2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

A Tutorial on the Non-Asymptotic Theory of System Identification

Ingvar Ziemann[1], Anastasios Tsiamis[2], Bruce Lee[1], Yassir Jedra[3], Nikolai Matni[1], and George J. Pappas[1]

[1]University of Pennsylvania
[2]ETH Zürich
[3]KTH Royal Institute of Technology

***Abstract*—This tutorial serves as an introduction to recently developed non-asymptotic methods in the theory of—mainly linear—system identification. We emphasize tools we deem particularly useful for a range of problems in this domain, such as the covering technique, the Hanson-Wright Inequality and the method of self-normalized martingales. We then employ these tools to give streamlined proofs of the performance of various least-squares based estimators for identifying the parameters in autoregressive models. We conclude by sketching out how the ideas presented herein can be extended to certain nonlinear identification problems. Note: For reasons of space, proofs have been omitted in this version and are available in an online version: https://arxiv.org/abs/2309.03873.**

Notation

Maxima (resp. minima) of two numbers $a, b \in \mathbb{R}$ are denoted by $a \vee b = \max(a, b)$ ($a \wedge b = \min(a, b)$). For two sequences $\{a_t\}_{t\in\mathbb{Z}}$ and $\{b_t\}_{t\in\mathbb{Z}}$ we introduce the shorthand $a_t \lesssim b_t$ if there exists a universal constant $C > 0$ and an integer t_0 such that $a_t \le Cb_t$ for every $t \ge t_0$. If $a_t \lesssim b_t$ and $b_t \lesssim a_t$ we write $a_t \asymp b_t$. Let $\mathsf{X} \subset \mathbb{R}^d$ and let $f, g \in \mathsf{X} \to R$. We write $f = O(g)$ if $\limsup_{x\to x_0} |f(x)/g(x)| < \infty$, where the limit point x_0 is typically understood from the context. We use $\tilde{O}$ to hide logarithmic factors and write $f = o(g)$ if $\limsup_{x\to x_0} |f(x)/g(x)| = 0$. We write $f = \Omega(g)$ if $\limsup_{x\to x_0} |f(x)/g(x)| > 0$. For an integer N, we also define the shorthand $[N] \triangleq \{1, \dots, N\}$.

Expectation (resp. probability) with respect to all the randomness of the underlying probability space is denoted by $\mathbf{E}$ (resp. $\mathbf{P}$).

The Euclidean norm on $\mathbb{R}^d$ is denoted $\|\cdot\|_2$, and the unit sphere in $\mathbb{R}^d$ is denoted $\mathbb{S}^{d-1}$. The standard inner product on $\mathbb{R}^d$ is denoted $\langle\cdot,\cdot\rangle$. We embed matrices $M \in \mathbb{R}^{d_1\times d_2}$ in Euclidean space by vectorization: $\mathsf{vec}\, M \in \mathbb{R}^{d_1 d_2}$, where vec is the operator that vertically stacks the columns of M (from left to right and from top to bottom). For a matrix M the Euclidean norm is the Frobenius norm, i.e., $\|M\|_F \triangleq \|\mathsf{vec}\, M\|_2$. We similarly define the inner product of two matrices M, N by $\langle M, N\rangle \triangleq \langle \mathsf{vec}\, M, \mathsf{vec}\, N\rangle$. The transpose of a matrix M is denoted by M^T and $\mathrm{tr}\, M$ denotes its trace. For a matrix $M \in \mathbb{R}^{d_1\times d_2}$, we order its singular values $\sigma_1(M), \dots, \sigma_{d_1\wedge d_2}(M)$ in descending order by magnitude. We also write $\|M\|_{\mathsf{op}}$ for its largest singular value: $\|M\|_{\mathsf{op}} \triangleq \sigma_1(M)$. To not carry dimensional notation, we will also use $\sigma_{\min}(M)$ for the smallest nonzero singular value. For square matrices $M \in \mathbb{R}^{d\times d}$ with real eigenvalues, we similarly order the eigenvalues of M in descending order as $\lambda_1(M), \dots, \lambda_d(M)$. In this case, $\lambda_{\min}(M)$ will also be used to denote the minimum (possibly zero) eigenvalue of M. For two symmetric matrices M, N, we write $M \succ N$ ($M \succeq N$) if $M - N$ is positive (semi-)definite.

I. Introduction

Machine learning methods are at an ever increasing pace being integrated into domains that have classically been within the purview of controls. There is a wide range of examples, including perception-based control, agile robotics, and autonomous driving and racing. As exciting as these developments may be, they have been most pronounced on the experimental and empirical sides. To deploy these systems safely, stably, and robustly into the real world, we argue that a principled and integrated theoretical understanding of a) fundamental limitations and b) statistical optimality is needed. Under the past few years, a host of new techniques have been introduced to our field. Unfortunately, existing results in this area are relatively inaccessible to a typical first or second year graduate student in control theory, as they require both sophisticated mathematical tools not typically included in a control theorist's training (e.g., high-dimensional statistics and learning theory).

This tutorial seeks to provide a streamlined exposition of some of these recent advances that are most relevant to the non-asymptotic theory of linear system identification. Our aim is not to be encyclopedic but rather to give simple proofs of the main developments and to highlight and collect the key technical tools to arrive at these results. For a broader—and less technical—overview of the literature we point the reader to our recent survey [31]. It is also worth to point out that the classical literature on system identification has done a formidable job at—often very accurately—characterizing the asymptotic performance of identification algorithms [14]. Our aim is not to supplant this literature but rather to complement the asymptotic picture with finite sample guarantees by relaying recently developed technical tools drawn from high-dimensional probability, statistics and learning theory [35, 37].

A. Problem Formulation

Let us now fix ideas. We are concerned with linear time-series models of the form:

$$Y_t = \theta^\star X_t + V_t \quad t = 1, 2, \dots, T \tag{1}$$

where $Y_{1:T}$ is a sequence of outputs (or targets) assuming values in $\mathbb{R}^{d_{\mathsf{Y}}}$ and $X_{1:T}$ is a sequence of inputs (or covariates) assuming values in $\mathbb{R}^{d_{\mathsf{X}}}$. The goal of the user (or learner) is to recover the a priori unknown linear map $\theta^\star \in \mathbb{R}^{d_{\mathsf{Y}} \times d_{\mathsf{X}}}$ using only the observations $X_{1:T}$ and $Y_{1:T}$. The linear relationship in the regression model (1) is perturbed by a stochastic noise sequence $V_{1:T}$ assuming values in $\mathbb{R}^{d_{\mathsf{Y}}}$. We refer to the regression model (1) as a time-series to emphasize the fact that the observations $X_{1:T}$ and $Y_{1:T}$ may arrive sequentially and in particular that past X_t and Y_t may influence future $X_{t'}$ and $Y_{t'}$ (i.e. with $t' > t$).

a) Example: Autoregressive Models.: For instance, a model class of particular interest to us which is subsumed by (1) are the (vector) autoregressive exogenous models of order p and q (briefly ARX(p, q)):

$$Y_t = \sum_{i=1}^{p} A_i^\star Y_{t-i} + \sum_{j=1}^{q} B_i^\star U_{t-j} + W_t \tag{2}$$

where typically $U_{1:T-1}$ is a sequence of user specified inputs taking values in $\mathbb{R}^{d_{\mathsf{U}}}$ and $W_{1:T}$ is an iid sequence of noise variables taking values in $\mathbb{R}^{d_{\mathsf{W}}}$. If we are only interested in the parameters $\begin{bmatrix} A^\star_{1:p} & B^\star_{1:q} \end{bmatrix}$, we obtain the model (2) by setting

$$X_t = \begin{bmatrix} Y^{\mathsf{T}}_{t-1:t-p} & U^{\mathsf{T}}_{t-1:t-q} \end{bmatrix}^{\mathsf{T}}; \quad \theta^\star = \begin{bmatrix} A^\star_{1:p} & B^\star_{1:q} \end{bmatrix}; \quad V_t = W_t. \tag{3}$$

We point out that that the above discussion presupposes that the order of the model, (p, q), is known (there are ways around this).

In this tutorial we will provide the necessary tools to tackle the following problem.

Problem I.1. *Fix $\varepsilon > 0$, $\delta \in (0, 1)$, and a norm $\|\cdot\|$. Fix also a 'reasonable' estimator $\widehat{\theta}$ of $\theta_\star$ using a sample $(X, Y)_{1:T}$ from (1). We seek to establish finite sample guarantees of the form*

$$\|\widehat{\theta} - \theta^\star\| \le \varepsilon \qquad \text{with probability at least } 1 - \delta \tag{4}$$

where ε controls the accuracy (or rate) and the failure parameter δ controls the confidence.

In the sequel, 'reasonable' estimator will typically mean some form of least squares estimator (7). These are introduced in Section I-B below. A bound of the form (4) is typically thought of as follows. We fix a priori the failure parameter δ and then provide guarantees of the form $\|\widehat{\theta} - \theta^\star\| \le \varepsilon(T, \delta, \mathsf{P}_{XY})$ where P_{XY} is the joint distribution of $(X, Y)_{1:T}$. Hence, the sample size T, the failure probability δ and the distribution of the samples all impact the performance guarantee ε we are able to establish. To be more specific, ε will typically be of the form

$$\varepsilon \propto (\text{Noise Scale}) \times \sqrt{\frac{\text{problem dimension} + \log(1/\delta)}{\text{sample size}}}. \tag{5}$$

Thus in principle, the best possible choice of ε^2 can be thought of as a high probability version of the (inverse) signal-to-noise ratio of the problem at hand. The fact that the confidence parameter δ typically affects (5) additively in $\log(1/\delta)$ is consistent with classical asymptotic normality theory of estimators. One often expects the normalized difference $T^{-1/2}(\widehat{\theta} - \theta^\star)$ to converge in law to a normal distribution [33]. In this tutorial we will provide tools that allow us to match such classical asymptotics but with a finite sample twist. Let us also remark that there often is a minimal requirement on the sample size necessary for a bound of the form (4)-(5) to hold. Such requirements are typically of the form

$$\text{sample size} \gtrsim \text{problem dimension} + \log(1/\delta). \tag{6}$$

Requirements such as (6) are called burn-in times and are related to the notion of persistence of excitation. They correspond to the rather minimal requirement that the parameter identification problem is feasible in the complete absence of observation noise.

B. Least Squares Regression and the Path Ahead

Let us now return to the general setting of (1). Fix a subset M of $\mathbb{R}^{d_{\mathsf{Y}} \times d_{\mathsf{X}}}$, called the model class. The estimator

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta \in \mathsf{M}} \frac{1}{T} \sum_{t=1}^{T} \|Y_t - \theta X_t\|_2^2 \tag{7}$$

is the least squares estimator (LSE) of $\theta^\star$ (with respect to M). Often we simply set $\mathsf{M} = \mathbb{R}^{d_{\mathsf{Y}} \times d_{\mathsf{X}}}$. In this case, equivalently:

$$\widehat{\theta} = \left(\sum_{t=1}^{T} Y_t X_t^{\mathsf{T}} \right) \left(\sum_{t=1}^{T} X_t X_t^{\mathsf{T}} \right)^{\dagger} \tag{8}$$

and the LSE reduces to the (minimum norm) ordinary least squares (OLS) estimator (8).

For simplicity, let us further assume that the (normalized) empirical covariance matrix:

$$\widehat{\Sigma} \triangleq \frac{1}{T} \sum_{t=1}^{T} X_t X_t^{\mathsf{T}}; \tag{9}$$

is full rank almost surely.

a) The Path Ahead.: Let us now briefly sketch the path ahead to solve Problem I.1. If (9) is full rank—as required above—the estimator (8) admits the convenient error representation:

$$\widehat{\theta} - \theta^\star = \left[\left(\sum_{t=1}^{T} V_t X_t^{\mathsf{T}} \right) \left(\sum_{t=1}^{T} X_t X_t^{\mathsf{T}} \right)^{-1/2} \right] \left(\sum_{t=1}^{T} X_t X_t^{\mathsf{T}} \right)^{-1/2}. \tag{10}$$

The leftmost term of (10) (in square brackets) can be shown to be (almost) time-scale invariant in many situations. For instance, if the noise $V_{1:T}$ is a sub-Gaussian martingale difference sequence with respect to the filtration generated by the covariates $X_{1:T}$, one can invoke methods from the theory of self-normalized processes to show this [18, 1]. These methods are the topic of Section IV.

Whenever this is the case, the dominant term in the rate of convergence of the least squares estimator is $\left(\sum_{t=1}^{T} X_t X_t^{\mathsf{T}}\right)^{-1/2}$. In other words, providing control of the smallest eigenvalue of (9) effectively yields control of the rate of convergence of the least squares estimator in many situations. Thus, to analyze the rate of convergence of (7) when $\mathsf{M} = \mathbb{R}^{d_{\mathsf{Y}} \times d_{\mathsf{X}}}$ it suffices to:

- Analyze the smallest eigenvalue (or lower tail) of (9). We provide such analyses in Section III
- Analyze the scale invariant term (in square brackets) of (10). This can in many situations be handled for instance by the self-normalized martingale method described in Section IV.

C. Overview

Before covering these more technical topics in Section III and Section IV, we also briefly review some preliminaries from probability theory in Section II. We then demonstrate how to apply these ideas in the setting of identifying the parameters of an ARX(p, q) model of the form (2) in Section V. An alternative perspective not based on the decomposition (10) for more general least squares algorithms is given in Section VI. We conclude with a brief discussion on how the tools in Section VI can be extended to study more general nonlinear phenomena in Section VII.

II. Preliminaries: Concentration Inequalities, Packing and Covering

Before we proceed to tackle the more advanced question of analyzing the LSE (7), let us discuss a few preliminary inequalities that control the tail of a random variable. Our first inequality is Markov's.

Lemma II.1. *Let X be a nonnegative random variable. For every $s > 0$ we have that*

$$\mathbf{P}(X \geq s) \leq s^{-1}\mathbf{E}[X]. \tag{11}$$

Proof. We have that $\mathbf{E}[X] \geq \mathbf{E}[\mathbf{1}_{X \geq s} X] \geq s\mathbf{E}[\mathbf{1}_{X \geq s}]$. Since $\mathbf{E}[\mathbf{1}_{X \geq s}] = \mathbf{P}(X \geq s)$ the result follows by rearranging. ■

Typically, Markov's inequality itself is insufficient for our goals: we seek deviation inequalities that taper of exponentially fast in s and not as s^{-1}. Such scaling is for instance predicted asymptotically by the central limit theorem by the asymptotic normality of renormalized sums of square integrable iid random variables; that is, sums of the form $S_n/\sqrt{n} = (X_1 + X_2 + \cdots + X_n)/\sqrt{n}$ where the $X_i, i \in [n]$ are independent and square integrable. For random variables possessing a moment generating function, Markov's inequality can be "boosted" by the so-called "Chernoff trick". Namely, we apply Markovs inequality to the moment generating function of the random variable instead of applying it directly to the random variable itself.

Corollary II.1 (Chernoff)**.** *Fix $s > 0$ and suppose that $\mathbf{E}\exp(\lambda X)$ exists. Then*

$$\mathbf{P}(X \geq s) \leq \min_{\lambda \geq 0} e^{-\lambda s}\mathbf{E}\exp(\lambda X). \tag{12}$$

Proof. Fix $\lambda \geq 0$. We have:

$$\begin{aligned}\mathbf{P}(X \geq s) &= \mathbf{P}(\exp(\lambda X) \geq \exp(\lambda s)) && \text{(mono: } x \mapsto e^{\lambda x}\text{)}\\ &\leq e^{-\lambda s}\mathbf{E}\exp(\lambda X) && \text{(Markov).}\end{aligned}$$

The result follows by optimizing. ■

Recall that the function $\psi_X(\lambda) \triangleq \mathbf{E}\exp(\lambda X)$ is the moment generating function of X. For instance, if X has univariate Gaussian distribution with mean zero and variance σ^2, the moment generating function appearing in (12) is just $\mathbf{E}\exp(\lambda X) = \exp\left(\lambda^2\sigma^2/2\right)$. Hence the probability that said Gaussian exceeds s is upper-bounded:

$$\mathbf{P}(X > s) \leq \min_{\lambda \geq 0} e^{-\lambda s}\exp\left(\lambda^2\sigma^2/2\right) = \exp\left(\frac{-s^2}{2\sigma^2}\right) \tag{13}$$

which (almost) exhibits the correct Gaussian tails as compared to (11).[1] It should be pointed out that assumptions stronger than those of the Central Limit Theorem (finite variance) are indeed needed for a non-asymptotic theory with sub-Gaussian tails as in (13). An assumption of this kind which is relatively standard in the literature is introduced next.

A. Sub-Gaussian Concentration and the Hanson-Wright Inequality

In the sequel, we will not want to impose the Gaussian assumption. Instead, we define a class of random variables that admit reasoning analogous to (13).

Definition II.1. *We say that a random vector W taking values in $\mathbb{R}^d$ is σ^2-sub-Gaussian (σ^2-subG) if for every $v \in \mathbb{R}^d$ we have that:*

$$\mathbf{E}\exp(\langle v, W\rangle) \leq \exp\left(\frac{\sigma^2\|v\|^2}{2}\right). \tag{14}$$

Similarly, we say that W is σ^2-conditionally sub-Gaussian with respect to a σ-field $\mathcal{F}$ if (14) holds with $\mathbf{E}[\cdot]$ replaced by $\mathbf{E}[\cdot|\mathcal{F}]$.

The term σ^2 appearing in (14) is called the variance proxy of a sub-Gaussian random variable. The significance of this definition is that the one-dimensional projections $X = \langle v, W\rangle$ (with $\|v\| = 1$) satisfy the tail inequality (13). While obviously Gaussian random variables are sub-Gaussian with their variance as variance-proxy, there are many examples beyond Gaussians that fit into this framework. It is for instance straightforward to show that bounded random

[1] We write almost because $\exp(-s^2/2\sigma^2) \approx \mathbf{P}(V > s)$ where $V \sim N(0, \sigma^2)$ but the expression is not exact.

variables have variance proxy proportional to the square of their width [see eg. 37, Examples 2.3 and 2.4]. Moreover, it is readily verified that the normalized sum mentioned above—$S_n/\sqrt{n} = (X_1 + \cdots + X_n)/\sqrt{n}$—satisfies the same bound (13) provided that the entries of $X_{1:n}$ are independent, mean zero and σ^2-sub-Gaussian. To see this, notice that the moment generating function "tensorizes" across products. Namely, for every $\lambda \in \mathbb{R}$:

$$\begin{aligned}\mathbf{E}\exp\left(\frac{\lambda}{\sqrt{n}}\sum_{i=1}^{n} X_i\right) &= \prod_{i=1}^{n}\mathbf{E}\exp\left(\frac{\lambda}{\sqrt{n}}X_i\right) \\ &\leq \prod_{i=1}^{n}\exp\left(\frac{\lambda^2\sigma^2}{2n}\right) = \exp\left(\frac{\lambda^2\sigma^2}{2}\right).\end{aligned}$$

Hence, by the exact same reasoning leading up to (13) such normalized sub-Gaussian sums satisfy the same tail bound (13).

When analyzing linear regression models, most variables of interest are typically either linear or quadratic in the variables of interest (cf. (10)). Hence, we also need to understand how squares of sub-Gaussian random variables behave. The next result shows that sub-Gaussian quadratic forms exhibit similar tail behavior to the Chi-squared distribution (often in the literature referred to as sub-exponential tails). It is known as the Hanson-Wright Inequality.

Theorem II.1 ([6, 22]). *Let $M \in \mathbb{R}^{d\times d}$. Fix a random variable $W = W_{1:d}$ where each $W_i, i \in [d]$ is a scalar, mean zero and independent σ^2-sub-Gaussian random variable. Then for every $s \in [0,\infty)$:*

$$\begin{aligned}&\mathbf{P}\left(|W^{\mathsf{T}}MW - \mathbf{E}W^{\mathsf{T}}MW| > s\right) \\ &\leq 2\exp\left(-\min\left(\frac{s^2}{144\sigma^4\|M\|_F^2}, \frac{s}{16\sqrt{2}\sigma^2\|M\|_{\mathsf{op}}}\right)\right).\end{aligned}$$

The proof of Theorem II.1 is rather long and technical and thus relegated to the appendix. There, the reader may also find further useful concentration inequalities for quadratic forms in sub-Gaussian variables. In fact, there are plethora of useful concentration inequalities not covered here and the interested reader is urged to consult the first few chapters of [35].

B. Covering and Discretization Arguments

We will often find ourselves in a situation where it is possible to obtain a scalar concentration bound but need this to hold uniformly for many random variables at once. The ε-net argument, which proceeds via the notion of covering numbers, is a relatively straightforward way of converting concentration inequalities for scalars into their counterparts for vectors, matrices and functions more generally.

The reader will for instance notice that the quantity being controlled by Theorem II.1 is a scalar quadratic form in sub-Gaussian random variables. By contrast, the empirical covariance matrix (9) is a matrix and so a conversion step is needed. This idea will be used frequently and in various forms throughout the manuscript, so we review it briefly here for the particular case of controlling the operator norm of a random matrix. To this end, we notice that for any matrix $M \in \mathbb{R}^{m\times d}$:

$$\|M\|_{\mathsf{op}}^2 = \max_{v\in\mathbb{S}^{d-1}} \langle Mv, Mv\rangle. \tag{15}$$

Hence, the operator norm of a random matrix is a maximum of scalar random variables indexed by the unit sphere $\mathbb{S}^{d-1}$.

Recall now that the union bound states that the probability that the maximum of a *finite collection* ($|S| < \infty$) $\{X_i\}_{i\in S}$ of random variables exceeds a certain threshhold can be bounded by the sum of their probabilities:

$$\mathbf{P}\left(\max_{i\in S} X_i > t\right) \leq \sum_{i\in S}\mathbf{P}\left(X_i > t\right). \tag{16}$$

Unfortunately, the unit sphere appearing (15) is not a finite set and so the union bound (16) cannot be directly applied. However, when the domain of optimization has geometric structure, one can often exploit this to leverage the union bound not directly but rather in combination with a discretization argument. Returning to our example of the operator norm of a matrix, the set S appearing in (16) will be a discretized version of the unit sphere $\mathbb{S}^{d-1}$.

The following notion is key.

Definition II.2. *Let (X, d) be a compact metric space and fix $\varepsilon > 0$. A subset $\mathcal{N}$ of X is called an ε-net of X if every point of X is within radius ε of a point of $\mathcal{N}$:*

$$\sup_{x\in\mathsf{X}}\inf_{x'\in\mathcal{N}} d(x,x') \leq \varepsilon. \tag{17}$$

Moreover, the minimal cardinality of $\mathcal{N}$ necessary such that (17) holds is called the covering number at resolution ε of (X, d) and is denoted $\mathcal{N}(\varepsilon, \mathsf{X}, d)$.

We will not explore this notion in full, but simply content ourselves to note that it plays very nicely with the notion of operator norm.

Lemma II.2 (Lemma 4.4.1 in [35]). *Let $M \in \mathbb{R}^{m\times d}$ and let $\varepsilon \in (0,1)$. Then for any ε-net $\mathcal{N}$ of $(\mathbb{S}^{d-1}, \|\cdot\|_2)$ we have that:*

$$\|M\|_{\mathsf{op}} \leq \frac{1}{1-\varepsilon}\sup_{v\in\mathcal{N}}\|Mv\|_2. \tag{18}$$

Hence at a small multiplicative cost, the computation of the operator norm can be restricted to the discretized sphere $\mathcal{N}$. Our intention is now to apply the union bound (16) to the right hand side of (18). To do so, we also need control of the size (cardinality) of the ε-net.

Lemma II.3 (Corollary 4.2.13 in [35]). *For any $\varepsilon > 0$ the covering numbers of $\mathbb{S}^{d-1}$ satisfy*

$$\mathcal{N}(\varepsilon, \mathbb{S}^{d-1}, \|\cdot\|) \leq \left(1 + \frac{1}{2\varepsilon}\right)^d. \tag{19}$$

We now provide two instances of this covering argument combined with the union bound. The second of these uses an

alternative variational characterization of the operator norm but otherwise similar ideas.

Lemma II.4. *Let M be an $m \times d$ random matrix, and $\epsilon \in (0,1)$. Furthermore, let $\mathcal{N}$ be an ϵ-net of $\mathbb{S}^{d-1}$ of minimal cardinality. Then for all $\rho > 0$, we have*

$$\mathbf{P}\left(\|M\|_{\mathsf{op}} > \rho\right) \leq \left(\frac{2}{\epsilon}+1\right)^d \max_{v\in\mathcal{N}} \mathbf{P}\left(\|Mv\|_2 > (1-\epsilon)\rho\right).$$

Lemma II.5. *Let M be an $d \times d$ symmetric random matrix, and let $\epsilon \in (0,1/2)$. Furthermore, let $\mathcal{N}$ be an ϵ-net of $\mathbb{S}^{d-1}$ with minimal cardinality. Then for all $\rho > 0$, we have*

$$\mathbf{P}\left(\|M\|_{\mathsf{op}} > \rho\right) \leq \left(\frac{2}{\epsilon}+1\right)^d \max_{v\in\mathcal{N}} \mathbf{P}\left(|v^\top M v| > (1-2\epsilon)\rho\right).$$

Lemma II.4 and Lemma II.5 exploit two different variational forms of the operator norm. Namely for any M we have that $\|M\|_{\mathsf{op}}^2 = \sup_{v\in\mathbf{S}^{d-1}} \|Mv\|^2$ and in addition, when M is symmetric we also have, $\|M\|_{\mathsf{op}} = \sup_{v\in S^{d-1}} |v^\top M v|$. The proof of these last two lemmas are standard and can be found for example in [35, Chapter 4].

C. Concentration of the Covariance Matrix of Linear Systems

To not get lost in the weeds, let us provide an example showcasing the use of Theorem II.1 due to [8]. Recall that the matrix $\widehat{\Sigma}$ appearing in (9) is crucial to the performance of the least squares estimator. We will now see that this matrix is well-conditioned when we consider stable first order auto-regressions of the form:

$$X_{t+1} = A^\star X_t + W_t \qquad t = 1,\dots,T \tag{20}$$

taking values in $\mathbb{R}^{d_{\mathsf{X}}}$ and with $W_{1:T}$ iid isotropic and K^2-subG. By stable we mean that the spectrum of $A^\star$ is contained in the unit disc.

The following result is a consequence of the Hanson-Wright inequality together with the discretization strategy outlined in Section II-B.

Theorem II.2. *Let $\varepsilon > 0$ and set $M \triangleq \left(\sum_{t=1}^T \sum_{k=0}^{t-1} (A^\star)^k (A^{\star,\mathsf{T}})^k\right)^{-\frac{1}{2}}$. Let also $\mathbf{L}$ be the linear operator such that $X_{1:T} = \mathbf{L} W_{1:T}$. Then simultaneously for every $i \in [d_{\mathsf{X}}]$:*

$$(1-\varepsilon)^2 \lambda_{\min}\left(\sum_{t=1}^T \sum_{k=0}^{t-1} (A^\star)^k (A^{\star,\mathsf{T}})^k\right) \leq \lambda_i\left(\sum_{t=1}^T X_t X_t^\mathsf{T}\right) \leq (1+\varepsilon)^2 \lambda_{\max}\left(\sum_{t=1}^T \sum_{k=0}^{t-1} (A^\star)^k (A^{\star,\mathsf{T}})^k\right)$$

holds with probability at least

$$1 - \exp\left(-\frac{\varepsilon^2}{576\, K^2 \|M\|_{\mathsf{op}}^2 \|\mathbf{L}\|_{\mathsf{op}}^2} + d_{\mathsf{X}} \log(18)\right). \tag{21}$$

Put differently, on the same event as in Theorem II.2, the spectrum of

$$\widehat{\Sigma} = \frac{1}{T}\sum_{t=1}^T X_t X_t^\mathsf{T} \tag{22}$$

is sandwiched by its population counterpart within a small multiplicative factor. The result holds with high probability for strictly stable systems.

The quantity $\|L\|_{\mathsf{op}}$ in (21) grows very quickly as the spectral radius of $A^\star$ tends to 1; Theorem II.2 becomes vacuous in the marginally stable regime. It turns out that requirement of two-sided concentration—the sandwiching of the entire spectrum—is too stringent a requirement to obtain bounds that degrade gracefully with the stability of the system. Fortunately, we only need sharp control of the lower half of the spectrum to control the error (10). This motivates Section III below, in which we will see how to relax the stability assumption and analyze more general linear systems.

D. Notes

The basic program carried out in Section II-C can be summarized as follows: (1) introduce a discretization of the problem considered—for matrices this is typically a discretization of the unit sphere; (2) prove an exponential inequality for a family of scalar random variables corresponding to one-dimensional projection of the discretization—in our case: prove bounds on the moment generating function of quadratic forms in random matrices; and (3) conclude to obtain a uniform bound by using the union bound across the discretization. This roughly summarizes the proof of Theorem II.2. These tools are thematic throughout this manuscript.

III. The Lower Spectrum of the Empirical Covariance

Recall that our outline of the analysis of the least squares estimator in Section I-B consists of two main components, one of which being the lower tail of the empirical covariance matrix (9). In this section we provide a self-contained analysis of this random matrix for a class of "causal" systems. Moreover, we will emphasize only the lower tail of this random matrix as to sidestep issues with bounds degrading with the stability of the system considered. This allows us to quantitatively separate the notions of persistence of excitation and stability.

Let us now carry out this program. Fix two integers T and k such that $T/k \in \mathbb{N}$. We consider causal processes of the form $X_{1:T} = (X_1^\mathsf{T}, \dots, X_T^\mathsf{T})^\mathsf{T}$ evolving on $\mathbb{R}^d$. More precisely, we assume the existence of an isotropic sub-Gaussian process evolving on $\mathbb{R}^p$, $W_{1:T}$ with $\mathbf{E} W_{1:T} W_{1:T}^\mathsf{T} = I_{pT}$ and a (block-) lower-triangular matrix $\mathbf{L} \in \mathbb{R}^{dT \times pT}$ such that

$$X_{1:T} = \mathbf{L} W_{1:T}. \tag{23}$$

We will assume that all the pT-many entries of $W_{1:T}$ are independent K^2-sub-Gaussian for some positive $K \in \mathbb{R}$.

We say that $X_{1:T}$ is k-causal if the matrix $\mathbf{L}$ has the block lower-triangular form:

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}_{1,1} & 0 & 0 & 0 & 0 \\ \mathbf{L}_{2,1} & \mathbf{L}_{2,2} & 0 & 0 & 0 \\ \mathbf{L}_{3,1} & \mathbf{L}_{3,2} & \mathbf{L}_{3,3} & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{L}_{T/k,1} & \dots & \dots & \dots & \dots \mathbf{L}_{T/k,T/k} \end{bmatrix} = \begin{bmatrix} \mathbf{L}_1 \\ \mathbf{L}_2 \\ \mathbf{L}_3 \\ \vdots \\ \mathbf{L}_{T/k} \end{bmatrix} \tag{24}$$

where each $\mathbf{L}_{ij} \in \mathbb{R}^{dk \times pk}, i, j \in [T/k] \triangleq \{1, 2, \dots, T/k\}$. In brief, we say that $X_{0:T-1}$ satisfying the above construction is k-causal with independent K^2-sub-Gaussian increments.

Obviously, every 1-causal process is k-causal for every $k \in \mathbb{N}$ as long as the divisibility condition holds. To analyze the lower tail of the empirical covariance of $X_{0:T-1}$ we will also associate a decoupled random process

$$\tilde{X}_{1:T} = \mathrm{blkdiag}(\mathbf{L}_{11}, \dots, \mathbf{L}_{T/k,T/k}) W_{1:T}.$$

Hence, the process $\tilde{X}_{1:T}$ is generated in much the same way as $X_{1:T}$ but by removing the sub-diagonal entries of $\mathbf{L}$:

$$\tilde{\mathbf{L}} \triangleq \begin{bmatrix} \mathbf{L}_{1,1} & 0 & 0 & 0 \\ 0 & \mathbf{L}_{2,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \mathbf{L}_{T/k,T/k} \end{bmatrix} \implies \tilde{X}_{1:T} = \tilde{\mathbf{L}} W_{1:T}.$$

We emphasize that by our assumptions on $W_{1:T}$ and the block-diagonal structure of $\tilde{\mathbf{L}}$ the variables $\tilde{X}_{1:k}, \tilde{X}_{k+1:2k}, \dots, \tilde{X}_{T-k+1:T}$ are all independent of each other; they have been decoupled. This decoupled process will effectively dictate our lower bound, and we will show under relatively mild assumptions that

$$\lambda_{\min}\left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t^{\mathsf{T}}\right) \gtrsim \lambda_{\min}\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{E}\tilde{X}_t \tilde{X}_t^{\mathsf{T}}\right) \tag{25}$$

with probability that approaches 1 at an exponential rate in the sample size T. More precisely, the following statement is the main result of this section.

Theorem III.1. *Fix an integer $k \in \mathbb{N}$, let $T \in N$ be divisible by k and suppose $X_{1:T}$ is a k-causal process taking values in $\mathbb{R}^d$ with K^2-sub-Gaussian increments. Suppose further that the diagonal blocks are all equal: $\mathbf{L}_{j,j} = \mathbf{L}_{1,1}$ for all $j \in [T/k]$. Suppose $\lambda_{\min}\left(\sum_{t=1}^{T} \mathbf{E}\tilde{X}_t\tilde{X}_t^{\mathsf{T}}\right) > 0$. We have that:*

$$\mathbf{P}\left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t^{\mathsf{T}} \not\succeq \frac{1}{8T}\sum_{t=1}^{T} \mathbf{E}\tilde{X}_t\tilde{X}_t^{\mathsf{T}}\right) \leq (C_{\mathsf{sys}})^d \exp\left(-\frac{T}{576K^2k}\right) \tag{26}$$

where

$$C_{\mathsf{sys}} \triangleq 1 + 2\sqrt{2}\frac{\left(\frac{T\|\mathbf{L}\mathbf{L}^{\mathsf{T}}\|_{\mathsf{op}}}{18k\lambda_{\min}\left(\sum_{t=1}^{T}\mathbf{E}X_tX_t^{\mathsf{T}}\right)} + 9\right)\lambda_{\max}\left(\sum_{t=1}^{T}\mathbf{E}X_tX_t^{\mathsf{T}}\right)}{\lambda_{\min}\left(\sum_{t=1}^{T}\mathbf{E}\tilde{X}_t\tilde{X}_t^{\mathsf{T}}\right)}. \tag{27}$$

To parse Theorem III.1, note that it simply informs us that there exist a a system-dependent constant C_{sys}—which itself has no more than polynomial dependence on relevant quantities—such that if

$$T/k \geq 576K^2(d\log C_{\mathsf{sys}} + \log(1/\delta)) \tag{28}$$

then on an event with probability mass at least $1-\delta$:

$$\frac{1}{T}\sum_{t=1}^{T} X_t X_t^{\mathsf{T}} \succeq \frac{1}{8T}\sum_{t=1}^{T} \mathbf{E}\tilde{X}_t\tilde{X}_t^{\mathsf{T}}.$$

Remark III.1. *Since the blocks of $\mathbf{L}$ can be regarded to specify the noise-to-output map, the assumption that the diagonal blocks are constant is for instance satisfied by linear time-invariant (LTI) systems. The assumption can be removed at the cost of a more complicated expression.*

The next example serves as the archetype for the reduction from $\mathbf{L}$ to $\tilde{\mathbf{L}}$.

Example III.1. *Suppose that (23) is specified via*

$$X_t = A_\star X_{t-1} + B_\star W_t \tag{29}$$

for $t \in [T]$ and where $(A_\star, B_\star) \in \mathbb{R}^{d_{\mathsf{X}} \times d_{\mathsf{X}} + d_{\mathsf{X}} \times d_{\mathsf{W}}}$. We set $d = d_{\mathsf{X}}$ and $p = d_{\mathsf{W}}$ in the theorem above. The reduction from $X_{1:T} = \mathbf{L}W_{1:T}$ to $\tilde{X}_{1:T} = \mathrm{blkdiag}(\mathbf{L}_{11}, \dots, \mathbf{L}_{T/k,T/k})W_{1:T}$ corresponds to replacing a single trajectory from the linear system (29) of length T by T/k trajectories of length k each and sampled independently of each other. The price we pay for decoupling these systems is that our lower bound is dictated by the gramians up to range k:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{E}\tilde{X}_t\tilde{X}_t^{\mathsf{T}} = \frac{1}{k}\sum_{t=1}^{k}\mathbf{E}\tilde{X}_t\tilde{X}_t^{\mathsf{T}} = \frac{1}{k}\sum_{t=1}^{k}\sum_{j=0}^{t-1}(A^\star)^j B^\star B^{\star,\mathsf{T}}(A^{\star,\mathsf{T}})^j$$

instead of the gramians up to range T:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{E}X_tX_t^{\mathsf{T}} = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=0}^{t-1}(A^\star)^j B^\star B^{\star,\mathsf{T}}(A^{\star,\mathsf{T}})^j.$$

Put differently, the reduction from $\mathbf{L}$ to $\tilde{\mathbf{L}}$ can be thought of as restarting the system every k steps.

Comparing with Theorem II.2, the advantage of Theorem III.1 is that it allows us to provide persistence-of-excitation type guarantees that do not rely strongly on the stability of the underlying system. While Theorem II.2 gives in

principle stronger two-sided concentration results, it comes at the cost of the guarantees becoming vacuous as the spectral radius of $A^\star$ in Example III.1 tends to marginal stability (tends to 1). By contrast, Theorem III.1 does not exhibit such a blow-up since the dependence on C_{sys} in (28) is logarithmic (instead of polynomial). The distinction might seem small, but it is qualitatively important as it (almost) decouples the phenomena of stability and persistence of excitation.

A. A Decoupling Inequality for sub-Gaussian Quadratic Forms

Our proof of Theorem III.1 will make heavy use of Proposition III.1 below. This is the crucial probalistic inequality that allows us to decouple—or restart as discussed in Example III.1.

Proposition III.1. *Fix $K \geq 1$, $x \in \mathbb{R}^n$ and a symmetric positive semidefinite $Q \in \mathbb{R}^{(n+m)\times(n+m)}$ of the form $Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$ with $Q_{22} \succ 0$. Let W be an m-dimensional mean zero, isotropic and K^2-sub-Gaussian random vector with independent entries. Then for every $\lambda \in \left[0, \frac{1}{8\sqrt{2}K^2\|Q_{22}\|_{\mathsf{op}}}\right]$ it holds true that:*

$$\mathbf{E} \exp\left(-\lambda \begin{bmatrix} x \\ W \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} x \\ W \end{bmatrix}\right) \leq \exp\left(-\lambda \operatorname{tr} Q_{22} + 36K^4\lambda^2 \operatorname{tr} Q_{22}^2\right). \quad (30)$$

By combining Lemma III.1 below with the exponential form of Hanson-Wright we obtain the exponential inequality (30), which in the sequel will allow us to control the lower tail of the conditionally random quadratic form

$$\begin{bmatrix} x \\ W \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} x \\ W \end{bmatrix}.$$

We point out that (30) is not the best possible if the entries of are W independent and Gaussian as opposed to just isotropic and sub-Gaussian. In this case, the factor $36K^4\lambda^2(\operatorname{tr} Q_{22})^2$ in (30) can be improved to $\frac{\lambda^2}{2} \operatorname{tr} Q_{22}^2$ and the inequality can be shown to hold for the entire range of non-negative λ [38, Lemma 2.1]. Irrespectively, we will see in the sequel that it captures the correct qualitative behavior.

Lemma III.1 (sub-Gaussian Decoupling)**.** *Fix $K \geq 1$, $x \in \mathbb{R}^n$ and a symmetric positive semidefinite $Q \in \mathbb{R}^{(n+m)\times(n+m)}$ of the form $Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$. Let W be an m-dimensional mean zero and K^2-sub-Gaussian random vector. Then for every $\lambda \in \left[0, \frac{1}{4K^2\|Q_{22}\|_{\mathsf{op}}}\right]$ it holds true that:*

$$\mathbf{E} \exp\left(-\lambda \begin{bmatrix} x \\ W \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} x \\ W \end{bmatrix}\right) \leq \sqrt{\mathbf{E} \exp\left(-2\lambda W^{\mathsf{T}} Q_{22} W\right)}. \quad (31)$$

Once equipped with (31), Proposition III.1 follows immediately.

B. The Lower Tail of the Empirical Covariance of Causal sub-Gaussian Processes

Repeated application of Proposition III.1 to the process $X_{1:T} = \mathbf{L}W_{1:T}$ in combination with the tower property of conditional expectation yields the following exponential inequality that controls the lower tail of (9) in any fixed direction.

Theorem III.2. *Fix an integer $k \in \mathbb{N}$, let $T \in N$ be divisible by k and suppose $X_{1:T}$ is a k-causal process driven by independent K^2-sub-Gaussian increments as described in Section III. Fix also a matrix $\Delta \in \mathbb{R}^{d'\times d}$. Let $Q_{\max} \triangleq \max_{j\in[T/k]} \|\mathbf{L}_{j,j}^{\mathsf{T}} \operatorname{blkdiag}(\Delta^{\mathsf{T}}\Delta)\mathbf{L}_{j,j}\|_{\mathsf{op}}$ Then for every $\lambda \in \left[0, \frac{1}{8\sqrt{2}K^2 Q_{\max}}\right]$:*

$$\mathbf{E} \exp\left(-\lambda \sum_{t=1}^{T} \|\Delta X_t\|_2^2\right) \leq \exp\left(-\lambda \sum_{j=1}^{T/k} \operatorname{tr}\left(\mathbf{L}_{j,j}^{\mathsf{T}} \operatorname{blkdiag}(\Delta^{\mathsf{T}}\Delta)\mathbf{L}_{j,j}\right) + 36K^4\lambda^2 \sum_{j=1}^{T/k} \operatorname{tr}\left(\mathbf{L}_{j,j}^{\mathsf{T}} \operatorname{blkdiag}(\Delta^{\mathsf{T}}\Delta)\mathbf{L}_{j,j}\right)^2 \right).$$

To appreciate the terms appearing in Theorem III.2, it is worth to point out that

$$\sum_{j=1}^{T/k} \operatorname{tr}\left(\mathbf{L}_{j,j}^{\mathsf{T}} \operatorname{blkdiag}(\Delta^{\mathsf{T}}\Delta)\mathbf{L}_{j,j}\right) = \sum_{t=1}^{T} \mathbf{E}\|\Delta \tilde{X}_t\|_2^2.$$

Hence Theorem III.2 effectively passes the expectation inside the exponential at the cost of working with the possibly less excited process $\tilde{X}_{1:T}$ and a quadratic correction term. Note also that the assumption that T is divisible by k is not particularly important. If not, let T' be the largest integer such that $T'/k \in \mathbb{N}$ and $T' \leq T$ and apply the result with T' in place of T.

The significance of Theorem III.2 is demonstrated by the following simple observation, which is just the Chernoff approach applied to the exponential inequality in Theorem III.2.

Lemma III.2. *Fix an integer $k \in \mathbb{N}$, let $T \in N$ be divisible by k and suppose $X_{1:T}$ is a k-causal process with independent K^2-sub-Gaussian increments. Suppose further that the diagonal blocks are all equal: $\mathbf{L}_{j,j} = \mathbf{L}_{1,1}$ for all $j \in [T/k]$. For every size-conforming matrix Δ we have that:*

$$\mathbf{P}\left(\sum_{t=1}^{T} \|\Delta X_t\|_2^2 \leq \frac{1}{2}\sum_{t=1}^{T} \mathbf{E}\|\Delta \tilde{X}_t\|_2^2\right) \leq \exp\left(-\frac{T}{576K^2k}\right). \quad (32)$$

Note that Lemma III.2 only yields *pointwise* control of the empirical covariance—i.e. pointwise on the sphere $\mathbb{S}^{d-1}$. By setting $\Delta = v \in \mathbb{S}^{d-1}$, the result holds for a fixed vector on the sphere, but not uniformly for all such vectors at once. Thus, returning to our over-arching goal of providing control

of the smallest eigenvalue of the empirical covariance matrix (9), we now combine (32) (using $d' = 1$) with a union bound. This approach yields Theorem III.1.[2]

C. Notes

In this manuscript we have chosen a perhaps less well-known but conceptually simpler approach to establishing lower bounds on the empirical covariance matrix Equation (25). The first proof of a statement similar to Theorem III.1 is due to [27] which in turn relies on a more advanced notion from probability theory known as the small-ball method, due to [15]. The emphasis therein is on anti-concentration—which can hold under milder moment assumptions—rather than concentration. However, the introduction of this tool is not necessary for Gaussian (or sub-Gaussian) system identification. For instance, Sarkar and Rakhlin [24] leverage the method of self-normalized martingales introduced in Section IV below.

Our motivation for providing a different proof is to streamline the exposition as to fit control of the lower tail into the "standard machinery", which roughly consists of: (1) prove a family of scalar exponential inequalities, (2) invoke the Chernoff method, and (3) conclude by a discretization argument and a union bound to port the result from scalars to matrices. Our proof here follows this outline and emphasizes the exponential inequality in Theorem III.2. We finally remark that the proof presented here is new to the literature and extends a result in [38] from the Gaussian setting to the sub-Gaussian setting.

IV. Self-Normalized Martingale Bounds

The objective in this section is to bound the operator and Frobenius norms of the self-normalized term of (10):

$$\left(\sum_{t=1}^{T} V_t X_t^{\mathsf{T}}\right)\left(\sum_{t=1}^{T} X_t X_t^{\mathsf{T}}\right)^{-1/2}. \tag{33}$$

This object has special structure. Firstly, in many cases of interest, e.g. the autoregressive model in (2), the noise V_t is independent of X_k for all $k \leq t$. This is what provides martingale structure, as will be made precise shortly. Secondly, it is self-normalized: if the covariates X_t are large for some t, then any increase in the left sum will be compensated by an increase in the sum in the term on the right. Together, these properties make the object above a *self-normalized martingale* term.

To express results generally and compactly, several definitions are in order.

Definition IV.1. *(Filtration and Adapted Process) A sequence of sub-σ-algebras $\{\mathcal{F}_t\}_{t=1}^T$ is said to be a* filtration *if $\mathcal{F}_t \subseteq \mathcal{F}_k$ for $t \leq k$. A stochastic process $\{W_t\}_{t=1}^T$ is said to be* adapted *to the filtration $\{\mathcal{F}_t\}_{t=1}^T$ if for all $t \geq 1$, W_t is $\mathcal{F}_t$-measurable.*

Conditioning on a sub-σ-algebra provides partial information about the total randomness. Therefore, the requirement that a filtration is non-decreasing captures the fact that information is not forgotten. An adapted process is one in which all the randomness at a particular time is explained by the information in the filtration up to that time.

Definition IV.2. *(Martingale) Consider a stochastic process $\{W_t\}_{t=1}^T$ which is adapted to a filtration $\{\mathcal{F}_t\}_{t=1}^T$. This process is called a* martingale *if for all $1 \leq t \leq T$, W_t is integrable and for all $1 \leq t < T$, $\mathbf{E}[W_{t+1}|\mathcal{F}_t] = W_t$.*

Martingales model causal or non-anticipative processes. To better appreciate this, note that the increments $W_{t+1} - W_t$ are mean zero and conditionally orthogonal to the past; they can be thought of as the "next step" in a random walk whose path is traced out by W_t.

In the context of the linear time-series model in (1), we may define the sub-σ-algebras $\mathcal{F}_t$ as those induced by the randomness up to time t: $\mathcal{F}_t = \sigma(X_1, \ldots, X_{t+1}, V_1, \ldots, V_t)$. In this case, the process $\{X_t\}_{t=1}^T$ is adapted to the filtration $\{\mathcal{F}_{t-1}\}_{t=1}^T$ and the process $\{V_t\}_{t=1}^T$ is adapted to the filtration $\{\mathcal{F}_t\}_{t=1}^T$.

Recall now again that the "numerator" in the ordinary least squares error is (33). We see that if we define the sum, $S_t \triangleq \sum_{s=1}^t V_s X_s^\top$, then the process $\{S_t\}_{t=1}^T$ is adapted to $\{\mathcal{F}_t\}_{t=1}^T$. Furthermore, $\mathbf{E}\left(S_{t+1}|\mathcal{F}_t\right) = S_t + \mathbf{E}\left(V_{t+1}|\mathcal{F}_t\right) X_{t+1}^\top$. In particular, as long as the noise has conditional mean zero ($\mathbf{E}\left(V_{t+1}|\mathcal{F}_t\right) = 0$), the process $\{S_t\}_{t=1}^T$ is a martingale.[3] Normalizing the sum S_t by the covariates as $S_t\left(\sum_{s=1}^t X_s X_s^\top\right)^{-1/2}$ *almost* preserves the martingale structure. Expressions of this type are called self-normalized martingales—although we stress that they are not strictly speaking martingales but only constructed from them.

We now state bounds on the operator and Frobenius norms of the self-normalized martingale. The main idea behind the result—the technique of pseudo-maximization—is due to [21]. The formulations presented here are a consequence of Theorem 3.4 in [1].

Theorem IV.1. *(Special cases of Theorem 3.4 in [1]) Let $\{\mathcal{F}_t\}_{t=0}^T$ be a filtration such that $\{X_t\}_{t=1}^T$ is adapted to $\{\mathcal{F}_{t-1}\}_{t=1}^T$ and $\{V_t\}_{t=1}^T$ is adapted to $\{\mathcal{F}_t\}_{t=1}^T$. Additionally, suppose that for all $1 \leq t \leq T$, V_t is σ^2-conditionally sub-Gaussian with respect to $\mathcal{F}_t$. Let Σ be a positive definite matrix in $\mathbb{R}^{d_{\mathsf{X}} \times d_{\mathsf{X}}}$. For a fixed $T \in \mathbb{N}_+$ and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left\|\sum_{t=1}^{T} V_t X_t^\top \left(\Sigma + \sum_{t=1}^{T} X_t X_t^\top\right)^{-1/2}\right\|_F^2 \leq d_{\mathsf{Y}}\sigma^2 \log\left(\frac{\det\left(\Sigma + \sum_{t=1}^T X_t X_t^\top\right)}{\det(\Sigma)}\right) + 2\sigma^2 \log\frac{1}{\delta}.$$

Additionally, for a fixed $T \in \mathbb{N}_+$ and $\delta \in (0,1)$, with

[2] Similar results can also be obtained for restricted eigenvalues.

[3] Indeed, S_t is a so-called martingale transform of $X_{1:t}$.

probability at least $1-\delta$,

$$\left\|\sum_{t=1}^{T} V_t X_t^\top \left(\Sigma + \sum_{t=1}^{T} X_t X_t^\top\right)^{-1/2}\right\|_{\mathsf{op}}^2 \le 4\sigma^2 \log\left(\frac{\det\left(\Sigma+\sum_{t=1}^{T} X_t X_t^\top\right)}{\det(\Sigma)}\right) + 8 d_{\mathsf{Y}} \sigma^2 \log 5 + 8\sigma^2 \log\frac{1}{\delta}.$$

Note that the quantities bounded above have a positive definite matrix Σ added to the normalization quantities that was not present in the original term of interest, (33). Furthermore, the covariates $\sum_{t=1}^{T} X_t X_t$ appear in the upper bound. Hence, one typically combines the self-normalized martingale bound with some weak form of concentration.[4] This is done in Section V.

In the sequel, we prove the above bounds. To obtain the bound on the Frobenius norm, we directly consider the object

$$\left\|\left(\sum_{t=1}^{T} V_t X_t^\mathsf{T}\right)\left(\sum_{t=1}^{T} X_t X_t^\mathsf{T}\right)^{-1/2}\right\|_F, \tag{34}$$

while to obtain the bound on the operator norm we consider the following vector norm for an arbitrary unit vector w as an intermediate step:

$$\left\|\left(w^\top \sum_{t=1}^{T} V_t X_t^\mathsf{T}\right)\left(\sum_{t=1}^{T} X_t X_t^\mathsf{T}\right)^{-1/2}\right\|_2 \tag{35}$$

and combine with a covering argument (recall Section II-B).

A. Exponential Inequalities via Pseudo-maximization

We begin by neglecting the details of the process that generated the data in (33). In particular, consider a random matrix P assuming values in $\mathbb{R}^{d_\eta \times d_\mathsf{X}}$ for $d_\eta \in \mathbb{N}_+$ and a random matrix Q assuming values in $\mathbb{R}^{d_\mathsf{X}\times d_\mathsf{X}}$ with Q almost surely nonsingular. Bounding the quantities in (34) and (35) are special cases of bounding $\|PQ^{-\frac{1}{2}}\|_F$. A naive first approach is to apply a Chernoff bound (12). Doing so results in the inequality

$$\mathbf{P}\left(\|PQ^{-1/2}\|_F \ge x\right) \le \min_{\lambda\ge 0} \exp\left(-\frac{\lambda}{2}x^2\right) \mathbf{E}\exp\left(\frac{\lambda}{2}\|PQ^{-1/2}\|_F^2\right).$$

If it is possible to bound the moment generating function $\mathbf{E}\exp\left(\frac{\lambda}{2}\|PQ^{-1/2}\|_F^2\right)$ by one for some $\lambda > 0$, then the above bound provides an exponential inequality. Obtaining a bound of the form $\mathbf{E}\exp\left(\frac{\lambda}{2}\|PQ^{-1/2}\|_F^2\right) \le 1$ requires very strong assumptions on P and Q which would not be suitable for our purposes. However, we may observe that $\frac{1}{2}\|PQ^{-1/2}\|_F^2 = \max_\Lambda \operatorname{tr}(P\Lambda - \frac{1}{2}\Lambda^\top Q\Lambda)$. This motivates the following canonical assumption in self-normalized process theory:

$$\max_{\Lambda\in\mathbb{R}^{d_\mathsf{X}\times d_\eta}} \mathbf{E}\exp\operatorname{tr}\left(P\Lambda - \frac{1}{2}\Lambda^\top Q\Lambda\right) \le 1. \tag{36}$$

This inequality is called the canonical assumption because a wide variety of self-normalized processes satisfy it. We will demonstrate that it is satisfied for (34) and (35) in Section IV-B. If we could exchange the order of the maximization with the expectation in (36), then the bound $\mathbf{E}\exp\left(\frac{1}{2}\|PQ^{-1/2}\|_F^2\right) \le 1$ would be satisfied, and the Chernoff bound above would provide a valuable exponential inequality. As this exchange is not possible, we instead lower bound the maximization over Λ by assigning a probability distribution to a random variable Ψ which takes values in $\mathbb{R}^{d_\mathsf{X}\times d_\eta}$, and taking the expectation over this distribution. Doing so preserves the inequality (36):

$$\mathbf{EE}\left[\exp\operatorname{tr}\left(P\Psi - \frac{1}{2}\Psi^\top Q\Psi\right)\Big|\Psi\right] \le 1.$$

The order of expectation over Ψ and over the random variables P and Q may then be exchanged by an appeal to Fubini's theorem:

$$1 \ge \mathbf{EE}\left[\exp\operatorname{tr}\left(P\Psi - \frac{1}{2}\Psi^\top Q\Psi\right)\Big|\Psi\right] = \mathbf{EE}\left[\exp\operatorname{tr}\left(P\Psi - \frac{1}{2}\Psi^\top Q\Psi\right)\Big|P,Q\right]. \tag{37}$$

By selecting the distribution over Ψ appropriately, the result is a so-called *pseudo-maximization*. In particular, by completing the square, the inner conditional expectation on the right may be expressed as

$$\mathbf{E}\left[\exp\operatorname{tr}\left(P\Psi - \frac{1}{2}\Psi^\top Q\Psi\right)\Big|P,Q\right] = \exp\operatorname{tr}\left(PQ^{-1}P^\top/2\right) \times \mathbf{E}\left[\exp\operatorname{tr}\left(-\frac{1}{2}(\Psi - Q^{-1}P^\top)^\top Q(\Psi - Q^{-1}P^\top)\right)\Big|P,Q\right].$$

For particular choices of the distribution of Ψ, the right side of the above expression approximates the maximimum value, $\exp\operatorname{tr}\left(PQ^{-1}P^\top/2\right)$, of $\exp\operatorname{tr}(P\Lambda - \frac{1}{2}\Lambda^\top Q\Lambda)$. This allows us to apply a Chernoff argument similar to the one sketched above to obtain an exponential bound on a quantity related to $\|PQ^{-1/2}\|_F$. The following lemma demonstrates one such bound that results by selecting the distribution of Ψ as a matrix normal distribution.

Lemma IV.1 (Extension of Theorem 14.7 in [18])**.** *Suppose that* (36) *is satisfied. Let* Σ *be a positive definite matrix in* $\mathbb{R}^{d_\mathsf{X}\times d_\mathsf{X}}$*. Then, for* $\delta > 0$*, with probability at least* $1-\delta$*,*

$$\|P(Q+\Sigma)^{-1/2}\|_F^2 \le 2\log\left(\frac{\det(Q+\Sigma)^{d_\eta/2}\det(\Sigma)^{-d_\eta/2}}{\delta}\right).$$

[4] Alternatively, in the analysis of ridge regression, Σ takes the role of the penalizing matrix which can be tuned.

B. Self-Normalized Martingales Satisfy the Canonical Assumption

In order to make use of Lemma IV.1 to bound (34) or (35), we must ensure that the condition (36) holds for

$$P = \sum_{t=1}^{T} \frac{\eta_t X_t^\top}{\sigma}, \quad Q = \sum_{t=1}^{T} X_t X_t^\top,$$

where $\eta_t \in \mathbb{R}^{d_\eta}$ is either the noise process V_t or the scalar process $w^\top V_t$ for some fixed unit vector w. The following lemma shows that it is sufficient for η_t to be σ^2-conditionally sub-Gaussian.

Lemma IV.2. *Fix $T \in \mathbb{N}_+$. Let $\{\mathcal{F}_t\}_{t=0}^T$ be a filtration such that $\{X_t\}_{t=1}^T$ is adapted to $\{\mathcal{F}_{t-1}\}_{t=1}^T$ and $\{\eta_t\}_{t=1}^T$ is adapted to $\{\mathcal{F}_t\}_{t=1}^T$. Additionally, suppose that for all $t \geq 1$, η_t is σ^2-conditionally sub-Gaussian with respect to $\mathcal{F}_t$. Let $\Lambda \in \mathbb{R}^{d_{\mathsf{X}} \times d_\eta}$ be arbitrary and consider for $t \in \{1, \dots, T\}$*

$$M_t(\Lambda) \triangleq \exp \operatorname{tr}\left(\sum_{s=1}^{t} \left[\frac{\eta_s X_s^\top \Lambda}{\sigma} - \frac{1}{2} \Lambda^\top X_s X_s^\top \Lambda \right] \right).$$

Then $\mathbf{E} M_T(\Lambda) \leq 1$.

Synthesizing the results in this section along with a covering argument, yields Theorem IV.1

C. Notes

Remark IV.1. *Consider the dimensional dependencies of the Frobenius and operator norm bounds in Theorem IV.1. The leading term in the Frobenius norm bound is d_{Y} multiplied by the $\log\det$ term, which scales with $d_{\mathsf{X}} \log T$ when the empirical covariance is well-conditioned. In particular, the leading term scales with $d_{\mathsf{X}} d_{\mathsf{Y}} \log T$. The factor of d_{Y} is no longer present on the $\log\det$ term for the operator norm. The term therefore scales as $d_{\mathsf{X}} \log T$ when the the empirical covariance matrix is well-conditioned. There is, however, a term $8 d_{\mathsf{Y}} \sigma^2 \log 5$ which results from the covering argument. The operator norm bound therefore scales as $\max\{d_{\mathsf{X}} \log T, d_{\mathsf{Y}}\}$.*

Remark IV.2. *Theorem IV.1 holds for a fixed $T \in \mathbb{N}_+$, which is sufficient for analyzing the system identification error. In contrast, the self-normalized margtingale bound in [1] holds for an arbitrary stopping time and thus uniformly for all $T \in \mathbb{N}_+$ by a stopping time construction. This uniform bound may be required in some settings, e.g. in error bounds for adapative control.*

V. System Identification

In this section, we analyze well-known linear system identification algorithms that rely on the least squares algorithm. Note that the problem formulation changes with the system parameterization (e.g., state space, ARMAX, etc.). However, a nice property of linear systems is that under certain conditions, we can obtain a linear non-parametric ARX model by regressing the system output to past outputs and inputs. Then, depending on the parameterization, we can recover a particular realization. In the following, we first review ARX identification, which can be seen as a fundamental building block for many linear system identification algorithms. Then, we analyze identification of Markov parameters in the case of state-space systems. We focus exclusively on the case of single trajectory data.

A. ARX Systems

Consider an unknown vector autoregressive system with exogenous inputs (ARX)

$$Y_t = \sum_{i=1}^{p} A_i^\star Y_{t-i} + \sum_{j=1}^{q} B_j^\star U_{t-j} + \Sigma_W^{1/2} W_t, \tag{38}$$

where $Y_t \in \mathbb{R}^{d_{\mathsf{Y}}}$ are the system outputs, $U_t \in \mathbb{R}^{d_{\mathsf{U}}}$ are the control (exogenous) inputs, and $W_t \in \mathbb{R}^{d_{\mathsf{Y}}}$ is the normalized process noise with $\Sigma_W \in \mathbb{R}^{d_{\mathsf{Y}} \times d_{\mathsf{Y}}}$ capturing the (non-normalized) noise covariance. Matrices $A_i^\star$, $i \leq p$ and $B_j^\star$, $j \leq q$ contain the unknown ARX coefficients. For the initial conditions, we assume $Y_{-1} = \cdots = Y_{-p} = 0$, $U_{-1} = \cdots = U_{-q} = 0$.

Assumption V.1 (System and Noise model)**.** *Let the noise covariance $\Sigma_W \succ 0$ be full rank. Let the normalized process noise $W_t, t \geq 0$ be independent and identically distributed, K^2-sub-Gaussian (see Definition II.1), with zero mean and unit covariance $\mathbf{E} W_t W_t^\top = I_{d_{\mathsf{Y}}}$. The orders p, q are known. System (38) is non-explosive, that is, the eigenvalues of matrix*

$$\mathcal{A}_{11} \triangleq \begin{bmatrix} A_1^\star & A_2^\star & \cdots & A_{p-1}^\star & A_p^\star \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & I & 0 \end{bmatrix}, \tag{39}$$

lie strictly on or inside the unit circle $\rho(\mathcal{A}_{11}) \leq 1$.

The techniques below can provide meaningful finite-sample bounds only when the system is non-explosive. Deriving finite sample guarantees for identifying open-loop, explosively unstable partially-observed systems from single trajectory data is open to the best of our knowledge [31].

In this tutorial, we focus solely on white-noise excitation inputs. For analyses of more advanced experimental designs—a crucial aspect of system identification—we refer to [36] as well as to the classical literature [14].

Assumption V.2 (White-noise excitation policy)**.** *We assume that the control input is generated by a random i.i.d. Gaussian process, that is, $U_t \sim \mathcal{N}(0, \sigma_u^2 I)$.*

Grouping all covariates into one vector and defining

$$X_t = \begin{bmatrix} Y_{t-1:t-p}^\top & U_{t-1:t-q}^\top \end{bmatrix}^\top, \; \theta^\star = \begin{bmatrix} A_{1:p}^\star & B_{1:q}^\star \end{bmatrix} \tag{40}$$

we can re-write (38) in terms of (1)

$$Y_t = \theta^\star X_t + \Sigma_W^{1/2} W_t,$$

where W_t is independent from X_t, but X_t has the special time-dependent structure induced by (38). Given samples $(Y_{1:T}, U_{0:T-1})$, the least-squares estimate is given by

$$\widehat{\theta}_T \triangleq \sum_{t=1}^{T} Y_t X_t^\top \left(\sum_{t=1}^{T} X_t X_t^\mathsf{T} \right)^\dagger, \tag{41}$$

where we purposely highlight the dependence of the estimate on the number of samples with the subscript T. Before we present the main result, let us define some quantities which are related to the quality of system estimates. The covariance at time $t \geq 0$ is defined as

$$\Sigma_t \triangleq \mathbf{E} X_t X_t^\top. \tag{42}$$

It captures the expected richness of the data, i.e., how excited the modes of the system are on average. In particular, the relative excitation of the data compared to the noise magnitude significantly affects the quality of system identification. This motivates the definition of signal-to-noise (SNR) as the ratio of the worst-case excitation over the worst-case noise magnitude

$$\mathsf{SNR}_t \triangleq \frac{\lambda_{\min}(\Sigma_t)}{\|\Sigma_W\|_{\mathsf{op}} K^2}. \tag{43}$$

The following theorem provides a finite-sample upper bound on the performance of the least-square estimator.

Theorem V.1 (ARX Finite-Sample Bound)**.** *Let $(Y_{1:T}, U_{0:T-1})$ be single trajectory input-output samples generated by system (38) under Assumptions V.1, V.2 for some horizon T. Fix a failure probability $0 < \delta < 1$ and a time index $\tau \geq \max\{p, q\}$. Let $T_{\mathsf{pe}}(\delta, \tau) \triangleq \min\{t : t \geq T_0(t, \delta/3, \tau)\}$, where T_0 is defined in (46). If $T \geq T_{\mathsf{pe}}(\delta, \tau)$, then with probability at least $1 - \delta$*

$$\|\widehat{\theta}_T - \theta^\star\|_{\mathsf{op}}^2 \leq \frac{C}{\mathsf{SNR}_\tau T} \left((pd_\mathsf{Y} + qd_\mathsf{U}) \log \frac{pd_\mathsf{Y} + qd_\mathsf{U}}{\delta} + \log\det\left(\Sigma_T \Sigma_\tau^{-1}\right) \right), \tag{44}$$

where C is a universal constant, i.e., it is independent of system, confidence δ and index τ.

For non-explosive systems, matrix $\Sigma_T \Sigma_\tau^{-1}$ increases at most polynomially with T in norm (in view of Lemma V.1). Hence, the identification error decays with a rate of $\tilde{O}(1/\sqrt{T})$.

Dimensional dependence. Ignoring logarithmic terms, the bound implies that the number of samples T should scale linearly with the dimension $pd_\mathsf{Y} + qd_\mathsf{U}$ of the covariates X_t. Since every sample Y_t contains at least d_Y measurements, this implies that the total number of measurements should be linear with $d_\mathsf{Y} \times (pd_\mathsf{Y} + qd_\mathsf{U})$. This scaling is qualitatively correct since $\theta^\star$ has $d_\mathsf{Y} \times (pd_\mathsf{Y} + qd_\mathsf{U})$ unknowns, requiring at least as many independent equations.

Logarithmic dependence on confidence. The error norm scales linearly with $\sqrt{\log 1/\delta}$. In the asymptotic regime we also recover the same order of $\sqrt{\log 1/\delta}$ by applying the Central Limit Theorem (CLT). However, in the regime of finite samples, obtaining the rate is non-trivial, see [31], and requires the analysis presented in this tutorial.

System theoretic constants. The identification error is directly affected by the SNR of the system. The more the system is excited and the smaller the noise, the better the SNR becomes. However, excitability varies heavily depending on the system and the choice of excitation policy. In particular, the system's controllability structure can affect the degree of excitation dramatically. Systems with poor controllability structure can exhibit SNR which suffers from curse of dimensionality, i.e., the smallest eigenvalue of Σ_τ degrades exponentially with the system dimension [30].

The upper bound also increases with the logarithm of the "condition number" $\det(\Sigma_T \Sigma_\tau^{-1})$. For stable systems, this condition number is bounded since Σ_T converges to a steady-state covariance as T increases; we can neglect it in this case. On the other hand, the term might be significant in the case of general non-explosive systems. Let κ be the size of the largest Jordan block of $\mathcal{A}_{11}$ with eigenvalues on the unit circle. Then, this term can be as large as $\kappa \log T$.

Burn-in time. The upper bound holds as soon as the number of samples exceeds a "burn-in" time $T_{\mathsf{pe}}(\delta, \tau)$. If the system is non-explosive, $T_{\mathsf{pe}}(\delta, \tau)$ is always finite for fixed τ. Exceeding the burn-in time guarantees that we have persistency of excitation, that is, all modes of the system are excited. The burn-in time increases as we require more confidence δ and we chose larger time indices τ. On the other had, larger τ leads to larger Σ_τ, which improves the SNR_τ. In other words, there is a tradeoff between improving the SNR and deteriorating the burn-in time. We analyze persistency of excitation in more detail, in the next subsection.

Proof outline We outline the proof below, the full proofs can be found in the online version. To analyze the least squares error, observe that

$$\widehat{\theta}_T - \theta^\star = \underbrace{\sum_{t=1}^{T} \Sigma_W^{1/2} W_t X_t^\top \left(\sum_{t=1}^{T} X_t X_t^\top \right)^{-1/2}}_{\text{noise}} \times \underbrace{\left(\sum_{t=1}^{T} X_t X_t^\top \right)^{-1/2}}_{\text{excitation}} \tag{45}$$

where we assumed that the inverse exists. To deal with the second term, we will prove persistency of excitation in finite time leveraging the techniques of Section III, which requires most of the work. To deal with the noise part we will apply the self-normalized martingale methods, which are reviewed in Section IV. We study both terms in the following subsections.

1) Persistency of Excitation in ARX Models: In this subsection, we leverage the result of Theorem III.1 to prove persistency of excitation. By persistency of excitation, we refer to the case when we have rich input-output data, that is, data which characterize all possible behaviors of the system. Recall the definition (9) of the empirical covariance matrix

$$\widehat{\Sigma}_T \triangleq \frac{1}{T} \sum_{t=1}^{T} X_t X_t^\top.$$

Using this definition, the excitation term in the least squares error can be re-written as $(T\widehat{\Sigma}_T)^{-1/2}$. We say that persistency of excitation holds if and only if the empirical covariance matrix is strictly positive definite (full rank). In the following, we show that the full rank condition holds with high probability, provided that the number of samples exceeds a certain threshold, i.e., the burn-in time.

Theorem V.2 (ARX PE)**.** *Let $(Y_{1:T}, U_{0:T-1})$ be input-output samples generated by system (38) under Assumptions V.1, V.2 for some fixed horizon T. Fix a failure probability $0 < \delta < 1$ and a time index $\tau \geq \max\{p, q\}$. Then, $\lambda_{\min}(\Sigma_\tau) > 0$. Moreover, if T is large enough*

$$T \geq T_0(T, \delta, \tau) \triangleq 1152\tau \max\{K^2, 1\} \times \left((pd_{\mathsf{Y}} + qd_{\mathsf{U}}) \log C_{\mathsf{sys}}(T, \tau) + \log(1/\delta)\right) \quad (46)$$

where

$$C_{\mathsf{sys}}(T, \tau) \triangleq \frac{T}{3\tau} \frac{\|\Sigma_T\|_{\mathsf{op}}^2}{\lambda_{\min}^2(\Sigma_\tau)},$$

then,

$$\mathbf{P}\left(\widehat{\Sigma}_T \succeq \frac{1}{16}\Sigma_\tau\right) \geq 1 - \delta.$$

The detailed proof can be found in the online version, we only sketch the main ideas here. The first step is to express the covariates $X_{1:T}$ as a causal linear combination of the noises and the inputs, mimicking (23). The evolution of the covariates follows a state-space recursion

$$X_{t+1} = \mathcal{A}X_t + \mathcal{B}V_t, \quad (47)$$

where we concatenate noises and inputs $V_t \triangleq \begin{bmatrix} W_t^\top & U_t^\top \end{bmatrix}^\top$. Matrices $\mathcal{A}$, $\mathcal{B}$ are given by

$$\mathcal{A} \triangleq \begin{bmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} \\ 0 & \mathcal{A}_{22} \end{bmatrix}, \quad \mathcal{B} \triangleq \begin{bmatrix} \mathcal{B}_1 & \mathcal{B}_2 \end{bmatrix}, \text{ where}$$

$\mathcal{A}_{11}$ is defined in (39),

$$\mathcal{A}_{12} \triangleq \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \otimes \begin{bmatrix} B_1^\star & \cdots & B_q^\star \end{bmatrix} \in \mathbb{R}^{pd_{\mathsf{Y}} \times qd_{\mathsf{U}}}$$

$$\mathcal{A}_{22} \triangleq \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 \\ I_{d_{\mathsf{U}}} & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & & \cdots & \\ 0 & \cdots & I_{d_{\mathsf{U}}} & 0 & 0 \\ 0 & \cdots & 0 & I_{d_{\mathsf{U}}} & 0 \end{bmatrix} \in \mathbb{R}^{qd_{\mathsf{U}} \times qd_{\mathsf{U}}}$$

$$\mathcal{B}_1 \triangleq \begin{bmatrix} \Sigma_W^{1/2} & 0_{((p-1)d_{\mathsf{Y}} + qd_{\mathsf{U}}) \times d_{\mathsf{Y}}} \end{bmatrix}^\top,$$

$$\mathcal{B}_2 \triangleq \begin{bmatrix} 0_{pd_{\mathsf{Y}} \times d_{\mathsf{U}}} & I_{d_{\mathsf{U}}} & 0_{(q-1)d_{\mathsf{U}} \times d_{\mathsf{U}}} \end{bmatrix}^\top.$$

The vector $X_{1:T}$ of all covariates satisfies the following causal linear relation

$$X_{1:T} = \underbrace{\begin{bmatrix} \mathcal{B} & 0 & \cdots & 0 \\ \mathcal{A}\mathcal{B} & \mathcal{B} & \cdots & 0 \\ \vdots & & \ddots & \\ \mathcal{A}^{T-1}\mathcal{B} & \mathcal{A}^{T-2}\mathcal{B} & \cdots & \mathcal{B} \end{bmatrix}}_{\mathbf{L}} V_{0:T-1}. \quad (48)$$

where the lower-block triangular matrix is the Toeplitz matrix generated by the Markov parameters matrices $\mathcal{A}$, $\mathcal{B}$. The second step is to apply Theorem III.1. The details can be found in the online version.

Remark V.1 (Existence of burn-in time.)**.** *For the above result to be meaningful, we need inequality (46) to be feasible. For non-explosive systems, the system theoretic term $\log C_{\mathsf{sys}}(T, \tau)$ increases at most logarithmically with T, since Σ_t increases polynomially with t in view of Lemma V.1. Hence, for any fixed τ, or, in general, any τ that increases mildly (sublinearly) with T, e.g. $O(\sqrt{T})$, it is possible to satisfy (46). Note that $\rho(\mathcal{A}) = \rho(\mathcal{A}_{11})$ due to the triangular structure of $\mathcal{A}$. Hence, by Assumption V.1, system $\mathcal{A}$ is also non-explosive.*

Remark V.2 (Unknown system orders p, q)**.** *The result of Theorem V.2 still holds if the orders p, q are unknown and we use the wrong orders $\hat{p}, \hat{q}$ in the covariates X_t. We just need to replace p, q with $\hat{p}, \hat{q}$ with $\hat{p}, \hat{q}$ and revise the size of Σ accordingly in (46). The finite-sample bounds of Theorem V.1 also hold (by revising accordingly), but only if we overestimate p, q, that is $\hat{p} \geq p$, $\hat{q} \geq q$. This also generalizes the single trajectory result of [3] to non-explosive systems.*

The following supporting lemma proves that the k-th powers of non-explosive matrices increase at most polynomially with k.

Lemma V.1 (Lemma 1 in [30])**.** *Let $\mathcal{A} \in \mathbb{R}^{d \times d}$ have all eigenvalues inside or on the unit circle, with $\|\mathcal{A}\|_{\mathsf{op}} \leq M$, for some $M > 0$. Then,*

$$\|\mathcal{A}^k\|_{\mathsf{op}} \leq (ek)^{d-1} \max\left\{M^d, 1\right\}. \quad (49)$$

As a corollary, the covariance matrices Σ_t also grow at most polynomially with the time t.

2) Dealing with the Noise Term: In this subsection, we modify the noise term so that we can leverage Theorem IV.1, which cannot be applied directly. We first manipulate the inverse of $T\widehat{\Sigma}_T$ to relate it to the inverse of $\Sigma + T\widehat{\Sigma}_T$, for some carefully selected Σ. Inspired by [24], we leverage the result of Theorem V.2. Under the event that persistency of excitation holds we have $\widehat{\Sigma}_T \succeq T\Sigma_\tau/16$. Thus, selecting $\Sigma = T\Sigma_\tau/16$ guarantees that

$$\left(T\widehat{\Sigma}_T\right)^{-1} \preceq 2\left(\Sigma + T\widehat{\Sigma}_T\right)^{-1}.$$

We can now apply Theorem IV.1. To finish the proof we need to upper-bound the determinant of $\log\det(\Sigma + T\widehat{\Sigma}_T)$. It is sufficient to establish a crude upper-bound on the empirical covariance $T\widehat{\Sigma}_T$ as in the following lemma.

Lemma V.2 (Matrix Markov's inequality)**.** *Fix a failure probability $\delta > 0$. With probability at least $1 - \delta$*

$$\widehat{\Sigma}_T \preceq \frac{pd_{\mathsf{Y}} + qd_{\mathsf{U}}}{\delta}\Sigma_T. \tag{50}$$

A more refined upper bound can also be applied (see e.g. the proof of Proposition VI.1 below or the results in [8]).

B. State-Space Systems

In this subsection, we derive finite-sample guarantees for learning Markov parameters of linear systems in state-space form. Consider the following state-space system in the so-called innovation form:

$$\begin{aligned} X_{t+1} &= A^\star X_t + B^\star U_t + F^\star \Sigma_E^{1/2} E_t \\ Y_t &= C^\star X_t + \Sigma_E^{1/2} E_t, \end{aligned} \tag{51}$$

where $A^\star \in \mathbb{R}^{d_{\mathsf{X}} \times d_{\mathsf{X}}}$, $B^\star \in \mathbb{R}^{d_{\mathsf{X}} \times d_{\mathsf{U}}}$, $F^\star \in \mathbb{R}^{d_{\mathsf{X}} \times d_{\mathsf{Y}}}$, and $C^\star \in \mathbb{R}^{d_{\mathsf{Y}} \times d_{\mathsf{X}}}$ are *unknown* state-space parameters. For the initial condition, we assume $X_0 = 0$. We call the normalized noise process E_t the innovation error process. Similar to the ARX case, we focus on white-noise excitation inputs, namely Assumption V.2 also holds here. Moreover, we assume the following.

Assumption V.3 (System and Noise model)**.** *Let the noise covariance $\Sigma_E \succ 0$ be full rank. Let the normalized innovation process E_t be independent, identically distributed, K^2-sub-Gaussian (see Definition II.1), with zero mean and unit covariance $\mathbf{E} E_t E_t^\top = I_{d_{\mathsf{Y}}}$. The order d_{X} is unknown. System (51) is non-explosive, that is, the eigenvalues of matrix $A^\star$ lie strictly on or inside the unit circle $\rho(A) \leq 1$. The system is also minimum-phase, i.e., the closed loop matrix*

$$A_{\mathsf{cl}}^\star \triangleq A^\star - F^\star C^\star \tag{52}$$

has all eigenvalues inside the unit circle $\rho(A_{\mathsf{cl}}^\star) < 1$.

The innovation form (51) might seem puzzling at first. In particular, the correlation between process and measurement noise via $F^\star$, and the requirement $\rho(A_{\mathsf{cl}}^\star) < 1$ seem restrictive. However, the representation (51) is standard in the system identification literature [34]. Moreover, as we review below, standard state-space models have input-output second-order statistics, which are equivalent to the ones generated by system (51) (for appropriate $F^\star$, Σ_E).

Remark V.3 (Generality of model)**.** *System class (51) captures general state-space systems driven by Gaussian noise. Consider the following state-space model*

$$\begin{aligned} S_{t+1} &= A^\star S_t + B^\star U_t + W_t \\ Y_t &= C^\star S_t + V_t, \end{aligned} \tag{53}$$

where W_t, V_t are i.i.d., independent of each other, mean-zero Gaussian, with covariances Σ_W and Σ_V respectively. Assume that $\Sigma_V \succ 0$ is full rank, the pair $(C^\star, A^\star)$ is detectable, and the pair $(A^\star, \Sigma_W)$ is stabilizable. These three assumptions imply that the Kalman filter of system (53) is well-defined [2]. In particular, define the Riccati operator as

$$\begin{aligned} \mathsf{RIC}(P) \triangleq\ & A^\star P (A^\star)^\top + \Sigma_W \\ & - A^\star P (C^\star)^\top (C^\star P (C^\star)^\top + \Sigma_V)^{-1} C^\star P (A^\star)^\top \end{aligned} \tag{54}$$

and let $P^\star$ be the unique positive semidefinite solution of $P^\star = \mathsf{RIC}(P^\star)$. Then the Kalman filter gain is equal to

$$F^\star = -A^\star P (C^\star)^\top (C^\star P (C^\star)^\top + \Sigma_V)^{-1}. \tag{55}$$

Assume that the initial state is also mean-zero Gaussian with covariance $P^\star$ and independent of the noises. Finally set

$$\Sigma_E = C^\star P^\star (C^\star)^\top + \Sigma_V. \tag{56}$$

Under the above assumptions and selection of $F^\star$, Σ_E systems (51) and (53) are statistically equivalent from an input-output perspective, see [19]. Both system descriptions lead to input-output trajectories with identical statistics. Moreover, due to the properties of Kalman filter, stability of $A_{\mathsf{cl}}^\star$ (minimum phase property) and independence of E_t are satisfied automatically [2].

In this tutorial we will only focus on recovering the first few (logarithmic in T-many) Markov parameters $C^\star (A_{\mathsf{cl}}^\star)^i B^\star$, $i \geq 0$ and $C^\star (A_{\mathsf{cl}}^\star)^j F^\star$, $j \geq 0$ of system (51). From a learning theory point of view, this is also known as improper learning, since the search space (finitely many Markov parameters) does not exactly, but only approximately, coincide with the hypothesis class (state space models). In principle, this forms the backbone of the SSARX method introduced by Jansson [7]. One can then proceed to recover the original state-space parameters (up to similarity transformation) from the Markov parameters by employing some realization method. We refer to [17, 31] for a discussion on this approach from a finite sample perspective.

1) Reduction to ARX Learning with Bias: Let $p > 0$ be a past horizon. Denote the Markov parameters up to time p by

$$\begin{aligned} \theta_p^\star \triangleq [\ & C^\star B^\star \quad \cdots \quad C^\star (A_{\mathsf{cl}}^\star)^{p-1} B^\star \\ & C^\star F^\star \quad \cdots \quad C^\star (A_{\mathsf{cl}}^\star)^{p-1} F^\star\ ]. \end{aligned} \tag{57}$$

Note that the innovation errors are equal to $\Sigma_E^{1/2} E_t = Y_t - C^\star X_t$. Replacing this expression into the state equation (51), we obtain

$$X_t = A_{\mathsf{cl}}^\star X_{t-1} + B^\star U_{t-1} + F^\star Y_{t-1}.$$

Unrolling the state equation p times, we get

$$Y_t = \underbrace{\theta_p^\star Z_t + \Sigma_E^{1/2} E_t}_{\text{ARX}} + \underbrace{C^\star (A_{\mathsf{cl}}^\star)^p X_{t-p}}_{\text{bias}}, \tag{58}$$

where Z_t includes the past p covariates

$$Z_t = \begin{bmatrix} Y_{t-1:t-p}^\top & U_{t-1:t-p}^\top \end{bmatrix}^\top. \tag{59}$$

The above recursion is an approximate ARX equation. There is an additive bias error term on top of the statistical noise. The least-squares solution is given by

$$\widehat{\theta}_{p,T} \triangleq \sum_{t=1}^{T} Y_t Z_t^\top \left(\sum_{t=1}^{T} Z_t Z_t^\mathsf{T} \right)^\dagger, \tag{60}$$

where we also highlight the dependence on the past p. By the minimum phase assumption, the bias term decays exponentially with the past horizon p. This follows from the fact that $A_{\mathsf{cl}}^\star$ is asymptotically stable, while X_t scales at most polynomially with t (in view of Lemma V.1). By selecting $p = \Omega(\log T)$, we can make the bias term decay very fast, making its contribution to the error $\theta_p^\star - \widehat{\theta}_{p,T}$ negligible. On the other hand, increasing the past horizon p increases the statistical error since the search space is larger.

2) Finite-Sample Guarantees: To derive a finite-sample rate for state space systems of the form (51), we follow the same steps as in the case of ARX systems. However, we have to account for the bias term and the fact that p grows with $\log T$. Let us define again the covariance at time $t \geq 0$

$$\Sigma_{p,t} \triangleq \mathbf{E} Z_t Z_t^\top, \tag{61}$$

where we highlight the dependence on both the past horizon p and the time t. The covariance of the state is defined similarly

$$\Sigma_{X,t} \triangleq \mathbf{E} X_t X_t^\top. \tag{62}$$

Define the SNR as

$$\mathsf{SNR}_{p,t} \triangleq \frac{\lambda_{\min}(\Sigma_{p,t})}{\|\Sigma_E\|_{\mathsf{op}} K^2}. \tag{63}$$

Unlike the ARX case, here the SNR might degrade since we allow p to grow with $\log T$. For this reason, we require the following additional assumption.

Assumption V.4 (Non-degenerate SNR)**.** *We assume that the SNR is uniformly lower bounded for all possible past horizons*

$$\liminf_{t \geq 0} \mathsf{SNR}_{t,t} > 0.$$

Later on, in Theorem V.4, we show that the above condition is non-vacuous and is satisfied for quite general systems.

Theorem V.3 (State Space Finite-Sample Bound)**.** *Let $(Y_{1:T}, U_{0:T-1})$ be single trajectory input-output samples generated by system (51) under Assumptions V.2, V.3, V.4, for some horizon T. Fix a failure probability $0 < \delta < 1$ and select $p = \beta \log T$, for β large enough such that*

$$\|C^\star (A_{\mathsf{cl}}^\star)^p\|_{\mathsf{op}} \|\Sigma_{X,T}\|_{\mathsf{op}} \leq T^{-3}. \tag{64}$$

Let $T_{\mathsf{pe}}^{\mathsf{ss}}(\delta, \beta) \triangleq \min\{t : t \geq T_0(t, \delta, \beta \log t)\}$, where T_0 is defined in (46). If $T \geq T_{\mathsf{pe}}^{\mathsf{ss}}(\delta, \beta)$, then with probability at least $1 - 2\delta$

$$\|\widehat{\theta}_{p,T} - \theta_p^\star\|_{\mathsf{op}}^2 \leq \frac{C_1}{\mathsf{SNR}_{p,p} T} \left(p(d_{\mathsf{Y}} + d_{\mathsf{U}}) \log \frac{p(d_{\mathsf{Y}} + d_{\mathsf{U}})}{\delta} + \log\det\left(\Sigma_{p,T} \Sigma_{p,p}^{-1}\right) \right), \tag{65}$$

where C_1 is a universal constant, i.e., it is independent of system, confidence δ and past horizon p.

For non-explosive systems, matrix $\Sigma_{p,T}\Sigma_{p,p}^{-1}$ increases at most polynomially with T in norm. Since the SNR is uniformly lower bounded, the identification error decays with a rate of $\tilde{O}(1/\sqrt{T})$. The bound seems similar to the one for ARX systems for $\tau = p$. However, since $p = \theta(\log T)$, we have an extra logarithmic term.

Role of β. Recall the approximate ARX relation (58). For the bias term to be small, the exponentially decaying $(A_{\mathsf{cl}}^\star)^p$ should counteract the magnitude of the state $\|X_{t-p}\|_{\mathsf{op}}$. Intuitively, the state grows as fast as $\|\Sigma_{X,t}\|_{\mathsf{op}}^{1/2}$, where $\Sigma_{X,t} = \mathbf{E} X_t X_t^\top$. Hence the state norm grows at most polynomially with T. Meanwhile, $\|(A_{\mathsf{cl}}^\star)^p\|_{\mathsf{op}} = O(\rho^p)$ for some $\rho > \rho(A_{\mathsf{cl}}^\star)$. With the choice $p = \beta \log T$, we get $\|(A_{\mathsf{cl}}^\star)^p\|_{\mathsf{op}} = O(T^{-\beta/\log(1/\rho)})$. Hence, if we select large enough β, we can make the bias term very small, even smaller than the dominant $\tilde{O}(1/\sqrt{T})$ term.

Burn-in time. Since the system is non-explosive, $T_{\mathsf{pe}}^{\mathsf{ss}}(\delta, \beta)$ is always finite under Assumption V.4, for any β. As before, exceeding the burn-in time guarantees that we have persistency of excitation. Naturally, larger β lead to larger past horizons p, which, in turn, increase the burn-in time.

Finally, we prove that Assumption V.4 is non-vacuous. It is sufficient for $F^\star$ and Σ_W to be generated by a Kalman filter as in (55), (56).

Theorem V.4. *Consider system (51) and the definition of $\mathsf{SNR}_{p,t}$ in (63). If matrices $F^\star$, Σ_E are generated as in (55), (56) with $(A^\star, \Sigma_W^{1/2})$ stabilizable, $(C^\star, A^\star)$ detectable and $\Sigma_V \succ 0$, then the SNR is uniformly lower bounded* $\liminf_{t \geq 0} \mathsf{SNR}_{t,t} > 0$.

Both conditions are sufficient. It is subject of future work to extend the result to more general non-exposive systems.

C. Notes

The exposition above is inspired by prior work on identifying fully-observed systems [4, 27, 24] and partially-observed systems [16, 28, 25, 29, 11, 10, 12]. For a wider overview of the literature, we refer the reader to [31].

Let us further remark that the guarantee for the ARX model in Theorem V.1 is almost optimal. The use of Matrix Markov's inequality yields extrananeous dependency on the problem dimension multiplying the deviation term $\log(1/\delta)$. This can in principle be removed by a more refined analysis (see e.g. the proof of Proposition VI.1 below or the results in [8]). The question of optimality in identifying partially observed state-space systems is more subtle, and while consistent, the bounds presented here are not (asymptotically) optimal.

VI. An Alternative Viewpoint: The Basic Inequality

In many situations, the choice of the model class $\mathsf{M} = \mathbb{R}^{d_{\mathsf{Y}} \times d_{\mathsf{X}}}$ leading to (8) is not appropriate. For instance physical or other modelling considerations might have already informed

us that the true $\theta^\star$ belongs to some smaller model class such as the family of low rank or sparse matrices which are strict subsets of M. Other properties one might wish to enforce include, stable, low norm, or even passivity-type properties. In either of the above examples no error expression of the form (10) is directly available. Instead, we observe by optimality of $\widehat{M}$ to the optimization program (7) that

$$\frac{1}{T}\sum_{t=1}^{T}\|Y_t-\widehat{\theta}X_t\|_2^2 \le \frac{1}{T}\sum_{t=1}^{T}\|Y_t-\theta^\star X_t\|_2^2. \qquad (66)$$

Expanding the squares and re-arranging terms we arrive at the so-called basic inequality of least squares:

$$\frac{1}{T}\sum_{t=1}^{T}\|(\widehat{\theta}-\theta^\star)X_t\|_2^2 \le \frac{2}{T}\sum_{t=1}^{T}\langle V_t,(\widehat{\theta}-\theta^\star)X_t\rangle. \qquad (67)$$

The inequality (67) serves as an alternative to the explicit error equation (10). To drive home this point, let us first re-arrange (67) slightly:

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\|(\widehat{\theta}-\theta^\star)X_t\|_2^2\\ &\le \frac{4}{T}\sum_{t=1}^{T}\langle V_t,(\widehat{\theta}-\theta^\star)X_t\rangle - \frac{1}{T}\sum_{t=1}^{T}\|(\widehat{\theta}-\theta^\star)X_t\|_2^2. \qquad (68)\end{aligned}$$

Note now that $\widehat{\theta}-\theta^\star$ are elements of $\mathsf{M}_\star \triangleq \mathsf{M}-\theta^\star$. Hence—by considering the worst-case (supremum) right hand side of (68)—we obtain:

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\|(\widehat{\theta}-\theta^\star)X_t\|_2^2\\ &\qquad\le \sup_{\theta\in\mathsf{M}_\star}\left\{\frac{4}{T}\sum_{t=1}^{T}\langle V_t,\theta X_t\rangle - \frac{1}{T}\sum_{t=1}^{T}\|\theta X_t\|_2^2\right\}. \qquad (69)\end{aligned}$$

In fact, if $\mathsf{M}=\mathbb{R}^{d_\mathsf{Y}\times d_\mathsf{X}}$, the optimization on the right hand side of (69) has an explicit solution. This implies that we always have the following upper-bound on the event that the design is nondegenerate:

$$\begin{aligned}&\sup_{\theta\in\mathsf{M}_\star}\left\{\frac{4}{T}\sum_{t=1}^{T}\langle V_t,\theta X_t\rangle - \frac{1}{T}\sum_{t=1}^{T}\|\theta X_t\|_2^2\right\}\\ &\le \sup_{\theta\in\mathbb{R}^{d_\mathsf{Y}\times d_\mathsf{X}}}\left\{\frac{4}{T}\sum_{t=1}^{T}\langle V_t,\theta X_t\rangle - \frac{1}{T}\sum_{t=1}^{T}\|\theta X_t\|_2^2\right\}\\ &= \frac{4}{T}\left\|\left(\sum_{t=1}^{T}V_tX_t^\mathsf{T}\right)\left(\sum_{t=1}^{T}X_tX_t^\mathsf{T}\right)^{-1/2}\right\|_F^2. \qquad \text{(opt.)}\end{aligned} \qquad (70)$$

Hence, we have in principle recovered an in-norm version of (10) with slightly worse constants. Put differently, we may regard (69) as a variational (or dual) form of the explicit error (10). Now, the advantage of (69) is twofold:

1) (69) and (70) hold for any $\mathsf{M}_\star \subset \mathbb{R}^{d_\mathsf{Y}\times d_\mathsf{X}}$ and hence allows us to analyze the LSE (7) beyond OLS ($\mathsf{M}_\star = \mathbb{R}^{d_\mathsf{Y}\times d_\mathsf{X}}$). This is important in identification problems where the parameter space is restricted.
2) We do not have to rely on (70) to control (69). In fact, for many reasonable classes of $\mathsf{M}_\star \subset \mathbb{R}^{d_\mathsf{Y}\times d_\mathsf{X}}$ we are able to give alternative arguments that are much sharper (in terms of e.g. dimensional scaling) than the naive bound (70). See Section VI-A below.

A third advantage of the variational form (69) is that it generalizes straightforwardly beyond linear least squares. In fact, none of the steps (66),(67), (68) and (69) relied on the linearity of $x\mapsto\widehat{\theta}x$ or that of $x\mapsto\theta^\star x$ ($x\in\mathbb{R}^{d_\mathsf{X}}$). We will explore this theme further in Section VI-A and Section VII.

A. Sparse Autoregressions

Before we proceed to sketch out how the basic inequality above extends to nonlinear problems in Section VII, let us use it to analyze a simple variation of the autoregression already encountered in Section V. Namely, the autoregressive model (38) which—for simplicity—is further assumed one-dimensional:

$$Y_t = \sum_{i=1}^{p}A_i^\star Y_{t-i} + W_t \qquad (71)$$

and assume in addition that it is known that only $s\in\mathbb{N}$ of the p entries of $\theta^\star=[A_1^\star,\dots,A_p^\star]$ are nonzero. Put differently, the vector $\theta^\star$ is known to be s-sparse and we write $\theta^\star\in\{\theta\in\mathbb{R}^p:\|\theta\|_0\le s\}\triangleq\mathsf{M}$. Hence, in this case the model class M is the union of $\binom{p}{s}$ subspaces. Clearly, we could use OLS (8) but this estimator does not take advantage of the additional information that $A^\star=\theta^\star$ lies in the s-dimensional submanifold M. Intuitively, if $s\ll p$ this set should be much smaller than $\mathbb{R}^p$ and so one expects that identification occurs at a faster rate.

In this section we demonstrate that the least squares estimator (7) in which the search is restricted to the low-dimensional manifold M outperforms the OLS. We stress that this is *not* a computationally efficient estimator and the results in this section should be thought of as an illustration of a proof technique.

Returning to the problem of controlling the error of this estimator, we note that in this case there is no closed form for the LSE and we do not have direct access to the error equation (10).[5] Hence, we instead use the offset basic inequality approach from Section VI. As before, it is convenient to set $X_t=[Y_{t-1},\dots,Y_{t-p}]^\mathsf{T}$. With this additional bit of notation in place, we recall from (69) that:

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\|(\widehat{\theta}-\theta^\star)X_t\|_2^2\\ &\qquad\le \max_{\theta\in\mathsf{M}_\star}\left\{\frac{4}{T}\sum_{t=1}^{T}W_t\theta X_t - \frac{1}{T}\sum_{t=1}^{T}|\theta X_t|_2^2\right\}\end{aligned}$$

where $\mathsf{M}_\star$ is the translation $\mathsf{M}-\theta^\star$. Since M is the union of $\binom{p}{s}$-many linear s-dimensional subspaces $S\subset\mathbb{R}^{d_\mathsf{X}\times d_\mathsf{X}}$,

[5]Although, in this particular case an alternative analysis based on this equation is possible.

$\mathsf{M}_\star$ is the union of $\binom{p}{s}$ affine subspaces s-dimensional affine subspaces of the form $S - \theta^\star$. Let us also note that $\mathsf{M}_\star \subset \mathsf{M} - \mathsf{M} = \{\theta \in \mathbb{R}^p : \|\theta\|_0 \le 2s\}$. Consequently:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|(\widehat{\theta}-\theta^\star)X_t\|_2^2 &\le \max_{\theta\in\mathsf{M}_\star}\left\{\frac{4}{T}\sum_{t=1}^{T}W_t\theta X_t - \frac{1}{T}\sum_{t=1}^{T}|\theta X_t|_2^2\right\}\\ &\le \max_{S}\max_{\theta\in S}\left\{\frac{4}{T}\sum_{t=1}^{T}W_t\theta X_t - \frac{1}{T}\sum_{t=1}^{T}|\theta X_t|_2^2\right\}. \quad (72)\end{aligned}$$

where maximization over S occurs over the $\binom{p}{2s}$-many sparse subspaces. Notice now that since θ in (72) is s-sparse, the products θX_t are just $\theta X_t = \sum_{i\in S}\theta_i(X_t)_i$ where we have abused notation and identified S with its support set. Hence, by the same direct calculation as in (70), if we denote $(X_t)_S$ the s-dimensional vector obtained by coordinate projection onto part of S not constrained to be identically zero (i.e. the image of the projection onto S represented as the s-dimensional Euclidean space) we find that:

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\|(\widehat{\theta}-\theta^\star)X_t\|_2^2\\ &\le \frac{4}{T}\max_{S}\left\|\left(\sum_{t=1}^{T}W_t(X_t)_S\right)\left(\sum_{t=1}^{T}(X_t)_S(X_t)_S^{\mathsf{T}}\right)^{-1/2}\right\|_2^2.\end{aligned} \quad (73)$$

The right hand side of (73) can be controlled by the self-normalized inequality in Theorem IV.1 for each fixed S. Moreover, there are only $\binom{p}{2s}$ such subspaces, so we can apply a union bound to control the maximum over these subspaces. Note also that the left hand side of (73) can be controlled by the tools developed in Section III. Carrying out these steps leads to the following guarantee.

Proposition VI.1. *Fix $T, k \in \mathbb{N}$ with T divisible by k and let $\mathbf{L}$ be the linear operator defined in (48). Let $\widehat{\theta}$ be the LSE (7) over the set $\mathsf{M} = \{\theta \in \mathbb{R}^p : \|\theta\|_0 \le s\}$ for the system (71). Define $\Sigma_j \triangleq \frac{1}{j}\sum_{t=1}^{j}\mathbf{E}X_tX_t^{\mathsf{T}}$ for $j \in [T]$ and*

$$\mathrm{cond}_{\mathsf{sys}}(T,k) \triangleq \left(1 + \frac{\|\mathbf{L}\mathbf{L}^{\mathsf{T}}\|_{\mathsf{op}}}{k\lambda_{\min}(\Sigma_T)}\right)\frac{\lambda_{\max}(\Sigma_T)}{\lambda_{\min}(\Sigma_k)}.$$

There exist univeral positive constants c, c' such that for any $\delta \in (0,1)$ it holds with probability at least $1-\delta$ that:

$$\|(\widehat{\theta}-\theta^\star)\sqrt{\Sigma_k}\|_2^2 \le c\sigma^2 \times \frac{s\log\left(\frac{p\times\mathrm{cond}_{\mathsf{sys}}(T,k)}{s}\right) + \log(1/\delta)}{T} \quad (74)$$

as long as

$$T/k \ge c'\sigma^2\left(s\left[\log\left(\mathrm{cond}_{\mathsf{sys}}(T,k)\right) + \log(p/s)\right] + \log(1/\delta)\right). \quad (75)$$

A few remarks are in order. The guarantee (74) depends on the dimension s of M, and not the total parameter dimension p. Similarly, the burn-in (75) exhibits a similar win, by depending linearly on s and only logarithmically on p. There is also the difference that the left hand side of (74) is given in the problem-dependent Mahalanobis norm induced by Σ_k and opposed to just the standard Euclidean 2-norm. This implies that if we actually want parameter identification in the sense of the previous section, a restricted eigenvalue condition on Σ_k is needed.[6] Indeed, for some positive number $\lambda_{\mathrm{restricted}}$, one requires that $v^{\mathsf{T}}\Sigma_k v \ge \lambda_{\mathrm{restricted}}$ for all $2s$-sparse vectors v on the unit sphere: $v \in \mathbb{S}^{p-1}$ and $\|v\|_0 \le 3s$. Obviously the requirements on $\lambda_{\mathrm{restricted}}$ are much milder than the corresponding ones on $\lambda_{\min}(\Sigma_k)$ and we always have $\lambda_{\mathrm{restricted}} \ge \lambda_{\min}(\Sigma_k)$.

The following lemma is central. Namely, we begin the proof of Proposition VI.1 by restricting to an event in which the designs $\sum_{t=1}^{T}(X_t)_S(X_t)_S^{\mathsf{T}}$ are sufficiently well-conditioned for all the subspaces S at once. The requirements on this event are relatively milder than the corresponding one over $\mathbb{R}^p$ and explains the "dimensional win" (when $s \ll p$) of the sparse estimator over OLS.

Lemma VI.1. *Let $\mathbf{L}$ be the linear operator defined in (48). Fix $\delta \in (0,1)$ and let T be divisible by $k \in \mathbb{N}$. There exist universal positive constants $c_1, c_2, c_3 \in \mathbb{R}$ such that the following two-sided control holds uniformly in S with probability $1-\delta$:*

$$\begin{aligned}\frac{c_1}{k}\sum_{t=1}^{k}\mathbf{E}\left[(X_t)_S(X_t)_S^{\mathsf{T}}\right] &\preceq \frac{1}{T}\sum_{t=1}^{T}(X_t)_S(X_t)_S^{\mathsf{T}}\\ &\preceq c_2\left(1 + \frac{T\|\mathbf{L}\mathbf{L}^{\mathsf{T}}\|_{\mathsf{op}}}{k\lambda_{\min}\left(\sum_{t=0}^{T-1}\mathbf{E}X_tX_t^{\mathsf{T}}\right)}\right)\\ &\quad\times\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{E}\left[(X_t)_S(X_t)_S^{\mathsf{T}}\right]\right)\end{aligned} \quad (76)$$

as long as

$$T \ge c_3K^2\left(s\left[\log C_{\mathsf{sys}} + \log(p/s)\right] + \log(1/\delta)\right). \quad (77)$$

Equation (77) is revealing about the advantage of using the sparse estimator searching over $\mathsf{M} = \{\theta \in \mathbb{R}^p : \|\theta\|_0 \le s\}$. The burn-in period in (77) is proportional to the dimension of the low-dimensional parameter manifold M instead of that of the latent space $\mathbb{R}^p$. Finally, as usual we have relegated the full proof of Proposition VI.1 to the appendix, see **??**.

B. Notes

The variational formulation of the least squares error—the basic inequality (67)—is standard in the nonparametric statistics literature [see e.g. 37, Chapters 13 and 14]. The idea to rewrite the basic inequality (67) as (68) was introduced to the statistical literature by [13], but has its roots in online learning [20].

[6] Note that $\widehat{\theta} - \theta^\star$ is $2s$-sparse.

VII. Beyond Linear Models

Let us now make another gradual shift of perspective. Instead of considering the linear model (1) introduced in Section I-A we consider the following *nonlinear* regression model:

$$Y_t = f^\star(X_t) + V_t, \qquad t \in [T]. \tag{78}$$

As before, $Y_{1:T}$,$X_{1:T}$ and $V_{1:T}$ are stochastic processes taking values in $\mathbb{R}^{d_{\mathsf{Y}}}$ and $\mathbb{R}^{d_{\mathsf{X}}}$ respectively. However, this time $f^\star$ is no longer constrained to be a linear map of the form $x \mapsto Ax$ for matrix A. Rather, we suppose that $f^\star$ in (78) belongs to some (square integrable) space of functions $\mathscr{F}$ such that $\mathscr{F} \ni f : x \mapsto f(x)$. It is perhaps now that the motivation behind the change of perspective from Section VI becomes most apparent: the basic inequality (68) remains valid. To be precise, let us define the *nonparametric* least squares estimator

$$\widehat{f} \in \operatorname*{argmin}_{f \in \mathscr{F}} \left\{ \frac{1}{T} \sum_{t=1}^{T} \|Y_t - f(X_t)\|_2^2 \right\}. \tag{79}$$

Let $\mathscr{F}_\star \triangleq \mathscr{F} - f^\star$. By the exact same optimality argument as in Section VI, the reader can now readily verify that:

$$\begin{aligned} &\frac{1}{T} \sum_{t=1}^{T} \|\widehat{f}(X_t) - f^\star(X_t)\|_2^2 \\ &\quad \le \sup_{f \in \mathscr{F}_\star} \frac{1}{T} \left(\sum_{t=1}^{T} 4\langle V_t, f(X_t)\rangle - \sum_{t=1}^{T} \|f(X_t)\|_2^2 \right). \end{aligned} \tag{80}$$

What does (80) entail in terms of estimating the unknown function $f^\star$? To answer this, we first need to define a performance criterion. The simplest one is small average L^2-norm-error, where

$$f \in \mathscr{F} : \quad \|f\|_{L^2}^2 \triangleq \frac{1}{T} \sum_{t=1}^{T} \mathbf{E}\|f(X_t)\|_2^2. \tag{81}$$

The program we have carried out in the previous sections now generalizes as follows:

- First, prove a so-called lower uniform law. That is to say, we wish to show that with overwhelming probability

$$\|f\|_{L^2}^2 \le \frac{C}{T} \sum_{t=1}^{T} \|f(X_t)\|^2 \quad \text{(simultaneously } \forall f \in \mathscr{F}_\star). \tag{82}$$

  for some universal positive constant C.
- Second, control the supremum of the *empirical process*:

$$f \mapsto \left(\sum_{t=1}^{T} 4\langle V_t, f(X_t)\rangle - \sum_{t=1}^{T} \|f(X_t)\|_2^2 \right) \tag{83}$$

  in terms of the noise level σ and some complexity measure $\mathrm{comp}(\mathscr{F})$.

By combining (82) and (83) we arrive at a high probability bound of the form:

$$\begin{aligned} \|\widehat{f} - f^\star\|_{L^2}^2 &\le \frac{C}{T} \sum_{t=1}^{T} \|f(X_t)\|^2 \\ &\le \frac{C \times \mathrm{comp}(\mathscr{F}, \sigma^2) + \text{deviation term}}{T}. \end{aligned} \tag{84}$$

A statement of this form is given as Theorem VII.1 below.

Remark VII.1. *It is worth to take pause and appreciate the analogy to the analysis of linear regression models. The first step (82) exactly corresponds to controlling the lower spectrum of the empirical covariance matrix. Suppose for simplicity that $d_{\mathsf{Y}} = 1$. Then for a linear map $\mathbb{S}^{d_{\mathsf{X}}-1} \ni f \mapsto \langle f, x\rangle$ we have:*

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^{T} \|f(X_t)\|_2^2 &= \frac{1}{T} \sum_{t=1}^{T} \langle f, (X_t X_t^{\mathsf{T}}) f\rangle \\ &= \left\langle f, \left[\frac{1}{T} \sum_{t=1}^{T} (X_t X_t^{\mathsf{T}}) \right] f \right\rangle \end{aligned} \tag{85}$$

which are just the one-dimensional projections of the empirical covariance matrix (9). In the context of linear models, establishing (82) was the topic of Section III. Analogously, for a linear predictor, the L^2-norm (81) becomes a Mahalanobis norm: $f \in \mathbb{R}^{d_{\mathsf{X}}} \Rightarrow \|f\|_{L^2}^2 = \langle f, \Sigma_X f\rangle$ for some $\Sigma_X = \frac{1}{T}\sum_{t=1}^{T} \mathbf{E} X_t X_t^{\mathsf{T}}$.

Moreover, For linear models, we had:

$$\begin{aligned} &\sup_{f \in \mathbb{R}^{d_{\mathsf{X}}}} \left(\sum_{t=1}^{T} 4\langle V_t, f(X_t)\rangle - \sum_{t=1}^{T} \|f(X_t)\|_2^2 \right) \\ &\qquad = 4 \left\| \left(\sum_{t=1}^{T} V_t X_t^{\mathsf{T}} \right) \left(\sum_{t=1}^{T} X_t X_t^{\mathsf{T}} \right)^{-1/2} \right\|_F^2. \end{aligned} \tag{86}$$

Analyzing terms of this form was the topic of Section IV.

In other words, the approach outlined above is very much in the same spirit as that in the rest of the manuscript. There are a few changes that need to be made since we less access to linearity in our argument, but in principle the key difference is that we will have to replace the indexing set $\mathbb{S}^{d-1}$ with a more general function class $\mathscr{F}_\star$.

A. Many Trajectories and Finite Hypothesis Classes

In order to make the exposition self-contained, we will now make two simplifying assumptions relating to the finiteness of the hypothesis class and the dependence structure of the covariate process $X_{1:T}$. A more general treatment without these can be found in [39]. Here, we impose the following:

- A1. The hypothesis class $\mathscr{F}$ is finite.
- A2. We have access to T/k-many independent trajectories from the same process: there exists an integer $k \in \mathbb{N}$ dividing T such that $X_{1:k}, X_{k+1:2k}, \ldots$ are drawn iid.

We will also impose the following rather minimal integrability condition:

A3. All functions $f \in \mathscr{F}$ are such that $\mathbf{E}\|f(X_t)\|_2^4 < \infty$ for all $t \in [T]$.

Moreover, as in Section V, we require the noise to be a sub-Gaussian martingale difference sequence:

A4. For each $t \in [T]$, $V_t|X_{1:t}$ is σ^2 conditionally-sub-Gaussian and mean zero.

Under these assumptions, the main result of [39] essentially simplifies to the following theorem.

Theorem VII.1. *Impose A1-A4, fix $\delta \in (0,1)$ and define*

$$\mathrm{cond}_{\mathscr{F}} \triangleq \max_{f \in \mathscr{F}_\star} \max_{t \in T} \frac{\sqrt{\mathbf{E}\|f(X_t)\|_2^4}}{\mathbf{E}\|f(X_t)\|_2^2}. \tag{87}$$

Suppose further that

$$T/k \geq 4\mathrm{cond}_{\mathscr{F}}^2 \left(\log|\mathscr{F}| + \log(2/\delta)\right)$$

then we have that:

$$\|\widehat{f} - f^\star\|_{L^2}^2 \leq 16\sigma^2 \left(\frac{\log(|\mathscr{F}|) + \log(2/\delta)}{T}\right). \tag{88}$$

A few remarks are in order. The structure of Theorem VII.1 is by now familiar and it is very much of the same structure as our previous results, cf. (5). The key differences are that: (1) we now control the L^2 norm of our estimator instead of the Euclidean or spectral norm; and (2) that the dimensional dependency has been replaced by the complexity term $\log|\mathscr{F}_\star|$. The proof is also structurally similar, as noted in Remark VII.1. We also caution the reader that (88) is strictly a statistical guarntee; we have said nothing—and will say nothing more—about the computational feasibility of the estimator (79).

Let us now discuss A1-A4. Assumption A1 informs us that the search space for the LSE (79) is finite. This is mainly imposed to avoid the introduction of the chaining technique which is the standard alternative to the bounds from Section IV. Using this technique, similar statements can for instance be derived for compact subsets of bounded function classes [39]. Assumption A2 controls the dependence structure of the process. Here, we assume that we are able to restart the process every k time steps. Again, a more general statement relying on stochastic stability can be found in [39]. Assumption A3 is relatively standard. Arguably the strongest assumption is A4, which in principle yields that the conditional expectation (given all past data) is a function in the search space $\mathscr{F}$. It is a so-called realizability assumption—the model (78) is well-specified—and it is not currently known how to remove it and still obtain sharp bounds beyond linear classes [for an analysis of linear misspecified models, see 40].

B. Notes

As noted in the previous section, the idea of using the "offset" basic inequality relied on here is due to [20, 13]. The "many trajectores"-style of analysis used here is due to [32] who introduced it in the linear setting. Here, we have extended their style of analysis to simplify the exposition of [39] who consider the single trajectory setting, but rely on a rather more advanced exponential inequality due to [23]. Note however that all the analyses above and in this section necessitate some degree of stability (mixing). This should be contrasted with the system identification bounds of Section V, which work even in the marginally regime. In principle, the consequence of this is that while the convergence rates for bounds such as Theorem VII.1 are correct, the burn-ins are deflated by various dependency measures.

There have also been other, more algorithmically focused, approaches to nonlinear identification problems in the recent literature. Noteably, gradient based methods in generalized linear models of the form $X_{t+1} = \phi(A^\star X_t) + V_t$ (with ϕ a known nonlinearity) have been the topic of a number of recent papers [see e.g. 5, 26]. The sharpest bounds for parameter recovery in this setting are due to [9].

References

[1] Yasin Abbasi-Yadkori. Online learning for linearly parametrized control problems. 2013.

[2] Brian DO Anderson and John B Moore. *Optimal Filtering*. Courier Corporation, 2012.

[3] Zhe Du, Zexiang Liu, Jack Weitze, and Necmiye Ozay. Sample complexity analysis and self-regularization in identification of over-parameterized ARX models. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 6026–6033. IEEE, 2022.

[4] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018.

[5] Dylan Foster, Tuhin Sarkar, and Alexander Rakhlin. Learning nonlinear dynamical systems from a single trajectory. In *Learning for Dynamics and Control*, pages 851–861. PMLR, 2020.

[6] David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

[7] Magnus Jansson. Subspace identification and ARX modeling. *IFAC Proceedings Volumes*, 36(16):1585–1590, 2003.

[8] Yassir Jedra and Alexandre Proutiere. Finite-time identification of linear systems: Fundamental limits and optimal algorithms. *IEEE Transactions on Automatic Control*, 2022.

[9] Suhas Kowshik, Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. Near-optimal offline and streaming algorithms for learning non-linear dynamical systems. *Advances in Neural Information Processing Systems*, 34, 2021.

[10] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Finite-time system identification and adaptive control in autoregressive exogenous systems. In *Learning for Dynamics and Control*, pages 967–979. PMLR, 2021.

[11] Bruce Lee and Andrew Lamperski. Non-asymptotic Closed-Loop System Identification using Autoregressive

Processes and Hankel Model Reduction. In *IEEE 59th Conference on Decision and Control (CDC)*, 2020.

[12] Holden Lee. Improved rates for prediction and identification of partially observed linear dynamical systems. In *International Conference on Algorithmic Learning Theory*, pages 668–698. PMLR, 2022.

[13] Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285. PMLR, 2015.

[14] Lennart Ljung. System identification: theory for the user. *PTR Prentice Hall, Upper Saddle River, NJ*, 28, 1999.

[15] Shahar Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39. PMLR, 2014.

[16] Samet Oymak and Necmiye Ozay. Non-asymptotic identification of LTI systems from a single trajectory. In *2019 American control conference (ACC)*, pages 5655–5661. IEEE, 2019.

[17] Samet Oymak and Necmiye Ozay. Revisiting Ho–Kalman-based system identification: Robustness and finite-sample analysis. *IEEE Transactions on Automatic Control*, 67(4):1914–1928, 2021.

[18] Victor H Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer, 2009.

[19] S Joe Qin. An overview of subspace identification. *Computers & chemical engineering*, 30(10-12):1502–1513, 2006.

[20] Alexander Rakhlin and Karthik Sridharan. Online nonparametric regression. In *Conference on Learning Theory*, pages 1232–1264. PMLR, 2014.

[21] Herbert Robbins and David Siegmund. Boundary crossing probabilities for the wiener process and sample sums. *The Annals of Mathematical Statistics*, pages 1410–1429, 1970.

[22] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

[23] Paul-Marie Samson. Concentration of measure inequalities for markov chains and ϕ-mixing processes. *The Annals of Probability*, 28(1):416–461, 2000.

[24] Tuhin Sarkar and Alexander Rakhlin. Near Optimal Finite Time Identification of Arbitrary Linear Dynamical Systems. In *International Conference on Machine Learning*, pages 5610–5618, 2019.

[25] Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. Finite time LTI system identification. *Journal of Machine Learning Research*, 22(26):1–61, 2021.

[26] Yahya Sattar and Samet Oymak. Non-asymptotic and accurate learning of nonlinear dynamical systems. *The Journal of Machine Learning Research*, 23(1):6248–6296, 2022.

[27] Max Simchowitz, Horia Mania, Stephen Tu, Michael I. Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

[28] Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning Linear Dynamical Systems with Semi-Parametric Least Squares. In *Conference on Learning Theory*, pages 2714–2802. PMLR, 2019.

[29] Anastasios Tsiamis and George J. Pappas. Finite sample analysis of stochastic system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3648–3654. IEEE, 2019.

[30] Anastasios Tsiamis and George J Pappas. Linear systems can be hard to learn. *arXiv preprint arXiv:2104.01120*, 2021.

[31] Anastasios Tsiamis, Ingvar Ziemann, Nikolai Matni, and George J Pappas. Statistical learning theory for control: A finite sample perspective. *to appear: IEEE Control Systems Magazine*, 2023.

[32] Stephen Tu, Roy Frostig, and Mahdi Soltanolkotabi. Learning from many trajectories. *arXiv preprint arXiv:2203.17193*, 2022.

[33] Aad W van der Vaart. *Asymptotic Statistics*. Cambridge university press, 2000.

[34] Michel Verhaegen and Vincent Verdult. *Filtering and system identification: a least squares approach*. Cambridge university press, 2007.

[35] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

[36] Andrew Wagenmaker and Kevin Jamieson. Active learning for identification of linear dynamical systems. In *Conference on Learning Theory*, pages 3487–3582. PMLR, 2020.

[37] Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[38] Ingvar Ziemann. A note on the smallest eigenvalue of the empirical covariance of causal gaussian processes. *arXiv preprint arXiv:2212.09508*, 2022.

[39] Ingvar Ziemann and Stephen Tu. Learning with little mixing. *arXiv preprint arXiv:2206.08269. NeurIPS'22*, 2022.

[40] Ingvar Ziemann, Stephen Tu, George J Pappas, and Nikolai Matni. The noise level in linear regression with dependent data. *arXiv preprint arXiv:2305.11165*, 2023.