How fast can linear dynamical systems be learned?

Tuhin Sarkar
Massachusetts Institute of Technology

Alexander Rakhlin
Massachusetts Institute of Technology

Abstract

We provide finite time error bounds for estimating general linear time-invariant systems from a single observed trajectory using the method of least squares. We derive sharp upper bounds for three separate regimes of eigenvalue distribution: stable, marginally stable, and explosive. 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. For the general case when there is an arbitrary distribution of eigenvalues and the covariate process is a combination of the three regimes, our techniques recover the exact behavior of the identification error. 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 \( \eta_t \) is the unobserved process noise. The goal is to learn \( A \) by observing only \( X_t \)'s. We will also discuss the extension of our results 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. [2]) 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. [18]). 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 [5, 15]. Specifically, [5] 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 [5] 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 [5] 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 [15], the bounds obtained in [5] 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 [15]. 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 [15] argue that a larger signal-to-noise ratio, measured by \( \lambda_{\min}(\sum_{t=0}^{T-1} A^t A^t') \), 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 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 \( \{A^{T-1}, A^{T-2}, \ldots\} \). 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 [9]. Some of the earlier work on finite time identification in systems theory include [3 17]. 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 [6] under the assumption that system is stable. Provable guarantees for system identification in general linear systems was also studied in [12]. However, the analysis there requires that $|\|X_{t}\|| < 1$. Generalization bounds for time series forecasting of non–stationary and non–mixing processes have been developed in [8].

2 Contributions

In this paper we offer a new statistical analysis of the ordinary least squares estimator of the dynamics $X_{t+1} = AX_t + \eta_t$ 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 $\sum_{t=1}^{T} X_tX_t'$ and a certain self normalized martingale difference term. Although such a coupled analysis is similar in flavor to [15], 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 show that the error decays as $T^{-1/2}$. This follows directly from a simple observation about the recurrence relation that governs $X_tX_t'$. For the case when all eigenvalues equal one the error bounds can be sharpened to $T^{-1}$. This is an improvement over previous results where either mixing time arguments prevented such an analysis or suboptimal bounds, $T^{-1/2}$, were known.

- 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 [5]9. By leveraging subgaussian tail inequalities we sharpen previous error bounds by removing polynomial factors. We also show that this analysis is indeed tight by providing a matching lower bound.

- We show that it is possible to improve results in [5] for heavy–tailed distributions by removing polynomial factors. Our analysis introduces only logarithmic factors in this case.

- For a given error probability, $\delta$, error threshold, $\epsilon$, we observe that, whenever possible, $T$ varies most favorably with $\epsilon$ when all eigenvalues are outside the unit circle and least favorably when all eigenvalues are inside the unit circle. However, surprisingly, the dependence of $T$ on $\delta$ is the same whether eigenvalues are inside, on or outside the unit circle.

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

$$X_{t+1} = AX_t + \eta_t.$$  

Here $\eta_t$ 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 $M$, $||M||_{op} = ||M||_2$.

**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

$$\mathbb{E}(X,x)^2 = ||x||_2^2$$

**Assumption 1.** $\{\eta_t\}_{t=1}^{\infty}$ are i.i.d isotropic subgaussian and coordinates of $\eta_t$ 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.

Define the data matrix $X$ and the noise matrix $E$ as

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

where the superscript $a'$ denotes the transpose. Then $X, E$ 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

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

(2)

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

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

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

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

- The perfectly stable matrix, $S_0$
  \[ \rho_t(A) \leq 1 - \frac{C}{T} \]
  for $1 \leq i \leq d$.

- The marginally stable matrix, $S_1$
  \[ 1 - \frac{C}{T} < \rho_t(A) \leq 1 + \frac{C}{T} \]
  for $1 \leq i \leq d$.

- The regular and explosive matrix, $S_2$
  \[ \rho_t > 1 + \frac{C}{T} \]
  for $1 \leq i \leq d$.

We will occasionally replace $X_t$ (or $X(t)$) with the lower-case 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

\[ \Gamma_t(A) = \sum_{k=0}^{t} A^k A^T \]  \hfill (3)

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

\[ J_d(\lambda) = \begin{bmatrix} \lambda & 1 & 0 & \ldots & 0 \\ 0 & \lambda & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & \lambda & 1 \\ 0 & 0 & \ldots & 0 & \lambda \end{bmatrix}_{d \times d} \]  \hfill (4)

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

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

We now state two auxiliary Propositions which will be used in the main proof.

**Proposition 3.1.** Let $1 - C/T \leq \rho_t(A) \leq 1 + C/T$. Then there exists $\alpha(d)$ depending only on $d$ such that for every $8d \leq t \leq T$ we have $\sigma_{\min}(\Gamma_t(A)) \geq t \alpha(d)$.

We prove this in the appendix as Proposition 7.5.

Critical to obtaining refined error rates, will be a result from the theory of self–normalized martingales. We let $F_t = \sigma(\eta_t, \eta_{t+1}, \ldots, \eta_T, X_1, \ldots, X_t)$ denote the filtration generated by the noise and covariate process.

**Proposition 3.2.** Let $V$ be a deterministic matrix with $V \succeq \rho$. For any $0 < \delta < 1$ and $\{\eta_t, X_t\}_{t=1}^{T}$ defined as before, we have with probability $1 - \delta$

\[ \|\bar{Y}_{T-1}^{-1/2} \sum_{i=0}^{T-1} X_i \eta_{t+1} \|_2 \leq R \sqrt{8d \log \left( \frac{5 \det(\bar{Y}_{T-1})^{1/2d} \det(V)^{-1/2d}}{\delta^{1/d}} \right)} \]  \hfill (5)

where $\bar{Y}_T^{-1} = (Y_T + V)^{-1}$ and $R^2$ is the variance proxy of $\eta_t$.

We prove this in the appendix as Proposition 8.2. It rests on Theorem 1 in [1] which is itself an application of the pseudo-maximization technique in [13] (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

\[ S_d(a) = \{ v | \min_{1 \leq i \leq d} | v_i | \geq a \} \]
How fast can linear dynamical systems be learned?

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

$$
\phi_{\min}(A) = \inf_{v \in S_d(1)} \sigma_{\min}\left( \sum_{i=1}^{T} A^{i+1} v P^{i-1} v^T A^{-1} \right) \tag{6}
$$

$$
\phi_{\max}(A) = \sup_{\|v\|_2 = 1} \sigma_{\max}\left( \sum_{i=1}^{T} A^{i+1} v P^{i-1} v^T A^{-1} \right) \tag{7}
$$

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

$\psi(A)$ is defined in Proposition 3.3 and is needed for error bounds for explosive matrices.

**Proposition 3.3** (Proposition 2 in [5]). Let $P^{-1} A P = A^{-T} \sum_{i=1}^{d} A^i - \eta_i$.

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

where $P = [P_1, P_2, \ldots, P_d]$. If $\rho_{\min}(A) > 1$, then

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

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

We summarize some notation in Table II for convenience in representing our results.

### 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 $A \in S_0 \cup S_1$

$$
\| A - \hat{A} \|_2 \leq \sqrt{C \frac{T}{T^*}} \gamma_s \left( A, \frac{\delta}{4} \right)
$$

whenever

$$
T \geq \max \left( T_\eta \left( \delta \frac{1}{4} \right), T_s \left( \delta \frac{1}{4} \right) \right)
$$

- For $A \in S_1$

$$
\| A - \hat{A} \|_2 \leq \frac{C \sigma_{\max}(A^{-1})}{\sqrt{T \sigma_{\min}(\frac{1}{\eta_{\min}}(A))}} \gamma_{ms} \left( A, \frac{\delta}{2} \right)^2
$$

whenever

$$
T \geq \max \left( 2 T_\eta \left( \delta \frac{3}{3T} \right), 2 T_s \left( \delta \frac{3}{3T} \right), T_{ms} \left( \delta \frac{1}{2} \right) \right)
$$

Since $\sigma_{\min}(\frac{1}{\eta_{\min}}(A)) \geq C(d)\frac{T}{\log T}$, we have that

$$
\| A - \hat{A} \|_2 \leq \sqrt{\frac{\log T \gamma_{\max}(A, \frac{\delta}{2})}{C(d) T}}
$$

- For $A \in S_2$

$$
\| A - \hat{A} \|_2 \leq C \sigma_{\max}(A^{-T}) \gamma_{s} \left( A, \frac{\delta}{5} \right)
$$

whenever

$$
T \in T_u \left( \delta \frac{1}{5} \right)
$$

Since $\sigma_{\max}(A^{-T}) \leq C(d) \rho_{\min}(A)^{-T}$, for $A \in S_2$, the identification error decays exponentially with $T$.

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

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

**Proof.** We start by writing an upper bound

$$
\| A - \hat{A} \|_{op} \leq \| Y_T^+ S_T \|_{op}
$$

$$
\leq \| (Y_T^+)^{1/2} \|_{op} \| (Y_T^+)^{1/2} S_T \|_{op} \tag{8}
$$

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 3.2. Here we sketch the basis of our approach. First, we find determinisitic $V_{up}, V_{dn}, T_0$ such that

$$
E_0 = \{ 0 \prec V_{dn} \preceq Y_T \preceq V_{up}, T \geq T_0 \} \tag{9}
$$

$$
P(E_0) \geq 1 - \delta \tag{10}
$$

The next step is to bound the self–normalized term. Under $E_0$, 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 $E_1$ in the following way

$$
E_1 = \left\{ \| S_T \| (Y_T + V_{dn})^{-1} \leq \sqrt{8 d \log \frac{5 \det(Y_T V_{dn}^{-1} + I)^1/2}{\delta^{1/d}}} \right\}
$$
It follows from Proposition 3.2 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

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

(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

$$
||A - \hat{A}||_p \leq \frac{1}{\sigma_{\min}(V_{dn})} \sqrt{16d \log \left( \frac{\det(V_{up} V_{dn}^{-1} + I)^{1/2d}}{\delta^1/d} \right)} 
$$

(12)

where $\alpha_T$ goes to zero with $T$ and $\beta_T$ 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 $\delta$. We now sketch the approach for finding these for each regime.

### $Y_T$ behavior when $A \in S_0 \cup S_1$

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

$$
Y_T = x_0 x_0' + AY_{T-1}A' + \sum_{t=0}^{T-1} (Ax_t x_{t+1}' + \eta_{t+1} x_t' A') + \sum_{t=1}^{T} \eta_t \eta_t'
$$

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

(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 $T I_{d\times d}$ with high probability. Using Proposition 3.2 once again, we will show that with high probability

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

where $\epsilon \leq 1/2$ whenever $\rho_1(A) \leq 1 + C/T$ and $T \geq T_0$ for some $T_0$ depending only on $A$. As a result with high probability we have

$$
Y_T \succeq (1 - \epsilon) AY_{T-1}A' + (1 - \epsilon) \sum_{t=1}^{T} \eta_t \eta_t'.
$$

(14)

The details of this proof are provided in appendix as Section 9. When $1 - C/T \leq \rho_1(A) \leq 1 + C/T$ we note that the bound in Eq. (14) is not tight. The key to sharpening the lower bound is the following observation: for $T > \max \left( 2T \eta \left( \frac{\delta}{577} \right), 2T \delta \left( \frac{\delta}{577} \right), T \max \left( \frac{\delta}{7} \right) \right)$ we can ensure with high probability

$$
\sum_{\tau=1}^{t} \eta_t \eta_t' = tI
$$

(15)

simultaneously for all $t \geq T/2$. Then we will show that $\epsilon = \beta_0(\delta)$ in Table 1. The sharpening of $\epsilon$ 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

$$
Y_T \succeq (1 - \epsilon)^{1/\beta_0(\delta)} AY_{T-2}A' + \frac{T}{2} (1 - \epsilon)^{4} A' A' + \sum_{t=1}^{T} (1 - \epsilon)^{4} A' A' + \frac{T}{2} (1 - \epsilon)^{4} A' A' + \frac{T}{2} \sum_{t=1}^{T} (1 - \epsilon)^{4} A' A'
$$

(16)

and then Eq. (16) can be reduced to

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

We show that

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

and by Proposition 7.5 $Y_T \succeq C(d) T^2$ for some function $C(\cdot)$ that depends only on $d$. The details of the proof are provided in appendix as Section 10.

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

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

Then we can use Hanson–Wright inequality or Markov inequality to get an upper bound as shown in appendix as Proposition 8.4.

### $Y_T$ behavior when $A \in S_2$

The concentration arguments used to show the convergence for stable systems do not work for unstable systems. As
We show that work by transforming \( X_T = \sum_{\tau=1}^T A^{t-\tau} \eta_t \) and, consequently, \( X_T \) depends strongly on \( X_1, X_2, \ldots \). Due to this dependence we are unable to use typical techniques where \( X_t \) are divided into roughly independent blocks of covariates, to obtain concentration results. Motivated by \([9]\), we instead work by transforming \( x_t \) as
\[
    z_t = A^{-t} x_t = x_0 + \sum_{\tau=1}^t A^{-\tau} \eta_\tau. \tag{17}
\]

The steps of the proof proceed as follows. Define
\[
    U_T = A^{-T} \sum_{t=1}^T x_t' A^{-T'} = A^{-T} Y_T A^{-T'} = \sum_{t=1}^T A^{-T+t} z_t' A^{-T+t},
\]
\[
    F_T = \sum_{t=0}^{T-1} A^{-t} z_t' A^{-t+t}. \tag{18}
\]
We show that
\[
    ||F_T - U_T||_\infty \leq \epsilon. \]
Here \( \epsilon \) 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\([11]\).

The analysis presented here is sharper than \([5]\) 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 \([9, 5]\) consider \( ||U_T - F_\infty|| \) instead and as a result they require a martingale concentration argument to show the existence of \( F_\infty \).

**Remark 2.** It is claimed in \([5]\) that techniques involving inequalities for subgaussian distributions cannot be used for the class of sub-Weibull distributions they consider. However, by bounding the noise process, as even \([5]\) does, we can convert the heavy tailed process into a zero mean independent subgaussian one. In such a case our techniques can still be applied, and they incur only an extra logarithmic factor. This is discussed in detail in appendix as Section\([14]\).

For the general case we use a well known fact for matrices, namely, that there exists a similarity transform \( \tilde{P} \) such that
\[
    A = \tilde{P}^{-1} \begin{bmatrix} A_x & 0 & 0 \\ 0 & A_{ms} & 0 \\ 0 & 0 & A_s \end{bmatrix} \tilde{P} \tag{19}
\]
Here \( A_x \in S_0, A_{ms} \in S_1, A_s \in S_2 \). 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\([13]\).

**Theorem 2.** For any regular matrix \( A \) we have with probability at least \( 1 - \delta \),
- For \( A \in S_1 \cup S_2 \)
  \[
  ||A - \hat{A}||_2 \leq \frac{\text{poly}(\log T, \log \frac{1}{\delta})}{T}
  \]
  whenever \( T \geq \text{poly}\left( \log \frac{1}{\delta} \right) \)
- For \( A \in S_0 \cup S_1 \cup S_2 \)
  \[
  ||A - \hat{A}||_2 \leq \frac{\text{poly}(\log T, \log \frac{1}{\delta})}{\sqrt{T}}
  \]
  whenever \( T \geq \text{poly}\left( \log \frac{1}{\delta} \right) \).

Here \( \text{poly}(\cdot) \) is a polynomial function.

**Proof.** Define the partition of \( A \) as Eq.\([19]\). Since
\[
    X_t = \sum_{\tau=1}^t A^{t-1} \eta_{t-\tau+1}
\]
\[
    \tilde{X}_t = \tilde{P}^{-1} X_t = \sum_{\tau=1}^t \tilde{A}^{t-1} \tilde{P}^{-1} \eta_{t-\tau+1}
\]
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 \( \tilde{X}_t, \tilde{\eta}_t \)
\[
    \tilde{X}_t = \begin{bmatrix} X_t^e \\ X_t^{ms} \end{bmatrix}, \tilde{\eta}_t = \begin{bmatrix} \eta_t^e \\ \eta_t^{ms} \end{bmatrix}
\]
Then we have
\[
    \sum_{t=1}^T \tilde{X}_t \tilde{X}_t' = \sum_{t=1}^T \begin{bmatrix} X_t^e (X_t^e)' \\ X_t^{ms} (X_t^{ms})' \end{bmatrix} \begin{bmatrix} X_t^e (X_t^e)' \\ X_t^{ms} (X_t^{ms})' \end{bmatrix}'
\]
\[
    = \begin{bmatrix} X_t^e (X_t^e)' \\ X_t^{ms} (X_t^{ms})' \end{bmatrix} \begin{bmatrix} X_t^e (X_t^e)' \\ X_t^{ms} (X_t^{ms})' \end{bmatrix}'
\]
\[
    \tag{22}
\]
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., $\{\tilde{\eta}_t^n, \tilde{\eta}_t^n, \eta_t^n\}$ are not independent. A major implication is that it is no longer obvious that the cross terms between different submatrices, such as $\sum_{t=1}^T X_t^r (X_t^{ms})'$, go to zero. Our proof will have four 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 $\eta_t$ are also independent for all $t$. The only change will be the appearance of additional $\sigma^2_t(P)$ subgaussian parameter (See Corollary 8.1).
- 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 are sufficiently small relative to the appropriate diagonal blocks.
- 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. Specifically, since $X_{mss}$ contains only stable and marginally stable components, it falls under $A \in S_0 \cup S_1$. 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, $\sum_{t=1}^T X_t^r (X_t^r)'$ is also invertible as it corresponds to $A \in S_2$.
- Finally we will show that the cross terms $||X_t^r (X_t^s)'||, ||X_t^r (X_t^{ms})'||$ are sufficiently small and as a result $\sum_{t=1}^T \tilde{X}_t \tilde{X}_t'$ becomes invertible. Since $X_t^s, X_t^s, X_t^{ms}$ are no longer independent blocks, the arguments require more work than before. Along the way we also obtain deterministic upper and lower 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 13.

Remark 3. Here we sketch how to extend our results to the general case when we also have a control input, i.e.,

$$X_{t+1} = AX_t + BU_t + \eta_{t+1}$$

(23)

Here $A, B$ are unknown but we can choose $U_t$. Pick independent vectors $\{U_t, \sim \mathcal{N}(0, I)\}_{t=1}^T$. We can represent this as a variant of Eq. (1) as follows

$$\begin{bmatrix} X_{t+1} \\ U_{t+1} \\ \tilde{X}_{t+1} \end{bmatrix} = \begin{bmatrix} A & B & 0 \\ 0 & 0 & 0 \\ \tilde{\eta}_{t+1} & \tilde{\eta}_{t+1} & \tilde{\eta}_{t+1} \end{bmatrix} \begin{bmatrix} X_t \\ U_t \\ \eta_{t+1} \end{bmatrix}$$







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

Since

$$\det \begin{bmatrix} A - \lambda I & B \\ 0 & -\lambda I \end{bmatrix} = 0$$

holds when $\lambda$ equals an eigenvalue of $A$ or $0$. The eigenvalues of $\hat{A}$ are the same as $A$ with some additional eigenvalues that are zero. Now we can simply use Theorem 2.

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

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

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

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 $\Gamma_r(A_o) = (1.1)^T I$, i.e., a high signal to noise ratio. In general, the relation between OLS identification of $A$ and its controllability Gramian, $\Gamma_r(A)$, is tenuous for unstable systems unlike what is suggested in [15]. To see this singularity observe that

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

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

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

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

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

$$A_o = \begin{bmatrix} 1.1 & 0 \\ 0 & 1.1 \end{bmatrix}$$
where $X_t^{(1)}, X_t^{(2)}$ are independent of each other. Define $a = 1.1$.

**Proposition 5.1.** We have with probability at least $1 - \delta$ that

$$||a^{-2T}Y_T - E|| \leq \left( 8T^2 \frac{a^{-T}}{1-a^{-2}} + \frac{16T a^{-2T-2}}{\delta(1-a^{-2})} \right)$$

whenever

$$T \geq \frac{2}{c} \left( \log \frac{8}{\delta} + \log 2 + 2 \log 5 \right)$$

where $E$ is singular with probability 1 and of the form

$$E = \begin{bmatrix} X^2 & XY \\ XY & Y^2 \end{bmatrix}$$

such that $X, Y$ are i.i.d. zero mean Gaussian random variables with $1/2 \leq E[X^2] \leq \frac{1}{1-a^2}$.

**Proof.** The proof technique is similar to the one for explosive systems in Theorem 1 Section 1. Define $Z_t^{(i)} = a^{-t}X_t^{(i)}$. Since $X_t^{(i)} = \sum_{\tau=1}^t a^{-\tau} \eta_t^{(i)}$ and it follows from the discussion in Theorem 1 for explosive systems that $U_T = \sum_{t=1}^T A^{-T} X_t^{(i)} (X_t^{(i)})' A^{-T} = \sum_{t=1}^T A^{-t+1} Z_t^{(i)} Z_t^{(i)} = F_T$ with high probability. Then observe that

$$(Z_T^{(i)})^2 \sum_{i=1}^T a^{-2i+2} = \frac{1-a^{-2T}}{1-a^{-2}} (Z_T^{(i)})^2$$

and we have, from Eq. (81) in appendix, with probability at least $1 - 2\delta$ that

$$\left| a^{-2T} \sum_{i=1}^T (X_t^{(i)})^2 - \frac{1-a^{-2T}}{1-a^{-2}} (Z_T^{(i)})^2 \right|$$

$$\leq \left( 4T^2 \frac{a^{-T}}{1-a^{-2}} + T a^{-2T-2} \frac{1}{\delta(1-a^{-2})} \right)$$

whenever

$$T \geq \frac{2}{c} \left( \log \frac{1}{\delta} + \log 2 + 2 \log 5 \right)$$

for some universal constant $c$. In a similar fashion(See Section 16 in appendix) as above we can show that with probability at least $1 - 3\delta$

$$\left| a^{-2T} \sum_{i=1}^T (X_t^{(1)}) (X_t^{(2)}) - \sum_{i=1}^T a^{-2i+2} Z_T^{(1)} Z_T^{(2)} \right|$$

$$= \left| a^{-2T} \sum_{i=1}^T (X_t^{(1)}) (X_t^{(2)}) - \frac{1-a^{-2T}}{1-a^{-2}} Z_T^{(1)} Z_T^{(2)} \right|$$

$$\leq 4T \frac{a^{-T}}{1-a^{-2}} \log \left( \frac{2T}{\delta} \right)$$

whenever

$$T \geq \frac{2}{c} \left( \log \frac{1}{\delta} + \log 2 + 2 \log 5 \right)$$

Define

$$F_T = \frac{1-a^{-2T}}{1-a^{-2}} \left[ (Z_T^{(1)})^2 Z_T^{(2)} \right]$$

Then it is clear from Eq. (23), (25) that

$$||a^{-2T}Y_T - F_T|| \leq \left( 8T^2 \frac{a^{-T}}{1-a^{-2}} + \frac{2T a^{-2T-2}}{\delta(1-a^{-2})} \right)$$

Observe that $\det(F_T) = 0$ and $F_T$ is singular. Further since

$$\mathbb{E}[(Z_t^{(i)})^2] = a^{-2T} \mathbb{E}[(X_t^{(i)})^2] \leq \frac{1}{1-a^2}$$

but $a^{-2T} \mathbb{E}[(X_t^{(i)})^2] \geq 1/2$. This shows that $a^{-2T}Y_T$ approaches a non-zero singular matrix at an exponentially fast rate.

**Proposition 5.1** suggests that the consistency of OLS estimate depends directly on the invertibility of the sample covariance matrix. This follows from the following observation

$$\hat{A}_o = \left( \sum_{t=0}^T X_{t+1} X_t^T \right) \left( \sum_{t=0}^T X_t X_t^T \right)^+$$

$$= \left( \sum_{t=0}^T a^{-2T} X_{t+1} X_t^T \right) \left( \sum_{t=0}^T a^{-2T} X_t X_t^T \right)^+$$

Now as we have shown $\sum_{t=0}^T a^{-2T} X_t X_t^T$ is a rank 1 matrix with high probability and so is $\hat{A}_o$. However, $A_o = aI$ and hence $||A_o - \hat{A}_o|| \geq a = 1.1$. 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” $\eta_t$s such as $\eta_t, \eta_{t+1}$ much more than the $\eta_{t-1}, \eta_T$ 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.

In Table 1 the error in identification for explosive matrices depends on $\delta$ as $1/2$ unlike stable and marginally stable matrices where the dependence is $\log \frac{1}{\delta}$. Typical minimax analyses are unable to capture this relation between error and $\delta$. Here we show that such a dependence is unavoidable:
Proposition 5.2. Let $A = a > 1$ be a $1-D$ matrix and $A = \hat{a}$ be its OLS estimate. Then whenever $T \in T_n(\delta)$ as in Table [7] with probability at least $\delta$ we have

$$|a - \hat{a}| \geq \frac{-Ca^{-T}}{\delta \ln(\delta)}(a^2 - 1)$$

where $C$ is a universal constant.

A proof of this result can be found in appendix as Section 15.

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 $\{S_0, S_1, S_2\}$. In contrast to past work we show that finite time error is at most $C(d)/T$ whenever $A \in S_1$. This observation becomes important when we provide finite time guarantees for the case when an arbitrary mixture of eigenvalues may be present. 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 $1 - \delta$

$$A \in S_0 \implies ||A - \hat{A}|| \leq \sqrt{\frac{C_0(d) \log \frac{1}{\delta}}{T}}$$

$$A \in S_1 \implies ||A - \hat{A}|| \leq \sqrt{\frac{C_1(d) \log \frac{1}{\delta}}{T}}$$

$$A \in S_2 \implies ||A - \hat{A}|| \leq \frac{C_2(d) \sigma_{\max}(A^{-T})}{\delta}$$

(26)

The lower bounds for $A \in S_0$ and $A \in S_1$ are given in [15] Appendix B, F.1 which are

$$A \in S_0 \implies ||A - \hat{A}|| \geq \sqrt{\frac{B_0(d) \log \frac{1}{\delta}}{T}}$$

$$A \in S_1 \implies ||A - \hat{A}|| \geq \sqrt{\frac{B_1(d) \log \frac{1}{\delta}}{T}}$$

(27)

with probability at least $\delta$. For $A \in S_2$ we provide a tighter lower bound in Proposition 5.2 i.e., with probability at least $\delta$

$$A \in S_2 \implies ||A - \hat{A}|| \geq \frac{B_2(d) \sigma_{\max}(A^{-T})}{\delta}$$

(28)

Now fix an error threshold $\epsilon$, from Eq. (20) we get with probability $\geq 1 - \delta$

$$A \in S_0 \implies ||A - \hat{A}|| \leq \epsilon \text{ if } T \geq \frac{\log \frac{1}{\delta}}{\epsilon^2 C_0(d)}$$

$$A \in S_1 \implies ||A - \hat{A}|| \leq \epsilon \text{ if } T \geq \frac{\log \frac{1}{\delta}}{\epsilon C_1(d)}$$

$$A \in S_2 \implies ||A - \hat{A}|| \leq \epsilon \text{ if } T \geq \frac{\log \frac{1}{\epsilon \delta} + \log C_2(d)}{\log \rho_{\min}}$$

From Eq. (27), (28) 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.

References

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

[2] Açıkmese, 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 (2013), 2104–2113.

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

[4] 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 (1994), 691 – 707.

[5] Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. Finite time identification in unstable linear systems. arXiv preprint arXiv:1710.01852 (2017).

[6] Hardt, M., Ma, T., and Recht, B. Gradient descent learns linear dynamical systems. arXiv preprint arXiv:1609.05191 (2016).

[7] Iksen, I. C., and Lee, D. J. Determinant approximations. arXiv preprint arXiv:1105.0437 (2011).

[8] Kuznetsov, V., and Mohri, M. Theory and algorithms for forecasting time series. CoRR abs/1803.05814 (2018).

[9] 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 (1983), 1–23.

[10] Liu, J. Eigenvalue and Singular Value Inequalities of Schur Complements. Springer US, Boston, MA, 2005, pp. 47–82.
How fast can linear dynamical systems be learned?

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

[12] Oymak, S., and Ozyay, N. Non-asymptotic identification of lti systems from a single trajectory. *arXiv preprint arXiv:1806.05722* (2018).

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

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

[15] 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).

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

[17] 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*. Springer, 2006, pp. 265–302.

[18] Yu, B. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability* (1994), 94–116.
7 Appendix

Proposition 7.1. Let $P, V$ be a psd and pd matrix respectively and define $\bar{P} = P + V$. Let there exist some matrix $Q$ for which we have the following relation

$$||\bar{P}^{-1/2}Q|| \leq \gamma$$

For any vector $v$ such that $v'Pv = \alpha, v'Vv = \beta$ it is true that

$$||v'Q|| \leq \sqrt{\beta + \alpha \gamma}$$

Proof. Since

$$||\bar{P}^{-1/2}Q||^2 \leq \gamma^2$$

for any vector $v \in \mathbb{S}_{d-1}$ we will have

$$\frac{v'\bar{P}^{1/2}\bar{P}^{-1/2}QQ'\bar{P}^{-1/2}\bar{P}^{1/2}v}{v'Pv} \leq \gamma^2$$

and substituting $v'\bar{P}v = \alpha + \beta$ gives us

$$v'QQ'v \leq \gamma^2 v'\bar{P}v = (\alpha + \beta)\gamma^2$$

Proposition 7.2. Consider a Jordan block matrix $J_d(\lambda)$ given by $[4]$, then $J_d(\lambda)^{-k}$ is a matrix where each off–diagonal (and the diagonal) has the same entries, i.e.,

$$J_d(\lambda)^{-k} = \begin{bmatrix} a_1 & a_2 & a_3 & \ldots & a_d \\ 0 & a_1 & a_2 & \ldots & a_{d-1} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & a_1 & a_2 \\ 0 & 0 & \ldots & 0 & a_1 \end{bmatrix}_{d \times d}$$

for some $\{a_i\}_{i=1}^d$.

Proof. $J_d(\lambda) = (\lambda I + N)$ where $N$ is the matrix with all ones on the 1st (upper) off-diagonal. $N^k$ is just all ones on the $k^{th}$ (upper) off-diagonal and $N$ is a nilpotent matrix with $N^d = 0$. Then

$$(\lambda I + N)^{-1} = \left( \sum_{l=0}^{d-1} (-1)^l \lambda^{-l-1} N^l \right)$$

$$(-1)^{k-1}(k-1)! (\lambda I + N)^{-k} = \left( \sum_{l=0}^{d-1} (-1)^l \frac{dk^{l-1}}{d\lambda^{k-1}} N^l \right)$$

and the proof follows in a straightforward fashion.

Proposition 7.3. Let $A$ be a regular matrix and $A = P^{-1} \Lambda P$ be its Jordan decomposition. Then

$$\inf_{||a||_2=1} \left|\sum_{i=1}^d a_i A^{-i+1} \right|_2 > 0$$

Further $\phi_\min(A) > 0$ where $\phi_\min(\cdot)$ is defined in Definition $[5]$. 

Proof. When $A$ is regular, the geometric multiplicity of each eigenvalue is 1. This implies that $A^{-1}$ is also regular. Regularity of a matrix $A$ is equivalent to the case when minimal polynomial of $A$ equals characteristic polynomial of $A$ (See Section [17] in appendix), i.e.,

$$\inf_{||a||_2=1} \left\| \sum_{i=1}^{d} a_i A^{-i+1} \right\|_2 > 0$$

Since $A^{-j} = P^{-1} \Lambda^{-j} P$ we have

$$\inf_{||a||_2=1} \left\| \sum_{i=1}^{d} a_i P^{-1} \Lambda^{-i+1} P \right\|_2 > 0$$

$$\inf_{||a||_2=1} \left\| \sum_{i=1}^{d} a_i P^{-1} \Lambda^{-i+1} \right\|_{2\sigma_{\min}(P)} > 0$$

$$\inf_{||a||_2=1} \left\| \sum_{i=1}^{d} a_i \Lambda^{-i+1} \right\|_{2\sigma_{\min}(P)\sigma_{\min}(P^{-1})} > 0$$

$$\inf_{||a||_2=1} \left\| \sum_{i=1}^{d} a_i \Lambda^{-i+1} \right\|_2 > 0$$

Since $\Lambda$ is Jordan matrix of the Jordan decomposition, it is of the following form

$$\Lambda = \begin{bmatrix}
J_{k_1}(\lambda_1) & 0 & \cdots & 0 & 0 \\
0 & J_{k_2}(\lambda_2) & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & J_{k_l}(\lambda_l) & 0 \\
0 & 0 & \cdots & 0 & J_{k_{l+1}}(\lambda_{l+1})
\end{bmatrix} \quad (30)$$

where $J_{k_i}(\lambda_i)$ is a $k_i \times k_i$ Jordan block corresponding to eigenvalue $\lambda_i$. Then

$$\Lambda^{-k} = \begin{bmatrix}
J_{k_1}^{-k}(\lambda_1) & 0 & \cdots & 0 & 0 \\
0 & J_{k_2}^{-k}(\lambda_2) & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & J_{k_l}^{-k}(\lambda_l) & 0 \\
0 & 0 & \cdots & 0 & J_{k_{l+1}}^{-k}(\lambda_{l+1})
\end{bmatrix} \quad (31)$$

Since $\left\| \sum_{i=1}^{d} a_i \Lambda^{-i+1} \right\|_2 > 0$, without loss of generality assume that there is a non–zero element in $k_1 \times k_1$ block. This implies

$$\left\| \sum_{i=1}^{d} a_i J_{k_1}^{-i+1}(\lambda_1) \right\|_2 > 0$$

By Proposition [7,2] we know that each off–diagonal (including diagonal) of $S$ will have same element. Let $j_0 = \inf \{ j | S_{ij} \neq 0 \}$ and in column $j_0$ pick the element that is non–zero and highest row number, $i_0$. By design $S_{i_0,j_0} > 0$ and further

$$S_{k_1-\left(j_0-i_0\right),k_1} = S_{i_0,j_0}$$

because they are part of the same off–diagonal (or diagonal) of $S$. Thus the row $k_1 - (j_0 - i_0)$ has only one non–zero element because of the minimality of $j_0$.

We proved that for any $||a|| = 1$ there exists a row with only one non–zero element in the matrix $\sum_{i=1}^{d} a_i \Lambda^{-i+1}$. This implies that if $v$ is a vector with all non–zero elements, then $\left\| \sum_{i=1}^{d} a_i \Lambda^{-i+1} v \right\|_2 > 0$, i.e.,

$$\inf_{||a||_2=1} \left\| \sum_{i=1}^{d} a_i \Lambda^{-i+1} v \right\|_2 > 0$$
This implies
\[
\inf_{||a||_2=1} \left| \left| [v, \Lambda^{-1}v, \ldots, \Lambda^{-d+1}v]a \right| \right|_2 > 0 \\
\sigma_{\min}( [v, \Lambda^{-1}v, \ldots, \Lambda^{-d+1}v]) > 0
\]

By Definition 3, we have
\[
\phi_{\min}(A) > 0
\]

**Proposition 7.4** (Corollary 2.2 in [7]). For any positive definite matrix \( M \) with diagonal entries \( m_{jj} \), \( 1 \leq j \leq d \) and \( \rho \) is the spectral radius of the matrix \( C \) with elements
\[
c_{ij} = 0 \quad \text{if} \quad i = j \\
= \frac{m_{ij}}{\sqrt{m_{ii}m_{jj}}} \quad \text{if} \quad i \neq j
\]
then
\[
0 < \frac{\prod_{j=1}^{d} m_{jj} - \det(M)}{\prod_{j=1}^{d} m_{jj}} \leq 1 - e^{-\frac{\rho^2}{\max_{\lambda_{\min}}}}
\]
where \( \lambda_{\min} = \min_{1 \leq j \leq d} \lambda_j(C) \).

**Proposition 7.5.** Let \( 1 - C/T \leq \rho_t(A) \leq 1 + C/T \) and \( A \) be a \( d \times d \) matrix. Then there exists \( \alpha(d) \) depending only on \( d \) such that for every \( 8d \leq t \leq T \)
\[
\sigma_{\min}(\Gamma_t(A)) \geq t\alpha(d)
\]

**Proof.** Since \( A = P^{-1}\Lambda P \) where \( \Lambda \) is the Jordan matrix. Since \( \Lambda \) can be complex we will assume that adjoint instead of transpose. This gives
\[
\Gamma_T(A) = I + \sum_{t=1}^{T} A^t(A^t)' \\
= I + P^{-1} \sum_{t=1}^{T} \Lambda^tPP'(\Lambda^t)'P^{-1} \\
\succeq I + \sigma_{\min}(P)^2P^{-1} \sum_{t=1}^{T} \Lambda^t(\Lambda^t)'P^{-1}
\]

Then this implies that
\[
\sigma_{\min}(\Gamma_T(A)) \geq 1 + \sigma_{\min}(P)^2\sigma_{\min}(P^{-1})^2 \sum_{t=1}^{T} \Lambda^t(\Lambda^t)'
\]
\[
\geq 1 + \frac{\sigma_{\min}(P)^2}{\sigma_{\max}(P)^2} \sigma_{\min}(\sum_{t=1}^{T} \Lambda^t(\Lambda^t)')
\]

Now
\[
\sum_{t=0}^{T} \Lambda^t(\Lambda^t)' = \begin{bmatrix}
\sum_{t=0}^{T} J^t_{k_1}(\lambda_1)(J^t_{k_1}(\lambda_1))^* & 0 & \ldots & 0 \\
0 & \sum_{t=1}^{T} J^t_{k_2}(\lambda_2)(J^t_{k_2}(\lambda_2))^* & 0 & \ldots \\
\vdots & \vdots & \ddots & \ddots \\
0 & \ldots & 0 & \sum_{t=1}^{T} J^t_{k_\ell}(\lambda_\ell)(J^t_{k_\ell}(\lambda_\ell))^*
\end{bmatrix}
\]
How fast can linear dynamical systems be learned?

Since \( \Lambda \) is block diagonal we only need to worry about the least singular value corresponding to some block. Let this block be the one corresponding to \( J_k(\lambda_1) \), i.e.,

\[
\sigma_{\text{min}} \left( \sum_{t=0}^{T} \Lambda^t (\Lambda^t)^* \right) = \sigma_{\text{min}} \left( \sum_{t=0}^{T} J_k^t (\lambda_1) (J_k^t (\lambda_1))^* \right)
\]

(32)

Define \( B = \sum_{t=0}^{T} J_k^t (\lambda_1) (J_k^t (\lambda_1))^* \). Note that \( J_k(\lambda_1) = (\lambda_1 I + N) \) where \( N \) is the nilpotent matrix that is all ones on the first off–diagonal and \( N^{k+1} = 0 \). Then

\[
(\lambda_1 I + N)^t = \sum_{j=0}^{t} \binom{t}{j} \lambda_1^{t-j} N^j
\]

\[
(\lambda_1 I + N)^t ((\lambda_1 I + N)^t)^* = \left( \sum_{j=0}^{t} \binom{t}{j} \lambda_1^{t-j} N^j \right) \left( \sum_{j=0}^{t} \binom{t}{j} (\lambda_1^*)^{t-j} N^j \right)
\]

\[
= \sum_{j=0}^{t} \binom{t}{j} |\lambda_1|^{2(t-j)} N^j (N^j)^* + \sum_{j \neq k} \binom{t}{j} \binom{t}{k} \lambda_1^j \lambda_1^{k-j} N^j (N^k)^*
\]

\[
= \sum_{j=0}^{t} \binom{t}{j} |\lambda_1|^{2(t-j)} N^j (N^j)^* + \sum_{j \neq k} \binom{t}{j} \binom{t}{k} |\lambda_1|^{2k} \lambda_1^{j-k} N^j (N^k)^*
\]

Let \( \lambda_1 = re^{i\theta} \), then similar to \( \Lambda \) there is \( D = \text{Diag}(1, e^{-i\theta}, e^{-2i\theta}, \ldots, e^{-i(k-1)\theta}) \) such that \( D(\lambda_1 I + N)^t ((\lambda_1 I + N)^t)^* D^* \) is a real matrix. Observe that any term on \( (j - k) \) upper off–diagonal of \( (\lambda_1 I + N)^t ((\lambda_1 I + N)^t)^* \) is of the form \( r_0 e^{i(j-k)\theta} \). In the product \( D(\lambda_1 I + N)^t ((\lambda_1 I + N)^t)^* D^* \) any term on the \( (j - k) \) upper off diagonal term now looks like \( e^{-ij\theta + ik\theta} r_0 e^{i(j-k)\theta} = r_0 \), which is real. Then we have

\[
D(\lambda_1 I + N)^t ((\lambda_1 I + N)^t)^* D^* = \sum_{j=0}^{t} \binom{t}{j} |\lambda_1|^{2(t-j)} N^j (N^j)^* + \sum_{j \neq k} \binom{t}{j} \binom{t}{k} |\lambda_1|^{2k} \lambda_1^{j-k} N^j (N^k)^*\]

\[
= \sum_{j=0}^{t} \binom{t}{j} |\lambda_1|^{2(t-j)} N^j (N^j)^* + \sum_{j \neq k} \binom{t}{j} \binom{t}{k} |\lambda_1|^{2k} \lambda_1^{j-k} N^j (N^k)^*\]

(33)

Since \( D \) is unitary and \( D(\lambda_1 I + N)^t ((\lambda_1 I + N)^t)^* D^* = (|\lambda_1 I + N|^t ((|\lambda_1 I + N|^t)^*)^* \), we can simply work with the case when \( \lambda_1 > 0 \) and real, as the singular values remain invariant under unitary transformations. Now we show the growth of \( ij^{th} \) term of the product \( D(\lambda_1 I + N)^t ((\lambda_1 I + N)^t)^* D^* \), Define \( B = \sum_{t=1}^{T} (|\lambda_1 I + N|^t ((|\lambda_1 I + N|^t)^*) \)

\[
B_{il} = \sum_{t=1}^{T} \frac{(|\lambda_1 I + N|^t ((|\lambda_1 I + N|^t)^*) |_{il}}{T}
\]

\[
= \sum_{t=1}^{T} \sum_{j=0}^{k_{i-l}} \binom{t}{j} |\lambda_1|^{2(t-j)}
\]

(34)
Since $1 - C/T \leq |\lambda_1| \leq 1 + C/T$, then for every $t \leq T$ we have

$$e^{-C} \leq |\lambda_1|^t \leq e^C \quad \text{for every } t \leq T$$

Then

$$B_{tl} = \sum_{t=1}^{T} \sum_{j=0}^{k_1 - l} \binom{t}{j}^2 (|\lambda_1|^{2(t-j)})$$

$$\geq e^{-2C} \sum_{t=1}^{T} \sum_{j=0}^{k_1 - l} \binom{t}{j}^2$$

$$\geq e^{-2C} \sum_{t=T/2}^{T} \sum_{j=0}^{k_1 - l} \binom{t}{j}^2 \geq e^{-2C} \sum_{t=T/2}^{T} c_{k_1} \frac{t^{2k_1 - 2l + 2} - 1}{t^2 - 1} \geq C(k_1)T^{2k_1 - 2l + 1} \quad (36)$$

An upper bound can be achieved in an equivalent fashion.

$$B_{tl} = \sum_{t=1}^{T} \sum_{j=0}^{k_1 - l} \binom{t}{j}^2 (|\lambda_1|^{2(t-j)})$$

$$\leq e^{2C} T \sum_{j=0}^{k_1 - l} T^{2j} \leq C(k_1)T^{2k_1 - 2l + 1} \quad (37)$$

Similarly, for any $B_{k,k+l}$ we have

$$B_{k,k+l} = \sum_{t=1}^{T} \sum_{j=0}^{k_1 - k - l} \binom{t}{j} \binom{t}{j + l} (|\lambda_1|^{2j} |\lambda_1|^l)$$

$$\geq e^{-2C} \sum_{t=T/2}^{T} \sum_{j=0}^{k_1 - k - l} \binom{t}{j} \binom{t}{j + l}$$

$$\geq e^{-2C} T \sum_{j=0}^{k_1 - k - l} T^{2j} \binom{T/2}{j} \binom{T/2}{j + l} \quad (39)$$

$$\geq C(k_1)T^{2k_1 - 2k - l + 1} \quad (41)$$

and by a similar argument as before we get $B_{jk} = C(k_1)T^{2k_1 - j - k + 1}$. For brevity we use the same $C(k_1)$ to indicate different functions of $k_1$ as we are interested only in the growth with respect to $T$. To summarize

$$B_{jk} = C(k_1)T^{2k_1 - j - k + 1} \quad (42)$$

whenever $T \geq 8d$. Recall Proposition 7.4 let the $M$ there be equal to $B$ then since

$$C_{ij} = C(k_1) \frac{B_{ij}}{B_{ii}B_{jj}} = C(k_1) \frac{T^{2k_1 - j - k + 1}}{T^{4k_1 - 2j - 2k + 2}}$$

it turns out that $C_{ij}$ is independent of $T$ and consequently $\lambda_{\min}(C), \rho$ are independent of $T$ and depend only on $k_1$: the Jordan block size. Then $\prod_{j=1}^{k_1} B_{jj} \geq \det(B) \geq \prod_{j=1}^{k_1} B_{jj} e^{-Ck_1^2} = C(k_1) \prod_{j=1}^{k_1} B_{jj}$. This means that $\det(B) = C(k_1) \prod_{j=1}^{k_1} B_{jj}$ for some function $C(k_1)$ depending only on $k_1$. Further using the values for $B_{jj}$ we get

$$\det(B) = C(k_1) \prod_{j=1}^{k_1} B_{jj} = \prod_{j=1}^{k_1} C(k_1)T^{2k_1 - 2l + 1} = C(k_1)T^{k_1^2} \quad (43)$$

Next we use Schur-Horn theorem, i.e., let $\sigma_i(B)$ be the ordered singular values of $B$ where $\sigma_i(B) \geq \sigma_{i+1}(B)$. Then $\sigma_i(B)$ majorizes the diagonal of $B$, i.e., for any $k \leq k_1$

$$\sum_{i=1}^{k} \sigma_i(B) \geq \sum_{i=1}^{k} B_{ii}$$
Observe that $B_{ii} \leq B_{jj}$ when $i \leq j$. Then from Eq. 42 it implies that

$$B_{k_1k_1} = C_1(k_1)T \geq \sigma_{k_1}(B)$$

$$\sum_{j=k_1-1}^{k_1} B_{jj} = C_2(k_1)T^3 + C_1(k_1)T \geq \sigma_{k_1-1}(A) + \sigma_{k_1}(A)$$

Since $k_1 \geq 1$ it can be checked that for $T \geq T_1 = 2k_1 \sqrt{\frac{C_1(k_1)}{C_2(k_1)}}$ we have $\sigma_{k_1-1}(A) \leq (1 + (2k_1)^{-2})C_2(k_1)T^3 \leq (1 + k_1^{-1})C_2(k_1)T^3$ as for every $T \geq T_1$ we have $C_2(k_1)T^3 \geq 4k_1^2C_1(k_1)T$. Again to upper bound $\sigma_{k_2-2}(A)$ we will use a similar argument

$$\sum_{j=k_2-2}^{k_2} B_{jj} = C_3(k_1)T^5 + C_2(k_1)T^3 + C_1(k_1)T \geq \sigma_{k_1-2}(A) + \sigma_{k_1-1}(A) + \sigma_{k_1}(A)$$

and show that whenever

$$T \geq \max \left(T_1, 2k_1 \sqrt{\frac{C_2(k_1)}{C_3(k_1)}} \right)$$

we get $\sigma_{k_1-2}(A) \leq (1 + (2k_1)^{-2} + (2k_1)^{-4})C_3(k_1)T^5 \leq (1 + k_1^{-1})C_3(k_1)T^5$ because $T \geq T_1$ ensures $C_2(k_1)T^3 \geq 4k_1^2C_1(k_1)T$ and $T \geq T_2 = 2k_1 \sqrt{\frac{C_2(k_1)}{C_3(k_1)}}$ ensures $C_3(k_1)T^5 \geq 4k_1^2C_2(k_1)T^3$. The $C_i(k_1)$ are not important, the goal is to show that for a sufficiently large $T$ we have an upper bound on each singular values (roughly) corresponding to the diagonal element. Similarly we can ensure for every $i$ we have $\sigma_i(A) \leq (1 + k_i^{-1})C_{k_i-i+1}(k_1)T^{2k_i-2i+1}$, whenever

$$T > T_i = \max \left(T_{i-1}, 2k_1 \sqrt{\frac{C_i(k_1)}{C_{i+1}(k_1)}} \right)$$

Recall Eq. 43 where $\det(B) = C(k_1)T^{k_1^2}$. Assume that $\sigma_{k_1}(B) < \frac{C(k_1)T}{\epsilon \prod_{i=1}^{k_1} C_{i+1}(k_1)}$. Then whenever $T \geq \max \left(8d, \sup_{k_1} 2k_1 \sqrt{\frac{C_i(k_1)}{C_{i+1}(k_1)}} \right)$

$$\det(B) = C(k_1)T^{k_1^2}$$

$$\prod_{i=1}^{k_1} \sigma_i = C(k_1)T^{k_1^2}$$

$$\sigma_{k_1}(B)(1 + k_1^{-1})k_1^{-1}T^{k_1^2-1} \prod_{i=2}^{k_1} C_{i+1} \geq C(k_1)T^{k_1^2}$$

$$\sigma_{k_1}(B) \geq \frac{C(k_1)T}{(1 + k_1^{-1})k_1^{-1} \prod_{i=2}^{k_1} C_{i+1}} \geq \epsilon \prod_{i=1}^{k_1} C_{i+1}(k_1)$$

which is a contradiction. This means that $\sigma_{k_1}(B) \geq \frac{C(k_1)T}{\epsilon \prod_{i=1}^{k_1} C_{i+1}(k_1)}$. This implies

$$\sigma_{\min}(T_A(A)) \geq 1 + \frac{\sigma_{\min}(P)^2}{\sigma_{\max}(P)^2} C(k_1)T$$

for some function $C(k_1)$ that depends only on $k_1$. □

It is possible that $\alpha(d)$ might be exponentially small in $d$, however for many cases such as orthogonal matrices or diagonal matrices $\alpha(A) = 1$ [As shown in [15]]. We are not interested in finding the best bound $\alpha(d)$ rather show that the bound of Proposition 7.5 exists and assume that such a bound is known.
Proposition 7.6. Let $t_1/t_2 = \beta > 1$ and $A$ be a $d \times d$ matrix. Then

$$\lambda_1(\Gamma_{t_1}(A)\Gamma_{t_2}^{-1}(A)) \leq C(d, \beta)$$

where $C(d, \beta)$ is a polynomial in $\beta$ of degree at most $d^2$ whenever $t_i \geq 8d$.

Proof. Since $\lambda_1(\Gamma_{t_1}(A)\Gamma_{t_2}^{-1}(A)) \geq 0$

$$\lambda_1(\Gamma_{t_1}(A)\Gamma_{t_2}^{-1}(A)) \leq \text{tr}(\Gamma_{t_1}(A)\Gamma_{t_2}^{-1}(A))$$

$$\leq \text{tr}(\Gamma_{t_2}^{-1/2}(A)\Gamma_{t_1}(A)\Gamma_{t_2}^{-1/2}(A))$$

$$\leq d\sigma_1(\Gamma_{t_2}^{-1/2}(A)\Gamma_{t_1}(A)\Gamma_{t_2}^{-1/2}(A))$$

$$\leq d \sup_{||x|| \neq 0} \frac{x'\Gamma_{t_1}(A)x}{x'\Gamma_{t_2}(A)x}$$

Now

$$\Gamma_{t_i}(A) = P^{-1}\sum_{l=0}^{t_i} \Lambda^iPP'(\Lambda^i)^*P^{-1}$$

$$\leq \sigma_{\text{max}}(P)^2P^{-1}\sum_{l=0}^{t_i} \Lambda^i(\Lambda^i)^*P^{-1}$$

$$\Gamma_{t_i}(A) \geq \sigma_{\text{min}}(P)^2P^{-1}\sum_{l=0}^{t_i} \Lambda^i(\Lambda^i)^*P^{-1}$$

Then this implies

$$\sup_{||x|| \neq 0} \frac{x'\Gamma_{t_1}(A)x}{x'\Gamma_{t_2}(A)x} \leq \frac{\sigma_{\text{max}}(P)^2}{\sigma_{\text{min}}(P)^2} \sup_{||x|| \neq 0} \frac{x'\sum_{l=0}^{t_1} \Lambda^l(\Lambda^l)^*x}{x'\sum_{l=0}^{t_2} \Lambda^l(\Lambda^l)^*x}$$

Then from Lemma 12 in [1] we get that

$$\sup_{||x|| \neq 0} \frac{x'\sum_{l=0}^{t_1} \Lambda^l(\Lambda^l)^*x}{x'\sum_{l=0}^{t_2} \Lambda^l(\Lambda^l)^*x} \leq \frac{\det(\sum_{l=0}^{t_1} \Lambda^l(\Lambda^l)^*)}{\det(\sum_{l=0}^{t_2} \Lambda^l(\Lambda^l)^*)}$$

Then

$$\frac{\det(\sum_{l=0}^{t_2} \Lambda^l(\Lambda^l)^*)}{\det(\sum_{l=0}^{t_1} \Lambda^l(\Lambda^l)^*)} \leq \frac{\det(\prod_{i=1}^{t_2} (\sum_{l=0}^{t_2} J_{k_i}(\lambda_i)^l(J_{k_i}(\lambda_i)^l)^*)))}{\det(\prod_{i=1}^{t_1} (\sum_{l=0}^{t_1} J_{k_i}(\lambda_i)^l(J_{k_i}(\lambda_i)^l)^*)))}$$

Here $l$ are the number of Jordan blocks of $A$. Then our assertion follows from Eq. (43) which implies that the determinant of $\sum_{l=0}^{t_2} J_{k_i}(\lambda_i)^l(J_{k_i}(\lambda_i)^l)^*$ is equal to the product of the diagonal elements (times a factor that depends only on Jordan block size), i.e., $C(k_i) t_2^{k_i^2}$. As a result the ratio is given by

$$\frac{\det(\prod_{i=1}^{t_2} (\sum_{l=0}^{t_2} J_{k_i}(\lambda_i)^l(J_{k_i}(\lambda_i)^l)^*)))}{\det(\prod_{i=1}^{t_1} (\sum_{l=0}^{t_1} J_{k_i}(\lambda_i)^l(J_{k_i}(\lambda_i)^l)^*)))} = \prod_{i=1}^{l} \beta_{k_i^2}$$

whenever $t_2, t_1 \geq 8d$. Summarizing we get

$$\sup_{||x|| \neq 0} \frac{x'\Gamma_{t_1}(A)x}{x'\Gamma_{t_2}(A)x} \leq \frac{\sigma_{\text{max}}(P)^2}{\sigma_{\text{min}}(P)^2} \prod_{i=1}^{l} \beta_{k_i^2}$$

□
8 Probabilistic Inequalities

Proposition 8.1 ([16]). Let $M$ be a random matrix. Then we have for any $\epsilon < 1$ and any $w \in S^{d-1}$ that

$$
\mathbb{P}(||M|| > z) \leq (1 + 2/\epsilon)^d \mathbb{P}(||Mw|| > (1 - \epsilon)z)
$$

The proof of the Proposition can be found, for instance, in [16].

Proposition 8.1 helps us in using the tools developed in de la Pena et al. and [1] for self–normalized martingales. We will define $\tilde{S}_t = \sum_{r=0}^t X_r \tilde{\eta}_{t+1}$ where $\tilde{\eta}_t = w^T \eta_t$ is standard normal when $w$ is a unit vector. Specifically, we use Lemma 9 of [1] which we state here for convenience:

Theorem 3 (Theorem 1 in [1]). Let $\{\mathcal{F}_t\}_{t=0}^\infty$ be a filtration. Let $\{\eta_t\}_{t=1}^\infty$ be a real valued stochastic process such that $\eta_t$ is $\mathcal{F}_t$ measurable and $\eta_t$ is conditionally $R$-sub-Gaussian for some $R > 0$, i.e.,

$$
\forall \lambda \in \mathbb{R} \quad \mathbb{E}[e^{\lambda \eta_t} | \mathcal{F}_{t-1}] \leq e^{\lambda^2 R^2}
$$

Let $\{X_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$–valued stochastic process such that $X_t$ is $\mathcal{F}_t$ measurable. Assume that $V$ is a $d \times d$ positive definite matrix. For any $t \geq 0$ define

$$
\tilde{V}_t = V + \sum_{s=1}^t X_s X_s^T S_t = \sum_{s=1}^t \eta_{s+1} X_s
$$

Then for any $\delta > 0$ with probability at least $1 - \delta$ for all $t \geq 0$

$$
||S_t||_{\tilde{V}_t}^2 \leq 2R^2 \log \left( \frac{\det(\tilde{V}_t)^{1/2} \det(V)^{-1/2}}{\delta} \right)
$$

Proposition 8.2. Let $P$ have full row rank and

$$
X_{t+1} = AX_t + P\eta_{t+1}
$$

where $\{\eta_t\}_{t=1}^\infty$ is an i.i.d. subGaussian process with variance proxy = 1 and each $\eta_t$ has independent elements. For any $0 < \delta < 1$, we have with probability $1 - \delta$

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

where $\tilde{Y}_{T-1} = (\sum_{t=1}^T X_t X_t^T + V)^{-1}$ and any deterministic $V$ with $V > 0$.

Proof. Note that $P\eta_t$ is a non–trivial subGaussian if $P$ has full rank.

Define $S_t = \sum_{s=1}^t X_s \eta_{s+1} P'$. Using Proposition 8.1 and setting $\epsilon = 1/2$, we have that

$$
\mathbb{P}(|||\tilde{Y}_{T-1}^{-1/2} S_{T-1} |||^2_2 \leq y) \leq 5^d \mathbb{P}(|||\tilde{Y}_{T-1}^{-1/2} S_{T-1} w |||^2_2 \leq \frac{y^2}{4}) = \mathbb{P}(|||\tilde{Y}_{T-1}^{-1/2} S_{T-1} w |||^2_2 \leq \frac{y^2}{4})
$$

Setting $S_{T-1} w = \sum_{s=1}^{T-1} X_s \eta_{s+1} P' w$ we observe that $\eta_{s+1} P' w$ satisfies the conditions of Theorem 3 with variance proxy $\sigma_{max}(P)^2$. Then replace in Eq. 45

$$
y^2 = 8R^2 \log \left( \frac{\det(\tilde{Y}_{T-1})^{1/2} \det(V)^{-1/2}}{5 - d \delta} \right)
$$

which gives us from Theorem 3

$$
\mathbb{P}(|||\tilde{Y}_{T-1}^{-1/2} S_{T-1} |||^2_2 \leq y) \leq \delta
$$
**Theorem 4** (Hanson–Wright Inequality). Given a subGaussian vector \( X = (X_1, X_2, \ldots, X_n) \in \mathbb{R}^n \) with \( \sup_i \| X_i \|_{\psi_2} \leq K \) and \( X_i \) are independent. Then for any \( B \in \mathbb{R}^{n \times n} \) and \( t \geq 0 \)

\[
\Pr(|X'BX - \mathbb{E}[X'BX]| \leq t) 
\leq 2 \exp \left\{ -c \min \left( \frac{t}{K^2 \| B \|}, \frac{t^2}{K^4 \| B \|_{HS}^2} \right) \right\}
\]  

(46)

**Proposition 8.3** (Theorem 5.39 [16]). Let \( E \) be an \( T \times d \) matrix whose rows \( \eta_i' \) are independent sub–Gaussian isotropic random vectors with variance proxy \( 1 \) in \( \mathbb{R}^d \). Then for every \( t \geq 0 \), with probability at least \( 1 - 2e^{-ct^2} \) one has

\[
\sqrt{T} - C \sqrt{d} - t \leq \sigma_{\min}(E) \leq \sqrt{T} + C \sqrt{d} + t
\]

(47)

The implication of Proposition 8.3 is as follows: \( E'E \geq (\sqrt{T} - C \sqrt{d} - t)^2 I \) with probability at least \( 1 - 2e^{-ct^2} \). Let \( t = \sqrt{\frac{1}{c} \log \frac{2}{\delta}} \), and ensure that

\[
T \geq T_\eta(\delta) = C \left( d + \log \frac{2}{\delta} \right)
\]

for some large enough universal constant \( C \). Then for \( T > T_\eta(\delta) \) we have, with probability at least \( 1 - \delta \), that

\[
\frac{3}{4} I \leq \frac{1}{T} \sum_{t=1}^{T} \eta_i' \eta_i \leq \frac{5}{4} I
\]

(48)

Further with the same probability

\[
\frac{3\sigma_{\min}^2(P)}{4} I \leq \frac{1}{T} \sum_{t=1}^{T} \eta_i' P \eta_i' P' \leq \frac{5\sigma_{\max}^2(P)}{4} I
\]

\[
T_\eta(\delta) = C \left( d + \log \frac{2}{\delta} \right)
\]

(49)

**Corollary 8.1** (Dependent Hanson–Wright Inequality). Given independent subGaussian vectors \( X_i \in \mathbb{R}^d \) such that \( X_{ij} \) are independent and \( \sup_{ij} \| X_{ij} \|_{\psi_2} \leq K \). Let \( P \) have full row rank. Define

\[
X = \begin{bmatrix}
PX_1 \\
PX_2 \\
\vdots \\
PX_n
\end{bmatrix} \in \mathbb{R}^{dn}
\]

Then for any \( B \in \mathbb{R}^{dn \times dn} \) and \( t \geq 0 \)

\[
\Pr\left(|X'BX - \mathbb{E}[X'BX]| \leq t\right)
\leq 2 \exp \left\{ -c \min \left( \frac{t}{K^2 \sigma_1^2(P) \| B \|}, \frac{t^2}{K^4 \sigma_1^2(P) \| B \|_{HS}^2} \right) \right\}
\]

(50)

**Proof.** Define

\[
\tilde{X} = \begin{bmatrix}
X_1 \\
X_2 \\
\vdots \\
X_n
\end{bmatrix}
\]

Now \( \tilde{X} \) is such that \( \tilde{X}_i \) are independent. Observe that \( X = (I_{n \times n} \otimes P) \tilde{X} \). Then \( X'BX = \tilde{X}(I_{n \times n} \otimes P)B(I_{n \times n} \otimes P') \tilde{X} \). Since

\[
\| (I_{n \times n} \otimes P)B(I_{n \times n} \otimes P') \| \leq \sigma_1^2(P) \| B \|
\]

\[
\text{tr}( (I_{n \times n} \otimes P)B(I_{n \times n} \otimes P')(I_{n \times n} \otimes P)B(I_{n \times n} \otimes P') ) \leq \sigma_1^2(P)^2 \text{tr}( (I_{n \times n} \otimes P)B^2(I_{n \times n} \otimes P') )
\]

\[
\leq \sigma_1^2(P)^2 \text{tr}(B^2)
\]
and now we can use Hanson–Wright in Theorem 4 and get the desired bound.

Let \( X_t = \sum_{j=0}^{t-1} A^j \eta_{t-j} \).

**Proposition 8.4.** Let \( P \) have full row rank and

\[
X_{t+1} = AX_t + P \eta_{t+1}
\]

where \( \{ \eta_t \} \) is an i.i.d. process and each \( \eta_t \) has independent elements. Then with probability at least \( 1 - \delta \), we have

\[
|| \sum_{t=1}^{T} X_t X_t'||_2 \leq \sigma_1(P)^2 \text{tr}(\Gamma_{T-1}(A)) \frac{T}{\delta}
\]

\[
|| \sum_{t=1}^{T} AX_t X_t' A'||_2 \leq \sigma_1(P)^2 \text{tr}(\Gamma_T(A) - I) \frac{T}{\delta}
\]

Let \( \delta \in (0, e^{-1}) \) then with probability at least \( 1 - \delta \)

\[
|| \sum_{t=1}^{T} X_t X_t'||_2 \leq \sigma_1(P)^2 \text{tr}(\sum_{t=0}^{T-1} \Gamma_t(A)) \left( 1 + \frac{1}{c} \log \left( \frac{1}{\delta} \right) \right)
\]

for some universal constant \( c \).

**Proof.** Define \( \tilde{\eta} = \begin{bmatrix} P \eta_1 \\ P \eta_2 \\ \vdots \\ P \eta_T \end{bmatrix} \). Then \( \tilde{\eta} \) is a non–trivial subGaussian whenever \( P \) has full row rank.

As in Corollary 8.1 by defining \( \tilde{A} \) as

\[
\tilde{A} = \begin{bmatrix}
I & 0 & 0 & \ldots & 0 \\
A & I & 0 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
A^{T-4} & A^{T-5} & \ldots & A^{T-3} & I \\
A^{T-2} & A^{T-3} & \ldots & A^{T-2} & I
\end{bmatrix}
\]

observe that

\[
\tilde{A} \tilde{\eta} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_T \end{bmatrix}.
\]

Since

\[
||X_t X_t'|| = X_t' X_t,
\]

we have that

\[
|| \sum_{t=1}^{T} X_t X_t'|| \leq \sum_{t=1}^{T} X_t' X_t = \tilde{\eta} \tilde{A} \tilde{\eta} = \text{tr}(\tilde{A} \tilde{\eta} \tilde{\eta}' \tilde{A}').
\]

The assertion of proposition follows by applying Markov’s Inequality to \( \text{tr}(\tilde{A} \tilde{\eta} \tilde{\eta}' \tilde{A}'). \) For the second part observe that each block matrix of \( \tilde{A} \) is scaled by \( A \), but the proof remains the same. Then in the notation of Theorem 4 \( B = \tilde{A}' \tilde{A}, X = \tilde{\eta} \)

\[
||B||_S = \text{tr}(\tilde{A}' \tilde{A}) = \sum_{t=0}^{T-1} \text{tr}(\Gamma_t(A))
\]

\[
||B||_F^2 \leq ||B||_S ||B||_2
\]
Define $c^* = \min(c, 1)$. Set $t = \frac{||B||_F^2}{c^*||B||} \log \left( \frac{1}{\delta} \right)$ and assume $\delta \in (0, e^{-1})$ then

$$\frac{t}{c^*||B||} \leq \frac{t^2}{c^*||B||_F^2} \log \left( \frac{1}{\delta} \right)$$

we get from Theorem 4 that

$$\tilde{\eta}' \tilde{A}' \tilde{A} \tilde{\eta} \leq \text{tr} \left( \sum_{t=0}^{T-1} \Gamma_t(A) \right) + \frac{||B||_F^2}{c^*||B||_2} \log \left( \frac{1}{\delta} \right)$$

$$\leq \text{tr} \left( \sum_{t=0}^{T-1} \Gamma_t(A) \right) + \frac{||B||_F^2}{c^*} \log \left( \frac{1}{\delta} \right)$$

$$\leq \text{tr} \left( \sum_{t=0}^{T-1} \Gamma_t(A) \right) \left( 1 + \frac{1}{c^*} \log \left( \frac{1}{\delta} \right) \right)$$

with probability at least $1 - \exp \left( - \frac{c||B||_F^2}{c^*||B||_2^2} \log \frac{1}{\delta} \right)$. Since

$$\frac{c||B||_F^2}{c^*||B||_2^2} \geq 1$$

it follows that

$$\exp \left( - \frac{c||B||_F^2}{c^*||B||_2^2} \log \frac{1}{\delta} \right) \leq \delta$$

and we can conclude that with probability at least $1 - \delta$

$$\tilde{\eta}' \tilde{A}' \tilde{A} \tilde{\eta} \leq \text{tr} \left( \sum_{t=0}^{T-1} \Gamma_t(A) \right) \left( 1 + \frac{1}{c^*} \log \left( \frac{1}{\delta} \right) \right)$$

Corollary 8.2. Whenever $\delta \in (0, e^{-1})$, we have with probability at least $1 - \delta$

$$|| \sum_{t=k+1}^{T} X_t X_t' ||_2 \leq \sigma_1^2(P) \text{tr} \left( \sum_{t=k}^{T-1} \Gamma_t(A) \right) \left( 1 + \frac{1}{c} \log \left( \frac{1}{\delta} \right) \right)$$

for some universal constant $c$.

Proof. The proof follows the same steps as Proposition 8.4. Define

$$\tilde{A} = \begin{bmatrix}
I & 0 & 0 & \ldots & 0 \\
A & I & 0 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
A^{T-1} & A^{T-2} & A^{T-3} & \ldots & I
\end{bmatrix} (I_{n \times n} \otimes P')$$

Define $\tilde{A}_k$ as the matrix formed by zeroing out all the rows of $\tilde{A}$ from $k + 1$ row onwards. Then observe that

$$\sum_{t=k+1}^{T} X_t X_t' \leq \text{tr} \left( \sum_{t=k+1}^{T} X_t X_t' \right) = \text{tr} \left( \sum_{t=1}^{T} X_t X_t' - \sum_{t=1}^{k} X_t X_t' \right)$$

$$= \tilde{\eta}' (\tilde{A}' \tilde{A} - \tilde{A}'_k \tilde{A}_k) \tilde{\eta}$$
Since \( \text{tr}(\sum_{t=1}^{T} X_tX'_t - \sum_{t=1}^{k} X_tX'_t) \geq 0 \) for any \( \hat{\eta} \) it implies \( B = (\hat{A}'\hat{A} - \hat{A}'_k\hat{A}_k) \succeq 0 \).

\[
\|B\|_S = \text{tr}(\hat{A}'\hat{A}) = \sum_{t=k}^{T-1} \text{tr}(\Gamma_t(A))
\]

\[
\|B\|_F^2 \leq \|B\|_S^2
\]

Define \( c^* = \min(c,1) \). Set \( t = \frac{\|B\|_F^2}{c^*\|B\|_S}\) and assume \( \delta \in (0, e^{-1}) \) then

\[
\frac{t}{c^*\|B\|_S} \leq \frac{t^2}{c^*\|B\|_F^2}
\]

we get from Theorem that

\[
\hat{\eta}'\hat{A}'\hat{A}\hat{\eta} \leq \|B\|_S + \frac{\|B\|_F^2}{c^*\|B\|_S}\log \left(\frac{1}{\delta}\right) \leq \|B\|_S + \frac{\|B\|_S}{c^*}\log \left(\frac{1}{\delta}\right) \leq \|B\|_S \left(1 + \frac{1}{c^*}\log \left(\frac{1}{\delta}\right)\right)
\]

with probability at least \( 1 - \exp\left\{ -\frac{c\|B\|_F^2}{c^*\|B\|_S}\log \frac{1}{\delta}\right\} \). Since

\[
\frac{c\|B\|_F^2}{c^*\|B\|_S} \geq 1
\]

it follows that

\[
\exp\left\{ -\frac{c\|B\|_F^2}{c^*\|B\|_S}\log \frac{1}{\delta}\right\} \leq \delta
\]

and we can conclude that with probability at least \( 1 - \delta \)

\[
\hat{\eta}'\hat{A}'\hat{A}\hat{\eta} \leq \text{tr}(\sum_{t=k}^{T-1} \Gamma_t(A)) \left(1 + \frac{1}{c^*}\log \left(\frac{1}{\delta}\right)\right)
\]

\[\square\]

**Proposition 8.5.** Whenever the pdf of \( X, f(\cdot) \), satisfies \( \text{ess sup}_x f(x) = C_X < \infty \) we have

\[
\Pr(|X| \leq \delta) \leq 2C_X\delta
\]

**Proof.** Since the essential supremum of \( f(\cdot) \) is bounded. Then

\[
\Pr(|X| \leq \delta) = \int_{x=-\delta}^{\delta} f(x)dx \leq 2C_X\delta
\]

\[\square\]

**Proposition 8.6** (Proposition 2 in [5]). Let \( P^{-1} \Lambda P = A \) be the Jordan decomposition of \( A \) and define \( z_T = A^{-T} \sum_{i=1}^{T} A^{T-1} \eta_i \). Further assume that \( \eta_i \) is continuous, subGaussian with variance proxy = 1 then

\[
\psi(A,\delta) = \sup\left\{ y \in \mathbb{R} : \Pr\left( \min_{1 \leq i \leq \delta} |P_i z_T| < y \right) \leq \delta \right\}
\]

where \( P = [P_1, P_2, \ldots, P_d]^T \). If \( \rho_{\min}(A) > 1 \), then

\[
\psi(A,\delta) \geq \psi(A)\delta > 0
\]

where \( \psi(A) \) depend only on \( A \).
Proof. Define the event $\mathcal{E} = \{\min_{1 \leq i \leq d}|P_i' z_T| < y\}$, $\mathcal{E}_i = \{|P_i' z_T| < y\}$. Clearly $\mathcal{E} \implies \bigcup_{i=1}^{d} \mathcal{E}_i$, then

$$P(\mathcal{E}) \leq P(\bigcup_{i=1}^{d} \mathcal{E}_i) \leq \sum_{i=1}^{d} P(\mathcal{E}_i)$$

From Proposition 8.5 and Assumption 1, we have $P(\mathcal{E}_i) \leq 2C_{|P_i' z_T|} y$. Then we get

$$P(\mathcal{E}) \leq \left(2 \sum_{i=1}^{d} C_{|P_i' z_T|} y \right) \leq 2d \sup_{1 \leq i \leq d} C_{|P_i' z_T|} y$$

where $C_{|P_i' z_T|}$ is the essential supremum of the pdf of $|P_i' z_T|$. Then $\psi(A) = \frac{1}{2d \sup_{1 \leq i \leq d} C_{|P_i' z_T|}}$. \hfill \qed

9 Lower Bound for $Y_T$ when $A \in S_0 \cup S_1$

Here we will prove our results when $\rho(A) \leq 1 + C/T$. Assume for this case that $\eta_t = L\tilde{\eta}_t$ where $\{\tilde{\eta}_t\}_{t=1}^{T}$ are i.i.d and all elements of $\tilde{\eta}_t$ are independent. Further $L$ is full row rank. Define $\sigma_{\min}(LL') = R^2 > 0$. Let $\sigma_{\max}(LL') = 1$ (this does not affect our result: $R$ is just the inverse of the condition number). Define

$$P = AY_{T-1} A'$$
$$Q = \sum_{t=0}^{T-1} Ax_t \eta_t'^{t+1}$$
$$V = TI$$
$$T_\eta = C \left( \log \frac{2}{\delta} + d \log 5 \right)$$

$$\mathcal{E}_1(\delta) = \left\{ ||Q||_{(P+V)^{-1}} \leq 8 \log \left( \frac{5^d \text{det}(P + V)^{1/2} \text{det}(V)^{-1/2}}{\delta} \right) \right\}$$

$$\mathcal{E}_2(\delta) = \left\{ |\sum_{t=0}^{T-1} Ax_t x_t' A'| \leq T \text{Tr}(\Gamma_T(A) - I) \right\}$$

$$\mathcal{E}_\eta(\delta) = \{ T > T_\eta(\delta), \frac{3R^2}{4} I \leq \frac{1}{T} \sum_{t=1}^{T} \eta_t \eta_t' \leq \frac{5}{4} I \}$$

$$\mathcal{E}(\delta) = \mathcal{E}_\eta(\delta) \cap \mathcal{E}_1(\delta) \cap \mathcal{E}_2(\delta)$$

Recall that

$$Y_T \succeq AY_{T-1} A' + \sum_{t=0}^{T-1} Ax_t \eta_t'^{t+1} + \eta_{t+1} x_t' A' + \sum_{t=1}^{T} \eta_t \eta_t' \tag{51}$$

Our goal here will be to control

$$||Q||_2 \tag{52}$$

Following Proposition 3.2, Proposition 8.4, it is true that $P(\mathcal{E}_1(\delta) \cap \mathcal{E}_2(\delta)) \geq 1 - 2\delta$. We will show that

$$\mathcal{E}(\delta) = \mathcal{E}_\eta(\delta) \cap \mathcal{E}_1(\delta) \cap \mathcal{E}_2(\delta) \implies \sigma_{\min}(\hat{Y}_T) \geq 1/4$$

Under $\mathcal{E}_\eta(\delta)$, we get

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

$$Y_T \succeq AY_{T-1} A' + \sum_{t=0}^{T-1} Ax_t \eta_t'^{t+1} + \eta_{t+1} x_t' A' + \frac{3}{4} R^2 TI$$

$$U' Y_T U \succeq U' AY_{T-1} A' U + U' \sum_{t=0}^{T-1} \left( Ax_t \eta_t'^{t+1} + \eta_{t+1} x_t' A' \right) U + \frac{3}{4} TR^2 \quad \forall U \in S^{d-1} \tag{53}$$
How fast can linear dynamical systems be learned?

Intersecting Eq. (53) with $\mathcal{E}_1(\delta) \cap \mathcal{E}_2(\delta)$, we find under $\mathcal{E}(\delta)$

\[
\|Q\|_{(P+V)^{-1}}^2 \leq 8 \log \left( \frac{5^d \det(P + V)^{1/2} \det(V)^{-1/2}}{\delta} \right) \\
\leq 8 \log \left( \frac{5^d \det(\frac{T}{\delta}(\Gamma_T(A) - I) + TI)^{1/2} \det(TI)^{-1/2}}{\delta} \right) \\
\leq 8 \log \left( \frac{5^d \det(\Gamma_T(A) - I + I)^{1/2}}{\delta^d} \right)
\]

Using Proposition 7.1 and letting $\kappa^2 = U^T P U$ then

\[
\|QU\|_2 \\
\leq \sqrt{\kappa^2 + T} \left( 8 \log \left( \frac{5^d \det(\Gamma_T(A) - I + I)^{1/2}}{\delta^d} \right) \right)
\]

So Eq. (53) implies

\[
U^T Y_T U \geq \kappa^2 - \sqrt{(\kappa^2 + T)} \sqrt{16d \log (\Gamma_T - I) + 1} + 32d \log \frac{5}{\delta} + \frac{3}{4} T R^2
\]

which gives us

\[
U^T Y_T U \geq \frac{\kappa^2}{T} - \sqrt{\left( \frac{\kappa^2}{T} + 1 \right)} \left( \sqrt{16d \log (\Gamma_T - I) + 1} + 32d \log \frac{5}{\delta} + \frac{3}{4} R^2 \right)
\]

If we can ensure

\[
\frac{TR^2}{128} \geq d \log (\Gamma_T - I) + 1 + d \log \frac{5}{\delta}
\]

then $\beta \leq R^2/2$, i.e.,

\[
\sqrt{\frac{16d}{T} \log (\Gamma_T - I) + 1} + \frac{32d}{T} \log \frac{5}{\delta} \leq \frac{R^2}{2}
\]

Let $T$ be large enough that Eq. (55) is satisfied then Eq. (54) implies

\[
U^T Y_T U \geq \frac{\kappa^2}{T} - \sqrt{\left( \frac{\kappa^2}{T} + 1 \right) R^2} + \frac{3R^2}{4} \geq \frac{R^2}{4} + \frac{\kappa^2}{2T}
\]

Since $U$ is arbitrarily chosen Eq. (56) implies

\[
Y_T \succeq \frac{TR^2}{4} I
\]

with probability at least $1 - 3\delta$ whenever

\[
\rho_t(A) \leq 1 + \frac{c}{T}
\]

\[
T \geq \max \left( C \left( \log \frac{2}{\delta} + d \log 5 \right), CR^2 \left( \frac{d}{2} \log (\Gamma_T - I) + 1 \right) + d \log \frac{5}{\delta} \right)
\]

**Remark 4.** Eq. (55) is satisfied whenever $\text{tr}(\Gamma_T - I)$ grows at most polynomially in $T$. This is true whenever $\rho(A) \leq 1 + \frac{c}{T}$. 
10 Sharpened bounds when $1 - \frac{c}{T} \leq \rho_t(A) \leq 1 + \frac{c}{T}$

Here we show that the bound for $Y_t$ in Eq. (57) can be sharpened to have quadratic growth in $T$. The key idea towards sharpening will be that we want Eq. (57) satisfied for every $t \geq \frac{T}{2}$ simultaneously, i.e., we need

$$Y_t \geq \frac{tR^2}{4} I$$

simultaneously for $t \geq \frac{T}{2}$ with high probability. By similar arguments as before as long as we have

$$\rho_t(A) I \leq 1$$

$$t \geq \max \left( C \left( \log \frac{2}{\delta} + d \log 5 \right), CR^2 \left( \frac{d}{2} \log (\Gamma_t - I) + 1 + d \log \frac{5}{\delta} \right) \right)$$

we can conclude with probability at least $1 - 2\delta$ that $Y_t \geq \frac{tR^2}{4} I$. This means that with probability at least $1 - 3\delta \frac{T}{2}$ we have for $t \geq \frac{T}{2}$ simultaneously

$$Y_t \geq \frac{tR^2}{4} I$$

when Eq. (60) is satisfied for each $t$. Since the LHS of Eq. (60) is least at $t = T/2$ and RHS is greatest at $t = T$, a sufficient condition for every $t \geq \frac{T}{2}$ satisfying Eq. (60) is the following

$$T \geq \max \left( C \left( \log \frac{2}{\delta} + d \log 5 \right), C \left( \frac{d}{2} \log (\Gamma_T - I) + 1 + d \log \frac{5}{\delta} \right) \right)$$

Then by substituting $\delta \rightarrow \frac{2\delta}{3T}$ we can conclude with probability at least $1 - \delta$ that

$$Y_t \geq \frac{tR^2}{4} I$$

simultaneously for every $t \geq \frac{T}{2}$ whenever

$$T \geq \max \left( C \left( \log \frac{3T}{2\delta} + d \log 5 \right), CR^2 \left( \frac{d}{2} \log (\Gamma_T - I) + 1 + d \log \frac{15T}{2\delta} \right) \right)$$

Define $\gamma_{t-1} = \sqrt{U^T A Y_{t-1} A U}$ and Eq. (56) becomes

$$U^T Y_{t-1} U \geq \gamma_{t-1}^2 - \sqrt{\gamma_{t-1}^2 \left( \frac{2\delta}{T} + t \right)} \sqrt{16d \log (\Gamma_t - I) + 1} + 32d \log \frac{15T}{2\delta} + \frac{3}{4} t R^2$$

Under Eq. (61) is $\leq \frac{R^2}{4}$

$$\geq \gamma_{t-1}^2 - (\gamma_{t-1} + \sqrt{t}) \sqrt{16d \log (\Gamma_t - I) + 1} + 32d \log \frac{15T}{2\delta} + \frac{3}{4} t R^2$$

$$\geq \gamma_{t-1}^2 - \gamma_{t-1} \sqrt{16d \log (\Gamma_t - I) + 1} + 32d \log \frac{15T}{2\delta} + \frac{3}{4} t R^2 - \sqrt{16d \log (\Gamma_t - I) + 1} + 32d \log \frac{15T}{2\delta} \leq \frac{R^2}{4}$$

$$\geq \gamma_{t-1}^2 \left( 1 - \frac{16d \log (\Gamma_t - I) + 1}{} \right) + \frac{t R^2}{4}$$

$$\geq \gamma_{t-1}^2 \left( 1 - \frac{16d \log (\Gamma_t - I) + 1}{} \right) + \frac{TR^2}{8}$$

$$\geq \frac{TR^2}{8}$$

(62)

Observe that

$$\gamma_{t-1} = \sqrt{U^T A Y_{t-1} A U} \geq \sigma_{\min}(A) \frac{TR^2}{8\epsilon}$$

(63)
Eq. (62) will give us a non–trivial bound only when \( \frac{c(A,\delta)}{\gamma_{t-1}} \leq 1/4 \) which is true whenever

\[
T \geq \frac{64ec(A,\delta)}{R^2\sigma_{\min}^2(A)} \tag{64}
\]

The scaling \( 1 - \sqrt{\frac{c(A,\delta)}{\gamma_{t-1}}} \) in Eq. (62) depends on \( \gamma_{t-1} \) itself. We will show that

\[
\gamma_{t-1}^2 = T\Omega(1) \implies \gamma_{t-1}^2 = T\Omega\left(\frac{T}{c(A,\delta)}\right)
\]

\[
\gamma_{t-1}^2 = T\Omega\left(\left(\frac{T}{c(A,\delta)}\right)^{1/2}\right) \implies \gamma_{t-1}^2 = T\Omega\left(\left(\frac{T}{c(A,\delta)}\right)^{3/4}\right)
\]

\[
\gamma_{t-1}^2 = T\Omega\left(\left(\frac{T}{c(A,\delta)}\right)^{2k+1}\right) \implies \gamma_{t-1}^2 = T\Omega\left(\left(\frac{T}{c(A,\delta)}\right)^{2k+1}\right)
\]

\[
\implies \ldots \implies \gamma_{t-1}^2 = T\Omega\left(\frac{T}{c(A,\delta)}\right)
\]

From Eq. (62),(63) since

\[
\sqrt{\frac{c(A,\delta)}{\gamma_{t-1}^2}} \leq \sqrt{\frac{16ec(A,\delta)}{\sigma_{\min}(AA')T}} = \beta_1
\]

it follows that

\[
Y_t \geq \left(1 - \sqrt{\frac{16ec(A,\delta)}{\sigma_{\min}(AA')TR^2}}\right)AY_{t-1}A' + \frac{R^2TI}{8}
\]

The goal here is to refine the upper bound for \( \sqrt{\frac{c(A,\delta)}{\gamma_{t-1}^2}} \) such that

\[
\sqrt{\frac{c(A,\delta)}{\gamma_{t-1}^2}} \leq \frac{C}{T}
\]

Eq. (65) implies that

\[
\begin{align*}
Y_t &\geq \frac{T \min\left(1, \frac{T}{4}\right)}{R^2} \sum_{k=1}^{\min\left(1, \frac{T}{4}\right)} (1 - \beta_1)^k A^k A^k A' + \frac{R^2TI}{16} \\
&\geq \frac{T \min\left(1, \frac{T}{4}\right)}{16e} A^k A^k + \frac{R^2TI}{16} \\
&\geq \frac{R^2T}{16e} \Gamma_{\frac{T}{4}}(A) + \frac{R^2TI}{16}
\end{align*}
\]

Here

\[
\beta_1 = \sqrt{\frac{16ec(A,\delta)}{\sigma_{\min}(AA')R^2T}} \tag{66}
\]

Due to the choice of \( T, d \) we will usually have \( \left|\frac{1}{\beta_1}\right| \leq \frac{T}{4} \). (a) follows by successively expanding Eq. (65), (b) follows because \( (1 - \beta_1)^\frac{T}{4} \geq \frac{e^{-1}}{2} \) since \( \beta_1 \leq 1/2 \). Then we can conclude that

\[
\gamma_{t-1}^2 \geq \sigma_{\min}(AY_tA')
\]

\[
\geq \frac{R^2T\sigma_{\min}(AA')\sigma_{\min}(\Gamma_{\frac{T}{4}}(A))}{16e}
\]

\[
\geq \frac{R^2T\sigma_{\min}(AA')\sigma_{\min}(\Gamma_{\frac{T}{4}}(A))}{16e} \tag{67}
\]
which gives us
\[
\sqrt{\frac{c(A, \delta)}{\gamma_{t-1}^2}} \leq \left( \frac{16ec(A, \delta)}{R^2T\sigma_{\min}(AA')\sigma_{\min}(\Gamma_{\frac{1}{\sigma_{m-1}}} (A))} \right)^{1/2} = \beta_2
\]  
(68)

It is clear from Eq. (68) that we get a recursion during the refinement process. Specifically at the \(k^{th}\) repetition of Eq. (68) up to Eq. (68) we get,
\[
\beta_k = \left( \frac{16ec(A, \delta)}{R^2T\sigma_{\min}(AA')\sigma_{\min}(\Gamma_{\frac{1}{\sigma_{m-1}}} (A))} \right)^{1/2}
\]  
(69)

Now \(\beta_k\) is a non-increasing sequence. We show this by induction. Since \(\sigma_{\min}(\Gamma_t(A)) \geq 1\) and
\[
\sqrt{\frac{16ec(A, \delta)}{\sigma_{\min}(AA')R^2T}} \leq 1
\]
it follows trivially that \(\beta_2 \leq \beta_1\). Assume our hypothesis holds for all \(k \leq m\). Then since \(\Gamma_{t_1}(A) \geq \Gamma_{t_2}(A)\) whenever \(t_1 \geq t_2\) we have
\[
\left( \frac{16ec(A, \delta)}{R^2T\sigma_{\min}(AA')\sigma_{\min}(\Gamma_{\frac{1}{\sigma_{m-1}}} (A))} \right)^{1/2} \leq \left( \frac{16ec(A, \delta)}{R^2T\sigma_{\min}(AA')\sigma_{\min}(\Gamma_{\frac{1}{\sigma_{m-1}}} (A))} \right)^{1/2}
\]
\[
\beta_{m+1} \leq \beta_m
\]
and we have proven our hypothesis. To now find the best upper bound for \(\sqrt{\frac{c(A, \delta)}{\gamma_{t-1}^2}}\) we find the steady state solution for Eq. (69), i.e.
\[
\beta_0^2\sigma_{\min}(\Gamma_{\frac{1}{\sigma_{m}}} (A)) = \left( \frac{16ec(A, \delta)}{R^2T\sigma_{\min}(AA')} \right)
\]  
(70)

Now a solution for \(\beta_0 \in (\frac{2C}{\sigma_{\min}(AA')R^2T}, 1)\). To see this set \(\beta_0 = 1\), then LHS > RHS. Next set \(\beta_0 = \frac{2C}{\sigma_{\min}(AA')R^2T}\) then since \(\rho_{\min}(A^t) \geq \sigma_{\min}(A^t)\) and \(\rho_t \leq 1 + C/T\) we see that
\[
\frac{4C^2\sigma_{\min}(\Gamma_{\frac{1}{\sigma_{m}}} (A))}{\sigma_{\min}(AA')^2R^2T} \leq \frac{4\sum_{t=0}^\infty \sigma_{\min}(A^t)2^t}{R^4\sigma_{\min}(AA')^2T^2/C^2} \leq \frac{2eC}{\sigma_{\min}(AA')^2T} \right) \leq \left( \frac{16ec(A, \delta)}{R^2T\sigma_{\min}(AA')} \right)
\]
and LHS < RHS because \(C\) is a constant but \(c(A, \delta)\) is growing logarithmically with \(T\) (and we can pick \(T\) accordingly). By ensuring that
\[
T \geq \frac{64ec(A, \delta)}{R^2\sigma_{\min}(A)^2}
\]
we also ensure that \(\beta_1 < 1/2\) and as a result all subsequent \(\beta_k < 1/2\). Now we can conclude that whenever \(T \geq \frac{64ec(A, \delta)}{\sigma_{\min}(AA')^2}\) we get Eq. (65)
\[
Y_t \geq (1 - \beta_0)AY_{t-1}A' + \frac{TR^2I}{8}
\]  
(71)

and following as before we get with probability at least \(1 - \delta\)
\[
Y_T \geq \frac{TR^2}{16c} + \frac{TR^2I}{16}
\]  
(72)

where \(\beta_0\) is solution to
\[
\beta_0^2\sigma_{\min}(\Gamma_{\frac{1}{\sigma_{m}}} (A)) = \left( \frac{16ec(A, \delta)}{TR^2\sigma_{\min}(AA')} \right)
\]
and
\[
c(A, \delta) = 16d \log (\text{tr}(\Gamma_T - I) + 1) + 32d \log \frac{15T}{2d}
\]
It should be noted that \( \frac{1}{\delta^2} \) will equal \( \frac{\sqrt{\alpha(d) R^2 \sigma_{\min}(AA')}}{16ec(A, \delta)} \), i.e., grow linearly with \( T \), as shown in Proposition 7.5. Then it can be seen from Eq. (72) that

\[
Y_T \geq \frac{TR^2}{16e} \Gamma_{[\frac{T}{\delta^2}]}(A) + \frac{TR^2 I}{16} \\
Y_T \geq \frac{TR^2}{16e} \sigma_{\min}(\Gamma_{[\frac{T}{\delta^2}]}(A)) + \frac{TR^2 I}{16} \\
\geq \frac{TR^2}{16e} \sqrt{\alpha(d) \sigma_{\min}(AA')} I = \frac{\sqrt{\alpha(d) R^2 \sigma_{\min}(AA')}}{256ec(A, \delta)}
\]  

(73)

11 Invertibility of \( Y_T \) in explosive systems

Assume for this case that \( \eta_t = L \bar{\eta}_t \) where \( \{\bar{\eta}_t\}_{t=1}^T \) are i.i.d and all elements of \( \bar{\eta}_t \) are independent. Further \( L \) is full row rank. Define \( \sigma_{\min}(LL') = R^2 > 0 \). Let \( \sigma_{\max}(LL') = 1 \). Recall that

\[
z_t = A^{-t} x_t = x_0 + \sum_{\tau=1}^t A^{-\tau} \eta_{\tau}
\]

Define

\[
z(T, t) = \left( \sum_{s=0}^{t-1} A^{-s} \eta_{T+1-t+s} \right)
\]

where \( z(T, t) = 0 \) for \( t \leq 0, t \geq T + 1 \). An observation that will be useful is that \( z(t) \) is statistically independent of \( z(T) - z(t) \). Recall that \( U_T = A^{-T} \sum_{t=1}^T x_t A^{-T'} \), \( F_T = \sum_{t=1}^T A^{-t+1} z_T z_t' A^{-t+1} \)

Bounding \( ||F_T - U_T||_{op} \)

Observe that

\[
z(T) - z(T - t) = A^{-T+t-1} \left( \sum_{s=0}^{t-1} A^{-s} \eta_{T+1-t+s} \right) = A^{-T+t-1} z(T, t)
\]

Then

\[
||U_T - F_T||_{op} = ||\sum_{t=1}^T A^{-t}(z(T - t) z(T - t)' - z(T) z(T)')(A^{-t})'||_2
\]

Let \( u = z(T - t) \), \( v = z(T) \) and since \( uu' - vv' = (u - v)u' + u(u - v)' - (u - v)(u - v)' \) we have

\[
||U_T - F_T||_{op} \leq ||\sum_{t=1}^T A^{-t}(z(T - t) - z(T))(z(T - t) - z(T))' A^{-t'}||_2
\]

\[
+ ||\sum_{t=1}^T A^{-t}((z(T - t) - z(T)) z(T - t)' + z(T - t) z(T - t) - z(T)')(A^{-t})'||_2
\]

(75)

The reason we decompose it in such a way is so that we can represent the cross terms \( (z(T - t) - z(T))(z(T - t)') \) as the product of independent terms. This will be useful in using Hanson–Wright bounds as we show later.

First we bound

\[
||\sum_{t=1}^T A^{-t}(z(T - t) - z(T))(z(T - t) - z(T))' A^{-t'}||_2
\]
From Eq. (74) we see that $A^{-t}(z(T-t) - z(T)) = -A^{-T-1}z(T,t)$, then

$$A^{-T-1}z(T,t) = A^{-T-1}[0,0\ldots, \sum_{t=1}^{T-t+1} \eta_1, A^{-1}, A^{-2}, \ldots, A^{-t+1}]$$

Since $\sum_{t=1}^{T} (z(T-t) - z(T))^t \leq \sum_{t=1}^{T} \text{trace}((z(T-t) - z(T))^t)I$. Based on these observations we have

$$||\sum_{t=1}^{T} A^{-t}(z(T-t) - z(T))(z(T-t) - z(T))^t A^{-t'}||_2 = ||\sum_{t=1}^{T} A^{-T-1}z(T,t)z(T,t)^t A^{-T-1'}||_2$$

$$\leq \text{trace}(A^{-T-1} \sum_{t=1}^{T} z(T,t)z(T,t)^t A^{-T-1'}) = \sum_{t=1}^{T} z(T,t)^t A^{-T-1'} A^{-T-1} z(t,T) = \tilde{\eta}' \tilde{A} \tilde{\eta}$$

where $\tilde{\eta} = \left[ \begin{array}{c} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_T \end{array} \right]$ and

$$\tilde{A} = \begin{bmatrix} 0 & 0 & \cdots & 0 & A^{-T-1} \\ 0 & 0 & \cdots & A^{-T-1} & A^{-T-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ A^{-T-1} & A^{-T-2} & \cdots & A^{-2T+1} & A^{-2T} \end{bmatrix}$$

Since $\text{tr}(\tilde{A} \tilde{A}') = T\text{tr}(A^{-T-1} \Gamma_T(A^{-1}) A^{-T-1'})$. Applying Markov’s Inequality (See Proposition 8.4), we have with probability at least $1 - \delta$ that

$$\tilde{\eta}' \tilde{A} \tilde{\eta} \leq \frac{\text{tr}(E[\tilde{A} \tilde{\eta}' \tilde{A}'])}{\delta} \leq \frac{\sigma_1(L)^2 T \text{tr}(A^{-T-1} \Gamma_T(A^{-1}) A^{-T-1'})}{\delta}$$

(76)

Although this bound can be tightened by dependent Hanson–Wright (See Corollary 8.1), there is no reason to do so as $\delta$ depends only logarithmically on $T$.

Next we analyze the second term

$$||\sum_{t=1}^{T} A^{-t}((z(T-t) - z(T))(z(T-t) - z(T))^t + z(T-t)(z(T-t) - z(T))^t) A^{-t'}||_2$$

Consider the summand $\sum_{t=1}^{T} A^{-t}((z(T-t) - z(T))^t A^{-t'})$, then

$$\sum_{t=1}^{T} A^{-t}((z(T-t) - z(T))^t A^{-t'}) = A^{-T-1} \sum_{t=1}^{T} z(T,t)(z(T-t)) A^{-t'}$$

(77)
We define scaled version of $z(T, t), z(T - t)$.

$$
\dot{z}(T, t) = A^{-T-1}z(T, t) = A^{-T-1} \left[ 0, 0, \ldots, \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_T \end{bmatrix} \right] \\
\dot{z}(T - t)' = z(T - t)' A^{-T} = \begin{bmatrix} \eta_1' \\ \eta_2' \\ \vdots \\ \eta_T' \end{bmatrix} + x_0
$$

Then the probability of the second term can be written as

$$
P(\| \sum_{t=1}^{T} (\dot{z}(T, t)\dot{z}(T - t)' + \dot{z}(T - t)\dot{z}(T, t)') \|_2 \geq z) \leq 2 \times 5^{2d} \times P\left( \sum_{t=1}^{T} 2uv^T \dot{z}(T, t)\dot{z}(T - t)'v \right) \geq z/4)$$

$$
\leq 2 \times 5^{2d} \times P\left( \sum_{t=1}^{T} A(T, t)'A^{-T-1}uv'A(T - t) + A(T - t)'vu'A^{-T-1}A(T, t) \right) \leq z/4)
$$

(78)

To Eq. (78) apply Hanson-Wright inequality. For any $u, v$, due to the statistical independence of $z(T - t), z(T)$ we have

$$
E\left[ \sum_{t=1}^{T} 2uv^T \dot{z}(T, t)\dot{z}(T - t)'v \right] = 0
$$

We now need an upper bound on $\|S\|_2, \|S\|_F$. Since $CD' + DC' \leq CC' + DD'$

$$
S = \sum_{t=1}^{T} A(T, t)'A^{-T-1}uv'A(T - t) + A(T - t)'vu'A^{-T-1}A(T, t)
$$

$$
\leq \sum_{t=1}^{T} A(T, t)'A^{-T-1}uv'A(T - t) + A(T - t)'vu'A^{-T-1}A(T, t)
$$

$$
\leq \sum_{t=1}^{T} A(T, t)'A^{-T-1}uv'A(T - t) + A(T - t)'vu'A^{-T-1}A(T, t)
$$

$$
\leq \sigma_1^2 (A^{-T-1})^\epsilon \sum_{t=1}^{T} A(T, t)'A(T - t) + u'A^{-T-1}A^{-T-1}A(T - t)
$$

$$
\leq \sigma_1^2 (A^{-T-1})^\epsilon \text{tr} \left( \sum_{t=1}^{T} A(T, t)'A(T, t) \right) I + \sigma_1^2 (A^{-T-1})^\epsilon \text{tr} \left( \sum_{t=1}^{T} A(T - t)'vu'A(T - t) \right) I
$$

\[ a \]

Here \[ a \] follows because

$$
A(T, t)A(T, t)' = \Gamma_{t-1}(A), A(T - t)A(T - t)' = \Gamma_{T-t}(A)
$$
Then whenever
\[ T \geq T_0 = \frac{2}{c} \left( \log \frac{1}{\delta} + \log 2 + 2d \log 5 \right) \]
(79)

Eq. (78) becomes with probability at least \(1 - \delta\) that
\[ \left\| \sum_{t=1}^{T} (z(T - t) - z(T))z(T - t)' + z(T - t)'(z(T - t) - z(T))' \right\|_2 \leq 4T^2 \sigma_1^2 (A^{(T+1)^r}) \text{tr} (\Gamma_T (A^{-1})) \]
(80)

Then combining Eq. (76), (80) we get for \(T \geq T_0\) given in Eq. (79),
\[ \left\| U_T - F_T \right\|_2 \leq \left( 4T^2 \sigma_1^2 (A^{(T+1)^r}) \text{tr} (\Gamma_T (A^{-1})) + \frac{T \text{tr} (A^{-T-1} \Gamma_T (A^{-1}) A^{T-1})}{\delta} \right) \]
(81)

with probability at least \(1 - 2\delta\). We pick \(\epsilon\) such that \((T + 1)\epsilon = \lceil \frac{T+1}{2} \rceil\).

**Bounding \(U_T\)**

To give lower and upper bounds on \(U_T\), we need to bound \(F_T\). The steps involve
\[ \left\| U_T - F_T \right\|_2 \leq \Delta \]
\[ F_T \geq V_{dn} > 0 \]
\[ \Rightarrow U_T \geq V_{dn} - \Delta I \]
\[ F_T \leq V_{up} \]
\[ \Rightarrow U_T \leq V_{up} + \Delta I \]

From Proposition [12.1] we get, with probability at least \(1 - 2\delta\),
\[ F_T \geq \phi_{\min}(A)^2 \psi(A)^2 \delta^2 \sigma_{\min}(P^{-1})^2 I \]
\[ F_T \leq \frac{\phi_{\max}(A)^2}{\sigma_{\min}(P)^2} (1 + \frac{1}{c} \log \frac{1}{\delta}) \text{tr} (P (\Gamma_T (A^{-1}) - I) P') I \]

Define
\[ \Delta = \frac{1}{2} \min \left( \frac{\phi_{\max}(A)^2}{\sigma_{\min}(P)^2} (1 + \frac{1}{c} \log \frac{1}{\delta}) \text{tr} (P (\Gamma_T (A^{-1}) - I) P') , \phi_{\min}(A)^2 \psi(A)^2 \delta^2 \sigma_{\min}(P^{-1})^2 \right) \]
\[ \Delta = \frac{\phi_{\min}(A)^2 \psi(A)^2 \delta^2 \sigma_{\min}(P^{-1})^2}{2} \]

Then in Eq. (81) by ensuring that
\[ \left( 4T^2 \sigma_1^2 (A^{(T+1)^r}) \text{tr} (\Gamma_T (A^{-1})) + \frac{T \text{tr} (A^{-T-1} \Gamma_T (A^{-1}) A^{T-1})}{\delta} \right) \leq \frac{\phi_{\min}(A)^2 \psi(A)^2 \delta^2}{2 \sigma_{\max}(P)^2} \]
we get with probability at least \(1 - 4\delta\) (since this is the intersection of events governed by Eq. (81), (87), (88))
\[ U_T \geq \phi_{\min}(A)^2 \psi(A)^2 \delta^2 \sigma_{\min}(P^{-1})^2 I - \frac{\phi_{\min}(A)^2 \psi(A)^2 \delta^2}{2 \sigma_{\max}(P)^2} I \geq \frac{\phi_{\min}(A)^2 \psi(A)^2 \delta^2}{2 \sigma_{\max}(P)^2} I \]
(82)

Similarly, for the upper bound
\[ U_T \leq \frac{3\phi_{\max}(A)^2}{2 \sigma_{\min}(P)^2} (1 + \frac{1}{c} \log \frac{1}{\delta}) \text{tr} (P (\Gamma_T (A^{-1}) - I) P') I \]
(83)
Thus with probability at least $1 - 4\delta$ we have
\[
Y_T \geq \frac{\phi_{\min}(A)^2 \psi(A)^2 \delta^2}{2\sigma_{\max}(P)^2} A^T A^T \gamma
\]
\[
Y_T \leq \frac{3\phi_{\max}(A)^2}{2\sigma_{\min}(P)^2} (1 + \frac{1}{c} \log \frac{1}{\delta}) \text{tr}(P(\Gamma(T)(A^{-1}) - I)P')A^T A^T \gamma
\]
whenever
\[
\left(4T^2 \sigma^2_1(A^{-(T+1)c}) \text{tr}(\Gamma(T)(A^{-1}))) + \frac{T \text{tr}(A^{-(T-1)}(A^{-1})A^{-(T-1)'} I}{\delta} \right) \leq \frac{\phi_{\min}(A)^2 \psi(A)^2 \delta^2}{2\sigma_{\max}(P)^2}
\]

12 Regularity and Invertibility

Through a counterexample in [11], Remark 4 in [14] it is shown that unless a matrix is regular, the estimation of the parameters may be statistically inconsistent.

Recall $F_T$ from Eq. [13]. Assume again that $\eta_t = L\bar{\eta}_t$ where $\{\bar{\eta}_t\}_{t=1}^T$ are i.i.d isotropic subGaussian and all elements of $\bar{\eta}_t$ are independent. Further $L$ is full row rank. Define $\sigma_{\min}(LL') = R^2 > 0$. Let $\sigma_{\max}(LL') = 1$ (this does not affect the main result as it appears only as a scaling). For the invertibility of $Y_T$ in explosive systems, it will be important that $F_T$ is invertible with high probability. It will turn out that invertibility of $F_T$ can be ensured by assuming regularity of $A$. This is Proposition 1 in [5] and has been presented here for completeness. It will be useful to recall the definitions of $\phi_{\min}(A), \phi_{\max}(A)$ from Definition [5]

We will show $F_T$ indeed has rank $d$ with probability 1. Formally,

**Proposition 12.1.** Let $A$ be regular, then we have with probability at least $1 - 2\delta$

\[
\sigma_{\min}(F_T) \geq \frac{\phi_{\min}(A)^2}{\sigma_{\max}(P)^2} \psi(A)^2 \delta^2
\]
\[
\sigma_{\max}(F_T) \leq \frac{\phi_{\max}(A)^2}{\sigma_{\min}(P)^2} (1 + \frac{1}{c} \log \frac{1}{\delta}) \text{tr}(P(\Gamma(T)(A^{-1}) - I)P')
\]

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

**Proof.** Let $S_k = [z_T, A^{-1}z_T, \ldots, A^{-k}z_T]$ where $z_T = A^{-1}x_T = A^{-1}(\sum_{k=0}^{T-1} A^k \bar{\eta}_{T-k})$. Note that $L\bar{\eta}_t$ is continuous whenever $L$ is full row rank. Then $F_T = S_T' S_T$. Observe that

\[
A^{-t} z_T = P^{-1} \Lambda^{-t} P z_T
\]

Define the event
\[
E_+(\delta) = \{ \min_{1 \leq i \leq d} |P'_i z_T| > \psi(A) \delta \}
\]

where $\psi(A)$ is the lower bound shown in Proposition [3.3] (which we can use due to the continuity of $L\bar{\eta}_t$) and $v = P z_T$. Under $E_+(\delta), |v_i| > 0$. Now we need a lower bound for $\sigma_{\min}(F_T)$ under $E_+(\delta)$

\[
F_T = P^{-1} \sum_{i=1}^T A^{-i+1} P z_T z_T' P' A^{i-1} P^{-1} = P^{-1} \sum_{i=1}^T A^{-i+1} \psi(A) \Lambda^{-i+1} P^{-1} P
\]
\[
\geq \phi_{\min}(A)^2 \psi(A)^2 \delta^2 P^{-1} P^{-1} \geq \frac{\phi_{\min}(A)^2}{\sigma_{\max}(P)^2} \psi(A)^2 \delta^2 I
\]

Further, since $A$ is regular we have that $\phi_{\min}(A) > 0$ from Proposition [7.3]. Then with probability at least $1 - \delta$ we have

\[
\sigma_{\min}(F_T) \geq \frac{\phi_{\min}(A)^2}{\sigma_{\max}(P)^2} \psi(A)^2 \delta^2 > 0
\]

For the upper bound, observe that $P z_T$ is a sub-Gaussian random variable. Since

\[
||P z_T z_T' P'|| \leq z_T' P' P z_T
\]
and recalling that

\[
\begin{bmatrix}
A^{-1}, A^{-2}, \ldots, A^{-T}
\end{bmatrix}
\]

we can use dependent Hanson Wright inequality (Corollary 8.1) to bound \(z_T' P^T P z_T\). In Theorem 4, \(B = \tilde{A}' P' P \tilde{A}\)

\[
\begin{align*}
\mathbb{E}[z_T' P^T P z_T] &= \text{tr}(P(\Gamma_T(A^{-1}) - I)P')\sigma_1(L)^2 = \text{tr}(P(\Gamma_T(A^{-1}) - I)P') \\
||B||_2, ||B||_F &\leq \text{tr}(\tilde{A}' P' P \tilde{A}) = \text{tr}(P(\Gamma_T(A^{-1}) - I)P')
\end{align*}
\]

Then with probability at least \(1 - \delta\) we have

\[
z_T' P^T P z_T \leq (1 + \frac{1}{c} \log \frac{1}{\delta}) \text{tr}(P(\Gamma_T(A^{-1}) - I)P')
\]

and we get from Eq. (86)

\[
F_T \leq \sum_{i=1}^{T} \Lambda^{-i+1} P z_T z_T' P' \Lambda^{-i+1}' P^{-1}'
\]

\[
\leq (z_T' P^T P z_T) \sup_{||v||_2=1} \sigma_{\max} \left( P^{-1} \sum_{i=1}^{T} \Lambda^{-i+1} v' \Lambda^{-i+1}' P^{-1}' \right) I
\]

\[
\leq \frac{\phi_{\max}(A)^2}{\sigma_{\min}(P)^2} (1 + \frac{1}{c} \log \frac{1}{\delta}) \text{tr}(P(\Gamma_T(A^{-1}) - I)P') I 
\]

Then we have with probability at least \(1 - 2\delta\)

\[
F_T \geq \frac{\phi_{\min}(A)^2}{\sigma_{\max}(P)^2} \psi(A)^2 \delta^2 I \quad (89)
\]

\[
F_T \leq \frac{\phi_{\max}(A)^2}{\sigma_{\min}(P)^2} (1 + \frac{1}{c} \log \frac{1}{\delta}) \text{tr}(P(\Gamma_T(A^{-1}) - I)P') I \quad (90)
\]

\[\square\]

### 13 Composite Result

In this section we discuss error rates for regular matrices which may have eigenvalues anywhere in the complex plane. The key step is to recall that for every matrix \(A\) it is possible to find \(\hat{P}\) such that

\[
A = \hat{P}^{-1} \begin{bmatrix}
A_e & 0 & 0 \\
0 & A_{ms} & 0 \\
0 & 0 & A_s
\end{bmatrix} \hat{P}
\]

Here \(A_e, A_{ms}, A_s\) are the purely explosive, marginally stable and stable portions of \(A\). This follows because any matrix \(A\) has a Jordan normal form \(A = P^{-1} \Lambda P\), where \(\Lambda\) is a block diagonal matrix and each block corresponds to an eigenvalue. We can always find \(Q\) (a rearrangement matrix) such that \(\Lambda\) is partitioned into two diagonal parts: explosive, marginally stable and stable, i.e.,

\[
A = P^{-1} Q^T \begin{bmatrix}
\Lambda_e & 0 & 0 \\
0 & \Lambda_{ms} & 0 \\
0 & 0 & \Lambda_s
\end{bmatrix} Q P
\]
Clearly, \( \tilde{P} = QP \). Since

\[
X_t = \sum_{\tau=1}^{\tau} A_{\tau}^{-1} \eta_{t-\tau+1}
\]

\[
\dot{X}_t = \tilde{P} X_t = \sum_{\tau=1}^{\tau} A_{\tau}^{-1} \tilde{P} \eta_{t-\tau+1}
\]

(93)

Now, the transformed dynamics are as follows:

\[
\dot{\tilde{X}}_{t+1} = \tilde{A} \tilde{X}_t + \tilde{\eta}_{t+1}
\]

where \( \tilde{A} \) has been partitioned into explosive and stable components as Eq. (91). Corresponding to \( \tilde{A} \) partition \( \tilde{X}_t, \tilde{\eta}_t \)

\[
\tilde{X}_t = \begin{bmatrix} X^e_t \\ X^{\text{mss}}_t \\ X^s_t \end{bmatrix}, \tilde{\eta}_t = \begin{bmatrix} \eta^e_t \\ \eta^{\text{mss}}_t \\ \eta^s_t \end{bmatrix}
\]

(94)

\[
\tilde{Y}_T = \sum_{t=1}^T \tilde{X}_t X_t = \sum_{t=1}^T \begin{bmatrix} X^e_t (X_t)' \\ X^{\text{mss}}_t (X_t)' \\ X^s_t (X_t)' \end{bmatrix} \begin{bmatrix} X^e_t (X^e_t)' \\ X^{\text{mss}}_t (X^{\text{mss}}_t)' \\ X^s_t (X^s_t)' \end{bmatrix}
\]

(95)

We analyze the error of identification in the transformed system instead and show how it relates to the actual error. Note that \( \tilde{P} \) is unknown, the transformation is done for ease of analysis. The invertibility of submatrix corresponding to stable and marginally stable components, \( i.e., \)

\[
X^{\text{mss}}_t = \begin{bmatrix} X^{\text{mss}}_t \\ X^s_t \end{bmatrix}
\]

follows from Theorem [1]. To see this let \( A_e \) be a \( d_e \times d_e \) matrix. Define

\[
P_{\text{mss}} = \tilde{P}[d_e + 1 : d, :]
\]

\( i.e., P_{\text{mss}} \) is the rectangular matrix formed by removing the rows of \( \tilde{P} \) corresponding to the explosive part. Then, by definition, we have that

\[
\begin{bmatrix} \eta^{\text{mss}}_t \\ \eta^s_t \end{bmatrix} = P_{\text{mss}} \eta_t
\]

and

\[
X^{\text{mss}}_{t+1} = \begin{bmatrix} A_{\text{mss}} & 0 \\ 0 & A_s \end{bmatrix} X^{\text{mss}}_t + \begin{bmatrix} \eta^{\text{mss}}_{t+1} \\ \eta^s_{t+1} \end{bmatrix}
\]

Further

\[
\mathbb{E}[P_{\text{mss}} \eta_t P_{\text{mss}}'] = P_{\text{mss}} P_{\text{mss}}' > 0
\]

Since all rows of \( \tilde{P} \) are independent then \( P_{\text{mss}} P_{\text{mss}}' \) is invertible and \( \{P_{\text{mss}} \eta_t\}_{t=1}^T \) are independent subGaussian vectors. Now this is the same set up as the general version of Theorem [1] discussed in Section [9]. Since \( A_{\text{mss}} \in S_0 \cup S_1 \) only has stable and marginally stable components, it follows from the Eq. (57) that

\[
\sum_{t=1}^T X^{\text{mss}}_t (X^{\text{mss}}_t)' \geq \frac{T}{4} \sigma_{\min}(P_{\text{mss}} P_{\text{mss}}') I
\]

with high probability. Then since \( \sigma_{\min}(P_{\text{mss}} P_{\text{mss}}') \geq \sigma_{\min}(\hat{P})^2 = R^2 \), we have that \( \sum_{t=1}^T X^{\text{mss}}_t (X^{\text{mss}}_t)' \geq \frac{T R^2}{4} I \). Let \( \sigma_{\max}(\hat{P}) = 1 \). (this makes no difference to the results and \( R \) can be interpreted as the inverse condition number)

Recall the definition of \( \beta_0(\delta) \)

\[
\beta_0(\delta) = \inf \left\{ \beta | \beta^2 \sigma_{\min}(\Gamma | \frac{1}{n})(A) \right\} \geq \left( \frac{\sec(A, \delta)}{TR^2 \sigma_{\min}(AA')} \right)
\]
we refer to $\beta_0(\delta)$ as $\beta_0$. Following our discussion in Proposition 7.5 we see that $\beta_0 > 0$ and since $\sigma_{\min}(\Gamma_t(A)) \geq \alpha(d)t$ we have that

$$\beta_0 \leq \frac{8cc(A, \delta)}{TR^2\sigma_{\min}^2(A)C(d)} \implies \frac{1}{\beta_0} \geq \frac{TR^2\sigma_{\min}^2(A)C(d)}{8cc(A, \delta)}$$

Define

$$V_e = \sum_{t=1}^{T} X_t^e (X_t^e)'$$, $V_\nu = TR^2 4I$, $V_{ms} = \left( \frac{TR^2}{8c} \Gamma_{\frac{1}{\gamma^2}}(A_{ms}) \right)$

where the invertibility in $V_e$ holds with high probability. Observe that $V_{ms} \lesssim \sum_{t=1}^{T} X_t^m (X_t^m)'$, $V_\nu \lesssim \sum_{t=1}^{T} X_t^e (X_t^e)'$ with high probability (follows from Eq. (57), (72)). This observation will be useful in proving the composite invertibility.

Although the technique to prove the invertibility of $\sum_{t=1}^{T} X_t^e X_t^e'$ is similar in spirit to that of [6], it addresses additional difficulties arising due to the presence of a marginally stable block.

$$B_{d \times d} = \begin{bmatrix} V_{e}^{-1/2} & 0 & 0 \\ 0 & V_{ms}^{-1/2} & 0 \\ 0 & 0 & V_{s}^{-1/2} \end{bmatrix} \tag{96}$$

We will show that $B \sum_{t=1}^{T} X_t^e X_t^e' B'$ is positive definite with high probability, i.e.,

$$\sum_{t=1}^{T} B X_t^e X_t^e' B' = \begin{bmatrix} 1 & \sum_{t=1}^{T} V_{ms}^{-1/2} X_t^m (X_t^m)' V_{ms}^{-1/2} & \sum_{t=1}^{T} V_{ms}^{-1/2} X_t^m (X_t^m)' V_{ms}^{-1/2} \\ \sum_{t=1}^{T} V_{ms}^{-1/2} X_t^m (X_t^m)' V_{ms}^{-1/2} & \sum_{t=1}^{T} V_{ms}^{-1/2} X_t^m (X_t^m)' V_{ms}^{-1/2} & \sum_{t=1}^{T} V_{ms}^{-1/2} X_t^m (X_t^m)' V_{ms}^{-1/2} \\ \sum_{t=1}^{T} V_{ms}^{-1/2} X_t^m (X_t^m)' V_{ms}^{-1/2} & \sum_{t=1}^{T} V_{ms}^{-1/2} X_t^m (X_t^m)' V_{ms}^{-1/2} & \sum_{t=1}^{T} V_{ms}^{-1/2} X_t^m (X_t^m)' V_{ms}^{-1/2} \end{bmatrix} \tag{97}$$

We already showed that lower submatrix is invertible. To show that the entire matrix is invertible we need to show

$$||V_{e}^{-1/2} \sum_{t=1}^{T} X_t^e (X_t^e)' V_{ms}^{-1/2}||, ||V_{e}^{-1/2} \sum_{t=1}^{T} X_t^e (X_t^e)' V_{s}^{-1/2}|| < \gamma/8$$

with high probability for some appropriate $\gamma$ and

$$\sigma_{\min}\left( \begin{bmatrix} V_{ms}^{-1/2} & 0 \\ 0 & V_{s}^{-1/2} \end{bmatrix} \sum_{t=1}^{T} X_t^{mass} (X_t^{mass})' \begin{bmatrix} V_{ms}^{-1/2} & 0 \\ 0 & V_{s}^{-1/2} \end{bmatrix} \right) \geq \gamma > 0$$

### 13.1 Cross Terms have low norm

Define the following quantities:

$$\alpha(A_c, \delta) = \frac{3\phi_{\max}(A_c)^2 \sigma_{\max}^2(A_c)}{\phi_{\min}(A_c)^2 \sigma_{\min}(A_c)^2} \left( 1 + \frac{1}{c} \log \frac{1}{\delta} \right) \text{tr}(P_e (\Gamma_T(A_c^{-1} - I)) P_e)$$ \tag{98}

$$T_{mc}(\delta) = \left\{ T \mid \alpha(A_c, \delta) \text{tr}(A_c^{-T+k_{mc}(T)} (A_c^{-T+k_{mc}(T)})') \leq \frac{\gamma^2}{256} \right\} \tag{99}$$

$$k_{mc} = k_{mc}(T) = T \left( 1 - \frac{R^2 \gamma^2}{2048d \lambda_1 (\Gamma_T(A_{ms}) \Gamma_{1/m}(A_{ms}) \left( 1 + \frac{1}{c} \log \frac{1}{\delta} \right))} \right) \tag{100}$$

$$T_{sc}(\delta) = \left\{ T \mid \alpha(A_c, \delta) \text{tr}(A_c^{-T+k_{sc}(T)} (A_c^{-T+k_{sc}(T)})') \leq \frac{\gamma^2}{256} \right\} \tag{101}$$

$$k_{sc} = k_{sc}(T) = T \left( 1 - \frac{R^2 \gamma^2}{1024d \lambda_1 (\Gamma_T(A_{c}) \left( 1 + \frac{1}{c} \log \frac{1}{\delta} \right))} \right) \tag{102}$$
Remark 5. Note that $T_{mc}(\delta)$ (and $T_{sc}(\delta)$) is a set where there exists a minimum $T_* < \infty$ such that $T \in T_{mc}(\delta)$ whenever $T \geq T_*$. However, there might be $T < T_*$ for which the inequality of $T_{mc}(\delta)$ holds. Whenever we write $T \in T_{mc}(\delta)$ we mean $T \geq T_*$. Second note that for every $T$, since $R, \gamma < 1$ we have

$$k_{sc}(T), k_{mc}(T) \geq \frac{T}{2}$$

These quantities will be useful in stating the error bounds. We have

$$||V^{-1/2} \sum_{t=1}^{T} X_t^e (X_{t}^{ms})' V_{ms}^{-1/2} || \leq ||V^{-1/2} \sum_{t=1}^{k} X_t^e (X_{t}^{ms})' V_{ms}^{-1/2} || + ||V^{-1/2} \sum_{t=k+1}^{T} X_t^e (X_{t}^{ms})' V_{ms}^{-1/2} ||$$

We will need a more nuanced argument than that provided in [5] (although it will be similar in flavor).

$$P(||V^{-1/2} \sum_{t=1}^{T} X_t^e (X_{t}^{ms})' V_{ms}^{-1/2} ||) \leq \lambda_1 (\sum_{t=k+1}^{T} X_t^{ms} (X_{t}^{ms})' V_{ms}^{-1/2})$$

(103)

For any $v_1, v_2$ we break $|v_1 V^{-1/2} \sum_{t=1}^{T} X_t^e (X_{t}^{ms})' V_{ms}^{-1/2} v_2|$ into two parts

$$|v_1 V^{-1/2} \sum_{t=1}^{k} X_t^e (X_{t}^{ms})' V_{ms}^{-1/2} v_2|$$

and

$$|v_1 V^{-1/2} \sum_{t=k+1}^{T} X_t^e (X_{t}^{ms})' V_{ms}^{-1/2} v_2|$$

. For $|v_1 V^{-1/2} \sum_{t=k+1}^{T} X_t^e (X_{t}^{ms})' V_{ms}^{-1/2} v_2|$ we have

$$|v_1 V^{-1/2} \sum_{t=k+1}^{T} X_t^e (X_{t}^{ms})' V_{ms}^{-1/2} v_2| \leq \sqrt{v_1 V^{-1/2} \sum_{t=k+1}^{T} X_t^e (X_{t}^{ms})' V_{ms}^{-1/2} v_1 \sum_{t=k+1}^{T} X_t^{ms} (X_{t}^{ms})' V_{ms}^{-1/2} v_2} \leq \sigma_1 (V^{-1/2} \sum_{t=k+1}^{T} X_t^{ms} (X_{t}^{ms})' V_{ms}^{-1/2}) \leq \lambda_1 (\sum_{t=k+1}^{T} X_t^{ms} (X_{t}^{ms})' V_{ms}^{-1/2})$$

(104)

To upper bound Eq. (104) we simply need to upper bound $V^{-1/2} \sum_{t=k+1}^{T} X_t^{ms} (X_{t}^{ms})' V_{ms}^{-1/2}$. We can use dependent Hanson–Wright inequality (Corollary 8.1) and Corollary 8.2. Then from Corollary 8.2 and since $V_{ms}$ is deterministic we can conclude that with probability at least $1 - \delta$ we get

$$V^{-1/2} \sum_{t=k+1}^{T} X_t^{ms} (X_{t}^{ms})' V_{ms}^{-1/2} \leq \sum_{t=k+1}^{T} \text{tr}(V^{-1/2} \Gamma_t (A_{ms}) V^{-1/2}) (1 + \frac{1}{e} \log \frac{1}{\delta}) I$$

(105)

We can upper bound the deterministic quantity in Eq. (105) as

$$\sum_{t=k+1}^{T} \text{tr}(V^{-1/2} \Gamma_t (A) V^{-1/2}) \leq d \lambda_1 (\sum_{t=k+1}^{T} \Gamma_t (A_{ms}) V_{ms}^{-1})$$

$$\leq d \lambda_1 \left( \frac{8e}{T R^2} \sum_{t=k+1}^{T} \Gamma_t (A_{ms}) \Gamma_t \frac{1}{\|X_t^{ms}\|} (A_{ms})^{-1} \right)$$

$$\leq d \lambda_1 \left( \frac{8(T-k)}{T R^2} \Gamma_T (A_{ms}) \Gamma_T \frac{1}{\|X_t^{ms}\|} (A_{ms})^{-1} \right)$$

(106)
The key is the choice of $k$. In Proposition 7.6 we showed that $\lambda_1(\Gamma_T(\Gamma_1^{-1}))$ only depends on the ratio of $t_1/t_2$ and $A_{ms}$ and not on the specific values of $t_1, t_2$. Note that due to Proposition 7.6 the normalized gramian term $\Gamma_T(A_{ms})\Gamma_{\overline{\Gamma_1^{-1}}}(A_{ms})$ has spectral radius that is at most polynomial in $T\beta_0(\delta)$. Since $\beta_0(\delta) \approx \frac{\log T}{T} \times \log \frac{1}{\delta}$, we get that

$$\lambda_1(\Gamma_T(\Gamma_1^{-1})) = \text{poly}(\log T, \log \frac{1}{\delta})$$

Our choices of $T_{mc}(\delta), k_{mc}(T)$ in Eq. (99), (100) are motivated by the preceding discussion. We set $k = k_{mc}(T)$ and we have that $d\lambda_1\left(\frac{\sigma(T-k)}{T}\Gamma_T(A_{ms})\Gamma_{\overline{\Gamma_1^{-1}}}(A_{ms})^{-1}\right) \leq \frac{\gamma}{256}$ (check by directly substituting $k = k_{mc}(T)$ in Eq. (106) and as a result from Eq. (104)

$$|v_e^i V_{e_{1/2}} \sum_{t=k+1}^{T} X^i_t (X^m_{i})^{-1/2} v_2 | \leq \frac{\gamma}{16}$$

for arbitrary $v_1, v_2$. Similarly for the second part

$$|v_e^i V_{e_{1/2}} \sum_{t=1}^{k} X^i_t (X^m_{i})^{-1/2} v_2 | \leq \frac{\gamma}{16}$$

For the choice of $k = k_{mc}$ the other term can be simplified as

$$a_1 = \left| \sqrt{v_e^i V_{e_{1/2}} \sum_{t=1}^{k} X^i_t (X^m_{i})^{-1/2} v_2 } \leq \frac{\gamma}{16} \right|$$

By ensuring that both $T, k = k_{mc}(T)$ is $\geq T/2 \in T_u(\delta)$ (from Table I) we have from Eq. (84) that

$$\sum_{t=1}^{k} X^i_t (X^m_{i})^{-1/2} \leq \frac{3}{2} \frac{\phi_{\min}(A_e)^2}{\phi_{\min}(P_e)^2} \left(1 + \frac{1}{c} \log \frac{1}{\delta} \right) \text{tr}(P_e(\Gamma_T(A_e^{-1}) - I)) A_e A_e^T$$

$$V_e \geq \frac{\phi_{\min}(A_e)^2 \phi(A_e)^2 \delta^2}{2 \sigma_{\min}(P_e)^2} A_e A_e^T$$

Define

$$\alpha(A_e, \delta) = \frac{3}{\phi_{\min}(A_e)^2 \phi_{\max}(A_e)^2} \left(1 + \frac{c}{2} \log \frac{1}{\delta} \right) \text{tr}(P_e(\Gamma_T(A_e^{-1}) - I)) P_e$$

and we can conclude

$$\sqrt{\text{tr}(\sum_{t=1}^{k} X^i_t (X^m_{i})^{-1/2})} \leq \sqrt{\alpha(A_e, \delta) \text{tr}(A_e^{-T+k} (A_e^{-T+k})') }$$

with probability at least $1 - 2\delta$. Since $T \in T_{mc}(\delta)$ we have

$$a_1 \leq \sqrt{\alpha(A_e, \delta) \text{tr}(A_e^{-T+k} (A_e^{-T+k})')} \leq \frac{\gamma}{16}$$

(109)
with probability at least $1 - 2\delta$. Then combining Eq. (104), (105), (107), (109) we get with probability at least $1 - 4\delta$ that

$$|t'_1 V^{-1/2}_e \sum_{t=1}^T X_t^e (X_t^{mss})' V^{-1/2}_{ms} v_2| \leq \frac{\gamma}{8}$$

(110)

This implies with probability at least $1 - 4\delta$ we have

$$|| V^{-1/2}_e \sum_{t=1}^T X_t^e (X_t^{mss})' V^{-1/2}_{ms} || \leq \frac{\gamma}{8}$$

(111)

We have a similar assertion for the stable–explosive block but with $T \in T_{sc}(\delta)$ and $k = k_{sc}(T)$.

$$|| V^{-1/2}_e \sum_{t=1}^T X_t^e (X_t^{mss})' V^{-1/2}_{s} || \leq \frac{\gamma}{8}$$

(112)

It should be noted that $T \in T_{sc}(\delta)$, $T_{mc}(\delta)$ are both poly logarithmic in $\delta$ because of $A^{-T + k_{mc}}$ (or $A^{-T + k_{sc}}$) term which is exponentially decaying.

**Remark 6.** Whenever $T \in T_{sc}(\delta)$, $T_{mc}(\delta)$, the other conditions on $T$ such as $T/2 \in T_{u}(\delta)$ or $T \geq T_s(\delta) \lor T_{ms}(\frac{\delta}{2T})$ for the invertibility of the individual stable, marginally stable blocks are satisfied simultaneously (or are trivial to satisfy) and we do not state them explicitly.

### 13.2 Norm of scaled $\sum_{t=1}^T X_t^{mss}(X_t^{mss})'$ is high

Now we need to check

$$\sigma_{min} \left( \begin{bmatrix} V_{ms}^{-1/2} & 0 \\ 0 & V_{s}^{-1/2} \end{bmatrix} \sum_{t=1}^T X_t^{mss}(X_t^{mss})' \begin{bmatrix} V_{ms}^{-1/2} & 0 \\ 0 & V_{s}^{-1/2} \end{bmatrix} \right) \geq \gamma > 0$$

Since from Theorem [1] and its extension in Section [9] it is known that with probability at least $1 - \delta$ we have $\sum_{t=1}^T X_t^{mss}(X_t^{mss})' \succeq R^2 T I$ for some fixed $R = \sigma_{min}(\tilde{P}) > 0$, then we know that the Schur complement of $\sum_{t=1}^T X_t^{mss}(X_t^{mss})'$ is invertible too. For shorthand let

$$M = \sum_{t=1}^T X_t^{mss}(X_t^{mss})' = \begin{bmatrix} M_{11} & Q' \\ Q & M_{22} \end{bmatrix}$$

Then the Schur complement is

$$M/M_{11} = M_{22} - Q M_{11}^{-1} Q'$$

Since $\sigma_{min}(M) \geq R^2 T I/4$ then from Corollary 2.3 in [10] we have that

$$\sigma_{min}(M/M_{11}) \geq R^2 T I/4$$

Since $M_{22} \succeq \sum_{t=0}^{T-1} tr(\Gamma_t(A_s)) \left(1 + \frac{1}{c} \log \frac{T}{\delta} \right) I$ with probability at least $1 - \delta$. We see that with probability at least $1 - \delta$

$$M_{22}^{-1/2}(M/M_{11}) M_{22}^{-1/2} = I - M_{22}^{-1/2} Q M_{11}^{-1/2} M_{22}^{-1/2} Q' M_{22}^{-1/2} \succeq \frac{R^2}{4 \text{tr}(\Gamma_T(A_s))(1 + \frac{1}{c} \log \frac{T}{\delta})} I$$

(113)

Since $A_s$ is stable $\text{tr}(\Gamma_T(A_s)) \leq \text{tr}(\Gamma_{\infty}(A_s)) < \infty$. Define

$$\omega(\delta) = \frac{R^2}{4 \text{tr}(\Gamma_T(A_s))(1 + \frac{1}{c} \log \frac{T}{\delta})} > 0$$

(114)

Then this implies that

$$\begin{bmatrix} M_{11}^{-1/2} & 0 \\ 0 & M_{22}^{-1/2} \end{bmatrix} M \begin{bmatrix} M_{11}^{-1/2} & 0 \\ 0 & M_{22}^{-1/2} \end{bmatrix} = \begin{bmatrix} I & M_{11}^{-1/2} Q M_{22}^{-1/2} \\ M_{22}^{-1/2} Q M_{11}^{-1/2} & I \end{bmatrix} \succeq \frac{\omega(\delta)}{4} I$$
because for any $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ we have

$$
\begin{bmatrix}
I & M_{11}^{-1/2}QM_{22}^{-1/2} \\
M_{22}^{-1/2}Q'M_{11}^{-1/2} & -D \\
\end{bmatrix}
\begin{bmatrix}
v' \\
\end{bmatrix}
= v'_1 v_1 + v'_2 Dv_2 + v'_2 D'v_1 + v'_2 v_2
$$

$$
v = v'_1 v_1 - 2\sqrt{1 - \omega(\delta)||v_2||}||v_1|| + v'_2 v_2
\geq v'_1 v_1 - 2\left(1 - \frac{\omega(\delta)}{2}\right)||v_2||||v_1|| + v'_2 v_2
$$

Since from Eq. (113) it follows that $||D||^2 \leq 1 - \omega(\delta)$ we obtain

$$
v'_1 v_1 - 2\sqrt{1 - \omega(\delta)||v_2||}||v_1|| + v'_2 v_2 = v'_1 v_1 - 2\left(1 - \frac{\omega(\delta)}{2}\right)||v_2||||v_1|| + v'_2 v_2
$$

$$
= \left(1 - \frac{\omega(\delta)}{2}\right)(||v_1||-||v_2||)^2 + \left(1 - \sqrt{1 - \frac{\omega(\delta)}{2}}\right)(||v_1||^2 + ||v_2||^2)
\geq \left(\frac{\omega(\delta)}{4}\right)(||v_1||^2 + ||v_2||^2)
$$

Combining these observations we get

$$
\begin{bmatrix}
I & M_{11}^{-1/2}QM_{22}^{-1/2} \\
M_{22}^{-1/2}Q'M_{11}^{-1/2} & -D \\
\end{bmatrix}
\begin{bmatrix}
v' \\
\end{bmatrix}
\geq \left(\frac{\omega(\delta)}{4}\right)
$$

We have that

$$
\sigma_{\min}(\begin{bmatrix}
M_{11}^{-1/2} & 0 \\
0 & M_{22}^{-1/2} \\
\end{bmatrix})
\sigma_{\min}(\begin{bmatrix}
M_{11}^{-1/2} & 0 \\
0 & M_{22}^{-1/2} \\
\end{bmatrix})
\geq \left(\frac{\omega(\delta)}{4}\right)
$$

Since $M_{22} \geq V_s, M_{11} \geq V_{ms}$ we have with probability at least $1 - \delta$

$$
\sigma_{\min}(\begin{bmatrix}
V_{ms}^{-1/2} & 0 \\
0 & V_{s}^{-1/2} \\
\end{bmatrix})
\geq \left(\frac{\omega(\delta)}{4}\right)
$$

Now we replace in Eq. (113) $\gamma \rightarrow \sqrt{\frac{\omega(\delta)}{32}}$. Then that implies

$$
||V_{e}^{-1/2}\sum_{t=1}^{T}X_t^c(X_t^m)^s/V_{s}^{-1/2}|| \geq \frac{\sqrt{\omega(\delta)}}{64}
$$

$$
||V_{e}^{-1/2}\sum_{t=1}^{T}X_t^c(X_t^m)^s/V_{ms}^{-1/2}|| \geq \frac{\sqrt{\omega(\delta)}}{64}
$$

### 13.3 Lower Bound on $\sum_{t=1}^{T} \tilde{X}_t \tilde{X}_t^c$

Recalling that

$$
\sum_{t=1}^{T} B\tilde{X}_t \tilde{X}_t^c B' = \begin{bmatrix}
\sum_{t=1}^{T} V_{ms}^{-1/2}X_t^c(X_t^m)^s/V_{ms}^{-1/2} \\
\sum_{t=1}^{T} V_{ms}^{-1/2}X_t^c(X_t^m)^s/V_{ms}^{-1/2} \\
\end{bmatrix}
\begin{bmatrix}
\sum_{t=1}^{T} V_{ms}^{-1/2}X_t^c(X_t^m)^s/V_{ms}^{-1/2} \\
\sum_{t=1}^{T} V_{ms}^{-1/2}X_t^c(X_t^m)^s/V_{ms}^{-1/2} \\
\end{bmatrix}
$$

then it follows from Eq. (115) that

$$
\sum_{t=1}^{T} B\tilde{X}_t \tilde{X}_t^c B' \geq \begin{bmatrix}
\sum_{t=1}^{T} V_{ms}^{-1/2}X_t^c(X_t^m)^s/V_{ms}^{-1/2} \\
\sum_{t=1}^{T} V_{ms}^{-1/2}X_t^c(X_t^m)^s/V_{ms}^{-1/2} \\
\end{bmatrix}
\begin{bmatrix}
\sum_{t=1}^{T} V_{ms}^{-1/2}X_t^c(X_t^m)^s/V_{ms}^{-1/2} \\
\sum_{t=1}^{T} V_{ms}^{-1/2}X_t^c(X_t^m)^s/V_{ms}^{-1/2} \\
\end{bmatrix}
$$
Let \( v = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \). Then \( v' \sum_{t=1}^{T} B \tilde{X}_t \tilde{X}_t' B' v = ||v_1||^2 + \frac{\omega(\delta)}{4} (||v_2||^2 + ||v_3||^2) + 2 v'_1 \sum_{t=1}^{T} v_e^{-1/2} X_t'(X_t')' v_e^{-1/2} v_2 + 2 v'_1 \sum_{t=1}^{T} v_e^{-1/2} X_t'(X_t')' v_e^{-1/2} v_3 \geq ||v_1||^2 + \frac{\omega(\delta)}{4} (||v_2||^2 + ||v_3||^2) - \frac{\sqrt{\omega(\delta)}}{32} ||v_1|| ||v_2|| - \frac{\sqrt{\omega(\delta)}}{32} ||v_1|| ||v_3||. \) Then we get

\[
v'_1 \sum_{t=1}^{T} B \tilde{X}_t \tilde{X}_t' B' v \geq ||v_1||^2 + \frac{\omega(\delta)}{4} (||v_2||^2 + ||v_3||^2) - \frac{\omega(\delta)}{64} (||v_1||^2 + ||v_2||^2) - \frac{\omega(\delta)}{64} (||v_1||^2 + ||v_3||^2)\]

Thus \( \sigma_{\min}(\sum_{t=1}^{T} B \tilde{X}_t \tilde{X}_t' B') \geq \frac{\omega(\delta)}{8}. \) Summarizing we have with probability at least \( 1 - C \delta \). The \( C \delta \) comes because we are considering the intersection of invertibility of \( \sum_{t=1}^{T} X_t'(X_t')' \) and \( \sum_{t=1}^{T} \tilde{X}_t'(X_t')' \).

\[
\sigma_{\min}(\sum_{t=1}^{T} B \tilde{X}_t \tilde{X}_t' B') \geq \frac{\omega(\delta)}{8}
\]

whenever

\[
T \in T_{mc}(\delta) \cap T_{se}(\delta)
\]

Replacing \( \delta \to \frac{\delta}{c} \) we get with probability at least \( 1 - \delta \) that

\[
\sigma_{\min}(\sum_{t=1}^{T} B \tilde{X}_t \tilde{X}_t' B') \geq \frac{\omega(\frac{\delta}{c})}{8}
\]

Define

\[
V_{dn}(\delta) = \frac{\phi_{\min}(A_e)^2 \psi(A_e)^2 \delta^2}{2 \sigma_{\max}(P)^2} A_e^T A_e^T, V_{dn}(\delta) = \frac{TR^2}{4} I, V_{dn}(\delta) = \left( \frac{TR^2}{8} e^{\frac{\delta}{c} \Gamma_{\frac{1}{2} \text{mc}}} (A_{ms}) \right)
\]

This implies that with probability at least \( 1 - 2 \delta \) we have that

\[
\sum_{t=1}^{T} B \tilde{X}_t \tilde{X}_t' B' \geq \frac{\omega(\frac{\delta}{c})}{8} I \implies \sum_{t=1}^{T} \tilde{X}_t \tilde{X}_t' \geq \frac{\omega(\frac{\delta}{c})}{8} B^{-2}
\]

\[
\sum_{t=1}^{T} \tilde{X}_t \tilde{X}_t' \geq \frac{\omega(\frac{\delta}{c})}{8} \begin{bmatrix} V_{dn}(\delta) & 0 \\ 0 & V_{dn}(\delta) \end{bmatrix} = V_{dn}(\delta)
\]

\( V_{dn}(\delta) \) depends differently than the rest because \( V_e \) was chosen to be data dependent and we only apply the lower bound on \( \sum_{t=1}^{T} X_t'(X_t')' \) at the very end.

### 13.4 Finding the Upper Bound \( \sum_{t=1}^{T} \tilde{X}_t \tilde{X}_t' \)

For the upper bound on \( \sum_{t=1}^{T} \tilde{X}_t \tilde{X}_t' \). We use Lemma A.5 of \[15\]. Consider an arbitrary matrix \( M = \begin{bmatrix} M_1 \\ M_2 \\ M_3 \end{bmatrix} \). Then

\[
\begin{bmatrix} 3M_1 & M_1' & 0 & 0 \\ 0 & 3M_2 & M_2' & 0 \\ 0 & 0 & 3M_3 & M_3' \end{bmatrix} \geq MM'. \text{ This is because}
\]

\[
\begin{bmatrix} 2M_1 & -M_1 & -M_1' \\ -M_2 & 2M_2 & -M_2' \\ -M_3 & -M_3' & 2M_3 \end{bmatrix} = \begin{bmatrix} M_1 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} M_2 \\ 0 \\ 0 \end{bmatrix} (\begin{bmatrix} M_1 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} M_2 \\ 0 \\ 0 \end{bmatrix})' + (\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} M_2 \\ 0 \\ 0 \end{bmatrix}) (\begin{bmatrix} M_1 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} M_2 \\ 0 \\ 0 \end{bmatrix})' \]
Define
\[ V_{up}^c(\delta) = \frac{3\phi_{\max}(A)^2\sigma_{\max}(\bar{P})^4}{\sigma_{\min}(\bar{P})^2} (1 + \frac{1}{c} \log \frac{1}{\delta}) \text{tr}(\Gamma_T(A^{-1}_c)A^T_c A'^T_c) \]
\[ V_{up}(\delta) = 3\sigma_{\max}(\bar{P})^2 T \text{tr}(\Gamma_T(A_s))(1 + \frac{1}{c} \log \frac{1}{\delta}) I \]
\[ V_{up}^{ms}(\delta) = 3\sigma_{\max}(\bar{P})^2 T \text{tr}(\Gamma_T(A_{ms}))(1 + \frac{1}{c} \log \frac{1}{\delta}) I \]

Then with probability at least \( 1 - 4\delta \) we have
\[
\begin{bmatrix}
\sum_{t=1}^{T} X^c(X^c_t)' & 0 & 0 \\
0 & \sum_{t=1}^{T} X^{ms}(X^{ms}_t)' & 0 \\
0 & 0 & \sum_{t=1}^{T} X^s(X^s_t)' 
\end{bmatrix} \leq \begin{bmatrix}
V_{up}^c(\delta) & 0 & 0 \\
0 & V_{up}^{ms}(\delta) & 0 \\
0 & 0 & V_{up}(\delta)
\end{bmatrix}
\]

We get these upper bounds for stable and marginally stable matrices from Proposition (8.4) and Eq. (84) for explosive matrices. Then with probability at least \( 1 - 4\delta \) we have
\[ \sum_{t=1}^{T} \tilde{X}_t \tilde{X}_t' \leq \begin{bmatrix}
3V_{up}^c(\delta) & 0 & 0 \\
0 & 3V_{up}^{ms}(\delta) & 0 \\
0 & 0 & 3V_{up}(\delta)
\end{bmatrix} \]

Note that the time requirement in Eq. (118) is sufficient to ensure the upper bounds with high probability and we do not state them explicitly.

### 13.5 Getting Error Bounds

For shorthand call \( \omega(\frac{\delta}{T}) = \mu(\delta) \). We recall the discussion for Theorem 1. We have \( V_{up}, V_{dn} \), so we compute \( V_{up} V_{dn}^{-1} \) which gives us
\[ V_{up} V_{dn}^{-1} = \frac{8}{\mu(\delta)} \begin{bmatrix}
3V_{up}^c(\delta)(V_{dn}^c(\delta))^{-1} & 0 & 0 \\
0 & 3V_{up}^{ms}(\delta)(V_{dn}^{ms}(\delta))^{-1} & 0 \\
0 & 0 & 3V_{up}(\delta)(V_{dn}(\delta))^{-1}
\end{bmatrix} \]
\[ \det(V_{up} V_{dn}^{-1}) = \left( \frac{24}{\mu(\delta)} \right)^d \det(V_{up}^c(\delta)(V_{dn}^c(\delta))^{-1}) \det(V_{up}^{ms}(\delta)(V_{dn}^{ms}(\delta))^{-1}) \det(V_{up}(\delta)(V_{dn}(\delta))^{-1}) \]

Further \( V_{dn}^c(\tilde{\Delta}) = V_{dn}^c(\delta) \) (only the time required to be greater than this with high probability changes). Then
\[ \log(\det(V_{up} V_{dn}^{-1})) = d(\log 24 - \log \mu(\delta)) + \log \det(V_{up}^c(\delta)(V_{dn}^c(\delta))^{-1}) \]
\[ + \log \det(V_{up}^{ms}(\delta)(V_{dn}^{ms}(\delta))^{-1}) + \log \det(V_{up}(\delta)(V_{dn}(\delta))^{-1}) \]

Following this the bounds are straightforward and can be computed as shown in Eq. (118). It should be noted that Proposition 3.2 works for a general case of noise process which \( \tilde{\eta}_t \) satisfies.

Now we only know the error of the transformed dynamics, i.e.,
\[ \sum_{t=1}^{T} \left( \sum_{t=1}^{T} \tilde{X}_t \tilde{X}_t \right) + \left( \sum_{t=1}^{T} \tilde{X}_t \tilde{\eta}_t \right) \]

Since \( \sum_{t=1}^{T} \tilde{X}_t \tilde{X}_t \) is invertible with high probability
\[ \sum_{t=1}^{T} \left( \sum_{t=1}^{T} \tilde{X}_t \tilde{X}_t \right) + \sum_{t=1}^{T} \tilde{X}_t \tilde{\eta}_t = \sum_{t=1}^{T} \tilde{X}_t \tilde{X}_t \eta_t = \tilde{X} \tilde{X} \eta = \tilde{X} \tilde{X} \eta \]
\[ = \tilde{X} \tilde{X} \tilde{\eta} \]
Then it is clear that
\[
\left\| \sum_{t=1}^{T} \left( \sum_{i=1}^{T} \tilde{X}_i \tilde{X}_i^{-1} \sum_{i=1}^{T} \tilde{X}_i \tilde{\eta}_{t+1} \right) \right\| \geq \sigma_{\min}(\tilde{P}^{-1}) \left\| \sum_{t=1}^{T} \left( \sum_{i=1}^{T} X_i X_i^{-1} X_i \eta_{t+1} \right) \right\| \sigma_{\min}(\tilde{P})
\]
and we have bounded the original error term in terms of the unknown \( \sigma_{\min}(\tilde{P}), \sigma_{\min}(\tilde{P}^{-1}) \). However this factor only depends on \( d \) and not \( T \).

### 14 Independence of noise

In this section we will show how the previous techniques can be applied to noise with heavy tailed distribution. We consider the class of distributions introduced in [5] called sub–Weibull distribution. Let \( \eta_{t,i} \) be the \( i^{th} \) element of \( \eta_t \) then \( \eta_{t,i} \) has sub–Weibull distribution if
\[
P(|\eta_{t,i}| > y) \leq b \exp \left( \frac{-y^\alpha}{m} \right)
\]
When \( \alpha = 2 \) it is subGaussian, \( \alpha = 1 \) it is subExponential and \( \alpha < 1 \) it is subWeibull. Assume for now that \( \eta_{t,i} \) has symmetric distribution. The extension to asymmetric case needs some computation in finding and is not discussed here. Consider the event
\[
W(\delta) = \left\{ \max_{1 \leq i \leq T} \| \eta_t \|_{\infty} \leq \nu_T(\delta) \right\}
\]
where \( \nu_T(\delta) = \left( m \log \left( \frac{\nu_T d}{\delta} \right) \right)^{1/\alpha} \). Then Proposition 3 in [5] shows that \( P(W(\delta)) \geq 1 - \delta \). Clearly because each \( \{ \eta_{t,i} \}_{t=1, i=1}^{T, i=d} \) are i.i.d and have symmetric distribution
\[
\mathbb{E}[\eta_{t,i}]|W(\delta)] = \mathbb{E}[\eta_{t,i}]|\{\eta_{t,i} \leq \nu_T(\delta)\}] = 0
\]
(120)
Then under \( W(\delta), \eta_{t,i} \) has mean zero and \( \{ \eta_{t,i} \}_{t=1, i=1}^{T, i=d} \) are independent under the event \( W(\delta) \). Further since under \( W(\delta) \) these are bounded, they are also subGaussian. The subGaussian parameter or variance proxy \( R^2 \leq \nu_T(\delta)^2 \) which is logarithmic in \( T \). This appears as simply a scaling factor in Theorem [3] Proposition [3].

### 15 Optimality of Bound

Let \( A = a \) be 1-D system. Assume that \( T \in T_a(\delta) \) (as in Table [1]). Then \( X_t, \eta_t \) are just numbers. Then let \( E \) be the error, i.e.,
\[
E = \left( \sum_{t=1}^{T} x_t^2 \right)^{-1} \left( \sum_{t=1}^{T} x_t \eta_{t+1} \right)
\]
\[
= a^{-T} \left( \sum_{t=1}^{T} a^{-2T} x_t^2 \right)^{-1} \left( \sum_{t=1}^{T} a^{-T} x_t \eta_{t+1} \right)
\]
In this section, we will show that the bound obtained for explosive systems is optimal in terms of \( \delta \). Assume \( \eta_t \sim \mathcal{N}(0, 1) \) i.i.d Gaussian. Let \( S_T = \sum_{t=1}^{T} a^{-T} x_t \eta_{t+1}, U_T = \sum_{t=1}^{T} a^{-2T} x_t^2 \). Now \( E = a^{-T} U_T^{-1} S_T \). Consider
\[
P \left( S_T U_T^{-1} \geq \frac{-a^{-T}}{C \delta \ln(\delta)} \right) \geq P \left( \sum_{t=1}^{T} x_t \eta_{t+1} + \frac{a^{-2T} x_T \eta_{T+1}}{\sum_{t=1}^{T} a^{-2T} x_t^2} \geq \frac{-a^{-T}}{C \delta \ln(\delta)} \right)
\]
\[
\geq \mathbb{P} \left( \sum_{t=1}^{T} x_t \eta_{t+1} \geq 0, \frac{a^{-T} x_T \eta_{T+1}}{\sum_{t=1}^{T} a^{-2T} x_t^2} \geq \frac{-1}{C \delta \ln(\delta)} \right)
\]
by symmetry
\[
\geq \frac{1}{2} \mathbb{P} \left( \frac{a^{-T} x_T \eta_{T+1}}{\sum_{t=1}^{T} a^{-2T} x_t^2} \geq \frac{-1}{C \delta \ln(\delta)} \right)
\]
Define \( F_T = \sum_{t=1}^{T} a^{-2t+2} (a^{-2T} x_t^2) = \frac{1-a^{-2T}}{1-a} a^{-2T} x_T^2 \), and \( \sigma^2 = \text{Var}(a^{-2T} x_T^2) \). It is clear that \( a^{-T} x_T \) is a Gaussian random variable. Note that \( F_T, U_T \) are the same as Eq. (18) and Section [1] when \( A = a \). We can easily calculate \( \sigma^2 \)
\[
a^{-2} \leq \sigma^2 \leq \frac{1}{a^2 - 1}
\]
Consider four events

\[\mathcal{E}_1(\delta) = \left\{ |U_T - F_T| \leq \frac{\delta^2 \sigma^2}{C} \right\}, \mathcal{E}_2(\delta) = \left\{ \left| a^{-T} x_T \eta_{T+1} \right| \geq \frac{-\delta}{\ln(\delta)} \right\}\]

\[\mathcal{E}_3(\delta) = \left\{ C_1 \delta^2 \sigma^2 \leq F_T \leq C_2 \delta^2 \sigma^2 \right\}, \mathcal{E}_4(\delta) = \left\{ (C_1 - 1/C) \delta^2 \sigma^2 \leq U_T \leq (C_2 + 1/C) \delta^2 \sigma^2 \right\}\]

Clearly \(\mathcal{E}_1(\delta) \cap \mathcal{E}_3(\delta) \implies \mathcal{E}_1(\delta) \cap \mathcal{E}_4(\delta)\). Assume \(\delta^2 \in (0, \frac{1}{128}]\) then

\[\mathbb{P}(\mathcal{E}_3(\delta)) = \frac{2}{\sqrt{2\pi}\sigma} \int_{2\delta}^{16\delta} e^{-\frac{x^2}{2\sigma^2}} dx \geq \frac{14\delta}{\sqrt{2\pi}} e^{-\frac{2\delta^2}{2\sigma^2}} \geq \frac{14\delta}{\sqrt{2\pi}} \geq 2\delta\]

For \(T \in T_u(\delta)\) (Note \(T_u(\delta)\) is the set of \(T\) that satisfies Eq. (85) when \(A = a\)) and from Eq. (81), we have with probability at least \(1 - \frac{\delta}{2}\) that

\[\|U_T - F_T\|_2 \leq \left( 4T^2 \sigma_1^2 (A^{-T+1}) \text{tr}(\Gamma_T(A^{-1})) + \frac{T \text{tr}(A^{-T-1} \Gamma_T(A^{-1}) A^{-T+1})}{\delta} \right) \leq \frac{2\sigma_{\max}(P)^2}{\phi_{\min}(A)^2 \psi(a)^2} \leq \frac{C\delta^2}{(a^2 - 1)}\]

The last equality follows because for 1-D systems \(\phi_{\min}(A), \psi(A), \sigma_{\max}(P)\) are just constants, for example \(P = 1, \phi_{\min}(a) = 1, \psi(a) = C \sigma^2 \leq \frac{C}{\sigma^2 - 1}\) which follows by definition. Thus, \(\mathbb{P}(\mathcal{E}_1(\delta)) \geq 1 - \frac{\delta}{2}\). We bound \(\mathbb{P}(\mathcal{E}_2(\delta))\) in Section 18 which gives \(\mathbb{P}(\mathcal{E}_2(\delta)) \geq 1 - \frac{\delta}{2}\) and then

\[\mathbb{P}(\mathcal{E}_1(\delta) \cap \mathcal{E}_2(\delta) \cap \mathcal{E}_3(\delta)) \geq \mathbb{P}(\mathcal{E}_1(\delta) \cap \mathcal{E}_2(\delta) \cap \mathcal{E}_3(\delta)) \geq \mathbb{P}(\mathcal{E}_1(\delta)) + \mathbb{P}(\mathcal{E}_2(\delta) \cap \mathcal{E}_3(\delta)) - 1 \geq \mathbb{P}(\mathcal{E}_1(\delta)) + \mathbb{P}(\mathcal{E}_2(\delta)) + \mathbb{P}(\mathcal{E}_3(\delta)) - 2 \geq \frac{\delta}{2}\]

Since

\[\mathbb{P}(\mathcal{E}_1(\delta) \cap \mathcal{E}_2(\delta) \cap \mathcal{E}_4(\delta)) \leq 2\mathbb{P}\left( \|S_T |U_T^{-1}| \geq \frac{-C}{\sigma^2 \delta \ln(\delta)} \right)\]

we have proved our claim that with probability at least \(\delta\) we have that

\[|E_T| \geq \frac{-Ca^{-T}}{\sigma^2 \delta \ln(\delta)}\]

16 Lemma A

We need to bound the following difference

\[\left| a^{-2T} \sum_{t=1}^{T} (X_t^{(1)})(X_t^{(2)}) - \sum_{t=1}^{T} a^{-2t+2} Z_t^{(1)} Z_t^{(2)} \right|\]
Recall $a^{-2T}(X^{(1)}_t) = a^{-2T} + 2T(Z^{(1)}_t)$ and the difference can be rewritten as

$$
\begin{align*}
& a^{-2T} \sum_{t=1}^{T} (X^{(1)}_t)(X^{(2)}_t) - \sum_{t=1}^{T} a^{-2T} + 2T(Z^{(1)}_t)Z^{(2)}_t = \sum_{t=1}^{T} a^{-t}(Z^{(1)}_{t-t} - Z^{(1)}_t)a^{-t}(Z^{(2)}_t) \\
& + \sum_{t=1}^{T} a^{-t}(Z^{(2)}_{t-t} - Z^{(2)}_t)a^{-t}(Z^{(1)}_t) \\
& = \sum_{t=1}^{T} a^{-2T}(Z^{(1)}_{t-t} - Z^{(1)}_t)(Z^{(2)}_t) + \sum_{t=1}^{T} a^{-2T}(Z^{(2)}_{t-t} - Z^{(2)}_t)(Z^{(1)}_t) \\
& = \sum_{t=1}^{T} a^{-2T}a^{-T+t-1}Z^{(1)}(T,t)(Z^{(2)}_t) + \sum_{t=1}^{T} a^{-2T}a^{-T+t-1}Z^{(2)}(T,t)(Z^{(1)}_t) \\
& = a^{-T} \sum_{t=1}^{T} a^{-t-1}Z^{(1)}(T,t)(Z^{(2)}_t) + a^{-T} \sum_{t=1}^{T} a^{-t-1}a^{-T+t-1}Z^{(2)}(T,t)(Z^{(1)}_t)
\end{align*}
$$

Here $Z^{(i)}(T,t) = \sum_{i=k}^{t} a^{-k+1}\eta^{(i)}_{t-k}$. We will use a very trivial bound here. Since $Z^{(1)}(T,t)$ is Gaussian for every $t$ we can just bound the maximum of $T$ Gaussians to upper bound. The variance of $Z^{(1)}(T,t)$ is $\leq \frac{1}{1-a^2} = \sigma^2$ for every $t$. Consider \( \{Z^{(1)}(T,t)\}_{t=1}^{T} \). Then a standard estimate on the maximum tells us that with probability at least $1 - \delta$ we have

$$
\sup_{1 \leq t \leq T} |Z^{(1)}(T,t)| \leq \sigma \sqrt{2T\log \frac{2T}{\delta}}.
$$

Similarly we do the same for the process \( \{Z^{(2)}(T,t)\}_{t=1}^{T} \) and we get with probability at least $1 - 3\delta$

$$
\begin{align*}
a^{-T} \sum_{t=1}^{T} a^{-t-1}Z^{(1)}(T,t)(Z^{(2)}_t) + a^{-T} \sum_{t=1}^{T} a^{-t-1}a^{-T+t-1}Z^{(2)}(T,t)(Z^{(1)}_t) & \leq 2a^{-T}T2\sigma^2 \log \frac{2T}{\delta} \\
\end{align*}
$$

### 17 Lemma B

Let the characteristic and minimal polynomial be $\chi(t), \mu(t)$ respectively.

$$
\chi(t) = \prod_{i=1}^{k}(t - \lambda_i)^{a_i}, \mu(t) = \prod_{i=1}^{k}(t - \lambda_i)^{b_i}
$$

where $b_i \leq a_i, b_i$ is the size of the largest Jordan block corresponding to $\lambda_i$ in the Jordan normal form. $a_i$ sum of size of all Jordan blocks corresponding to $\lambda_i$. Now, if $\chi(t) = \mu(t)$ then $a_i = b_i$, i.e., there is only Jordan block corresponding to each $\lambda_i$. On the other if there is only one Jordan block (geometric multiplicity = 1) corresponding to each eigenvalue $\Rightarrow a_i = b_i$ and $\chi(t) = \mu(t)$.

### 18 Distribution of $x_T^T \eta_{T+1}$

We only need to check the distribution of $a^{-T}x_T^T \eta_{T+1}$ where $a^{-T}x_T$ is independent of $\eta_{T+1}$. Now $a^{-T}x_T$ is Gaussian with variance $\leq \frac{1}{1-a^2}$ but $\geq 1$. Now

$$
\begin{align*}
\mathbb{P}(|a^{-T}x_T^T \eta_{T+1}| \leq \delta) &= \int_{\delta}^{\infty} \int_{-\delta/y}^{\delta/y} Ce^{-\frac{y^2}{2}}e^{-\frac{\delta^2}{2}} dx dy + \int_{0}^{\delta} \int_{-\delta/y}^{\delta/y} Ce^{-\frac{y^2}{2}}e^{-\frac{\delta^2}{2}} dx dy \\
& \leq C\delta + \int_{\delta}^{\infty} \int_{-\delta/y}^{\delta/y} Ce^{-\frac{y^2}{2}}e^{-\frac{\delta^2}{2}} dx dy \leq C\delta + C\delta \int_{\delta}^{\infty} e^{-\frac{y^2}{2}} dy \\
& \leq C\delta + C\delta \int_{\delta}^{1} e^{-\frac{y^2}{2}} dy + C\delta \int_{1}^{\infty} e^{-\frac{y^2}{2}} dy \\
& \leq C\delta - C\delta \ln(\delta)
\end{align*}
$$

Let $x = \frac{-C\delta}{\ln(\delta)}$, then $\mathbb{P}(|a^{-T}x_T^T \eta_{T+1}| \leq x) \leq \frac{\delta}{2}$. 

---

**How fast can linear dynamical systems be learned?**