Near optimal finite time identification of arbitrary linear dynamical systems

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

Abstract

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

1 Introduction

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

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

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

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

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

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

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

2 Contributions

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

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

3 Notation and Definitions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

One can show that

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

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

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

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

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

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

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

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

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

- The perfectly stable matrix class, $\mathcal{S}_0$
$$\rho_i(A) \leq 1 - \frac{C}{T}$$
for $1 \leq i \leq d$.
- The marginally stable matrix, $\mathcal{S}_1$
$$1 - \frac{C}{T} < \rho_i(A) \leq 1 + \frac{C}{T}$$
for $1 \leq i \leq d$.
- The regular and explosive matrix, $\mathcal{S}_2$
$$\rho_i > 1 + \frac{C}{T}$$
for $1 \leq i \leq d$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table 1. Definitions of key quantities in the paper

4 Main Results

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

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

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

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

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

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

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

whenever

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

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

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

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

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

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

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

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

Proof. We start by writing an upper bound

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned} Y_T &= x_0x_0^{'} + AY_{T-1}A^{'} + \\ &+ \sum_{t=0}^{T-1}(Ax_t\eta_{t+1}^{'} + \eta_{t+1}x_t^{'}A^{'}) + \sum_{t=1}^{T}\eta_t\eta_t^{'} \\ &\succeq AY_{T-1}A^{'} + \\ &+ \sum_{t=0}^{T-1}(Ax_t\eta_{t+1}^{'} + \eta_{t+1}x_t^{'}A^{'}) + \sum_{t=1}^{T}\eta_t\eta_t^{'}. \end{aligned} \tag{13}$$

Since $\{\eta_t\}_{t=1}^T$ are i.i.d. subgaussian we can show that $\sum_{t=1}^T \eta_t\eta_t^{\prime}$ concentrates near $TI_{d\times d}$ with high probability. Using Proposition 3.1 once again, we will show that with high probability

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

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

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

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

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

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

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

and then Eq. (16) can be reduced to

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

We show that

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

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

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

$$Y_T \preceq \text{tr}\left(\sum_{t=1}^{T}X_tX_t^{\prime}\right)I.$$

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

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

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

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

The steps of the proof proceed as follows. Define

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

We show that

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$X_t = \sum_{\tau=1}^{t} A^{\tau-1}\eta_{t-\tau+1}$$

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

then the transformed dynamics are as follows:

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

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

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

Then we have

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

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

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

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

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

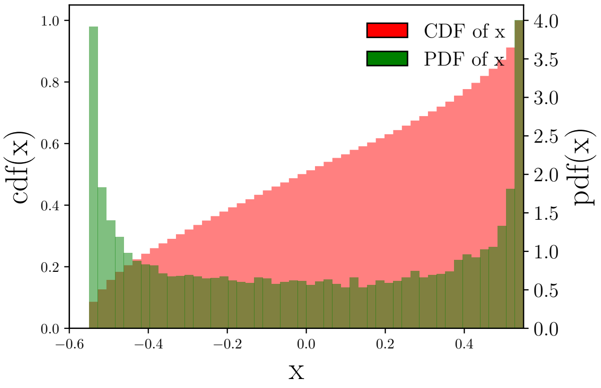

Figure 1. CDF and PDF of $\hat{\beta}_o$

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

The details are in appendix as Section 14. □

5 Inconsistency of OLS

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

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

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

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

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

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

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

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

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

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

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

The proof is given in Section 20 and Proposition 20.2. Proposition 5.1 suggests that the consistency of OLS estimate depends directly on the condition number of the sample covariance matrix. In fact, OLS is inconsistent when condition number grows exponentially fast in T (as in the case of A_o). The proof requires a careful expansion of the (appropriately scaled) sample covariance matrix inverse using Woodbury's identity. Since the sample covariance matrix is highly ill–conditioned, it magnifies the noise-covariate cross terms so that the identification error no longer decays as time increases. Although for stable and marginally stable A this invertibility can be characterized $\sigma_{\min}(\Gamma_T(A))$ such an intuition does not extend to explosive systems. This is because the behavior of Y_T is dominated by "past" η_ts such as η_1, η_2 much more than the η_{T-1}, η_T etc. When A is explosive, all singular values of $||A^T||$ grow exponentially fast. Since $X_T = A^{T-1}\eta_1 + A^{T-2}\eta_2 + \ldots + A\eta_{T-1} + \eta_T$ the behavior of X_T is dominated by $A^{T-1}\eta_1$. This causes a very strong dependence between X_T and X_{T+1} and some structural constraints (such as regularity) are necessary for OLS identification.

6 Discussion

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

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

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

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

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

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

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

7 Road Map of Results

We sketch the road map of our results here. Critical to the results our finding matrices V_{up}, V_{dn} that satisfy Eq. (9), (10).

- For the case when $\rho_i(A) \leq 1 + C/T$, we find these matrices in Section 10. We show that $V_{up}, V_{dn} = \Theta(T)I$.
- The bound for V_{dn} can be sharpened to $V_{dn} = \Omega(T^2)I$ when all the eigenvalues of A lie in $(1 - c/T, 1 + c/T)$. This result is proven as part of Section 11.
- Section 12 (specifically Proposition 12.2) discusses the proof technique for finding V_{up}, V_{dn} for explosive systems. Non-trivial bounds on the matrix rely critically on the regularity of the explosive matrix.
- When the regularity condition is violated, we show via a simple construction of a scaled identity matrix that OLS is inconsistent in Section 20. This involves explicitly showing that the error is a random variable which has a non-zero norm even when $T \to \infty$. These are Propositions 20.1, 20.2.
- We then combine the separate cases of stable, marginally stable and explosive matrices to show that even with an arbitrary distribution of eigen values (albeit regular), OLS is consistent. Furthermore, the rate of convergence is limited by the slowest component. The proof requires a careful transformation of the matrix into blocks of stable, marginally stable and explosive and showing that the cross terms zero out. This is proven in Section 14.
- Other minor extensions of our results can be found in Section 15 (when there is an additional control input) and Section 16 (when the noise is heavy-tailed).

8 Matrix Inequalities

In this section we present some probabilistic and matrix inequalities that will be used in our main results.

Proposition 8.1. *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

$$||\bar{P}^{-1/2}Q||_2^2 \leq \gamma^2$$

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

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

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 8.2. *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.,*

$$J_d(\lambda)^{-k} = \begin{bmatrix} a_1 & a_2 & a_3 & \dots & a_d \\ 0 & a_1 & a_2 & \dots & a_{d-1} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & a_1 & a_2 \\ 0 & 0 & \dots & 0 & a_1 \end{bmatrix}_{d \times d} \tag{26}$$

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

Proof. $J_d(\lambda) = (\lambda I + N)$ where N is the matrix with all ones on the 1^{st} (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

$$(\lambda I + N)^{-1} = (\sum_{l=0}^{d-1} (-1)^l \lambda^{-l-1} N^l)$$
$$(-1)^{k-1}(k-1)!\,(\lambda I + N)^{-k} = \Big(\sum_{l=0}^{d-1} (-1)^l \frac{d^{k-1}\lambda^{-l-1}}{d\lambda^{k-1}} N^l\Big) = \Big(\sum_{l=0}^{d-1} (-1)^l c_{l,k} N^l\Big)$$

and the proof follows in a straightforward fashion. □

Proposition 8.3. *Let A be a regular matrix and $A = P^{-1}\Lambda P$ be its Jordan decomposition. Then*

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

Further $\phi_{\min}(A) > 0$ where $\phi_{\min}(\cdot)$ is defined in Definition 3.

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 19 in appendix), *i.e.*,

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

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

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

Since Λ is Jordan matrix of the Jordan decomposition, it is of the following form

$$\Lambda = \begin{bmatrix} J_{k_1}(\lambda_1) & 0 & \ldots & 0 & 0 \\ 0 & J_{k_2}(\lambda_2) & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & J_{k_l}(\lambda_l) & 0 \\ 0 & 0 & \ldots & 0 & J_{k_{l+1}}(\lambda_{l+1}) \end{bmatrix} \tag{27}$$

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

$$\Lambda^{-k} = \begin{bmatrix} J^{-k}_{k_1}(\lambda_1) & 0 & \ldots & 0 & 0 \\ 0 & J^{-k}_{k_2}(\lambda_2) & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & J^{-k}_{k_l}(\lambda_l) & 0 \\ 0 & 0 & \ldots & 0 & J^{-k}_{k_{l+1}}(\lambda_{l+1}) \end{bmatrix} \tag{28}$$

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

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

By Proposition 8.2 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

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

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 $||\sum_{i=1}^{d} a_i \Lambda^{-i+1} v||_2 > 0$, *i.e.*,

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

This implies

$$\inf_{||a||_2 = 1} ||[v, \Lambda^{-1} v, \ldots, \Lambda^{-d+1} v] a||_2 > 0$$
$$\sigma_{\min}([v, \Lambda^{-1} v, \ldots, \Lambda^{-d+1} v]) > 0$$

By Definition 3 we have

$$\phi_{\min}(A) > 0$$

□

Proposition 8.4 (Corollary 2.2 in (Ipsen & Lee, 2011))**.** *For any positive definite matrix M with diagonal entries m_{jj}, $1 \leq j \leq d$ and ρ is the spectral radius of the matrix C with elements*

$$\begin{aligned} c_{ij} &= 0 \quad \textit{if } i = j \\ &= \frac{m_{ij}}{\sqrt{m_{ii} m_{jj}}} \quad \textit{if } i \neq j \end{aligned}$$

then

$$0 < \frac{\prod_{j=1}^{d} m_{jj} - det(M)}{\prod_{j=1}^{d} m_{jj}} \leq 1 - e^{-\frac{d\rho^2}{1+\lambda_{\min}}}$$

where $\lambda_{\min} = \min_{1 \leq j \leq d} \lambda_j(C)$.

Proposition 8.5. *Let $1 - C/T \leq \rho_i(A) \leq 1 + C/T$ 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 Λ is the Jordan matrix. Since Λ can be complex we will assume that adjoint instead of transpose. This gives

$$\begin{aligned} \Gamma_T(A) &= I + \sum_{t=1}^{T} A^t (A^t)^{\prime} \\ &= I + P^{-1} \sum_{t=1}^{T} \Lambda^t P P^{\prime} (\Lambda^t)^* P^{-1\prime} \succeq I + \sigma_{\min}(P)^2 P^{-1} \sum_{t=1}^{T} \Lambda^t (\Lambda^t)^* P^{-1\prime} \end{aligned}$$

Then this implies that

$$\begin{aligned}\sigma_{\min}(\Gamma_T(A)) &\geq 1 + \sigma_{\min}(P)^2 \sigma_{\min}(P^{-1} \sum_{t=1}^{T} \Lambda^t (\Lambda^t)' P^{-1\prime}) \geq 1 + \sigma_{\min}(P)^2 \sigma_{\min}(P^{-1})^2 \sigma_{\min}(\sum_{t=1}^{T} \Lambda^t (\Lambda^t)') \\ &\geq 1 + \frac{\sigma_{\min}(P)^2}{\sigma_{\max}(P)^2} \sigma_{\min}(\sum_{t=1}^{T} \Lambda^t (\Lambda^t)')\end{aligned}$$

Now

$$\sum_{t=0}^{T} \Lambda^t (\Lambda^t)^* = \begin{bmatrix} \sum_{t=0}^{T} J_{k_1}^t(\lambda_1)(J_{k_1}^t(\lambda_1))^* & 0 & \ldots & 0 \\ 0 & \sum_{t=1}^{T} J_{k_2}^t(\lambda_2)(J_{k_2}^t(\lambda_2))^* & 0 & \ldots \\ \vdots & \vdots & \ddots & \ddots \\ 0 & \ldots & 0 & \sum_{t=1}^{T} J_{k_l}^t(\lambda_l)(J_{k_l}^t(\lambda_l))^* \end{bmatrix}$$

Since Λ 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 $J_{k_1}(\lambda_1)$, *i.e.*,

$$\sigma_{\min}(\sum_{t=0}^{T} \Lambda^t (\Lambda^t)^*) = \sigma_{\min}(\sum_{t=0}^{T} J_{k_1}^t(\lambda_1)(J_{k_1}^t(\lambda_1))^*) \tag{29}$$

Define $B = \sum_{t=0}^{T} J_{k_1}^t(\lambda_1)(J_{k_1}^t(\lambda_1))^*$. Note that $J_{k_1}(\lambda_1) = (\lambda_1 I + N)$ where N is the nilpotent matrix that is all ones on the first off–diagonal and $N^{k_1} = 0$. Then

$$\begin{aligned}(\lambda_1 I + N)^t &= \sum_{j=0}^{t} \binom{t}{j} \lambda_1^{t-j} N^j \\ (\lambda_1 I + N)^t ((\lambda_1 I + N)^t)^* &= \Big(\sum_{j=0}^{t} \binom{t}{j} \lambda_1^{t-j} N^j\Big)\Big(\sum_{j=0}^{t} \binom{t}{j} (\lambda_1^*)^{t-j} N^{j\prime}\Big) \\ &= \sum_{j=0}^{t} \binom{t}{j}^2 |\lambda_1|^{2(t-j)} \underbrace{N^j (N^j)'}_{\text{Diagonal terms}} + \sum_{j \neq k}^{j=t,k=t} \binom{t}{k}\binom{t}{j} \lambda_1^j (\lambda_1^*)^k N^j (N^k)' \\ &= \sum_{j=0}^{t} \binom{t}{j}^2 |\lambda_1|^{2(t-j)} \underbrace{N^j (N^j)'}_{\text{Diagonal terms}} + \sum_{j > k}^{j=t,k=t} \binom{t}{k}\binom{t}{j} \lambda_1^j (\lambda_1^*)^k N^j (N^k)' \\ &+ \sum_{j < k}^{j=t,k=t} \binom{t}{k}\binom{t}{j} \lambda_1^j (\lambda_1^*)^k N^j (N^k)' \\ &= \sum_{j=0}^{t} \binom{t}{j}^2 |\lambda_1|^{2(t-j)} \underbrace{N^j (N^j)'}_{\text{Diagonal terms}} + \sum_{j > k}^{j=t,k=t} \binom{t}{k}\binom{t}{j} \underbrace{|\lambda_1|^{2k} \lambda_1^{j-k} N^{j-k} N^k (N^k)'}_{\text{On } (j-k) \text{ upper off–diagonal}} \\ &+ \sum_{j < k}^{j=t,k=t} \binom{t}{k}\binom{t}{j} \underbrace{|\lambda_1|^{2j} (\lambda_1^*)^{k-j} N^j (N^j)' (N^{j-k})'}_{\text{On } (k-j) \text{ lower off–diagonal}}\end{aligned}$$

Let $\lambda_1 = re^{i\theta}$, then similar to (Erxiong, 1994), there is $D = \text{Diag}(1, e^{-i\theta}, e^{-2i\theta}, \ldots, e^{-i(k_1-1)\theta})$ 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 $e^{-ij\theta+ik\theta}r_0e^{i(j-k)\theta} = r_0$, which is real. Then we have

$$D(\lambda_1 I+N)^t((\lambda_1 I+N)^t)^*D^* = \sum_{j=0}^{t}\binom{t}{j}^2|\lambda_1|^{2(t-j)}\underbrace{N^j(N^j)'}_{\text{Diagonal terms}} + \sum_{j>k}^{j=t,k=t}\binom{t}{k}\binom{t}{j}\underbrace{|\lambda_1|^{2k}|\lambda_1|^{j-k}N^{j-k}N^k(N^k)'}_{\text{On } (j\,-\,k) \text{ upper off–diagonal}}$$
$$+ \sum_{j<k}^{j=t,k=t}\binom{t}{k}\binom{t}{j}\underbrace{|\lambda_1|^{2j}|\lambda_1|^{k-j}N^j(N^j)'(N^{k-j})'}_{\text{On } (k\,-\,j) \text{ lower off–diagonal}} \tag{30}$$

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 $B = \sum_{t=1}^{T}(|\lambda_1|I+N)^t((|\lambda_1|I+N)^t)'$

$$B_{ll} = \sum_{t=1}^{T}[(\lambda_1 I+N)^t((\lambda_1 I+N)^t)^*]_{ll} \tag{31}$$
$$= \sum_{t=1}^{T}\sum_{j=0}^{k_1-l}\binom{t}{j}^2|\lambda_1|^{2(t-j)} \tag{32}$$

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

$$e^{-C} \leq |\lambda_1|^t \leq e^C$$

Then

$$B_{ll} = \sum_{t=1}^{T}\sum_{j=0}^{k_1-l}\binom{t}{j}^2|\lambda_1|^{2(t-j)} \geq e^{-2C}\sum_{t=1}^{T}\sum_{j=0}^{k_1-l}\binom{t}{j}^2$$
$$\geq e^{-2C}\sum_{t=T/2}^{T}\sum_{j=0}^{k_1-l}\binom{t}{j}^2 \geq e^{-2C}\sum_{t=T/2}^{T}c_{k_1}\frac{t^{2k_1-2l+2}-1}{t^2-1} \geq C(k_1)T^{2k_1-2l+1} \tag{33}$$

An upper bound can be achieved in an equivalent fashion.

$$B_{ll} = \sum_{t=1}^{T}\sum_{j=0}^{k_1-l}\binom{t}{j}^2|\lambda_1|^{2(t-j)} \leq e^{2C}T\sum_{j=0}^{k_1-l}T^{2j} \leq C(k_1)T^{2k_1-2l+1} \tag{34}$$

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

$$B_{k,k+l} = \sum_{t=1}^{T}\sum_{j=0}^{k_1-k-l}\binom{t}{j}\binom{t}{j+l}|\lambda_1|^{2j}|\lambda_1|^{l} \geq \sum_{t=1}^{T}e^{-2C}\sum_{t=T/2}^{T}\sum_{j=0}^{k_1-k-l}\binom{t}{j}\binom{t}{j+l} \tag{35}$$
$$\geq e^{-2C}\frac{T}{2}\sum_{j=0}^{k_1-k-l}\binom{T/2}{j}\binom{T/2}{j+l} \geq C(k_1)T^{2k_1-2k-l+1} \tag{36}$$

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

$$B_{jk} = C(k_1)T^{2k_1-j-k+1} \tag{37}$$

whenever $T \geq 8d$. Recall Proposition 8.4, let the M there be equal to B then since

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

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 $\prod_{j=1}^{k_1} B_{jj} \geq \det(B) \geq \prod_{j=1}^{k_1} B_{jj} e^{-\frac{d\rho^2}{1+\lambda_{\min}}} = C(k_1) \prod_{j=1}^{k_1} B_{jj}$. 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

$$\det(B) = C(k_1) \prod_{j=1}^{k_1} B_{jj} = \prod_{j=1}^{k_1} C(k_1) T^{2k_1 - 2l + 1} = C(k_1) T^{k_1^2} \tag{38}$$

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. (37) it implies that

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

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 $C_2(k_1) T^3 \geq 4k_1^2 C_1(k_1) T$. Again to upper bound $\sigma_{k_1-2}(A)$ we will use a similar argument

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

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 $C_2(k_1) T^3 \geq 4k_1^2 C_1(k_1) T$ and $T \geq T_2 = 2k_1 \sqrt{\frac{C_2(k_1)}{C_3(k_1)}}$ ensures $C_3(k_1) T^5 \geq 4k_1^2 C_2(k_1) T^3$. 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 $\sigma_i(A) \leq (1 + k_1^{-1}) C_{k_1 - i + 1}(k_1) T^{2k_1 - 2i + 1}$, 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. (38) where $\det(B) = C(k_1) T^{k_1^2}$. Assume that $\sigma_{k_1}(B) < \frac{C(k_1) T}{e \prod_{i=1}^{d} C_{i+1}(k_1)}$. Then whenever $T \geq \max\left(8d, \sup_i 2k_1 \sqrt{\frac{C_i(k_1)}{C_{i+1}(k_1)}}\right)$

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

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

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

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 (Simchowitz et al., 2018)]. We are not interested in finding the best bound $\alpha(d)$ rather show that the bound of Proposition 8.5 exists and assume that such a bound is known.

Proposition 8.6. *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 β of degree at most d^2 whenever $t_i \geq 8d$.

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

$$\begin{aligned}\lambda_1(\Gamma_{t_1}(A)\Gamma_{t_2}^{-1}(A)) &\leq \text{tr}(\Gamma_{t_1}(A)\Gamma_{t_2}^{-1}(A)) \leq \text{tr}(\Gamma_{t_2}^{-1/2}(A)\Gamma_{t_1}(A)\Gamma_{t_2}^{-1/2}(A)) \\ &\leq d\sigma_1(\Gamma_{t_2}^{-1/2}(A)\Gamma_{t_1}(A)\Gamma_{t_2}^{-1/2}(A)) \leq d \sup_{||x|| \neq 0} \frac{x'\Gamma_{t_1}(A)x}{x'\Gamma_{t_2}(A)x}\end{aligned}$$

Now

$$\Gamma_{t_i}(A) = P^{-1}\sum_{t=0}^{t_i} \Lambda^t PP'(\Lambda^t)^* P^{-1\prime} \preceq \sigma_{\max}(P)^2 P^{-1}\sum_{t=0}^{t_i} \Lambda^t(\Lambda^t)^* P^{-1\prime}$$

$$\Gamma_{t_i}(A) \succeq \sigma_{\min}(P)^2 P^{-1}\sum_{t=0}^{t_i} \Lambda^t(\Lambda^t)^* P^{-1\prime}$$

Then this implies

$$\sup_{||x|| \neq 0} \frac{x'\Gamma_{t_1}(A)x}{x'\Gamma_{t_2}(A)x} \leq \frac{\sigma_{\max}(P)^2}{\sigma_{\min}(P)^2} \sup_{||x|| \neq 0} \frac{x'\sum_{t=0}^{t_1}\Lambda^t(\Lambda^t)^* x}{x'\sum_{t=0}^{t_2}\Lambda^t(\Lambda^t)^* x}$$

Then from Lemma 12 in (Abbasi-Yadkori et al., 2011) we get that

$$\sup_{||x|| \neq 0} \frac{x'\sum_{t=0}^{t_1}\Lambda^t(\Lambda^t)^* x}{x'\sum_{t=0}^{t_2}\Lambda^t(\Lambda^t)^* x} \leq \frac{\det(\sum_{t=0}^{t_1}\Lambda^t(\Lambda^t)^*)}{\det(\sum_{t=0}^{t_2}\Lambda^t(\Lambda^t)^*)}$$

Then

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

Here l are the number of Jordan blocks of A. Then our assertion follows from Eq. (38) which implies that the determinant of $\sum_{t=0}^{t_2} J_{k_i}(\lambda_i)^t(J_{k_i}(\lambda_i)^t)^*$ 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

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

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

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

□

9 Probabilistic Inequailities

Proposition 9.1 ((Vershynin, 2010))**.** *Let M be a random matrix. Then we have for any $\epsilon < 1$ and any $w \in \mathcal{S}^{d-1}$ 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 (Vershynin, 2010).

Proposition 9.1 helps us in using the tools developed in de la Pena et. al. and (Abbasi-Yadkori et al., 2011) for self–normalized martingales. We will define $\tilde{S}_t = \sum_{\tau=0}^{t-1} X_\tau \tilde{\eta}_{\tau+1}$ where $\tilde{\eta}_t = w^T \eta_t$ is standard normal when w is a unit vector. Specifically, we use Lemma 9 of (Abbasi-Yadkori et al., 2011) which we state here for convenience:

Theorem 3 (Theorem 1 in (Abbasi-Yadkori et al., 2011))**.** *Let $\{\boldsymbol{\mathcal{F}}_t\}_{t=0}^\infty$ be a filtration. Let $\{\eta_t\}_{t=1}^\infty$ be a real valued stochastic process such that η_t is $\boldsymbol{\mathcal{F}}_t$ measurable and η_t is conditionally R-sub-Gaussian for some $R > 0$., i.e.,*

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

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

$$\bar{V}_t = V + \sum_{s=1}^t X_s X_s' \;\; S_t = \sum_{s=1}^t \eta_{s+1} X_s$$

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

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

Proposition 9.2. *Let P have full row rank and*

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

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

$$||(\bar{Y}_{T-1})^{-1/2} \sum_{t=0}^{T-1} X_t \eta_{t+1}' P'||_2 \leq R \sqrt{8d \log\left(\frac{5det(\bar{Y}_{T-1})^{1/2d} det(V)^{-1/2d}}{\delta^{1/d}}\right)} \tag{39}$$

where $\bar{Y}_\tau^{-1} = (\sum_{t=1}^\tau X_t X_t' + V)^{-1}$ and any deterministic V with $V \succ 0$.

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 9.1 and setting $\epsilon = 1/2$, we have that

$$\mathbb{P}(||\bar{Y}_{T-1}^{-1/2} S_{T-1}||_2 \leq y) \leq 5^d \mathbb{P}(||\bar{Y}_{T-1}^{-1/2} S_{T-1} w||_2 \leq \frac{y}{2}) = 5^d \mathbb{P}(||\bar{Y}_{T-1}^{-1/2} S_{T-1} w||_2^2 \leq \frac{y^2}{4}) \tag{40}$$

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. (40)

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

which gives us from Theorem 3

$$\mathbb{P}(||\bar{Y}_{T-1}^{-1/2} S_{T-1}||_2 \leq y) \leq \delta$$

□

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 $t \geq 0$*

$$\Pr(|X'BX - \mathbb{E}[X'BX]| \leq t) \leq 2\exp\left\{ -c\min\Big(\frac{t}{K^2||B||}, \frac{t^2}{K^4||B||_{HS}^2}\Big)\right\} \tag{41}$$

Proposition 9.3 (Theorem 5.39 (Vershynin, 2010)). *Let E be an $T \times d$ matrix whose rows η_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*

$$\sqrt{T} - C\sqrt{d} - t \leq \sigma_{\min}(E) \leq \sqrt{T} + C\sqrt{d} + t \tag{42}$$

The implication of Proposition 9.3 is as follows: $E'E \succeq (\sqrt{T} - C\sqrt{d} - t)^2 I$ with probability at least $1 - 2e^{-ct^2}$. Let $t = \sqrt{\frac{1}{c}\log\frac{2}{\delta}}$, and ensure that

$$T \geq T_\eta(\delta) = C\Big(d + \log\frac{2}{\delta}\Big)$$

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

$$\frac{3}{4}I \preceq \underbrace{\frac{1}{T}\sum_{t=1}^{T}\eta_t\eta_t'}_{E'E} \preceq \frac{5}{4}I \tag{43}$$

Further with the same probability

$$\frac{3\sigma_{\min}^2(P)}{4}I \preceq \frac{1}{T}\sum_{t=1}^{T}P\eta_t\eta_t'P' \preceq \frac{5\sigma_{\max}^2(P)}{4}I$$
$$T_\eta(\delta) = C\Big(d + \log\frac{2}{\delta}\Big) \tag{44}$$

Corollary 9.1 (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 $t \geq 0$

$$\Pr(|X'BX - \mathbb{E}[X'BX]| \leq t) \leq 2\exp\left\{ -c\min\Big(\frac{t}{K^2\sigma_1^2(P)||B||}, \frac{t^2}{K^4\sigma_1^4(P)||B||_{HS}^2}\Big)\right\} \tag{45}$$

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

$$\begin{aligned} ||(I_{n\times n} \otimes P)B(I_{n\times n} \otimes P')|| &\leq \sigma_1^2(P)||B|| \\ \text{tr}((I_{n\times n} \otimes P)B(I_{n\times n} \otimes P')(I_{n\times n} \otimes P)B(I_{n\times n} \otimes P')) &\leq \sigma_1^2(P)\text{tr}((I_{n\times n} \otimes P)B^2(I_{n\times n} \otimes P')) \\ &\leq \sigma_1^4(P)\text{tr}(B^2) \end{aligned}$$

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 9.4. *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 η_t has independent elements. Then with probability at least $1-\delta$, we have

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

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

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

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 9.1 by defining $\tilde{A}$ as

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

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

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

The assertion of proposition follows by applying Markov's Inequality to $\text{tr}(\tilde{A}\tilde{\eta}\eta'\tilde{A}')$. 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 $B = \tilde{A}'\tilde{A}, X = \tilde{\eta}$

$$\begin{aligned} ||B||_S &= \text{tr}(\tilde{A}'\tilde{A}) \\ &= \sum_{t=0}^{T-1} \text{tr}(\Gamma_t(A)) \\ ||B||_F^2 &\leq ||B||_S ||B||_2 \end{aligned}$$

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

$$\frac{t}{c^*||B||} \leq \frac{t^2}{c^*||B||_F^2}$$

we get from Theorem 4 that

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

with probability at least $1 - \exp\Big\{\Big(-\frac{c||B||_F^2}{c^*||B||_2^2}\log\frac{1}{\delta}\Big)\Big\}$. Since $\frac{c||B||_F^2}{c^*||B||_2^2} \geq 1$ it follows that

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

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

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

□

Corollary 9.2. *Whenever $\delta \in (0, e^{-1})$, we have with probability at least $1 - \delta$*

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

for some universal constant c.

Proof. The proof follows the same steps as Proposition 9.4. Define

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

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

$$\begin{aligned} ||\sum_{t=k+1}^{T} X_t X_t'|| \leq \text{tr}(\sum_{t=k+1}^{T} X_t X_t') &= \text{tr}(\sum_{t=1}^{T} X_t X_t' - \sum_{t=1}^{k} X_t X_t') \\ &= \tilde{\eta}'(\tilde{A}'\tilde{A} - \tilde{A}_k'\tilde{A}_k)\tilde{\eta} \end{aligned}$$

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

$$\begin{aligned} ||B||_S &= \text{tr}(\tilde{A}'\tilde{A}) = \sum_{t=k}^{T-1}\text{tr}(\Gamma_t(A)) \\ ||B||_F^2 &\leq ||B||_S ||B||_2 \end{aligned}$$

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

$$\frac{t}{c^*||B||} \leq \frac{t^2}{c^*||B||_F^2}$$

we get from Theorem 4 that

$$\tilde{\eta}'\tilde{A}'\tilde{A}\tilde{\eta} \leq ||B||_S + \frac{||B||_F^2}{c^*||B||}\log\Big(\frac{1}{\delta}\Big) \leq ||B||_S + \frac{||B||_S}{c^*}\log\Big(\frac{1}{\delta}\Big) \leq ||B||_S\Big(1 + \frac{1}{c^*}\log\Big(\frac{1}{\delta}\Big)\Big)$$

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

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

it follows that

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

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

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

□

Proposition 9.5. *Whenever the pdf of X, $f(\cdot)$, satisfies ess sup$_x f(x) = C_X < \infty$ we have*

$$\mathbb{P}(|X|\leq \delta) \leq 2C_X\delta$$

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

$$\mathbb{P}(|X|\leq \delta) = \int_{x=-\delta}^{\delta} f(x)dx \leq 2C_X\delta$$

□

Proposition 9.6 (Proposition 2 in (Faradonbeh et al., 2017))**.** *Let $P^{-1}\Lambda P = A$ be the Jordan decomposition of A and define $z_T = A^{-T}\sum_{i=1}^{T}A^{T-i}\eta_i$. Further assume that η_t is continuous, subGaussian with variance proxy $= 1$ then*

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

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

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

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 $\mathcal{E} = \cup_{i=1}^d\mathcal{E}_i$, then

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

From Proposition 9.5 and Assumption 1, we have $\mathbb{P}(\mathcal{E}_i) \leq 2C_{|P_i^{'}z_T|}y$. Then we get

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

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

10 Lower Bound for Y_T when $A \in \mathcal{S}_0 \cup \mathcal{S}_1$

Here we will prove our results when $\rho(A) \leq 1 + C/T$. 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. This result is a generalization from the case when $\{\bar{\eta}_t\}_{t=1}^T$ are i.i.d., *i.e.*, $L = I$. Recall from Eq. (13) that

$$Y_T \succeq AY_{T-1}A' + \sum_{t=0}^{T-1} Ax_t\eta_{t+1}' + \eta_{t+1}x_t'A' + \sum_{t=1}^{T} \eta_t\eta_t' \tag{46}$$

In this section we will find V_{dn}, V_{up} such that $V_{dn} \preceq Y_T \preceq V_{up}$. The way we will approach this is by first controlling the error cross terms, *i.e.*, $||\sum_{t=0}^{T-1} Ax_t\eta_{t+1}' + \eta_{t+1}x_t'A'||_2 = O(\sqrt{T})$ and then showing that $\sum_{t=1}^T \eta_t\eta_t' = \Omega(T)I$ with high probability. By the inequality in Eq. (46) we will then conclude that $Y_T \succeq \Omega(T)I$.

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

$$\begin{aligned}
P &= AY_{T-1}A' \\
Q &= \sum_{\tau=0}^{T-1} Ax_t\eta_{t+1}' \\
V &= TI \\
T_\eta &= C\Big(\log\frac{2}{\delta} + d\log 5\Big) \\
\mathcal{E}_1(\delta) &= \left\{ ||Q||^2_{(P+V)^{-1}} \leq 8\log\left(\frac{5^d \det(P+V)^{1/2}\det(V)^{-1/2}}{\delta}\right)\right\} \\
\mathcal{E}_2(\delta) &= \left\{ ||\sum_{\tau=0}^{T-1} Ax_\tau x_\tau' A'|| \leq \frac{T\text{tr}(\Gamma_T(A) - I)}{\delta}\right\} \\
\mathcal{E}_\eta(\delta) &= \{T > T_\eta(\delta), \frac{3R^2}{4}I \preceq \frac{1}{T}\sum_{t=1}^T \eta_t\eta_t^{'} \preceq \frac{5}{4}I\} \\
\mathcal{E}(\delta) &= \mathcal{E}_\eta(\delta) \cap \mathcal{E}_1(\delta) \cap \mathcal{E}_2(\delta)
\end{aligned}$$

Proposition 10.1. *If $\rho_i(A) \leq 1 + c/T$ and*

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

then with probability at least $1 - 3\delta$ we have $Y_T \succeq \frac{TR^2}{4}I$.

Proof. Our goal here will be to control $||Q||_2$. Following Proposition 3.1, Proposition 9.4, it is true that $\mathbb{P}(\mathcal{E}_1(\delta) \cap \mathcal{E}_2(\delta)) \geq 1 - 2\delta$. 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

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

$$U'Y_TU \geq U'AY_{T-1}A'U + U'\sum_{t=0}^{T-1}\Big(Ax_t\eta_{t+1}' + \eta_{t+1}x_t'A'\Big)U + \frac{3}{4}TR^2 \quad \forall U \in \mathcal{S}^{d-1} \tag{47}$$

Intersecting Eq. (47) with $\mathcal{E}_1(\delta) \cap \mathcal{E}_2(\delta)$, we find under $\mathcal{E}(\delta)$

$$\begin{aligned}
||Q||^2_{(P+V)^{-1}} &\leq 8\log\left(\frac{5^d \text{det}(P+V)^{1/2}\text{det}(V)^{-1/2}}{\delta}\right) \\
&\leq 8\log\left(\frac{5^d \text{det}(\frac{TI\text{tr}(\Gamma_T(A)-I)}{\delta} + TI)^{1/2}\text{det}(TI)^{-1/2}}{\delta}\right) \\
&\leq 8\log\left(\frac{5^d \text{det}(\text{tr}(\Gamma_T(A)-I)I+I)^{1/2}}{\delta^d}\right)
\end{aligned}$$

Using Proposition 8.1 and letting $\kappa^2 = U'PU$ then

$$\begin{aligned}
&||QU||_2 \\
&\leq \sqrt{\kappa^2+T}\sqrt{8\log\left(\frac{5^d \text{det}(\text{tr}(\Gamma_T(A)-I)I+I)^{1/2}}{\delta^d}\right)}
\end{aligned}$$

So Eq. (47) implies

$$U'Y_TU \geq \kappa^2 - \sqrt{(\kappa^2+T)}\sqrt{16d\log\left(\text{tr}(\Gamma_T-I)+1\right) + 32d\log\frac{5}{\delta}} + \frac{3}{4}TR^2$$

which gives us

$$U'\frac{Y_T}{T}U \geq \frac{\kappa^2}{T} - \sqrt{(\frac{\kappa^2}{T}+1)}\underbrace{\sqrt{\frac{16d}{T}\log\left(\text{tr}(\Gamma_T-I)+1\right) + \frac{32d}{T}\log\frac{5}{\delta}}}_{=\beta} + \frac{3}{4}R^2 \tag{48}$$

If we can ensure

$$\frac{TR^4}{128} \geq \frac{d}{2}\log\left(\text{tr}(\Gamma_T-I)+1\right) + d\log\frac{5}{\delta} \tag{49}$$

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

$$\sqrt{\frac{16d}{T}\log\left(\text{tr}(\Gamma_T-I)+1\right) + \frac{32d}{T}\log\frac{5}{\delta}} \leq \frac{R^2}{2}$$

Let T be large enough that Eq. (49) is satisfied then Eq. (48) implies

$$U'\frac{Y_T}{T}U \geq \frac{\kappa^2}{T} - \frac{\sqrt{(\frac{\kappa^2}{T}+1)}R^2}{2} + \frac{3R^2}{4} \geq \frac{R^2}{4} + \frac{\kappa^2}{2T} \tag{50}$$

Since U is arbitrarily chosen Eq. (50) implies

$$Y_T \succeq \frac{TR^2}{4}I \tag{51}$$

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

$$\begin{aligned}
&\rho_i(A) \leq 1 + \frac{c}{T} \\
&\quad T \geq \max\Big(C\Big(\log\frac{2}{\delta} + d\log 5\Big), CR^2\Big(\frac{d}{2}\log\left(\text{tr}(\Gamma_T-I)+1\right) + d\log\frac{5}{\delta}\Big)\Big)
\end{aligned} \tag{52}$$

□

Remark 2. *Eq. (49) is satisfied whenever tr$(\Gamma_T - I)$ grows at most polynomially in T. This is true whenever $\rho(A) \leq 1 + \frac{c}{T}$.*

11 Sharpened bounds when $1 - \frac{c}{T} \leq \rho_i(A) \leq 1 + \frac{c}{T}$

Here we show that the bound for Y_T in Eq. (51) can be sharpened to have quadratic growth in T. The key idea towards sharpening will be that

$$\begin{aligned} Y_T &\succeq \underbrace{AY_{T-1}A'}_{\approx (1-\frac{c}{T})Y_T} + \underbrace{\sum_{t=0}^{T-1} Ax_t\eta_{t+1}' + \eta_{t+1}x_t'A' + \sum_{t=1}^{T} \eta_t\eta_t'}_{\approx CTI} \\ Y_T &\succeq CT^2 I \end{aligned}$$

Formally,

Proposition 11.1. *Let* $1 - c/T \leq \rho_i(A) \leq 1 + c/T$ *and*

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

then with probability at least $1 - \delta$ *we have*

$$Y_T \succeq \frac{\sqrt{\alpha(d)}T^2R^4\sigma_{\min}(AA')}{256e^2c(A,\delta)}I$$

where $\alpha(\cdot)$ *is a function of only d,* R *is an absolute constant and*

$$c(A,\delta) = 16d\log\left(tr(\Gamma_T - I) + 1\right) + 32d\log\frac{15T}{2\delta}$$

Proof. For this we want Eq. (51) satisfied for every $t \geq \frac{T}{2}$ simultaneously, *i.e.*, we need

$$Y_t \succeq \frac{tR^2}{4}I \tag{53}$$

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

$$\begin{aligned} &\rho_i(A) \leq 1 \\ &\quad t \geq \max\Big(C\Big(\log\frac{2}{\delta} + d\log 5\Big), CR^2\Big(\frac{d}{2}\log\left(\text{tr}(\Gamma_t - I) + 1\right) + d\log\frac{5}{\delta}\Big)\Big) \end{aligned} \tag{54}$$

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

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

when Eq. (54) is satisfied for each t. Since the LHS of Eq. (54) 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. (54) is the following

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

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

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

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

$$T \geq \max\Big(C\Big(\log\frac{3T}{2\delta} + d\log 5\Big), CR^2\Big(\frac{d}{2}\log\left(\text{tr}(\Gamma_T - I) + 1\right) + d\log\frac{15T}{2\delta}\Big)\Big) \tag{55}$$

Define $\gamma_{t-1} = \sqrt{U'A'Y_{t-1}AU}$ and Eq. (50) becomes

$$\begin{aligned}
U'Y_tU &\geq \gamma_{t-1}^2 - \sqrt{(\gamma_{t-1}^2 + t)}\underbrace{\sqrt{16d\log\left(\text{tr}(\Gamma_t - I) + 1\right) + 32d\log\frac{15T}{2\delta}}}_{\text{Under Eq. (55) is}\leq \frac{R^2\sqrt{t}}{2}} + \frac{3}{4}tR^2 \\
&\geq \gamma_{t-1}^2 - (\gamma_{t-1} + \sqrt{t})\sqrt{16d\log\left(\text{tr}(\Gamma_t - I) + 1\right) + 32d\log\frac{15T}{2\delta}} + \frac{3t}{4}R^2 \\
&\geq \gamma_{t-1}^2 - \gamma_{t-1}\sqrt{16d\log\left(\text{tr}(\Gamma_t - I) + 1\right) + 32d\log\frac{15T}{2\delta}} + \frac{3tR^2}{4} - \sqrt{t}\underbrace{\sqrt{16d\log\left(\text{tr}(\Gamma_t - I) + 1\right) + 32d\log\frac{15T}{2\delta}}}_{\leq R^2\frac{\sqrt{t}}{2}} \\
&\geq \gamma_{t-1}^2\Big(1 - \sqrt{\frac{16d\log\left(\text{tr}(\Gamma_t - I) + 1\right) + 32d\log\frac{15T}{2\delta}}{\gamma_{t-1}^2}}\Big) + \frac{tR^2}{4} \\
&\geq \gamma_{t-1}^2\Big(1 - \underbrace{\sqrt{\frac{16d\log\left(\text{tr}(\Gamma_T - I) + 1\right) + 32d\log\frac{15T}{2\delta}}{\gamma_{t-1}^2}}\Big)}_{=\sqrt{\frac{c(A,\delta)}{\gamma_{t-1}^2}}} + \frac{TR^2}{8} \qquad (56)
\end{aligned}$$

Observe that

$$\gamma_{t-1} = \sqrt{U'A'Y_{t-1}AU} \geq \sigma_{\min}(A)\sqrt{\frac{TR^2}{8e}} \qquad (57)$$

Eq. (56) will give us a non–trivial bound only when $\frac{c(A,\delta)}{\gamma_{t-1}^2} \leq 1/4$ which is true whenever

$$T \geq \frac{64ec(A,\delta)}{R^2\sigma_{\min}^2(A)} \qquad (58)$$

The scaling $1 - \sqrt{\frac{c(A,\delta)}{\gamma_{t-1}^2}}$ in Eq. (56) depends on γ_{t-1} itself. We will show that

$$\begin{aligned}
&\gamma_{t-1}^2 = T\Omega(1) \implies \gamma_{t-1}^2 = T\Omega\Big(\sqrt{\frac{T}{c(A,\delta)}}\Big) \\
&\gamma_{t-1}^2 = T\Omega\Big(\Big(\frac{T}{c(A,\delta)}\Big)^{1/2}\Big) \implies \gamma_{t-1}^2 = T\Omega\Big(\Big(\frac{T}{c(A,\delta)}\Big)^{3/4}\Big) \\
&\gamma_{t-1}^2 = T\Omega\Big(\Big(\frac{T}{c(A,\delta)}\Big)^{\frac{2^k-1}{2^k}}\Big) \implies \gamma_{t-1}^2 = T\Omega\Big(\Big(\frac{T}{c(A,\delta)}\Big)^{\frac{2^{k+1}-1}{2^{k+1}}}\Big) \\
&\implies \ldots \implies \gamma_{t-1}^2 = T\Omega\Big(\frac{T}{c(A,\delta)}\Big)
\end{aligned}$$

From Eq. (56),(57) 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

$$Y_t \succeq \left(1 - \underbrace{\sqrt{\frac{16ec(A,\delta)}{\sigma_{\min}(AA')TR^2}}}_{=\beta_1}\right)AY_{t-1}A' + \frac{R^2TI}{8} \qquad (59)$$

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. (59) implies that

$$\begin{aligned}
Y_t &\overset{(a)}{\succeq} \frac{TR^2}{8}\sum_{k=1}^{\min(\lfloor\frac{1}{\beta_1}\rfloor, \frac{T}{4})}(1-\beta_1)^k A^k A^{k\prime} + \frac{R^2TI}{16} \\
&\overset{(b)}{\succeq} \frac{TR^2}{16e}\sum_{k=1}^{\min(\lfloor\frac{1}{\beta_1}\rfloor, \frac{T}{4})} A^k A^{k\prime} + \frac{R^2TI}{16} \\
&\succeq \frac{R^2T}{16e}\Gamma_{\lfloor\frac{1}{\beta_1}\rfloor}(A) + \frac{R^2TI}{16}
\end{aligned}$$

Here

$$\beta_1 = \sqrt{\frac{16ec(A,\delta)}{\sigma_{\min}(AA')R^2T}} \tag{60}$$

Due to the choice of T, d we will usually have $\lfloor\frac{1}{\beta_1}\rfloor^2 \leq \frac{T}{4}$. (a) follows by successively expanding Eq. (59), (b) follows because $(1-\beta_1)^{\lfloor\frac{1}{\beta_1}\rfloor} \geq \frac{e^{-1}}{2}$ since $\beta_1 \leq 1/2$ by Eq. (58). Then we can conclude that

$$\begin{aligned}
\gamma_{t-1}^2 &\geq \sigma_{\min}(AY_tA') \\
&\geq \frac{R^2T\sigma_{\min}(AA')\sigma_{\min}(\Gamma_{\lfloor\frac{1}{\beta_1}\rfloor}(A))}{16e}
\end{aligned} \tag{61}$$

which gives us

$$\sqrt{\frac{c(A,\delta)}{\gamma_{t-1}^2}} \leq \Big(\frac{16ec(A,\delta)}{R^2T\sigma_{\min}(AA')\sigma_{\min}(\Gamma_{\lfloor\frac{1}{\beta_1}\rfloor}(A))}\Big)^{1/2} = \beta_2 \tag{62}$$

It is clear from Eq. (62) that we get a recursion during the refinement process. Specifically at the k^{th} repetition of Eq. (59) up to Eq. (62) we get,

$$\beta_k = \Big(\frac{16ec(A,\delta)}{R^2T\sigma_{\min}(AA')\sigma_{\min}(\Gamma_{\lfloor\frac{1}{\beta_{k-1}}\rfloor}(A))}\Big)^{1/2} \tag{63}$$

Now β_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 $\Gamma_{t_1}(A) \succeq \Gamma_{t_2}(A)$ whenever $t_1 \geq t_2$ we have

$$\Big(\frac{16ec(A,\delta)}{R^2T\sigma_{\min}(AA')\sigma_{\min}(\Gamma_{\lfloor\frac{1}{\beta_m}\rfloor}(A))}\Big)^{1/2} \leq \Big(\frac{16ec(A,\delta)}{R^2T\sigma_{\min}(AA')\sigma_{\min}(\Gamma_{\lfloor\frac{1}{\beta_{m-1}}\rfloor}(A))}\Big)^{1/2}$$

$$\beta_{m+1} \leq \beta_m$$

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. (63), *i.e.*

$$\beta_0^2\sigma_{\min}(\Gamma_{\lfloor\frac{1}{\beta_0}\rfloor}(A)) = \Big(\frac{16ec(A,\delta)}{R^2T\sigma_{\min}(AA')}\Big) \tag{64}$$

Now a solution for $\beta_0 \in (\frac{2C}{\sigma_{\min}(AA')TR^2}, 1)$. To see this set $\beta_0 = 1$, then LHS > RHS. Next set $\beta_0 = \frac{2C}{\sigma_{\min}(AA')TR^2}$ then since $\rho_{\min}(A^t) \geq \sigma_{\min}(A^t)$ and $\rho_i \leq 1 + C/T$ we see that

$$\begin{aligned}
\frac{4C^2\sigma_{\min}(\Gamma_{\lfloor\frac{1}{\beta_0}\rfloor}(A))}{\sigma_{\min}(AA')^2T^2} &\leq \frac{4\sum_{t=0}^{\sigma_{\min}(A)^2R^2T/2C}\rho_{\min}(A)^{2t}}{R^4\sigma_{\min}(AA')^2T^2/C^2} \\
&\leq \frac{2eC}{\sigma_{\min}(A)^2T} \leq \Big(\frac{16ec(A,\delta)}{R^2T\sigma_{\min}(AA')}\Big)
\end{aligned}$$

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 $T \geq \frac{64ec(A,\delta)}{\sigma_{\min}(A)^2}$ we get Eq. (59)

$$Y_t {\succeq} (1 - \beta_0) A Y_{t-1} A' + \frac{TR^2 I}{8} \tag{65}$$

and following as before we get with probability at least $1 - \delta$

$$Y_T \succeq \frac{TR^2}{16e} \Gamma_{\lfloor \frac{1}{\beta_0} \rfloor}(A) + \frac{TR^2 I}{16} \tag{66}$$

where β_0 is solution to

$$\beta_0^2 \sigma_{\min}(\Gamma_{\lfloor \frac{1}{\beta_0} \rfloor}(A)) = \Big(\frac{16ec(A, \delta)}{TR^2 \sigma_{\min}(AA')}\Big)$$

and

$$c(A, \delta) = 16d \log\left(\mathrm{tr}(\Gamma_T - I) + 1\right) + 32d \log \frac{15T}{2\delta}$$

It should be noted that $\frac{1}{\beta_0}$ will equal $\frac{\sqrt{\alpha(d)} TR^2 \sigma_{\min}(AA')}{16ec(A,\delta)}$, *i.e.*, grow linearly with T, as shown in Proposition 8.5. Then it can be seen from Eq. (66) that

$$\begin{aligned}
Y_T &\succeq \frac{TR^2}{16e} \Gamma_{\lfloor \frac{1}{\beta_0} \rfloor}(A) + \frac{TR^2 I}{16} \\
Y_T &\succeq \frac{TR^2}{16e} \sigma_{\min}(\Gamma_{\lfloor \frac{1}{\beta_0} \rfloor}(A)) + \frac{TR^2 I}{16} \\
&\succeq \frac{TR^2}{16e} \frac{TR^2 \sqrt{\alpha(d)} \sigma_{\min}(AA')}{16ec(A, \delta) C(d)} I = \frac{\sqrt{\alpha(d)} T^2 R^4 \sigma_{\min}(AA')}{256e^2 c(A, \delta)} I
\end{aligned} \tag{67}$$

□

12 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 $\sigma_{\min}(LL') = R^2 > 0$. In this section we show the invertibility (with high probability) of Y_T when A is regular and explosive.

Let $\sigma_{\max}(LL') = 1$. Recall that

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

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 from Eq. (18) that $U_T = A^{-T} \sum_{t=1}^T x_t x_t' A^{-T'}, F_T = \sum_{t=1}^T A^{-t+1} z_T z_T' A^{-t+1'}$. U_T is a scaled version of Y_T and we will show that $||U_T - F_T||_2 \leq c$ with high probability. Then we show that $F_T \succeq 2cI$ as a result $U_T \succeq cI$ with high probability. This behavior is only possible due to the regularity of the matrix A and significantly different from Section 10.

Bounding $||F_T - U_T||_{\text{op}}$

Proposition 12.1. *We have with probability at least* $1 - 2\delta$,

$$||U_T - F_T||_2 \leq \Bigg(4T^2\sigma_1^2(A^{-(T+1)\epsilon})tr(\Gamma_T(A^{-1})) + \Big(1 + \frac{1}{c}\log\frac{1}{\delta}\Big)T tr(A^{-T-1}\Gamma_T(A^{-1})A^{-T-1\prime})\Bigg)$$

whenever

$$T \geq T_0 = \frac{2}{c}\Big(\log\frac{1}{\delta} + \log 2 + 2d\log 5\Big)$$

Proof. Observe that

$$z(T) - z(T-t) = A^{-T+t-1}\left(\sum_{s=0}^{t-1} A^{-s}\eta_{T+1-t+s}\right) = A^{-T+t-1}z(T,t) \tag{68}$$

Then

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

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

$$\begin{aligned} ||U_T - F_T||_{\text{op}} \leq\, & ||\sum_{t=1}^{T} A^{-t}(z(T-t) - z(T))(z(T-t) - z(T))^{'}A^{-t^{'}}||_2 \\ & + ||\sum_{t=1}^{T} A^{-t}((z(T-t) - z(T))z(T-t)^{'} + z(T-t)(z(T-t)^{'} - z(T)^{'})A^{-t^{'}}||_2 \end{aligned} \tag{69}$$

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. (68) we see that $A^{-t}(z(T-t) - z(T)) = -A^{-T-1}z(T,t)$, then

$$A^{-T-1}z(T,t) = A^{-T-1}[0, 0, \ldots, \underbrace{I}_{T-t+1 \text{ term}}, A^{-1}, A^{-2}, \ldots, A^{-t+1}] \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_T \end{bmatrix}$$

Since $\sum_{t=1}^{T}(z(T-t) - z(T))(z(T-t) - z(T))^{'} \preceq \sum_{t=1}^{T} \text{trace}((z(T-t) - z(T))(z(T-t) - z(T))^{'})I$. Based on these observations we have

$$\begin{aligned} & ||\sum_{t=1}^{T} A^{-t}(z(T-t) - z(T))(z(T-t) - z(T))^{'}A^{-t^{'}}||_2 = ||\sum_{t=1}^{T} A^{-T-1}z(T,t)z(T,t)^{'}A^{-T-1^{'}}||_2 \\ & \leq \text{trace}(A^{-T-1}\sum_{t=1}^{T} z(T,t)z(T,t)^{'}A^{-T-1^{'}}) = \sum_{t=1}^{T} z(T,t)^{'}A^{-T-1^{'}}A^{-T-1}z(T,t) = \tilde{\eta}^{'}\tilde{A}^{'}\tilde{A}\tilde{\eta} \end{aligned}$$

where $\tilde{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_T \end{bmatrix}$ and

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

Since $\text{tr}(\tilde{A}\tilde{A}') = T\text{tr}(A^{-T-1}\Gamma_T(A^{-1})A^{-T-1\prime})$. Applying Markov's Inequality (See Proposition 9.4), we have with probability at least $1-\delta$ that

$$\tilde{\eta}^{'}\tilde{A}^{'}\tilde{A}\tilde{\eta} \leq \frac{\text{tr}(\mathbb{E}[\tilde{A}\tilde{\eta}\tilde{\eta}^{'}\tilde{A}^{'}])}{\delta} \leq \frac{\sigma_1(L)^2 T\text{tr}(A^{-T-1}\Gamma_T(A^{-1})A^{-T-1\prime})}{\delta} \tag{70}$$

Although this bound can be tightened by dependent Hanson–Wright (See Corollary 9.1), there is no reason to do so as δ depends only logarithmically on T. In fact we get with probability at least $1-\delta$ that

$$\tilde{\eta}^{'}\tilde{A}^{'}\tilde{A}\tilde{\eta} \leq \Big(1 + \frac{1}{c}\log\frac{1}{\delta}\Big)(\sigma_1(L)^2 T\text{tr}(A^{-T-1}\Gamma_T(A^{-1})A^{-T-1\prime})) \tag{71}$$

Next we analyze the second term

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

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

$$\sum_{t=1}^{T} A^{-t}((z(T-t) - z(T))z(T-t)^{'}A^{-t\prime} = A^{-T-1}\sum_{t=1}^{T} z(T,t)z(T-t)^{'}A^{-t\prime} \tag{72}$$

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

$$\tilde{z}(T,t) = A^{-T-1}z(T,t) = A^{-T-1}\underbrace{[0,0,\ldots,\underbrace{I}_{T-t+1 \text{ term}},A^{-1},A^{-2},\ldots,A^{-t+1}]}_{A(T,t)}\begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_T \end{bmatrix}$$

$$\tilde{z}(T-t)' = z(T-t)'A^{-t\prime} = \underbrace{[\eta_1', \eta_2', \ldots, \eta_T']}_{\tilde{\eta}'}\underbrace{\begin{bmatrix} A^{-t-1\prime} \\ A^{-t-2\prime} \\ \vdots \\ A^{-T\prime} \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{A(T-t)'} + x_0$$

Then the probability of the second term can be written as

$$\mathbb{P}(||\sum_{t=1}^{T}(\tilde{z}(T,t)\tilde{z}(T-t)^{'} + \tilde{z}(T-t)\tilde{z}(T,t)^{'})||_2 \geq z) \underbrace{\leq}_{\frac{1}{2}-\text{net}} 2 \times 5^{2d} \times \mathbb{P}(\Big|\sum_{t=1}^{T} 2u^{'}\tilde{z}(T,t)\tilde{z}(T-t)^{'}v\Big|) \geq z/4)$$

$$\leq 2 \times 5^{2d} \times \mathbb{P}\left(\left|\tilde{\eta}'\Big(\sum_{t=1}^{T} A(T,t)'A^{-T-1\prime}uv'A(T-t) + A(T-t)'vu'A^{-T-1}A(T,t)\Big)\tilde{\eta}\right| \leq z/4\right) \tag{73}$$

To Eq. (73) apply Hanson-Wright inequality. For any u, v, due to the statistical independence of $z(T-t), z(T,t)$ we have

$$\mathbb{E}[\sum_{t=1}^{T} 2u^{'}\tilde{z}(T,t)\tilde{z}(T-t)^{'}v] = 0$$

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

$$\begin{aligned}
S &= \sum_{t=1}^{T} A(T,t)'A^{-T-1\prime}uv'A(T-t) + A(T-t)'vu'A^{-T-1}A(T,t) \\
&= \sum_{t=1}^{T} \underbrace{A(T,t)'A^{-(T+1)\epsilon\prime}}_{=C} \underbrace{A^{-(T+1)(1-\epsilon)\prime}uv'A(T-t)}_{=D'} + A(T-t)'vu'A^{-(T+1)(1-\epsilon)}A^{-(T+1)\epsilon}A(T,t) \\
&\preceq \sum_{t=1}^{T} \underbrace{A(T,t)'A^{-(T+1)\epsilon\prime}A^{-(T+1)\epsilon}A(T,t)}_{=CC'} + \sum_{t=1}^{T} \underbrace{A(T-t)'vu'A^{-(T+1)(1-\epsilon)}A^{-(T+1)(1-\epsilon)\prime}uv'A(T-t)}_{=DD'} \\
&\preceq \sigma_1^2(A^{-(T+1)\epsilon}) \sum_{t=1}^{T} A(T,t)'A(T,t) + u'A^{-(T+1)(1-\epsilon)}A^{-(T+1)(1-\epsilon)\prime}u \sum_{t=1}^{T} A(T-t)'vv'A(T-t) \\
&\preceq \sigma_1^2(A^{-(T+1)\epsilon})\text{tr}\Big(\sum_{t=1}^{T} A(T,t)'A(T,t)\Big)I + \sigma_1^2(A^{-(T+1)(1-\epsilon)})\text{tr}\Big(\sum_{t=1}^{T} A(T-t)'vv'A(T-t)\Big)I \\
&\overset{(a)}{\preceq} 2T\sigma_1^2(A^{-(T+1)\epsilon})\text{tr}(\Gamma_T(A^{-1}))I
\end{aligned}$$

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

$$T \geq T_0 = \frac{2}{c}\Big(\log\frac{1}{\delta} + \log 2 + 2d\log 5\Big) \tag{74}$$

Eq. (73) becomes with probability at least $1-\delta$ that

$$||\sum_{t=1}^{T}((z(T-t) - z(T))z(T-t)^{'} + z(T-t)(z(T-t)^{'} - z(T)^{'})||_2 \leq 4T^2\sigma_1^2(A^{-(T+1)\epsilon})\text{tr}(\Gamma_T(A^{-1})) \tag{75}$$

Then combining Eq. (70),(75) we get for $T \geq T_0$ given in Eq. (74),

$$||U_T - F_T||_2 \leq \left(4T^2\sigma_1^2(A^{-(T+1)\epsilon})\text{tr}(\Gamma_T(A^{-1})) + \frac{T\text{tr}(A^{-T-1}\Gamma_T(A^{-1})A^{-T-1\prime})}{\delta}\right) \tag{76}$$

with probability at least $1-2\delta$. We pick ϵ such that $(T+1)\epsilon = \lfloor\frac{T+1}{2}\rfloor$. In fact using Eq. (71) instead of Eq. (70) we get

$$||U_T - F_T||_2 \leq \left(4T^2\sigma_1^2(A^{-(T+1)\epsilon})\text{tr}(\Gamma_T(A^{-1})) + \Big(1 + \frac{1}{c}\log\frac{1}{\delta}\Big)T\text{tr}(A^{-T-1}\Gamma_T(A^{-1})A^{-T-1\prime})\right) \tag{77}$$

□

Bounding U_T

Proposition 12.2. *We have with probability at least $1-4\delta$*

$$\begin{aligned}
Y_T &\succeq \frac{\phi_{\min}(A)^2\psi(A)^2\delta^2}{2\sigma_{\max}(P)^2}A^TA^{T\prime} \\
Y_T &\preceq \frac{3\phi_{\max}(A)^2}{2\sigma_{\min}(P)^2}(1 + \frac{1}{c}\log\frac{1}{\delta})tr(P(\Gamma_T(A^{-1}) - I)P')A^TA^{T\prime}
\end{aligned}$$

whenever

$$\left(4T^2\sigma_1^2(A^{-(T+1)\epsilon})tr(\Gamma_T(A^{-1})) + \frac{Ttr(A^{-T-1}\Gamma_T(A^{-1})A^{-T-1\prime})}{\delta}\right) \leq \frac{\phi_{\min}(A)^2\psi(A)^2\delta^2}{2\sigma_{\max}(P)^2}$$

Proof. To give lower and upper bounds on U_T, we need to bound F_T. The steps involve

$$\begin{aligned}
||U_T - F_T||_2 &\leq \Delta \\
F_T &\succeq V_{dn} \succ 0 \\
\implies U_T &\geq V_{dn} - \Delta I \\
F_T &\preceq V_{up} \\
\implies U_T &\preceq V_{up} + \Delta I
\end{aligned}$$

From Proposition 13.1 we get, with probability at least $1 - 2\delta$,

$$\begin{aligned}
F_T &\succeq \phi_{\min}(A)^2\psi(A)^2\delta^2\sigma_{\min}(P^{-1})^2 I \\
F_T &\preceq \frac{\phi_{\max}(A)^2}{\sigma_{\min}(P)^2}(1 + \frac{1}{c}\log\frac{1}{\delta})\text{tr}(P(\Gamma_T(A^{-1}) - I)P')I
\end{aligned}$$

Define

$$\begin{aligned}
\Delta &= \frac{1}{2}\min\left(\frac{\phi_{\max}(A)^2}{\sigma_{\min}(P)^2}(1 + \frac{1}{c}\log\frac{1}{\delta})\text{tr}(P(\Gamma_T(A^{-1}) - I)P'), \phi_{\min}(A)^2\psi(A)^2\delta^2\sigma_{\min}(P^{-1})^2\right) \\
&= \frac{\phi_{\min}(A)^2\psi(A)^2\delta^2\sigma_{\min}(P^{-1})^2}{2}
\end{aligned}$$

Then in Eq. (76) by ensuring that

$$\left(4T^2\sigma_1^2(A^{-(T+1)\epsilon})\text{tr}(\Gamma_T(A^{-1})) + \frac{T\text{tr}(A^{-T-1}\Gamma_T(A^{-1})A^{-T-1\prime})}{\delta}\right) \leq \frac{\phi_{\min}(A)^2\psi(A)^2\delta^2}{2\sigma_{\max}(P)^2}$$

we get with probability at least $1 - 4\delta$ (since this is the intersection of events governed by Eq. (76),(83),(84))

$$U_T \succeq \phi_{\min}(A)^2\psi(A)^2\delta^2\sigma_{\min}(P^{-1})^2 I - \frac{\phi_{\min}(A)^2\psi(A)^2\delta^2}{2\sigma_{\max}(P)^2}I \succeq \frac{\phi_{\min}(A)^2\psi(A)^2\delta^2}{2\sigma_{\max}(P)^2}I \tag{78}$$

Similarly, for the upper bound

$$U_T \preceq \frac{3\phi_{\max}(A)^2}{2\sigma_{\min}(P)^2}(1 + \frac{1}{c}\log\frac{1}{\delta})\text{tr}(P(\Gamma_T(A^{-1}) - I)P')I \tag{79}$$

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

$$\begin{aligned}
Y_T &\succeq \frac{\phi_{\min}(A)^2\psi(A)^2\delta^2}{2\sigma_{\max}(P)^2}A^T A^{T\prime} \\
Y_T &\preceq \frac{3\phi_{\max}(A)^2}{2\sigma_{\min}(P)^2}(1 + \frac{1}{c}\log\frac{1}{\delta})\text{tr}(P(\Gamma_T(A^{-1}) - I)P')A^T A^{T\prime}
\end{aligned} \tag{80}$$

whenever

$$\left(4T^2\sigma_1^2(A^{-(T+1)\epsilon})\text{tr}(\Gamma_T(A^{-1})) + \frac{T\text{tr}(A^{-T-1}\Gamma_T(A^{-1})A^{-T-1\prime})}{\delta}\right) \leq \frac{\phi_{\min}(A)^2\psi(A)^2\delta^2}{2\sigma_{\max}(P)^2} \tag{81}$$

□

13 Regularity and Invertibility

Through a counterexample in (Nielsen, 2008), Remark 4 in (Phillips & Magdalinos, 2013) it is shown that unless a matrix is regular, the estimation of the parameters maybe asymptotically inconsistent.

Recall F_T from Eq. (18). 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 (Faradonbeh et al., 2017) and has been presented here for completeness. It will be useful to recall the definitions of $\phi_{\min}(A), \phi_{\max}(A)$ from Definition 3.

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

Proposition 13.1. *Let A be regular, then we have with probability at least $1-2\delta$*

$$\sigma_{\min}(F_T) \geq \frac{\phi_{\min}(A)^2}{\sigma_{\max}(P)^2}\psi(A)^2\delta^2$$
$$\sigma_{\max}(F_T) \leq \frac{\phi_{\max}(A)^2}{\sigma_{\min}(P)^2}(1+\frac{1}{c}\log\frac{1}{\delta})tr(P(\Gamma_T(A^{-1})-I)P')$$

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

$$A^{-t}z_T = P^{-1}\Lambda^{-t}Pz_T$$

Define the event

$$\mathcal{E}_+(\delta) = \{\min_{1\leq i\leq d}|P_i' z_T|> \psi(A)\delta\}$$

where $\psi(A)$ is the lower bound shown in Proposition 3.2 (which we can use due to the continuity of $L\bar{\eta}_t$) and $v = Pz_T$. Under $\mathcal{E}_+(\delta)$, $|v_i|> 0$. Now we need a lower bound for $\sigma_{\min}(F_T)$ under $\mathcal{E}_+(\delta)$

$$F_T = P^{-1}\sum_{i=1}^T \Lambda^{-i+1}Pz_Tz_T'P'\Lambda^{-i+1'}P^{-1'} = P^{-1}\sum_{i=1}^T \Lambda^{-i+1}vv'\Lambda^{-i+1'}P^{-1'} \tag{82}$$
$$\succeq \phi_{\min}(A)^2\psi(A)^2\delta^2P^{-1}P^{-1'}\succeq \frac{\phi_{\min}(A)^2}{\sigma_{\max}(P)^2}\psi(A)^2\delta^2 I \tag{83}$$

Further, since A is regular we have that $\phi_{\min}(A) > 0$ from Proposition 8.3. 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 Pz_T is a sub-Gaussian random variable. Since

$$||Pz_Tz_T'P'||\leq z_T'P'Pz_T$$

and recalling that

$$z_T = \underbrace{[A^{-1}, A^{-2}, \ldots, A^{-T}]}_{\tilde{A}}\begin{bmatrix}\eta_1\\ \eta_2\\ \vdots\\ \eta_T\end{bmatrix}$$

we can use dependent Hanson Wright inequality (Corollary 9.1) to bound $z_T'P'Pz_T$. In Theorem 4,

$$B = \tilde{A}'P'P\tilde{A}$$
$$\mathbb{E}[z_T'P'Pz_T] = \text{tr}(P(\Gamma_T(A^{-1})-I)P')\sigma_1(L)^2 = \text{tr}(P(\Gamma_T(A^{-1})-I)P')$$
$$||B||_2, ||B||_F\leq \text{tr}(\tilde{A}'P'P\tilde{A}) = \text{tr}(P(\Gamma_T(A^{-1})-I)P')$$

Then with probability at least $1 - \delta$ we have

$$z_T' P' P z_T \leq (1 + \frac{1}{c} \log \frac{1}{\delta}) \text{tr}(P(\Gamma_T(A^{-1}) - I)P')$$

and we get from Eq. (82)

$$\begin{aligned} F_T &\preceq P^{-1} \sum_{i=1}^{T} \Lambda^{-i+1} P z_T z_T' P' \Lambda^{-i+1\prime} P^{-1\prime} \preceq (z_T' P' P z_T) \sup_{||v||_2=1} \sigma_{\max}\Big(P^{-1} \sum_{i=1}^{T} \Lambda^{-i+1} v v' \Lambda^{-i+1\prime} P^{-1\prime}\Big) I \\ &\preceq \frac{\phi_{\max}(A)^2}{\sigma_{\min}(P)^2} (1 + \frac{1}{c} \log \frac{1}{\delta}) \text{tr}(P(\Gamma_T(A^{-1}) - I)P') I \end{aligned} \tag{84}$$

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

$$F_T \succeq \frac{\phi_{\min}(A)^2}{\sigma_{\max}(P)^2} \psi(A)^2 \delta^2 I \tag{85}$$

$$F_T \preceq \frac{\phi_{\max}(A)^2}{\sigma_{\min}(P)^2} (1 + \frac{1}{c} \log \frac{1}{\delta}) \text{tr}(P(\Gamma_T(A^{-1}) - I)P') I \tag{86}$$

□

14 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 $\tilde{P}$ such that

$$A = \tilde{P}^{-1} \underbrace{\begin{bmatrix} A_e & 0 & 0 \\ 0 & A_{ms} & 0 \\ 0 & 0 & A_s \end{bmatrix}}_{=\tilde{A}} \tilde{P} \tag{87}$$

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 Λ is a block diagonal matrix and each block corresponds to an eigenvalue. We can always find Q (a rearrangement matrix) such that Λ is partitioned into two diagonal parts: explosive, marginally stable and stable, *i.e.*,

$$A = P^{-1} Q^T \begin{bmatrix} \Lambda_e & 0 & 0 \\ 0 & \Lambda_{ms} & 0 \\ 0 & 0 & \Lambda_s \end{bmatrix} QP \tag{88}$$

Clearly, $\tilde{P} = QP$. Since

$$\begin{aligned} X_t &= \sum_{\tau=1}^{t} A^{\tau-1} \eta_{t-\tau+1} \\ \tilde{X}_t = \tilde{P} X_t &= \sum_{\tau=1}^{t} \tilde{A}^{\tau-1} \underbrace{\tilde{P} \eta_{t-\tau+1}}_{\tilde{\eta}_{t-\tau+1}} \end{aligned} \tag{89}$$

Now, the transformed dynamics are as follows:

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

where $\tilde{A}$ has been partitioned into explosive and stable components as Eq. (87). Corresponding to $\tilde{A}$ partition $\tilde{X}_t, \tilde{\eta}_t$

$$\tilde{X}_t = \begin{bmatrix} X_t^e \\ X_t^{ms} \\ X_t^s \end{bmatrix}, \tilde{\eta}_t = \begin{bmatrix} \eta_t^e \\ \eta_t^{ms} \\ \eta_t^s \end{bmatrix} \tag{90}$$

$$\tilde{Y}_T = \sum_{t=1}^{T} \tilde{X}_t \tilde{X}_t' = \sum_{t=1}^{T} \begin{bmatrix} X_t^e (X_t^e)' & X_t^e (X_t^{ms})' & X_t^e (X_t^s)' \\ X_t^{ms} (X_t^e)' & X_t^{ms} (X_t^{ms})' & X_t^{ms} (X_t^s)' \\ X_t^s (X_t^e)' & X_t^s (X_t^{ms})' & X_t^s (X_t^s)' \end{bmatrix} \tag{91}$$

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.*,

$$X_t^{mss} = \begin{bmatrix} X_t^{ms} \\ X_t^s \end{bmatrix}$$

follows from Theorem 1. To see this let A_e be a $d_e \times d_e$ matrix. Define

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

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

$$\begin{bmatrix} \eta_t^{ms} \\ \eta_t^s \end{bmatrix} = P_{mss} \eta_t$$

and

$$X_{t+1}^{mss} = \underbrace{\begin{bmatrix} A_{ms} & 0 \\ 0 & A_s \end{bmatrix}}_{A_{mss}} X_t^{mss} + \begin{bmatrix} \eta_{t+1}^{ms} \\ \eta_{t+1}^s \end{bmatrix}$$

Further

$$\mathbb{E}[P_{mss} \eta_t \eta_t' P_{mss}'] = P_{mss} P_{mss}' \succ 0$$

Since all rows of $\tilde{P}$ are independent then $P_{mss} P_{mss}'$ is invertible and $\{P_{mss} \eta_t\}_{t=1}^T$ are independent subGaussian vectors. Now this is the same set up as the general version of Theorem 1 discussed in Section 10. Since $A_{mss} \in \mathcal{S}_0 \cup \mathcal{S}_1$ only has stable and marginally stable components, it follows from the Eq. (51) that

$$\sum_{t=1}^{T} X_t^{mss} (X_t^{mss})' \succeq \frac{T}{4} \sigma_{\min}(P_{mss} P_{mss}') I$$

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

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

$$\beta_0(\delta) = \inf \left\{ \beta | \beta^2 \sigma_{\min}(\Gamma_{\lfloor \frac{1}{\beta} \rfloor}(A)) \geq \Big(\frac{8ec(A, \delta)}{TR^2 \sigma_{\min}(AA')} \Big) \right\}$$

we refer to $\beta_0(\delta)$ as β_0. Following our discussion in Proposition 8.5 we see that $\beta_0 > 0$ and since $\sigma_{\min}(\Gamma_t(A)) \geq \alpha(d) t$ we have that

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

Define

$$V_e = (\sum_{t=1}^{T} X_t^e (X_t^e)'), V_s = \frac{TR^2}{4} I, V_{ms} = \Big(\frac{TR^2}{8e} \Gamma_{\lfloor \frac{1}{\beta_0} \rfloor}(A_{ms}) \Big)$$

where the invertibility in V_e holds with high probability. Observe that $V_{ms} \preceq (\sum_{t=1}^T X_t^{ms}(X_t^{ms})'), V_s \preceq (\sum_{t=1}^T X_t^s (X_t^s)')$ with high probability (follows from Eq. (51),(66)). This observation will be useful in proving the composite invertibility.

Although the technique to prove the invertibility of $\sum_{t=1}^T \tilde{X}_t \tilde{X}_t'$ is similar in spirit to that of (Faradonbeh et al., 2017), it addresses additional difficulties arising due to the presence of a marginally stable block.

$$B_{d \times d} = \begin{bmatrix} V_e^{-1/2} & 0 & 0 \\ 0 & V_{ms}^{-1/2} & 0 \\ 0 & 0 & V_s^{-1/2} \end{bmatrix} \tag{92}$$

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

$$\sum_{t=1}^T B\tilde{X}_t\tilde{X}_t' B' = \begin{bmatrix} I & \sum_{t=1}^T V_e^{-1/2} X_t^e (X_t^{ms})' V_{ms}^{-1/2\prime} & \sum_{t=1}^T V_e^{-1/2} X_t^e (X_t^{s})' V_{s}^{-1/2\prime} \\ \sum_{t=1}^T V_{ms}^{-1/2} X_t^{ms} (X_t^{e})' V_{e}^{-1/2\prime} & \sum_{t=1}^T V_{ms}^{-1/2} X_t^{ms} (X_t^{ms})' V_{ms}^{-1/2\prime} & \sum_{t=1}^T V_{ms}^{-1/2} X_t^{ms} (X_t^{s})' V_{s}^{-1/2\prime} \\ \sum_{t=1}^T V_{s}^{-1/2} X_t^{s} (X_t^{e})' V_{e}^{-1/2\prime} & \sum_{t=1}^T V_{s}^{-1/2} X_t^{s} (X_t^{ms})' V_{ms}^{-1/2\prime} & \sum_{t=1}^T V_{s}^{-1/2} X_t^{s} (X_t^{s})' V_{ms}^{-1/2\prime} \end{bmatrix} \tag{93}$$

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

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

with high probability for some appropriate γ and

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

14.1 Cross Terms have low norm

Define the following quantities:

$$\alpha(A_e,\delta) = \frac{3\phi_{\max}(A_e)^2\sigma^2_{\max}(A_e)}{\phi_{\min}(A_e)^2\sigma_{\min}(A_e)^2}\frac{\Big(1+\frac{1}{c}\log\frac{1}{\delta}\Big)\text{tr}(P_e(\Gamma_T(A_e^{-1}-I))P_e')}{\psi(A_e)^2\delta^2} \tag{94}$$

$$T_{mc}(\delta) = \left\{T \middle| \alpha(A_e,\delta)\text{tr}(A_e^{-T+k_{mc}(T)}(A_e^{-T+k_{mc}(T)})') \leq \frac{\gamma^2}{256}\right\} \tag{95}$$

$$k_{mc} = k_{mc}(T) = T\left(1 - \frac{R^2\gamma^2}{2048de\lambda_1\Big(\Gamma_T(A_{ms})\Gamma^{-1}_{\lfloor\frac{1}{\beta_0(\delta)}\rfloor}(A_{ms})\Big(1+\frac{1}{c}\log\frac{1}{\delta}\Big)\Big)}\right) \tag{96}$$

$$T_{sc}(\delta) = \left\{T \middle| \alpha(A_e,\delta)\text{tr}(A_e^{-T+k_{sc}(T)}(A_e^{-T+k_{sc}(T)})') \leq \frac{\gamma^2}{256}\right\} \tag{97}$$

$$k_{sc} = k_{sc}(T) = T\left(1 - \frac{R^2\gamma^2}{1024d\lambda_1\Big(\Gamma_T(A_{s})\Big(1+\frac{1}{c}\log\frac{1}{\delta}\Big)\Big)}\right) \tag{98}$$

Remark 3. *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

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

We will need a more nuanced argument to upper bound Eq. (99) than that provided in (Faradonbeh et al., 2017) (although it will be similar in flavor).

$$\mathbb{P}(||V_e^{-1/2}\sum_{t=1}^T X_t^e (X_t^{ms})' V_{ms}^{-1/2}||) \tag{99}$$

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

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

and

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

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

$$\begin{aligned}
|v_1' V_e^{-1/2} \sum_{t=k+1}^{T} X_t^e (X_t^{ms})' V_{ms}^{-1/2} v_2| &\leq \underbrace{\sqrt{v_1' V_e^{-1/2} \sum_{t=k+1}^{T} X_t^e (X_t^e)' V_e^{-1/2} v_1}}_{\leq 1} \sqrt{v_2' V_{ms}^{-1/2} \sum_{t=k+1}^{T} X_t^{ms} (X_t^{ms})' V_{ms}^{-1/2} v_2} \\
&\leq \sqrt{v_2' V_{ms}^{-1/2} \sum_{t=k+1}^{T} X_t^{ms} (X_t^{ms})' V_{ms}^{-1/2} v_2} \leq \sqrt{\sigma_1 (V_{ms}^{-1/2} \sum_{t=k+1}^{T} X_t^{ms} (X_t^{ms})' V_{ms}^{-1/2})} \\
&\leq \sqrt{\lambda_1 (\sum_{t=k+1}^{T} X_t^{ms} (X_t^{ms})' V_{ms}^{-1})} \qquad (100)
\end{aligned}$$

To upper bound Eq. (100) we simply need to upper bound $V_{ms}^{-1/2} \sum_{t=k+1}^T X_t^{ms} (X_t^{ms})' V_{ms}^{1/2}$. We can use dependent Hanson–Wright inequality (Corollary 9.1) and Corollary 9.2. Then from Corollary 9.2 and since V_{ms} is deterministic we can conclude that with probability at least $1 - \delta$ we get

$$V_{ms}^{-1/2} \sum_{t=k+1}^{T} X_t^{ms} (X_t^{ms})' V_{ms}^{-1/2} \preceq \sum_{t=k+1}^{T} \text{tr}(V_{ms}^{-1/2} \Gamma_t(A_{ms}) V_{ms}^{-1/2}) \Big(1 + \frac{1}{c} \log \frac{1}{\delta}\Big) I \qquad (101)$$

We can upper bound the deterministic quantity in Eq. (101) as

$$\begin{aligned}
\sum_{t=k+1}^{T} \text{tr}(V_{ms}^{-1/2} \Gamma_t(A) V_{ms}^{-1/2}) &\leq d\lambda_1 (\sum_{t=k+1}^{T} \Gamma_t(A_{ms}) V_{ms}^{-1}) \\
&= d\lambda_1 \Big(\frac{8e}{TR^2} \sum_{t=k+1}^{T} \Gamma_t(A_{ms}) \Gamma_{\lfloor \frac{1}{\beta_0(\delta)} \rfloor}(A_{ms})^{-1} \Big) \\
&\leq d\lambda_1 \Big(\frac{8e(T-k)}{TR^2} \Gamma_T(A_{ms}) \Gamma_{\lfloor \frac{1}{\beta_0(\delta)} \rfloor}(A_{ms})^{-1} \Big) \qquad (102)
\end{aligned}$$

The last inequality holds because the eigenvalues of $P^{-1/2} Q P^{-1/2}$ are the same as QP^{-1} and non–negative whenever P, Q are psd matrices. The normalized gramian term, $\Gamma_t(A_{ms}) \Gamma_{\lfloor \frac{1}{\beta_0(\delta)} \rfloor}(A_{ms})^{-1}$, appears in Eq. (102) only because V_{ms} is deterministic. This will help us in getting non–trivial upper bounds for the cross terms of explosive and marginally stable pair. The key is the choice of k. In Proposition 8.6 we showed that $\lambda_1(\Gamma_{t_1} \Gamma_{t_2}^{-1})$ 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 8.6 the normalized gramian term $\Gamma_T(A_{ms}) \Gamma^{-1}_{\lfloor \frac{1}{\beta_0(\delta)} \rfloor}(A_{ms})$ 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

$$\lambda_1(\Gamma_T(A_{ms}) \Gamma^{-1}_{\lfloor \frac{1}{\beta_0(\delta)} \rfloor}(A_{ms})) = \text{poly}\Big(\log T, \log \frac{1}{\delta} \Big)$$

Our choices of $T_{mc}(\delta), k_{mc}(T)$ in Eq. (95),(96) are motivated by the preceding discussion. We set $k = k_{mc}(T)$ and we have that $d\lambda_1 \Big(\frac{8e(T-k)}{TR^2} \Gamma_T(A_{ms}) \Gamma_{\lfloor \frac{1}{\beta_0(\delta)} \rfloor}(A_{ms})^{-1} \Big) \leq \frac{\gamma^2}{256}$ (check by directly substituting $k = k_{mc}(T)$ in Eq. (102)) and as a result from Eq. (100)

$$|v_1' V_e^{-1/2} \sum_{t=k+1}^{T} X_t^e (X_t^{ms})' V_{ms}^{-1/2} v_2| \leq \frac{\gamma}{16}$$

for arbitrary v_1, v_2. Similarly for the second part

$$|v_1' V_e^{-1/2} \sum_{t=1}^{k} X_t^e (X_t^{ms})' V_{ms}^{-1/2} v_2| \leq \underbrace{\sqrt{v_1' V_e^{-1/2} \sum_{t=1}^{k} X_t^e (X_t^e)' V_e^{-1/2} v_1}}_{a_1} \underbrace{\sqrt{v_2' V_{ms}^{-1/2} \sum_{t=1}^{k} X_t^{ms} (X_t^{ms})' V_{ms}^{-1/2} v_2}}_{\leq 1} \tag{103}$$

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

$$\begin{aligned} a_1 &= \sqrt{v_1' V_e^{-1/2} \sum_{t=1}^{k} X_t^e (X_t^e)' V_e^{-1/2} v_1} \leq \sqrt{\sigma_1 (V_e^{-1/2} \sum_{t=1}^{k} X_t^e (X_t^e)' V_e^{-1/2})} \leq \sqrt{\lambda_1 (\sum_{t=1}^{k} X_t^e (X_t^e)' V_e^{-1})} \\ &\leq \sqrt{\text{tr}(\sum_{t=1}^{k} X_t^e (X_t^e)' V_e^{-1})} \end{aligned} \tag{104}$$

By ensuring that both $T, k = k_{mc}$(which is $\geq T/2) \in T_u(\delta)$ (from Table 1) we have from Eq. (80) that

$$\begin{aligned} \sum_{l=1}^{k} X_t^e (X_t^e)' &\preceq \frac{3\phi_{\max}(A_e)^2}{2\sigma_{\min}(P_e)^2} (1 + \frac{1}{c} \log \frac{1}{\delta}) \text{tr}(P_e (\Gamma_T (A_e^{-1}) - I) P_e') A_e^k A_e^{k\prime} \\ V_e &\succeq \frac{\phi_{\min}(A_e)^2 \psi(A_e)^2 \delta^2}{2\sigma_{\max}(P_e)^2} A_e^T A_e^{T\prime} \end{aligned}$$

Define

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

and we can conclude

$$\sqrt{\text{tr}(\sum_{t=1}^{k} X_t^e (X_t^e)' V_e^{-1})} \leq \sqrt{\alpha(A_e, \delta) \text{tr}(A_e^{-T+k} (A_e^{-T+k})')}$$

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

$$a_1 \leq \sqrt{\alpha(A_e, \delta) \text{tr}(A_e^{-T+k} (A_e^{-T+k})')} \leq \frac{\gamma}{16} \tag{105}$$

with probability at least $1 - 2\delta$. Then combining Eq. (100),(101),(103),(105) we get with probability at least $1 - 4\delta$ that

$$|v_1' V_e^{-1/2} \sum_{t=1}^{T} X_t^e (X_t^{ms})' V_{ms}^{-1/2} v_2| \leq \frac{\gamma}{8} \tag{106}$$

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

$$||V_e^{-1/2} \sum_{t=1}^{T} X_t^e (X_t^{ms})' V_{ms}^{-1/2}|| \leq \frac{\gamma}{8} \tag{107}$$

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

$$||V_e^{-1/2} \sum_{t=1}^{T} X_t^e (X_t^s)' V_s^{-1/2}|| \leq \frac{\gamma}{8} \tag{108}$$

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

Remark 4. *Whenever $T \in T_{sc}(\delta), T_{mc}(\delta)$, the other conditions on T such as $T/2 \in T_u(\delta)$ or $T \geq T_s(\delta) \vee T_{ms}(\frac{\delta}{2T})$ 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.*

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

Now we need to check

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

Since from Theorem 1 and its extension in Section 10 it is known that with probability at least $1-\delta$ we have $\sum_{t=1}^{T} X_t^{mss}(X_t^{mss})' \succeq R^2 \frac{TI}{4}$ 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

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

Since $\sigma_{\min}(M) \geq R^2 \frac{TI}{4}$ then from Corollary 2.3 in (Liu, 2005) we have that

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

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

$$M_{22}^{-1/2}(M/M_{11})M_{22}^{-1/2} = I - M_{22}^{-1/2}QM_{11}^{-1/2}M_{11}^{-1/2}Q'M_{22}^{-1/2} \succeq \frac{R^2}{4\text{tr}(\Gamma_T(A_s))(1+\frac{1}{c}\log\frac{1}{\delta})}I \tag{109}$$

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

$$\omega(\delta) = \frac{R^2}{4\text{tr}(\Gamma_T(A_s))(1+\frac{1}{c}\log\frac{1}{\delta})} > 0 \tag{110}$$

Then this implies that

$$\begin{bmatrix} M_{11}^{-1/2} & 0 \\ 0 & M_{22}^{-1/2} \end{bmatrix} M \begin{bmatrix} M_{11}^{-1/2} & 0 \\ 0 & M_{22}^{-1/2} \end{bmatrix} = \begin{bmatrix} I & M_{11}^{-1/2}Q'M_{22}^{-1/2} \\ M_{22}^{-1/2}QM_{11}^{-1/2} & I \end{bmatrix} \succeq \frac{\omega(\delta)}{4}I$$

because for any $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ we have

$$\begin{aligned} v'\begin{bmatrix} I & \underbrace{M_{11}^{-1/2}QM_{22}^{-1/2}}_{=D'} \\ M_{22}^{-1/2}Q'M_{11}^{-1/2} & I \end{bmatrix} v &= v_1'v_1 + v_1'Dv_2 + v_2'D'v_1 + v_2'v_2 \\ &= v_1'v_1 - 2\sqrt{1-\omega(\delta)}||v_2||||v_1|| + v_2'v_2 \\ &\geq v_1'v_1 - 2\Big(1 - \frac{\omega(\delta)}{2}\Big)||v_2||||v_1|| + v_2'v_2 \end{aligned}$$

Since from Eq. (109) it follows that $||D||^2 \leq 1 - \omega(\delta)$ we obtain

$$\begin{aligned} v_1'v_1 - 2\sqrt{1-\omega(\delta)}||v_2||||v_1|| + v_2'v_2 &= v_1'v_1 - 2\Big(1 - \frac{\omega(\delta)}{2}\Big)||v_2||||v_1|| + v_2'v_2 \\ &= \Big(1 - \frac{\omega(\delta)}{2}\Big)(||v_1|| - ||v_2||)^2 + \Big(1 - \sqrt{1 - \frac{\omega(\delta)}{2}}\Big)(||v_1||^2 + ||v_2||^2) \\ &\geq \Big(\frac{\omega(\delta)}{4}\Big)(||v_1||^2 + ||v_2||^2) \end{aligned}$$

Combining these observations we get

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

We have that

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

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

$$\sigma_{\min}\Bigg(\begin{bmatrix} V_{ms}^{-1/2} & 0 \\ 0 & V_s^{-1/2} \end{bmatrix} \sum_{t=1}^{T} X_t^{mss} (X_t^{mss})' \begin{bmatrix} V_{ms}^{-1/2} & 0 \\ 0 & V_s^{-1/2} \end{bmatrix} \Bigg) \geq \Big(\frac{\omega(\delta)}{4}\Big) > 0 \tag{111}$$

Now we replace in Eq. (107),(108) $\gamma \to \frac{\sqrt{\omega(\delta)}}{32}$. Then that implies

$$|| \, V_e^{-1/2} \sum_{t=1}^{T} X_t^e (X_t^s)' V_s^{-1/2} || \geq \frac{\sqrt{\omega(\delta)}}{64}$$

$$|| \, V_e^{-1/2} \sum_{t=1}^{T} X_t^e (X_t^{ms})' V_{ms}^{-1/2} || \geq \frac{\sqrt{\omega(\delta)}}{64}$$

14.3 Lower Bound on $\sum_{t=1}^T \tilde{X}_t \tilde{X}_t'$

Recalling that

$$\sum_{t=1}^{T} B \tilde{X}_t \tilde{X}_t' B' = \begin{bmatrix} I & \sum_{t=1}^T V_e^{-1/2} X_t^e (X_t^{ms})' V_{ms}^{-1/2\prime} & \sum_{t=1}^T V_e^{-1/2} X_t^e (X_t^{s})' V_{s}^{-1/2\prime} \\ \sum_{t=1}^T V_{ms}^{-1/2} X_t^{ms} (X_t^{e})' V_{e}^{-1/2\prime} & \sum_{t=1}^T V_{ms}^{-1/2} X_t^{ms} (X_t^{ms})' V_{ms}^{-1/2\prime} & \sum_{t=1}^T V_{ms}^{-1/2} X_t^{ms} (X_t^{s})' V_{s}^{-1/2\prime} \\ \sum_{t=1}^T V_{s}^{-1/2} X_t^{s} (X_t^{e})' V_{e}^{-1/2\prime} & \sum_{t=1}^T V_{s}^{-1/2} X_t^{s} (X_t^{ms})' V_{ms}^{-1/2\prime} & \sum_{t=1}^T V_{s}^{-1/2} X_t^{s} (X_t^{s})' V_{ms}^{-1/2\prime} \end{bmatrix}$$

then it follows from Eq. (111) that

$$\sum_{t=1}^{T} B \tilde{X}_t \tilde{X}_t' B' \succeq \begin{bmatrix} I & \sum_{t=1}^T V_e^{-1/2} X_t^e (X_t^{ms})' V_{ms}^{-1/2\prime} & \sum_{t=1}^T V_e^{-1/2} X_t^e (X_t^{s})' V_{s}^{-1/2\prime} \\ \sum_{t=1}^T V_{ms}^{-1/2} X_t^{ms} (X_t^{e})' V_{e}^{-1/2\prime} & \frac{\omega(\delta)}{4} I & 0 \\ \sum_{t=1}^T V_{s}^{-1/2} X_t^{s} (X_t^{e})' V_{e}^{-1/2\prime} & 0 & \frac{\omega(\delta)}{4} I \end{bmatrix}$$

Let $v = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$ Then $v' \sum_{t=1}^T B \tilde{X}_t \tilde{X}_t' B' v = ||v_1||^2 + \frac{\omega(\delta)}{4}(||v_2||_2^2 + ||v_3||_2^2) + 2 v_1' \sum_{t=1}^T V_e^{-1/2} X_t^e (X_t^{ms})' V_{ms}^{-1/2\prime} v_2 + 2 v_1' \sum_{t=1}^T V_e^{-1/2} X_t^e (X_t^{s})' V_{s}^{-1/2\prime} v_3 \geq ||v_1||^2 + \frac{\omega(\delta)}{4}(||v_2||_2^2 + ||v_3||_2^2) - \frac{\sqrt{\omega(\delta)}}{32} ||v_1|| ||v_2|| - \frac{\sqrt{\omega(\delta)}}{32} ||v_1|| ||v_3||$. Then we get

$$v' \sum_{t=1}^{T} B \tilde{X}_t \tilde{X}_t' B' v \geq ||v_1||^2 + \frac{\omega(\delta)}{4}(||v_2||_2^2 + ||v_3||_2^2) - \frac{\omega(\delta)}{64}(||v_1||^2 + ||v_2||^2) - \frac{\omega(\delta)}{64}(||v_1||^2 + ||v_3||^2)$$

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 $\sum_{t=1}^T X_t^{mss} (X_t^{mss})'$ and $\sum_{t=1}^T X_t^e (X_t^e)', \sum_{t=1}^T X_t^s (X_t^s)', \sum_{t=1}^T X_t^{ms} (X_t^{ms})'$.

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

whenever

$$T \in T_{mc}(\delta) \cap T_{sc}(\delta) \tag{112}$$

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

$$V^e_{dn}(\delta) = \frac{\phi_{\min}(A_e)^2\psi(A_e)^2\delta^2}{2\sigma_{\max}(P)^2}A_e^T A_e^{T\prime}, V^s_{dn}(\delta) = \frac{TR^2}{4}I, V^{ms}_{dn}(\delta) = \Big(\frac{TR^2}{8e}\Gamma_{\lfloor \frac{1}{\beta_0(\delta)}\rfloor}(A_{ms})\Big)$$

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

$$\begin{aligned}\sum_{t=1}^{T} B\tilde{X}_t\tilde{X}_t'B' \succeq \frac{\omega(\frac{\delta}{C})}{8}I &\implies \sum_{t=1}^{T}\tilde{X}_t\tilde{X}_t' \succeq \frac{\omega(\frac{\delta}{C})}{8}B^{-2} \\ \sum_{t=1}^{T}\tilde{X}_t\tilde{X}_t' \succeq & \underbrace{\frac{\omega(\frac{\delta}{C})}{8}\begin{bmatrix} V^e_{dn}(\delta) & 0 & 0 \\ 0 & V^{ms}_{dn}(\frac{\delta}{C}) & 0 \\ 0 & 0 & V^s_{dn}(\frac{\delta}{C})\end{bmatrix}}_{=V_{dn}}\end{aligned} \tag{113}$$

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

14.4 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 (Simchowitz et al., 2018). Consider an arbitrary matrix $M = \begin{bmatrix} M_1 \\ M_2 \\ M_3 \end{bmatrix}$. Then $\begin{bmatrix} 3M_1M_1' & 0 & 0 \\ 0 & 3M_2M_2' & 0 \\ 0 & 0 & 3M_3M_3'\end{bmatrix} \succeq MM'$. This is because

$$\begin{aligned}\begin{bmatrix} 2M_1M_1' & -M_1M_2' & -M_1M_3' \\ -M_2M_1' & 2M_2M_2' & -M_2M_3' \\ -M_3M_1' & -M_3M_2' & 2M_3M_3'\end{bmatrix} &= (\begin{bmatrix} M_1 \\ 0 \\ 0\end{bmatrix} - \begin{bmatrix} 0 \\ M_2 \\ 0\end{bmatrix})(\begin{bmatrix} M_1 \\ 0 \\ 0\end{bmatrix} - \begin{bmatrix} 0 \\ M_2 \\ 0\end{bmatrix})' \\ &+ (\begin{bmatrix} M_1 \\ 0 \\ 0\end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ M_3\end{bmatrix})(\begin{bmatrix} M_1 \\ 0 \\ 0\end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ M_3\end{bmatrix})' + (\begin{bmatrix} 0 \\ 0 \\ M_3\end{bmatrix} - \begin{bmatrix} 0 \\ M_2 \\ 0\end{bmatrix})(\begin{bmatrix} 0 \\ 0 \\ M_3\end{bmatrix} - \begin{bmatrix} 0 \\ M_2 \\ 0\end{bmatrix})'\end{aligned}$$

Define

$$\begin{aligned}V^e_{up}(\delta) &= \frac{3\phi_{\max}(A)^2\sigma_{\max}(\tilde{P})^4}{\sigma_{\min}(\tilde{P})^2}(1+\frac{1}{c}\log\frac{1}{\delta})\text{tr}(\Gamma_T(A_e^{-1}))A_e^TA_e^{T\prime} \\ V^s_{up}(\delta) &= 3\sigma_{\max}(\tilde{P})^2T\text{tr}(\Gamma_T(A_s))\Big(1+\frac{1}{c}\log\Big(\frac{1}{\delta}\Big)\Big)I \\ V^{ms}_{up}(\delta) &= 3\sigma_{\max}(\tilde{P})^2T\text{tr}(\Gamma_T(A_{ms}))\Big(1+\frac{1}{c}\log\Big(\frac{1}{\delta}\Big)\Big)I\end{aligned}$$

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

$$\begin{bmatrix} \sum_{t=1}^T X^e(X_t^e)' & 0 & 0 \\ 0 & \sum_{t=1}^T X^{ms}(X_t^{ms})' & 0 \\ 0 & 0 & \sum_{t=1}^T X^s(X_t^s)'\end{bmatrix} \preceq \begin{bmatrix} V^e_{up}(\delta) & 0 & 0 \\ 0 & V^{ms}_{up}(\delta) & 0 \\ 0 & 0 & V^s_{up}(\delta)\end{bmatrix}$$

We get these upper bounds for stable and marginally stable matrices from Proposition (9.4) and Eq. (80) for explosive matrices. Then with probability at least $1-4\delta$ we have

$$\sum_{t=1}^{T}\tilde{X}_t\tilde{X}_t^{\prime} \preceq \underbrace{\begin{bmatrix}3V_{up}^{e}(\delta) & 0 & 0\\ 0 & 3V_{up}^{ms}(\delta) & 0\\ 0 & 0 & 3V_{up}^{s}(\delta)\end{bmatrix}}_{=V_{up}} \tag{114}$$

Note that the time requirement in Eq. (112) is sufficient to ensure the upper bounds with high probability and we do not state them explicitly.

14.5 Getting Error Bounds

We recall the discussion for Theorem 1. We have V_{up}, V_{dn}, so we compute $V_{up}V_{dn}^{-1}$ which gives us

$$V_{up}V_{dn}^{-1} = \frac{8}{\omega(\frac{\delta}{C})}\begin{bmatrix}3V_{up}^{e}(\delta)(V_{dn}^{e}(\delta))^{-1} & 0 & 0\\ 0 & 3V_{up}^{ms}(\delta)(V_{dn}^{ms}(\frac{\delta}{C}))^{-1} & 0\\ 0 & 0 & 3V_{up}^{s}(\delta)(V_{dn}^{s})^{-1}(\frac{\delta}{C})\end{bmatrix}$$
$$\text{det}(V_{up}V_{dn}^{-1}) = \Big(\frac{24}{\omega(\frac{\delta}{C})}\Big)^{d}\text{det}(V_{up}^{e}(\delta)(V_{dn}^{e}(\delta))^{-1})\text{det}(V_{up}^{ms}(\delta)(V_{dn}^{ms}(\frac{\delta}{C}))^{-1})\text{det}(V_{up}^{s}(\delta)(V_{dn}^{s}(\frac{\delta}{C}))^{-1})$$

Further $V_{dn}^{s}(\frac{\delta}{C}) = V_{dn}^{s}(\delta)$ (only the time required to be greater than this with high probability changes). Then

$$\begin{aligned}\log\left(\text{det}(V_{up}V_{dn}^{-1})\right) &= d(\log 24 - \log\omega(\frac{\delta}{C})) + \log\det(V_{up}^{e}(\delta)(V_{dn}^{e}(\delta))^{-1})\\ &\quad + \log\det(V_{up}^{ms}(\delta)(V_{dn}^{ms}(\frac{\delta}{C}))^{-1}) + \log\det(V_{up}^{s}(\delta)(V_{dn}^{s}(\frac{\delta}{C}))^{-1})\end{aligned}$$

Following this the bounds are straightforward and can be computed as shown in Eq. (12). It should be noted that Proposition 3.1 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.*,

$$\sum_{t=1}^{T}(\sum_{t=1}^{T}\tilde{X}_t\tilde{X}_t)^{+}(\sum_{t=1}^{T}\tilde{X}_t\tilde{\eta}_{t+1})$$

Since $(\sum_{t=1}^{T}\tilde{X}_t\tilde{X}_t)$ is invertible with high probability

$$\begin{aligned}\sum_{t=1}^{T}(\sum_{t=1}^{T}\tilde{X}_t\tilde{X}_t)^{+}(\sum_{t=1}^{T}\tilde{X}_t\tilde{\eta}_{t+1}) &= (\sum_{t=1}^{T}\tilde{X}_t\tilde{X}_t)^{-1}(\sum_{t=1}^{T}\tilde{X}_t\tilde{\eta}_{t+1})\\ &= \sum_{t=1}^{T}\tilde{P}^{-1\prime}(\sum_{t=1}^{T}X_tX_t)^{-1}\tilde{P}^{-1}\tilde{P}X_t\eta_{t+1}\tilde{P}^{\prime}\\ &= \tilde{P}^{-1\prime}\sum_{t=1}^{T}(\sum_{t=1}^{T}X_tX_t)^{-1}X_t\eta_{t+1}\tilde{P}^{\prime}\end{aligned}$$

Then it is clear that

$$\Bigg|\Bigg|\sum_{t=1}^{T}(\sum_{t=1}^{T}\tilde{X}_t\tilde{X}_t)^{-1}(\sum_{t=1}^{T}\tilde{X}_t\tilde{\eta}_{t+1})\Bigg|\Bigg| \geq \sigma_{\min}(\tilde{P}^{-1})\Bigg|\Bigg|\sum_{t=1}^{T}(\sum_{t=1}^{T}X_tX_t)^{-1}X_t\eta_{t+1}\Bigg|\Bigg|\sigma_{\min}(\tilde{P})$$

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.

15 Extension to presence of control input

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

$$X_{t+1} = AX_t + BU_t + \eta_{t+1} \tag{115}$$

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

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

Since

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

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

16 Extension to heavy tailed noise

It is claimed in (Faradonbeh et al., 2017) 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 (Faradonbeh et al., 2017) 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. We consider the class of distributions introduced in (Faradonbeh et al., 2017) called sub–Weibull distribution. Let $\eta_{t,i}$ be the i^{th} element of η_t then $\eta_{t,i}$ has sub–Weibull distribution if

$$\mathbb{P}(|\eta_{t,i} > y|) \leq b \exp\left\{ \left(\frac{-y^{\alpha}}{m} \right) \right\} \tag{116}$$

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

$$\mathcal{W}(\delta) = \left\{ \max_{1 \leq t \leq T} ||\eta_t||_{\infty} \leq \nu_T(\delta) \right\}$$

where $\nu_T(\delta) = \left(m \log\left(\frac{bTd}{\delta}\right)^{1/\alpha} \right)$. Then Proposition 3 in (Faradonbeh et al., 2017) shows that $\mathbb{P}(\mathcal{W}(\delta)) \geq 1 - \delta$. Clearly because each $\{\eta_{t,i}\}_{t=1,i=1}^{t=T,i=d}$ are i.i.d and have symmetric distribution

$$\mathbb{E}[\eta_{t,i}|\mathcal{W}(\delta)] = \mathbb{E}[\eta_{t,i}|\{|\eta_{t,i}| \leq \nu_T(\delta)\}] = 0 \tag{117}$$

Then under $\mathcal{W}(\delta)$, $\eta_{t,i}$ has mean zero and $\{\eta_{t,i}\}_{t=1,i=1}^{t=T,i=d}$ are independent under the event $\mathcal{W}(\delta)$. Further since under $\mathcal{W}(\delta)$ 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 3.1. We can now use all our techniques from before.

17 Optimality of Bound

In this section we show that the upper bound for explosive systems in Theorem 1 is optimal. To that end, we analyze a 1-D system. By explicitly calculating the propbability distribution of the error term we provide an (almost) optimal lower bound.

Let $A = a$ be 1-D system. Assume that $T \in T_u(\delta)$ (as in Table 1). Then X_t, η_t are just numbers. Then let E be the error, *i.e.*,

$$\begin{aligned} E &= (\sum_{t=1}^{T} x_t^2)^{-1} (\sum_{t=1}^{T} x_t \eta_{t+1}) \\ &= a^{-T} (\sum_{t=1}^{T} a^{-2T} x_t^2)^{-1} (\sum_{t=1}^{T} a^{-T} x_t \eta_{t+1}) \end{aligned}$$

In this section, we will show that the bound obtained for explosive systems is optimal in terms of δ. Assume $\eta_t \sim \boldsymbol{\mathcal{N}}(0,1)$ 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 $E = a^{-T} U_T^{-1} S_T$ and S_T has the following form

$$2S_T = [\eta_{T+1}, \ldots, \eta_1] \underbrace{\begin{bmatrix} 0 & a^{-T} & a^{-T+1} & \ldots & a^{-1} \\ a^{-T} & 0 & a^{-T} & \ldots & a^{-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots \\ a^{-1} & a^{-2} & a^{-3} & \ldots & 0 \end{bmatrix}}_{=M} \underbrace{\begin{bmatrix} \eta_{T+1} \\ \vdots \\ \eta_1 \end{bmatrix}}_{=\tilde{\eta}} \tag{118}$$

Define $F_T = \sum_{i=1}^T a^{-2i+2}(a^{-2T} x_T^2) = \frac{1-a^{-2T}}{1-a^{-2}} a^{-2T} x_T^2$. 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 12 when $A = a$. We can easily calculate σ^2

$$a^{-2} \leq \sigma^2 \leq \frac{1}{a^2 - 1}$$

Consider four events

$$\mathcal{E}_1(\delta) = \left\{ |U_T - F_T| \leq \frac{\delta^2 \sigma^2}{C} \vee \Big(\frac{CT^2 a^{-T}}{1-a^{-2}} + \Big(1 + \frac{1}{c} \log \frac{1}{\delta}\Big) \frac{Ta^{-2T}}{(1-a^{-2})} \Big) \right\}, \mathcal{E}_2(\delta) = \left\{ |S_T| \geq \frac{\delta}{-Ca^2 \log \delta} \right\}$$

$$\mathcal{E}_3(\delta) = \left\{ 0 \leq F_T \leq C_2 \delta^2 \sigma^2 \right\}, \mathcal{E}_4(\delta) = \left\{ 0 \leq U_T \leq \Big((C_2 + 1/C) \delta^2 \sigma^2 \Big) \vee \Big(\frac{CT^2 a^{-T}}{1-a^{-2}} + \Big(1 + \frac{1}{c} \log \frac{1}{\delta}\Big) \frac{Ta^{-2T}}{(1-a^{-2})} \Big) \right\}$$

From Eq. (77) we have with probability at least $1 - \frac{\delta}{2}$ that

$$\begin{aligned} ||U_T - F_T||_2 &\underbrace{\leq}_{\text{Eq. (77)}} \Big(4T^2 \sigma_1^2 (A^{-\frac{(T+1)}{2}}) \text{tr}(\Gamma_T(A^{-1})) + \Big(T + \frac{T}{c} \log \frac{1}{\delta}\Big) \sigma_1^2 (A^{-T-1}) \text{tr}(\Gamma_T(A^{-1})) \Big) \\ &\leq \frac{4T^2 a^{-T}}{1-a^{-2}} + \Big(1 + \frac{1}{c} \log \frac{1}{\delta}\Big) \frac{Ta^{-2T}}{(1-a^{-2})} \end{aligned}$$

Assume $\delta^2 \in (0, \frac{1}{128}]$ then

$$\begin{aligned} \mathbb{P}(\mathcal{E}_3(\delta)) &= \frac{2}{\sqrt{2\pi}\sigma} \int_{2\delta\sigma}^{16\delta\sigma} e^{-\frac{x^2}{2\sigma^2}} dx \\ &\geq \frac{14\delta}{\sqrt{2\pi}} e^{-\frac{256\delta^2}{2}} \\ &\geq \frac{14\delta}{\sqrt{2\pi e}} \geq 2\delta \end{aligned}$$

Recall $T_u(\delta)$ is the set of T that satisfies Eq. (81) when $A = a$.

17.1 $T \in T_u(\delta)$

For $T \in T_u(\delta)$ and from Eq. (76), we have with probability at least $1 - \frac{\delta}{2}$ that

$$||U_T - F_T||_2 \leq \frac{4T^2 a^{-T}}{1-a^{-2}} + \frac{Ta^{-2T}}{\delta(1-a^{-2})} \underbrace{\leq}_{T \in T_u(\delta), \text{Eq. (81)}} \frac{\phi_{\min}(a)^2 \psi(a)^2 \delta^2}{2\sigma_{\max}(P)^2} \leq \frac{C\delta^2}{(a^2-1)}$$

The last inequality follows because for 1-D systems $\phi_{\min}(A), \psi(A), \sigma_{\max}(P)$ are just constants, for example $P = 1, \phi_{\min}(a) = 1, \psi(a)^2 = C\sigma^2 \leq \frac{C}{a^2-1}$ which follows by definition. Note $T \in T_u(\delta)$ if and only if we have

$$\delta^2 \sigma^2 > \frac{CT^2 a^{-T}}{1-a^{-2}}$$

Thus, $\mathbb{P}(\mathcal{E}_1(\delta)) \geq 1 - \frac{\delta}{2}$. Clearly $\mathcal{E}_1(\delta) \cap \mathcal{E}_3(\delta) \implies \mathcal{E}_1(\delta) \cap \mathcal{E}_4(\delta)$ and

$$\mathcal{E}_2(\delta) \cap \mathcal{E}_4(\delta) \implies \Big\{|S_T|U_T^{-1} \geq \frac{C}{-\sigma^2 a^2 \delta \log \delta}\Big\}$$

We bound $\mathbb{P}(\mathcal{E}_2(\delta))$ in Section 18 and Eq. (121), which gives $\mathbb{P}(\mathcal{E}_2(\delta)) \geq 1 - \frac{\delta}{2}$ and then

$$\begin{aligned}\mathbb{P}(\mathcal{E}_1(\delta) \cap \mathcal{E}_2(\delta) \cap \mathcal{E}_4(\delta)) &\geq \mathbb{P}(\mathcal{E}_1(\delta) \cap \mathcal{E}_2(\delta) \cap \mathcal{E}_3(\delta)) \\ &\geq \mathbb{P}(\mathcal{E}_1(\delta)) + \mathbb{P}(\mathcal{E}_2(\delta) \cap \mathcal{E}_3(\delta)) - 1 \\ &\geq \mathbb{P}(\mathcal{E}_1(\delta)) + \mathbb{P}(\mathcal{E}_2(\delta)) + \mathbb{P}(\mathcal{E}_3(\delta)) - 2 \\ &\geq \frac{\delta}{2}\end{aligned}$$

Since $\mathcal{E}_2(\delta) \cap \mathcal{E}_4(\delta) \implies \{|S_T|U_T^{-1} \geq \frac{C}{-\sigma^2 a^2 \delta \log \delta}\}$ when $T \in T_u(\delta)$ then

$$\mathbb{P}(\{|S_T|U_T^{-1} \geq \frac{C}{-\sigma^2 a^2 \delta \log \delta}\}) \geq \frac{\delta}{2}$$

we have proved our claim that with probability at least δ we have that

$$|E_T| \geq \Big(\frac{C}{-\sigma^2 a^2 \delta \log \delta}\Big) a^{-T} \geq \frac{C(1-a^{-2})}{-\delta \log \delta} a^{-T} \tag{119}$$

whenever $Ca^2T^2a^{-T} \leq \delta^2$.

17.2 $T \not\in T_u(\delta)$

If $Ca^2T^2a^{-T} > \delta^2$, then with probability at least $1 - \frac{\delta}{2}$

$$|U_T - F_T| \leq \underbrace{\frac{CT^2 a^{-T}}{1 - a^{-2}}}_{\text{Follows by direct computation}}$$

and we have with probability at least δ that

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

and we can conclude with probability at least δ

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

where $Ca^2T^2a^{-T} \geq \delta^2 \implies T \leq -\log \delta$.

17.3 Comparison to existing bounds

Theorem 5 (Theorem B.2 (Simchowitz et al., 2018))**.** *Fix an $a_* \in \mathbb{R}$ and define $\Gamma_T = \sum_{t=1} a_*^{2t}$. Fix an alternative $a' \in \{a_* - 2\epsilon, a_* + 2\epsilon\}$ and $\delta \in (0, 1/4)$. Then for any estimator $\hat{a}$*

$$\sup_{a \in \{a_*, a'\}} \mathbb{P}(|\hat{a}(T) - a_*| \geq \epsilon) \geq \delta$$

for any T such that $T\Gamma_T \leq \frac{\log(1/2\delta)}{8\epsilon^2}$.

Note $\Gamma_T = \frac{a^{2T+2}-1}{a^2-1}$. Theorem 5 suggests that for a given T, δ if $\epsilon \leq a^{-T}\sqrt{\frac{-C \log \delta}{T}}$ then $\mathbb{P}(|a_* - \hat{a}(T)| \geq \epsilon) \geq \delta$. However we show that whenever $Ca^2T^2a^{-T} \leq \delta^2$, we have that

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

Since $a^{-T}\sqrt{\frac{-C \log \delta}{T}} \leq a^{-T}\frac{C(1-a^{-2})}{-\delta \log \delta}$ our lower bound is tighter.

Theorem 6 (Theorem B.1 (Simchowitz et al., 2018))**.** *Let $\epsilon \in (0,1)$ and $\delta \in (0, 1/2)$. Then $\mathbb{P}(|\hat{a}(T) - a_*|\leq \epsilon) \geq 1-\delta$ as long as*

$$T \geq \max\Big\{\frac{8}{(|a_* - \epsilon|)^2 - 1}\log\frac{2}{\delta}, \frac{4\log\frac{1}{\epsilon}}{\log{(|a_*| - \epsilon)}} + 8\log\frac{2}{\delta}\Big\}$$

We now compare Eq. (119) to the upper bound in Theorem 6. Eq. (119) gives us that if

$$\epsilon \leq \frac{C(1-a^{-2})}{-\delta\log\delta}a^{-T}$$

we have with probability at least δ that $|E_T|\geq \epsilon$. This reduces to whenever

$$T_{-} \leq \frac{\log\frac{1}{\epsilon}}{\log a} + \frac{\log\frac{C(1-a^{-2})}{\delta}}{\log a} \tag{120}$$

we have with probability at least δ that $|E_T|\geq \epsilon$. We focus on the case $a_* > 1 + \epsilon$ of Theorem 6. Let $a_* = 1+\epsilon+\gamma$, then the bounds in Theorem 6 indicate that whenever

$$T_{+} \geq \frac{8}{2\gamma + \gamma^2}\log\frac{2}{\delta} + \frac{4\log\frac{1}{\epsilon}}{\log{(\gamma+1)}} + \log\frac{2}{\delta}$$

we have with probability at least $1-\delta$ $|E_T|\leq \epsilon$. If $\gamma = o(\epsilon)$, then the requirement on T reduces to

$$T_{+} \geq \frac{8}{o(\epsilon)}\log\frac{2}{\delta} + \frac{4\log\frac{1}{\epsilon}}{o(\epsilon)} + \text{ smaller terms}$$

By substituting $\log a \approx \epsilon$ in T_{-} we note that $T_{-} \leq T_{+}$. For the case when $\gamma = \Omega(\epsilon)$ for T_{+} we get

$$T_{+} \geq \Big(\frac{8}{\Omega(\epsilon)} \vee 1\Big)\log\frac{2}{\delta} + \frac{4\log\frac{1}{\epsilon}}{\log{(1+\Omega(\epsilon))}} \approx \underbrace{\Big(\frac{8}{\Omega(\epsilon)} \vee 1\Big)}_{\geq (\log a)^{-1}}\log\frac{2}{\delta} + \frac{2\log\frac{1}{\epsilon}}{\log a}$$

In either cases $T_{-} \leq T_{+}$.

18 Distribution of S_T

Recall S_T from Eq. (118). Since $\sum_{i,j}|M|_{i,j}\geq ||M||_*$ (the nuclear norm), we have that $||M||_*\leq \frac{2a^{-1}}{1-a^{-1}}$ and it is obvious that $||M||_2\geq a^{-1}$. Since $M = U^{\top}\Lambda U$ (because it is symmetric) and η_t are i.i.d Gaussian then $U\tilde{\eta}$ is also Gaussian with each of its entries being i.i.d Gaussian. This implies that $2S_T = \sum_{j=1}^{T+1}\lambda_j g_j^2$ where λ_j are eigenvalues of M and g_j are i.i.d Gaussian with $\sum_j \lambda_j = 0, \sum_j|\lambda_j|\leq \frac{2a^{-1}}{1-a^{-1}}$. The characteristic function of S_T is

$$\phi_{S_T}(t) = \prod_{j=1}^{T+1}\Big(\frac{1}{1-2it\lambda_j}\Big)^{1/2} = \Big(\frac{1}{1-4t^2(\sum_{l\neq j}\lambda_l\lambda_j) - i8t^3(\sum_{l\neq j\neq k}\lambda_l\lambda_j\lambda_k) + 16t^4(\sum_{l\neq j\neq k\neq p}\lambda_l\lambda_j\lambda_k\lambda_p)\ldots}\Big)^{1/2}$$

where the coefficient of t vanishes because $\sum_{j=1}^{T+1}\lambda_j = 0$. Further since $\sum_{l\neq j}2\lambda_l\lambda_j = -\sum_j\lambda_j^2$ we have and

$$\begin{aligned}(\sum_{l\neq j\neq k\neq m}\lambda_l\lambda_j\lambda_k\lambda_m) &= \sum_l\lambda_l(\sum_{l\neq j\neq k\neq m}\lambda_j\lambda_k\lambda_m) = \sum_l\lambda_l(\sum_{l\neq j\neq k\neq m}\lambda_j\lambda_k\lambda_m + \sum_{l\neq p\neq m}\lambda_l\lambda_p\lambda_m - \sum_{l\neq p\neq m}\lambda_l\lambda_p\lambda_m) \\ &= \sum_l\lambda_l(\sum_{j\neq k\neq m}\lambda_j\lambda_k\lambda_m - \sum_{l\neq p\neq m}\lambda_l\lambda_p\lambda_m - \sum_{l\neq m}\lambda_l^2\lambda_m + \sum_{l\neq m}\lambda_l^2\lambda_m) \\ &= \sum_l\lambda_l(-\lambda_l\sum_{p\neq m}\lambda_p\lambda_m + \sum_{l\neq m}\lambda_l^2\lambda_m) = \frac{(\sum_l\lambda_l^2)^2}{2} - \sum_l\lambda_l^4 = \frac{\text{tr}(M^2)^2}{2} - \text{tr}(M^4)\end{aligned}$$

The coefficients of even powers of t can be obtained in a similar fashion. Then recall by Levy's theorem that

$$f_{S_T}(x) = \int_{-\infty}^{\infty} e^{-itx}\phi_{S_T}(t)dt \implies \sup_x f_{S_T}(x) \leq \int_{-\infty}^{\infty} |\phi_{S_T}(t)|dt \leq \int_{-\infty}^{\infty} \frac{1}{\sqrt{1+c_1t^2+c_2t^4+\ldots}}dt$$

Now whenever $c_k > 0$ (and not decaying asymptotically to zero) for some $k \geq 2$, we get $\sup_x f_{S_T}(x) \leq C$ for some universal constant C and we can use Proposition 9.5 to get $\mathbb{P}(|S_T|\leq \delta) \leq C\delta$. But since that may not be always be true we can explicitly calculate the integral

$$f_{S_T}(x) = \int_{-\infty}^{\infty} e^{-itx}\phi_{S_T}(t)dt \approx \underbrace{\int_{-\infty}^{\infty} \frac{e^{itx}}{\sqrt{1+2a^{-2}t^2}}dt}_{\text{Modified Bessel Function of the Second Kind}}$$

$$\begin{aligned}\int_{-\delta}^{\delta} f_{S_T}(x)dx &= \int_{-\delta}^{\delta}\int_{-\infty}^{\infty} \frac{e^{itx}}{\sqrt{1+2a^{-2}t^2}}dtdx = 2\int_{-\infty}^{\infty}\int_{-\delta}^{\delta} \frac{\cos(tx)}{\sqrt{1+2a^{-2}t^2}}dxdt \\ &= C\delta\int_{0}^{\infty} \frac{\sin(t\delta)}{\delta t\sqrt{1+2a^{-2}t^2}}dt = C\delta\int_{0}^{\delta} \frac{\sin(t\delta)}{\delta t\sqrt{1+2a^{-2}t^2}}dt + C\delta\int_{\delta}^{\infty} \frac{\sin(t\delta)}{\delta t\sqrt{1+2a^{-2}t^2}}dt \\ &\leq C\delta^2 - Ca\delta\log(\delta)\end{aligned}$$

Thus

$$\mathbb{P}(|S_T|\leq \delta) \leq -Ca\delta\log\delta$$

and replacing $\delta \to \frac{-C\delta}{2a\log\delta}$ we get

$$\mathbb{P}\Big(|S_T|\leq \frac{-C\delta}{a\log\delta}\Big) \leq \frac{\delta}{2} \tag{121}$$

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

20 Inconsistency of explosive systems

In this section we provide proof that OLS is inconsistent when the regularity assumption of explosive matrices is violated. In fact we show that even the simple scale identity matrix cannot be correctly learned. The proof proceeds by analyzing the scaled sample covariance matrix $a^{-2T}Y_T$. Using tools from matrix analysis, we show that the error term does not decay to zero as $T \to \infty$

$$\hat{A}_o - A_o = \Big(\sum_{t=1}^{T} a^{-2T}\eta_{t+1}X_t^{\prime}\Big) \underbrace{(a^{-2T}\sum_{t=1}^{T}X_tX_t^{\top})^{-1}}_{=\text{Scaled Sample Covariance matrix}}$$

The key insight in the result is that although $\Big(\sum_{t=1}^{T} a^{-2T}\eta_{t+1}X_t^{\prime}\Big)$ decays as $O(a^{-T})$, $(a^{-2T}\sum_{t=1}^{T}X_tX_t^{\top})$ has a singular value $o(a^{-T})$ due to which the error is a non-decaying. Let $A = aI$ where $a \geq 1.1$ and η_t are i.i.d. Gaussian random vectors. Then

Proposition 20.1. *Let $\{\eta_t\}_{t=1}^{T}$ be i.i.d standard Gaussian then whenever $T^2 \leq a^T$, we have that*

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

where γ_T is a random variable that admits a continuous pdf and does not decay to zero as $T \to \infty$.

Proof.

$$\begin{bmatrix} X^{(1)}_{t+1} \\ X^{(2)}_{t+1} \end{bmatrix} = A \begin{bmatrix} X^{(1)}_{t} \\ X^{(2)}_{t} \end{bmatrix} + \begin{bmatrix} \eta^{(1)}_{t+1} \\ \eta^{(2)}_{t+1} \end{bmatrix}$$

Since A is scaled identity we have that $X_t^{(1)} = \sum_{t=1}^{T} a^{T-t}\eta_t^{(1)}, X_t^{(2)} = \sum_{t=1}^{T} a^{T-t}\eta_t^{(2)}$. The scaled sample covariance matrix $a^{-2T}Y_T = a^{-2T}\sum_{t=1}^{T} X_t X_t^{\top}$ is of the following form

$$a^{-2T}Y_T = \begin{bmatrix} a^{-2T}\sum_{t=1}^{T}(X_t^{(1)})^2 & a^{-2T}\sum_{t=1}^{T} X_t^{(1)}X_t^{(2)} \\ a^{-2T}\sum_{t=1}^{T} X_t^{(1)}X_t^{(2)} & a^{-2T}\sum_{t=1}^{T}(X_t^{(2)})^2 \end{bmatrix} \tag{122}$$

Define $a^{-T}X_T = Z_T$ with $Z_T^{(i)}$ corresponding to appropriate coordinates, and recall that $Z_T^{(i)}$ is a Gaussian random variable with variance in $(a^{-2}, \frac{a^{-2}}{1-a^{-2}})$ and each $a^{-T}X_t = \langle a^{-T}X_t, Z_T\rangle Z_T + \langle a^{-T}X_t, Z_T^{\perp}\rangle Z_T^{\perp}$. This implies

$$\begin{aligned}
a^{-2T}\sum_{t=1}^{T} X_t X_t^{\top} &= \sum_{t=1}^{T}(\underbrace{a^{-T}\langle X_t, Z_T\rangle}_{=\alpha_t})^2 Z_T Z_T^{\top} + \sum_{t=1}^{T} a^{-2T}\langle X_t, Z_T\rangle\langle X_t, Z_T^{\perp}\rangle Z_T (Z_T^{\perp})^{\top} \\
&+ \sum_{t=1}^{T}\underbrace{\langle a^{-T}X_t, Z_T\rangle}_{=\alpha_t}\underbrace{\langle a^{-T}X_t, Z_T^{\perp}\rangle}_{=\beta_t} Z_T^{\perp}Z_T^{\top} + \sum_{t=1}^{T}(\underbrace{a^{-T}\langle X_t, Z_T^{\perp}\rangle}_{=\beta_t})^2 Z_T^{\perp}(Z_T^{\perp})^{\top} \\
&= \underbrace{||\alpha||^2 Z_T Z_T^{\top} + ||\beta||^2 Z_T^{\perp}(Z_T^{\perp})^{\top}}_{=M} + \langle\alpha,\beta\rangle(Z_T^{\perp}Z_T^{\top} + Z_T(Z_T^{\perp})^{\top}) \\
&= M + \underbrace{\langle\alpha,\beta\rangle[Z_T Z_T^{\perp}]}_{=U}\underbrace{\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}}_{=C}\underbrace{\begin{bmatrix} Z_T^{\top} \\ (Z_T^{\perp})^{\top} \end{bmatrix}}_{=V}
\end{aligned}$$

By using Woodbury's matrix identity and since $M^{-1} = ||\alpha||^{-2}Z_T Z_T^{\top} + ||\beta||^{-2}Z_T^{\perp}(Z_T^{\perp})^{\top}, C = C^{-1}$ we get

$$\begin{aligned}
(a^{-2T}\sum_{t=1}^{T} X_t X_t^{\top})^{-1} &= M^{-1} - \langle\alpha,\beta\rangle M^{-1}U(C + \langle\alpha,\beta\rangle U^{\top}M^{-1}U)^{-1}U^{\top}M^{-1} \\
&= M^{-1} - \langle\alpha,\beta\rangle[||\alpha||^{-2}Z_T \;\; ||\beta||^{-2}Z_T^{\perp}]\Big(\begin{bmatrix} \langle\alpha,\beta\rangle||\alpha||^{-2} & 1 \\ 1 & ||\beta||^{-2}\langle\alpha,\beta\rangle \end{bmatrix}\Big)^{-1}\begin{bmatrix} ||\alpha||^{-2}Z_T^{\top} \\ ||\beta||^{-2}(Z_T^{\perp})^{\top} \end{bmatrix}
\end{aligned}$$

Then the error term is

$$\begin{aligned}
\hat{A}_o - A_o &= \Big(\sum_{t=1}^{T} a^{-2T}\eta_{t+1}X_t'\Big)(a^{-2T}\sum_{t=1}^{T} X_t X_t^{\top})^{-1} \\
&= \Big(\sum_{t=1}^{T}\langle a^{-T}X_t, Z_T\rangle a^{-T}\eta_{t+1}Z_T' + \sum_{t=1}^{T}\langle a^{-T}X_t, Z_T^{\perp}\rangle a^{-T}\eta_{t+1}(Z_T^{\perp})'\Big)(a^{-2T}\sum_{t=1}^{T} X_t X_t^{\top})^{-1}
\end{aligned}$$

We now check the projection of $Z_T, Z_T^{\perp}$ on $(a^{-2T}\sum_{t=1}^{T} X_t X_t^{\top})^{-1}$

$$\begin{aligned}
Z_T^{\top}(a^{-2T}\sum_{t=1}^{T} X_t X_t^{\top})^{-1} &= ||\alpha||^{-2}Z_T^{\top} - \langle\alpha,\beta\rangle[||\alpha||^{-2} \;\; 0]\Big(\begin{bmatrix} \langle\alpha,\beta\rangle||\alpha||^{-2} & 1 \\ 1 & \langle\alpha,\beta\rangle||\beta||^{-2} \end{bmatrix}\Big)^{-1}\begin{bmatrix} ||\alpha||^{-2}Z_T^{\top} \\ ||\beta||^{-2}(Z_T^{\perp})^{\top} \end{bmatrix} \\
&= \frac{-||\alpha||^{-2}Z_T^{\top} + \langle\alpha,\beta\rangle||\alpha||^{-2}||\beta||^{-2}(Z_T^{\perp})^{\top}}{\langle\alpha,\beta\rangle^2||\alpha||^{-2}||\beta||^{-2}-1} \qquad (123) \\
(Z_T^{\perp})^{\top}(a^{-2T}\sum_{t=1}^{T} X_t X_t^{\top})^{-1} &= ||\beta||^{-2}(Z_T^{\perp})^{\top} - \langle\alpha,\beta\rangle[0 \;\; ||\beta||^{-2}]\Big(\begin{bmatrix} \langle\alpha,\beta\rangle||\alpha||^{-2} & 1 \\ 1 & \langle\alpha,\beta\rangle||\beta||^{-2} \end{bmatrix}\Big)^{-1}\begin{bmatrix} ||\alpha||^{-2}Z_T^{\top} \\ ||\beta||^{-2}(Z_T^{\perp})^{\top} \end{bmatrix} \\
&= \frac{-||\beta||^{-2}(Z_T^{\perp})^{\top} + \langle\alpha,\beta\rangle||\alpha||^{-2}||\beta||^{-2}Z_T^{\top}}{\langle\alpha,\beta\rangle^2||\alpha||^{-2}||\beta||^{-2}-1} \qquad (124)
\end{aligned}$$

We will show that with high probability $||\alpha||^{-2}= \Theta(1), ||\beta||^{-2}= \Omega(a^{2T}), \langle \alpha, \beta \rangle = O(a^{-T})$ as a result Eq. (123) is $\Omega(a^T)$ and Eq. (124) is $\Omega(a^{2T})$. Note that $Z_T^{\perp} = \begin{bmatrix} Z_T^{(2)} \\ -Z_T^{(1)} \end{bmatrix}$ where we have ignored the scaling (as these will be of constant order with high probability). First taking a closer look at $\alpha_t = a^{-2T} X_t^{(1)} Z_T^{(1)} + a^{-2T} X_t^{(2)} Z_T^{(2)}$ reveals the following behaviour

$$\begin{aligned}
a^{-2T} X_{T-1}^{(1)} Z_T^{(1)} &= a^{-1}(Z_T^{(1)})^2 - a^{-T-1} Z_T^{(1)} \eta_T^{(1)} \\
\alpha_{T-1} &= a^{-1}((Z_T^{(1)})^2 + (Z_T^{(2)})^2) - a^{-T-1}(Z_T^{(1)} \eta_T^{(1)} + Z_T^{(2)} \eta_T^{(2)}) \\
a^{-2T} X_{T-2}^{(1)} Z_T^{(1)} &= a^{-2}(Z_T^{(1)})^2 - a^{-T-1} Z_{T-1}^{(1)} \eta_T^{(1)} - a^{-T-2} Z_T^{(1)} \eta_T^{(1)} \\
\alpha_{T-2} &= a^{-2}((Z_T^{(1)})^2 + (Z_T^{(2)})^2) - a^{-T-1}(Z_{T-1}^{(1)} \eta_T^{(1)} + Z_{T-1}^{(2)} \eta_T^{(2)}) - a^{-T-2}(Z_T^{(1)} \eta_T^{(1)} + Z_T^{(2)} \eta_T^{(2)})
\end{aligned}$$

Since $Z_T^{(1)}$ is a Gaussian random variable with bounded variance, we see that α_t decays exponentially as t decreases (up to some a^{-T} additive terms). In a similar fashion one can show that $\sum_{t=1}^T \alpha_t^2 = \frac{1-a^{-2T}}{1-a^{-2}}((Z_T^{(1)})^2 + (Z_T^{(2)})^2)^2 + O(T^2 a^{-T})$ with high probability. Clearly $||\alpha||^{-2}= \Theta(1)$ with high probability. For β, note that $Z_T^{(2)}$ is independent of $X_t^{(1)}$ and observe that $\{a^T \beta_t\}_{t=1}^{T-1}$ are non–decaying and non–trivial random variables. Specifically these are subexponential random variables with $||\cdot||_{\psi_1}$ norm as $||a^T \beta_t||_{\psi_1}= C a^{-1}$. Here $||\cdot||_{\psi_1}$ norm is the same Definition 2.7.5 in (Vershynin, 2018). To see this consider for example $t = T-1, T-2$, then

$$\begin{aligned}
a^T \beta_{T-1} &= \langle X_{T-1}, Z_T^{\perp} \rangle = X_{T-1}^{(1)} Z_T^{(2)} - X_{T-1}^{(2)} Z_T^{(1)} = a^{-1}(\eta_T^{(2)} Z_T^{(1)} - \eta_T^{(1)} Z_T^{(2)}) \\
a^T \beta_{T-2} &= \langle X_{T-1}, Z_T^{\perp} \rangle = X_{T-1}^{(1)} Z_T^{(2)} - X_{T-1}^{(2)} Z_T^{(1)} = a^{-1}((\eta_{T-1}^{(2)} + a^{-1} \eta_T^{(2)}) Z_T^{(1)} - (\eta_{T-1}^{(1)} + a^{-1} \eta_T^{(1)}) Z_T^{(2)}) \quad (125)
\end{aligned}$$

Clearly, $a^{2T} ||\beta||_2^2= \Omega(1)$ and $a^{2T} ||\beta||_2^2= O(T)$ with high probability. Recall the error term

$$\begin{aligned}
\hat{A}_o - A_o &= \Big(\sum_{t=1}^T a^{-2T} \eta_{t+1} X_t'\Big)(a^{-2T} \sum_{t=1}^T X_t X_t^{\top})^{-1} \\
&= \Big(\sum_{t=1}^T \langle a^{-T} X_t, Z_T \rangle a^{-T} \eta_{t+1} Z_T' + \sum_{t=1}^T \langle a^{-T} X_t, Z_T^{\perp} \rangle a^{-T} \eta_{t+1} (Z_T^{\perp})'\Big)(a^{-2T} \sum_{t=1}^T X_t X_t^{\top})^{-1} \\
(\hat{A}_o - A_o) Z_T^{\perp} &= (\sum_{t=1}^T \langle a^{-T} X_t, Z_T \rangle a^{-T} \eta_{t+1} Z_T')(a^{-2T} \sum_{t=1}^T X_t X_t^{\top})^{-1} Z_T^{\perp} \\
&+ (\sum_{t=1}^T \langle a^{-T} X_t, Z_T^{\perp} \rangle a^{-T} \eta_{t+1} (Z_T^{\perp})')(a^{-2T} \sum_{t=1}^T X_t X_t^{\top})^{-1} Z_T^{\perp} \\
&= \frac{\langle \alpha, \beta \rangle ||\alpha||^{-2} ||\beta||^{-2}}{\langle \alpha, \beta \rangle^2 ||\alpha||^{-2} ||\beta||^{-2} - 1} \sum_{t=1}^T \langle a^{-T} X_t, Z_T \rangle a^{-T} \eta_{t+1} - \frac{-||\beta||^{-2}}{\langle \alpha, \beta \rangle^2 ||\alpha||^{-2} ||\beta||^{-2} - 1} \sum_{t=1}^T \langle a^{-T} X_t, Z_T^{\perp} \rangle a^{-T} \eta_{t+1} \\
&= \frac{||\alpha||^{-2} ||a^T \beta||^{-2}}{\langle \alpha, a^T \beta \rangle^2 ||\alpha||^{-2} ||a^T \beta||^{-2} - 1} \Big(\sum_{t=1}^T (\langle \alpha, a^T \beta \rangle \alpha_t - a^T \beta_t ||\alpha||^2) \eta_{t+1}\Big) = \gamma_T \quad (126)
\end{aligned}$$

Observe the term $a^T\beta_t||\alpha||^2\eta_{t+1}$

$$a^T\beta_t||\alpha||^2\eta_{t+1} = ||\alpha||^2\begin{bmatrix}(a^{-1}(\eta^{(2)}_{t+1}Z^{(1)}_T - \eta^{(1)}_{t+1}Z^{(2)}_T) + a^{-2}(\eta^{(2)}_{t+2}Z^{(1)}_T - \eta^{(1)}_{t+2}Z^{(2)}_T) + \ldots)\eta^{(1)}_{t+1} \\ (a^{-1}(\eta^{(2)}_{t+1}Z^{(1)}_T - \eta^{(1)}_{t+1}Z^{(2)}_T) + a^{-2}(\eta^{(2)}_{t+2}Z^{(1)}_T - \eta^{(1)}_{t+2}Z^{(2)}_T) + \ldots)\eta^{(2)}_{t+1}\end{bmatrix}$$

$$= ||\alpha||^2\begin{bmatrix}a^{-1}(\eta^{(2)}_{t+1}\eta^{(1)}_{t+1}Z^{(1)}_T - (\eta^{(1)}_{t+1})^2Z^{(2)}_T) + (a^{-2}(\eta^{(2)}_{t+2}Z^{(1)}_T - \eta^{(1)}_{t+2}Z^{(2)}_T) + \ldots)\eta^{(1)}_{t+1} \\ a^{-1}((\eta^{(2)}_{t+1})^2Z^{(1)}_T - \eta^{(2)}_{t+1}\eta^{(1)}_{t+1}Z^{(2)}_T) + (a^{-2}(\eta^{(2)}_{t+2}Z^{(1)}_T - \eta^{(1)}_{t+2}Z^{(2)}_T) + \ldots)\eta^{(2)}_{t+1}\end{bmatrix}$$

$$\sum_{t=1}^T a^T\beta_t||\alpha||^2\eta_{t+1} = a^{-1}||\alpha||^2\Big(\underbrace{\begin{bmatrix}-\sum_{t=1}^T(\eta^{(1)}_{t+1})^2Z^{(2)}_T \\ \sum_{t=1}^T(\eta^{(2)}_{t+1})^2Z^{(1)}_T\end{bmatrix}}_{=\Theta(T)} + \sum_{t=1}^T\begin{bmatrix}\eta^{(2)}_{t+1}\eta^{(1)}_{t+1}Z^{(1)}_T + (a^{-1}(\eta^{(2)}_{t+2}Z^{(1)}_T - \eta^{(1)}_{t+2}Z^{(2)}_T) + \ldots)\eta^{(1)}_{t+1} \\ \eta^{(2)}_{t+1}\eta^{(1)}_{t+1}Z^{(2)}_T + (a^{-1}(\eta^{(2)}_{t+2}Z^{(1)}_T - \eta^{(1)}_{t+2}Z^{(2)}_T) + \ldots)\eta^{(2)}_{t+1}\end{bmatrix}\Big)$$

$$\begin{aligned}&= a^{-1}||\alpha||^2\Big(\Theta(T) \\ &\underbrace{+\sum_{t=1}^T\begin{bmatrix}\eta^{(2)}_t\eta^{(1)}_tZ^{(1)}_T + a^{-1}\eta^{(2)}_{t+1}\eta^{(1)}_tZ^{(1)}_T + a^{-2}\eta^{(2)}_{t+2}\eta^{(1)}_tZ^{(1)}_T + \ldots - a^{-1}\eta^{(1)}_{t+1}\eta^{(1)}_tZ^{(2)}_T - a^{-2}\eta^{(1)}_{t+2}\eta^{(1)}_tZ^{(2)}_T - \ldots \\ \eta^{(2)}_t\eta^{(1)}_tZ^{(2)}_T + a^{-1}\eta^{(2)}_{t+1}\eta^{(2)}_tZ^{(1)}_T + a^{-2}\eta^{(2)}_{t+2}\eta^{(2)}_tZ^{(1)}_T + \ldots - a^{-1}\eta^{(1)}_{t+1}\eta^{(2)}_tZ^{(2)}_T - a^{-2}\eta^{(1)}_{t+2}\eta^{(2)}_tZ^{(1)}_T - \ldots\end{bmatrix}\Big)}_{=O(\sqrt{T}\log\frac{T}{\delta})}\end{aligned}$$

The $O(\sqrt{T}\log T)$ follows by applying Hanson-Wright inequality to each of $a^{-j}\sum_{t=1}^T\eta^{(2)}_{t+j}\eta^{(1)}_t$ terms where we get with probability at least $1-\delta/T$ that $a^{-j}\sum_{t=1}^T\eta^{(2)}_{t+j}\eta^{(1)}_t \leq ca^{-j}O(\sqrt{T}\log\frac{T}{\delta})$. Therefore simultaneously for all $j \leq T$ we have with probability at least $1-\delta$ (using union bound) that $a^{-j}\sum_{t=1}^T\eta^{(2)}_{t+j}\eta^{(1)}_t \leq ca^{-j}O(\sqrt{T}\log\frac{T}{\delta}) \implies \sum_{j=1}^T a^{-j}\sum_{t=1}^T\eta^{(2)}_{t+j}\eta^{(1)}_t \leq O(\sqrt{T}\log\frac{T}{\delta})$. Plugging this in Eq. (126) we get that

$$\gamma_T = \frac{||\alpha||^{-2}||a^T\beta||^{-2}}{\langle\alpha, a^T\beta\rangle^2||\alpha||^{-2}||a^T\beta||^{-2}-1}\Big(\underbrace{\sum_{t=1}^T(\langle\alpha, a^T\beta\rangle\alpha_t}_{=O(\sqrt{T})} - \underbrace{a^T\beta_t||\alpha||^2)\eta_{t+1}}_{=\Theta(T)}\Big)$$

Clearly then γ_T in Eq. (126) satisfies a non–trivial pdf, *i.e.*, error does not decay to zero. □

Another interesting observation is that $\sum_{t=1}^T a^{-2T}\eta_{t+1}X_t^\top$ decays $O(a^{-T})$ with high probability, however the error is a non–decaying random variable. This immediately gives us that

Proposition 20.2. *The sample covariance matrix $\sum_{t=1}^T X_tX_t^\top$ has the following singular values*

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

Proof. The largest singular values of $\sum_{t=1}^T X_tX_t^\top = \Theta(a^{2T})$ this follows because

$$||\sum_{t=1}^T a^{-2T}X_tX_t^\top - \frac{1-a^{-2T}}{1-a^{-2}}Z_TZ_T^\top||_2 \leq O(a^{-T})$$

with high probability, which follows from the claims of Eq. (17), (18) in Theorem 1 and discussion in Section 12. The second claim follows because $\sum_{t=1}^T a^{-2T}\eta_{t+1}X_t^\top$ decays $\Omega(a^{-T})$ with high probability. To see this

$$\sum_{t=1}^T a^{-2T}\eta_{t+1}X_t^\top \leq a^{-T}\sqrt{\sum_{t=1}^T\eta_t^\prime\eta_t}\sqrt{\sum_{t=1}^T a^{-2T}X_t^\prime X_t} \approx \sqrt{T}a^{-T}$$

The $\sqrt{T}$ factor can be removed by similar arguments as above. However the identification error is a random variable which implies that $\sigma_2(\sum_{t=1}^T a^{-2T}X_tX_t^\top) = O(\sqrt{T}a^{-T})$. □