Near optimal finite time identification of arbitrary linear dynamical systems

Tuhin Sarkar [1] **Alexander Rakhlin** [2]

Abstract

We derive finite time error bounds for estimating general linear time-invariant (LTI) systems from a single observed trajectory using the method of least squares. We provide the first analysis of the general case when eigenvalues of the LTI system are arbitrarily distributed in three regimes: stable, marginally stable, and explosive. Our analysis yields sharp upper bounds for each of these cases separately. We observe that although the underlying process behaves quite differently in each of these three regimes, the systematic analysis of a self–normalized martingale difference term helps bound identification error up to logarithmic factors of the lower bound. On the other hand, we demonstrate that the least squares solution may be statistically inconsistent under certain conditions even when the signal-to-noise ratio is high.

1 Introduction

Finite time system identification—the problem of estimating the parameters of an unknown dynamical system given a finite time series of its output—is an important problem in the context of time-series analysis, control theory, economics and reinforcement learning. In this work we will focus on obtaining sharp non–asymptotic bounds for *linear* dynamical system identification using the ordinary least squares (OLS) method. Such a system is described by $X_{t+1} = AX_t + \eta_{t+1}$ where $X_t \in \mathbb{R}^d$ is the state of the system and η_t is the unobserved process noise. The goal is to learn A by observing only X_t's. Our techniques can easily be extended to the more general case when there is a control input U_t, *i.e.*, $X_{t+1} = AX_t + BU_t + \eta_{t+1}$. In this case (A, B) are unknown, and we can choose U_t.

Linear systems are ubiquitous in control theory. For example, proportional-integral-derivative (PID) controller is a popular linear feedback control system found in a variety of devices, from planetary soft landing systems for rockets (see e.g. (Açıkmeşe et al., 2013)) to coffee machines. Further, linear approximations to many non–linear systems have been known to work well in practice. Linear systems also appear as auto–regressive (AR) models in time series analysis and econometrics. Despite its importance, sharp non–asymptotic characterization of identification error in such models was relatively unknown until recently.

In the statistics literature, correlated data is often dealt with using mixing–time arguments (see e.g. (Yu, 1994)). However, a fundamental limitation of the mixing-time method is that bounds deteriorate when the underlying process mixes slowly. For discrete linear systems, this happens when $\rho(A)$—the spectral radius of A—approaches 1. As a result these methods cannot extend to the case when $\rho(A) \geq 1$. More recently there has been renewed effort in obtaining sharp non–asymptotic error bounds for linear system identification (Faradonbeh et al., 2017; Simchowitz et al., 2018). Specifically, (Faradonbeh et al., 2017) analyzed the case when the system is either stable ($\rho(A) < 1$) or purely explosive ($\rho(A) > 1$). For the case when $\rho(A) < 1$ the techniques in (Faradonbeh et al., 2017) are similar to the standard mixing time arguments and, as a result, suffer from the same limitations. When the system is purely explosive, the authors of (Faradonbeh et al., 2017) show that finite time identification is only possible if the system is regular, *i.e.*, if the geometric multiplicity of eigenvalues greater than unity is one. However, as discussed in (Simchowitz et al., 2018), the bounds obtained in (Faradonbeh et al., 2017) are suboptimal due to a decoupled analysis of the sample covariance, $\sum_{t=1}^{T} X_t X_t'$, and the martingale difference term $\sum_{t=1}^{T} X_t \eta_{t+1}'$. A second approach, based on Mendelson's small–ball method, was studied in (Simchowitz et al., 2018). Such a technique eschewed the need for mixing-time arguments and sharper error bounds for $1 - C/T \leq \rho(A) \leq 1 + C/T$ could be obtained. The authors in (Simchowitz et al., 2018) argue that a larger signal-to-noise ratio, measured by $\lambda_{\min}(\sum_{t=0}^{T-1} A^t A^{t\prime})$, makes it easier to estimate A. Although this intuition is consistent for the case when $\rho(A) \leq 1$, it does not extend to the case when eigenvalues are far outside the unit circle. Since $X_T = \sum_{t=1}^{T} A^{T-t}\eta_t$, the behavior of X_T is dominated by $\{\eta_1, \eta_2, \ldots\}$, *i.e.*, the past, due to exponential scaling by

[1]Department of Electrical Engineering and Computer Sciences, MIT [2]Department of Brain and Cognitive Sciences, MIT. Correspondence to: Tuhin Sarkar <tsarkar@mit.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

$\{A^{T-1}, A^{T-2}, \ldots\}$. As a result, X_1 depends strongly on $\{X_2, \ldots, X_T\}$ and standard techniques of creating "independent" blocks of covariates fail.

The problem of system identification has received a lot of attention. Asymptotic results on identification of AR models can be found in (Lai & Wei, 1983). Some of the earlier work on finite time identification in systems theory include (Campi & Weyer, 2002; Vidyasagar & Karandikar, 2006). A more general setting of the problem considered here is when X_t is observed indirectly via its filtered version, *i.e.*, $Y_t = CX_t$ where C is unknown. The single input single output (SISO) version of this problem, *i.e.*, when Y_t, U_t are numbers, has been studied in (Hardt et al., 2016) under the assumption that system is stable. Provable guarantees for system identification in general linear systems was also studied in (Oymak & Ozay, 2018). However, the analysis there requires that $||A|| < 1$. Generalization bounds for time series forecasting of non–stationary and non–mixing processes have been developed in (Kuznetsov & Mohri, 2018).

2 Contributions

In this paper we offer a new statistical analysis of the ordinary least squares estimator of the dynamics $X_{t+1} = AX_t + \eta_{t+1}$ with no inputs. Unlike previous work, we do not impose any restrictions on the spectral radius of A and provide nearly optimal rates (up to logarithmic factors) for every regime of $\rho(A)$. The contributions of our paper can be summarized as follows

- At the center of our techniques is a systematic analysis of the sample covariance $\sum_{t=1}^{T} X_t X_t'$ and a certain self normalized martingale difference term. Although such a coupled analysis is similar in flavor to (Simchowitz et al., 2018), it comes without the overhead of choosing a block size and applies to a general case when covariates grow exponentially in time.
- Specifically, for the case when $\rho(A) \leq 1$, we recover the optimal finite time identification error rates previously derived in (Simchowitz et al., 2018). For the case when all eigenvalues are outside the unit circle, we argue that small ball methods cannot be used. Instead we use anti–concentration arguments discussed in (Faradonbeh et al., 2017; Lai & Wei, 1983). By leveraging subgaussian tail inequalities we sharpen previous error bounds by removing polynomial factors. We also show that this analysis is indeed tight by deriving a matching lower bound.
- We provide the first analysis of the general case when eigenvalues of A are arbitrarily distributed in three regimes: stable, marginally stable and explosive. This involves a careful analysis of the noise-covariate cross terms as the underlying process behaves differently in each of these regimes.
- We show that when A does not satisfy certain regularity conditions, OLS identification is statistically inconsistent, even when signal-to-noise ratio is high. Our result indicates that consistency of OLS identification depends on the condition number of the sample covariance matrix, rather than the signal-to-noise ratio itself.

3 Notation and Definitions

A linear time invariant system (LTI) is parametrized by a matrix, A, where the observed variable, X_t, indexed by t evolves as

$$X_{t+1} = AX_t + \eta_{t+1}. \tag{1}$$

Here η_t is the noise process. Denote by $\rho_i(A)$ the absolute value of the i^{th} eigenvalue of the $d \times d$ matrix A. Then

$$\rho_{\max}(A) = \rho_1(A) \geq \rho_2(A) \geq \ldots \geq \rho_d(A) = \rho_{\min}(A).$$

Similarly the singular values of A are denoted by $\sigma_i(A)$. For any matrix M, $||M||_{\text{op}} = ||M||_2$.

Definition 1. *A stable LTI system is that where $\rho_{\max}(A) < 1$. An explosive LTI system is that where $\rho_{\min}(A) > 1$.*

For simplicity of exposition, we assume that $X_0 = 0$ with probability 1. All the results can be obtained by assuming X_0 to be some bounded vector.

Definition 2. *A random vector $X \in \mathbb{R}^d$ is called isotropic if for all $x \in \mathbb{R}^d$ we have*

$$\mathbb{E}\langle X, x\rangle^2 = ||x||_2^2$$

Assumption 1. *$\{\eta_t\}_{t=1}^{\infty}$ are i.i.d isotropic subgaussian and coordinates of η_t are i.i.d. Further, let $f(x)$ be the pdf of each noise coordinate then the essential supremum of $f(\cdot)$ is bounded above by $C < \infty$.*

We will deal with only regular systems, *i.e.*, LTI systems where eigenvalues of A with absolute value greater than unity have geometric multiplicity one. We will show that when A is not regular, OLS is statistically inconsistent.

Define the data matrix $\mathbf{X}$ and the noise matrix E as

$$\mathbf{X} = \begin{bmatrix} X_0' \\ X_1' \\ \vdots \\ X_T' \end{bmatrix}, \quad E = \begin{bmatrix} \eta_1' \\ \eta_2' \\ \vdots \\ \eta_{T+1}', \end{bmatrix}$$

where the superscript a' denotes the transpose. Then $\mathbf{X}$, E are $(T+1) \times d$ matrices. Consider the OLS solution

$$\hat{A} = \arg\min_{B} \sum_{t=0}^{T} ||X_{t+1} - BX_t||_2^2.$$

One can show that

$$A - \hat{A} = ((\mathbf{X}'\mathbf{X})^{+}\mathbf{X}'E)' \tag{2}$$

where M^+ is the pseudo inverse of M. We define

$$Y_T = \mathbf{X}'\mathbf{X} = \sum_{t=0}^{T} X_t X_t', \quad S_T = \mathbf{X}'E = \sum_{t=0}^{T} X_t \eta_{t+1}'.$$

To analyze the error in estimating A, we will aim to bound the norm of $(\mathbf{X}'\mathbf{X})^+\mathbf{X}'$.

We will occasionally replace X_t (or $X(t)$) with the lowercase counterparts x_t (or $x(t)$) to denote state at time t, whenever this does not cause confusion. Further, we will use C, c to indicate universal constants that can change from line to line. Define the *Gramian* as

$$\Gamma_t(A) = \sum_{k=0}^{t} A^k A^{k\prime} \tag{3}$$

and a Jordan block matrix $J_d(\lambda)$ as

$$J_d(\lambda) = \begin{bmatrix} \lambda & 1 & 0 & \dots & 0 \\ 0 & \lambda & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \lambda & 1 \\ 0 & 0 & \dots & 0 & \lambda \end{bmatrix}_{d\times d} \tag{4}$$

We present the three classes of matrices that will be of interest to us:

- The perfectly stable matrix class, $\mathcal{S}_0$

$$\rho_i(A) \le 1 - \frac{C}{T}$$

  for $1 \le i \le d$.
- The marginally stable matrix, $\mathcal{S}_1$

$$1 - \frac{C}{T} < \rho_i(A) \le 1 + \frac{C}{T}$$

  for $1 \le i \le d$.
- The regular and explosive matrix, $\mathcal{S}_2$

$$\rho_i > 1 + \frac{C}{T}$$

  for $1 \le i \le d$.

Slightly abusing the notation, whenever we write $A \in \mathcal{S}_i \cup \mathcal{S}_j$ we mean that A has eigenvalues in both $\mathcal{S}_i, \mathcal{S}_j$.

Critical to obtaining refined error rates, will be a result from the theory of self–normalized martingales. We let $\boldsymbol{\mathcal{F}}_t = \sigma(\eta_1, \eta_2, \dots, \eta_t, X_1, \dots, X_t)$ to denote the filtration generated by the noise and covariate process.

Proposition 3.1. *Let V be a deterministic matrix with $V \succ 0$. For any $0 < \delta < 1$ and $\{\eta_t, X_t\}_{t=1}^T$ defined as before, we have with probability $1-\delta$*

$$||(\bar{Y}_{T-1})^{-1/2} \sum_{t=0}^{T-1} X_t \eta_{t+1}'||_2 \le R\sqrt{8d \log\left(\frac{5det(\bar{Y}_{T-1})^{1/2d} det(V)^{-1/2d}}{\delta^{1/d}}\right)} \tag{5}$$

where $\bar{Y}_\tau^{-1} = (Y_\tau + V)^{-1}$ and R^2 is the subGaussian parameter of η_t.

The proof can be found in appendix as Proposition 9.2. It rests on Theorem 1 in (Abbasi-Yadkori et al., 2011) which is itself an application of the pseudo-maximization technique in (Peña et al., 2008) (see Theorem 14.7).

Finally, we define several A-dependent quantities that will appear in time complexities in the next section.

Definition 3 (Outbox Set)**.** *For the space $\mathbb{R}^d$ define the a–outbox, $S_d(a)$, as the following set*

$$S_d(a) = \{v | \min_{1\le i\le d} |v_i| \ge a\}$$

$S_d(a)$ will be used to quantify the following norm–like quantities of a matrix:

$$\phi_{\min}(A) = \sqrt{\inf_{v\in S_d(1)} \sigma_{\min}\Big(\sum_{i=1}^{T} \Lambda^{-i+1} v v' \Lambda^{-i+1\prime}\Big)} \tag{6}$$

$$\phi_{\max}(A) = \sqrt{\sup_{||v||_2=1} \sigma_{\max}\Big(\sum_{i=1}^{T} \Lambda^{-i+1} v v' \Lambda^{-i+1\prime}\Big)} \tag{7}$$

where $A = P^{-1}\Lambda P$ is the Jordan normal form of A.

$\psi(A)$ is defined in Proposition 3.2 and is needed for error bounds for explosive matrices.

Proposition 3.2 (Proposition 2 in (Faradonbeh et al., 2017))**.** *Let $\rho_{\min}(A) > 1$ and $P^{-1}\Lambda P = A$ be the Jordan decomposition of A. Define $z_T = A^{-T}\sum_{i=1}^{T} A^{T-i}\eta_i$ and*

$$\psi(A,\delta) = \sup\left\{y \in \mathbb{R} : \mathbb{P}\Big(\min_{1\le i\le d} |P_i' z_T| < y\Big) \le \delta\right\}$$

where $P = [P_1, P_2, \dots, P_d]'$. Then

$$\psi(A,\delta) \ge \psi(A)\delta > 0$$

Here $\psi(A) = \frac{1}{2d \sup_{1\le i\le d} C_{|P_i' z_T|}}$ where C_X is the essential supremum of the pdf of X.

We summarize some definitions in Table 1 for convenience in representing our results.

$$
\begin{aligned}
&T_\eta(\delta) = C\Big(\log\tfrac{2}{\delta} + d\log 5\Big)\\
&T_s(\delta) = C\Big(d\log\left(\mathrm{tr}(\Gamma_T(A)\right) + 1) + 2d\log\tfrac{5}{\delta}\Big)\\
&c(A,\delta) = T_s(\tfrac{2\delta}{3T})\\
&\beta_0(\delta) = \inf\Big\{\beta | \beta^2\sigma_{\min}(\Gamma_{\lfloor\frac{1}{\beta}\rfloor}(A)) \geq \Big(\tfrac{16ec(A,\delta)}{T\sigma_{\min}(AA')}\Big)\Big\}\\
&T_{ms}(\delta) = \inf\Big\{T\Big|T \geq \tfrac{Cc(A,\delta)}{\sigma_{\min}(AA')}\Big\}\\
&T_u(\delta) = \Big\{T\Big|\Big(4T^2\sigma_1^2(A^{-\lfloor\frac{T+1}{2}\rfloor})\mathrm{tr}(\Gamma_T(A^{-1})) + \tfrac{T\mathrm{tr}(A^{-T-1}\Gamma_T(A^{-1})A^{-T-1\prime})}{\delta}\Big) \leq \tfrac{\phi_{\min}(A)^2\psi(A)^2\delta^2}{2\sigma_{\max}(P)^2}\Big\}\\
&\gamma(A,\delta) = \tfrac{4\phi_{\max}(A)^2\sigma_{\max}^2(A)}{\phi_{\min}(A)^2\sigma_{\min}^2(A)\psi(A)^2\delta^2}(1+\tfrac{1}{c}\log\tfrac{1}{\delta})\mathrm{tr}(P(\Gamma_T(A^{-1}))P')I\\
&\gamma_s(A,\delta) = \sqrt{8d\Big(\log\Big(\tfrac{5}{\delta}\Big) + \tfrac{1}{2}\log\Big(4\mathrm{tr}(\Gamma_T(A)) + 1\Big)\Big)}\\
&\gamma_{ms}(A,\delta) = \sqrt{16d\log\left(\mathrm{tr}(\Gamma_T(A)\right) + 1) + 32d\log\Big(\tfrac{15T}{2\delta}\Big)}\\
&\gamma_e(A,\delta) = \tfrac{\sqrt{d}\sigma_{\max}(P)}{\phi_{\min}(A)\psi(A)\delta}\sqrt{\log\tfrac{2}{\delta} + 2\log 5 + \log\left(1+\gamma(A,\delta)\right)}
\end{aligned}
$$

Table 1. Definitions of key quantities in the paper

4 Main Results

We will first show non–asymptotic rates for the three separate regimes, followed by the case when A has a general eigenvalue distribution.

Theorem 1. *The following non-asymptotic bounds hold, with probability at least $1-\delta$, for the least squares estimator:*

- *For $A \in \mathcal{S}_0 \cup \mathcal{S}_1$*

$$||A - \hat{A}||_2 \leq \sqrt{\frac{C}{T}}\underbrace{\gamma_s\Big(A, \frac{\delta}{4}\Big)}_{=O(\sqrt{\log(\frac{1}{\delta})})}$$

  whenever $T \geq \max\Big(T_\eta\Big(\frac{\delta}{4}\Big), T_s\Big(\frac{\delta}{4}\Big)\Big)$.

- *For $A \in \mathcal{S}_1$*

$$||A - \hat{A}||_2 \leq \frac{C\sigma_{\max}(A^{-1})}{\sqrt{T\sigma_{\min}(\Gamma_{\lfloor\frac{1}{\beta_0(\delta)}\rfloor}(A))}}\underbrace{\gamma_{ms}\Big(A, \frac{\delta}{2}\Big)^2}_{=O(\log(\frac{T}{\delta}))}$$

  whenever

$$T \geq \max\Big(\underbrace{2T_\eta\Big(\frac{\delta}{3T}\Big)}_{=O(\log T)}, \underbrace{2T_s\Big(\frac{\delta}{3T}\Big)}_{=O(\log T)}, \underbrace{T_{ms}\Big(\frac{\delta}{2}\Big)}_{=O(\log T)}\Big)$$

  Since $\sigma_{\min}(\Gamma_{\lfloor\frac{1}{\beta_0(\delta)}\rfloor}(A)) \geq \alpha(d)\frac{T}{\log T}$, we have that

$$||A - \hat{A}||_2 \leq \sqrt{\frac{\log T}{\alpha(d)}}\frac{\gamma_{ms}\Big(A, \frac{\delta}{2}\Big)^2}{T}$$

- *For $A \in \mathcal{S}_2$*

$$||A - \hat{A}||_2 \leq C\sigma_{\max}(A^{-T})\underbrace{\gamma_e\Big(A, \frac{\delta}{5}\Big)}_{=O(\frac{1}{\delta})}$$

  whenever $T \in T_u\Big(\frac{\delta}{5}\Big)$. Since $\sigma_{\max}(A^{-T}) \leq \alpha(d)(\rho_{\min}(A))^{-T}$ for $A \in \mathcal{S}_2$, the identification error decays exponentially with T.

Here C, c are absolute constants and $\alpha(d)$ is a function that depends only on d.

Remark 1. *$T_u(\delta)$ is a set where there exists a minimum $T_* < \infty$ such that $T \in T_u(\delta)$ whenever $T \geq T_*$. However, there might be $T < T_*$ for which the inequality of $T_u(\delta)$ holds. Whenever we write $T \in T_u(\delta)$ we mean $T \geq T_*$.*

Proof. We start by writing an upper bound

$$\begin{aligned}||A - \hat{A}||_{\mathrm{op}} &\leq ||Y_T^+ S_T||_{\mathrm{op}}\\ &\leq ||(Y_T^+)^{1/2}||_{\mathrm{op}}||(Y_T^+)^{1/2}S_T||_{\mathrm{op}}. \end{aligned} \tag{8}$$

The rest of the proof can be broken into two parts:

- Showing invertibility of Y_T and lower bounds on the least singular value
- Bounding the self-normalized martingale term given by $(Y_T^+)^{1/2}S_T$

The invertibility of Y_T is where most of the work lies. Once we have a tight characterization of Y_T, one can simply obtain the error bound by using Proposition 3.1. Here we sketch the basis of our approach. First, we find deterministic V_{up}, V_{dn}, T_0 such that

$$\mathcal{E}_0 = \{0 \prec V_{dn} \preceq Y_T \preceq V_{up}, T \geq T_0\} \tag{9}$$

$$\mathbb{P}(\mathcal{E}_0) \geq 1 - \delta \tag{10}$$

The next step is to bound the self–normalized term. Under $\mathcal{E}_0$, it is clear that Y_T is invertible and we have

$$(Y_T^+)^{1/2}S_T = Y_T^{-1/2}S_T.$$

Define event $\mathcal{E}_1$ in the following way

$$\mathcal{E}_1 = \left\{ ||S_T||_{(Y_T+V_{dn})^{-1}} \leq \sqrt{8d \log\left(\frac{5\det(Y_T V_{dn}^{-1}+I)^{1/2d}}{\delta^{1/d}}\right)} \right\}$$

It follows from Proposition 3.1 that $\mathbb{P}(\mathcal{E}_1) \geq 1-\delta$. Then

$$\mathcal{E}_0 \implies Y_T + V_{dn} \preceq 2Y_T \implies (Y_T+V_{dn})^{-1} \succeq \frac{1}{2}Y_T^{-1},$$

and we have that under $\mathcal{E}_0$

$$||S_T||_{Y_T^{-1}} \leq \sqrt{2}||S_T||_{(Y_T+V_{dn})^{-1}}.$$

Now considering the intersection $\mathcal{E}_0 \cap \mathcal{E}_1$, we get

$$\mathcal{E}_0 \cap \mathcal{E}_1 \implies \mathcal{E}_0 \cap \left\{ ||S_T||_{Y_T^{-1}} \leq \sqrt{16d \log\left(\frac{5\det(V_{up} V_{dn}^{-1}+I)^{1/2d}}{\delta^{1/d}}\right)} \right\} \tag{11}$$

We replaced the LHS of $\mathcal{E}_1$ by the lower bound obtained above and in the RHS replaced Y_T by its upper bound under $\mathcal{E}_0$, V_{up}. Further, observe that $\mathbb{P}(\mathcal{E}_0 \cap \mathcal{E}_1) \geq 1-2\delta$. Under $\mathcal{E}_0 \cap \mathcal{E}_1$ we get

$$||A-\hat{A}||_{\text{op}} \leq \underbrace{\frac{1}{\sigma_{\min}(V_{dn})}}_{\alpha_T} \underbrace{\sqrt{16d \log\left(\frac{5\det(V_{up} V_{dn}^{-1}+I)^{1/2d}}{\delta^{1/d}}\right)}}_{\beta_T} \tag{12}$$

where α_T goes to zero with T and β_T is typically a constant. This shows that OLS learns A with increasing accuracy as T grows. The deterministic V_{up}, V_{dn}, T_0 differ for each regime of $\rho(A)$ and typically depend on the probability threshold δ. We now sketch the approach for finding these for each regime.

Y_T behavior when $A \in \mathcal{S}_0 \cup \mathcal{S}_1$

The key step here is to characterize Y_T in terms of Y_{T-1}.

$$\begin{aligned} Y_T &= x_0 x_0^{'} + AY_{T-1}A^{'} + \\ &\quad + \sum_{t=0}^{T-1}(Ax_t\eta_{t+1}^{'} + \eta_{t+1}x_t^{'}A^{'}) + \sum_{t=1}^{T}\eta_t\eta_t^{'} \\ &\succeq AY_{T-1}A^{'} + \\ &\quad + \sum_{t=0}^{T-1}(Ax_t\eta_{t+1}^{'} + \eta_{t+1}x_t^{'}A^{'}) + \sum_{t=1}^{T}\eta_t\eta_t^{'}. \end{aligned} \tag{13}$$

Since $\{\eta_t\}_{t=1}^T$ are i.i.d. subgaussian we can show that $\sum_{t=1}^T \eta_t\eta_t'$ concentrates near $TI_{d\times d}$ with high probability. Using Proposition 3.1 once again, we will show that with high probability

$$\sum_{t=0}^{T-1}(Ax_t\eta_{t+1}^{'} + \eta_{t+1}x_t^{'}A^{'}) \succeq -\epsilon(AY_{T-1}A^{'} + \sum_{t=1}^{T}\eta_t\eta_t')$$

where $\epsilon \leq 1/2$ whenever $\rho_i(A) \leq 1 + C/T$ and $T \geq T_0$ for some T_0 depending only on A. As a result with high probability we have

$$\begin{aligned} Y_T &\succeq (1-\epsilon)AY_{T-1}A^{'} + (1-\epsilon)\sum_{t=1}^{T}\eta_t\eta_t' \\ &\succeq (1-\epsilon)\sum_{t=1}^{T}\eta_t\eta_t'. \end{aligned} \tag{14}$$

The details of this proof are provided in appendix as Section 10. When $1 - C/T \leq \rho_i(A) \leq 1 + C/T$ we note that the bound in Eq. (14) is not tight. The key to sharpening the lower bound is the following observation: for $T > \max\left(2T_\eta\left(\frac{\delta}{3T}\right), 2T_s\left(\frac{\delta}{3T}\right), T_{ms}\left(\frac{\delta}{2}\right)\right)$ we can ensure with high probability

$$\begin{aligned} \sum_{\tau=1}^{t}\eta_\tau\eta_\tau' &= tI \\ Y_t &\succeq (1-\epsilon)AY_{t-1}A^{'} + (1-\epsilon)tI \end{aligned} \tag{15}$$

simultaneously for all $t \geq T/2$. Then we will show that $\epsilon = \beta_0(\delta)$ in Table 1. The sharpening of ϵ from $1/2$ to $\beta_0(\delta)$ is only possible because all the eigenvalues of A are close to unity. In that case by successively expanding Eq. (15) we get

$$Y_T \succeq (1-\epsilon)^{1/\beta_0(\delta)}AY_{T/2-1}A^{'} + \frac{T}{2}\sum_{t=1}^{1/\beta_0(\delta)}(1-\epsilon)^t A^t A^{t\prime} \tag{16}$$

and then Eq. (16) can be reduced to

$$Y_T \succeq (1-\epsilon)^{1/\beta_0(\delta)}AY_{T/2-1}A^{'} + \frac{T(\Gamma_{1/\beta_0(\delta)}(A) - I)}{4e}.$$

We show that

$$1/\beta_0(\delta) \geq \frac{\alpha(d)TR^2\sigma_{\min}(AA')}{8ec(A,\delta)}$$

and by Proposition 8.5, $Y_T \succeq \alpha(d)T^2$ for some function $\alpha(\cdot)$ that depends only on d. The details of the proof are provided in appendix as Section 11.

To get deterministic upper bounds for Y_T with high probability, we note that

$$Y_T \preceq \text{tr}\left(\sum_{t=1}^{T}X_tX_t'\right)I.$$

Then we can use Hanson–Wright inequality or Markov inequality to get an upper bound as shown in appendix as Proposition 9.4.

Y_T behavior when $A \in \mathcal{S}_2$

The concentration arguments used to show the convergence for stable systems do not work for unstable systems. As discussed before $X_t = \sum_{\tau=1}^{T} A^{t-\tau}\eta_t$ and, consequently, X_T depends strongly on $X_1, X_2, \ldots$. Due to this dependence we are unable to use typical techniques where X_is are divided into roughly independent blocks of covariates. to obtain concentration results. Motivated by (Lai & Wei, 1983), we instead work by transforming x_t as

$$\begin{aligned} z_t &= A^{-t}x_t \\ &= x_0 + \sum_{\tau=1}^{t} A^{-\tau}\eta_\tau. \end{aligned} \tag{17}$$

The steps of the proof proceed as follows. Define

$$\begin{aligned} U_T &= A^{-T}\sum_{t=1}^{T} x_t x_t' A^{-T\prime} = A^{-T}Y_T A^{-T\prime} \\ &= \sum_{t=1}^{T} A^{-T+t} z_t z_t' A^{-T+t\prime} \\ F_T &= \sum_{t=0}^{T-1} A^{-t} z_T z_T^{'} A^{-t\prime} \end{aligned} \tag{18}$$

We show that

$$||F_T - U_T||_{\text{op}} \leq \epsilon.$$

Here ϵ decays exponentially fast with T. Then the lower and upper bounds of U_T can be shown by proving corresponding bounds for F_T. A necessary condition for invertibility of F_T is that the matrix A should be regular (in a later section we show that it is also sufficient). If A is regular, the deterministic lower bound for F_T is fairly straightforward and depends on $\phi_{\min}(A)$ defined in Definition 3. The upper bound can be obtained by using Hanson–Wright inequality. The complete steps are given in appendix as Section 12. □

The analysis presented here is sharper than (Faradonbeh et al., 2017) as we use subgaussian matrix inequalities such as Hanson–Wright Inequality (Theorem 4) to bound the error terms in contrast to uniformly bounding each noise variable and applying a less efficient Bernstein inequality. Another minor difference is that (Lai & Wei, 1983),(Faradonbeh et al., 2017) consider $||U_T - F_\infty||$ instead and as a result they require a martingale concentration argument to show the existence of z_∞.

Lower bounds for identification error when $\rho(A) \leq 1$ have been derived in (Simchowitz et al., 2018). In Table 1 and Theorem 1, the error in identification for explosive matrices depends on δ as $\frac{1}{\delta}$ unlike stable and marginally stable matrices where the dependence is $\log \frac{1}{\delta}$. Typical minimax analyses, such as the one in (Simchowitz et al., 2018), are unable to capture this relation between error and δ. Here we show that such a dependence is unavoidable:

Proposition 4.1. *Let $A = a \geq 1.1$ be a 1–D matrix and $\hat{A} = \hat{a}$ be its OLS estimate. Then whenever $Ca^2T^2a^{-T} > \delta^2$, we have with probability at least δ that*

$$|a - \hat{a}| \geq \frac{C(1-a^{-2})\delta}{-a^2(\log \delta)^3}$$

where C is a universal constant. If $Ca^2T^2a^{-T} \leq \delta^2$ then with probability at least δ we have

$$|a - \hat{a}| \geq \Big(\frac{C(1-a^{-2})}{-\delta \log \delta}\Big) a^{-T}$$

Our lower bounds indicate that $\frac{1}{\delta}$ is inevitable in Theorem 1, *i.e.*, when $Ca^2T^2a^{-T} \leq \delta^2$. Second, when $Ca^2T^2a^{-T} > \delta^2$, our bound sharpens Theorem B.2 in (Simchowitz et al., 2018). The proof and an explicit comparison is provided in Section 17.

For the general case we use a well known fact for matrices, namely, that there exists a similarity transform $\tilde{P}$ such that

$$A = \tilde{P}^{-1}\begin{bmatrix} A_e & 0 & 0 \\ 0 & A_{ms} & 0 \\ 0 & 0 & A_s \end{bmatrix}\tilde{P} \tag{19}$$

Here $A_e \in \mathcal{S}_0, A_{ms} \in \mathcal{S}_1, A_s \in \mathcal{S}_2$. Although one might be tempted to use Theorem 1 to provide error bounds, mixing between different components due to the transformation $\tilde{P}$ requires a careful analysis of identification error. We show that error bounds are limited by the slowest component as we describe below. We do not provide the exact characterization due to a shortage of space. The details are given in appendix as Section 14.

Theorem 2. *For any regular matrix A we have with probability at least $1-\delta$,*

- *For $A \in \mathcal{S}_1 \cup \mathcal{S}_2$, $||A - \hat{A}||_2 \leq \frac{poly(\log T, \log \frac{1}{\delta})}{T}$ whenever*

$$T \geq poly\Big(\log \frac{1}{\delta}\Big)$$

- *For $A \in \mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2$, $||A - \hat{A}||_2 \leq \frac{poly(\log T, \log \frac{1}{\delta})}{\sqrt{T}}$ whenever*

$$T \geq poly\Big(\log \frac{1}{\delta}\Big)$$

Here poly$(\cdot)$ is a polynomial function.

Proof. Define the partition of A as Eq. (19). Since

$$X_t = \sum_{\tau=1}^{t} A^{\tau-1}\eta_{t-\tau+1}$$
$$\tilde{X}_t = \tilde{P}^{-1}X_t = \sum_{\tau=1}^{t} \tilde{A}^{\tau-1}\underbrace{\tilde{P}^{-1}\eta_{t-\tau+1}}_{\tilde{\eta}_{t-\tau+1}} \tag{20}$$

then the transformed dynamics are as follows:

$$\tilde{X}_{t+1} = \tilde{A}\tilde{X}_t + \tilde{\eta}_{t+1}.$$

Here $\{\tilde{\eta}_t\}_{t=1}^T$ are still independent. Correspondingly we also have a partition for $\tilde{X}_t, \tilde{\eta}_t$

$$\tilde{X}_t = \begin{bmatrix} X_t^e \\ X_t^{ms} \\ X_t^s \end{bmatrix}, \tilde{\eta}_t = \begin{bmatrix} \eta_t^e \\ \eta_t^{ms} \\ \eta_t^s \end{bmatrix} \tag{21}$$

Then we have

$$\sum_{t=1}^{T} \tilde{X}_t\tilde{X}_t' = \sum_{t=1}^{T} \begin{bmatrix} X_t^e(X_t^e)' & X_t^e(X_t^{ms})' & X_t^e(X_t^s)' \\ X_t^{ms}(X_t^e)' & X_t^{ms}(X_t^{ms})' & X_t^{ms}(X_t^s)' \\ X_t^e(X_t^s)' & X_t^s(X_t^{ms})' & X_t^s(X_t^s)' \end{bmatrix} \tag{22}$$

The next step is to show the invertibility of $\sum_{t=1}^T \tilde{X}_t\tilde{X}_t'$. Although reminiscent of our previous set up, there are some critical differences. First, unlike before, coordinates of $\tilde{\eta}_t$, *i.e.*, $\{\eta_t^e, \eta_t^{ms}, \eta_t^s\}$ are not independent. A major implication is that it is no longer obvious that the cross terms between different submatrices, such as $\sum_{t=1}^T X_t^e(X_t^{ms})'$, go to zero. Our proof will have three major steps:

- First we will show that the diagonal submatrices are invertible. This follows from Theorem 1 by arguing that the result can be extended to a noise process $\{P\eta_t\}_{t=1}^T$ where $\{\eta_t\}_{t=1}^T$ are independent subgaussian and elements of η_t are also independent for all t. The only change will be the appearance of additional $\sigma_1^2(P)$ subgaussian parameter (See Corollary 9.1). We will then show that

$$X_{mss} = \sum_{t=1}^{T} \begin{bmatrix} X_t^{ms}(X_t^{ms})' & X_t^{ms}(X_t^s)' \\ X_t^s(X_t^{ms})' & X_t^s(X_t^s)' \end{bmatrix}$$

  is invertible. This will follow from Theorem 1 (its dependent extension). Specifically, since X_{mss} contains only stable and marginally stable components, it falls under $A \in \mathcal{S}_0 \cup \mathcal{S}_1$. It should be noted that since X_t^{ms}, X_t^s are not independent in general, the invertibility of X_{mss} can be shown only through Theorem 1. In a similar fashion, $\sum_{t=1}^T X_t^e(X_t^e)'$ is also invertible as it corresponds to $A \in \mathcal{S}_2$.

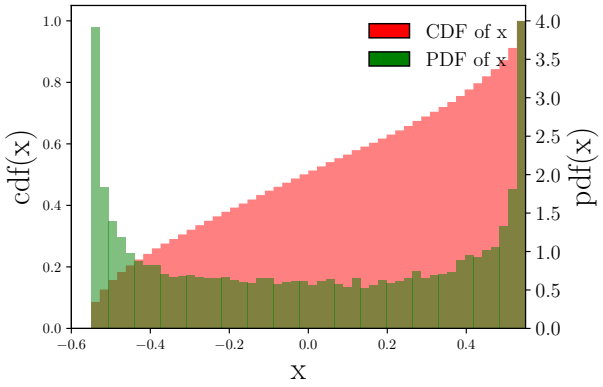

Figure 1. CDF and PDF of $\hat{\beta}_o$

- Since invertibility of block diagonal submatrices in $\sum_{t=1}^T \tilde{X}_t\tilde{X}_t'$ does not imply the invertibility of the entire matrix we also need to show that the cross terms $||X_t^e(X_t^{ms})'||_2, ||X_t^e(X_t^s)'||_2$ are sufficiently small relative to the appropriate diagonal blocks.
- Along the way we also obtain deterministic lower and upper bounds for the sample covariance matrix following which the steps for bounding the error are similar to Theorem 1.

The details are in appendix as Section 14. □

5 Inconsistency of OLS

We will now show that when a matrix is irregular, then it cannot be learned despite a high signal-to-noise ratio. Consider the two cases

$$A_r = \begin{bmatrix} 1.1 & 1 \\ 0 & 1.1 \end{bmatrix}, A_o = \begin{bmatrix} 1.1 & 0 \\ 0 & 1.1 \end{bmatrix}$$

Here A_r is a regular matrix and A_o is not. Now we run Eq. (1) for $A = A_r, A_o$ for $T = 10^3$. Let the OLS estimate of A_r, A_o be $\hat{A}_r, \hat{A}_o$ respectively. Define

$$\beta_r = [A_r]_{1,2}, \beta_o = [A_o]_{1,2}$$
$$\hat{\beta}_r = [\hat{A}_r]_{1,2}, \hat{\beta}_o = [\hat{A}_o]_{1,2}$$

Although $\beta_r \approx \hat{\beta}_r$, $\hat{\beta}_o$ does not equal zero. Instead Fig. 1 shows that $\hat{\beta}_o$ has a non–trivial distribution which is bimodal at $\{-0.55, 0.55\}$ and as a result OLS is inconsistent for A_o. This happens because the sample covariance matrix for A_o is singular despite the fact that $\Gamma_T(A_o) = (1.1)^T I$, *i.e.*, a high signal to noise ratio. In general, the relation between OLS identification of A and its controllability Gramian, $\Gamma_T(A)$, is tenuous for unstable systems unlike what is suggested in (Simchowitz et al., 2018). To see this singularity observe

that

$$X_{t+1} = A_o \begin{bmatrix} X_t^{(1)} \\ X_t^{(2)} \end{bmatrix} + \begin{bmatrix} \eta_{t+1}^{(1)} \\ \eta_{t+1}^{(2)} \end{bmatrix}$$

$$Y_T = \begin{bmatrix} \sum_{t=1}^T (X_t^{(1)})^2 & \sum_{t=1}^T (X_t^{(1)})(X_t^{(2)}) \\ \sum_{t=1}^T (X_t^{(1)})(X_t^{(2)}) & \sum_{t=1}^T (X_t^{(2)})^2 \end{bmatrix}$$

where $X_t^{(1)}, X_t^{(2)}$ are independent of each other. Define $a = 1.1$.

Proposition 5.1. *Let $\{\eta_t\}_{t=1}^T$ be i.i.d standard Gaussian then whenever $T^2 \leq a^T$, we have that*

$$||\hat{A}_o - A_o|| = \gamma_T$$

where γ_T is a random variable that admits a continuous pdf and does not decay to zero as $T \to \infty$. Further, the sample covariance matrix has the following singular values

$$\sigma_1(\sum_{t=1}^T X_t X_t^\top) = \Theta(a^{2T}), \sigma_2(\sum_{t=1}^T X_t X_t^\top) = O(\sqrt{T} a^T)$$

The proof is given in Section 20 and Proposition 20.1. Proposition 5.1 suggests that the consistency of OLS estimate depends directly on the condition number of the sample covariance matrix. In fact, OLS is inconsistent when condition number grows exponentially fast in T (as in the case of A_o). The proof requires a careful expansion of the (appropriately scaled) sample covariance matrix inverse using Woodbury's identity. Since the sample covariance matrix is highly ill–conditioned, it magnifies the noise-covariate cross terms so that the identification error no longer decays as time increases. Although for stable and marginally stable A this invertibility can be characterized $\sigma_{\min}(\Gamma_T(A))$ such an intuition does not extend to explosive systems. This is because the behavior of Y_T is dominated by "past" η_ts such as η_1, η_2 much more than the η_{T-1}, η_T etc. When A is explosive, all singular values of $||A^T||$ grow exponentially fast. Since $X_T = A^{T-1}\eta_1 + A^{T-2}\eta_2 + \ldots + A\eta_{T-1} + \eta_T$ the behavior of X_T is dominated by $A^{T-1}\eta_1$. This causes a very strong dependence between X_T and X_{T+1} and some structural constraints (such as regularity) are necessary for OLS identification.

6 Discussion

In this work we provided finite time guarantees for OLS identification for LTI systems. We show that whenever A is regular, with an otherwise arbitrary distribution of eigenvalues, OLS can be used for identification. More specifically we give sharpest possible rates when A belongs to one of $\{\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2\}$. When the assumption of regularity is violated, we show that OLS is statistically inconsistent. This suggests that statistical consistency relies on the conditioning of the sample covariance matrix and *not* so much on the signal-to-noise ratio for explosive matrices. Despite substantial differences between the distributional properties of the covariates we find that time taken to reach a given error threshold scales the same (up to some constant that depends only on A) across all regimes in terms of the probability of error. To see this, observe that Theorem 1 gives us with probability at least $1 - \delta$

$$\begin{aligned} A \in \mathcal{S}_0 &\implies ||A - \hat{A}|| \leq \sqrt{\frac{C_0(d) \log \frac{1}{\delta}}{T}} \\ A \in \mathcal{S}_1 &\implies ||A - \hat{A}|| \leq \frac{C_1(d)}{T} \log\left(\frac{T}{\delta}\right) \\ A \in \mathcal{S}_2 &\implies ||A - \hat{A}|| \leq \frac{C_2(d)\sigma_{\max}(A^{-T})}{\delta} \end{aligned} \tag{23}$$

The lower bounds for $A \in \mathcal{S}_0$ and $A \in \mathcal{S}_1$ are given in (Simchowitz et al., 2018) Appendix B, F.1 which are

$$\begin{aligned} A \in \mathcal{S}_0 &\implies ||A - \hat{A}|| \geq \sqrt{\frac{B_0(d) \log \frac{1}{\delta}}{T}} \\ A \in \mathcal{S}_1 &\implies ||A - \hat{A}|| \geq \frac{B_1(d)}{T} \log\left(\frac{1}{\delta}\right) \end{aligned} \tag{24}$$

with probability at least δ. For $A \in \mathcal{S}_2$ we provide a tighter lower bound in Proposition 4.1, *i.e.*, with probability at least δ

$$A \in \mathcal{S}_2 \implies ||A - \hat{A}|| \geq \frac{B_2(d)\sigma_{\max}(A^{-T})}{-\delta \log \delta} \tag{25}$$

Now fix an error threshold ϵ, from Eq. (23) we get with probability $\geq 1 - \delta$

$$\begin{aligned} A \in \mathcal{S}_0 &\implies ||A - \hat{A}|| \leq \epsilon \text{ if } T \geq \frac{\log \frac{1}{\delta}}{\epsilon^2 C_0(d)} \\ A \in \mathcal{S}_1 &\implies ||A - \hat{A}|| \leq \epsilon \text{ if } T \geq \frac{\log \frac{T}{\delta}}{\epsilon C_1(d)} \\ A \in \mathcal{S}_2 &\implies ||A - \hat{A}|| \leq \epsilon \text{ if } T \geq \frac{\log \frac{1}{\delta\epsilon} + \log C_2(d)}{\log \rho_{\min}} \end{aligned}$$

From Eq. (24),(25) we also know this is tight. In summary to reach a certain error threshold, T must be at least as large as $\log \frac{1}{\delta}$ for every regime.

Another key contribution of this work is providing finite time guarantees for a general distribution of eigenvalues. A major hurdle towards applying Theorem 1 to the general case is the mixing between separate components (corresponding to stable, marginally stable or explosive). Despite these difficulties we provide error bounds where each component, stable, marginally stable or explosive, has (almost) the same behavior as Theorem 1. The techniques introduced here can be used to analyze extensions such as identification in the presence of a control input U_t or heavy tailed distribution of noise (See Sections 15 and 16).

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Açıkmeşe, B., Carson, J. M., and Blackmore, L. Lossless convexification of nonconvex control bound and pointing constraints of the soft landing optimal control problem. *IEEE Transactions on Control Systems Technology*, 21 (6):2104–2113, 2013.

Campi, M. C. and Weyer, E. Finite sample properties of system identification methods. *IEEE Transactions on Automatic Control*, 47(8):1329–1334, 2002.

Erxiong, J. Bounds for the smallest singular value of a jordan block with an application to eigenvalue perturbation. *Linear Algebra and its Applications*, 197-198:691 – 707, 1994. ISSN 0024-3795. doi: https://doi.org/10.1016/0024-3795(94)90510-X. URL http://www.sciencedirect.com/science/article/pii/002437959490510X.

Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. Finite time identification in unstable linear systems. *arXiv preprint arXiv:1710.01852*, 2017.

Hardt, M., Ma, T., and Recht, B. Gradient descent learns linear dynamical systems. *arXiv preprint arXiv:1609.05191*, 2016.

Ipsen, I. C. and Lee, D. J. Determinant approximations. *arXiv preprint arXiv:1105.0437*, 2011.

Kuznetsov, V. and Mohri, M. Theory and algorithms for forecasting time series. *CoRR*, abs/1803.05814, 2018. URL http://arxiv.org/abs/1803.05814.

Lai, T. and Wei, C. Asymptotic properties of general autoregressive models and strong consistency of least-squares estimates of their parameters. *Journal of multivariate analysis*, 13(1):1–23, 1983.

Liu, J. *Eigenvalue and Singular Value Inequalities of Schur Complements*, pp. 47–82. Springer US, Boston, MA, 2005.

Nielsen, B. Singular vector autoregressions with deterministic terms: Strong consistency and lag order determination. 2008.

Oymak, S. and Ozay, N. Non-asymptotic identification of lti systems from a single trajectory. *arXiv preprint arXiv:1806.05722*, 2018.

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

Phillips, P. C. and Magdalinos, T. Inconsistent var regression with common explosive roots. *Econometric Theory*, 29 (4):808–837, 2013.

Simchowitz, M., Mania, H., Tu, S., Jordan, M. I., and Recht, B. Learning without mixing: Towards a sharp analysis of linear system identification. *arXiv preprint arXiv:1802.08334*, 2018.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Vershynin, R. High-dimensional probability: An introduction with applications in data science. 47, 2018. URL https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.pdf.

Vidyasagar, M. and Karandikar, R. L. A learning theory approach to system identification and stochastic adaptive control. In *Probabilistic and randomized methods for design under uncertainty*, pp. 265–302. Springer, 2006.

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