each output element $y_{(l,k)-i}$ can be decomposed as follows using the representation in (4): for all $j \in [\bar{n}_\alpha]$,

$$y_{(l,k)-j} = g_{L-j}^{x}\mathbf{x}_{(l,k)-L} + \sum_{i=1}^{L-j}\left(g_i^u u_{(l,k)-j-i} + g_i^\eta \eta_{(l,k)-j-i}\right).$$

We define the decoupled output and sub-trajectories, which do not include the initial state terms $\mathbf{x}_{(l,k)-L}$:

$$\begin{aligned}
\bar{y}_{(l,k)-j} &:= \sum_{i=1}^{L-j}\left(g_i^u u_{(l,k)-j-i} + g_i^\eta \eta_{(l,k)-j-i}\right) \\
\bar{\mathbf{z}}_{(l,k)} &:= \left[\bar{y}_{(l,k)-1}, \dots, \bar{y}_{(l,k)-\bar{n}_\alpha}, u_{(l,k)-1}, \dots, u_{(l,k)-\bar{n}_\beta}\right]^\intercal \\
\bar{\mathbf{Z}}_{(l)} &:= \left[\bar{\mathbf{z}}_{(l,1)}, \bar{\mathbf{z}}_{(l,2)}, \dots, \bar{\mathbf{z}}_{(l,K)}\right]^\intercal.
\end{aligned}$$

In matrix $\bar{\mathbf{Z}}_{(l)}$, by definition, row $\bar{\mathbf{z}}_{(l,k)}^\intercal$ depends only on $u_{(l,k)-1:(l,k)-L}$ and $\eta_{(l,k)-2:(l,k)-L}$ while $\bar{\mathbf{z}}_{(l,k+1)}^\intercal$ depends only on $u_{(l,k+1)-1:(l,k+1)-L}$ and $\eta_{(l,k+1)-2:(l,k+1)-L}$. Since $\bar{\mathbf{z}}_{(l,k)}$ and $\bar{\mathbf{z}}_{(l,k+1)}$ do not depend on any common u and η

terms, and the mapping relations from their respective u and η terms to $\bar{\mathbf{z}}_{(l,k)}$ and $\bar{\mathbf{z}}_{(l,k+1)}$ are the same, we see $\bar{\mathbf{z}}_{(l,k)}$ and $\bar{\mathbf{z}}_{(l,k+1)}$ are i.i.d.. This shows that the originally dependent rows in matrix $\mathbf{Z}_{(l)}$ are completely decoupled in the newly constructed matrix $\bar{\mathbf{Z}}_{(l)}$, whose rows are i.i.d.. We refer to $\bar{\mathbf{Z}}_{(l)}$ as the decoupled sub-trajectory. Similarly, we see each row $\bar{\mathbf{z}}_{(l,k)}^\intercal$ in $\bar{\mathbf{Z}}_{(l)}$ is independent of the corresponding element $\eta_{(l,k)-1}$ in matrix $\mathbf{E}_{(l)}$.

Bounding $\|\hat{\theta} - \theta\|$: It is easy to verify that the estimation error is given by $\hat{\theta} - \theta = (\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal\mathbf{E}$, which further gives $\|\hat{\theta} - \theta\| \leq \frac{\|\mathbf{Z}^\intercal\mathbf{E}\|}{\underline{\lambda}(\mathbf{Z}^\intercal\mathbf{Z})}$. We can bound these factors in terms of sub-trajectories and their decoupled counterparts:

$$\begin{aligned}\|\mathbf{Z}^\intercal\mathbf{E}\| &\leq \sum_l \|\mathbf{Z}_{(l)}^\intercal\mathbf{E}_{(l)}\| \\ &\leq \sum_l \left(\|\bar{\mathbf{Z}}_{(l)}^\intercal\mathbf{E}_{(l)}\| + \|(\mathbf{Z}_{(l)} - \bar{\mathbf{Z}}_{(l)})^\intercal\mathbf{E}_{(l)}\|\right) \qquad (15)\end{aligned}$$

$$\begin{aligned}\underline{\lambda}(\mathbf{Z}^\intercal\mathbf{Z}) &\geq \sum_l \underline{\lambda}(\mathbf{Z}_{(l)}^\intercal\mathbf{Z}_{(l)}) \\ &\geq \sum_l \left(\underline{\lambda}(\bar{\mathbf{Z}}_{(l)}^\intercal\bar{\mathbf{Z}}_{(l)}) - \|\mathbf{Z}_{(l)}^\intercal\mathbf{Z}_{(l)} - \bar{\mathbf{Z}}_{(l)}^\intercal\bar{\mathbf{Z}}_{(l)}\|\right). \qquad (16)\end{aligned}$$

Hence, to bound $\|\hat{\theta} - \theta\|$, it suffices to lower bound $\underline{\lambda}(\bar{\mathbf{Z}}_{(l)}^\intercal\bar{\mathbf{Z}}_{(l)})$ and upper bound $\|\mathbf{Z}_{(l)}^\intercal\mathbf{Z}_{(l)} - \bar{\mathbf{Z}}_{(l)}^\intercal\bar{\mathbf{Z}}_{(l)}\|$, $\|\bar{\mathbf{Z}}_{(l)}^\intercal\mathbf{E}_{(l)}\|$, and $\|(\mathbf{Z}_{(l)} - \bar{\mathbf{Z}}_{(l)})^\intercal\mathbf{E}_{(l)}\|$.

Since all K rows in matrix $\bar{\mathbf{Z}}_{(l)}$ are i.i.d., bounding $\underline{\lambda}(\bar{\mathbf{Z}}_{(l)}^\intercal\bar{\mathbf{Z}}_{(l)})$ and $\|\bar{\mathbf{Z}}_{(l)}^\intercal\mathbf{E}_{(l)}\|$ is no different from assuming the rows in $\bar{\mathbf{Z}}_{(l)}$ are generated by K i.i.d. trajectories with zero initial conditions. Therefore, following the ideas in [38, Sec 2.2], we can show

$$\underline{\lambda}(\bar{\mathbf{Z}}_{(l)}^\intercal\bar{\mathbf{Z}}_{(l)}) \geq \mathcal{O}(K), \quad \|\bar{\mathbf{Z}}_{(l)}^\intercal\mathbf{E}_{(l)}\| \leq \mathcal{O}(\sqrt{\bar{n}K}). \qquad (17)$$

Now we consider the residual terms $\|\mathbf{Z}_{(l)}^\intercal\mathbf{Z}_{(l)} - \bar{\mathbf{Z}}_{(l)}^\intercal\bar{\mathbf{Z}}_{(l)}\|$ and $\|(\mathbf{Z}_{(l)} - \bar{\mathbf{Z}}_{(l)})^\intercal\mathbf{E}_{(l)}\|$ that capture the dependencies that are removed in the decoupling. Expressing matrix multiplications as the summation of outer products and using triangle inequality give

$$\begin{aligned}&\|\mathbf{Z}_{(l)}^\intercal\mathbf{Z}_{(l)} - \bar{\mathbf{Z}}_{(l)}^\intercal\bar{\mathbf{Z}}_{(l)}\| \leq \\ &\qquad \sum_{k=1}^{K} 2\|\bar{\mathbf{z}}_{(l,k)}\|\|\mathbf{z}_{(l,k)} - \bar{\mathbf{z}}_{(l,k)}\| + \|\mathbf{z}_{(l,k)} - \bar{\mathbf{z}}_{(l,k)}\|^2\end{aligned}$$

$$\|(\mathbf{Z}_{(l)} - \bar{\mathbf{Z}}_{(l)})^\intercal\mathbf{E}_{(l)}\| \leq \sum_{k=1}^{K} \|\mathbf{z}_{(l,k)} - \bar{\mathbf{z}}_{(l,k)}\||\eta_{(l,k)-1}|.$$

Thus, it suffices to bound (i) the error in ignoring state in the decoupling $\|\mathbf{z}_{(l,k)} - \bar{\mathbf{z}}_{(l,k)}\|$, (ii) the decoupled data $\|\bar{\mathbf{z}}_{(l,k)}\|$, and (iii) the noise term $|\eta_{(l,k)-1}|$.

Since the ARX process is stable, i.e. $\rho^\star < 1$, the decoupled data $\bar{\mathbf{z}}_{(l,k)}$ is essentially Gaussian with bounded covariance. Hence, (i) $\|\bar{\mathbf{z}}_{(l,k)}\|$ and (ii) $|\eta_{(l,k)-1}|$ can be bounded with high probability, i.e. $\|\bar{\mathbf{z}}_{(l,k)}\|, |\eta_{(l,k)-1}| \leq \mathcal{O}(1)$.

Finally, we can again exploit stability to bound (iii) $\|\mathbf{z}_{(l,k)} - \bar{\mathbf{z}}_{(l,k)}\|$. By definition,

$$\begin{aligned}\mathbf{z}_{(l,k)} - \bar{\mathbf{z}}_{(l,k)} &= [(g_{L-1}^x)^\intercal, \ldots, (g_{L-\bar{n}_\alpha}^x)^\intercal]^\intercal \cdot \mathbf{x}_{(l,k)-L} \\ &= [(\mathbf{C}\mathbf{A}^{L-1})^\intercal, \ldots, (\mathbf{C}\mathbf{A}^{L-\bar{n}_\alpha})^\intercal]^\intercal \cdot \mathbf{x}_{(l,k)-L}.\end{aligned}$$

Since $\|\mathbf{A}^k\| \leq \mathcal{O}(\rho^k)$ and $\|\mathbf{x}_{(l,k)-L}\| < \mathcal{O}(1)$ by stability, we can show $\|\mathbf{z}_{(l,k)} - \bar{\mathbf{z}}_{(l,k)}\| \leq \mathcal{O}(\rho^{\frac{L}{2}})$. Note that the choice of sub-sampling period L guarantees $\rho^{\frac{L}{2}} \leq 1/\sqrt{T}$, thus $\|\mathbf{z}_{(l,k)} - \bar{\mathbf{z}}_{(l,k)}\| \leq \mathcal{O}(1/\sqrt{T})$.

Combining the bounds for (i-iii), we obtain

$$\|\mathbf{Z}_{(l)}^\intercal\mathbf{Z}_{(l)} - \bar{\mathbf{Z}}_{(l)}^\intercal\bar{\mathbf{Z}}_{(l)}\| \leq \mathcal{O}(K/\sqrt{T}) \qquad (18)$$

$$\|(\mathbf{Z}_{(l)} - \bar{\mathbf{Z}}_{(l)})^\intercal\mathbf{E}_{(l)}\| \leq \mathcal{O}(K/\sqrt{T}). \qquad (19)$$

Finally, plugging (17), (18) (19) into (15) (16) gives

$$\|\mathbf{Z}^\intercal\mathbf{E}\| \leq \mathcal{O}(\sqrt{\bar{n}K}) \quad \underline{\lambda}(\mathbf{Z}^\intercal\mathbf{Z}) \geq \mathcal{O}(K). \qquad (20)$$

Therefore, $\|\hat{\theta} - \theta\| \leq \underline{\lambda}(\mathbf{Z}^\intercal\mathbf{Z})^{-1}\|\mathbf{Z}^\intercal\mathbf{E}\| \leq \mathcal{O}(\sqrt{\bar{n}/K}) = \mathcal{O}(\sqrt{\bar{n}L/T}) = \mathcal{O}(\sqrt{\bar{n}\log(T)/T})$. ■

V. Multiple-trajectory Case

In this section, we present results for the OLS estimator $\hat{\theta}$ computed from N i.i.d. trajectories. Consider the random vector $\mathbf{z}$ of regressors in (7), defined by

$$\mathbf{z} := [y_{T-1}, \ldots, y_{T-\bar{n}_\alpha}, u_{T-1}, \ldots, u_{T-\bar{n}_\beta}]^\intercal \quad \in \mathbb{R}^{\bar{n}}. \qquad (21)$$

Note that $\mathbf{z}$ is a zero-mean Gaussian random vector, and the vectors $\mathbf{z}^{(i)}$ in (7) for all $i \in [N]$ are realizations of $\mathbf{z}$. Denote the covariance of $\mathbf{z}$ by $\mathbf{\Sigma}_\mathbf{z}$. The OLS estimator $\hat{\theta} := (\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal\mathbf{Y}$ is well-defined when $\mathbf{Z}^\intercal\mathbf{Z} = \sum_i \mathbf{z}^{(i)}\mathbf{z}^{(i)\intercal}$ is invertible, which is true only if $\mathbf{\Sigma}_\mathbf{z}$ is positive definite. This is guaranteed by the following proposition.

Proposition 2: The covariance matrix $\mathbf{\Sigma}_\mathbf{z}$ is positive definite if the trajectory length $T \geq \max(\bar{n}_\alpha, \bar{n}_\beta) + 1$ and the standard deviations satisfy $\sigma_\eta > 0$ and $\sigma_u > 0$.

Proof: Since it is assumed in Section II that the system starts from rest, we have $y_t = \sum_{i=1}^t g_i^u u_{t-i} + \sum_{i=1}^t g_i^\eta \eta_{t-i}$ according to (4). Using this relation, the vector $\mathbf{z}$ in (21) can be expressed as the product of a matrix $\mathbf{G} \in \mathbb{R}^{\bar{n}\times(2T-1)}$ constructed with the Markov parameters g_i^u's and g_i^η's, and the vector $[u_0, \ldots, u_{T-1}, \eta_0, \ldots, \eta_{T-2}]^\intercal$. One can verify that $\mathbf{G}$ is given by

$$\mathbf{G} := \left[\begin{array}{c|c|c} 0_{\bar{n}_\alpha\times 1} & \mathbf{G}^u(T-1) & \mathbf{G}^\eta(T-1) \\ \hline \multicolumn{2}{c|}{\mathbf{I}_{\bar{n}_\beta}} & 0_{\bar{n}_\beta\times(2T-1-\bar{n}_\beta)} \end{array}\right], \qquad (22)$$

where $\mathbf{G}^u(T-1)$ and $\mathbf{G}^\eta(T-1)$ are as in (13). Thus, the covariance matrix $\mathbf{\Sigma}_\mathbf{z}$ is equal to

$$\mathbf{\Sigma}_\mathbf{z} = \mathbb{E}[\mathbf{z}\mathbf{z}^\intercal] = \mathbf{G}\begin{bmatrix}\sigma_u^2\mathbf{I}_T & 0 \\ 0 & \sigma_\eta^2\mathbf{I}_{T-1}\end{bmatrix}\mathbf{G}^\intercal. \qquad (23)$$

From the special structure of $\mathbf{G}$, one can see it has full row rank when $T \geq \max(\bar{n}_\alpha, \bar{n}_\beta) + 1$. It follows that matrix $\mathbf{\Sigma}_\mathbf{z}$ is positive definite since $\sigma_u, \sigma_\eta > 0$. ■

With $\mathbf{\Sigma}_\mathbf{z} \succ 0$, using concentration results of Gaussian matrices, e.g. [20, Theorem 6.1], we can show $\mathbf{Z}^\intercal\mathbf{Z}$ is invertible with high probability, which indicates $\hat{\theta}$ is well-defined. The analysis for special cases $\sigma_\eta = 0$ or $\sigma_u = 0$ is similar to the single-trajectory case, thus left to the readers to fill in. We then have the following result regarding the convergence of multiple-trajectory OLS. Its proof can be found in [1].

Theorem 2 (OLS Convergence — Multiple Trajectories): Suppose $T \geq \max(\bar{n}_\alpha, \bar{n}_\beta) + 1$, $\sigma_\eta > 0$, and $\sigma_u > 0$. If the sample size N satisfies that

$$N \geq \max\left(\bar{n}, 4\Big(\sqrt{\mathbf{tr}(\boldsymbol{\Sigma}_\mathbf{z})/\underline{\sigma}(\boldsymbol{\Sigma}_\mathbf{z})} + \sqrt{2\log(1/\delta)}\Big)^2\right), \quad (24)$$

with probability at least $1-\delta$,

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| \leq \frac{16\sigma_\eta\sqrt{(1+\bar{n})\|\boldsymbol{\Sigma}_\mathbf{z}\|\log(18/\delta)}}{\underline{\sigma}(\boldsymbol{\Sigma}_\mathbf{z})\sqrt{N}}. \quad (25)$$

By Theorem 2, the OLS estimator $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}$ with rate $\mathcal{O}(1/\sqrt{N})$, which improves upon the rate $\mathcal{O}(\sqrt{\log(T)/T})$ in the single-trajectory case since i.i.d. trajectories are used. Furthermore, the analysis here no longer requires the system to be stable, i.e., the result holds for arbitrary $\rho^\star$.

The presence of terms involving the covariance matrix $\Sigma_\mathbf{z}$, i.e., $\|\boldsymbol{\Sigma}_\mathbf{z}\|$, $\mathbf{tr}(\boldsymbol{\Sigma}_\mathbf{z})$, and $\underline{\sigma}(\boldsymbol{\Sigma}_\mathbf{z})$, in Theorem 2 can make the bound hard to evaluate. However, more explicit bounds can be obtained by relaxing those terms with more informative quantities. From (23), we can obtain

$$\begin{aligned} \|\boldsymbol{\Sigma}_\mathbf{z}\| \leq& \|\mathbf{G}\|^2 \max(\sigma_u^2, \sigma_\eta^2), \\ \mathbf{tr}(\boldsymbol{\Sigma}_\mathbf{z}) \leq& \sigma_u^2(\bar{n}_\beta + \|\mathbf{G}^u(T-1)\|_\mathrm{F}^2) + \sigma_\eta^2\|\mathbf{G}^\eta(T-1)\|_\mathrm{F}^2, \\ \underline{\sigma}(\boldsymbol{\Sigma}_\mathbf{z}) \geq& \underline{\sigma}^2(\mathbf{G})\min(\sigma_u^2, \sigma_\eta^2). \end{aligned} \quad (26)$$

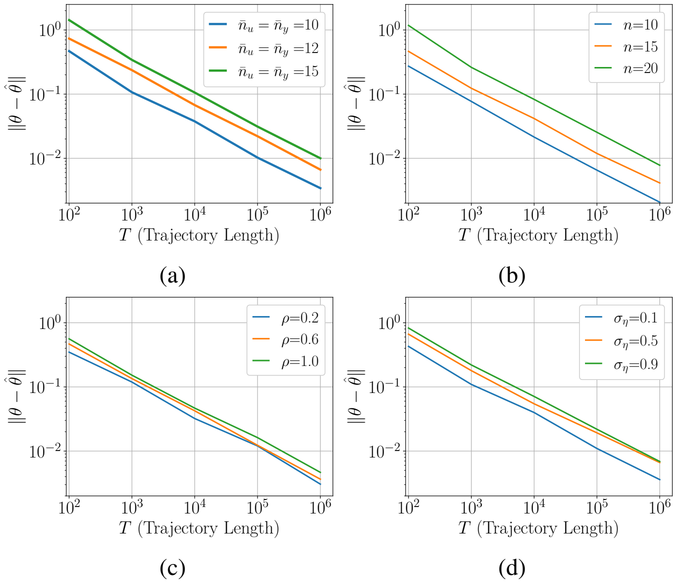
By replacing $\|\boldsymbol{\Sigma}_\mathbf{z}\|$, $\mathbf{tr}(\Sigma_\mathbf{z})$, and $\underline{\sigma}(\boldsymbol{\Sigma}_\mathbf{z})$ in (24) and (25) with their corresponding bounds in (26), we obtain sample complexity results that explicitly depend on σ_u, σ_η, and the matrices $\mathbf{G}$.

VI. EXPERIMENTS

In this section, we present the experimental results on the OLS estimator and compare those results with our theoretical findings.

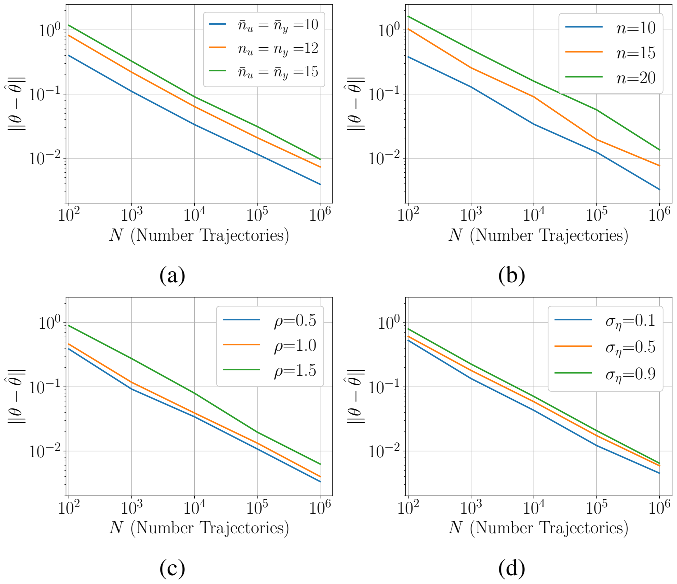
We evaluate the estimation error $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|$ by averaging over 50 Monte Carlo runs. In each run, an ARX model is generated randomly[1] with order $n = n_\alpha = n_\beta$ and spectral radius $\rho^\star$. We set $n = 10$, $\rho^\star = 0.85$, and $\sigma_u^2 = \sigma_\eta^2 = 1$ unless otherwise specified. In the multiple-trajectory case, we fix the trajectory length $T = 50$. Fig. 2 (resp. Fig. 3) shows the averaged estimation error against the trajectory length T in the single-trajectory case (resp. the number of trajectories N in the multiple-trajectory case).

Let us first analyze Fig. 2. In Fig. 2a, we vary the over-parameterization order $\bar{n}_\alpha = \bar{n}_\beta$. We see $\hat{\boldsymbol{\theta}}$ under over-parameterization converges to the true $\boldsymbol{\theta}$, and the performance degrades when $\bar{n}_\alpha = \bar{n}_\beta$ increases, i.e., more redundancy in the parameters. These manifest the self-regularization property and agree with Theorem 1. Since the y-axis has logarithmic scale, the almost-constant distances between the three lines imply that the performance degradation caused by large over-parameterization can be overcome quickly with sufficient amount of data. We also see in Fig. 2b-2d that estimation is harder with larger ground truth order n, $\rho^\star$, and noise level σ_η, which all agree with Theorem 1.



(a) (b) (c) (d)

Fig. 2. Single-trajectory OLS estimation error vs trajectory length vs (a) over-parameterization order $\bar{n}_\alpha, \bar{n}_\beta$; (b) true order n; (c) $\rho^\star$; (d) noise level σ_η. In (b-d), true orders are used, i.e., $\bar{n}_\alpha = n_\alpha, \bar{n}_\beta = n_\beta$.



(a) (b) (c) (d)

Fig. 3. Multiple-trajectory OLS estimation error vs number of trajectories vs (a) over-parameterization order $\bar{n}_\alpha, \bar{n}_\beta$; (b) true order n; (c) $\rho^\star$; (d) noise level σ_η. In (b-d), true orders are used, i.e., $\bar{n}_\alpha = n_\alpha, \bar{n}_\beta = n_\beta$.

In Fig. 3, we repeat similar experiments for learning from multiple trajectories, showing estimation error trends follow the upper bound in Theorem 2. In Theorem 2 we do not require the underlying model to be stable, which is reflected in Fig. 3c where learning is still possible with $\rho^\star = 1.5$.

VII. CONCLUSIONS AND FUTURE WORKS

In this work, we provide a non-asymptotic analysis for using OLS to learn ARX models. A side product of our

[1] The parameters $\beta_{1:n_\beta}$ are generated by i.i.d. standard Gaussian samples. To generate $\alpha_{1:n_\alpha}$, we first sample n_α roots of $q(z)$ following i.i.d. standard Gaussian, followed by uniform scaling so that the largest root has magnitude $\rho^\star$; then the coefficients of $q(z)$ are extracted and taken as $\alpha_{1:n_\alpha}$.

analysis is that even if the OLS is applied to learn an over-parameterized model with orders larger than the true ones, it still converges to the true parameters, which also reveals the true orders. Since our method does not require any regularization, there are no hyper-parameters to be tuned either.

There are a few interesting future directions: (i) One can generalize this analysis framework to vector ARX models. Indeed, linear time-invariant systems in full-observation setting would be a special case of vector ARX model with order one. (ii) It is worth investigating if the self-regularization property holds for estimators other than the OLS. (iii) It would be interesting to extend this analysis to the more general Auto-Regressive Moving Average eXogenous (ARMAX) models, with the main challenge being that the OLS is known to be inconsistent for this class of models.

References

[1] Z. Du, Z. Liu, J. Weitze, and N. Ozay, “Sample complexity analysis and self-regularization in identification of over-parameterized ARX models,” University of Michigan, Tech. Rep., 2022, doi: 10.7302/5876.

[2] G. Michailidis and F. d’Alché Buc, “Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues,” *Mathematical biosciences*, vol. 246, no. 2, pp. 326–334, 2013.

[3] D. P. Burke, S. P. Kelly, P. De Chazal, R. B. Reilly, and C. Finucane, “A parametric feature extraction and classification strategy for brain-computer interfacing,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 13, no. 1, pp. 12–17, 2005.

[4] B. Fetics, E. Nevo, C.-H. Chen, and D. A. Kass, “Parametric model derivation of transfer function for noninvasive estimation of aortic pressure by radial tonometry,” *IEEE Transactions on Biomedical Engineering*, vol. 46, no. 6, pp. 698–706, 1999.

[5] J. C. Wu and F. D. Xia, “Measuring the macroeconomic impact of monetary policy at the zero lower bound,” *Journal of Money, Credit and Banking*, vol. 48, no. 2-3, pp. 253–291, 2016.

[6] C. Hsiao, “Autoregressive modeling of canadian money and income data,” *Journal of the American Statistical Association*, vol. 74, no. 367, pp. 553–560, 1979.

[7] C. Hiemstra and J. D. Jones, “Testing for linear and nonlinear granger causality in the stock price-volume relation,” *The Journal of Finance*, vol. 49, no. 5, pp. 1639–1664, 1994.

[8] L. Lennart, “System identification: theory for the user,” *PTR Prentice Hall, Upper Saddle River, NJ*, vol. 28, p. 540, 1999.

[9] L. Ljung, “Consistency of the least-squares identification method,” *IEEE Transactions on Automatic Control*, vol. 21, no. 5, pp. 779–781, 1976.

[10] T. L. Lai and C. Z. Wei, “Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems,” *The Annals of Statistics*, vol. 10, no. 1, pp. 154–166, 1982.

[11] G. C. Tiao and R. S. Tsay, “Consistency properties of least squares estimates of autoregressive parameters in ARMA models,” *The Annals of Statistics*, pp. 856–871, 1983.

[12] I. Choi, “Asymptotic normality of the least-squares estimates for higher order autoregressive integrated processes with some applications,” *Econometric Theory*, vol. 9, no. 2, pp. 263–282, 1993.

[13] F. J. Breid, R. A. Davis, K.-S. Lh, and M. Rosenblatt, “Maximum likelihood estimation for noncausal autoregressive processes,” *Journal of Multivariate Analysis*, vol. 36, no. 2, pp. 175–198, 1991.

[14] H. Lee and C. Zhang, “Robust guarantees for learning an autoregressive filter,” in *Algorithmic Learning Theory*. PMLR, 2020, pp. 490–517.

[15] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, “Finite-time system identification and adaptive control in autoregressive exogenous systems,” in *Learning for Dynamics and Control*. PMLR, 2021, pp. 967–979.

[16] B. Yu, “Rates of convergence for empirical processes of stationary mixing sequences,” *The Annals of Probability*, pp. 94–116, 1994.

[17] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*. Cambridge University Press, 2012, p. 210–268.

[18] S. Mendelson, “Learning without concentration,” in *Conference on Learning Theory*. PMLR, 2014, pp. 25–39.

[19] V. H. Peña, T. L. Lai, and Q.-M. Shao, *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008.

[20] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019, vol. 48.

[21] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, “Learning without mixing: Towards a sharp analysis of linear system identification,” in *Conference On Learning Theory*. PMLR, 2018, pp. 439–473.

[22] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, “Finite time identification in unstable linear systems,” *Automatica*, vol. 96, pp. 342–353, 2018.

[23] T. Sarkar, A. Rakhlin, and M. A. Dahleh, “Finite time LTI system identification,” *Journal of Machine Learning Research*, vol. 22, pp. 1–61, 2021.

[24] S. Oymak and N. Ozay, “Revisiting Ho-Kalman based system identification: robustness and finite-sample analysis,” *IEEE Transactions on Automatic Control*, 2021.

[25] T. Sarkar and A. Rakhlin, “Near optimal finite time identification of arbitrary linear dynamical systems,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 5610–5618.

[26] Y. Jedra and A. Proutiere, “Finite-time identification of stable linear systems optimality of the least-squares estimator,” in *2020 59th IEEE Conference on Decision and Control (CDC)*, 2020, pp. 996–1001.

[27] E. J. Candes, “Modern statistical estimation via oracle inequalities,” *Acta numerica*, vol. 15, pp. 257–325, 2006.

[28] Y. Sun, S. Oymak, and M. Fazel, “Finite sample system identification: Optimal rates and the role of regularization,” in *Learning for Dynamics and Control*. PMLR, 2020, pp. 16–25.

[29] H. Mania, S. Tu, and B. Recht, “Certainty equivalence is efficient for linear quadratic control,” in *NeurIPS*, 2019.

[30] L. Ljung and B. Wahlberg, “Asymptotic properties of the least-squares method for estimating transfer functions and disturbance spectra,” *Advances in Applied Probability*, vol. 24, no. 2, pp. 412–440, 1992.

[31] A. B. Kock and L. Callot, “Oracle inequalities for high dimensional vector autoregressions,” *Journal of Econometrics*, vol. 186, no. 2, pp. 325–344, 2015.

[32] A. C. Singer, S. S. Kozat, and M. Feder, “Universal linear least squares prediction: Upper and lower bounds,” *IEEE Transactions on Information Theory*, vol. 48, no. 8, pp. 2354–2362, 2002.

[33] R. A. González and C. R. Rojas, “A finite-sample deviation bound for stable autoregressive processes,” in *Learning for Dynamics and Control*. PMLR, 2020, pp. 191–200.

[34] K. C. Wong, Z. Li, and A. Tewari, “Lasso guarantees for β-mixing heavy-tailed time series,” *The Annals of Statistics*, vol. 48, no. 2, pp. 1124–1142, 2020.

[35] Y. Sattar, Z. Du, D. A. Tarzanagh, L. Balzano, N. Ozay, and S. Oymak, “Identification and adaptive control of markov jump systems: Sample complexity and regret bounds,” *arXiv preprint arXiv:2111.07018*, 2021.

[36] Y. Sattar and S. Oymak, “Non-asymptotic and accurate learning of nonlinear dynamical systems,” *arXiv preprint arXiv:2002.08538*, 2020.

[37] S. Oymak, “Stochastic gradient descent learns state equations with nonlinear activations,” in *Conference on Learning Theory*. PMLR, 2019, pp. 2551–2579.

[38] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, “On the sample complexity of the linear quadratic regulator,” *Foundations of Computational Mathematics*, vol. 20, no. 4, pp. 633–679, 2020.