Journal of Machine Learning Research 22 (2021) 1-61 Submitted 8/19; Revised 4/20; Published 1/21

Finite Time LTI System Identification

Tuhin Sarkar TUHIN91@GMAIL.COM
Department of Electrical Engineering and Computer Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Alexander Rakhlin RAKHLIN@MIT.EDU
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Munther A. Dahleh DAHLEH@MIT.EDU
Department of Electrical Engineering and Computer Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Editor: Benjamin Recht

Abstract

We address the problem of learning the parameters of a stable linear time invariant (LTI) system with unknown latent space dimension, or order, from a single time–series of noisy input-output data. We focus on learning the best lower order approximation allowed by finite data. Motivated by subspace algorithms in systems theory, where the doubly infinite system Hankel matrix captures both order and good lower order approximations, we construct a Hankel-like matrix from noisy finite data using ordinary least squares. This circumvents the non-convexities that arise in system identification, and allows accurate estimation of the underlying LTI system. Our results rely on careful analysis of self-normalized martingale difference terms that helps bound identification error up to logarithmic factors of the lower bound. We provide a data-dependent scheme for order selection and find an accurate realization of system parameters, corresponding to that order, by an approach that is closely related to the Ho-Kalman subspace algorithm. We demonstrate that the proposed model order selection procedure is not overly conservative, i.e., for the given data length it is not possible to estimate higher order models or find higher order approximations with reasonable accuracy.

Keywords: Linear Dynamical Systems, System Identification, Non–parametric statistics, control theory, Statistical Learning theory

1. Introduction

Finite-time system identification—the problem of estimating the system parameters given a finite single time series of its output—is an important problem in the context of control theory, time series analysis, robotics, and economics, among many others. In this work, we focus on parameter estimation and model approximation of linear time invariant (LTI)

systems or linear dynamical system (LDS), which are described by

$$\begin{aligned} X_{t+1} &= AX_t + BU_t + \eta_{t+1} \\ Y_t &= CX_t + w_t. \end{aligned} \tag{1}$$

Here $C \in \mathbb{R}^{p\times n}, A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}$; $\{\eta_t, w_t\}_{t=1}^{\infty}$ are process and output noise, U_t is an external control input, X_t is the latent state variable and Y_t is the observed output. The goal here is parameter estimation, *i.e.*, learning (C, A, B) from a single finite time series of $\{Y_t, U_t\}_{t=1}^{T}$ when the order, n, is unknown. Since typically $p, m < n$, it becomes challenging to find suitable parametrizations of LTI systems for provably efficient learning. When $\{X_j\}_{j=1}^{\infty}$ are observed (or, C is known to be the identity matrix), identification of (C, A, B) in Eq. (1) is significantly easier, and ordinary least squares (OLS) is a statistically optimal estimator. It is, in general, unclear how (or if) OLS can be employed in the case when X_t's are not observed.

To motivate the study of a lower-order approximation of a high-order system, consider the following example:

Example 1 *Consider $M_1 = (A_1, B_1, C_1)$ with*

$$A_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \\ -a & 0 & 0 & 0 & \dots & 0 \end{bmatrix}_{n\times n} \quad B_1 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}_{n\times 1} \quad C_1 = B_1^{\top} \tag{2}$$

where $na \ll 1$ and $n > 20$. Here the order of M_1 is n. However, it can be approximated well by M_2 which is of a much lower order and given by

$$A_2 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \quad B_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad C_2 = B_2^{\top}. \tag{3}$$

For the same input U_t, if $Y_t^{(1)}, Y_t^{(2)}$ be the output generated by M_1 and M_2 respectively then a simple computation shows that

$$\sup_{U} \sum_{t=1}^{\infty} \frac{(Y_t^{(1)} - Y_t^{(2)})^2}{U_t^2} \leq 4n^2a^2 \ll 1$$

This suggests that the actual value of n is not important; rather there exists an effective order, r (which is 2 in this case). This lower order model captures "most" of the LTI system.

Since the true model order is not known in many cases, we emphasize a nonparametric approach to identification: one which adaptively selects the best model order for the given data and approximates the underlying LTI system better as T (length of data) grows. The key to this approach will be designing an estimator $\hat{M}$ from which we obtain a realization $(\hat{C}, \hat{A}, \hat{B})$ of the selected order.

1.1 Related Work

Linear time invariant systems are an extensively studied class of models in control and systems theory. These models are used in feedback control systems (for example in planetary soft landing systems for rockets (Açıkmeşe et al., 2013)) and as linear approximations to many non–linear systems that nevertheless work well in practice. In the absence of process and output noise, subspace-based system identification methods are known to learn (C, A, B) (up to similarity transformation)(Ljung, 1987; Van Overschee and De Moor, 2012). These typically involve constructing a Hankel matrix from the input–output pairs and then obtaining system parameters by a singular value decomposition. Such methods are inspired by the celebrated Ho-Kalman realization algorithm (Ho and Kalman, 1966). The correctness of these methods is predicated on the knowledge of n or presence of infinite data. Other approaches include rank minimization-based methods for system identification (Fazel et al., 2013; Grussler et al., 2018), further relaxing the rank constraint to a suitable convex formulation. However, there is a lack of statistical guarantees for these algorithms, and it is unclear how much data is required to obtain accurate estimates of system parameters from finite noisy data. Empirical methods such as the EM algorithm (Dempster et al., 1977) are also used in practice; however, these suffer from non-convexity in problem formulation and can get trapped in local minima. Learning simpler approximations to complex models in the presence of finite noisy data was studied in Venkatesh and Dahleh (2001) where identification error is decomposed into error due to approximation and error due to noise; however the analysis assumes the knowledge of a "good" parametrization and does not provide statistical guarantees for learning the system parameters of such an approximation.

More recently, there has been a resurgence in the study of statistical identification of LTI systems from a single time series in the machine learning community. In cases when $C = I$, *i.e.*, X_t is observed directly, sharp finite time error bounds for identification of A, B from a single time series are provided in Faradonbeh et al. (2017); Simchowitz et al. (2018); Sarkar and Rakhlin (2018). The approach to finding A, B is based on a standard ordinary least squares (OLS) given by

$$(\hat{A}, \hat{B}) = \arg\min_{A,B} \sum_{t=1}^{T} ||X_{t+1} - [A, B][X_t^\top, U_t^\top]^\top||_2^2.$$

Another closely related area is that of online prediction in time series Hazan et al. (2018); Agarwal et al. (2018). Finite time regret guarantees for prediction in linear time series are provided in Hazan et al. (2018). The approach there circumvents the need for system identification and instead uses a filtering technique that convolves the time series with eigenvectors of a specific Hankel matrix.

Closest to our work is that of Oymak and Ozay (2018). Their algorithm, which takes inspiration from the Kalman–Ho algorithm, assumes the knowledge of model order n. This limits the applicability of the algorithm in two ways: first, it is unclear how the techniques can be extended to the case when n is unknown—as is usually the case—and, second, in many cases n is very large and a much lower order LTI system can be a very good approximation of the original system. In such cases, constructing the order n estimate might be unnecessarily conservative (See Example 1). Consequently, the error bounds do not reflect accurate dependence on the system parameters.

When n is unknown, it is unclear when a singular value decomposition should be performed to obtain the parameter estimates via Ho-Kalman algorithm. This leads to the question of model order selection from data. For subspace based methods, such problems have been addressed in Shibata (1976) and Bauer (2001). These papers address the question of estimating order in the context of subspace methods. Specifically, order estimation is achieved by analyzing the information contained in the estimated singular values and/or estimated innovation variance. Furthermore, they provide guarantees for asymptotic consistency of the methods described. It is unclear, however, if these techniques and guarantees can be extended to the case when only finite data is available. Another line of literature studied in Ljung et al. (2015) for example, approaches the identification of systems with unknown order by first learning the largest possible model that fits the data and then performing model reduction to obtain the final system. Although one can show that asymptotically this method outputs the true model, we show that such a two step procedure may underperform in a finite time setting. A possible explanation for this could be that learning the largest possible model with finite data over-fits on the exogenous noise and therefore gives poor model estimates. The key difference from prior work is that we provide a direct approach to model selection, instead of learning the largest possible model from data and subsequent model truncation, and provide finite time guarantees.

Other related work on identifying finite impulse response approximations include Goldenshluger (1998); Tu et al. (2017); but they do not discuss parameter estimation or reduced order modeling. Several authors Campi and Weyer (2002); Shah et al. (2012); Hardt et al. (2016) and references therein have studied the problem of system identification in different contexts. However, they fail to capture the correct dependence of system parameters on error rates. More importantly, they suffer from the same limitation as Oymak and Ozay (2018) that they require the knowledge of n.

2. Mathematical Preliminaries

Throughout the paper, we will refer to an LTI system with dynamics as Eq. (1) by $M = (C, A, B)$. For a matrix A, let $\sigma_i(A)$ be the i^{th} singular value of A with $\sigma_i(A) \geq \sigma_{i+1}(A)$. Further, $\sigma_{\max}(A) = \sigma_1(A) = \sigma(A)$. Similarly, we define $\rho_i(A) = |\lambda_i(A)|$, where $\lambda_i(A)$ is an eigenvalue of A with $\rho_i(A) \geq \rho_{i+1}(A)$. Again, $\rho_{\max}(A) = \rho_1(A) = \rho(A)$.

Definition 1 *A matrix A is Schur stable if $\rho_{\max}(A) < 1$.*

We will only be interested in the class of LTI systems that are Schur stable. Fix $\gamma > 0$ (and possibly much greater than 1). The model class $\mathcal{M}_r$ of LTI systems parametrized by $r \in \mathbb{Z}_+$ is defined as

$$\mathcal{M}_r = \{(C, A, B) \mid C \in \mathbb{R}^{p \times r}, A \in \mathbb{R}^{r \times r}, B \in \mathbb{R}^{r \times m}, \rho(A) < 1, \sigma(A) \leq \gamma\}. \tag{4}$$

Definition 2 *The (k,p,q)–dimensional Hankel matrix for $M=(C,A,B)$ as*

$$\mathcal{H}_{k,p,q}(M)=\begin{bmatrix} CA^kB & CA^{k+1}B & \dots & CA^{q+k-1}B \\ CA^{k+1}B & CA^{k+2}B & \dots & CA^{q+k}B \\ \vdots & \vdots & \ddots & \vdots \\ CA^{p+k-1}B & \dots & \dots & CA^{p+q+k-2}B \end{bmatrix}$$

and its associated Toeplitz matrix as

$$\mathcal{T}_{k,d}(M)=\begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ CA^kB & 0 & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & 0 \\ CA^{d+k-3}B & \dots & CA^kB & 0 & 0 \\ CA^{d+k-2}B & CA^{d+k-3}B & \dots & CA^kB & 0 \end{bmatrix}.$$

We will slightly abuse notation by referring to $\mathcal{H}_{k,p,q}(M)=\mathcal{H}_{k,p,q}$. Similarly for the Toeplitz matrices $\mathcal{T}_{k,d}(M)=\mathcal{T}_{k,d}$. The matrix $\mathcal{H}_{0,\infty,\infty}(M)$ is known as the *system Hankel matrix* corresponding to M, and its rank is known as the *model order* (or simply *order*) of M. The system Hankel matrix has two well-known properties that make it useful for system identification. First, the rank of $\mathcal{H}_{0,\infty,\infty}$ has an upper bound n. Second, it maps the "past" inputs to "future" outputs. These properties are discussed in detail in appendix as Section 9.2. For infinite matrices $\mathcal{H}_{0,\infty,\infty}$, $||\mathcal{H}_{0,\infty,\infty}||_2 \triangleq ||\mathcal{H}_{0,\infty,\infty}||_{\text{op}}$, *i.e.*, the operator norm.

Definition 3 *The transfer function of $M=(C,A,B)$ is given by $G(z)=C(zI-A)^{-1}B$ where $z\in\mathbb{C}$.*

The transfer function plays a critical role in control theory as it relates the input to the output. Succinctly, the transfer function of an LTI system is the Z–transform of the output in response to a unit impulse input. Since for any invertible S the LTI systems $M_1=(CS^{-1},SAS^{-1},SB), M_2=(C,A,B)$ have identical transfer functions, identification may not be unique, but equivalent up to a transformation S, *i.e.*, $(C,A,B)\equiv(CS,S^{-1}AS,S^{-1}B)$. Next, we define a system norm that will be important from the perspective of model identification and approximation.

Definition 4 *The $\mathcal{H}_\infty$–system norm of a Schur stable LTI system M is given by*

$$||M||_\infty = \sup_{\omega\in\mathbb{R}} \sigma_{\max}(G(e^{j\omega})).$$

Here, $G(\cdot)$ is the transfer function of M. The r–truncation of the transfer function is defined as

$$G_r \coloneqq [CB, CAB, \dots, CA^{r-1}B]. \tag{5}$$

For a stable LTI system M we have

Proposition 2.1 (Lemma 2.2 Glover (1987)) *Let M be a LTI system then*

$$||M||_H = \sigma_1 \le ||M||_\infty \le 2(\sigma_1+\dots+\sigma_n)$$

where σ_i are the singular values of $\mathcal{H}_{0,\infty,\infty}(M)$.

For any matrix Z, define $Z_{m:n,p:q}$ as the submatrix including row m to n and column p to q. Further, $Z_{m:n,:}$ is the submatrix including row m to n and all columns and a similar notion exists for $Z_{:,p:q}$. Finally, we define balanced truncated models which will play an important role in our algorithm.

Definition 5 (Kung and Lin (1981)) *Let $\mathcal{H}_{0,\infty,\infty}(M) = U\Sigma V^\top$ where $\Sigma \in \mathbb{R}^{n\times n}$ (n is the model order). Then for any $r \leq n$, the r–order balanced truncated model parameters are given by*

$$C_r = [U\Sigma^{1/2}]_{1:p,1:r}, A_r = \Sigma^{-1/2}_{1:r,1:r} U^\top_{:,1:r}[U\Sigma^{1/2}]_{p+1:,1:r}, B_r = [\Sigma^{1/2}V^\top]_{1:r,1:m}.$$

For $r > n$, the r–order balanced truncated model parameters are the n–order truncated model parameters.

Definition 6 *We say a random vector $v \in \mathbb{R}^d$ is subgaussian with variance proxy τ^2 if*

$$\sup_{||\theta||_2=1} \sup_{p\geq 1} \left\{ p^{-1/2} \left(\mathbb{E}[|\langle v, \theta\rangle|^p]\right)^{1/p} \right\} = \tau$$

and $\mathbb{E}[v] = \boldsymbol{0}$. We denote this by $v \sim \mathsf{subg}(\tau^2)$.

A fundamental result in model reduction from systems theory is the following

Theorem 2.1 (Theorem 21.26 Zhou et al. (1996)) *Let $M = (C, A, B)$ be the true model of order n and $M_r = (C_r, A_r, B_r)$ be its balance truncated model of order $r < n$. Assume that $\sigma_r \neq \sigma_{r+1}$. Then*

$$||M - M_r||_\infty \leq 2(\sigma_{r+1} + \sigma_{r+2} + \ldots + \sigma_n)$$

where σ_i are the Hankel singular values of M.

Critical to obtaining refined error rates, will be a result from the theory of self–normalized martingales, an application of the pseudo-maximization technique in (Peña et al., 2008, Theorem 14.7):

Theorem 2.2 *Let $\{\boldsymbol{\mathcal{F}}_t\}_{t=0}^\infty$ be a filtration. Let $\{\eta_t \in \mathbb{R}^m, X_t \in \mathbb{R}^d\}_{t=1}^\infty$ be stochastic processes such that η_t, X_t are $\boldsymbol{\mathcal{F}}_t$ measurable and η_t is $\boldsymbol{\mathcal{F}}_{t-1}$-conditionally $\mathsf{subg}(L^2)$ for some $L > 0$. For any $t \geq 0$, define $V_t = \sum_{s=1}^t X_s X_s'$, $S_t = \sum_{s=1}^t \eta_{s+1} X_s$. Then for any $\delta > 0, V \succ 0$ and all $t \geq 0$ we have with probability at least $1 - \delta$*

$$S_t^\top (V + V_t)^{-1} S_t \leq 4L^2 \left(\log \frac{1}{\delta} + \log \frac{det(V + V_t)}{det(V)} + m \right).$$

The proof of this result can be found as Theorem 8.5.

We denote by c universal constants which can change from line to line. For numbers a, b, we define $a \wedge b \triangleq \min(a, b)$ and $a \vee b \triangleq \max(a, b)$.

Finally, for two matrices $M_1 \in \mathbb{R}^{l_1\times l_1}, M_2 \in \mathbb{R}^{l_2\times l_2}$ with $l_1 < l_2$, $M_1 - M_2 \triangleq \tilde{M}_1 - M_2$ where $\tilde{M}_1 = \begin{bmatrix} M_1 & 0_{l_1\times l_2 - l_1} \\ 0_{l_2-l_1\times l_1} & 0_{l_2-l_1\times l_2-l_1} \end{bmatrix}$.

Proposition 2.2 (System Reduction) *Let* $||S-P|| \leq \epsilon$ *and the singular values of* S *be arranged as follows:*

$$\sigma_1(S) > \ldots > \sigma_{r-1}(S) > \sigma_r(S) \geq \sigma_{r+1}(S) \geq \ldots \geq \sigma_s(S) > \sigma_{s+1}(S) > \ldots \sigma_n(S) > \sigma_{n+1}(S) = 0$$

Furthermore, let ϵ *be such that*

$$\epsilon \leq \inf_{\{1 \leq i \leq r-1\} \cup \{s+1 \leq i \leq n\}} \Big(\frac{\sigma_i(P) - \sigma_{i+1}(P)}{2} \Big). \tag{6}$$

Define $K_0 = [1, 2, \ldots, r-1] \cup [s+1, s+2, \ldots, n]$*, then*

$$||U^S_{K_0}(\Sigma^S_{K_0})^{1/2} - U^P_{K_0}(\Sigma^P_{K_0})^{1/2}||_2 \leq 2\sqrt{\sum_{i=1}^{r-1} \frac{\sigma_i \epsilon^2}{(\sigma_i - \sigma_{i+1})^2 \wedge (\sigma_{i-1} - \sigma_i)^2}} + 2\sqrt{\frac{\sigma_s \epsilon^2}{((\sigma_{r-1} - \sigma_s) \wedge (\sigma_r - \sigma_{s+1}))^2}} + \sup_{1 \leq i \leq s} |\sqrt{\sigma_i} - \sqrt{\hat{\sigma}_i}|$$

and $\sigma_i = \sigma_i(S), \hat{\sigma}_i = \sigma_i(P)$*.*

The proof is provided in Proposition 12.4 in the appendix. This is an extension of Wedin's result that allows us to scale the recovery error of the r^{th} singular vector by only condition number of that singular vector. This is useful to represent the error of identifying a r-order approximation as a function of the r^{th}-singular value only.

We briefly summarize our contributions below.

3. Contributions

In this paper we provide a purely data-driven approach to system identification from a single time–series of finite noisy data. Drawing from tools in systems theory and the theory of self–normalized martingales, we offer a nearly optimal OLS-based algorithm to learn the system parameters. We summarize our contributions below:

- The central theme of our approach is to estimate the infinite system Hankel matrix (to be defined below) with increasing accuracy as the length T of data grows. By utilizing a specific reformulation of the input–output relation in Eq. (1) we reduce the problem of Hankel matrix identification to that of regression between appropriately transformed versions of output and input. The OLS solution is a matrix $\hat{\mathcal{H}}$ of size $\hat{d}$. More precisely, we show that with probability at least $1-\delta$,

$$\Big|\Big|\hat{\mathcal{H}} - \mathcal{H}_{0,\hat{d},\hat{d}}\Big|\Big|_2 \lesssim \sqrt{\frac{\beta^2 \hat{d}}{T}} \sqrt{p\hat{d} + \log \frac{T}{\delta}}$$

  for T above a certain threshold, where $\mathcal{H}_{0,\hat{d},\hat{d}}$ is the $p\hat{d} \times m\hat{d}$ principal submatrix of the system Hankel. Here β is the $\mathcal{H}_\infty$–system norm.

- We show that by growing $\hat{d}$ with T in a specific fashion, $\hat{\mathcal{H}}$ becomes the minimax optimal estimator of the system Hankel matrix. The choice of $\hat{d}$ for a fixed T is purely data-dependent and does not depend on spectral radius of A or n.

- It is well known in systems theory that SVD of the doubly infinite system Hankel matrix gives us A, B, C. However, the presence of finite noisy data prevents learning these parameters accurately. We show that it is always possible to learn the parameters of a lower-order approximation of the underlying system. This is achieved by selecting the top k singular vectors of $\hat{\mathcal{H}}$. The estimation guarantee corresponds to *model selection* in Statistics. More precisely, for every $k \leq \hat{d}$ if (A_k, B_k, C_k) are the parameters of a k-order balanced approximation of the original LTI system and $(\hat{A}_k, \hat{B}_k, \hat{C}_k)$ are the estimates of our algorithm then for T above a certain threshold we have

$$||C_k - \hat{C}_k||_2 + ||A_k - \hat{A}_k||_2 + ||B_k - \hat{B}_k||_2 \lesssim \sqrt{\frac{\beta^2 \hat{d}}{\hat{\sigma}_k^2 T}} \sqrt{p\hat{d} + \log \frac{T}{\delta}}$$

  with probability at least $1 - \delta$ where $\hat{\sigma}_i$ is the i^{th} largest singular value of $\hat{\mathcal{H}}$.

4. Problem Formulation and Discussion

4.1 Data Generation

Assume there exists an unknown $M = (C, A, B) \in \mathcal{M}_n$ for some unknown n. Let the transfer function of M be $G(z)$. Suppose we observe the noisy output time series $\{Y_t \in \mathbb{R}^{p\times 1}\}_{t=1}^T$ in response to user chosen input series, $\{U_t \in \mathbb{R}^{m\times 1}\}_{t=1}^T$. We refer to this data generated by M as $Z_T = \{(U_t, Y_t)\}_{t=1}^T$. We enforce the following assumptions on M.

Assumption 1 *The noise process $\{\eta_t, w_t\}_{t=1}^\infty$ in the dynamics of M given by Eq. (1) are i.i.d. and η_t, w_t are isotropic with subGaussian parameter 1. Furthermore, $X_0 = 0$ almost surely. We will only select inputs, $\{U_t\}_{t=1}^T$, that are isotropic subGaussian with subGaussian parameter 1.*

The input–output map of Eq. (1) can be represented in multiple alternate ways. One commonly used reformulation of the input–output map in systems and control theory is the following

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_T \end{bmatrix} = \mathcal{T}_{0,T} \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_T \end{bmatrix} + \mathcal{TO}_{0,T} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_T \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_T \end{bmatrix}$$

where $\mathcal{TO}_{k,d}$ is defined as the Toeplitz matrix corresponding to process noise η_t (similar to Definition 2):

$$\mathcal{TO}_{k,d} = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ CA^k & 0 & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & 0 \\ CA^{d+k-3} & \dots & CA^k & 0 & 0 \\ CA^{d+k-2} & CA^{d+k-3} & \dots & CA^k & 0 \end{bmatrix}.$$

$||\mathcal{T}_{0,T}||_2, ||\mathcal{TO}_{0,T}||_2$ denote observed amplifications of the control input and process noise respectively. Note that stability of A ensures $||\mathcal{T}_{0,\infty}||_2, ||\mathcal{TO}_{0,\infty}||_2 < \infty$. Suppose both

m: Input dimension, p: Output dimension
γ: Known upper bound on $\|\|A\|\|_2$
δ: Error probability
$c, \mathcal{C}$: Known absolute constants
R: Known noise to signal ratio, or, $\frac{\|\|\mathcal{TO}_{0,\infty}\|\|_2}{\|\|\mathcal{T}_{0,\infty}\|\|_2}$
β: Known upper bound on $\mathcal{H}_\infty$-norm of LTI system
$\mathcal{D}(T) = \{d \vert T \geq cm^2 d \log^2(d) \log^2(m^2/\delta) + cd \log^3(2d)\}$
$\sigma_A = \sum_{l=1}^{d} \|\|CA^l B\|\|_2$, $\sigma_B = \sum_{l=1}^{d} \|\|CA^l\|\|_2$
$\sigma_C = \sqrt{\sigma\left(\sum_{k=1}^{d} \mathcal{T}_{d+k,T}^\top \mathcal{T}_{d+k,T}\right)}$, $\sigma_D = \sqrt{\sigma\left(\sum_{k=1}^{d} \mathcal{TO}_{d+k,T}^\top \mathcal{TO}_{d+k,T}\right)}$
$\alpha(l) = \sqrt{l}\left(\sqrt{\frac{lp + \log(T/\delta) + m}{T}}\right)$

Table 1: Summary of constants

$\eta_t, w_t = 0$ in Eq. (1). Then it is a well-known fact that

$$\|\|M\|\|_\infty = \sup_{U_t} \sqrt{\frac{\sum_{t=0}^{\infty} Y_t^\top Y_t}{\sum_{t=0}^{\infty} U_t^\top U_t}} \implies \|\|M\|\|_\infty = \|\|\mathcal{T}_{0,\infty}\|\|_2 \geq \|\|\mathcal{H}_{0,\infty,\infty}\|\|_2. \tag{7}$$

Assumption 2 *There exist universal constants $\beta, R \geq 1$ such that $\|\|\mathcal{T}_{0,\infty}\|\|_2 \leq \beta$, $\frac{\|\|\mathcal{TO}_{0,\infty}\|\|_2}{\|\|\mathcal{T}_{0,\infty}\|\|_2} \leq R$.*

Remark 7 ($\mathcal{H}_\infty$-norm estimation) *Assumption 2 implies that an upper bound to the $\mathcal{H}_\infty$–norm of the system. It is possible to estimate $\|\|M\|\|_\infty$ from data (See Tu et al. (2018a) and references therein). It is reasonable to expect that error rates for identification of the parameters (C, A, B) depend on the noise-to-signal ratio $\frac{\|\|\mathcal{TO}_{0,\infty}\|\|_2}{\|\|\mathcal{T}_{0,\infty}\|\|_2}$, i.e., identification is much harder when the ratio is large.*

Remark 8 (R estimation) *The noise to signal ratio hyperparameter can also be estimated from data, by allowing the system to run with $U_t = 0$ and taking the average ℓ_2 norm of the output Y_t, i.e., $(1/T)\sum_{t=1}^{T} \|Y_t\|_2^2$. For the purpose of the results of the paper we simply assume an upper bound on R. If U_t was $\mathsf{subg}(L)$ instead of $\mathsf{subg}(1)$, the noise-to-signal ratio is modified to R/L instead.*

5. Algorithmic Details

We will now represent the input–output relationship in terms of the Hankel and Toeplitz matrices defined before. Fix a d, then for any l we have

$$\begin{bmatrix} Y_l \\ Y_{l+1} \\ \vdots \\ Y_{l+d-1} \end{bmatrix} = \mathcal{H}_{0,d,d} \begin{bmatrix} U_{l-1} \\ U_{l-2} \\ \vdots \\ U_{l-d} \end{bmatrix} + \mathcal{T}_{0,d} \begin{bmatrix} U_l \\ U_{l+1} \\ \vdots \\ U_{l+d-1} \end{bmatrix} + \mathcal{O}_{0,d,d} \begin{bmatrix} \eta_{l-1} \\ \eta_{l-2} \\ \vdots \\ \eta_{l-d+1} \end{bmatrix} + \mathcal{T}\mathcal{O}_{0,d} \begin{bmatrix} \eta_l \\ \eta_{l+1} \\ \vdots \\ \eta_{l+d-1} \end{bmatrix} + \mathcal{H}_{d,d,l-d-1} \begin{bmatrix} U_{l-d-1} \\ U_{l-d-1} \\ \vdots \\ U_1 \end{bmatrix} + \mathcal{O}_{d,d,l-d-1} \begin{bmatrix} \eta_{l-d-1} \\ \eta_{l-d-1} \\ \vdots \\ \eta_1 \end{bmatrix} + \begin{bmatrix} w_l \\ w_{l+1} \\ \vdots \\ w_{l+d-1} \end{bmatrix} \tag{8}$$

or, succinctly,

$$\begin{aligned} \tilde{Y}^+_{l,d} = \mathcal{H}_{0,d,d}\tilde{U}^-_{l-1,d} + \mathcal{T}_{0,d}\tilde{U}^+_{l,d} + \mathcal{H}_{d,d,l-d-1}\tilde{U}^-_{l-d-1,l-d-1} \\ + \mathcal{O}_{0,d,d}\tilde{\eta}^-_{l-1,d} + \mathcal{T}\mathcal{O}_{0,d}\tilde{\eta}^+_{l,d} + \mathcal{O}_{d,d,l-d-1}\tilde{\eta}^-_{l-d-1,l-d-1} + \tilde{w}^+_{l,d} \end{aligned} \tag{9}$$

Here

$$\mathcal{O}_{k,p,q} = \begin{bmatrix} CA^k & CA^{k+1} & \dots & CA^{q+k-1} \\ CA^{k+1} & CA^{k+2} & \dots & CA^{d+k} \\ \vdots & \vdots & \ddots & \vdots \\ CA^{p+k-1} & \dots & \dots & CA^{p+q+k-2} \end{bmatrix}, \tilde{Y}^-_{l,d} = \begin{bmatrix} Y_l \\ Y_{l-1} \\ \vdots \\ Y_{l-d+1} \end{bmatrix}, \tilde{Y}^+_{l,d} = \begin{bmatrix} Y_l \\ Y_{l+1} \\ \vdots \\ Y_{l+d-1} \end{bmatrix}.$$

Furthermore, $\tilde{U}^-_{l,d}, \tilde{\eta}^-_{l,d}$ are defined similar to $\tilde{Y}^-_{l,d}$ and $\tilde{U}^+_{l,d}, \tilde{\eta}^+_{l,d}, \tilde{w}^+_{l,d}$ are similar to $\tilde{Y}^+_{l,d}$. The $+$ and $-$ signs indicate moving forward and backward in time respectively. This representation will be at the center of our analysis.

There are three key steps in our algorithm which we describe in the following sections:

(a) Hankel submatrix estimation: Estimating $\mathcal{H}_{0,l,l}$ for every $1 \le l \le T$. We refer to the estimators as $\{\hat{\mathcal{H}}_{0,l,l}\}_{l=1}^T$.

(b) Model Selection: From the estimators $\{\hat{\mathcal{H}}_{0,l,l}\}_{l=1}^T$ select $\hat{\mathcal{H}}_{0,\hat{d},\hat{d}}$ in a data dependent way such that it "best" estimates $\mathcal{H}_{0,\infty,\infty}$.

(c) Parameter Recovery: For every $k \le \hat{d}$, we do a singular value decomposition of $\hat{\mathcal{H}}_{0,\hat{d},\hat{d}}$ to obtain parameter estimates for a "good" k-order approximation of the true model.

5.1 Hankel Submatrix Estimation

The goal of our systems identification is to estimate either $\mathcal{H}_{0,n,n}$ or $\mathcal{H}_{0,\infty,\infty}$. Since we only have finite data and no apriori knowledge of n it is not possible to directly estimate the unknown matrices. The first step then is to estimate all possible Hankel submatrices that are "allowed" by data, *i.e.*, $\mathcal{H}_{0,d,d}$ for $d \le T$. For a fixed d, Algorithm 1 estimates the $d \times d$ principal submatrix $\mathcal{H}_{0,d,d}$.

Algorithm 1 LearnSystem(T, d, m, p)

Input T = Horizon for learning
d = Hankel Size
m = Input dimension
p = Output dimension
Output System Parameters: $\hat{\mathcal{H}}_{0,d,d}$

1: Generate $2T$ i.i.d. inputs $\{U_j \sim \mathcal{N}(0, I_{m\times m})\}_{j=1}^{2T}$.
2: Collect $2T$ input–output pairs $\{U_j, Y_j\}_{j=1}^{2T}$.
3: $\hat{\mathcal{H}}_{0,d,d} = \arg\min_{\mathcal{H}} \sum_{l=0}^{T-1} ||\tilde{Y}^+_{l+d+1,d} - \mathcal{H}\tilde{U}^-_{l+d,d}||_2^2$
4: **return** $\hat{\mathcal{H}}_{0,d,d}$

It can be shown that

$$\hat{\mathcal{H}}_{0,d,d} = \Big(\sum_{l=0}^{T-1} \tilde{Y}^+_{l+d+1,d}(\tilde{U}^-_{l+d,d})^\top\Big)\Big(\sum_{l=0}^{T-1} \tilde{U}^-_{l+d,d}(\tilde{U}^-_{l+d,d})^\top\Big)^+ \tag{10}$$

and by running the algorithm T times, we obtain $\{\hat{\mathcal{H}}_{0,d,,d}\}_{d=1}^{T}$. A key step in showing that $\hat{\mathcal{H}}_{0,d,d}$ is a good estimator for $\mathcal{H}_{0,d,d}$ is to prove the finite time isometry of $V_T = \sum_{l=0}^{T-1} \tilde{U}^-_{l+d,d}(\tilde{U}^-_{l+d,d})^\top$, *i.e.*, the sample covariance matrix.

Lemma 5.1 *Define*

$$T_0(\delta, d) = cm^2 d \log^2(d) \log^2(m^2/\delta) + cd \log^3(2d)$$

where c is some universal constant. Define the sample covariance matrix $V_T \coloneqq \sum_{l=0}^{T-1} \tilde{U}^-_{l+d,d}(\tilde{U}^-_{l+d,d})^\top$. We have with probability $1-\delta$ and for $T > T_0(\delta, d)$

$$\frac{1}{2}TI \preceq V_T \preceq \frac{3}{2}TI \tag{11}$$

Lemma 5.1 allows us to write Eq. (10) as $\hat{\mathcal{H}}_{0,d,d} = \Big(\sum_{l=0}^{T-1} \tilde{Y}^+_{l+d+1,d}(\tilde{U}^-_{l+d,d})^\top\Big)\Big(\sum_{l=0}^{T-1} \tilde{U}^-_{l+d,d}(\tilde{U}^-_{l+d,d})^\top\Big)^{-1}$ with high probability and upper bound estimation error for $d \times d$ principal submatrix.

Theorem 5.1 *Fix d and let $\hat{\mathcal{H}}_{0,d,d}$ be the output of Algorithm 1. Then for any $0 < \delta < 1$ and $T \geq T_0(\delta, d)$, we have with probability at least $1-\delta$*

$$\Big|\Big|\hat{\mathcal{H}}_{0,d,d} - \mathcal{H}_{0,d,d}\Big|\Big|_2 \leq 4\sigma\sqrt{\frac{1}{T}}\sqrt{pd + \log\frac{1}{\delta} + m}.$$

Here $T_0(\delta, d) = cm^2 d \log^2(d) \log^2(m^2/\delta) + cd \log^3(2d)$, c is a universal constant and $\sigma = \max(\sigma_A, \sigma_B, \sigma_C, \sigma_D)$ from Table 1.

Proof We outline the proof here. Recall Eq. (8), (9). Then for a fixed d

$$\hat{\mathcal{H}}_{0,d,d} = \Big(\sum_{l=0}^{T-1} \tilde{Y}^+_{l+d+1,d}(\tilde{U}^-_{l+d,d})^\top\Big)V_T^+.$$

Then the identification error is

$$\begin{aligned}\left|\left|\hat{\mathcal{H}}_{0,d,d}-\mathcal{H}_{0,d,d}\right|\right|_2 &= \Big|\Big|V_T^+\Big(\sum_{l=0}^{T-1}\tilde{U}^-_{l+d,d}\tilde{U}^{+\top}_{l+d+1,d}\mathcal{T}_{0,d}^\top+\tilde{U}^-_{l+d,d}\tilde{U}^{-\top}_{l,l}\mathcal{H}^\top_{d,d,l}+\tilde{U}^-_{l+d,d}\tilde{w}^{+\top}_{l+d+1,d}\\ &\quad+\tilde{U}^-_{l+d,d}\tilde{\eta}^{-\top}_{l+d,d}\mathcal{O}^\top_{0,d,d}+\tilde{U}^-_{l+d,d}\tilde{\eta}^{+\top}_{l+d+1,d}\mathcal{T}\mathcal{O}^\top_{0,d}+\tilde{U}^-_{l+d,d}\tilde{\eta}^{-\top}_{l,l}\mathcal{O}^\top_{d,d,l}\Big)\Big|\Big|_2\\ &= ||V_T^+E||_2 \end{aligned} \tag{12}$$

with

$$\begin{aligned}E=&\sum_{l=0}^{T-1}\tilde{U}^-_{l+d,d}\tilde{U}^{+\top}_{l+d+1,d}\mathcal{T}_{0,d}^\top+\tilde{U}^-_{l+d,d}\tilde{U}^{-\top}_{l,l}\mathcal{H}^\top_{d,d,l}+\tilde{U}^-_{l+d,d}\tilde{w}^{+\top}_{l+d+1,d}\\ &+\tilde{U}^-_{l+d,d}\tilde{\eta}^{-\top}_{l+d,d}\mathcal{O}^\top_{0,d,d}+\tilde{U}^-_{l+d,d}\tilde{\eta}^{+\top}_{l+d+1,d}\mathcal{T}\mathcal{O}^\top_{0,d}+\tilde{U}^-_{l+d,d}\tilde{\eta}^{-\top}_{l,l}\mathcal{O}^\top_{d,d,l}.\end{aligned}$$

By Lemma 5.1 we have, whenever $T \geq T_0(\delta, d)$, with probability at least $1-\delta$

$$\frac{TI}{2} \preceq V_T \preceq \frac{3TI}{2}. \tag{13}$$

This ensures that, with high probability, that V_T^{-1} exists and decays as $O(T^{-1})$. The next step involves showing that $||E||_2$ grows at most as $\sqrt{T}$ with high probability. This is reminiscent of Theorem 2.2 and the theory of self–normalized martingales. However, unlike that cases the conditional sub-Gaussianity requirements do not hold here. For example, let $\boldsymbol{\mathcal{F}}_l = \sigma(\eta_1, \ldots, \eta_l)$ then $\mathbb{E}[v^\top \tilde{\eta}^-_{l+1,l+1}|\boldsymbol{\mathcal{F}}_l] \neq 0$ for all v since $\{\tilde{\eta}^-_{l+1,l+1}\}_{l=0}^{T-1}$ is not an independent sequence. As a result it is not immediately obvious on how to apply Theorem 2.2 to our case. Under the event when Eq. (13) holds (which happens with high probability), a careful analysis of the normalized cross terms, *i.e.*, $V_T^{-1/2}E$ shows that $||V_T^{-1/2}E||_2 = O(1)$ with high probability. This is summarized in Propositions 11.1-11.3. The idea is to decompose E into a linear combination of independent subgaussians and reduce it to a form where we can apply Theorem 2.2. This comes at the cost of additional scaling in the form of system dependent constants – such as the $\mathcal{H}_\infty$–norm. Then we can conclude with high probability that $||\hat{\mathcal{H}} - \mathcal{H}_{0,d,d}||_2 \leq ||V_T^{-1/2}||_2||V_T^{-1/2}E||_2 \leq T^{-1/2}O(1)$. The full proof has been deferred to Section 11.1 in Appendix 11. ■

Remark 9 *Recall $\mathcal{D}(T)$ from Table 1. Since*

$$d \in \mathcal{D}(T) \implies T \geq T_0(\delta, d)$$

we can restate Theorem 5.1 as follows: for a fixed T, we have with probability at least $1-\delta$ that

$$\left|\left|\hat{\mathcal{H}}_{0,d,d}-\mathcal{H}_{0,d,d}\right|\right|_2 \leq 4\sigma\sqrt{\frac{1}{T}}\sqrt{pd+\log\frac{1}{\delta}+m}$$

when $d \in \mathcal{D}(T)$.

We next present bounds on σ in Theorem 5.1. From the perspective of model selection in later sections, we require that σ be known. In the next proposition we present two bounds on σ, the first one depends on unknown parameters and recovers the precise dependence on d. The second bound is an apriori known upper bound and incurs an additional factor of $\sqrt{d}$.

Proposition 5.1 *σ upper bound independent of d:*

$$\sigma \leq \frac{c_n}{(1-\rho(A))^2}$$

where c_n depends only on n.

σ upper bound dependent on d:

$$\sigma \leq \beta R\sqrt{d}.$$

where R is the noise-to-signal ratio as in Table 1

Proof

By Gelfand's formula, since $||A^d||_2 \leq c(n)\rho_{\max}(A)^d$ where $\rho_{\max}(A) < 1$ and $c(n)$ is a constant that only depends on n, it implies that

$$\sigma_A = \sum_{l=0}^{d} ||CA^lB||_2 \leq \sum_{l=0}^{\infty} ||CA^lB||_2 \leq \sum_{l=0}^{\infty} c(n)\rho(A)^l = \frac{c(n)}{1-\rho(A)},$$

and

$$||\mathcal{T}_{d+k,T}||_2 \leq \sum_{l=0}^{T-1} ||CA^{d+k+l}B||_2 \leq \frac{c(n)\rho(A)^{d+k}}{1-\rho(A)}.$$

Then

$$\sigma_C = \sqrt{\sigma\left(\sum_{k=1}^{d} \mathcal{T}_{d+k,T}^{\top}\mathcal{T}_{d+k,T}\right)} \leq \frac{c(n)\rho(A)^d}{(1-\rho(A))^2} \leq \frac{c(n)}{(1-\rho(A))^2}.$$

Similarly, there exists a finite upper bound on σ_B, σ_D by replacing CA^lB and $\mathcal{T}_{d+k,T}$ with CA^l and $\mathcal{TO}_{d+k,T}$ respectively. For the d independent upper bound, we have

$$\sigma_A = \sum_{l=0}^{d} ||CA^lB||_2 \leq \sqrt{d}\sqrt{\sum_{l=0}^{d} ||CA^lB||_2^2} \leq \sqrt{d}||M||_H \leq \sqrt{d}\beta.$$

Since $\sigma\left(\mathcal{T}_{d+k,T}^{\top}\mathcal{T}_{d+k,T}\right) \leq \beta$, then

$$\sigma_C = \sqrt{\sigma\left(\sum_{k=1}^{d} \mathcal{T}_{d+k,T}^{\top}\mathcal{T}_{d+k,T}\right)} \leq \beta\sqrt{d}.$$

For the σ_B, σ_D we get an extra R because $\mathcal{TO}_{0,\infty} \leq \beta R$. ■

The key feature of the data dependent upper bound is that it only depends on β and R which are known apriori.

Recall that $G_d = [CB, CAB, \ldots, CA^{d-1}B]$, *i.e.*, the d-order FIR truncation of $G(z)$. Since the p rows of the $\mathcal{H}_{0,d,d}$ matrix corresponds to G_d we can obtain estimators for any d-order FIR.

Corollary 5.1 *Let $\widehat{G}_d = \hat{\mathcal{H}}_{0,d,d}[1:p,:]$ denote the first p-rows of $\hat{\mathcal{H}}_{0,d,d}$. Then for any $0 < \delta < 1$ and $T \geq T_0(\delta, d)$, we have with probability at least $1-\delta$,*

$$||\widehat{G}_d - G_d||_2 \leq 4\sigma\sqrt{\frac{1}{T}}\sqrt{pd + \log\frac{1}{\delta} + m}.$$

Proof Proof follows because $G_d = \mathcal{H}_{0,d,d}[1:p,:]$ and Theorem 5.1. ■

Next, we show that the error in Theorem 5.1 is minimax optimal (up to logarithmic factors) and cannot be improved by any estimation method.

Proposition 5.2 *Let $\sqrt{T} \geq c$ where c is an absolute constant. Then for any estimator $\hat{\mathcal{H}}$ of $\mathcal{H}_{0,\infty,\infty}$ we have*

$$\sup_{\hat{\mathcal{H}}} \mathbb{E}[||\hat{\mathcal{H}} - \mathcal{H}_{0,\infty,\infty}||_2] \geq c_n \cdot \sqrt{\frac{\log T}{T}}$$

where $c_n > 0$ is a constant that is independent of T but can depend on system level parameters.

Proof Assume the contrary that

$$\sup_{\hat{\mathcal{H}}} \mathbb{E}[||\hat{\mathcal{H}} - \mathcal{H}_{0,\infty,\infty}||_2] = o\left(\sqrt{\frac{\log T}{T}}\right).$$

Then recall that $[\mathcal{H}_{0,\infty,\infty}]_{1:p,:} = [CB, CAB, \ldots,]$ and $G(z) = z^{-1}CB + z^{-2}CAB + \ldots$. Similarly we have $\hat{G}(z)$. Define

$$||G - \hat{G}||_2 = \sqrt{\sum_{k=0}^{\infty} ||CA^kB - \hat{C}\hat{A}^k\hat{B}||_2^2}.$$

If $\sup_{\hat{\mathcal{H}}} \mathbb{E}[||\hat{\mathcal{H}} - \mathcal{H}_{0,\infty,\infty}||_2] = o\Big(\sqrt{\frac{\log T}{T}}\Big)$, then since $||\hat{\mathcal{H}} - \mathcal{H}_{0,\infty,\infty}||_2 \geq ||G - \hat{G}||_2$ we can conclude that

$$\mathbb{E}[||G - \hat{G}||_2] = o\left(\sqrt{\frac{\log T}{T}}\right)$$

which contradicts Theorem 5 in (Goldenshluger, 1998). Thus, $\sup_{\hat{\mathcal{H}}} \mathbb{E}[||\hat{\mathcal{H}} - \mathcal{H}_{0,\infty,\infty}||_2] \geq c_n \cdot \sqrt{\frac{\log T}{T}}$. ■

5.2 Model Selection

At a high level, we want to choose $\hat{\mathcal{H}}_{0,\hat{d},\hat{d}}$ from $\{\hat{\mathcal{H}}_{0,d,d}\}_{d=1}^{T}$ such that $\hat{\mathcal{H}}_{0,\hat{d},\hat{d}}$ is a good estimator of $\mathcal{H}_{0,\infty,\infty}$. Our idea of model selection is motivated by (Goldenshluger, 1998). For any $\hat{\mathcal{H}}_{0,d,d}$, the error from $\mathcal{H}_{0,\infty,\infty}$ can be broken as:

$$||\hat{\mathcal{H}}_{0,d,d} - \mathcal{H}_{0,\infty,\infty}||_2 \leq \underbrace{||\hat{\mathcal{H}}_{0,d,d} - \mathcal{H}_{0,d,d}||_2}_{=\text{Estimation Error}} + \underbrace{||\mathcal{H}_{0,d,d} - \mathcal{H}_{0,\infty,\infty}||_2}_{=\text{Truncation Error}}.$$

We would like to select a $d = \hat{d}$ such that it balances the truncation and estimation error in the following way:

$$c_2 \cdot \text{Data dependent upper bound } \geq c_1 \cdot \text{Estimation Error} \geq \text{Truncation Error}$$

where c_i are absolute constants. Such a balancing ensures that

$$||\hat{\mathcal{H}}_{0,\hat{d},\hat{d}} - \mathcal{H}_{0,\infty,\infty}||_2 \leq c_2 \cdot (1/c_1 + 1) \cdot \text{Data dependent upper bound}\,. \tag{14}$$

Note that such a balancing is possible because the estimation error increases as d grows and truncation error decreases with d. Furthermore, a data dependent upper bound for estimation error can be obtained from Theorem 5.1. Unfortunately (C, A, B) are unknown and it is not immediately clear on how to obtain such a bound for truncation error.

To achieve this, we first define a truncation error proxy, *i.e.*, how much do we truncate if a specific $\hat{\mathcal{H}}_{0,d,d}$ is used. For a given d, we look at $||\hat{\mathcal{H}}_{0,d,d} - \hat{\mathcal{H}}_{0,l,l}||_2$ for $l \in \mathcal{D}(T) \geq d$. This measures the additional error incurred if we choose $\hat{\mathcal{H}}_{0,d,d}$ as an estimator for $\mathcal{H}_{0,\infty,\infty}$ instead of $\hat{\mathcal{H}}_{0,l,l}$ for $l > d$. Then we pick $\hat{d}$ as follows:

$$\hat{d} := \inf\left\{ d \middle| ||\hat{\mathcal{H}}_{0,d,d} - \hat{\mathcal{H}}_{0,l,l}||_2 \leq 16\beta R \cdot \alpha(l) \quad \forall l \in \mathcal{D}(T) \geq d \right\}. \tag{15}$$

Recall that $\alpha(l) = \sqrt{\frac{l\log(l/\delta)+pl^2+ml}{T}}$, where $\sqrt{\frac{\log(l/\delta)+pl+m}{T}}$ denotes how much estimation error is incurred in learning $l \times l$ Hankel submatrix, the extra $\beta\sqrt{l}$ is incurred because we need a data dependent, albeit coarse, upper bound on the estimation error.

A key step will be to show that for any $l \geq d$, whenever

$$||\hat{\mathcal{H}}_{0,d,d} - \hat{\mathcal{H}}_{0,l,l}||_2 \leq c\beta R \cdot \alpha(l)$$

ensures that

$$||\hat{\mathcal{H}}_{0,d,d} - \mathcal{H}_{0,\infty,\infty}||_2 \leq c\beta R \cdot \alpha(l) \quad \text{and} \quad ||\hat{\mathcal{H}}_{0,l,l} - \mathcal{H}_{0,\infty,\infty}||_2 \leq c\beta R \cdot \alpha(l)$$

and there is no gain in choosing a larger Hankel submatrix estimate. By picking the smallest d for which such a property holds for all larger Hankel submatrices, we ensure that a regularized model is estimated that "agrees" with the data.

Algorithm 2 Choice of d

Output $\hat{\mathcal{H}}_{0,\hat{d},\hat{d}}, \quad \hat{d}$

1: $\mathcal{D}(T) = \left\{d \middle| d \leq \frac{T}{cm^2 \log^3(Tm/\delta)}\right\}, \alpha(h) = \sqrt{h}\left(\sqrt{\frac{m+hp+\log\left(\frac{T}{\delta}\right)}{T}}\right).$

2: $d_0(T,\delta) = \inf\left\{l \middle| ||\hat{\mathcal{H}}_{0,l,l} - \hat{\mathcal{H}}_{0,h,h}||_2 \leq 16\beta R(\alpha(h) + 2\alpha(l)) \ \ \forall h \in \mathcal{D}(T), h \geq l\right\}.$

3: $\hat{d} = \max\left(d_0(T,\delta), \log\left(\frac{T}{\delta}\right)\right)$

4: **return** $\hat{\mathcal{H}}_{0,\hat{d},\hat{d}}, \quad \hat{d}$

We now state the main estimation result for $\mathcal{H}_{0,\infty,\infty}$ for $d = \hat{d}$ as chosen in Algorithm 2. Define

$$T_*(\delta) = \inf\left\{T \middle| d_*(T,\delta) \in \mathcal{D}(T), \ \ d_*(T,\delta) \leq 2d_*\left(\frac{T}{256}, \delta\right)\right\} \tag{16}$$

where

$$d_*(T,\delta) = \inf\left\{d \middle| 16\beta R\alpha(d) \geq ||\mathcal{H}_{0,d,d} - \mathcal{H}_{0,\infty,\infty}||_2\right\}. \tag{17}$$

A close look at Eq. (17) reveals that picking $d = d_*(T,\delta)$ ensures the balancing of Eq. (14). However, $d_*(T,\delta)$ depends on unknown quantities and is unknown. In such a case, $\hat{d}$ in Eq. (15) becomes a proxy for $d_*(T,\delta)$. From an algorithmic stand point, we no longer need any unknown information; the unknown parameter only appear in $T_*(\delta)$, which is only required to make the theoretical guarantee of Theorem 5.2 below.

Theorem 5.2 *Whenever we have $T \geq T_*(\delta)$ we have with probability at least $1-\delta$ that*

$$||\hat{\mathcal{H}}_{0,\hat{d},\hat{d}} - \mathcal{H}_{0,\infty,\infty}||_2 \leq 12c\beta R\left(\sqrt{\frac{m\hat{d} + p\hat{d}^2 + \hat{d}\log\frac{T}{\delta}}{T}}\right).$$

The proof of Theorem 5.2 can be found as Proposition 13.8 in Appendix 13. We see that the error between $\hat{\mathcal{H}}_{0,\hat{d},\hat{d}}$ and $\mathcal{H}_{0,\infty,\infty}$ can be upper bounded by a purely data dependent quantity. The next proposition shows that $\hat{d}$ does not grow more that logarithmically in T.

Proposition 5.3 *Let $T \geq T_*(\delta)$, $d_*(T,\delta)$ be as in Eq. (17). Then with probability at least $1-\delta$ we have*

$$\hat{d} \leq d_*(T,\delta) \vee \log\left(\frac{T}{\delta}\right).$$

Furthermore,

$$d_*(T,\delta) \leq \frac{c\log\left(cT + \log\frac{1}{\delta}\right) - \log R + \log\beta}{\log\frac{1}{\rho(A)}}.$$

The effect of unknown quantities, such as the spectral radius, are subsumed in the finite time condition $T \geq T_*(\delta)$ and appear in an upper bound for $\hat{d}$; however this information is not needed from an algorithmic perspective as the selection of $\hat{d}$ is agnostic to the knowledge of $\rho(A)$. The proof of proposition can be found as Propositions 13.7 and 13.4.

5.3 Parameter Recovery

Next we discuss finding the system parameters. To obtain system parameters we use a balanced truncation algorithm on $\hat{\mathcal{H}}_{0,\hat{d},\hat{d}}$ where $\hat{d}$ is the output of Algorithm 2. The details are summarized in Algorithm 3 where $\mathcal{H} = \hat{\mathcal{H}}_{0,\hat{d},\hat{d}}$.

Algorithm 3 Hankel2Sys$(T, \hat{d}, k, m, p)$

Input T = Horizon for Learning
$\hat{d}$ = Hankel Size
m = Input dimension
p = Output dimension
Output System Parameters: $(\hat{C}_{\hat{d}}, \hat{A}_{\hat{d}}, \hat{B}_{\hat{d}})$

1: $\mathcal{H} = \mathcal{H}_{0,\hat{d},\hat{d}}$
2: Pad $\mathcal{H}$ with zeros to make of dimension $4p\hat{d} \times 4m\hat{d}$
3: $U, \Sigma, V \leftarrow$ SVD of $\mathcal{H}$
4: $U_{\hat{d}}, V_{\hat{d}} \leftarrow$ top $\hat{d}$ singular vectors
5: $\hat{C}_{\hat{d}} \leftarrow$ first p rows of $U_{\hat{d}}\Sigma_{\hat{d}}^{1/2}$
6: $\hat{B}_{\hat{d}} \leftarrow$ first m columns of $\Sigma_{\hat{d}}^{1/2}V_{\hat{d}}^\top$
7: $Z_0 = [U_{\hat{d}}\Sigma_{\hat{d}}^{1/2}]_{1:4p\hat{d}-p,:}, Z_1 = [U_{\hat{d}}\Sigma_{\hat{d}}^{1/2}]_{p+1:,:}$
8: $\hat{A}_{\hat{d}} \leftarrow (Z_0^\top Z_0)^{-1} Z_0^\top Z_1$.
9: **return** $(\hat{C}_{\hat{d}}, \hat{A}_{\hat{d}}, \hat{B}_{\hat{d}})$

To state the main result we define a quantity that measures the singular value weighted subspace gap of a matrix S:

$$\Gamma(S, \epsilon) = \sqrt{\sigma^1_{\max}/\zeta_1^2 + \sigma^2_{\max}/\zeta_2^2 + \ldots + \sigma^l_{\max}/\zeta_l^2},$$

where $S = U\Sigma V^\top$ and Σ is arranged into blocks of singular values such that in each block i we have $\sup_j \sigma^i_j - \sigma^i_{j+1} \leq \epsilon$, *i.e.*,

$$\Sigma = \begin{bmatrix} \Lambda_1 & 0 & \ldots & 0 \\ 0 & \Lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \ldots & \Lambda_l \end{bmatrix}$$

where Λ_i are diagonal matrices, σ^i_j is the j^{th} singular value in the block Λ_i and $\sigma^i_{\min}, \sigma^i_{\max}$ are the minimum and maximum singular values of block i respectively. Furthermore,

$$\zeta_i = \min\left(\sigma^{i-1}_{\min} - \sigma^i_{\max}, \sigma^i_{\min} - \sigma^{i+1}_{\max}\right)$$

for $1 < i < l$, $\zeta_1 = \sigma^1_{\min} - \sigma^2_{\max}$ and $\zeta_l = \min\left(\sigma^{l-1}_{\min} - \sigma^l_{\max}, \sigma^l_{\min}\right)$. Informally, the ζ_i measure the singular value gaps between each blocks. It should be noted that l, the number of separated blocks, is a function of ϵ itself. For example: if $\epsilon = 0$ then the number of blocks correspond to the number of distinct singular values. On the other hand, if ϵ is very large then $l = 1$.

Theorem 5.3 *Let M be the true unknown model and*

$$\epsilon = 12c\beta R\left(\sqrt{\frac{m\hat{d} + p\hat{d}^2 + \hat{d}\log\frac{T}{\delta}}{T}}\right).$$

Then whenever $T \geq T_*(\delta)$, we have with probability at least $1-\delta$:

$$\left.\begin{array}{l} ||C_{\hat{d}} - \hat{C}_{\hat{d}}||_2 \\ ||B_{\hat{d}} - \hat{B}_{\hat{d}}||_2 \\ ||A_{\hat{d}} - \hat{A}_{\hat{d}}||_2 \end{array}\right\} \leq \bar{\gamma}\epsilon\Gamma(\hat{\mathcal{H}}_{0,\hat{d},\hat{d}}, 2\epsilon) + \bar{\gamma}\sup_{1\leq i\leq\hat{d}}\left(\sqrt{\hat{\sigma}^i_{\max}} - \sqrt{\hat{\sigma}^i_{\min}}\right) + \bar{\gamma}\cdot\frac{\epsilon\wedge\sqrt{\hat{\sigma}_{\hat{d}}\epsilon}}{\sqrt{\hat{\sigma}_{\hat{d}}}}$$

where $\sup_{1\leq i\leq\hat{d}}\sqrt{\hat{\sigma}^i_{\max}} - \sqrt{\hat{\sigma}^i_{\min}} \leq \frac{2}{\sqrt{\hat{\sigma}_{\hat{d}}}}\epsilon\hat{d}\wedge\sqrt{2\hat{d}\epsilon}$ *and* $\bar{\gamma} = \max(4\gamma, 8)$.

Theorem 5.3 holds for all $k \leq \hat{d}$ and proof follows directly from Theorem 13.8 where we show

$$||\hat{\mathcal{H}}_{0,\hat{d},\hat{d}} - \mathcal{H}_{0,\infty,\infty}||_2 \leq \epsilon$$

and Proposition 14.2. Theorem 5.3 provides an error bound between parameters (of model order $\hat{d}$) when true order is unknown. The subspace gap measure, $\Gamma(\hat{\mathcal{H}}_{0,\hat{d},\hat{d}}, 2\epsilon)$, is bounded even when $\epsilon = 0$. To see this, note that when $\epsilon = 0$, $\hat{\mathcal{H}}_{0,\hat{d},\hat{d}}$ corresponds exactly to $\mathcal{H}_{0,\hat{d},\hat{d}}$. In that case, the number of blocks correspond to the number of distinct singular values of $\mathcal{H}_{0,\hat{d},\hat{d}}$, and ζ_{n_i} then corresponds to singular value gap between the unequal singular values. As a result $\Gamma(\hat{\mathcal{H}}_{0,\hat{d},\hat{d}}, 2\epsilon) = \Delta < \infty$. Then the bound decays as $\epsilon = O\left(\sqrt{\hat{d}^2/T}\right)$ for singular values $\hat{\sigma}_{\hat{d}} > \hat{d}\epsilon$, but for much smaller singular values the bound decays as $\sqrt{\epsilon} = O\left(\left(d^2/T\right)^{1/4}\right)$.

To shed more light on the behavior of our bounds, we consider the special case of known order. If n is the model order, then we can set $\hat{d} = n$. If $\sigma_i = \sigma_i(\mathcal{H}_{0,\infty,\infty})$, then for large enough T one can ensure that

$$\min_{\sigma_i\neq\sigma_{i+1}}(\sigma_i - \sigma_{i+1})/2 > \epsilon,$$

i.e., ϵ is less than the singular value gap and small enough that the spectrum of $\hat{\mathcal{H}}_{0,n,n}$ is very close to that of $\mathcal{H}_{0,\infty,\infty}$. Consequently $\hat{\sigma}_n \geq \sigma_n/2$ and we have that

$$\left.\begin{array}{l} ||C_n - \hat{C}_n||_2 \\ ||B_n - \hat{B}_n||_2 \\ ||A_n - \hat{A}_n||_2 \end{array}\right\} \leq \bar{\gamma}\epsilon\Delta + \bar{\gamma}\epsilon/\sqrt{\sigma_n} = c\beta\bar{\gamma}R\left(\sqrt{\frac{pn^2 + n\log\frac{T}{\delta}}{\sigma_n T}}\right). \tag{18}$$

This upper bound is (nearly) identical to the bounds obtained in Oymak and Ozay (2018) for the known order case. We get an improvement in the bounds when $\sigma_n \leq \frac{1}{n}$, which is a consequence of the fact that we know where to threshold our Hankel matrix. The major advantage of our result is that we do not require any information/assumption on the LTI system besides β. Nonparametric approaches to estimating β have been studied in Tu et al. (2017).

5.4 Order Estimation Lower Bound

In Theorem 5.3 it is shown that whenever $T = \Omega\left(\frac{1}{\sigma^2_{\hat{d}}}\right)$ we can find an accurate $\hat{d}$–order approximation. Now we show that if $T = O\left(\frac{1}{\sigma^2_{\hat{d}}}\right)$ then there is always some non–zero probability with which we can not recover the singular vector corresponding to the $\sigma_{\hat{d}+1}$. We prove the following lower bound for model order estimation when inputs $\{U_t\}_{t=1}^T$ are active and bounded which we define below

Definition 10 *An input sequence $\{U_t\}_{t=1}^T$ is said to be active if U_t is allowed to depend on past history $\{U_l, Y_l\}_{l=1}^{t-1}$. The input sequence is bounded if $\mathbb{E}[U_t^\top U_t] \leq 1$ for all t.*

Active inputs allow for the case when input selection can be adaptive due to feedback.

Theorem 5.4 *Fix $\delta > 0, \zeta \in (0, 1/2)$. Let M_1, M_2 be two LTI systems and $\sigma_i^{(1)}, \sigma_i^{(2)}$ be the i^{th}-Hankel singular values respectively. Let $\frac{\sigma_1^{(1)}}{\sigma_2^{(1)}} \leq \frac{2}{\zeta}$ and $\sigma_2^{(2)} = 0$. Then whenever $T \leq \frac{CR^2}{\zeta^2} \log \frac{2}{\delta}$ we have*

$$\sup_{M \in \{M_1, M_2\}} \mathbb{P}_{Z_T \sim M}(\textit{order}(\hat{M}(Z_T)) \neq \textit{order}(M)) \geq \delta$$

Here $Z_T = \{U_t, Y_t\}_{t=1}^T \sim M$ means M generates T data points $\{Y_t\}_{t=1}^T$ in response to active and bounded inputs $\{U_t\}_{t=1}^T$ and $\hat{M}(Z_T)$ is any estimator.

Proof The proof can be found in appendix in Section 15 and involves using Fano's (or Birge's) inequality to compute the minimax risk between the probability density functions generated by two different LTI systems:

$$A_0 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ \zeta & 0 & 0 \end{bmatrix}, A_1 = A_0, B_0 = \begin{bmatrix} 0 \\ 0 \\ \sqrt{\beta}/R \end{bmatrix}, B_1 = \begin{bmatrix} 0 \\ \sqrt{\beta}/R \\ \sqrt{\beta}/R \end{bmatrix}, C_0 = \begin{bmatrix} 0 & 0 & \sqrt{\beta}R \end{bmatrix}, C_1 = C_0. \tag{19}$$

A_0, A_1 are Schur stable whenever $|\zeta| < 1$. ■

Theorem 5.4 shows that when the time required to recover higher order models depends inversely on the condition number, where the condition number is the ratio of largest and least singular values of the Hankel matrix. Specifically, to correctly distinguish between an order 1 and order 2 model $T \geq \Omega(2/\zeta^2)$ where ζ is the condition number of the 2-order model. We compare this to our upper bound in Theorem 5.3 and Eq. (18), assume $\Gamma(\hat{\mathcal{H}}_{0,\hat{d},\hat{d}}, 2\epsilon) \leq \Delta$ for all $\epsilon \in [0, 1]$ and $\hat{d}\epsilon \leq \hat{\sigma}_{\hat{d}}$, then since parameter error, $\mathcal{E}$, is upper bounded as

$$\mathcal{E} \leq c\beta\Delta R\left(\sqrt{\frac{m\hat{d} + p\hat{d}^2 + \hat{d}\log\frac{T}{\delta}}{\sigma_{\hat{d}} T}}\right),$$

we need

$$\frac{T}{\log\frac{T}{\delta}} \geq \Omega\left(\frac{\beta^2\Delta^2 R^2 \hat{d}^2}{\sigma^2_{\hat{d}}}\right)$$

to correctly identify $\hat{d}$-order model. The ratio $(\beta/\sigma_{\hat{d}})$ is equal to the condition number of the Hankel matrix. In this sense, the model selection followed by singular value thresholding is not too conservative in terms of R (the signal-to-noise ratio) and conditioning of the Hankel matrix.

6. Experiments

The experiments in this paper are for the single trajectory case. A detailed analysis for system identification from multiple trajectories can be found in Tu et al. (2017). Suppose that the LTI system generating data, M, has transfer function given by

$$G(z) = \alpha_0 + \sum_{l=1}^{149} \alpha_l \rho^l z^{-l}, \quad \rho < 1 \tag{20}$$

where $\alpha_i \sim \mathcal{N}(0,1)$. M is a finite dimensional LTI system or order 150 with parameters as $M = (C \in \mathbb{R}^{1\times 150}, A \in \mathbb{R}^{150\times 150}, B \in \mathbb{R}^{150\times 1})$. For these illustrations, we assume a balanced system and choose $R = 1, \delta = 0.05$. We estimate $\beta_{0.6} = 15, \beta_{0.9} = 40, \beta_{0.99} = 140$, pick $U_t \sim \mathcal{N}(0,1)$ and $\{w_t, \eta_t\} \sim \{\mathcal{N}(0,1), \mathcal{N}(0,I)\}$ respectively. We note that our algorithm requires the knowledge of universal constant c. Theoretically, it can be shown that $c < 100$ but in practice a value $c \leq 16$ works well for simulations.

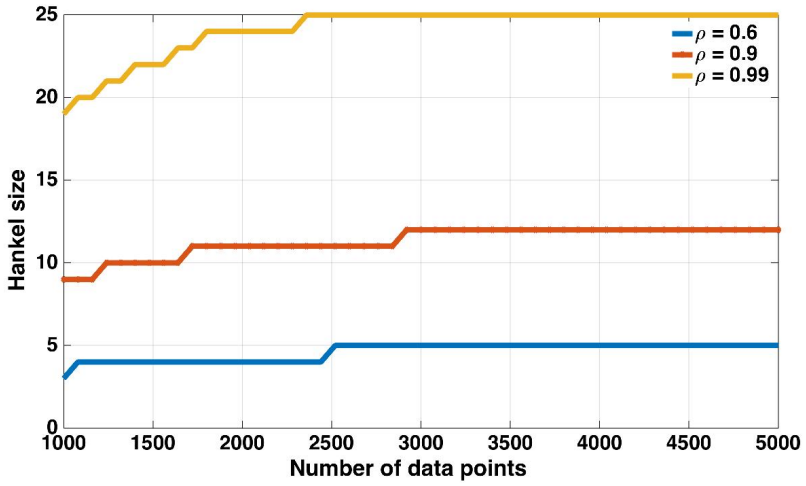

Figure 1: Variation of Hankel size $= \hat{d}$ with T for different values of ρ

Fig. 1 shows how $d = \hat{d}$ change with the number of data points for different values of ρ. When $\rho = 0.6$, *i.e.*, small, $\hat{d}$ does not grow too big with T even when the number of data points is increased. This shows that a small model order is sufficient to specify system dynamics. On the other hand, when $\rho = 0.99$, *i.e.*, closer to instability the $\hat{d}$ required is much larger, indicating the need for a higher order. Although $\hat{d}$ implicitly captures the effect of spectral radius, the knowledge of ρ is not required for $\hat{d}$ selection.

In principle, our algorithm increases the Hankel size to the "appropriate" size as the data increases. We compare this to a deterministic growth policy $d = \log(T)$ and the SSREGEST

algorithm Ljung et al. (2015). The SSREGEST algorithm first learns a large model from data and then performs model reduction to obtain a final model. In contrast, we go to reduced model directly by picking a small $\hat{d}$. This reduces the sensitivity to noise.

In Fig. 2 shows the model errors for a deterministic growth policy $d = \log(T)$ and our algorithm. Although the difference is negligible when $\rho = 0.6$ (small), we see that our algorithm does better $\rho = 0.99$ due to its adaptive nature, *i.e.*, $\hat{d}$ responds faster for our algorithm.

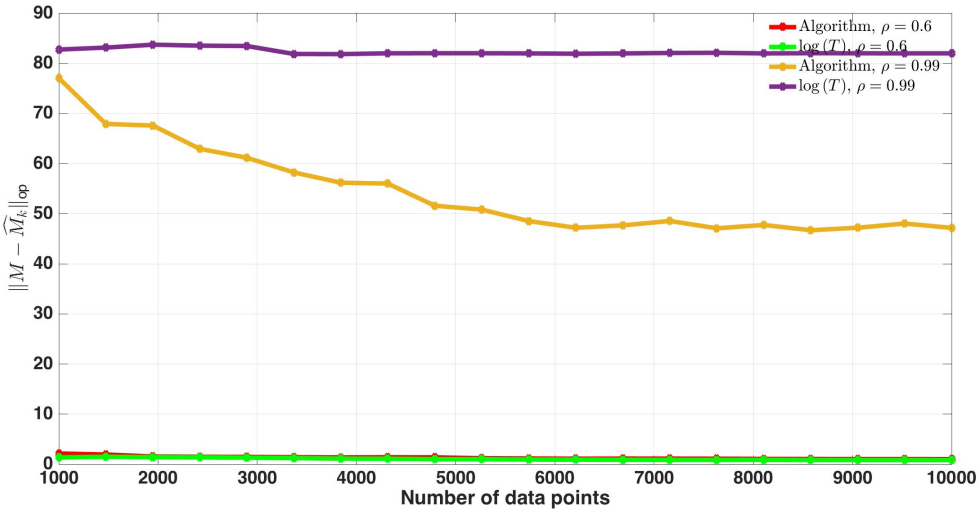

Figure 2: Variation of $||M - \widehat{M}_k||_{\text{op}}$ for different values of ρ. Here $k = \hat{d}$ for our algorithm and $k = \log(T)$. Furthermore, $||\cdot||_{\text{op}}$ is the Hankel norm.

Finally, for the case when $\rho = 0.9, \beta = 40$, we show the model errors for SSREGEST and our algorithm as T increases. Although asymptotically both algorithms perform the same, it is clear that for small T our algorithm is more robust to the presence of noise.

T	SSREGEST	Our Algorithm
500	6.21 ± 1.35	13.37 ± 3.7
≈ 850	30.20 ± 7.55	11.25 ± 2.89
≈ 1200	26.80 ± 8.94	9.83 ± 2.60
1500	23.27 ± 10.65	9.17 ± 2.30
2000	26.38 ± 12.88	7.70 ± 1.60

7. Discussion

We propose a new approach to system identification when we observe only finite noisy data. Typically, the order of an LTI system is large and unknown and a priori parametrizations may fail to yield accurate estimates of the underlying system. However, our results suggest that there always exists a lower order approximation of the original LTI system that can be learned with high probability. The central theme of our approach is to recover a good lower order approximation that can be accurately learned. Specifically, we show that identification

of such approximations is closely related to the singular values of the system Hankel matrix. In fact, the time required to learn a $\hat{d}$–order approximation scales as $T = \Omega(\frac{\beta^2}{\sigma_{\hat{d}}^2})$ where $\sigma_{\hat{d}}$ is the $\hat{d}$–the singular value of system Hankel matrix. This means that system identification does not explicitly depend on the model order n, rather depends on n through σ_n. As a result, in the presence of finite data it is preferable to learn only the "significant" (and perhaps much smaller) part of the system when n is very large and $\sigma_n \ll 1$. Algorithm 1 and 3 provide a guided mechanism for learning the parameters of such significant approximations with optimal rules for hyperparameter selection given in Algorithm 2.

Future directions for our work include extending the existing low–rank optimization-based identification techniques, such as (Fazel et al., 2013; Grussler et al., 2018), which typically lack statistical guarantees. Since Hankel based operators occur quite naturally in general (not necessarily linear) dynamical systems, exploring if our methods could be extended for identification of such systems appears to be an exciting direction.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Behçet Açıkmeşe, John M Carson, and Lars Blackmore. 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.

Anish Agarwal, Muhammad Jehangir Amjad, Devavrat Shah, and Dennis Shen. Time series analysis via matrix estimation. *arXiv preprint arXiv:1802.09064*, 2018.

Zeyuan Allen-Zhu and Yuanzhi Li. Lazysvd: Even faster svd decomposition yet without agonizing pain. In *Advances in Neural Information Processing Systems*, pages 974–982, 2016.

Dietmar Bauer. Order estimation for subspace methods. *Automatica*, 37(10):1561–1573, 2001.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

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

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems. *arXiv preprint arXiv:1710.01852*, 2017.

Maryam Fazel, Ting Kei Pong, Defeng Sun, and Paul Tseng. Hankel matrix rank minimization with applications to system identification and realization. *SIAM Journal on Matrix Analysis and Applications*, 34(3):946–977, 2013.

Keith Glover. All optimal hankel-norm approximations of linear multivariable systems and their l_∞-error bounds. *International journal of control*, 39(6):1115–1193, 1984.

Keith Glover. Model reduction: a tutorial on hankel-norm methods and lower bounds on l2 errors. *IFAC Proceedings Volumes*, 20(5):293–298, 1987.

Alexander Goldenshluger. Nonparametric estimation of transfer functions: rates of convergence and adaptation. *IEEE Transactions on Information Theory*, 44(2):644–658, 1998.

Christian Grussler, Anders Rantzer, and Pontus Giselsson. Low-rank optimization with convex constraints. *IEEE Transactions on Automatic Control*, 2018.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *arXiv preprint arXiv:1609.05191*, 2016.

Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Spectral filtering for general linear dynamical systems. *arXiv preprint arXiv:1802.03981*, 2018.

BL Ho and Rudolph E Kalman. Effective construction of linear state-variable models from input/output functions. *at-Automatisierungstechnik*, 14(1-12):545–548, 1966.

S Kung and D Lin. Optimal hankel-norm model reductions: Multivariable systems. *IEEE Transactions on Automatic Control*, 26(4):832–852, 1981.

Lennart Ljung. *System identification: theory for the user*. Prentice-hall, 1987.

Lennart Ljung, Rajiv Singh, and Tianshi Chen. Regularization features in the system identification toolbox. *IFAC-PapersOnLine*, 48(28):745–750, 2015.

Mark Meckes et al. On the spectral norm of a random toeplitz matrix. *Electronic Communications in Probability*, 12:315–325, 2007.

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

Victor H Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008.

Tuhin Sarkar and Alexander Rakhlin. How fast can linear dynamical systems be learned? *arXiv preprint arXiv:1812.0125*, 2018.

Parikshit Shah, Badri Narayan Bhaskar, Gongguo Tang, and Benjamin Recht. Linear system identification via atomic norm regularization. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 6265–6270. IEEE, 2012.

Ritei Shibata. Selection of the order of an autoregressive model by akaike's information criterion. *Biometrika*, 63(1):117–126, 1976.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. *arXiv preprint arXiv:1802.08334*, 2018.

Stephen Tu, Ross Boczar, Andrew Packard, and Benjamin Recht. Non-asymptotic analysis of robust control from coarse-grained identification. *arXiv preprint arXiv:1707.04791*, 2017.

Stephen Tu, Ross Boczar, and Benjamin Recht. On the approximation of toeplitz operators for nonparametric $\mathcal{H}_\infty$–norm estimation. In *2018 Annual American Control Conference (ACC)*, pages 1867–1872. IEEE, 2018a.

Stephen Tu, Ross Boczar, and Benjamin Recht. Minimax lower bounds for $\mathcal{H}_\infty$-norm estimation. *arXiv preprint arXiv:1809.10855*, 2018b.

Eugene E Tyrtyshnikov. *A brief introduction to numerical analysis*. Springer Science & Business Media, 2012.

Sara van de Geer and Johannes Lederer. The bernstein–orlicz norm and deviation inequalities. *Probability theory and related fields*, 157(1-2):225–250, 2013.

Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

Peter Van Overschee and BL De Moor. *Subspace identification for linear systems: Theory—Implementation—Applications*. Springer Science & Business Media, 2012.

Saligrama R Venkatesh and Munther A Dahleh. On system identification of complex systems from finite data. *IEEE Transactions on Automatic Control*, 46(2):235–257, 2001.

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

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

K Zhou, JC Doyle, and K Glover. Robust and optimal control, 1996.

8. Preliminaries

Theorem 8.1 (Theorem 5.39 Vershynin (2010)) *if E is a $T \times md$ matrix with independent sub–Gaussian isotropic rows with subGaussian parameter 1 then with probability at least $1 - 2e^{-ct^2}$ we have*

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

Proposition 8.1 (Vershynin (2010)) *We have for any $\epsilon < 1$ and any $w \in \mathcal{S}^{d-1}$ that*

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

Theorem 8.2 (Theorem 1 Meckes et al. (2007)) *] Suppose $\{X_i \in \mathbb{R}^m\}_{i=1}^{\infty}$ are independent, $\mathbb{E}[X_j] = \mathbf{0}$ for all j, and X_{ij} are independent* subg(1) *random variables. Then $\mathbb{P}(||T_d|| \geq cm\sqrt{d \log 2d} + t) \leq e^{-t^2/d}$ where*

$$T_n = \begin{bmatrix} X_0 & X_1 & \dots & X_{d-1} \\ X_1 & X_0 & \dots & X_{d-2} \\ \vdots & \ddots & \ddots & \vdots \\ X_{d-1} & \dots & \dots & X_0 \end{bmatrix}$$

Theorem 8.3 (Hanson–Wright Inequality) *Given a subgaussian vector $X = [X_1, X_2, \dots, X_n] \in \mathbb{R}^n$ with $\sup_i \|X_i\|_{\psi_2} \leq K$. Then for any $B \in \mathbb{R}^{n \times n}$ and $t \geq 0$*

$$\mathbb{P}\Big(\|XBX^\top - \mathbb{E}[XBX^\top]\| \leq t \Big) \leq 2 \exp\left(\max\left(\frac{-ct}{K^2\|B\|}, \frac{-ct^2}{K^4\|B\|_{HS}^2} \right) \right).$$

Proposition 8.2 (Lecture 2 Tyrtyshnikov (2012)) *Suppose that L is the lower triangular part of a matrix $A \in \mathbb{R}^{d \times d}$. Then*

$$\|L\|_2 \leq \log_2(2d) \|A\|_2.$$

Let ψ be a nondecreasing, convex function with $\psi(0) = 0$ and X a random variable. Then the Orlicz norm $||X||_\psi$ is defined as

$$||X||_\psi = \inf\Big\{ \alpha > 0 : \mathbb{E}[\psi(|X|/\alpha)] \leq 1 \Big\}.$$

Let (B, d) be an arbitrary semi–metric space. Denote by $N(\epsilon, d)$ is the minimal number of balls of radius ϵ needed to cover B.

Theorem 8.4 (Corollary 2.2.5 in Van Der Vaart and Wellner (1996)) *The constant K can be chosen such that*

$$||\sup_{s,t} |X_s - X_t|||_\psi \leq K \int_0^{diam(B)} \psi^{-1}(N(\epsilon/2, d)) d\epsilon$$

where $diam(B)$ is the diameter of B and $d(s,t) = ||X_s - X_t||_\psi$.

Theorem 8.5 (Theorem 1 in Abbasi-Yadkori et al. (2011)) *Let $\{\boldsymbol{\mathcal{F}}_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\eta_t \in \mathbb{R}^m, X_t \in \mathbb{R}^d\}_{t=1}^{\infty}$ be stochastic processes such that η_t, X_t are $\boldsymbol{\mathcal{F}}_t$ measurable and η_t is $\boldsymbol{\mathcal{F}}_{t-1}$-conditionally $\mathsf{subg}(L^2)$ for some $L > 0$. For any $t \geq 0$, define $V_t = \sum_{s=1}^{t} X_s X_s'$, $S_t = \sum_{s=1}^{t} X_s \eta_{s+1}^{\top}$. Then for any $\delta > 0, V \succ 0$ and all $t \geq 0$ we have with probability at least $1 - \delta$*

$$S_t^{\top}(V + V_t)^{-1} S_t \leq 2L^2 \bigg(\log \frac{1}{\delta} + \log \frac{det(V + V_t)}{det(V)} + m \bigg).$$

Proof Define $M = (V + V_t)^{-1/2} S_t$. Now we use Proposition 8.1 and setting $\epsilon = 1/2$,

$$\mathbb{P}(||M||_2 > z) \leq 5^m \mathbb{P}(||Mw||_2 > 2z)$$

for $w \in \mathcal{S}^{m-1}$. Then we can use Theorem 1 in Abbasi-Yadkori et al. (2011), and with probability at least $1 - \delta$ we have

$$||Mw||_2^2 \leq 2L^2 \bigg(\log \frac{1}{\delta} + \log \frac{\det(V + V_t)}{\det(V)} \bigg).$$

By $\delta \to 5^{-m}\delta$, we have with probability at least $1 - 5^{-m}\delta$

$$||Mw||_2 \leq \sqrt{2} L \sqrt{\bigg(m \log(5) + \log \frac{1}{\delta} + \log \frac{\det(V + V_t)}{\det(V)} \bigg)}.$$

Then with probability at least $1 - \delta$,

$$||M||_2 \leq \sqrt{\frac{\log(5)}{2}} L \sqrt{\bigg(m + \log \frac{1}{\delta} + \log \frac{\det(V + V_t)}{\det(V)} \bigg)}.$$

■

Lemma 8.1 *For any $M = (C, A, B)$, we have that*

$$||\mathcal{B}_{T \times mT}^{v}|| = \sqrt{\sigma\Big(\sum_{k=1}^{d} \mathcal{T}_{d+k,T}^{\top} \mathcal{T}_{d+k,T} \Big)}$$

Here $\mathcal{B}_{T \times mT}^{v}$ is defined as follows: $\beta = \mathcal{H}_{d,d,T}^{\top} v = [\beta_1^{\top}, \beta_2^{\top}, \ldots, \beta_T^{\top}]^{\top}$.

$$\mathcal{B}_{T \times mT}^{v} = \begin{bmatrix} \beta_1^{\top} & 0 & 0 & \ldots \\ \beta_2^{\top} & \beta_1^{\top} & 0 & \ldots \\ \vdots & \vdots & \ddots & \vdots \\ \beta_T^{\top} & \beta_{T-1}^{\top} & \ldots & \beta_1^{\top} \end{bmatrix}$$

and $||v||_2 = 1$.

Proof For the matrix $\mathcal{B}^v$ we have

$$\mathcal{B}^v u = \begin{bmatrix} \beta_1^\top u_1 \\ \beta_1^\top u_2 + \beta_2^\top u_1 \\ \beta_1^\top u_3 + \beta_2^\top u_2 + \beta_3^\top u_1 \\ \vdots \\ \beta_1^\top u_T + \beta_2^\top u_{T-1} + \ldots + \beta_T^\top u_1 \end{bmatrix} = \begin{bmatrix} v^\top \begin{bmatrix} CA^{d+1}Bu_1 \\ CA^{d+2}Bu_1 \\ \vdots \\ CA^{2d}Bu_1 \end{bmatrix} \\ v^\top \begin{bmatrix} CA^{d+2}Bu_1 + CA^{d+1}Bu_2 \\ CA^{d+3}Bu_1 + CA^{d+2}Bu_2 \\ \vdots \\ CA^{2d+1}Bu_1 + CA^{2d}Bu_2 \end{bmatrix} \\ \vdots \\ v^\top \begin{bmatrix} CA^{T+d}Bu_1 + \ldots + CA^{d+1}Bu_T \\ CA^{T+d+2}Bu_1 + \ldots + CA^{d+2}Bu_T \\ \vdots \\ CA^{T+2d-1}Bu_1 + \ldots + CA^{2d}Bu_T \end{bmatrix} \end{bmatrix}$$

$$= \mathcal{V} \begin{bmatrix} \begin{bmatrix} CA^{d+1}Bu_1 \\ CA^{d+2}Bu_1 \\ \vdots \\ CA^{2d}Bu_1 \end{bmatrix} \\ \begin{bmatrix} CA^{d+2}Bu_1 + CA^{d+1}Bu_2 \\ CA^{d+3}Bu_1 + CA^{d+2}Bu_2 \\ \vdots \\ CA^{2d+1}Bu_1 + CA^{2d}Bu_2 \end{bmatrix} \\ \vdots \\ \begin{bmatrix} CA^{T+d}Bu_1 + \ldots + CA^{d+1}Bu_T \\ CA^{T+d+2}Bu_1 + \ldots + CA^{d+2}Bu_T \\ \vdots \\ CA^{T+2d-1}Bu_1 + \ldots + CA^{2d}Bu_T \end{bmatrix} \end{bmatrix}$$

$$= \mathcal{V} \underbrace{\begin{bmatrix} CA^{d+1}B & 0 & 0 & \ldots & 0 \\ CA^{d+2}B & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ CA^{2d}B & 0 & 0 & \ldots & 0 \\ CA^{d+2}B & CA^{d+1}B & 0 & \ldots & 0 \\ CA^{d+3}B & CA^{d+2}B & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ CA^{2d+1}B & CA^{2d}B & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ CA^{T+d-1}B & CA^{T+d}B & CA^{T+d-1}B & \ldots & CA^{d+1}B \\ CA^{T+d+2}B & CA^{T+d+1}B & CA^{T+d}B & \ldots & CA^{d+2}B \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ CA^{T+2d-1}B & CA^{T+2d-1}B & CA^{T+2d-2}B & \ldots & CA^{2d}B \end{bmatrix}}_{=S} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix}$$

It is clear that $||\mathcal{V}||_2, ||u||_2 = 1$ and for any matrix S, $||S||$ does not change if we interchange rows of S. Then we have

$$||S||_2 = \sigma\left(\begin{bmatrix} CA^{d+1}B & 0 & 0 & \dots & 0 \\ CA^{d+2}B & CA^{d+1}B & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ CA^{T+d+1}B & CA^{T+d}B & CA^{T+d-1}B & \dots & CA^{d+1}B \\ CA^{d+2}B & 0 & 0 & \dots & 0 \\ CA^{d+3}B & CA^{d+2}B & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ CA^{T+d+2}B & CA^{T+d+1}B & CA^{T+d}B & \dots & CA^{d+2}B \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ CA^{2d}B & 0 & 0 & \dots & 0 \\ CA^{2d+1}B & CA^{2d}B & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ CA^{T+2d-1}B & CA^{T+2d-1}B & CA^{T+2d-2}B & \dots & CA^{2d}B \end{bmatrix}\right)$$

$$= \sigma\left(\begin{bmatrix} \mathcal{T}_{d+1,T} \\ \mathcal{T}_{d+2,T} \\ \vdots \\ \mathcal{T}_{2d,T} \end{bmatrix}\right) = \sqrt{\sigma\Big(\sum_{k=1}^{d} \mathcal{T}_{d+k,T}^{\top}\mathcal{T}_{d+k,T}\Big)}$$

■

Proposition 8.3 (Lemma 4.1 Simchowitz et al. (2018)) *Let S be an invertible matrix and $\kappa(S)$ be its condition number. Then for a $\frac{1}{4\kappa}$–net of $\mathcal{S}^{d-1}$ and an arbitrary matrix A, we have*

$$||SA||_2 \leq 2 \sup_{v \in \mathcal{N}_{\frac{1}{4\kappa}}} \frac{||v'A||_2}{||v'S^{-1}||_2}$$

Proof For any vector $v \in \mathcal{N}_{\frac{1}{4\kappa}}$ and w be such that $||SA||_2 = \frac{||w'A||_2}{||w'S^{-1}||_2}$ we have

$$\begin{aligned} ||SA||_2 - \frac{||v'A||_2}{||v'S^{-1}||_2} &\leq \Big|\frac{||w'A||_2}{||w'S^{-1}||_2} - \frac{||v'A||_2}{||v'S^{-1}||_2}\Big| \\ &= \Big|\frac{||w'A||_2}{||w'S^{-1}||_2} - \frac{||v'A||_2}{||w'S^{-1}||_2} + \frac{||v'A||_2}{||w'S^{-1}||_2} - \frac{||v'A||_2}{||v'S^{-1}||_2}\Big| \\ &\leq ||SA||_2 \frac{\frac{1}{4\kappa}||S^{-1}||_2}{||w'S^{-1}||_2} + ||SA||_2 \Big|\frac{||v'S^{-1}||_2}{||w'S^{-1}||_2} - 1\Big| \\ &\leq \frac{||SA||_2}{2} \end{aligned}$$

■

9. Control and Systems Theory Preliminaries

9.1 Sylvester Matrix Equation

Define the discrete time Sylvester operator $S_{A,B} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$

$$\mathcal{L}_{A,B}(X) = X - AXB \tag{21}$$

Then we have the following properties for $\mathcal{L}_{A,B}(\cdot)$.

Proposition 9.1 *Let λ_i, μ_i be the eigenvalues of A, B then $\mathcal{L}_{A,B}$ is invertible if and only if for all i, j*

$$\lambda_i \mu_j \neq 1$$

Define the discrete time Lyapunov operator for a matrix A as $\mathcal{L}_{A,A'}(\cdot) = S^{-1}_{A,A'}(\cdot)$. Clearly it follows from Proposition 9.1 that whenever $\lambda_{\max}(A) < 1$ we have that the $S_{A,A'}(\cdot)$ is an invertible operator.

Now let $Q \succeq 0$ then

$$\begin{aligned} & S_{A,A'}(Q) = X \\ \implies & X = AXA' + Q \\ \implies & X = \sum_{k=0}^{\infty} A^k Q A'^k \end{aligned} \tag{22}$$

Eq. (22) follows directly by substitution and by Proposition 9.1 is unique if $\rho(A) < 1$. Further, let $Q_1 \succeq Q_2 \succeq 0$ and X_1, X_2 be the corresponding solutions to the Lyapunov operator then from Eq. (22) that

$$\begin{aligned} X_1, X_2 &\succeq 0 \\ X_1 &\succeq X_2 \end{aligned}$$

9.2 Properties of System Hankel matrix

- **Rank of system Hankel matrix**: For $M = (C, A, B) \in \mathcal{M}_n$, the system Hankel matrix, $\mathcal{H}_{0,\infty,\infty}(M)$, can be decomposed as follows:

$$\mathcal{H}_{0,\infty,\infty}(M) = \underbrace{\begin{bmatrix} C \\ CA \\ \vdots \\ CA^d \\ \vdots \end{bmatrix}}_{=\mathcal{O}} \underbrace{\begin{bmatrix} B & AB & \dots & A^dB & \dots \end{bmatrix}}_{=\mathcal{R}} \tag{23}$$

  It follows from definition that $\text{rank}(\mathcal{O}), \text{rank}(\mathcal{R}) \leq n$ and as a result $\text{rank}(\mathcal{O}\mathcal{R}) \leq n$. The system Hankel matrix rank, or $\text{rank}(\mathcal{O}\mathcal{R})$, which is also the model order(or

simply order), captures the complexity of M. If $\text{SVD}(\mathcal{H}_{0,\infty,\infty}) = U\Sigma V^\top$, then $\mathcal{O} = U\Sigma^{1/2}S, \mathcal{R} = S^{-1}\Sigma^{1/2}V^\top$. By noting that

$$CA^lS = CS(S^{-1}AS)^l, S^{-1}A^lB = (S^{-1}AS)^lS^{-1}B$$

we have obtained a way of recovering the system parameters (up to similarity transformations). Furthermore, $\mathcal{H}_{0,\infty,\infty}$ uniquely (up to similarity transformation) recovers (C, A, B).

- **Mapping Past to Future**: $\mathcal{H}_{0,\infty,\infty}$ can also be viewed as an operator that maps "past" inputs to "future" outputs. In Eq. (1) assume that $\{\eta_t, w_t\} = 0$. Then consider the following class of inputs U_t such that $U_t = 0$ for all $t \geq T$ but U_t may not be zero for $t < T$. Here T is chosen arbitrarily. Then

$$\underbrace{\begin{bmatrix} Y_T \\ Y_{T+1} \\ Y_{T+2} \\ \vdots \end{bmatrix}}_{\text{Future}} = \mathcal{H}_{0,\infty,\infty} \underbrace{\begin{bmatrix} U_{T-1} \\ U_{T-2} \\ U_{T-3} \\ \vdots \end{bmatrix}}_{\text{Past}} \tag{24}$$

9.3 Model Reduction

Given an LTI system $M = (C, A, B)$ of order n with its doubly infinite system Hankel matrix as $\mathcal{H}_{0,\infty,\infty}$. We are interested in finding the best k order lower dimensional approximation of M, *i.e.*, for every $k < n$ we would like to find M_k of model order k such that $||M - M_k||_\infty$ is minimized. Systems theory gives us a class of model approximations, known as balanced truncated approximations, that provide strong theoretical guarantees (See Glover (1984) and Section 21.6 in Zhou et al. (1996)). We summarize some of the basics of model reduction below. Assume that M has distinct Hankel singular values.

Recall that a model $M = (C, A, B)$ is equivalent to $\tilde{M} = (CS, S^{-1}AS, S^{-1}B)$ with respect to its transfer function. Define

$$Q = A^\top QA + C^\top C$$
$$P = APA^\top + BB^\top$$

For two positive definite matrices P, Q it is a known fact that there exist a transformation S such that $S^\top QS = S^{-1}PS^{-1^\top} = \Sigma$ where Σ is diagonal and the diagonal elements are decreasing. Further, σ_i is the i^{th} singular value of $\mathcal{H}_{0,\infty,\infty}$. Then let $\tilde{A} = S^{-1}AS, \tilde{C} = CS, \tilde{B} = S^{-1}B$. Clearly $\widetilde{M} = (\tilde{A}, \tilde{B}, \tilde{C})$ is equivalent to M and we have

$$\Sigma = \tilde{A}^\top \Sigma \tilde{A} + \tilde{C}^\top \tilde{C}$$
$$\Sigma = \tilde{A}\Sigma\tilde{A}^\top + \tilde{B}\tilde{B}^\top \tag{25}$$

Here $\tilde{C}, \tilde{A}, \tilde{B}$ is a balanced realization of M.

Proposition 9.2 *Let $\mathcal{H}_{0,\infty,\infty} = U\Sigma V^\top$. Here $\Sigma \succeq 0 \in \mathbb{R}^{n\times n}$. Then*

$$\begin{aligned}\tilde{C} &= [U\Sigma^{1/2}]_{1:p,:}\\ \tilde{A} &= \Sigma^{-1/2}U^\top [U\Sigma^{1/2}]_{p+1:,:}\\ \tilde{B} &= [\Sigma^{1/2}V^\top]_{:,1:m}\end{aligned}$$

The triple $(\tilde{C}, \tilde{A}, \tilde{B})$ is a balanced realization of M. For any matrix L, $L_{:,m:n}$ (or $L_{m:n,:}$) denotes the submatrix with only columns (or rows) m through n.

Proof Let the SVD of $\mathcal{H}_{0,\infty,\infty} = U\Sigma V^\top$. Then M can constructed as follows: $U\Sigma^{1/2}, \Sigma^{1/2}V^\top$ are of the form

$$U\Sigma^{1/2} = \begin{bmatrix} CS \\ CAS \\ CA^2S \\ \vdots \end{bmatrix}, \Sigma^{1/2}V^\top = \begin{bmatrix} S^{-1}B & S^{-1}AB & S^{-1}A^2B \dots \end{bmatrix}$$

where S is the transformation which gives us Eq. (25). This follows because

$$\begin{aligned}\Sigma^{1/2}U^\top U\Sigma^{1/2} &= \sum_{k=0}^{\infty} S^\top A^{k\top} C^\top C A^k S\\ &= \sum_{k=0}^{\infty} S^\top A^{k\top} S^{-1\top} S^\top C^\top C S S^{-1} A^k S\\ &= \sum_{k=0}^{\infty} \tilde{A}^{k\top}\tilde{C}^\top\tilde{C}\tilde{A}^k = \tilde{A}^\top\Sigma\tilde{A} + \tilde{C}^\top\tilde{C} = \Sigma\end{aligned}$$

Then $\tilde{C} = U\Sigma^{1/2}_{1:p,:}$ and

$$\begin{aligned}U\Sigma^{1/2}\tilde{A} &= [U\Sigma^{1/2}]_{p+1:,:}\\ \tilde{A} &= \Sigma^{-1/2}U^\top[U\Sigma^{1/2}]_{p+1:,:}\end{aligned}$$

We do a similar computation for B. ■

It should be noted that a balanced realization $\tilde{C}, \tilde{A}, \tilde{B}$ is unique except when there are some Hankel singular values that are equal. To see this, assume that we have

$$\sigma_1 > \ldots > \sigma_{r-1} > \sigma_r = \sigma_{r+1} = \ldots = \sigma_s > \sigma_{s+1} > \ldots \sigma_n$$

where $s - r > 0$. For any unitary matrix $Q \in \mathbb{R}^{(s-r+1)\times(s-r+1)}$, define Q_0

$$Q_0 = \begin{bmatrix} I_{(r-1)\times(r-1)} & 0 & 0 \\ 0 & Q & 0 \\ 0 & 0 & I_{(n-s)\times(n-s)} \end{bmatrix} \tag{26}$$

Then every triple $(\tilde{C}Q_0, Q_0^\top \tilde{A} Q_0, Q_0^\top \tilde{B})$ satisfies Eq. (25) and is a balanced realization. Let $M_k = (\tilde{C}_k, \tilde{A}_{kk}, \tilde{B}_k)$ where

$$\tilde{A} = \begin{bmatrix} \tilde{A}_{kk} & \tilde{A}_{0k} \\ \tilde{A}_{k0} & \tilde{A}_{00} \end{bmatrix}, \tilde{B} = \begin{bmatrix} \tilde{B}_k \\ \tilde{B}_0 \end{bmatrix}, \tilde{C} = \begin{bmatrix} \tilde{C}_k & \tilde{C}_0 \end{bmatrix} \tag{27}$$

Here $\tilde{A}_{kk}$ is the $k \times k$ submatrix and corresponding partitions of $\tilde{B}, \tilde{C}$. The realization $M_k = (\tilde{C}_k, \tilde{A}_{kk}, \tilde{B}_k)$ is the k–order balanced truncated model. Clearly $M \equiv M_n$ which gives us $\tilde{C} = \tilde{C}_{nn}, \tilde{A} = \tilde{A}_{nn}, \tilde{B} = \tilde{B}_{nn}$, *i.e.*, the balanced version of the true model. We will show that for the balanced truncation model we only need to care about the top k singular vectors and not the entire model.

Proposition 9.3 *For the k order balanced truncated model M_k, we only need top k singular values and singular vectors of $\mathcal{H}_{0,\infty,\infty}$.*

Proof From the preceding discussion in Proposition 9.2 and Eq. (27) it is clear that the first $p \times k$ block submatrix of $U\Sigma^{1/2}$ (corresponding to the top k singular vectors) gives us $\tilde{C}_k$. Since

$$\tilde{A} = \Sigma^{-1/2} U^\top [U\Sigma^{1/2}]_{p+1:,:}$$

we observe that $\tilde{A}_{kk}$ depend only on the top k singular vectors U_k and corresponding singular values. This can be seen as follows: $[U\Sigma^{1/2}]_{p+1:,:}$ denotes the submatrix of $U\Sigma^{1/2}$ with top p rows removed. Now in $U\Sigma^{1/2}$ each column of U is scaled by the corresponding singular value. Then the $\tilde{A}_{kk}$ submatrix depends only on top k rows of $\Sigma^{-1/2}U^\top$ and the top k columns of $[U\Sigma^{1/2}]_{p+1:,:}$ which correspond to the top k singular vectors. ■

10. Isometry of Input Matrix: Proof of Lemma 5.1

Theorem 11 *Define*

$$U \coloneqq \begin{bmatrix} U_d & U_{d+1} & \dots & U_{T+d-1} \\ U_{d-1} & U_d & \dots & U_{T+d-2} \\ \vdots & \vdots & \ddots & \vdots \\ U_1 & U_2 & \dots & U_T \end{bmatrix}$$

where each $U_i \sim \mathsf{subg}(1)$ and isotropic. Then there exists an absolute constant c such that U satisfies:

$$(1/2)T \le \sigma_{\min}(UU^\top) \le \sigma_{\max}(UU^\top) \le (3/2)T$$

whenever $T \ge cm^2 d(\log^2(d)\log^2(m^2/\delta) + \log^3(2d))$ with probability at least $1-\delta$.

Proof

Define

$$A_{md\times md} \coloneqq \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ I & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \dots & I & 0 & 0 \\ 0 & \dots & 0 & I & 0 \end{bmatrix}, B_{md\times m} \coloneqq \begin{bmatrix} I \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \widehat{U}_k \coloneqq U_{d+k}$$

Since

$$U = \begin{bmatrix} U_d & U_{d+1} & \dots & U_{T+d-1} \\ U_{d-1} & U_d & \dots & U_{T+d-2} \\ \vdots & \vdots & \dots & \vdots \\ U_1 & U_2 & \dots & U_T \end{bmatrix}$$

we can reformulate it so that each column is the output of an LTI system in the following sense:

$$x_{k+1} = Ax_k + B\widehat{U}(k+1) \tag{28}$$

where $UU^\top = \sum_{k=0}^{T-1} x_k x_k^\top$ and $x_0 = \begin{bmatrix} U_d \\ U_{d-1} \\ \vdots \\ U_1 \end{bmatrix}$. From Theorem 8.1 we have that

$$\frac{3}{4}TI \preceq \sum_{k=0}^{T-1} \widehat{U}_k \widehat{U}_k^\top \preceq \frac{5}{4}TI$$

with probability at least $1-\delta$ whenever $T \geq c\Big(m + \log\frac{2}{\delta}\Big)$. Define $V_t = \sum_{l=0}^{t-1} x_k x_k^\top$ then,

$$V_T = AV_{T-1}A^\top + B\left(\sum_{k=0}^{T-1} \widehat{U}_k \widehat{U}_k^\top\right)B^\top + \sum_{k=0}^{T-2}\Big(Ax_k\widehat{U}_{k+1}^\top B^\top + B\widehat{U}_{k+1}x_k^\top A^\top\Big) \tag{29}$$

It can be easily checked that $x_k = \begin{bmatrix} U_{d+k} \\ U_{d+k-1} \\ \vdots \\ U_{k+1} \end{bmatrix}$ and consequently

$$\sum_{k=0}^{T-2} Ax_k\widehat{U}_{k+1}^\top B^\top = \sum_{k=0}^{T-2} \begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ U_{d+k}U_{d+k+1}^\top & 0 & \dots & 0 & 0 \\ U_{d+k-1}U_{d+k+1}^\top & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ U_{k+2}U_{d+k+1}^\top & 0 & \dots & 0 & 0 \end{bmatrix}.$$

Define $L_j \coloneqq \sum_{k=0}^{T-2} U_{d+k-j+1}U_{d+k+1}^\top$ and L_j is a $m \times m$ block matrix. Then

$$T_d = \sum_{l=0}^{d-1} A^l\left(\sum_{k=0}^{T-2} Ax_k\widehat{U}_{k+1}^\top B^\top\right)A^{l\top} = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 & 0 \\ L_1 & 0 & \dots & 0 & 0 & 0 \\ L_2 & L_1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ L_{d-1} & 0 & \dots & 0 & L_1 & 0 \end{bmatrix}.$$

Use Lemma 10.1 to show that

$$\|T_d\| \leq cm\sqrt{Td}\log(d)\log(m^2/\delta) \tag{30}$$

with probability at least $1-\delta$. Then

$$V_T = \sum_{l=0}^{d-1} A^l B\left(\sum_{k=0}^{T-1}\widehat{U}_k\widehat{U}_k^\top\right)B^\top A^{l\top} + T_d - \sum_{l=0}^{d-1} A^l x_{T-1}x_{T-1}^\top A^{l\top}.$$

From Theorem 8.1 we have with probability atleast $1-\delta$ that

$$(3/4)TI \preceq \sum_{l=0}^{d-1} A^l B\left(\sum_{k=0}^{T-1}\widehat{U}_k\widehat{U}_k^\top\right)B^\top A^{l\top} \preceq (5/4)TI \tag{31}$$

whenever $T \geq c\Big(m + \log\frac{2}{\delta}\Big)$. Observe that

$$\left\|\sum_{l=1}^{d} A^l x_{T-1}x_{T-1}^\top A^{l\top}\right\| = \sigma_1^2([Ax_{T-1}, A^2x_{T-1}, \ldots, A^dx_{T-1}])$$

The matrix $[Ax_{T-1}, A^2x_{T-1}, \ldots, A^dx_{T-1}]$ is the lower triangular submatrix of a random Toeplitz matrix with i.i.d $\mathsf{subg}(1)$ entries as in Theorem 8.2. Then using Theorem 8.2 and Proposition 8.2 we get that with probability at least $1-\delta$ we have

$$\left\|[Ax_{T-1}, A^2x_{T-1}, \ldots, A^dx_{T-1}]\right\| \leq cm(\sqrt{d\log(2d)}\log(2d) + \sqrt{d\log(1/\delta)}). \tag{32}$$

Then $\left\|\sum_{l=1}^{d} A^l x_{T-1}x_{T-1}^\top A^{l\top}\right\| \leq cm^2d(\log^3(2d) + \log(1/\delta) + \log(2d)\sqrt{\log(2d)\log(1/\delta)})$ with probability at least $1-\delta$. By ensuring that Eqs. (30), (31) and (32) hold simultaneously we can ensure that $cm\sqrt{Td}\log(d)\log(m^2/\delta) \leq T/8$ and $cm^2d(\log^3(2d) + \log(1/\delta) + \log(2d)\sqrt{\log(2d)\log(1/\delta)}) \leq T/8$ for large enough T and absolute constant c. ■

Lemma 10.1 *Let $\{U_j \in \mathbb{R}^{m\times 1}\}_{j=1}^{T+d}$ be independent $\mathsf{subg}(1)$ random vectors. Define $L_j \coloneqq \sum_{k=0}^{T-2} U_{d+k-j+1}U_{d+k+1}^\top$ for all $j \geq 1$ and*

$$T_d \coloneqq \begin{bmatrix} 0 & 0 & \ldots & 0 & 0 & 0 \\ L_1 & 0 & \ldots & 0 & 0 & 0 \\ L_2 & L_1 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ L_{d-1} & 0 & \ldots & 0 & L_1 & 0 \end{bmatrix}.$$

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

$$\|T_d\| \leq cm\sqrt{Td}\log(d)\log(m/\delta).$$

Proof Since L_js are block matrices, the techniques in Meckes et al. (2007) cannot be directly applied. However, by noting that E can be broken into a sum of m matrices where the norm of each matrix can be bounded by a Toeplitz matrix we can use the result from Meckes et al. (2007). For instance if $m = 2$ and $\{u_i\}_{i=1}^{\infty}$ are independent $\mathsf{subg}(1)$ random variables then we have

$$T_d = \begin{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \cdots \\ \begin{bmatrix} u_1 & u_2 \\ u_3 & u_4 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \cdots \\ \begin{bmatrix} u_5 & u_6 \\ u_7 & u_8 \end{bmatrix} & \begin{bmatrix} u_1 & u_2 \\ u_3 & u_4 \end{bmatrix} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}.$$

Now,

$$T_d = \underbrace{\begin{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \cdots \\ \begin{bmatrix} u_1 & 0 \\ u_3 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \cdots \\ \begin{bmatrix} u_5 & 0 \\ u_7 & 0 \end{bmatrix} & \begin{bmatrix} u_1 & 0 \\ u_3 & 0 \end{bmatrix} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}}_{=M_1} + \underbrace{\begin{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \cdots \\ \begin{bmatrix} 0 & u_2 \\ 0 & u_4 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \cdots \\ \begin{bmatrix} 0 & u_6 \\ 0 & u_8 \end{bmatrix} & \begin{bmatrix} 0 & u_2 \\ 0 & u_4 \end{bmatrix} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}}_{=M_2},$$

then $||T_d|| \leq \sup_{1 \leq i \leq 2} ||M_i||$. Furthermore for each M_i we have

$$M_1 = \underbrace{\begin{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \cdots \\ \begin{bmatrix} u_1 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \cdots \\ \begin{bmatrix} u_5 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} u_1 & 0 \\ 0 & 0 \end{bmatrix} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}}_{=M_{11}} + \underbrace{\begin{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \cdots \\ \begin{bmatrix} 0 & 0 \\ u_3 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \cdots \\ \begin{bmatrix} 0 & 0 \\ u_7 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ u_3 & 0 \end{bmatrix} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}}_{=M_{12}},$$

and $||M_1|| \leq ||M_{11}|| + ||M_{12}||$. The key idea is to show that M_{i1} are Toeplitz matrices (after removing the zeros in the blocks) and we can use the standard techniques described in proof of Theorem 1 in Meckes et al. (2007). Then we will show that each $||M_{ij}|| \leq C$ with high probability and $||T_d|| \leq mC$.

For brevity, we will assume for now that U_i are scalars and at the end we will scale by m. By standard techniques described in proof of Theorem 1 in Meckes et al. (2007), we have that the finite Toeplitz matrix $T_d + T_d^{\top}$ is $d \times d$ submatrix of the infinite Laurent matrix

$$M = [L_{|j-k|} \mathbf{1}_{|j-k|<d-1}]_{j,k \in \mathbb{Z}}.$$

Consider M as an operator on $\ell^2(\mathbb{Z})$ in the canonical way, and let $\psi : \ell^2(\mathbb{Z}) \to L^2[0,1]$ denote the usual linear trigonometric isometry $\psi(e_j)(x) = e^{2\pi ijx}$. Then $\psi M_d \psi^{-1} : L^2 \to L^2$ is the operator correpsonding to

$$f(x) = \sum_{j=-(d-1)}^{d-1} L_{|j|} e^{2\pi ijx} = L_0 + 2\sum_{j=1}^{d-1} \cos(2\pi jx) L_j$$

Therefore,

$$\left\| T_d + T_d^\top \right\| \leq \|M\| = \|f\|_\infty = \sup_{0 \leq x \leq 1} |Y_x|$$

where $Y_x = 2\sum_{j=1}^{d-1} \cos(2\pi jx) L_j$. Furthermore note that Y_x has the following form

$$Y_x = U^\top \underbrace{\begin{bmatrix} 0 & c_1^x & c_2^x & \dots & c_{d-1}^x & 0 & \dots & 0 \\ 0 & 0 & c_1^x & \dots & c_{d-1}^x & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \ddots & \dots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 0 & c_1^x & \dots & c_{d-1}^x \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}}_{=C_x} U. \tag{33}$$

Here $U = \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_{T+d} \end{bmatrix}$ and $c_j^x = 2\cos(2\pi jx)$. For any x and assuming $U_j \sim \mathsf{subg}(1)$, we have from Theorem 8.3

$$\mathbb{P}\Big(\Big| Y_x / \sqrt{Td} \Big| \leq t \Big) \leq 2\exp\big\{ -c(t \wedge t^2) \big\} \tag{34}$$

The tail behavior of $Y_x/\sqrt{Td}$ is not strictly subgaussian and we need to use Theorem 8.4. The function ψ can be found as Eq. 1 of van de Geer and Lederer (2013) (equivalent upto universal constants) with $L = 2$ and its inverse being

$$\psi^{-1}(t) = \sqrt{\log(1+t)} + \log(1+t).$$

We have that

$$\left\| \sup_t |Y_t| \right\|_\psi \leq \|Y_0\|_\psi + K\sqrt{Td} \int_0^1 \psi^{-1}(N(\epsilon/2, d)) d\epsilon,$$

where $d(s,t) = \left\| (Y_s - Y_t)/\sqrt{Td} \right\|_\psi$ and $N(\epsilon, d)$ is the minimal number of balls of radius ϵ needed to cover $[0,1]$ where $d(\cdot,\cdot)$ is the pseudometric. Since Y_s has distribution as in Eq. (34), it follows that $d(s,t) \leq c|s-t|$ for some absolute constant c. Then

$$\int_0^1 \psi^{-1}(N(\epsilon/2, d)) d\epsilon \leq c$$

for some universal constant $c > 0$. This ensures that $\|\sup_t |Y_t|\|_\psi \leq c\sqrt{Td}$. Since $\mathbb{E}[X] \leq \|X\|_\psi$ we have that $\mathbb{E}[\sup_{0\leq x\leq 1}|Y_x|] \leq \sqrt{Td}$. This implies $\mathbb{E}[\left\|T_d + T_d^\top\right\|] \leq \sqrt{Td}$, and using Proposition 8.2 we have $\mathbb{E}[\|T_d\|] \leq c\sqrt{Td}\log(d)$. Furthermore, we can make a stronger statement because $\|\sup_t |Y_t|\|_\psi \leq c\sqrt{Td}$ which implies that

$$\|T_d\| \leq c\sqrt{Td}\log(d)\log(1/\delta)$$

with probability at least $1-\delta$. Then recalling that in the general case that L_js of T_d were $m \times m$ block matrices we scale by m and get with probability at least $1-\delta$

$$\|T_d\| \leq cm\sqrt{Td}\log(d)\log(m^2/\delta)$$

where the union is over all m^2 elements being less that $c\sqrt{Td}\log(d)\log(m^2/\delta)$. Note that c hides the universal constant K from Theorem 8.4. ■

11. Error Analysis for Theorem 5.1

For this section we assume that $U_t \sim \mathsf{subg}(L^2)$.

11.1 Proof of Theorem 5.1

Recall Eq. (8) and (9), *i.e.*,

$$\begin{aligned}\tilde{Y}_{l,d}^+ &= \mathcal{H}_{0,d,d}\tilde{U}_{l-1,d}^- + \mathcal{T}_{0,d}\tilde{U}_{l,d}^+ + \mathcal{H}_{d,d,l-d-1}\tilde{U}_{l-d-1,l-d-1}^- \\ &\quad + \mathcal{O}_{0,d,d}\tilde{\eta}_{l-1,d}^- + \mathcal{T}\mathcal{O}_{0,d}\tilde{\eta}_{l,d}^+ + \mathcal{O}_{d,d,l-d-1}\tilde{\eta}_{l-d-1,l-d-1}^- + \tilde{w}_{l,d}^+ \end{aligned} \tag{35}$$

Assume for now that we have $T+2d$ data points instead of T. It is clear that

$$\hat{\mathcal{H}}_{0,d,d} = \arg\min_{\mathcal{H}} \sum_{l=0}^{T-1} ||\tilde{Y}_{l+d+1,d}^+ - \mathcal{H}\tilde{U}_{l+d,d}^-||_2^2 = \left(\sum_{l=0}^{T-1} \tilde{Y}_{l+d+1,d}^+ \left(\tilde{U}_{l+d,d}^-\right)^\top\right) V_T^+$$

where

$$V_T = \sum_{l=0}^{T-1} \tilde{U}_{l+d,d}^- \tilde{U}_{l+d,d}^{-\prime}, \tag{36}$$

or

$$V_T = UU'$$

where

$$U \coloneqq \begin{bmatrix} U_d & U_{d+1} & \dots & U_{T+d-1} \\ U_{d-1} & U_d & \dots & U_{T+d-2} \\ \vdots & \vdots & \ddots & \vdots \\ U_1 & U_2 & \dots & U_T \end{bmatrix}.$$

It is show in Theorem 11 that V_T is invertible with probability at least $1-\delta$. So in our analysis we can write this as

$$\left(\sum_{l=0}^{T-1}\tilde{U}^{-}_{l+d,d}\tilde{U}^{-\top}_{l+d,d}\right)^{+}=\left(\sum_{l=0}^{T-1}\tilde{U}^{-}_{l+d,d}\tilde{U}^{-\top}_{l+d,d}\right)^{-1}$$

From this one can conclude that

$$\begin{aligned}\left|\left|\hat{\mathcal{H}}-\mathcal{H}_{0,d,d}\right|\right|_2=\Big|\Big|\Big(\sum_{l=0}^{T-1}\tilde{U}^{-}_{l+d,d}\tilde{U}^{-\top}_{l+d,d}\Big)^{-1}\Big(\sum_{l=0}^{T-1}\tilde{U}^{-}_{l+d,d}\tilde{U}^{+\top}_{l+d+1,d}\mathcal{T}^{\top}_{0,d}\\ +\tilde{U}^{-}_{l+d,d}\tilde{U}^{-\top}_{l,l}\mathcal{H}^{\top}_{d,d,l}+\tilde{U}^{-}_{l+d,d}\tilde{\eta}^{-\top}_{l+d,d}\mathcal{O}^{\top}_{0,d,d}\\ +\tilde{U}^{-}_{l+d,d}\tilde{\eta}^{+\top}_{l+d+1,d}\mathcal{T}\mathcal{O}^{\top}_{0,d}+\tilde{U}^{-}_{l+d,d}\tilde{\eta}^{-\top}_{l,l}\mathcal{O}^{\top}_{d,d,l}+\tilde{U}^{-}_{l+d,d}\tilde{w}^{+\top}_{l+d+1,d}\Big)\Big|\Big|_2\end{aligned}\tag{37}$$

Here as we can observe $\tilde{U}^{-\top}_{l,l},\tilde{\eta}^{-\top}_{l,l}$ grow with T in dimension. Based on this we divide our error terms in two parts:

$$E_1=\Big(\sum_{l=0}^{T-1}\tilde{U}^{-}_{l+d,d}\tilde{U}^{-\top}_{l+d,d}\Big)^{-1}\left(\tilde{U}^{-}_{l+d,d}\tilde{U}^{-\top}_{l,l}\mathcal{H}^{\top}_{d,d,l}+\tilde{U}^{-}_{l+d,d}\tilde{\eta}^{-\top}_{l,l}\mathcal{O}^{\top}_{d,d,l}\right)\tag{38}$$

and

$$\begin{aligned}E_2=\Big(\sum_{l=0}^{T-1}\tilde{U}^{-}_{l+d,d}\tilde{U}^{-\top}_{l+d,d}\Big)^{-1}\Big(\tilde{U}^{-}_{l+d,d}\tilde{\eta}^{+\top}_{l+d+1,d}\mathcal{T}\mathcal{O}^{\top}_{0,d}+\tilde{U}^{-}_{l+d,d}\tilde{U}^{+\top}_{l+d+1,d}\mathcal{T}^{\top}_{0,d}+\\ \tilde{U}^{-}_{l+d,d}\tilde{\eta}^{+\top}_{l+d+1,d}\mathcal{T}\mathcal{O}^{\top}_{0,d}+\tilde{U}^{-}_{l+d,d}\tilde{w}^{+\top}_{l+d+1,d}\Big)\end{aligned}\tag{39}$$

Then the proof of Theorem 5.1 will reduce to Propositions 11.1–11.3. We first analyze

$$\left|\left|V_T^{-1/2}\Big(\sum_{l=0}^{T-1}\tilde{U}^{-}_{l+d,d}\tilde{U}^{-\top}_{l,l}\mathcal{H}^{\top}_{d,d,l}\Big)\right|\right|_2$$

The analysis of $||V_T^{-1/2}(\sum_{l=0}^{T-1}\tilde{U}^{-}_{l+d,d}\tilde{\eta}^{-\top}_{l,l}O^{\top}_{d,d,l})||$ will be almost identical and will only differ in constants.

Proposition 11.1 *For $0<\delta<1$, we have with probability at least $1-2\delta$*

$$\left|\left|V_T^{-1/2}\Big(\sum_{l=0}^{T-1}\tilde{U}^{-}_{l+d,d}\tilde{U}^{-\top}_{l,l}\mathcal{H}^{\top}_{d,d,l}\Big)\right|\right|_2\le 4\sigma\sqrt{\log\frac{1}{\delta}+pd+m}$$

where $\sigma=\sqrt{\sigma(\sum_{k=1}^{d}\mathcal{T}^{\top}_{d+k,T}\mathcal{T}_{d+k,T})}$.

Proof We proved that $\frac{TI}{2} \preceq V_T \preceq \frac{3TI}{2}$ with high probability, then

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\Big|V_T^{-1/2}\Big(\sum_{l=0}^{T-1}\tilde{U}_{l+d,d}^{-}\tilde{U}_{l,l}^{-\prime}\mathcal{H}'_{d,d,l}\Big)\Big|\Big|_2 \geq a, \frac{TI}{2} \preceq V_T \preceq \frac{3TI}{2}\Big)\\
&\leq \mathbb{P}\Big(\Big|\Big|\sqrt{\frac{2}{T}}\Big(\sum_{l=0}^{T-1}\tilde{U}_{l+d,d}^{-}\tilde{U}_{l,l}^{-\prime}\mathcal{H}'_{d,d,l}\Big)\Big|\Big|_2 \geq a, \frac{TI}{2} \preceq V_T \preceq \frac{3TI}{2}\Big)\\
&\leq \mathbb{P}\Big(2\sup_{v\in\mathcal{N}_{\frac{1}{2}}}\Big|\Big|\sqrt{\frac{2}{T}}\Big(\sum_{l=0}^{T-1}\tilde{U}_{l+d,d}^{-}\tilde{U}_{l,l}^{-\prime}\mathcal{H}'_{d,d,l}v\Big)\Big|\Big|_2 \geq a\Big) + \mathbb{P}\Big(\frac{TI}{2} \preceq V_T \preceq \frac{3TI}{2}\Big) - 1\\
&\leq 5^{pd}\mathbb{P}\Big(2\Big|\Big|\sqrt{\frac{2}{T}}\Big(\sum_{l=0}^{T-1}\tilde{U}_{l+d,d}^{-}\tilde{U}_{l,l}^{-\prime}\mathcal{H}'_{d,d,l}v\Big)\Big|\Big|_2 \geq a\Big) - \delta. \qquad (40)
\end{aligned}$$

Define the following $\eta_{l,d} = \tilde{U}_{l,l}^{-\top}\mathcal{H}_{d,d,l}^{\top}v, X_{l,d} = \sqrt{\frac{2}{T}}\tilde{U}_{l+d,d}^{-}$. Observe that $\eta_{l,d}, \eta_{l+1,d}$ have contributions from U_{l-1}, U_{l-2} etc. and do not immediately satisfy the conditions of Theorem 2.2. Instead we will use the fact that $X_{i,d}$ is independent of U_j for all $j \leq i$.

$$\begin{aligned}
\Big|\Big|V_T^{-1/2}\Big(\sum_{l=0}^{T-1}\tilde{U}_{l+d,d}^{-}\tilde{U}_{l,l}^{-\prime}\mathcal{H}'_{d,d,l}\Big)\Big|\Big|_2 &\leq 2\sup_{v\in\mathcal{N}_{\frac{1}{2}}}||\sqrt{\frac{2}{T}}\sum_{l=0}^{T-1}\tilde{U}_{l+d,d}^{-}\tilde{U}_{l,l}^{-\prime}\mathcal{H}'_{d,d,l}v||\\
&\leq 2\sup_{v\in\mathcal{N}_{\frac{1}{2}}}||\sum_{l=0}^{T-1}X_{l,d}\eta_{l,d}||.
\end{aligned}$$

Define $\mathcal{H}_{d,d,l}^{\top}v = [\beta_1^{\top}, \beta_2^{\top}, \ldots, \beta_l^{\top}]^{\top}$. β_i are $m \times 1$ vectors when LTI system is MIMO. Then $\eta_{l,d} = \sum_{k=0}^{l-1} U_{l-k}^{\top}\beta_{k+1}$. Let $\alpha_l = X_{l,d}$. Then consider the matrix

$$\mathcal{B}_{T\times mT} = \begin{bmatrix} \beta_1^{\top} & 0 & 0 & \ldots \\ \beta_2^{\top} & \beta_1^{\top} & 0 & \ldots \\ \vdots & \vdots & \ddots & \vdots \\ \beta_T^{\top} & \beta_{T-1}^{\top} & \ldots & \beta_1^{\top} \end{bmatrix}.$$

Observe that the matrix $||\mathcal{B}_{T\times mT}||_2 = \sqrt{\sigma(\sum_{k=1}^{d}\mathcal{T}_{d+k,T}^{\top}\mathcal{T}_{d+k,T})} \leq \sqrt{d}||\mathcal{T}_{d,\infty}||_2 < \infty$ which follows from Lemma 8.1. Then

$$\begin{aligned}\sum_{l=0}^{T-1} X_{l,d}\eta_{l,d} &= [\alpha_1,\dots,\alpha_T]\mathcal{B}\begin{bmatrix}U_1\\U_2\\\vdots\\U_T\end{bmatrix}\\ &= [\sum_{k=1}^{T}\alpha_k\beta_k^{\top}, \sum_{k=2}^{T}\alpha_k\beta_{k-1}^{\top},\dots,\alpha_T\beta_1^{\top}]\begin{bmatrix}U_1\\U_2\\\vdots\\U_T\end{bmatrix}\\ &= \sum_{j=1}^{T}\Big(\sum_{k=j}^{T}\alpha_k\beta_k^{\top}U_j\Big).\end{aligned}$$

Here $\alpha_i = X_{i,d}$ and recall that $X_{i,d}$ is independent of U_j for all $i \geq j$. Let $\gamma' = \alpha'\mathcal{B}$. Define $\mathcal{G}_{T+d-k} = \tilde{\sigma}(\{U_{k+1}, U_{k+2},\dots,U_{T+d}\})$ where $\tilde{\sigma}(A)$ is the sigma algebra containing the set A with $\mathcal{G}_0 = \phi$. Then $\mathcal{G}_{k-1} \subset \mathcal{G}_k$. Furthermore, since γ_{j-1}, U_j are $\mathcal{G}_{T+d+1-j}$ measurable and U_j is conditionally (on $\mathcal{G}_{T+d-j}$) subGaussian, we can use Theorem 2.2 on $\gamma'U = \alpha'\mathcal{B}U$ (where $\gamma_j = X_{T+d-j}, U_j = \eta_{T+d-j+1}$ in the notation of Theorem 2.2). Then with probability at least $1-\delta$ we have

$$\Big|\Big|\Big(\alpha'\mathcal{B}\mathcal{B}'\alpha + V\Big)^{-1/2}\gamma'U\Big|\Big| \leq L\sqrt{\Big(\log\frac{1}{\delta} + \log\frac{\det(\alpha'\mathcal{B}\mathcal{B}'\alpha+V)}{\det(V)}\Big)}. \tag{41}$$

For any fixed $V > 0$. With probability at least $1-\delta$, we know from Theorem 11 that $\alpha'\alpha \preceq \frac{3I}{2} \implies \alpha'\mathcal{B}\mathcal{B}'\alpha \preceq \frac{3\sigma_1^2(\mathcal{B})I}{2}$. By combining this event and the event in Eq. (41) and setting $V = \frac{3\sigma_1^2(\mathcal{B})I}{2}$, we get with probability at least $1-2\delta$ that

$$||\alpha'\mathcal{B}U||_2 = ||\gamma'U||_2 \leq \sqrt{3}\sigma_1(\mathcal{B})L\sqrt{\Big(\log\frac{1}{\delta} + pd\log 3 + m\Big)}. \tag{42}$$

Replacing $\delta \to 5^{-pd}\frac{\delta}{2}$, we get from Eq. (40)

$$\Big|\Big|V_T^{-1/2}\Big(\sum_{l=0}^{T-1}\tilde{U}_{l+d,d}^{-}\tilde{U}_{l,l}^{-\prime}\mathcal{H}'_{d,d,l}\Big)\Big|\Big|_2 \leq \sqrt{6}\log(5)L\sigma_1(\mathcal{B})\sqrt{\log\frac{1}{\delta}+pd+m}$$

with probability at least $1-\delta$. Since $L = 1$ we get our desired result. ■

Then similar to Proposition 11.1, we analyze $\Big|\Big|V_T^{-1/2}\Big(\sum_{l=0}^{T-1}\tilde{U}_{l+d,d}^{-}\tilde{U}_{l+d+1,d}^{+\top}\mathcal{T}_{0,d}^{\top}\Big)\Big|\Big|_2$

Proposition 11.2 *For $0 < \delta < 1$ and large enough T, we have with probability at least $1-\delta$*

$$\Big|\Big|V_T^{-1/2}\Big(\sum_{l=0}^{T-1}\tilde{U}_{l+d,d}^{-}\tilde{U}_{l+d+1,d}^{+\top}\mathcal{T}_{0,d}^{\top}\Big)\Big|\Big|_2 \leq 4\sigma\sqrt{\log\frac{1}{\delta}+pd+m}$$

where

$$\sigma \leq \sup_{||v||_2=1} \Big|\Big| \begin{bmatrix} v^\top CA^dB & v^\top CA^{d-1}B & v^\top CA^{d-2}B & \dots & v^\top CB & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & v^\top CA^dB & v^\top CA^{d-1}B & \dots & \dots & v^\top CB \end{bmatrix} \Big|\Big|_2 \leq \sum_{j=0}^{d} ||CA^jB||_2 \leq \beta\sqrt{d}.$$

Proof

Note $\Big|\Big|V_T^{-1/2}\Big(\sum_{l=0}^{T-1} \tilde{U}^-_{l+d,d}\tilde{U}^{+\top}_{l+d+1,d}\mathcal{T}_{0,d}^\top\Big)\Big|\Big|_2 \leq \Big|\Big|\sqrt{\frac{2}{T}}\Big(\sum_{l=0}^{T-1} \tilde{U}^-_{l+d,d}\tilde{U}^{+\top}_{l+d+1,d}\mathcal{T}_{0,d}^\top\Big)\Big|\Big|_2$ with probability at least $1-\delta$ for large enough T. Here $\mathcal{T}_{0,d}^\top$ is $md \times pd$ matrix. Then define $X_l = \sqrt{\frac{2}{T}}\tilde{U}^-_{l+d,d}$ and the vector $M_l \in \mathbb{R}^{pd}$ as $M_l^\top = \tilde{U}^{+\top}_{l+d+1,d}\mathcal{T}_{0,d}^\top$. Then

$$\mathbb{P}(||\sum_{l=0}^{T-1} X_l M_l^\top||_2 \geq t) \underbrace{\leq}_{\frac{1}{2}-\text{net}} 5^{pd}\mathbb{P}(||\sum_{l=0}^{T-1} X_l M_l^\top v||_2 \geq t/2)$$

where $M_l^\top v$ is a real value. Let $\beta \coloneqq \mathcal{T}_{0,d}^\top v$, then $M_l^\top v = \tilde{U}^{+\top}_{l+d+1,d}\beta$. This allows us to write $X_l M_l^\top v$ in a form that will enable us to apply Theorem 2.2.

$$\sum_{l=0}^{T-1} X_l M_l^\top v = \underbrace{[X_0, X_1, \dots, X_{T-1}]}_{=X} \underbrace{\begin{bmatrix} \beta_1^\top & \beta_2^\top & \dots & \beta_d^\top & \dots & 0 \\ 0 & \beta_1^\top & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & \dots & 0 & \beta_1^\top & \dots & \beta_d^\top \end{bmatrix}}_{=\mathcal{I}} \underbrace{\begin{bmatrix} U_{d+1} \\ U_{d+2} \\ \vdots \\ U_{T+2d} \end{bmatrix}}_{=N} \tag{43}$$

Here $\mathcal{I}$ is $\mathbb{R}^{T\times(mT+md)}$. It is known from Theorem 11 that $XX^\top \preceq \frac{3I}{2}$ with high probability and consequently $X\mathcal{I}\mathcal{I}^\top X^\top \preceq \frac{3\sigma_1^2(\mathcal{I})I}{2}$. Define $\boldsymbol{\mathcal{F}}_l = \tilde{\sigma}(\{U_l\}_{j=1}^{d+l})$ as the sigma field generated by $(\{U_l\}_{j=1}^{d+l}$. Furthermore N_l is $\boldsymbol{\mathcal{F}}_l$ measurable, and $[X\mathcal{I}]_l$ is $\boldsymbol{\mathcal{F}}_{l-1}$ measurable and we can apply Theorem 2.2. Now the proof is similar to Proposition 11.1. Following the same steps as before we get with probability at least $1-\delta$

$$||\sum_{l=0}^{T-1} X_l M_l^\top v||_2 = ||\sum_{l=0}^{T-1} [X\mathcal{I}]_l N_l||_2 \leq \sqrt{3}\sigma_1(\mathcal{I})L\sqrt{\log\frac{1}{\delta} + pd\log 3 + m}$$

and substituting $\delta \to 5^{-pd}\delta$ we get

$$||\sum_{l=0}^{T-1} X_l M_l^\top||_2 \leq \sqrt{6}\log(5)\sigma_1(\mathcal{I})L\sqrt{\log\frac{1}{\delta} + pd + m}$$

and

$$||\sum_{l=0}^{T-1} X_l M_l||_2 \leq 4\sigma_1(\mathcal{I})L\sqrt{\log\frac{1}{\delta} + pd + m}. \tag{44}$$

■

The proof for noise and covariate cross terms is almost identical to Proposition 11.2 but easier because of independence. Finally note that $\sigma_1(\mathcal{I}) \leq \sqrt{\sum_{i=1}^{d}\|\beta_i\|_2^2}\sqrt{d} = \sqrt{\|\mathcal{T}_{0,d}^{\top}v\|_2^2}\sqrt{d} \leq \beta\sqrt{d}$.

Proposition 11.3 *For $0 < \delta < 1$, we have with probability at least $1-\delta$*

$$\Big|\Big|V_T^{-1/2}\Big(\sum_{k=0}^{T}\tilde{U}^{-}_{l+d,d}\tilde{\eta}^{+\prime}_{l+1+d,d}\mathcal{T}\mathcal{O}'_{0,d}\Big)\Big|\Big|_2 \leq 4\sigma_A\sqrt{\log\frac{1}{\delta}+pd+m}$$

$$\Big|\Big|V_T^{-1/2}\Big(\sum_{k=0}^{T}\tilde{U}^{-}_{l+d,d}\tilde{\eta}^{-\prime}_{l,l}\mathcal{O}'_{d,d,l}\Big)\Big|\Big|_2 \leq 4\sigma_B\sqrt{\log\frac{1}{\delta}+pd+m}$$

$$\Big|\Big|V_T^{-1/2}\Big(\sum_{k=0}^{T}\tilde{U}^{-}_{l+d,d}\tilde{\eta}^{-\prime}_{l+d,d}\mathcal{O}'_{0,d,d}\Big)\Big|\Big|_2 \leq 4\sigma_C\sqrt{\log\frac{1}{\delta}+pd+m}$$

$$\Big|\Big|V_T^{-1/2}\Big(\sum_{k=0}^{T}\tilde{U}^{-}_{l+d,d}\tilde{w}^{+\prime}_{l+1+d,d}\Big)\Big|\Big|_2 \leq 4\sigma_D\sqrt{\log\frac{1}{\delta}+pd+m}$$

Here $\sigma = \max(\sigma_A, \sigma_B, \sigma_C, \sigma_D)$ where

$$\sigma_A \vee \sigma_C \leq \sup_{||v||_2=1}\Big|\Big|\begin{bmatrix} v^{\top}CA^d & v^{\top}CA^{d-1} & v^{\top}CA^{d-2} & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots \\ 0 & \dots & v^{\top}CA^d & \dots & v^{\top}C \end{bmatrix}\Big|\Big|_2 \leq \sum_{j=0}^{d}||CA^j||_2 \leq \beta R\sqrt{d}$$

$\sigma_B = \sqrt{\sigma(\sum_{k=1}^{d}\mathcal{T}\mathcal{O}_{d+k,T}^{\top}\mathcal{T}\mathcal{O}_{d+k,T})} \leq \beta R\sqrt{d}, \sigma_D \leq c.$

By taking the intersection of all the aforementioned events for a fixed δ we then have with probability at least $1-\delta$

$$\Big|\Big|\hat{\mathcal{H}}_{0,d,d} - \mathcal{H}_{0,d,d}\Big|\Big|_2 \leq 16\sigma\sqrt{\frac{1}{T}}\sqrt{m+pd+\log\frac{d}{\delta}}$$

12. Subspace Perturbation Results

In this section we present variants of the famous Wedin's theorem (Section 3 of Wedin (1972)) that depends on the distribution of Hankel singular values. These will be "sign free" generalizations of the gap–Free Wedin Theorem from Allen-Zhu and Li (2016). The major difference from the traditional Wedin's theorem is that the Frobenius error bound can include the dimension of the matrix; however in our case the Hankel matrix is allow to grow with T and such a bound may not be ideal. To address we introduce this mild variant of Wedin's theorem.

First we define the Hermitian dilation of a matrix.

$$\mathcal{H}(S) = \begin{bmatrix} 0 & S \\ S' & 0 \end{bmatrix}$$

The Hermitian dilation has the property that $||S_1 - S_2|| \leq \epsilon \iff ||\mathcal{H}(S_1) - \mathcal{H}(S_2)|| \leq \epsilon$. Hermitian dilations will be useful in applying Wedin's theorem for general (not symmetric) matrices.

Proposition 12.1 *Let $S, \hat{S}$ be symmetric matrices and $||S - \hat{S}|| \leq \epsilon$. Further, let $v_j, \hat{v}_j$ correspond to the j^{th} eigenvector of $S, \hat{S}$ respectively such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ and $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_n$. Then we have*

$$|\langle v_j, \hat{v}_k \rangle| \leq \frac{\epsilon}{|\lambda_j - \hat{\lambda}_k|} \tag{45}$$

if either λ_j or $\hat{\lambda}_k$ is not zero.

Proof Let $S = \lambda_j v_j v_j' + V \Lambda_{-j} V'$ and $\hat{S} = \hat{\lambda}_k \hat{v}_k \hat{v}_k' + \hat{V} \hat{\Lambda}_{-k} \hat{V}'$, wlog assume $|\lambda_j| \leq |\hat{\lambda}_k|$. Define $R = S - \hat{S}$

$$S = \hat{S} + R$$
$$v_j' S \hat{v}_k = v_j' \hat{S} \hat{v}_k + v_j' R \hat{v}_k$$

Since $v_j, \hat{v}_k$ are eigenvectors of S and $\hat{S}$ respectively.

$$\lambda_j v_j' \hat{v}_k = \hat{\lambda}_k v_j' \hat{v}_k + v_j' R \hat{v}_k$$
$$|\lambda_j - \hat{\lambda}_k||v_j' \hat{v}_k| \leq \epsilon$$

■

Proposition 12.1 gives an eigenvector subjective Wedin's theorem. Next, we show how to extend these results to arbitrary subsets of eigenvectors.

Proposition 12.2 *For $\epsilon > 0$, let S, P be two symmetric matrices such that $||S - P||_2 \leq \epsilon$. Let*

$$S = U \Sigma^S U^\top, P = V \Sigma^P V^\top$$

Let V_+ correspond to the eigenvectors of singular values $\geq \beta$, V_- correspond to the eigenvectors of singular values $\leq \alpha$ and $\bar{V}$ are the remaining ones. Define a similar partition for S. Let $\alpha < \beta$

$$||U_-^\top V_+|| \leq \frac{\epsilon}{\beta - \alpha}$$

Proof The proof is similar to before. S, P have a spectral decomposition of the form

$$S = U_+ \Sigma_+^S U_+' + U_- \Sigma_-^S U_-' + \bar{U} \Sigma^S{}_0 \bar{U}'$$
$$P = V_+ \Sigma_+^P V_+' + V_- \Sigma_-^P V_-' + \bar{V} \Sigma^P{}_0 \bar{V}'$$

Let $R = S - P$ and since U_+ is orthogonal to $U_-, \bar{U}$ and similarly for V

$$U_-' S = \Sigma_-^S U_-' = U_-' P + U_-' R$$
$$\Sigma_-^S U_-' V_+ = U_-' V_+ \Sigma_+^P + U_-' R V_+$$

Diving both sides by Σ^P

$$\begin{aligned}\Sigma_-^S U_-' V_+ (\Sigma_+^P)^{-1} &= U_-' V_+ + U_-' R V_+ (\Sigma_+^P)^{-1} \\ ||\Sigma_-^S U_-' V_+ (\Sigma_+^P)^{-1}|| &\geq ||U_-' V_+|| - ||U_-' R V_+ (\Sigma_+^P)^{-1}|| \\ \frac{\alpha}{\beta}||U_-' V_+|| &\geq ||U_-' V_+|| - \frac{\epsilon}{\beta} \\ ||U_-' V_+|| &\leq \frac{\epsilon}{\beta - \alpha}\end{aligned}$$

■

Let S_k, P_k be the best rank k approximations of S, P respectively. We develop a sequence of results to see how $||S_k - P_k||$ varies when $||S - P|| \leq \epsilon$ as a function of k.

Proposition 12.3 *Let S, P be such that*

$$||S - P|| \leq \epsilon$$

Furthermore, let ϵ be such that

$$\epsilon \leq \inf_{\{1 \leq i \leq r-1\} \cup \{s+1 \leq i \leq n\}} \Big(\frac{\sigma_i(P) - \sigma_{i+1}(P)}{2} \Big) \tag{46}$$

and U_j^S, V_j^S be the left and right singular vectors of S corresponding to $\sigma_j(S)$. There exists a unitary transformation Q such that

$$\begin{aligned}\sigma_{\max}([U_r^P, \ldots, U_s^P]Q - [U_r^S, \ldots, U_s^S]) &\leq \frac{2\epsilon}{\min\Big(\sigma_{r-1}(P) - \sigma_r(S), \sigma_s(S) - \sigma_{s+1}(P)\Big)} \\ \sigma_{\max}([V_r^P, \ldots, V_s^P]Q - [V_r^S, \ldots, V_s^S]) &\leq \frac{2\epsilon}{\min\Big(\sigma_{r-1}(P) - \sigma_r(S), \sigma_s(S) - \sigma_{s+1}(P)\Big)}.\end{aligned}$$

Proof Let $r \leq k \leq s$. First divide the indices $[1, n]$ into 3 parts $K_1 = [1, r-1], K_2 = [r, s], K_3 = [s+1, n]$. Although we focus on only three groups extension to general case will be a straight forward extension of this proof. Define the Hermitian dilation of S, P as $\mathcal{H}(S), \mathcal{H}(P)$ respectively. Then we know that the eigenvalues of $\mathcal{H}(S)$ are

$$\cup_{i=1}^n \{\sigma_i(S), -\sigma_i(S)\}$$

Further the eigenvectors corresponding to these are

$$\cup_{i=1}^n \left\{ \frac{1}{\sqrt{2}} \begin{bmatrix} u_i^S \\ v_i^S \end{bmatrix}, \frac{1}{\sqrt{2}} \begin{bmatrix} u_i^S \\ -v_i^S \end{bmatrix} \right\}$$

Similarly define the respective quantities for $\mathcal{H}(P)$. Now clearly, $||\mathcal{H}(S) - \mathcal{H}(P)|| \leq \epsilon$ since $||S - P|| \leq \epsilon$. Then by Weyl's inequality we have that

$$|\sigma_i(S) - \sigma_i(P)| \leq \epsilon$$

Now we can use Proposition 12.1. To ease notation, define $\sigma_i(S) = \lambda_i(\mathcal{H}(S))$ and $\lambda_{-i}(\mathcal{H}(S)) = -\sigma_i(S)$ and let the corresponding eigenvectors be a_i, a_{-i} for S and b_i, b_{-i} for P respectively. Note that we can make the assumption that $\langle a_i, b_i \rangle \geq 0$ for every i. This does not change any of our results because a_i, b_i are just stacking of left and right singular vectors and $u_i v_i^\top$ is identical for u_i, v_i and $-u_i, -v_i$.

Then using Proposition 12.1 we get for every $(i,j) \notin K_2 \times K_2$ and $i \neq j$

$$|\langle a_i, b_j \rangle| \leq \frac{\epsilon}{|\sigma_i(S) - \sigma_j(P)|} \tag{47}$$

similarly

$$|\langle a_{-i}, b_j \rangle| \leq \frac{\epsilon}{|\sigma_i(S) + \sigma_j(P)|} \tag{48}$$

Since

$$a_i = \frac{1}{\sqrt{2}} \begin{bmatrix} u_i^S \\ v_i^S \end{bmatrix}, a_{-i} = \frac{1}{\sqrt{2}} \begin{bmatrix} u_i^S \\ -v_i^S \end{bmatrix}, b_j = \frac{1}{\sqrt{2}} \begin{bmatrix} u_i^P \\ v_i^P \end{bmatrix}$$

and $\sigma_i(S), \sigma_i(P) \geq 0$ we have by adding Eq. (47),(48) that

$$\max\left(|\langle u_i^S, u_j^P \rangle|, |\langle v_i^S, v_j^P \rangle|\right) \leq \frac{\epsilon}{|\sigma_i(S) - \sigma_j(P)|}$$

Define $U_{K_i}^S$ to be the matrix formed by the orthornormal vectors $\{a_j\}_{j \in K_i}$ and $U_{K_{-i}}^S$ to be the matrix formed by the orthonormal vectors $\{a_j\}_{j \in -K_i}$. Define similar quantities for P. Then

$$\begin{aligned}
(U_{K_2}^S)^\top U_{K_2}^P (U_{K_2}^P)^\top U_{K_2}^S &= (U_{K_2}^S)^\top (I - \sum_{j \neq 2} U_{K_j}^P (U_{K_j}^P)^\top) U_{K_2}^S \\
&= (U_{K_2}^S)^\top (I - \sum_{|j| \neq 2} U_{K_j}^P (U_{K_j}^P)^\top - U_{K_{-2}}^P (U_{K_{-2}}^P)^\top) U_{K_2}^S \\
&= I - (U_{K_2}^S)^\top \sum_{|j| \neq 2} U_{K_j}^P (U_{K_j}^P)^\top U_{K_2}^S - (U_{K_2}^S)^\top U_{K_{-2}}^P (U_{K_{-2}}^P)^\top U_{K_2}^S
\end{aligned} \tag{49}$$

Now K_1, K_{-1} corresponds to eigenvectors where singular values $\geq \sigma_{r-1}(P)$, K_3, K_{-3} corresponds to eigenvectors where singular values $\leq \sigma_{s+1}(P)$. We are in a position to use Proposition 12.2. Using that on Eq. (49) we get the following relation

$$\begin{aligned}
(U_{K_2}^P)^\top U_{K_2}^S (U_{K_2}^S)^\top U_{K_2}^P \succeq I \Bigg(1 - \frac{\epsilon^2}{(\sigma_{r-1}(P) - \sigma_s(S))^2} - \frac{\epsilon^2}{(\sigma_s(S) - \sigma_{s+1}(P))^2} \Bigg) \\
- (U_{K_2}^S)^\top U_{K_{-2}}^P (U_{K_{-2}}^P)^\top U_{K_2}^S
\end{aligned} \tag{50}$$

In the Eq. (50) we need to upper bound $(U_{K_2}^S)^\top U_{K_{-2}}^P (U_{K_{-2}}^P)^\top U_{K_2}^S$. To this end we will exploit the fact that all singular values corresponding to $U_{K_2}^S$ are the same. Since $||\mathcal{H}(S) - \mathcal{H}(P)|| \leq \epsilon$, then

$$\begin{aligned}
\mathcal{H}(S) &= U_{K_2}^S \Sigma_{K_2}^S (U_{K_2}^S)^\top + U_{K_{-2}}^S \Sigma_{K_{-2}}^S (U_{K_{-2}}^S)^\top + U_{K_0}^S \Sigma_{K_0}^S (U_{K_0}^S)^\top \\
\mathcal{H}(P) &= U_{K_2}^P \Sigma_{K_2}^P (U_{K_2}^P)^\top + U_{K_{-2}}^P \Sigma_{K_{-2}}^P (U_{K_{-2}}^P)^\top + U_{K_0}^P \Sigma_{K_0}^P (U_{K_0}^P)^\top
\end{aligned}$$

Then by pre–multiplying and post–multiplying we get

$$(U^S_{K_2})^\top \mathcal{H}(S) U^P_{K_{-2}} = \Sigma^S_{K_2} (U^S_{K_2})^\top U^P_{K_{-2}}$$
$$(U^S_{K_2})^\top \mathcal{H}(P) U^P_{K_{-2}} = (U^S_{K_2})^\top U^P_{K_{-2}} \Sigma^P_{K_{-2}}$$

Let $\mathcal{H}(S) - \mathcal{H}(P) = R$ then

$$(U^S_{K_2})^\top (\mathcal{H}(S) - \mathcal{H}(P)) U^P_{K_{-2}} = (U^S_{K_2})^\top R U^P_{K_{-2}}$$
$$\Sigma^S_{K_2} (U^S_{K_2})^\top U^P_{K_{-2}} - (U^S_{K_2})^\top U^P_{K_{-2}} \Sigma^P_{K_{-2}} = (U^S_{K_2})^\top R U^P_{K_{-2}}$$

Since $\Sigma^S_{K_2} = \sigma_s(A) I$ then

$$||(U^S_{K_2})^\top U^P_{K_{-2}} (\sigma_s(S) I - \Sigma^P_{K_{-2}})|| = ||(U^S_{K_2})^\top R U^P_{K_{-2}}||$$
$$||(U^S_{K_2})^\top U^P_{K_{-2}}|| \leq \frac{\epsilon}{\sigma_s(S) + \sigma_s(P)}$$

Similarly

$$||(U^P_{K_2})^\top U^S_{K_{-2}}|| \leq \frac{\epsilon}{\sigma_s(P) + \sigma_s(S)}$$

Since $\sigma_s(P) + \sigma_s(S) \geq \sigma_s(S) - \sigma_{s+1}(P)$ combining this with Eq. (50) we get

$$\sigma_{\min}((U^S_{K_2})^\top U^P_{K_2}) \geq 1 - \frac{3\epsilon^2}{\min\Big(\sigma_{r-1}(P) - \sigma_s(S), \sigma_s(S) - \sigma_{s+1}(P)\Big)^2} \tag{51}$$

Since

$$\epsilon \leq \inf_i \Big(\frac{\sigma_i(P) - \sigma_{i+1}(P)}{2}\Big),$$

for Eq. (51), we use the inequality $\sqrt{1-x^2} \geq 1 - x^2$ whenever $x < 1$ which is true when Eq. (46) is true. This means that there exists unitary transformation Q such that

$$||U^S_{K_2} - U^P_{K_2} Q|| \leq \frac{2\epsilon}{\min\Big(\sigma_{r-1}(P) - \sigma_s(S), \sigma_s(S) - \sigma_{s+1}(P)\Big)}$$

■

Remark 12 *Note that S, P will be Hermitian dilations of $\mathcal{H}_{0,\infty,\infty}, \hat{\mathcal{H}}_{0,\hat{d},\hat{d}}$ respectively in our case. Since the singular vectors of S (and P) are simply stacked version of singular vectors of $\mathcal{H}_{0,\infty,\infty}$ (and $\hat{\mathcal{H}}_{0,\hat{d},\hat{d}}$), our results hold directly for the singular vectors of $\mathcal{H}_{0,\infty,\infty}$ (and $\hat{\mathcal{H}}_{0,\hat{d},\hat{d}}$)*

Let $r \leq k \leq s$. First divide the indices $[1, n]$ into 3 parts $K_1 = [1, r-1], K_2 = [r, s], K_3 = [s+1, n]$.

Proposition 12.4 (System Reduction) *Let $||S-P|| \leq \epsilon$ and the singular values of S be arranged as follows:*

$$\sigma_1(S) > \ldots > \sigma_{r-1}(S) > \sigma_r(S) \geq \sigma_{r+1}(S) \geq \ldots \geq \sigma_s(S) > \sigma_{s+1}(S) > \ldots \sigma_n(S) > \sigma_{n+1}(S) = 0$$

Furthermore, let ϵ be such that

$$\epsilon \leq \inf_{\{1 \leq i \leq r-1\} \cup \{s+1 \leq i \leq n\}} \Big(\frac{\sigma_i(P) - \sigma_{i+1}(P)}{2} \Big). \tag{52}$$

Define $K_0 = K_1 \cup K_2$, then

$$||U^S_{K_0}(\Sigma^S_{K_0})^{1/2} - U^P_{K_0}(\Sigma^P_{K_0})^{1/2}||_2 \leq 2\epsilon \sqrt{\sum_{i=1}^{r-1} \sigma_i/\zeta_i^2 + \sigma_r/\zeta_r^2} + \sup_{1 \leq i \leq s} |\sqrt{\sigma_i} - \sqrt{\hat{\sigma}_i}|$$

and $\sigma_i = \sigma_i(S), \hat{\sigma}_i = \sigma_i(P)$. Here $\zeta_i = \min(\sigma_i - \sigma_{i+1}, \sigma_i - \sigma_{i+1})$ and $\zeta_r = \min(\sigma_{r-1} - \sigma_r, \sigma_s - \sigma_{s+1})$.

Proof

Since $U^S_{K_0} = [U^S_{K_1} U^S_{K_2}]$ and likewise for B, we can separate the analysis for K_1, K_2 as follows

$$\begin{aligned} ||U^S_{K_0}(\Sigma^S_{K_0})^{1/2} - U^P_{K_0}(\Sigma^P_{K_0})^{1/2}|| &\leq ||(U^S_{K_0} - U^P_{K_0})(\Sigma^S_{K_0})^{1/2}|| + ||U^P_{K_0}((\Sigma^S_{K_0})^{1/2} - (\Sigma^P_{K_0})^{1/2})|| \\ &\leq \sqrt{||(U^S_{K_1} - U^P_{K_1})(\Sigma^S_{K_1})^{1/2}||_2^2 + ||(U^S_{K_2} - U^P_{K_2})(\Sigma^S_{K_2})^{1/2}||_2^2} \\ &\quad + ||(\Sigma^S_{K_0})^{1/2} - (\Sigma^P_{K_0})^{1/2}|| \end{aligned}$$

Now $||(\Sigma^S_{K_0})^{1/2} - (\Sigma^P_{K_0})^{1/2}|| = \sup_l |\sqrt{\sigma_l(S)} - \sqrt{\sigma_l(P)}|$. Recall that $\sigma_r(S) = \ldots = \sigma_k(S) = \ldots = \sigma_{s-1}(S)$ and by conditions on ϵ we are guaranteed that $\frac{\epsilon}{\sigma_i - \sigma_j} < 1/2$ for all $1 \leq i \neq j \leq r$. We will combine our previous results in Proposition 12.1–12.3 to prove this claim. Specifically from Proposition 12.3 we have

$$||(U^S_{K_2} - U^P_{K_2})(\Sigma^S_{K_2})^{1/2}|| \leq \frac{2\epsilon\sqrt{\sigma_r(S)}}{\min\Big(\sigma_{r-1}(P) - \sigma_r(S), \sigma_r(S) - \sigma_{s+1}(P)\Big)}$$

On the remaining term we will use Proposition 12.3 on each column

$$\begin{aligned} ||(U^S_{K_1} - U^P_{K_1})(\Sigma^S_{K_1})^{1/2}|| \leq |||[\sqrt{\sigma_1(S)}c_1, \ldots, \sqrt{\sigma_{|K_1|}(S)}c_{|K_1|}]||| &\leq \sqrt{\sum_{j=1}^{r-1} \sigma_j^2 ||c_j||^2} \\ &\leq \epsilon \sqrt{\sum_{j=1}^{r-1} \frac{2\sigma_j(S)}{\min\Big(\sigma_{j-1}(P) - \sigma_j(S), \sigma_j(S) - \sigma_{j+1}(P)\Big)^2}} \end{aligned}$$

■

In the context of our system identification, $S = \mathcal{H}_{0,\infty,\infty}$ and $P = \hat{\mathcal{H}}_{0,\hat{d},\hat{d}}$. P will be made

compatible by padding it with zeros to make it doubly infinite. Then $U^S_{K_0}, U^P_{K_0}$ (after padding) has infinite rows. Define $Z_0 = U^S_{K_0}(\Sigma^S_{K_0})^{1/2}(1:,:), Z_1 = U^S_{K_0}(\Sigma^S_{K_0})^{1/2}(p+1:,:)$ (both infinite length) and similarly we will have $\hat{Z}_0, \hat{Z}_1$. Note that from a computational perspective we do not need to Z_0, Z_1; we only need to work with $\hat{Z}_0 = U^P_{K_0}(\Sigma^P_{K_0})^{1/2}(1:,:), \hat{Z}_1 = U^P_{K_0}(\Sigma^P_{K_0})^{1/2}(p+1:,:)$ and since most of it is just zero padding we can simply compute on $\hat{Z}_0(1:pd,:), \hat{Z}_1(1:pd,:)$.

Proposition 12.5 *Assume $Z_1 = Z_0 A$. Furthermore, $||S-P||_2 \leq \epsilon$ and let ϵ be such that*

$$\epsilon \leq \inf_{\{1\leq i\leq r-1\}\cup\{s+1\leq i\leq n\}} \Big(\frac{\sigma_i(P)-\sigma_{i+1}(P)}{2}\Big) \tag{53}$$

then

$$||(Z_0'Z_0)^{-1}Z_0'Z_1 - (\hat{Z}_0'\hat{Z}_0)^{-1}\hat{Z}_0'\hat{Z}_1|| \leq \frac{\mathcal{C}\epsilon(\gamma+1)}{\sigma_s}\left(\sqrt{\frac{\sigma_s^2}{((\sigma_s-\sigma_{s+1})\wedge(\sigma_{r-1}-\sigma_s))^2}} + \sqrt{\sum_{i=1}^{r-1}\frac{\sigma_i\sigma_s}{(\sigma_i-\sigma_{i+1})^2\wedge(\sigma_{i-1}-\sigma_i)^2}}\right)$$

where $\sigma_1(A) \leq \gamma$.

Proof Note that $Z_1 = Z_0 A$, then

$$\begin{aligned}
&||(Z_0'Z_0)^{-1}Z_0'Z_1 - (\hat{Z}_0'\hat{Z}_0)^{-1}\hat{Z}_0'\hat{Z}_1||_2 \\
=&||A - (\hat{Z}_0'\hat{Z}_0)^{-1}\hat{Z}_0'\hat{Z}_1||_2 = ||(\hat{Z}_0'\hat{Z}_0)^{-1}\hat{Z}_0'\hat{Z}_0 A - (\hat{Z}_0'\hat{Z}_0)^{-1}\hat{Z}_0'\hat{Z}_1||_2 \\
=&||(\hat{Z}_0'\hat{Z}_0)^{-1}\hat{Z}_0'\hat{Z}_0 A - (\hat{Z}_0'\hat{Z}_0)^{-1}\hat{Z}_0'Z_0 A + (\hat{Z}_0'\hat{Z}_0)^{-1}\hat{Z}_0'Z_0 A - (\hat{Z}_0'\hat{Z}_0)^{-1}\hat{Z}_0'\hat{Z}_1||_2 \\
\leq&||(\hat{Z}_0'\hat{Z}_0)^{-1}\hat{Z}_0'\hat{Z}_0 A - (\hat{Z}_0'\hat{Z}_0)^{-1}\hat{Z}_0'Z_0 A||_2 + ||(\hat{Z}_0'\hat{Z}_0)^{-1}\hat{Z}_0'Z_0 A - (\hat{Z}_0'\hat{Z}_0)^{-1}\hat{Z}_0'\hat{Z}_1||_2 \\
\leq&||(\hat{Z}_0'\hat{Z}_0)^{-1}\hat{Z}_0'||_2\Big(||Z_0 A - \hat{Z}_0 A||_2 + ||\underbrace{Z_0 A}_{\text{Shifted version of } Z_0} - \hat{Z}_1||_2\Big)
\end{aligned}$$

Now, $||(\hat{Z}_0'\hat{Z}_0)^{-1}\hat{Z}_0'||_2 \leq (\sqrt{\sigma_s-\epsilon})^{-1}$, $||Z_0 A - \hat{Z}_1||_2 \leq ||Z_0 - \hat{Z}_0||_2$ since $Z_1 = Z_0 A$ is a submatrix of Z_0 and $\hat{Z}_1$ is a submatrix of $\hat{Z}_0$ we have $||Z_0 A - \hat{Z}_1||_2 \leq ||Z_0 - \hat{Z}_0||_2$ and $||Z_0 A - \hat{Z}_0 A||_2 \leq ||A||_2||Z_0 - \hat{Z}_0||_2$

$$\leq \frac{c\epsilon(\gamma+1)}{\sigma_s}\left(\sqrt{\frac{\sigma_s^2}{((\sigma_s-\sigma_{s+1})\wedge(\sigma_{r-1}-\sigma_s))^2}} + \sqrt{\sum_{i=1}^{r-1}\frac{\sigma_i\sigma_s}{(\sigma_i-\sigma_{i+1})^2\wedge(\sigma_{i-1}-\sigma_i)^2}}\right)$$

■

13. Hankel Matrix Estimation Results

In this section we provide the proof for Theorem 5.2. For any matrix P, we define its doubly infinite extension $\bar{P}$ as

$$\bar{P} = \begin{bmatrix} P & 0 & \dots \\ 0 & 0 & \dots \\ \vdots & \vdots & \vdots \end{bmatrix} \tag{54}$$

Proposition 13.1 *Fix $d > 0$. Then we have*

$$||\mathcal{H}_{d,\infty,\infty}||_2 \leq ||\mathcal{H}_{0,\infty,\infty} - \bar{\mathcal{H}}_{0,d,d}||_2 \leq \sqrt{2}||\mathcal{H}_{d,\infty,\infty}||_2 \leq \sqrt{2}||\mathcal{T}_{d,\infty}||_2$$

Proof Define $\tilde{C}_d, \tilde{B}_d$ as follows

$$\tilde{C}_d = \begin{bmatrix} 0_{md\times n} \\ C \\ CA \\ \vdots \end{bmatrix}$$
$$\tilde{B}_d = \begin{bmatrix} 0_{n\times pd} & B & AB & \dots \end{bmatrix}$$

Now pad $\mathcal{H}_{0,d,d}$ with zeros to make it a doubly infinite matrix and call it $\bar{\mathcal{H}}_{0,d,d}$ and we get that

$$||\bar{\mathcal{H}}_{0,d,d} - \mathcal{H}_{0,\infty,\infty}|| = \begin{bmatrix} 0 & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$$

Note here that M_{21} and $M_0 = \begin{bmatrix} M_{12} \\ M_{22} \end{bmatrix}$ are infinite matrices. Further $||\mathcal{H}_{d,\infty,\infty}||_2 = ||M_0||_2 \geq ||M_{21}||_2$. Then

$$||\bar{\mathcal{H}}_{0,d,d} - \mathcal{H}_{0,\infty,\infty}|| \leq \sqrt{||M_{12}||_2^2 + ||M_0||_2^2} \leq \sqrt{2}||\mathcal{H}_{d,\infty,\infty}||_2$$

Further $||\bar{\mathcal{H}}_{0,d,d} - \mathcal{H}_{0,\infty,\infty}|| \geq ||M_0|| = ||\mathcal{H}_{d,\infty,\infty}||_2$. ■

Proposition 13.2 *For any $d_1 \geq d_2$, we have*

$$||\mathcal{H}_{0,\infty,\infty} - \bar{\mathcal{H}}_{0,d_1,d_1}||_2 \leq \sqrt{2}||\mathcal{H}_{0,\infty,\infty} - \bar{\mathcal{H}}_{0,d_2,d_2}||_2$$

Proof Since $||\mathcal{H}_{d_1,\infty,\infty}||_2 \leq ||\mathcal{H}_{0,\infty,\infty} - \bar{\mathcal{H}}_{0,d_1,d_1}||_2 \leq \sqrt{2}||\mathcal{H}_{d_1,\infty,\infty}||_2$ from Proposition 13.1. It is clear that $||\mathcal{H}_{d_1,\infty,\infty}||_2 \leq ||\mathcal{H}_{d_2,\infty,\infty}||_2$. Then

$$\frac{1}{\sqrt{2}}||\mathcal{H}_{0,\infty,\infty} - \bar{\mathcal{H}}_{0,d_1,d_1}||_2 \leq ||\mathcal{H}_{d_1,\infty,\infty}||_2 \leq ||\mathcal{H}_{d_2,\infty,\infty}||_2 \leq ||\mathcal{H}_{0,\infty,\infty} - \bar{\mathcal{H}}_{0,d_2,d_2}||_2$$

■

Proposition 13.3 *Fix $d > 0$. Then*

$$||\mathcal{T}_{d,\infty}(M)||_2 \leq \frac{\tilde{M}\rho(A)^d}{1-\rho(A)}$$

Proof Recall that

$$\mathcal{T}_{d,\infty}(M) = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ CA^dB & 0 & 0 & \dots & 0 \\ CA^{d+1}B & CA^dB & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \end{bmatrix}$$

Then $||\mathcal{T}_{d,\infty}(M)||_2 \leq \sum_{j=d}^{\infty} ||CA^jB||_2$. Now from Eq. 4.1 and Lemma 4.1 in Tu et al. (2017) we get that $||CA^jB||_2 \leq \tilde{M}\rho(A)^j$. Then

$$\sum_{j=d}^{\infty} ||CA^jB||_2 \leq \frac{\tilde{M}\rho(A)^d}{1-\rho(A)}$$

■

Remark 13 *Proposition 13.3 is just needed to show exponential decay and is not precise. Please refer to Tu et al. (2017) for explicit rates.*

Next we show that $T_*^{(\kappa)}(\delta)$ and $d_*(T,\delta)$ defined in Eq. (16) given by

$$d_*(T,\delta) = \inf\left\{ d \middle| 16\beta R\sqrt{d}\sqrt{\frac{m+pd+\log\frac{T}{\delta}}{T}} \geq ||\mathcal{H}_{0,d,d} - \mathcal{H}_{0,\infty,\infty}||_2 \right\}$$

$$T_*^{(\kappa)}(\delta) = \inf\left\{ T \middle| \frac{T}{cm^2\log^3(Tm/\delta)} \geq d_*(T,\delta),\ \ d_*(T,\delta) \leq \frac{\kappa d_*(\frac{T}{\kappa^2},\delta)}{8} \right\} \tag{55}$$

The existence of $d_*(T,\delta)$ is predicated on the finiteness of $T_*^{(\kappa)}(\delta)$ which we discuss below.

13.1 Existence of $T_*^{(\kappa)}(\delta) < \infty$

Construct two sets

$$T_1(\delta) = \inf\left\{ T \middle| d_*(T,\delta) \in \mathcal{D}(T) \right\} \tag{56}$$

$$T_2(\delta) = \inf\left\{ T \middle| d_*(t,\delta) \leq \frac{\kappa d_*(\frac{t}{\kappa^2},\delta)}{8},\quad \forall t \geq T \right\} \tag{57}$$

Clearly, $T_*^{(\kappa)}(\delta) < T_1(\delta) \vee T_2(\delta)$. A key assumption in the statement of our results is that $T_*^{(\kappa)}(\delta) < \infty$. We will show that it is indeed true. Let $\kappa \geq 16$.

Proposition 13.4 *For a fixed $\delta > 0$, $T_1(\delta) < \infty$ with $d_*(T,\delta) \leq \frac{c\log(cT+\log\frac{1}{\delta}) - \log R + \log(\tilde{M}/\beta)}{\log\frac{1}{\rho}}$. Here $\rho = \rho(A)$.*

Proof Note the form for $d_*(T,\delta)$, it is the minimum d that satisfies

$$16\beta R\sqrt{d}\sqrt{\frac{m+pd+\log\frac{T}{\delta}}{T}} \geq ||\mathcal{H}_{0,d,d} - \mathcal{H}_{0,\infty,\infty}||_2$$

Since from Proposition 13.1 and 13.3 we have $||\mathcal{H}_{0,d,d} - \mathcal{H}_{0,\infty,\infty}||_2 \leq \frac{3\tilde{M}\rho^d}{1-\rho(A)}$, then $d_*(T,\delta) \leq d$ that satisfies

$$16\beta R\sqrt{d}\sqrt{\frac{m+pd+\log\frac{T}{\delta}}{T}} \geq \frac{3\tilde{M}\rho^d}{1-\rho(A)}$$

which immediately implies $d_*(T,\delta) \leq d = \frac{c\log(cT - \log R + \log\frac{1}{\delta}) + \log(\tilde{M}/\beta)}{\log\frac{1}{\rho}}$, *i.e.*, $d_*(T,\delta)$ is at most logarithmic in T. As a result, for a large enough T

$$cm^2 d\log^2(d)\log^2(m^2/\delta) + cd\log^3(2d) \geq \frac{c\log(cT + \log\frac{1}{\delta}) - \log R + \log(\tilde{M}/\beta)}{\log\frac{1}{\rho}}$$

■

The intuition behind $T_2(\delta)$ is the following: $d_*(T,\delta)$ grows at most logarithmically in T, as is clear from the previous proof. Then $T_2(\delta)$ is the point where $d_*(T,\delta)$ is still growing as $\sqrt{T}$ (*i.e.*, "mixing" has not happened) but at a slightly reduced rate.

Proposition 13.5 *For a fixed $\delta > 0$, $T_2(\delta) < \infty$.*

Proof Recall from the proof of Proposition 13.1 that $||\mathcal{H}_{d,\infty,\infty}|| \leq ||\mathcal{H}_{0,\infty,\infty} - \mathcal{H}_{0,d,d}|| \leq \sqrt{2}||\mathcal{H}_{d,\infty,\infty}||$. Now $\mathcal{H}_{d,\infty,\infty}$ can be written as

$$\mathcal{H}_{d,\infty,\infty} = \underbrace{\begin{bmatrix} C \\ CA \\ \vdots \end{bmatrix}}_{=\tilde{C}} A^d \underbrace{[B, AB, \ldots]}_{=\tilde{B}}$$

Define $P_d = A^d\tilde{B}\tilde{B}^\top(A^d)^\top$. Let d_κ be such that for every $d \geq d_\kappa$ and $\kappa \geq 16$

$$P_d \preceq \frac{1}{4\kappa}P_0 \tag{58}$$

Clearly such a $d_\kappa < \infty$ would exist because $P_0 \neq 0$ but $\lim_{d\to\infty} P_d = 0$. Then observe that $P_{2d} \preceq \frac{1}{4\kappa}P_d$. Then for every $d \geq d_\kappa$ we have that

$$||\mathcal{H}_{d,\infty,\infty}|| \geq 4\kappa||\mathcal{H}_{2d,\infty,\infty}||$$

Let

$$T \geq \frac{4d_\kappa \cdot (16)^2 \cdot \beta^2 R^2}{\sigma_0^2}(d_\kappa p + \log(T/\delta)) \tag{59}$$

where $\sigma_0 = ||\mathcal{H}_{d_\kappa,\infty,\infty}||$. Assume that $\sigma_0 > 0$ (if not then are condition is trivially true). Then simple computation shows that

$$||\mathcal{H}_{0,d_\kappa,d_\kappa} - \mathcal{H}_{0,\infty,\infty}|| \geq ||\mathcal{H}_{d_\kappa,\infty,\infty}|| \geq \underbrace{16\beta R\sqrt{d_\kappa}\sqrt{\frac{m + pd_\kappa + \log\frac{T}{\delta}}{T}}}_{<\frac{\sigma_0}{2}}$$

This implies that $d_* = d_*(T,\delta) \geq d_\kappa$ for T prescribed as above (ensured by Proposition 13.2). But from our discussion above we also have

$$||\mathcal{H}_{0,d_*,d_*} - \mathcal{H}_{0,\infty,\infty}|| \geq ||\mathcal{H}_{d_*,\infty,\infty}|| \geq 4\kappa||\mathcal{H}_{2d_*,\infty,\infty}|| \geq 2\kappa||\mathcal{H}_{0,2d_*,2d_*} - \mathcal{H}_{0,\infty,\infty}||$$

This means that if

$$||\mathcal{H}_{0,d_*,d_*} - \mathcal{H}_{0,\infty,\infty}|| \leq 16\beta R\sqrt{d_*}\sqrt{\frac{m + pd_* + \log\frac{T}{\delta}}{T}}$$

then

$$||\mathcal{H}_{0,2d_*,2d_*} - \mathcal{H}_{0,\infty,\infty}|| \leq \frac{16}{2\kappa}\beta R\sqrt{d_*}\sqrt{\frac{m + pd_* + \log\frac{T}{\delta}}{T}} \leq 16\beta R\sqrt{2d_*}\sqrt{\frac{m + 2pd_* + \log\frac{\kappa^2 T}{\delta}}{\kappa^2 T}}$$

which implies that $d_*(\kappa^2 T,\delta) \leq 2d_*(T,\delta)$. The inequality follows from the definition of $d_*(\kappa^2 T,\delta)$. Furthermore, if $\kappa \geq 16$, $2d_*(T,\delta) \leq \frac{\kappa}{8}d_*(T,\delta)$ whenever T is greater than a certain finite threshold of Eq. (59). ■

Eq. (58) happens when $\sigma(A^d)^2 \leq \frac{1}{4\kappa} \implies d_\kappa = \mathcal{O}\Big(\frac{\log\kappa}{\log\frac{1}{\rho}}\Big)$ where $\rho = \rho(A)$ and $T_2(\delta) \leq cT_1(\delta)$. It should be noted that the dependence of $T_i(\delta)$ on $\log\frac{1}{\rho}$ is worst case, *i.e.*, there exists some "bad" LTI system that gives this dependence and it is quite likely $T_i(\delta)$ is much smaller. The condition $T \geq T_1(\delta) \vee T_2(\delta)$ simply requires that we capture some reasonable portion of the dynamics and not necessarily the entire dynamics.

13.2 Proof of Theorem 5.2

Proposition 13.6 *Let $T \geq T_*^{(\kappa)}(\delta)$ and $d_* = d_*(T,\delta)$ then*

$$||\mathcal{H}_{0,\infty,\infty} - \hat{\mathcal{H}}_{0,d_*,d_*}|| \leq 2c\beta R\sqrt{\frac{d_*}{T}}\sqrt{m + pd_* + \log\frac{T}{\delta}}$$

Proof Consider the following error

$$||\mathcal{H}_{0,\infty,\infty} - \hat{\mathcal{H}}_{0,d_*,d_*}||_2 \leq ||\mathcal{H}_{0,d_*,d_*} - \hat{\mathcal{H}}_{0,d_*,d_*}||_2 + ||\mathcal{H}_{0,\infty,\infty} - \mathcal{H}_{0,d_*,d_*}||_2$$

From Proposition 13.1 and Eq. (55) we get that

$$||\mathcal{H}_{0,\infty,\infty} - \mathcal{H}_{0,d_*,d_*}||_2 \leq 16\beta R\sqrt{\frac{d_*}{T}}\sqrt{m + pd_* + \log\frac{T}{\delta}}$$

Since from Theorem 5.1

$$||\mathcal{H}_{0,d_*,d_*} - \hat{\mathcal{H}}_{0,d_*,d_*}||_2 \leq 16\beta R\sqrt{\frac{d_*}{T}}\sqrt{m + pd_* + \log\frac{T}{\delta}}$$
$$||\mathcal{H}_{0,\infty,\infty} - \hat{\mathcal{H}}_{0,d_*,d_*}||_2 \leq 32\beta R\sqrt{\frac{d_*}{T}}\sqrt{m + pd_* + \log\frac{T}{\delta}} \tag{60}$$

■

Recall the adaptive rule to choose d in Algorithm 1. From Theorem 5.1 we know that for every $d \in \mathcal{D}(T)$ we have with probability at least $1-\delta$.

$$||\mathcal{H}_{0,d,d} - \hat{\mathcal{H}}_{0,d,d}||_2 \leq 16\beta R\sqrt{d}\left(\sqrt{m + \frac{dp}{T} + \frac{\log\frac{T}{\delta}}{T}}\right)$$

Let $\alpha(l) = \sqrt{l}\left(\sqrt{\frac{lp}{T} + \frac{\log\frac{T}{\delta}}{T}}\right)$. Then consider the following adaptive rule

$$d_0(T,\delta) = \inf\left\{l \middle| ||\hat{\mathcal{H}}_{0,l,l} - \hat{\mathcal{H}}_{0,h,h}||_2 \leq 16\beta R(2\alpha(l) + \alpha(h)) \ \ \forall h \in \mathcal{D}(T), h \geq l\right\} \tag{61}$$
$$\hat{d} = \hat{d}(T,\delta) = d_0(T,\delta) \vee \log\left(\frac{T}{\delta}\right) \tag{62}$$

for the same universal constant c as Theorem 5.1. Let $d_*(T,\delta)$ be as Eq. (55). Recall that $d_* = d_*(T,\delta)$ is the point where estimation error dominates the finite truncation error. Unfortunately, we do not have apriori knowledge of $d_*(T,\delta)$ to use in the algorithm. Therefore, we will simply use Eq. (62) as our proxy. The goal will be to bound $||\hat{\mathcal{H}}_{0,\hat{d},\hat{d}} - \mathcal{H}_{0,\infty,\infty}||_2$

Proposition 13.7 *Let $T \geq T_*^{(\kappa)}(\delta)$, $d_*(T,\delta)$ be as in Eq. (55) and $\hat{d}$ be as in Eq. (62). Then with probability at least $1-\delta$ we have*

$$\hat{d} \leq d_*(T,\delta) \vee \log\left(\frac{T}{\delta}\right)$$

Proof Let $d_* = d_*(T,\delta)$. First for all $h \in \mathcal{D}(T) \geq d_*$, we note

$$||\hat{\mathcal{H}}_{0,d_*,d_*} - \hat{\mathcal{H}}_{0,h,h}||_2 \leq ||\hat{\mathcal{H}}_{0,d_*,d_*} - \mathcal{H}_{0,d_*,d_*}|| + ||\mathcal{H}_{0,h,h} - \hat{\mathcal{H}}_{0,h,h}||_2 + ||\mathcal{H}_{0,h,h} - \mathcal{H}_{0,d_*,d_*}||_2$$
$$\underbrace{\leq}_{\infty > h \geq d_*} ||\hat{\mathcal{H}}_{0,d_*,d_*} - \mathcal{H}_{0,d_*,d_*}||_2 + ||\mathcal{H}_{0,h,h} - \hat{\mathcal{H}}_{0,h,h}||_2 + ||\mathcal{H}_{0,\infty,\infty} - \mathcal{H}_{0,d_*,d_*}||_2.$$

We use the property that $||\mathcal{H}_{0,\infty,\infty} - \mathcal{H}_{0,d_*,d_*}||_2 \geq ||\mathcal{H}_{0,h,h} - \mathcal{H}_{0,d_*,d_*}||_2$. Furthermore, because of the properties of d_* we have

$$||\mathcal{H}_{0,\infty,\infty} - \mathcal{H}_{0,d_*,d_*}||_2 \leq 16\beta R\alpha(d_*)$$

and

$$||\hat{\mathcal{H}}_{0,d_*,d_*} - \mathcal{H}_{0,d_*,d_*}||_2 \leq 16\beta R\alpha(d_*), \quad ||\mathcal{H}_{0,h,h} - \hat{\mathcal{H}}_{0,h,h}||_2 \leq 16\beta R\alpha(h). \tag{63}$$

and

$$||\hat{\mathcal{H}}_{0,d_*,d_*} - \hat{\mathcal{H}}_{0,h,h}||_2 \leq 16\beta R(2\alpha(d_*) + \alpha(h)).$$

This implies that $d_0(T,\delta) \leq d_*$ and the assertion follows. ■

We have the following key lemma about the behavior of $\hat{\mathcal{H}}_{0,\hat{d},\hat{d}}$.

Lemma 13.1 *For a fixed $\kappa \geq 20$, whenever $T \geq T_*^{(\kappa)}(\delta)$ we have with probability at least $1-\delta$*

$$||\mathcal{H}_{0,\infty,\infty} - \hat{\mathcal{H}}_{0,\hat{d},\hat{d}}||_2 \leq 3c\beta R\alpha\left(\max\left(d_*(T,\delta), \log\left(\frac{T}{\delta}\right)\right)\right) \tag{64}$$

Furthermore, $\hat{d} = O(\log \frac{T}{\delta})$.

Proof Let $d_* > \hat{d}$ then

$$\begin{aligned}||\mathcal{H}_{0,\infty,\infty} - \hat{\mathcal{H}}_{0,\hat{d},\hat{d}}||_2 &\leq ||\mathcal{H}_{0,\infty,\infty} - \mathcal{H}_{0,d_*,d_*}||_2 + ||\hat{\mathcal{H}}_{0,\hat{d},\hat{d}} - \mathcal{H}_{0,\hat{d},\hat{d}}||_2 + ||\hat{\mathcal{H}}_{0,d_*,d_*} - \mathcal{H}_{0,d_*,d_*}||_2 \\ &\leq 3c\beta R\alpha(d_*)\end{aligned}$$

If $\hat{d} > d_*$ then

$$\begin{aligned}||\mathcal{H}_{0,\infty,\infty} - \hat{\mathcal{H}}_{0,\hat{d},\hat{d}}||_2 &\leq ||\mathcal{H}_{0,\infty,\infty} - \mathcal{H}_{0,\hat{d},\hat{d}}||_2 + ||\hat{\mathcal{H}}_{0,\hat{d},\hat{d}} - \mathcal{H}_{0,\hat{d},\hat{d}}||_2 = 2||\hat{\mathcal{H}}_{0,\hat{d},\hat{d}} - \mathcal{H}_{0,\hat{d},\hat{d}}||_2 \\ &\leq 2c\beta R\alpha(\hat{d}) = 2c\beta R\alpha\left(\log\left(\frac{T}{\delta}\right)\right)\end{aligned}$$

where the equality follows from Proposition 13.7. The fact that $\hat{d} = O(\log \frac{T}{\delta})$ follows from Proposition 13.1. ■

In the following we will use $\mathcal{H}_l = \mathcal{H}_{0,l,l}$ for shorthand.

Proposition 13.8 *Fix $\kappa \geq 16$, and $T \geq T_*^{(\kappa)}(\delta)$. Then*

$$||\hat{\mathcal{H}}_{0,\hat{d}(T,\delta),\hat{d}(T,\delta)} - \mathcal{H}_{0,\infty,\infty}||_2 \leq 12c\beta R\sqrt{\hat{d}(T,\delta)}\sqrt{\frac{m + p\hat{d}(T,\delta) + \log\frac{T}{\delta}}{T}}$$

with probability at least $1-\delta$.

Proof Assume that $\log\left(\frac{T}{\delta}\right) \leq d_*(T,\delta)$. Recall the following functions

$$\begin{aligned}d_*(T,\delta) &= \inf\left\{d \middle| c\beta R\sqrt{d}\sqrt{\frac{m + pd + \log\frac{T}{\delta}}{T}} \geq ||\mathcal{H}_d - \mathcal{H}_\infty||_2\right\} \\ d_0(T,\delta) &= \inf\left\{l \middle| ||\hat{\mathcal{H}}_l - \hat{\mathcal{H}}_h||_2 \leq c\beta R(\alpha(h) + 2\alpha(l)) \ \ \forall h \geq l, \ \ h \in \mathcal{D}(T)\right\} \\ \hat{d}(T,\delta) &= d_0(T,\delta) \vee \log\left(\frac{T}{\delta}\right)\end{aligned}$$

It is clear that $d_*(\kappa^2 T,\delta) \leq (1 + \frac{1}{2p})\kappa d_*(T,\delta)$ for any $\kappa \geq 16$. Assume the following

- $d_*(T,\delta) \leq \frac{\kappa}{8} d_*(\kappa^{-2}T,\delta)$ (This relation is true whenever $T \geq T_*^{(\kappa)}(\delta)$),

- $||\mathcal{H}_{\hat{d}(T,\delta)} - \mathcal{H}_\infty||_2 \geq 6c\beta R\sqrt{\hat{d}(T,\delta)}\sqrt{\frac{m+p\hat{d}(T,\delta)+\log\frac{T}{\delta}}{T}}$,
- $\hat{d}(T,\delta) < d_*(\kappa^{-2}T,\delta) - 1$.

The key will be to show that with high probability that all three assumptions can *not* hold with high probability. For shorthand we define $d_*^{(1)} = d_*(T,\delta), d_*^{(\kappa^2)} = d_*(\kappa^{-2}T,\delta), \hat{d}^{(1)} = \hat{d}(T,\delta), \hat{d}^{(\kappa^2)} = \hat{d}(\kappa^{-2}T,\delta)$ and $\mathcal{H}_l = \mathcal{H}_{0,l,l}, \hat{\mathcal{H}}_l = \hat{\mathcal{H}}_{0,l,l}$. Let $\tilde{T} = \kappa^{-2}T$. Then this implies that

$$\frac{c\beta R(\sqrt{d_*^{(1)}} + 2\sqrt{\hat{d}^{(1)}})}{\kappa}\sqrt{\frac{m+pd_*^{(1)}+\log\frac{\kappa^2\tilde{T}}{\delta}}{\tilde{T}}} \geq ||\hat{\mathcal{H}}_{\hat{d}^{(1)}} - \hat{\mathcal{H}}_{d_*^{(1)}}||_2$$

$$||\hat{\mathcal{H}}_{\hat{d}^{(1)}} - \hat{\mathcal{H}}_{d_*^{(1)}}||_2 \geq ||\hat{\mathcal{H}}_{\hat{d}^{(1)}} - \mathcal{H}_\infty||_2 - ||\hat{\mathcal{H}}_{d_*^{(1)}} - \mathcal{H}_\infty||_2$$

$$||\hat{\mathcal{H}}_{d_*^{(1)}} - \mathcal{H}_\infty||_2 + ||\hat{\mathcal{H}}_{\hat{d}^{(1)}} - \hat{\mathcal{H}}_{d_*^{(1)}}||_2 \geq ||\hat{\mathcal{H}}_{\hat{d}^{(1)}} - \mathcal{H}_\infty||_2$$

$$||\hat{\mathcal{H}}_{d_*^{(1)}} - \mathcal{H}_{d_*^{(1)}}||_2 + ||\mathcal{H}_{d_*^{(1)}} - \mathcal{H}_\infty||_2 + ||\hat{\mathcal{H}}_{\hat{d}^{(1)}} - \hat{\mathcal{H}}_{d_*^{(1)}}||_2 \geq ||\hat{\mathcal{H}}_{\hat{d}^{(1)}} - \mathcal{H}_\infty||_2$$

Since by definition of $d_*(\cdot,\cdot)$ we have

$$||\hat{\mathcal{H}}_{d_*^{(1)}} - \mathcal{H}_{d_*^{(1)}}||_2 + ||\mathcal{H}_{d_*^{(1)}} - \mathcal{H}_\infty||_2 \leq \frac{2c\beta R}{\kappa}\sqrt{d_*^{(1)}}\sqrt{\frac{m+pd_*^{(1)}+\log\frac{\kappa^2\tilde{T}}{\delta}}{\tilde{T}}}$$

and by assumptions $d_*^{(1)} \leq \frac{\kappa}{8}d_*^{(\kappa^2)}, \hat{d}^{(1)} \leq d_*^{(\kappa^2)}$ then as a result $(\sqrt{d_*^{(1)}} + 2\sqrt{\hat{d}^{(1)}})\sqrt{d_*^{(1)}} \leq (\frac{2\kappa}{8}+1)d_*^{(\kappa^2)}$

$$||\hat{\mathcal{H}}_{\hat{d}^{(1)}} - \mathcal{H}_\infty||_2 \leq ||\hat{\mathcal{H}}_{d_*^{(1)}} - \mathcal{H}_{d_*^{(1)}}||_2 + ||\mathcal{H}_{d_*^{(1)}} - \mathcal{H}_\infty||_2 + \underbrace{||\hat{\mathcal{H}}_{\hat{d}^{(1)}} - \hat{\mathcal{H}}_{d_*^{(1)}}||_2}_{\Downarrow}$$

$$\leq \underbrace{\frac{2c\beta R\sqrt{d_*^{(1)}}}{\kappa}\sqrt{\frac{m+pd_*^{(1)}+\log\frac{\kappa^2\tilde{T}}{\delta}}{\tilde{T}}}}_{\text{Prop. 13.6}} + \underbrace{\frac{c\beta R(\sqrt{d_*^{(1)}} + 2\sqrt{\hat{d}^{(1)}})}{\kappa}\sqrt{\frac{m+pd_*^{(1)}+\log\frac{\kappa^2\tilde{T}}{\delta}}{\tilde{T}}}}_{\text{Definition of } \hat{d}^{(1)}}$$

$$||\hat{\mathcal{H}}_{\hat{d}^{(1)}} - \mathcal{H}_\infty||_2 \leq \Big(\frac{1}{2} + \frac{1}{\kappa}\Big)c\beta R\sqrt{d_*^{(\kappa^2)}}\sqrt{\frac{m+pd_*^{(\kappa^2)}+\log\frac{\tilde{T}}{\delta}}{\tilde{T}}}$$

where the last inequality follows from $(\sqrt{d_*^{(1)}} + 2\sqrt{\hat{d}^{(1)}})\sqrt{d_*^{(1)}} \leq (\frac{2\kappa}{8}+1)d_*^{(\kappa^2)}$. Now by assumption

$$||\mathcal{H}_{\hat{d}^{(1)}} - \mathcal{H}_\infty||_2 \geq 6c\beta R\sqrt{\hat{d}^{(1)}}\sqrt{\frac{m+p\hat{d}^{(1)}+\log\frac{\kappa^2\tilde{T}}{\delta}}{\kappa^2\tilde{T}}}$$

it is clear that

$$||\hat{\mathcal{H}}_{\hat{d}^{(1)}} - \mathcal{H}_\infty||_2 \geq \frac{5}{6}||\mathcal{H}_{\hat{d}^{(1)}} - \mathcal{H}_\infty||_2$$

and we can conclude that, since $\frac{6}{5}\Big(\frac{1}{2}+\frac{1}{\kappa}\Big) < \frac{1}{\sqrt{2}}$,

$$||\mathcal{H}_{\hat{d}^{(1)}} - \mathcal{H}_\infty||_2 < c\beta R\sqrt{\frac{d_*^{(\kappa^2)}}{2}}\sqrt{\frac{m+pd_*^{(\kappa^2)}+\log\frac{\tilde{T}}{\delta}}{\tilde{T}}}$$

which implies that $\hat{d}^{(1)} \geq d_*^{(\kappa^2)} - 1$. This is because by definition of $d_*^{(\kappa^2)}$ we know that $d_*^{(\kappa^2)}$ is the minimum such that

$$||\mathcal{H}_{d_*^{(\kappa^2)}} - \mathcal{H}_\infty||_2 \leq c\beta R\sqrt{\frac{d_*^{(\kappa^2)}}{2}}\sqrt{\frac{m + pd_*^{(\kappa^2)} + \log\frac{\tilde{T}}{\delta}}{\tilde{T}}}$$

and furthermore from Proposition 13.2 we have for any $d_1 \leq d_2$

$$||\mathcal{H}_{0,\infty,\infty} - \mathcal{H}_{0,d_1,d_1}|| \geq \frac{1}{\sqrt{2}}||\mathcal{H}_{0,\infty,\infty} - \mathcal{H}_{0,d_2,d_2}||.$$

This contradicts Assumption 3. So, this means that one of three assumptions do not hold. Clearly if assumption 3 is invalid then we have a suitable lower bound on the chosen $\hat{d}(\cdot,\cdot)$, *i.e.*, since $d_*(\kappa^{-2}T,\delta) \leq d_*(T,\delta) \leq \frac{\kappa}{8}d_*(\kappa^{-2}T,\delta)$ we get

$$\frac{\kappa}{8}\hat{d}(\kappa^2\tilde{T},\delta) \geq \frac{\kappa}{8}d_*(\tilde{T},\delta) - \frac{\kappa}{8} \geq d_*(\kappa^2\tilde{T},\delta) - \frac{\kappa}{8} \geq \hat{d}(\kappa^2\tilde{T},\delta) - \frac{\kappa}{8} \geq d_*(\tilde{T},\delta) - \frac{\kappa}{8}$$

which implies from Lemma 13.1 that (since we pick $\kappa = 16$, for large enough T $d_*(\tilde{T},\delta) \geq 4$) and we have

$$\begin{aligned}||\hat{\mathcal{H}}_{\hat{d}(\kappa^2\tilde{T},\delta)} - \mathcal{H}_\infty||_2 &\leq 3c\beta R\sqrt{d_*(\kappa^2\tilde{T},\delta)}\sqrt{\frac{pd_*(\kappa^2\tilde{T},\delta) + \log\frac{\kappa^2\tilde{T}}{\delta}}{\kappa^2\tilde{T}}} \\ &\leq \frac{3\kappa}{8}c\beta R\sqrt{\hat{d}(\kappa^2\tilde{T},\delta)}\sqrt{\frac{p\hat{d}(\kappa^2\tilde{T},\delta) + \log\frac{\kappa^2\tilde{T}}{\delta}}{\kappa^2\tilde{T}}}\end{aligned}$$

Similarly, if assumption 2 is invalid then we get that

$$||\mathcal{H}_{\hat{d}(\kappa^2\tilde{T},\delta)} - \mathcal{H}_\infty||_2 < 6c\beta R\sqrt{\hat{d}(\kappa^2\tilde{T},\delta)}\sqrt{\frac{p\hat{d}(\kappa^2\tilde{T},\delta) + \log\frac{\kappa^2\tilde{T}}{\delta}}{\kappa^2\tilde{T}}}$$

and because $\hat{d}(\kappa^2\tilde{T},\delta) \leq d_*(\kappa^2\tilde{T},\delta)$ and $||\hat{\mathcal{H}}_{\hat{d}(\kappa^2\tilde{T},\delta)} - \mathcal{H}_\infty||_2 \leq ||\mathcal{H}_{\hat{d}(\kappa^2\tilde{T},\delta)} - \mathcal{H}_\infty||_2 + ||\hat{\mathcal{H}}_{\hat{d}(\kappa^2\tilde{T},\delta)} - \mathcal{H}_\infty||_2$ we get in a similar fashion to Proposition 13.6

$$||\hat{\mathcal{H}}_{\hat{d}(\kappa^2\tilde{T},\delta)} - \mathcal{H}_\infty||_2 \leq 12c\beta R\sqrt{\hat{d}(\kappa^2\tilde{T},\delta)}\sqrt{\frac{p\hat{d}(\kappa^2\tilde{T},\delta) + \log\frac{\kappa^2\tilde{T}}{\delta}}{\kappa^2\tilde{T}}}$$

Replacing $\kappa^2\tilde{T} = T$ it is clear that for any $\kappa \geq 16$

$$||\hat{\mathcal{H}}_{\hat{d}(T,\delta)} - \mathcal{H}_\infty||_2 \leq 12c\beta R\sqrt{\hat{d}(T,\delta)}\sqrt{\frac{p\hat{d}(T,\delta) + \log\frac{T}{\delta}}{T}} \tag{65}$$

If $d_*(T,\delta) \leq \log\left(\frac{T}{\delta}\right)$ then we can simply apply Lemma 13.1 and our assertion holds. ■

14. Model Selection Results

Proposition 14.1 *Let* $\mathcal{H}_{0,\infty,\infty} = U\Sigma V^\top, \hat{\mathcal{H}}_{0,\hat{d},\hat{d}} = \hat{U}\hat{\Sigma}\hat{V}^\top$ *and*

$$||\mathcal{H}_{0,\infty,\infty} - \hat{\mathcal{H}}_{0,\hat{d},\hat{d}}|| \leq \epsilon.$$

Let $\hat{\Sigma}$ *be arranged into blocks of singular values such that in each block* i *we have*

$$\sup_j \hat{\sigma}_j^i - \hat{\sigma}_{j+1}^i \leq \chi\epsilon$$

for some $\chi \geq 2$*, i.e.,*

$$\hat{\Sigma} = \begin{bmatrix} \Lambda_1 & 0 & \dots & 0 \\ 0 & \Lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \Lambda_l \end{bmatrix}$$

where Λ_i *are diagonal matrices and* $\hat{\sigma}_j^i$ *is the* j^{th} *singular value in the block* Λ_i*. Then there exists an orthogonal transformation,* Q*, such that*

$$\begin{aligned} ||\hat{U}\hat{\Sigma}^{1/2}Q - U\Sigma^{1/2}||_2 \leq 2\epsilon\sqrt{\hat{\sigma}_1/\zeta_{n_1}^2 + \hat{\sigma}_{n_1+1}/\zeta_{n_2}^2 + \ldots + \hat{\sigma}_{\sum_{i=1}^{l-1} n_i+1}/\zeta_{n_l}^2} \\ + 2\sup_{1\leq i\leq l} \sqrt{\hat{\sigma}_{\max}^i} - \sqrt{\hat{\sigma}_{\min}^i} + \frac{\epsilon}{\sqrt{\hat{\sigma}_{\hat{d}}}} \wedge \sqrt{\epsilon}. \end{aligned}$$

Here $\sup_{1\leq i\leq l} \sqrt{\hat{\sigma}_{\max}^i} - \sqrt{\hat{\sigma}_{\min}^i} \leq \frac{\chi}{\sqrt{\hat{\sigma}_{\max}^i}}\epsilon\hat{d} \wedge \sqrt{\chi\hat{d}\epsilon}$ *and*

$$\zeta_{n_i} = \min\left(\hat{\sigma}_{\min}^{n_{i-1}} - \hat{\sigma}_{\max}^{n_i}, \hat{\sigma}_{\min}^{n_i} - \hat{\sigma}_{\max}^{n_{i+1}}\right)$$

for $1 < i < l$*,* $\zeta_{n_1} = \hat{\sigma}_{\min}^{n_1} - \hat{\sigma}_{\max}^{n_2}$ *and* $\zeta_{n_l} = \min\left(\hat{\sigma}_{\min}^{n_{l-1}} - \hat{\sigma}_{\max}^{n_l}, \hat{\sigma}_{\min}^{n_l}\right)$*.*

Proof Let $\hat{U}\hat{\Sigma}\hat{V}^\top = \text{SVD}(\hat{\mathcal{H}}_{0,\hat{d},\hat{d}})$ and $U\Sigma V^\top = \text{SVD}(\mathcal{H}_{0,\infty,\infty})$ where $||\hat{\mathcal{H}}_{0,\hat{d},\hat{d}} - \mathcal{H}_{0,\infty,\infty}||_2 \leq \epsilon$. $\hat{\Sigma}$ is arranged into blocks of singular values such that in each block i we have $\hat{\sigma}_j^i - \hat{\sigma}_{j+1}^i \leq \chi\epsilon$, *i.e.*,

$$\hat{\Sigma} = \begin{bmatrix} \Lambda_1 & 0 & \dots & 0 \\ 0 & \Lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \Lambda_l \end{bmatrix}$$

where Λ_i are diagonal matrices and $\hat{\sigma}_j^i$ is the j^{th} singular value in the block Λ_i. Furthermore, $\hat{\sigma}_{\min}^{i-1} - \hat{\sigma}_{\max}^i > \chi\epsilon$. From $\hat{\Sigma}$ define $\bar{\hat{\Sigma}}$ as follows:

$$\bar{\hat{\Sigma}} = \begin{bmatrix} \bar{\hat{\sigma}}_1 I_{n_1\times n_1} & 0 & \dots & 0 \\ 0 & \bar{\hat{\sigma}}_2 I_{n_2\times n_2} & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \bar{\hat{\sigma}}_l I_{n_l\times n_l} \end{bmatrix} \tag{66}$$

where Λ_i is a $n_i \times n_i$ matrix and $\bar{\hat{\sigma}}_i = \frac{1}{n_i}\sum_j \hat{\sigma}_j^i$. The key idea of the proof is the following: $(A, B, C) \equiv (QAQ^\top, QB, CQ^\top)$ where Q is a orthogonal transformation and we will show that there exists a block diagonal unitary matrix Q of the form

$$Q = \begin{bmatrix} Q_{n_1\times n_1} & 0 & \dots & 0 \\ 0 & Q_{n_2\times n_2} & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & Q_{n_l\times n_l} \end{bmatrix} \tag{67}$$

such that each block $Q_{n_i\times n_i}$ corresponds to a orthogonal matrix of dimensions $n_i \times n_i$ and that $||\hat{U}\hat{\Sigma}^{1/2}Q - U\Sigma^{1/2}||_2$ is small if $||\hat{\mathcal{H}}_{0,\hat{d},\hat{d}} - \mathcal{H}_{0,\infty,\infty}||_2$ is small. Each of the blocks correspond to the set of singular values where the inter-singular value distance is "small". To start off, note that from Propositon 12.4 there must exist a Q that is block diagonal with orthogonal entries such that

$$||\hat{U}Q\hat{\Sigma}^{1/2} - U\Sigma^{1/2}||_2 \le c\epsilon\sqrt{\hat{\sigma}_1/\zeta_{n_1}^2 + \hat{\sigma}_{n_1+1}/\zeta_{n_2}^2 + \ldots + \hat{\sigma}_{\sum_{i=1}^{l-1} n_i+1}/\zeta_{n_l}^2} + \sup_{1\le i\le \hat{d}} |\sqrt{\sigma_i} - \sqrt{\hat{\sigma}_i}| \tag{68}$$

Here

$$\zeta_{n_i} = \min\left(\hat{\sigma}_{\min}^{n_{i-1}} - \hat{\sigma}_{\max}^{n_i}, \hat{\sigma}_{\min}^{n_i} - \hat{\sigma}_{\max}^{n_{i+1}}\right)$$

for $1 < i < l$, $\zeta_{n_1} = \hat{\sigma}_{\min}^{n_1} - \hat{\sigma}_{\max}^{n_2}$ and $\zeta_{n_l} = \min\left(\hat{\sigma}_{\min}^{n_{l-1}} - \hat{\sigma}_{\max}^{n_l}, \hat{\sigma}_{\min}^{n_l}\right)$. Informally, the ζ_i measure the singular value gaps between each blocks.

Furthermore, it can be shown that for any Q of the form in Eq. (67)

$$||\hat{U}Q\hat{\Sigma}^{1/2} - \hat{U}\hat{\Sigma}^{1/2}Q||_2 \le ||\hat{U}Q\bar{\hat{\Sigma}}^{1/2} - \hat{U}Q\hat{\Sigma}^{1/2}||_2 + ||\hat{U}\hat{\Sigma}^{1/2}Q - \hat{U}\bar{\hat{\Sigma}}^{1/2}Q||_2 \le 2||\hat{\Sigma}^{1/2} - \bar{\hat{\Sigma}}^{1/2}||_2$$

because $\hat{U}Q\bar{\hat{\Sigma}}^{1/2} = \hat{U}\bar{\hat{\Sigma}}^{1/2}Q$. Note that $||\hat{\Sigma}^{1/2} - \bar{\hat{\Sigma}}^{1/2}||_2 \le \sup_{1\le i\le l}\sqrt{\hat{\sigma}_{\max}^i} - \sqrt{\hat{\sigma}_{\min}^i}$. Now, when $\hat{\sigma}_{\max}^i \ge \chi n_i\epsilon$, then $\sqrt{\hat{\sigma}_{\max}^i} - \sqrt{\hat{\sigma}_{\min}^i} \le \frac{\chi\epsilon}{\sqrt{\hat{\sigma}_{\max}^i}}$; on the other hand when $\hat{\sigma}_{\max}^i < \chi n_i\epsilon$ then $\sqrt{\hat{\sigma}_{\max}^i} - \sqrt{\bar{\hat{\sigma}}^i} \le \sqrt{\chi n_i\epsilon}$ and this implies that

$$\sup_{1\le i\le l}\sqrt{\hat{\sigma}_{\max}^i} - \sqrt{\hat{\sigma}_{\min}^i} \le \frac{\chi n_i}{\sqrt{\hat{\sigma}_{\max}^i}}\epsilon \wedge \sqrt{\chi n_i\epsilon}.$$

Finally,

$$\begin{aligned} ||\hat{U}\hat{\Sigma}^{1/2}Q - U\Sigma^{1/2}||_2 &\le ||\hat{U}Q\hat{\Sigma}^{1/2} - U\Sigma^{1/2}||_2 + ||\hat{U}Q\hat{\Sigma}^{1/2} - \hat{U}\hat{\Sigma}^{1/2}Q||_2 \\ &= 2\epsilon\sqrt{\hat{\sigma}_1/\zeta_{n_1}^2 + \hat{\sigma}_{n_1+1}/\zeta_{n_2}^2 + \ldots + \hat{\sigma}_{\sum_{i=1}^{l-1} n_i+1}/\zeta_{n_l}^2} + \sup_{1\le i\le \hat{d}} |\sqrt{\sigma_i} - \sqrt{\hat{\sigma}_i}| \\ &\quad + \frac{\chi\epsilon}{\sqrt{\hat{\sigma}_{\max}^i}} \wedge \sqrt{\chi\epsilon}. \end{aligned}$$

Our assertion follows since $\sup_{1\le i\le \hat{d}}|\sqrt{\sigma_i} - \sqrt{\hat{\sigma}_i}| \le \frac{\epsilon}{\sqrt{\hat{\sigma}_{\hat{d}}}} \wedge \sqrt{\epsilon}$. ■

Proposition 14.2 *Let* $\mathcal{H}_{0,\infty,\infty} = U\Sigma V^\top, \hat{\mathcal{H}}_{0,\hat{d},\hat{d}} = \hat{U}\hat{\Sigma}\hat{V}^\top$ *and*

$$||\mathcal{H}_{0,\infty,\infty} - \hat{\mathcal{H}}_{0,\hat{d},\hat{d}}|| \leq \epsilon.$$

Let $\hat{\Sigma}$ *be arranged into blocks of singular values such that in each block* i *we have*

$$\sup_j \hat{\sigma}_j^i - \hat{\sigma}_{j+1}^i \leq \chi\epsilon$$

for some $\chi \geq 2$*, i.e.,*

$$\hat{\Sigma} = \begin{bmatrix} \Lambda_1 & 0 & \dots & 0 \\ 0 & \Lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \Lambda_l \end{bmatrix}$$

where Λ_i *are diagonal matrices and* $\hat{\sigma}_j^i$ *is the* j^{th} *singular value in the block* Λ_i*. Then there exists an orthogonal transformation,* Q*, such that*

$$\begin{aligned} \max\Big(||\hat{C} - C||_2, ||\hat{B} - B||_2\Big) &\leq 2\epsilon\sqrt{\hat{\sigma}_1/\zeta_{n_1}^2 + \hat{\sigma}_{n_1+1}/\zeta_{n_2}^2 + \ldots + \hat{\sigma}_{\sum_{i=1}^{l-1} n_i+1}/\zeta_{n_l}^2} \\ &\quad + 2\sup_{1\leq i\leq l} \sqrt{\hat{\sigma}_{\max}^i} - \sqrt{\hat{\sigma}_{\min}^i} + \frac{\epsilon}{\sqrt{\hat{\sigma}_{\hat{d}}}} \wedge \sqrt{\epsilon} = \zeta, \\ ||A - \hat{A}||_2 &\leq 4\gamma \cdot \zeta/\sqrt{\hat{\sigma}_{\hat{d}}}. \end{aligned}$$

Here $\sup_{1\leq i\leq l} \sqrt{\hat{\sigma}_{\max}^i} - \sqrt{\hat{\sigma}_{\min}^i} \leq \frac{\chi}{\sqrt{\hat{\sigma}_{\max}^i}}\epsilon\hat{d} \wedge \sqrt{\chi\hat{d}\epsilon}$ *and*

$$\zeta_{n_i} = \min\left(\hat{\sigma}_{\min}^{n_{i-1}} - \hat{\sigma}_{\max}^{n_i}, \hat{\sigma}_{\min}^{n_i} - \hat{\sigma}_{\max}^{n_{i+1}}\right)$$

for $1 < i < l$*,* $\zeta_{n_1} = \hat{\sigma}_{\min}^{n_1} - \hat{\sigma}_{\max}^{n_2}$ *and* $\zeta_{n_l} = \min\left(\hat{\sigma}_{\min}^{n_{l-1}} - \hat{\sigma}_{\max}^{n_l}, \hat{\sigma}_{\min}^{n_l}\right)$*.*

Proof The proof follows because all parameters are equivalent up to a orthogonal transform (See discussion preceding Proposition 9.2). Following that we use Propositions 12.4 and 12.5. ■

15. Order Estimation Lower Bound

Lemma 14 (Theorem 4.21 in Boucheron et al. (2013)) *Let* $\{\mathbb{P}_i\}_{i=0}^N$ *be probability laws over* $(\Sigma, \mathcal{A})$ *and let* $\{A_i \in \mathcal{A}\}_{i=0}^N$ *be disjoint events. If* $a = \min_{i=0,\ldots,N} \mathbb{P}_i(A_i) \geq 1/(N+1)$*,*

$$a \leq a\log\Big(\frac{Na}{1-a}\Big) + (1-a)\log\Big(\frac{1-a}{1-\frac{1-a}{N}}\Big) \leq \frac{1}{N}\sum_{i=1}^N KL(P_i||P_0) \tag{69}$$

Lemma 15 (Le Cam's Method) *Let* P_0, P_1 *be two probability laws then*

$$\sup_{\theta\in\{0,1\}} \mathbb{P}_\theta[M \neq \hat{M}] \geq \frac{1}{2} - \frac{1}{2}\sqrt{\frac{1}{2}KL(P_0||P_1)}$$

Proposition 15.1 *Let $\mathcal{N}_0, \mathcal{N}_1$ be two multivariate Gaussians with mean $\mu_0 \in \mathbb{R}^T, \mu_1 \in \mathbb{R}^T$ and covariance matrix $\Sigma_0 \in \mathbb{R}^{T\times T}, \Sigma_1 \in \mathbb{R}^{T\times T}$ respectively. Then the $KL(\mathcal{N}_0, \mathcal{N}_1) = \frac{1}{2}\Big(tr(\Sigma_1^{-1}\Sigma_0) - T + \log\frac{det(\Sigma_1)}{det(\Sigma_0)} + \mathbb{E}_{\mu_1,\mu_0}[(\mu_1-\mu_0)^\top\Sigma_1^{-1}(\mu_1-\mu_0)]\Big)$.*

In this section we will prove a lower bound on the finite time error for model approximation. In systems theory subspace based methods are useful in estimating the true system parameters. Intuitively, it should be harder to correctly estimate the subspace that corresponds to lower Hankel singular values, or "energy" due to the presence of noise. However, due to strong structural constraints on Hankel matrix finding a minimax lower bound is a much harder proposition for LTI systems. Specifically, it is not clear if standard subspace identification lower bounds can provide reasonable estimates for a structured and non i.i.d. setting such as our case. To alleviate some of the technical difficulties that arise in obtaining the lower bounds, we will focus on a small set of LTI systems which are simply parametrized by a number ζ. Consider the following canonical form order 1 and 2 LTI systems respectively with $m = p = 1$ and let R be the noise-to-signal ratio bound.

$$A_0 = \begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 0\\ \zeta & 0 & 0\end{bmatrix}, A_1 = A_0, B_0 = \begin{bmatrix}0\\ 0\\ \sqrt{\beta}/R\end{bmatrix}, B_1 = \begin{bmatrix}0\\ \sqrt{\beta}/R\\ \sqrt{\beta}/R\end{bmatrix}, C_0 = \begin{bmatrix}0 & 0 & \sqrt{\beta}R\end{bmatrix}, C_1 = C_0 \tag{70}$$

A_0, A_1 are Schur stable whenever $|\zeta| < 1$.

$$\mathcal{H}_{\zeta,0} = \beta\begin{bmatrix}1 & 0 & 0 & 0 & 0 & \dots\\ 0 & 0 & 0 & 0 & 0 & \dots\\ 0 & 0 & 0 & 0 & 0 & \dots\\ 0 & 0 & 0 & 0 & 0 & \dots\\ 0 & 0 & 0 & 0 & 0 & \dots\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\end{bmatrix}$$
$$\mathcal{H}_{\zeta,1} = \beta\begin{bmatrix}1 & 0 & \zeta & 0 & 0 & \dots\\ 0 & \zeta & 0 & 0 & 0 & \dots\\ \zeta & 0 & 0 & 0 & 0 & \dots\\ 0 & 0 & 0 & 0 & 0 & \dots\\ 0 & 0 & 0 & 0 & 0 & \dots\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\end{bmatrix} \tag{71}$$

Here $\mathcal{H}_{\zeta,0}, \mathcal{H}_{\zeta,1}$ are the Hankel matrices generated by $(C_0, A_0, B_0), (C_1, A_1, B_1)$ respectively. It is easy to check that for $\mathcal{H}_{\zeta,1}$ we have $\frac{1}{\zeta} \le \frac{\sigma_1}{\sigma_2} \le \frac{1+\zeta}{\zeta}$ where σ_i are Hankel singular values. Further the rank of $\mathcal{H}_{\zeta,0}$ is 1 and that of $\mathcal{H}_{\zeta,1}$ is at least 2. Also, $\frac{||\mathcal{T}\mathcal{O}_{0,\infty}((C_i,A_i,B_i))||_2}{||\mathcal{T}_{0,\infty}((C_i,A_i,B_i))||_2} \le R$.

This construction will be key to show that identification of a particular rank realization depends on the condition number of the Hankel matrix. An alternate representation of the

input–output behavior is

$$\begin{bmatrix} y_T \\ y_{T-1} \\ \vdots \\ y_1 \end{bmatrix} = \underbrace{\begin{bmatrix} CB & CA_iB & \dots & CA_i^{T-1}B \\ 0 & CB & \dots & CA_i^{T-2}B \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & CB \end{bmatrix}}_{\Pi_i} \underbrace{\begin{bmatrix} u_{T+1} \\ u_T \\ \vdots \\ u_2 \end{bmatrix}}_{U} + \underbrace{\begin{bmatrix} C & CA_i & \dots & CA_i^{T-1} \\ 0 & C & \dots & CA_i^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & C \end{bmatrix}}_{O_i} \begin{bmatrix} \eta_{T+1} \\ \eta_T \\ \vdots \\ \eta_2 \end{bmatrix} + \begin{bmatrix} w_T \\ w_{T-1} \\ \vdots \\ w_1 \end{bmatrix} \tag{72}$$

where $A_i \in \{A_0, A_1\}$. We will prove this result for a general class of inputs, *i.e.*, active inputs. Then we will follow the same steps as in proof of Theorem 2 in Tu et al. (2018b).

$$\begin{aligned} \mathrm{KL}(P_0||P_1) &= \mathbb{E}_{P_0}\left[\log \prod_{t=1}^{T} \frac{\gamma_t(u_t|\{u_l, y_l\}_{l=1}^{t-1})P_0(y_t|\{u_l\}_{l=1}^{t-1})}{\gamma_t(u_t|\{u_l, y_l\}_{l=1}^{t-1})P_1(y_t|\{u_l\}_{l=1}^{t-1})}\right] \\ &= \mathbb{E}_{P_0}\left[\log \prod_{t=1}^{T} \frac{P_0(y_t|\{u_l\}_{l=1}^{t-1})}{P_1(y_t|\{u_l\}_{l=1}^{t-1})}\right] \end{aligned}$$

Here $\gamma_t(\cdot|\cdot)$ is the active rule for choosing u_t from past data. From Eq. (72) it is clear that conditional on $\{u_l\}_{l=1}^T$, $\{y_l\}_{l=1}^T$ is Gaussian with mean given by $\Pi_i U$. Then we use Birge's inequality (Lemma 14). In our case $\Sigma_0 = O_0 O_0^\top + I, \Sigma_1 = O_1 O_1^\top + I$ where O_i is given in Eq. (72). We will apply a combination of Lemma 14, Proposition 15.1 and assume η_i are i.i.d Gaussian to obtain our desired result. Note that $O_1 = O_0$ but $\Pi_1 \neq \Pi_0$. Therefore, from Proposition 15.1 $KL(\mathcal{N}_0, \mathcal{N}_1) = \mathbb{E}_{\mu_1,\mu_0}[(\mu_1 - \mu_0)^\top \Sigma_1^{-1}(\mu_1 - \mu_0)] \leq T\frac{\zeta^2}{R^2}$ where $\mu_i = \Pi_i U$. For any $\delta \in (0, 1/4)$, set $a = 1 - \delta$ in Proposition 14, then we get whenever

$$\delta \log\left(\frac{\delta}{1-\delta}\right) + (1-\delta)\log\left(\frac{1-\delta}{\delta}\right) \geq \frac{T\zeta^2}{R^2} \tag{73}$$

we have $\sup_{i\neq j} \mathbb{P}_{A_i}(A_j) \geq \delta$. For $\delta \in [1/4, 1)$ we use Le Cam's method in Lemma 15 and show that if $8\delta^2 \geq \frac{T\zeta^2}{R^2}$ then $\sup_{i\neq j} \mathbb{P}_{A_i}(A_j) \geq \delta$. Since $\delta^2 \geq c \log\frac{1}{\delta}$ when $\delta \in [1/4, 1)$ for an absolute constant, our assertion holds.