

## MIT Open Access Articles

*Nonparametric System identification  
of Stochastic Switched Linear Systems*

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

**Citation:** Sarkar, Tuhin et al. "Nonparametric System identification of Stochastic Switched Linear Systems." Paper in the Proceedings of the IEEE Conference on Decision and Control, 2019-December, 2019 IEEE 58th Conference on Decision and Control (CDC), Nice, France, 11-13 Dec. 2019, IEEE © 2019 The Author(s)

**As Published:** 10.1109/CDC40024.2019.9030173

**Publisher:** Institute of Electrical and Electronics Engineers (IEEE)

**Persistent URL:** <https://hdl.handle.net/1721.1/129977>

**Version:** Original manuscript: author's manuscript prior to formal peer review

**Terms of use:** Creative Commons Attribution-Noncommercial-Share Alike



# Nonparametric System identification of Stochastic Switched Linear Systems

Tuhin Sarkar    Alexander Rakhlin    Munther Dahleh

**Abstract**—We address the problem of learning the parameters of a mean square stable switched linear systems(SLS) with unknown latent space dimension, or *order*, from its noisy input–output data. In particular, we focus on learning a good lower order approximation of the underlying model allowed by finite data. This is achieved by constructing Hankel-like matrices from data and obtaining suitable approximations via SVD truncation where the threshold for SVD truncation is purely data dependent. By exploiting tools from theory of model reduction for SLS, we find that the system parameter estimates are close to a balanced truncated realization of the underlying system with high probability.

## I. INTRODUCTION

Finite time system identification is an important problem in the context of control theory, times series analysis and robotics among many others. In this work, we focus on parameter estimation and model approximation of switched linear systems (SLS), which are described by

$$\begin{aligned} x_{k+1} &= A_{\theta_k} x_k + B u_k + \eta_{k+1} \\ y_k &= C x_k + w_k \end{aligned} \quad (1)$$

Here at time  $k$ ,  $x_k \in \mathbb{R}^n, y_k \in \mathbb{R}^p, u_k \in \mathbb{R}^m$  are the latent state, output and input respectively.  $\theta_k \in \{1, 2, \dots, s\}$  is the discrete state, mode or switch with  $\eta_k, w_k$  being the process and output noise respectively. We assume that  $\{\theta_k\}_{k=1}^{\infty}$  is an i.i.d process with  $\mathbb{P}(\theta_k = i) = p_i$ . The goal is to learn  $(C, \{p_i, A_i\}_{i=1}^s, B)$  from observed data  $\{y_k, u_k, \theta_k\}_{k=1}^N$  when the latent space dimension  $n$  is unknown. In many cases  $n > p, m$  and it becomes difficult to find suitable parametrizations that allow for provably efficient learning. For the special case of LTI systems, *i.e.*,  $s = 1$ , these issues were discussed in detail in [1]. It was suggested there that one can learn lower order approximations of the original system from finite noisy data. To motivate the study of such approximations, consider the following example:

**Example 1.** Let  $s = 2, p_i = 0.5, |\gamma| < 1$ . Consider  $M_1 =$

$\{C, A_1 \in \mathbb{R}^{n \times n}, B\}, M_2 = \{C, A_2 \in \mathbb{R}^{n \times n}, B\}$  given by

$$\begin{aligned} B &= \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, C = B^\top, A_1 = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \gamma \end{bmatrix} \\ A_2 &= \begin{bmatrix} 0 & 1 & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots \\ a & 0 & \dots & 0 \end{bmatrix} \end{aligned} \quad (2)$$

Assume that  $na \ll 1$ . This SLS is of order  $n$  which may be large. However, it can be suitably modeled by a lower dimensional SLS (“effective” order is  $\leq 2$  and can be checked by a simple computation of  $\{C A_i A_j B\}_{i,j=1}^2$ ).

The previous example suggests that in many cases the true order is not important; rather a lower order model exists that approximates the true system well. Furthermore, finite noisy data limits the complexity of models that can be effectively learned (See discussion in [2]). The existence of an “effective” lower order and finite data length motivate the question of finding “good” lower dimensional approximations of the underlying model from finite noisy data.

### A. Related Work

The study of switched linear systems has attracted a lot of attention [3], [4], [5] to name a few. These have been used in neuroscience to model neuron firing [6], modeling the stock index [7] and more generally approximate non-linear processes [8] with reasonable accuracy. The problem of realization, *i.e.*, whether there exists a SLS that satisfies the given data (in the noiseless case), has been studied in [9], [10], [11] and references therein. Specifically, [9] provides a purely algebraic view of realization where the switching is a function of discrete input symbols. The authors in [10] consider the case when discrete events are external inputs and there are linear reset maps that reset the state after switching. Finally, the theory of realization for generalized bilinear systems is studied in [11] and typically relies on the finite rank property of a certain Hankel-like matrix. Identification of a special class of SLS known as switched ARX systems has been widely studied [8], [12], [13], [14], [15], [16]. Under the assumption that an upper bound on the model order is known, an algebro-geometric approach to system identification is proposed under the assumption that  $\{\theta_k\}_{k=1}^{\infty}$  are not observed. The algorithms there typically involve clustering and as a result suffer exponential in order

TS,AR,MD are with Massachusetts Institute of Technology, Cambridge, MA 02139 (email: tsarkar, rakhlin, dahleh@mit.edu)

sample complexity [17]. From a system theory perspective, model approximation of SLS has been very well studied [18], [19], [20]. These methods mimic balanced truncation-like methods for model reduction and provide error guarantees between the original and reduced system. Despite substantial work on realization theory, identification and model reduction of SLS, there is little work on purely data driven approaches to model approximation. More recently, [1], [21] study data driven approaches to learning reduced order approximations of the original model. However, [21] does not assume any noise in the data generating process. This work is an extension of the work in [1] to the case of SLS.

## B. Contributions

In our work we study the case when  $\{y_k, u_k, \theta_k\}_{k=1}^N$  is observed and we would like to learn  $(C, \{A_i, p_i\}_{i=1}^s, B)$  from observed data. Such a case is relevant when the switches are exogenous but not a control input; for example traffic congestion (continuous state) as a function of weather conditions (discrete switches: snow, heavy rains etc.). The contributions of this paper can be summarized as follows:

- We extend the techniques introduced in [1] for SLS identification. Specifically, central to our approach is finding a system Hankel-like matrix for the SLS. We show that, similar to LTI systems, an appropriate SVD of the doubly infinite system Hankel matrix gives the individual system parameters (up to similarity transformation).
- Due to the presence of noisy finite data, we provide a  $p(\frac{s^N-1}{s-1}) \times m(\frac{s^N-1}{s-1})$  dimensional estimate of the doubly infinite system Hankel matrix. We show that if we let  $N$  grow carefully with the number of samples, we can obtain an accurate (with PAC guarantees) estimate of the system Hankel matrix.
- By leveraging tools from the theory of model order reduction of SLS, we provide an algorithm to obtain “good” lower order approximations of the original system directly from data. To this end, we also provide a model order selection rule to choose the best approximation of the underlying SLS than can be learned from data with high probability. The model selection rule essentially involves a hard singular value thresholding and can be shown to be minimax optimal.

## II. SYSTEM MODEL AND ALGORITHM

Recall the SLS dynamics in Eq. (1). Denote by  $\mathbf{I}_i^j = \{\theta_j, \theta_{j-1}, \dots, \theta_i\} \in [s]^{j-1+1}$  an arbitrary sequence of switches from  $i$  to  $j$  and  $A_{\mathbf{I}_i^j} = A_{\theta_j} A_{\theta_{j-1}} \dots A_{\theta_i}$ . For two switch sequences  $\{\theta_2, \theta_1\}, \{\phi_2, \phi_1\}$  define a concatenation operator ‘:’ as  $\{\theta_2, \theta_1\} : \{\phi_2, \phi_1\} = \{\theta_2, \theta_1, \phi_2, \phi_1\}$ . Then  $\mathbf{I}_1^i : \mathbf{I}_1^j$  is concatenation of  $\mathbf{I}_1^i, \mathbf{I}_1^j$ . We state our assumptions below

- We have

$$\sup_{N \geq 0} \left\{ \sum_{\mathbf{I}_1^N \in [s]^N} \|CA_{\mathbf{I}_1^N} B\|_F^2, \sum_{\mathbf{I}_1^N \in [s]^N} \|CA_{\mathbf{I}_1^N}\|_F^2 \right\} \leq \beta^2$$

- $\{\eta_t, w_t\}_{t=1}^\infty$  are i.i.d zero mean subGaussian noise process with subGaussian norm 1 (see Def. 2.5.6 in [22]).
- SLS is mean-square stable or equivalently  $\sum_{i=1}^s p_i A_i \otimes A_i$  is Schur stable. (Theorem 2.1 in [18])
- There exist  $X_1, X_2 \succ 0$  such that  $\sum_{i=1}^s p_i A_i X_1 A_i^\top + B B^\top = X_1$  and  $\sum_{i=1}^s p_i A_i^\top X_2 A_i + C^\top C = X_2$ .
- Let  $\sigma_i = \lambda_i(X_1 X_2)$  and assume that  $\sigma_{i+1} < \sigma_i$ . Furthermore, let  $\tau_+ = \inf_{1 \leq i \leq n-1} (1 - \frac{\sigma_{i+1}}{\sigma_i})$ .

The third assumption ensures minimality, *i.e.*, controllability and observability, of the data generating SLS (See [11]). The goal is to identify  $\{C, \{p_i, A_i\}_{i=1}^s, B\}$  from data  $\{y_k, u_k, \theta_k\}_{k=1}^\infty$  when  $n$  (or its upper bound) is unknown. The final assumption mimics the distinct Hankel singular value assumption for LTI systems. For simplicity we call  $p_{\max} = \sup_{1 \leq i \leq s} p_i$ .

It is clear that for any sequence of observed switches  $\mathbf{I}_1^N$ , we have the corresponding output  $y_N$  as

$$y_N = \sum_{j=2}^{N-1} C A_{\theta_{N-1}} A_{\theta_{N-2}} \dots A_{\theta_j} B u_{j-1} + C B u_{N-1} + \sum_{j=2}^{N-1} C A_{\theta_{N-1}} A_{\theta_{N-2}} \dots A_{\theta_j} \eta_{j-1} + C \eta_{N-1} + w_N \quad (3)$$

Finally a measure of distance between two switched linear systems with probabilistic switches is the stochastic  $L_2$  gain given by

**Definition 1** (Definition 2.2 in [18]). *Let the noise  $\{\eta_k, w_k\}_{k=1}^\infty = 0$ . Let  $\theta = (\theta_1, \theta_2, \dots) \in [s]^\infty, u = (u_1, u_2, \dots) \in \mathbb{R}^\infty$  and  $y_M^{(\theta, u)}, y_{M_r}^{(\theta, u)} \in \mathbb{R}^\infty$  be the output sequence, in response to input  $u$  and switch sequence  $\theta$ , of system  $M$  and  $M_r$  respectively. Then the stochastic  $L_2$  distance between  $M$  and  $M_r$  denoted by  $\Delta_{M, M_r}$  is*

$$\Delta_{M, M_r}^2 = \sup_{\|u\|_2 \leq 1} \mathbb{E}_\theta [\|y_M^{(\theta, u)} - y_{M_r}^{(\theta, u)}\|_2^2]$$

The first question we pose is if there exists a Hankel matrix based representation for SLS as in the case of LTI systems that captures important properties about the system. In particular, whether it is possible to find the system parameters from input–output data in the ideal case of infinite noiseless data. We will now construct a system Hankel-like matrix that indeed answers this question positively. First, we will arrange  $\{\mathbf{I}_1^i \in [s]^i\}_{i=0}^N$  in a lexicographic order. This can be done for example as in [23]. To summarize, every sequence  $\mathbf{I}_1^N$  has a unique index  $L(\mathbf{I}_1^N) = \theta_N s^{N-1} + \dots + \theta_1$  with  $N = 0 \implies L(\mathbf{I}_1^N) = 0$ . Then the  $p(\frac{s^{N+1}-1}{s-1}) \times m(\frac{s^{N+1}-1}{s-1})$  Hankel-like matrix,

$$[\mathcal{H}^{(N)}]_{pL(\mathbf{I}_1^i)+1:pL(\mathbf{I}_1^i)+p, mL(\mathbf{I}_1^j)+1:mL(\mathbf{I}_1^j)+m} = \sqrt{p_{\mathbf{I}_1^i: \mathbf{I}_1^j}} C A_{\mathbf{I}_1^i: \mathbf{I}_1^j} B \quad (4)$$

$\forall 0 \leq i, j \leq N-1$ . Define  $\mathcal{H}_k^{(N)}$  as

$$[\mathcal{H}_k^{(N)}]_{pL(\mathbf{I}_1^i)+1:pL(\mathbf{I}_1^i)+p, mL(\mathbf{I}_1^j)+1:mL(\mathbf{I}_1^j)+m} = [\mathcal{H}^{(N)}]_{pL(\mathbf{I}_1^i: \{k\})+1:pL(\mathbf{I}_1^i: \{k\})+p, mL(\mathbf{I}_1^j)+1:mL(\mathbf{I}_1^j)+m} \quad (5)$$

Note that if  $s \rightarrow 1$ , i.e., LTI system, then  $\mathcal{H}^{(N)}$  becomes  $p(N+1) \times m(N+1)$  matrix and becomes the standard Hankel matrix for LTI systems. Let  $\mathcal{H}^{(\infty)} = \lim_{N \rightarrow \infty} \mathcal{H}^{(N)}$ , i.e., its doubly infinite extension. To give some intuition we present an example below

**Example 2.** Let  $s = 2$ . Then  $L(\phi) = 0, L(\{1\}) = 1, L(\{2\}) = 2, L(\{1, 1\}) = 3, L(\{1, 2\}) = 4, \dots$  As a result

$$\underbrace{\begin{bmatrix} CB & \sqrt{p_1}CA_1B & \sqrt{p_2}CA_2B & \dots \\ \sqrt{p_1}CA_1B & \sqrt{p_1^2}CA_1^2B & \sqrt{p_1p_2}CA_1A_2B & \dots \\ \sqrt{p_2}CA_2B & \sqrt{p_1p_2}CA_2A_1B & \sqrt{p_2^2}CA_2^2B & \dots \\ \sqrt{p_1^2}CA_1^2B & \sqrt{p_1^3}CA_1^3B & \sqrt{p_1^2p_2}CA_1^2A_2B & \dots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}}_{=\mathcal{H}^{(\infty)}} \quad (6)$$

$$\underbrace{\sqrt{p_k} \begin{bmatrix} CA_kB & \sqrt{p_1}CA_kA_1B & \dots \\ \sqrt{p_1}CA_1A_kB & \sqrt{p_1^2}CA_1A_kA_1B & \dots \\ \sqrt{p_2}CA_2A_kB & \sqrt{p_1p_2}CA_2A_kA_1B & \dots \\ \sqrt{p_1^2}CA_1^2A_kB & \sqrt{p_1^3}CA_1^2A_kA_1B & \dots \\ \vdots & \vdots & \vdots \end{bmatrix}}_{=\mathcal{H}_k^{(\infty)}} \quad (7)$$

**Proposition 1.**  $\mathcal{H}^{(\infty)}$  is a well defined operator with  $\text{rank}(\mathcal{H}^{(\infty)}) = n$ . Let  $\mathcal{H}^{(\infty)} = U\Sigma V^\top$  and  $\mathcal{H}_k^{(\infty)}$  be as Eq. (5). Then  $[U\Sigma^{1/2}]_{1:p,:} = C, [\Sigma^{1/2}V^\top]_{:,1:m} = B$

$$\sqrt{p_k}A_k = \Sigma^{-1/2}U^\top \mathcal{H}_k^{(\infty)} V \Sigma^{-1/2}$$

for every  $1 \leq k \leq s$ .

*Proof.* Note that

$$\mathcal{H}^{(\infty)} = \underbrace{\begin{bmatrix} C \\ \sqrt{p_1}CA_1 \\ \vdots \\ \sqrt{p_{i_1}}CA_{i_1} \\ \vdots \end{bmatrix}}_{=\mathcal{O}} \underbrace{[B, \sqrt{p_1}A_1B, \dots, \sqrt{p_{i_1}}A_{i_1}B, \dots]}_{=\mathcal{R}} \quad (8)$$

Then  $\mathcal{O}^\top \mathcal{O} = X_2, \mathcal{R}\mathcal{R}^\top = X_1$  and the result follows by Sylvester rank inequality. Now,  $\mathcal{H}_k^{(\infty)}$  is such that we only choose the block matrices in  $\mathcal{O}$  that end in  $A_k$ , i.e.,  $\mathcal{H}_k^{(\infty)} = \tilde{\mathcal{O}}\mathcal{R}$  where each of the submatrices in  $\tilde{\mathcal{O}}$  end in  $A_k$ . Since the occurrence of a switch is independent we get the desiderata by noting that  $\tilde{\mathcal{O}} = \sqrt{p_k}\mathcal{O}A_k$ .  $\square$   $\square$

Proposition 1 indicates that  $\mathcal{H}^{(\infty)}$  plays the role of traditional Hankel matrix in LTI systems theory for SLS. Similar subspace based methods for system identification has been discovered in mildly different forms for HMM parameter recovery in [23], [24] or weighted automaton parameter identification in [25].

Unfortunately, we do not have access to  $\mathcal{H}^{(\infty)}$ ; rather we only possess finite noisy data and consequently need

to obtain an accurate estimate  $\hat{\mathcal{H}}^{(N)}$  of  $\mathcal{H}^{(\infty)}$ . In order to find an estimate for the system Hankel matrix we assume that the switched linear system can be restarted multiple times. Although we believe that it is possible to relax this requirement, we enforce this assumption to ease exposition. Define the number of restarts as  $N_S$ , also referred as the sample complexity. In each restart, we let the SLS run for  $N$  time steps, also known as rollout length. Let  $\theta_k^{(t)}, y_k^{(t)}, u_k^{(t)}$  denote the switch, output and input respectively at rollout time  $k$  for sample  $t$ . Clearly  $t \leq N_S, k \leq N$ . Now define the set  $\mathcal{N}_{m_i}$  as

$$\mathcal{N}_{m_i} = \{(t, k) | (\theta_{k+i-1}^{(t)}, \theta_{k+i-2}^{(t)}, \dots, \theta_k^{(t)}) = m_i \in [s]^i\} \quad (9)$$

$\mathcal{N}_{m_i}$  is the set of occurrences of the switch sequence  $m_i$  with  $N_{m_i} = |\mathcal{N}_{m_i}|$ . Our next result bounds the error rates obtained from the regression. The proof of the following result follows standard analysis in statistical learning literature such as [26].

**Proposition 2.** Fix  $\delta > 0$  and sequence  $\mathbf{l}_1^i \in [s]^i$ . Let  $\hat{\Theta}_i$  be the following solution

$$\hat{\Theta}_{\mathbf{l}_1^i} = \arg \inf_{\Theta} \sum_{(t,k) \in \mathcal{N}_{\mathbf{l}_1^i}} \|y_k^{(t)} - \Theta u_k^{(t)}\|_F^2$$

where  $\{u_k^{(t)}\}_{t,k=1}^\infty$  are i.i.d isotropic Gaussian (or subGaussian) random variables. Then whenever  $N_{\mathbf{l}_1^i} \geq \alpha(m + \log \frac{1}{\delta})$  we have with probability at least  $1 - \delta$  that

$$\|CA_{\mathbf{l}_1^i}B - \hat{\Theta}_{\mathbf{l}_1^i}\|_F \leq \alpha\beta \sqrt{\frac{m + \log \frac{1}{\delta}}{N_{\mathbf{l}_1^i}}} \quad (10)$$

An important thing to note about the bound above is that it does not hold when  $N_{\mathbf{l}_1^i} < \alpha(m + \log \frac{1}{\delta})$  we set  $\hat{\Theta}_{\mathbf{l}_1^i} = 0$ , i.e., when we have scarce data for a certain sequence we can not use the regression estimate as it becomes unreliable. In such cases (and some others) we set  $\hat{\Theta}_{\mathbf{l}_1^i} = 0$ ; the exact details are specified below.

#### A. Regression Estimates

Recall Proposition 2, for any sequence  $\mathbf{l}_1^i$  of length  $i$  the result holds with probability at least  $1 - \delta$  only if we have  $N_{\mathbf{l}_1^i} \geq \alpha(m + \log \frac{2}{\delta})$ . The regression estimate for  $\mathbf{l}_1^i$  is unreliable when we do not have enough occurrences. In such a case we propose a simple estimate, i.e., we set the regression estimate to 0. Let us assume we have roll out length of  $\hat{N}$ , then we need to ensure that for all sequences of length at most  $\hat{N}$  the regression estimates hold; in that case by applying a union bound we have that for all sequences simultaneously we have with probability at

least  $1 - \delta$  that  $\|CA_{\mathbf{l}_1^j}B - \hat{\Theta}_{\mathbf{l}_1^j}\|_2 \leq \alpha\beta_{\mathbf{l}_1^j} \sqrt{\frac{m + \log \frac{s^{\hat{N}+1}-1}{(s-1)\delta}}{N_{\mathbf{l}_1^j}}}$  if

$N_{\mathbf{l}_1^j} \geq \alpha(m + \log \frac{2(s^{\hat{N}+1}-1)}{(s-1)\delta})$ . The  $\frac{s^{\hat{N}+1}-1}{s-1}$  appears because we are taking a union bound over  $\frac{s^{\hat{N}+1}-1}{s-1}$  sequences. One observation is that we cannot ensure the high probability bound simultaneously over all sequences up to length  $N_S$  because if  $\hat{N} = \Theta(N_S)$  then the regression estimate error

bound becomes trivial. As a result, we define  $N_{up}$  an upper bound for rollout length up to which we can ensure high probability bound. Define a sequence length dependent threshold  $\gamma_k = \alpha(m + \log \frac{2(s^{k+1}-1)}{(s-1)\delta})$  then

$$N_{up} = \inf \{N | N_{\mathbf{I}_1^{2N}} < \gamma_{2N} \ \forall \mathbf{I}_1^{2N} \in [s]^{2N}\} \quad (11)$$

Intuitively,  $2N_{up}$  is the least sequence length such that none of the sequences of that length can be reliably learned by regression, *i.e.*, all sequences with length up to  $N_{up}$  occur often enough. Furthermore, since the probability decays as the length of the sequence it suggests that no longer sequence can be learned reliably either. We show in Proposition 9 that  $N_{up}$  is logarithmic in  $N_S$  with high probability. With this we can construct an estimate of the system Hankel-like matrix as follows. Let  $\hat{N}$  be the rollout length then define

$$\hat{p}_{\mathbf{I}_1^i : \mathbf{I}_1^j} = \begin{cases} \frac{N_{\mathbf{I}_1^i : \mathbf{I}_1^j}}{N_S(\hat{N}-i-j+1)}, & \text{if } i+j > 0 \\ 1, & \text{otherwise} \end{cases}$$

$\hat{p}_{\mathbf{I}_1^i : \mathbf{I}_1^j}$  is an unbiased estimator for  $p_{\mathbf{I}_1^i : \mathbf{I}_1^j}$ . To see this, recall the experiment set up: we run  $N_S$  identical samples of the SLS for length  $\hat{N}$ . Then for each sample  $i \leq N_S$ , any sequence  $\mathbf{I}_1^k$  can start at position  $1, 2, \dots, \hat{N} - k + 1$ . Thus for sample  $i$  the number of occurrences of  $\mathbf{I}_1^k$  is given by  $\sum_{l=1}^{\hat{N}-k+1} \mathbf{1}_{\{\mathbf{I}_1^k \text{ starts at position } l\}}^{(i)}$ , then  $N_{\mathbf{I}_1^k}$  is given by

$$N_{\mathbf{I}_1^k} = \sum_{i=1}^{N_S} \sum_{l=1}^{\hat{N}-k+1} \mathbf{1}_{\{\mathbf{I}_1^k \text{ starts at position } l\}}^{(i)} \quad (12)$$

and it is clear that  $\mathbb{E}[N_{\mathbf{I}_1^k}] = p_{\mathbf{I}_1^k} N_S (\hat{N} - k + 1)$ . For the estimates of  $CA_{\mathbf{I}_1^i}B$  we have

$$\hat{\Theta}_{\mathbf{I}_1^i} = \begin{cases} \text{Regression estimate of Prop. 2,} & \text{if } N_{\mathbf{I}_1^i} \geq \gamma_{2N_{up}} \\ 0, & \text{otherwise} \end{cases}$$

$$[\hat{\mathcal{H}}^{(N)}]_{pL(\mathbf{I}_1^i)+1:pL(\mathbf{I}_1^i)+p, mL(\mathbf{I}_1^j)+1:mL(\mathbf{I}_1^j)+m} = \sqrt{\hat{p}_{\mathbf{I}_1^i : \mathbf{I}_1^j}} \hat{\Theta}_{\mathbf{I}_1^i : \mathbf{I}_1^j} \quad (13)$$

The road map for system identification can be summarized as follows.

- For a given  $N_S$  we do model order selection by choosing two numbers  $\hat{N}, r$  which are functions of  $N_S$ .
- Following that we create a finite dimensional estimate  $\hat{\mathcal{H}}^{\hat{N}}$  of  $\mathcal{H}^{(\infty)}$ , and from  $\hat{\mathcal{H}}^{\hat{N}}$  we obtain the system parameters of  $r$ -dimensional approximation of the original SLS using a balanced truncation procedure.
- The error between the estimated  $r$ -dimensional approximation and the true  $r$ -dimensional approximation can be bounded by subspace perturbation bounds [1].

We now describe details of the balanced truncation below.

### B. Balanced Truncation

Given the parameters of SLS in Eq. (1) define the following SLS

$$\begin{aligned} \tilde{x}(t+1) &= \sqrt{p_{\theta_t}} A_{\theta_t} \tilde{x}(t) + \sqrt{p_{\theta_t}} B u(t) \\ \tilde{y}_t &= \sqrt{p_{\theta_t}} C \tilde{x}(t) \end{aligned} \quad (14)$$

By our assumption the SLS in Eq. (14) is strongly stable (See [20]). Now we can use the results in [19] (specifically Eq. (25a), (25b)). To summarize there exists a linear transformation  $S$  such that

$$S^\top X_1 S = S^{-1} X_2 S^{-\top} = \Sigma \quad (15)$$

where  $\Sigma$  is diagonal with entries arranged in descending order, then note that  $\Sigma$  satisfies (by definition of  $X_1, X_2$ )

$$p_i \tilde{A}_i^\top \Sigma \tilde{A}_i - \Sigma + p_i \tilde{C}^\top \tilde{C} \preceq 0, p_i \tilde{A}_i \Sigma \tilde{A}_i^\top - \Sigma + p_i \tilde{B} \tilde{B}^\top \preceq 0$$

where  $\tilde{A}_i = S^\top A_i S^{-\top}, \tilde{C} = C S^{-\top}, \tilde{B} = S^\top B$ . Partition  $\tilde{A}_i, \tilde{B}, \tilde{C}$  as follows

$$\tilde{A}_i = \begin{bmatrix} A_i^{(r)} & \tilde{A}_i^{(12)} \\ \tilde{A}_i^{(21)} & \tilde{A}_i^{(22)} \end{bmatrix}, \tilde{B}_i = \begin{bmatrix} B^{(r)} \\ \tilde{B}^{(2)} \end{bmatrix}, \tilde{C} = [C^{(r)}, \tilde{C}^{(2)}] \quad (16)$$

where  $A_i^{(r)} \in \mathbb{R}^{r \times r}, B^{(r)} \in \mathbb{R}^{r \times 1}, C^{(r)} \in \mathbb{R}^{1 \times r}$  and are the  $r$ -order balanced truncated version of the true SLS. Then the discussion in Section 4.2 in [19] provides error guarantees between the true model and its approximation. Note that in the setting of Section 4.2 in [19]  $\alpha_i(k) = \sqrt{p_i} \mathbf{1}_{\{\theta_k=i\}}$ . This observation combined with some linear algebra similar to Section 21.6 of [27] gives us the following proposition.

**Proposition 3.** Assume  $\{\eta_t, w_t\}_{t=1}^\infty = 0$ . Denote by  $M$  the SLS in Eq. (1) and the reduced model corresponding  $(C^{(r)}, \{A_i^{(r)}\}_{i=1}^s, B^{(r)})$  described in Eq. (16) as  $M_r$ .  $M_r$  is mean square stable and furthermore

$$\Delta_{M, M_r} \leq 2s \sum_{l=r+1}^n \sigma_l$$

where  $\sigma_l$  are the singular values of the Gramian in Eq. (15).

Next, we show how to obtain the reduced order models  $(C^{(r)}, \{A_i^{(r)}\}_{i=1}^s, B^{(r)})$  directly from system Hankel matrix  $\mathcal{H}^{(\infty)}$ . Recall the SVD from Proposition 1,  $\mathcal{H}^{(\infty)} = U \Sigma V^\top$  where

$$\begin{aligned} U \Sigma^{1/2} &= \begin{bmatrix} \tilde{C} \\ \sqrt{p_1} \tilde{C} \tilde{A}_1 \\ \vdots \\ \sqrt{p_{\mathbf{I}_1^i}} \tilde{C} \tilde{A}_{\mathbf{I}_1^i} \\ \vdots \end{bmatrix} \\ &= \mathcal{O} \\ \Sigma^{1/2} V^\top &= \underbrace{[\tilde{B}, \sqrt{p_1} \tilde{A}_1 \tilde{B}, \dots, \sqrt{p_{\mathbf{I}_1^i}} \tilde{A}_{\mathbf{I}_1^i} \tilde{B}, \dots]}_{=\mathcal{R}} \end{aligned} \quad (17)$$

with  $\mathcal{O}^\top \mathcal{O} = \mathcal{R} \mathcal{R}^\top = \Sigma$ .

**Proposition 4.** Fix  $r \leq n$  and  $\mathcal{H}^{(\infty)} = U \Sigma V^\top$ . Then the  $r$ -order balanced truncated model  $(C^{(r)}, \{A_i^{(r)}\}_{i=1}^s, B^{(r)})$  is given by

$$C^{(r)} = [U \Sigma^{1/2}]_{1:p, 1:r}, B^{(r)} = [\Sigma^{1/2} V^\top]_{1:r, 1:m}$$

and

$$\sqrt{p_i} A_i^{(r)} = \Sigma_r^{-1/2} U_r^\top \mathcal{H}_i^{(\infty)} V_r \Sigma_r^{-1/2}$$

where  $U_r, V_r, \Sigma_r$  correspond to top  $r$  left singular vectors, right singular vectors and singular values respectively.

Proposition 4 makes it clear that to find  $r$ -order balanced truncated models we only need top  $r$ -singular vectors (and singular values). This observation is important because in the presence of finite noisy data estimating singular vectors corresponding to very low singular values typically requires a lot of data. Instead one could focus on simply estimating the significant singular vectors via balanced truncation. Furthermore, the stochastic  $L_2$  distance between the original SLS and its  $r$ -order balanced truncated version is given by Proposition 3. We can now summarize our algorithm below.

---

### Algorithm 1 Learning SLS

---

**Input**  $m$ : Input dimension,  $p$ : Output dimension

$N_S$ : Sample complexity,  $\hat{N}$ : Rollout length,  $r$ : order

**Output** System Parameters  $(\tilde{C}, \{\tilde{A}_i, \hat{p}_i\}_{i=1}^s, \tilde{B})$

---

- 1: **for**  $j = 1$  to  $N_S$  **do**
  - 2:   Sample  $\{u_i^{(j)} \sim \mathcal{N}(0, I_{m \times m})\}_{i=1}^{\hat{N}}$
  - 3:   Collect output  $(y_1^{(j)}, \dots, y_{\hat{N}}^{(j)})$  in response to input  $(u_1^{(j)}, \dots, u_{\hat{N}}^{(j)})$
  - 4:   Collect switch sequence  $(\theta_1^{(j)}, \dots, \theta_{\hat{N}}^{(j)})$
  - 5: Estimate  $\hat{\mathcal{H}}^{(\hat{N})}$  by regression as in Eq. (13).
  - 6:  $\hat{\mathcal{H}}^{(\hat{N})} = \hat{U} \hat{\Sigma} \hat{V}^\top$  and estimate  $\hat{p}_i = \frac{\sum_{l,j} \mathbf{1}_{\{\theta_l^{(j)} = i\}}}{N_S \hat{N}}$
  - 7:
  - 8:  $\tilde{C} = [\hat{U}_r \hat{\Sigma}_r^{1/2}]_{1:p, 1:r}$ ,  $\tilde{B} = [\hat{\Sigma}_r^{1/2} \hat{V}_r^\top]_{1:r, 1:m}$
  - 9:  $\tilde{A}_i = \hat{p}_i^{-1/2} \hat{\Sigma}_r^{-1/2} \hat{U}_r^\top \hat{\mathcal{H}}_i^{(\hat{N})} \hat{V}_r \hat{\Sigma}_r^{-1/2}$
  - 10: **Return:**  $(\tilde{C}, \{\tilde{A}_i, \hat{p}_i\}_{i=1}^s, \tilde{B})$
- 

### III. MODEL SELECTION

Algorithm 1 has two hyperparameters  $\hat{N}, r$ . In this section we discuss how to choose these hyperparameters as a function of  $N_S$ .

#### A. Selecting $\hat{N}$

Since the Hankel matrix is  $p \binom{s^{\hat{N}+1}-1}{s-1} \times m \binom{s^{\hat{N}+1}-1}{s-1}$ ,  $\hat{N}$  cannot be too large as it will make any algorithm infeasible (and estimation error will suffer) and indeed it cannot be too small as that will mean we only learn a small part of the dynamics (high truncation error). The key idea is to grow  $\hat{N}$  in a controlled fashion with respect to  $N_S$ . Formally, let  $\bar{\mathcal{H}}^{(\hat{N})}$  be  $\mathcal{H}^{(\hat{N})}$  padded with zeros to make it doubly infinite and define

$$T_{\hat{N}}^2 = \underbrace{\|\bar{\mathcal{H}}^{(\hat{N})} - \mathcal{H}^{(\infty)}\|_F^2}_{\text{Truncation Error}}, E_{\hat{N}}^2 = \underbrace{\|\mathcal{H}^{(\hat{N})} - \hat{\mathcal{H}}^{(\hat{N})}\|_F^2}_{\text{Estimation Error}} \quad (18)$$

Observe that the Frobenius norm of the difference  $\bar{\mathcal{H}}^{(\hat{N})} - \mathcal{H}^{(\infty)}$  can be represented as  $\|\bar{\mathcal{H}}^{(\hat{N})} - \mathcal{H}^{(\infty)}\|_F^2 \leq \sum_{i+j \geq 2\hat{N}-1} p_{\mathbf{i}_i, \mathbf{i}_j} \|CA_{\mathbf{i}_i, \mathbf{i}_j} B\|_F^2$ . Clearly as  $\hat{N}$  increases the truncation error decreases. For the case of estimation error we can use Proposition 2. Intuitively, it would make sense that  $E_{\hat{N}}$  grows with  $\hat{N}$  (keeping  $N_S$  fixed) as we are trying

to estimate a larger matrix. As a result, for large enough  $N_S$ , there exists  $\hat{N} < \infty$  such that  $T_{\hat{N}} \leq \alpha E_{\hat{N}}$  for some absolute constant  $\alpha \geq 1$ . The key idea will be to choose  $\hat{N}$  such that  $T_{\hat{N}} \leq \alpha E_{\hat{N}}$ . This idea is formalized in Proposition 10. Furthermore, for such a choice of  $\hat{N}$  we have  $\|\mathcal{H}^{(\infty)} - \hat{\mathcal{H}}^{(\hat{N})}\|_F^2 \leq (1 + \alpha^2) T_{\hat{N}}^2$  implying that we can estimate the system Hankel matrix well if  $T_{\hat{N}}$  is low.

**Proposition 5.** Fix  $\hat{N}, N_S$  and  $\delta$ . Then with probability at least  $1 - \delta$  we have

$$E_{\hat{N}} = \|\mathcal{H}^{(\hat{N})} - \hat{\mathcal{H}}^{(\hat{N})}\|_F^2 \leq 2\mathcal{E}^{\delta, \hat{N}}(N_S)$$

with  $\mathcal{E}^{\delta, \hat{N}} = \alpha^2 \hat{N}^2 \frac{\beta^2(m+s_0)}{N_S} \left(\frac{s^{2\hat{N}+1}-1}{s-1}\right)$ . Here  $\alpha \geq 1$  is a known absolute constant and  $s_0 = \log \frac{(s^{N_S p+1}-1)}{(s-1)^\delta}$ .

*Proof.* Let  $s_0 = \log \frac{(s^{N_S p+1}-1)}{(s-1)^\delta}$ . By definition we have

$$\begin{aligned} & [\mathcal{H}^{(\hat{N})} - \hat{\mathcal{H}}^{(\hat{N})}]_{L(\mathbf{i}_i^j), L(\mathbf{i}_j^i)} = \sqrt{p_{\mathbf{i}_i^j, \mathbf{i}_j^i}} CA_{\mathbf{i}_i^j, \mathbf{i}_j^i} B - \sqrt{\hat{p}_{\mathbf{i}_i^j, \mathbf{i}_j^i}} \hat{\Theta}_{\mathbf{i}_i^j, \mathbf{i}_j^i} \\ &= \underbrace{\left(\sqrt{p_{\mathbf{i}_i^j, \mathbf{i}_j^i}} - \sqrt{\hat{p}_{\mathbf{i}_i^j, \mathbf{i}_j^i}}\right) CA_{\mathbf{i}_i^j, \mathbf{i}_j^i} B}_{=\mathcal{E}_{1, \mathbf{i}_i^j, \mathbf{i}_j^i}} + \underbrace{\sqrt{\hat{p}_{\mathbf{i}_i^j, \mathbf{i}_j^i}} (CA_{\mathbf{i}_i^j, \mathbf{i}_j^i} B - \hat{\Theta}_{\mathbf{i}_i^j, \mathbf{i}_j^i})}_{=\mathcal{E}_{0, \mathbf{i}_i^j, \mathbf{i}_j^i}} \end{aligned}$$

First we analyze  $\mathcal{E}_{1, \mathbf{i}_i^j, \mathbf{i}_j^i}$ . It is clear that

$$\begin{aligned} & \sum_{0 \leq i, j \leq \hat{N}-1} \|\mathcal{E}_{1, \mathbf{i}_i^j, \mathbf{i}_j^i}\|_F^2 \\ & \leq \sum_{0 \leq i, j \leq \hat{N}-1} \left(\sqrt{p_{\mathbf{i}_i^j, \mathbf{i}_j^i}} - \sqrt{\hat{p}_{\mathbf{i}_i^j, \mathbf{i}_j^i}}\right)^2 \|CA_{\mathbf{i}_i^j, \mathbf{i}_j^i} B\|_F^2 \\ & \leq \sum_{k=0}^{2\hat{N}-2} (k+1) \sum_{\mathbf{i}_1^k \in [s]^k} \left(\sqrt{p_{\mathbf{i}_1^k}} - \sqrt{\hat{p}_{\mathbf{i}_1^k}}\right)^2 \|CA_{\mathbf{i}_1^k} B\|_F^2 \end{aligned}$$

The  $k+1$  is the number of times a  $k$ -length sequence appears in the Hankel-like matrix. Using Proposition 8 with probability at least  $1 - \delta$ ,  $\sum_{0 \leq i, j \leq \hat{N}-1} \|\mathcal{E}_{1, \mathbf{i}_i^j, \mathbf{i}_j^i}\|_F^2 \leq \frac{2\alpha \hat{N}^2 \beta^2 s_0}{N_S}$ . For  $\mathcal{E}_{0, \mathbf{i}_i^j, \mathbf{i}_j^i}$  we have

$$\sum \|\mathcal{E}_{0, \mathbf{i}_i^j, \mathbf{i}_j^i}\|_F^2 \leq \sum_{0 \leq i, j \leq \hat{N}-1} \hat{p}_{\mathbf{i}_i^j, \mathbf{i}_j^i} \|CA_{\mathbf{i}_i^j, \mathbf{i}_j^i} B - \hat{\Theta}_{\mathbf{i}_i^j, \mathbf{i}_j^i}\|_F^2$$

Recall that whenever  $N_{\mathbf{i}_i^j, \mathbf{i}_j^i} < s_0$ , i.e., scarce data, we set  $\hat{\Theta}_{\mathbf{i}_i^j, \mathbf{i}_j^i} = 0$ . Then we get

$$\begin{aligned} & \sum \hat{p}_{\mathbf{i}_i^j, \mathbf{i}_j^i} \|CA_{\mathbf{i}_i^j, \mathbf{i}_j^i} B - \hat{\Theta}_{\mathbf{i}_i^j, \mathbf{i}_j^i}\|_F^2 \leq \sum_{N_{\mathbf{i}_i^j, \mathbf{i}_j^i} < s_0} \hat{p}_{\mathbf{i}_i^j, \mathbf{i}_j^i} \|CA_{\mathbf{i}_i^j, \mathbf{i}_j^i} B\|_F^2 \\ & + \sum_{N_{\mathbf{i}_i^j, \mathbf{i}_j^i} \geq s_0} \hat{p}_{\mathbf{i}_i^j, \mathbf{i}_j^i} \|CA_{\mathbf{i}_i^j, \mathbf{i}_j^i} B - \hat{\Theta}_{\mathbf{i}_i^j, \mathbf{i}_j^i}\|_F^2 \end{aligned} \quad (19)$$

We now use Proposition 2 (applied with union bound to all

sequences) with probability at least  $1 - \delta$

$$\begin{aligned}
& \sum_{N_{\hat{i}_1, \hat{i}_1^j} \geq s_0} \frac{N_{\hat{i}_1, \hat{i}_1^j} \|CA_{\hat{i}_1, \hat{i}_1^j} B - \hat{\Theta}_{\hat{i}_1, \hat{i}_1^j}\|_F^2}{N_S(\hat{N} - i - j + 1)} \\
& \leq \alpha^2 \sum_{N_{\hat{i}_1, \hat{i}_1^j} \geq s_0} \frac{\beta^2(m + s_0)}{N_S(\hat{N} - i - j + 1)} \\
& \leq \alpha^2 \sum_{k=0}^{2\hat{N}-2} (k+1) \sum_{\mathbf{l}_1^k \in [s]^k} \frac{\beta^2(m + s_0)}{N_S(\hat{N} - k + 1)} \\
& \leq 2\alpha^2 \hat{N}^2 \sup_{k \leq 2\hat{N}} \sum_{\mathbf{l}_1^k \in [s]^k} \frac{\beta^2 s^k (m + s_0)}{N_S} \\
& \leq 2\alpha^2 \hat{N}^2 \frac{\beta^2(m + s_0)}{N_S} \left( \frac{s^{2\hat{N}+1} - 1}{s - 1} \right) \quad (20)
\end{aligned}$$

From first part in Proposition 7 we get with probability at least  $1 - \delta$

$$\begin{aligned}
\sum_{N_{\hat{i}_1, \hat{i}_1^j} < s_0} \|CA_{\hat{i}_1, \hat{i}_1^j} B\|_F^2 \hat{p}_{\hat{i}_1, \hat{i}_1^j} & \leq \frac{2\alpha^2 \hat{N}^2 s_0}{N_S} \sup_k \sum_{\mathbf{l}_1^k \in [s]^k} \|CA_{\mathbf{l}_1^k} B\|_F^2 \\
& \leq \frac{2\alpha^2 \beta^2 \hat{N}^2 s_0}{N_S} \quad (21)
\end{aligned}$$

Then combining these observations we get  $\sum \hat{p}_{\hat{i}_1, \hat{i}_1^j} \|CA_{\hat{i}_1, \hat{i}_1^j} B - \hat{\Theta}_{\hat{i}_1, \hat{i}_1^j}\|_F^2 \leq \mathcal{E}^{\delta, \hat{N}}$ .  $\square$   $\square$

Proposition 5 provides an upper bound on  $E_{\hat{N}}^2$  almost entirely in terms of data dependent quantities. From here on we will use  $\mathcal{E}^{\delta, \hat{N}}(N_S)$  as a proxy for  $E_{\hat{N}}^2$ . For shorthand  $\mathcal{E}^{\delta, \hat{N}} = \mathcal{E}^{\delta, \hat{N}}(N_S)$ . Given this dependence of estimation error on  $\hat{N}, N_S$ , we find that if we set  $\hat{N}$  in a data dependent fashion as follows:

$$\hat{N} \triangleq \hat{N}(N_S) = \inf \left\{ l \mid \|\hat{\mathcal{H}}^{(l)} - \hat{\mathcal{H}}^{(h)}\|_F^2 \leq \alpha_0 \mathcal{E}^{\delta, h} \right. \\
\left. \forall N_{up} \geq h \geq l \right\} \quad (22)$$

where  $\alpha_0$  is a known absolute constant and  $N_{up}$  is given in Eq. (11).

**Theorem 1.** Fix  $\delta > 0$ . For large enough  $N_S$ , pick  $\hat{N}$  as in Eq. (22). Then with probability at least  $1 - \delta$  we have

$$\|\tilde{\mathcal{H}}^{(\hat{N})} - \mathcal{H}^{(\infty)}\|_F^2 \leq 2\alpha^2 \frac{\hat{N}^2 s_{2\hat{N}} \beta^2 (m + \log \frac{s_{N_{up}}}{\delta})}{N_S}$$

where  $\tilde{\mathcal{H}}^{(\hat{N})}$  is the zero padded version of  $\hat{\mathcal{H}}^{(\hat{N})}$  to make it compatible with  $\mathcal{H}^{(\infty)}$  and  $\alpha \geq 1$  is an absolute constant. Here  $s_k = \frac{s^{k+1} - 1}{s - 1}$ .

*Proof.* We sketch the details of the proof here. We assume all matrices are size compatible by padding with zeros. The large enough  $N_S$  is required only to ensure that there exists  $\hat{N} < \infty$  such that  $T_{\hat{N}} \leq \mathcal{E}^{\delta, \hat{N}}$ . Define  $\hat{N}^* = \inf \{N | T_N \leq \mathcal{E}^{\delta, N}\}$ . In general  $\hat{N}^*$  is unknown as it is complex function of unknown system parameters (because of  $T_N$ ). By Proposition 10 such  $\hat{N}^*$  exists. However, by

leveraging results from [1] specifically Proposition 12.1 and 12.2 we can show that

$$\hat{N}(N_S) \leq \hat{N}^* \leq \log(\alpha_0) \hat{N}(N_S)$$

with probability at least  $1 - \delta$ . We show  $\hat{N}^* \geq \hat{N}(N_S)$  in Proposition 11. The other inequality follows the same steps as Prop 12.2 in [1]. Based on this observation we note for any  $l \geq \hat{N}$

$$\begin{aligned}
\sqrt{\alpha_0 \mathcal{E}^{\delta, \hat{N}^*}} & \geq \|\hat{\mathcal{H}}^{(\hat{N})} - \hat{\mathcal{H}}^{(\hat{N}^*)}\|_F \\
& \geq \|\hat{\mathcal{H}}^{(\hat{N})} - \mathcal{H}^{(\infty)}\|_F - \|\mathcal{H}^{(\infty)} - \hat{\mathcal{H}}^{(\hat{N}^*)}\|_F
\end{aligned}$$

This gives

$$\|\hat{\mathcal{H}}^{(\hat{N})} - \mathcal{H}^{(\infty)}\|_F \leq \sqrt{\alpha_0 \mathcal{E}^{\delta, \hat{N}^*}} + \|\mathcal{H}^{(\infty)} - \hat{\mathcal{H}}^{(\hat{N}^*)}\|_F$$

Since  $E_{\hat{N}^*} \geq T_{\hat{N}^*}$  we have that  $\|\hat{\mathcal{H}}^{(\hat{N}^*)} - \hat{\mathcal{H}}^{(\hat{N}^*)}\|_F \leq 2\mathcal{E}^{\delta, \hat{N}^*}$ . Then it implies that

$$\|\hat{\mathcal{H}}^{(\hat{N})} - \mathcal{H}^{(\infty)}\|_F^2 \leq (\alpha_0 + 2) \mathcal{E}^{\delta, \hat{N}^*} \quad (23)$$

Our claim follows by noting that  $\mathcal{E}^{\delta, \hat{N}^*} \leq \mathcal{E}^{\delta, \log(\alpha_0) \hat{N}}$ .  $\square$   $\square$

The key insight of Theorem 1 is that for the choice of  $\hat{N}(N_S)$  in Eq. (22) we can get a good upper bound on the error between the true system Hankel matrix,  $\mathcal{H}^{(\infty)}$ , and its estimate  $\hat{\mathcal{H}}^{(\hat{N})}$ . Furthermore this bound does not depend on the system order,  $n$ , but only data dependent quantities and some energy metrics which can be measured easily. The result in Proposition 10 (and Eq. (29)) shows that

$$\|\tilde{\mathcal{H}}^{(\hat{N})} - \mathcal{H}^{(\infty)}\|_F^2 = O(N^{-\delta_s}) \quad (24)$$

where  $\delta_s = \left( \log \frac{1}{p_{\max}} \right) \left( \log \frac{s}{p_{\max}} \right)^{-1}$ . Eq. (24) shows that decay in error between the true system Hankel-like matrix and its estimator is roughly  $\frac{1}{\sqrt{N^{-\delta_s}}}$  (ignoring the log factors) and the error between  $\tilde{\mathcal{H}}^{(\hat{N})}, \mathcal{H}^{(\infty)}$  goes to zero asymptotically as  $N_S \rightarrow \infty$ .

## B. Selecting $r$

Now that we have a consistent statistical estimator for  $\mathcal{H}^{(\infty)}$ . We provide a way to choose  $r$  such that we can find a  $r$ -order balanced representation of the SLS. For shorthand, we will refer to the data dependent error  $\epsilon^2 = 4\alpha_0 \mathcal{E}^{\delta, \hat{N}}$ . This implies  $\|\mathcal{H}^{(\infty)} - \hat{\mathcal{H}}^{(\hat{N})}\|_F \leq \epsilon$  and we can use Wedin-type subspace perturbation bounds [28]. Consider the following rule for selecting  $r$

$$r = \sup \left\{ l \mid \tau_+ \sigma_l(\hat{\mathcal{H}}^{(\hat{N})}) \geq 4\epsilon \right\} \quad (25)$$

The existence of  $\tau_+$  is not required for our results as the same discussion of Section 11.3 in [1] would apply here. Furthermore, we can also substitute  $\tau_+$  by  $\hat{\tau}_+ = \inf_{1 \leq i \leq n} \left( 1 - \frac{\sigma_{i+1}(\hat{\mathcal{H}}^{(\hat{N})})}{\sigma_i(\hat{\mathcal{H}}^{(\hat{N})})} \right)$  and that performs sufficiently well.

**Theorem 2** (Theorem 5.2 [1]). *For large enough  $N_S$ , we have for the choice of  $\tilde{N}, r$  in Eq. (22), (25) respectively that*

$$\sup \{ \|\tilde{C} - C^{(r)}\|_2, \|\tilde{B} - B^{(r)}\|_2 \} \leq O\left(\frac{\epsilon}{\sqrt{\sigma_r}}\right)$$

$$\sup_{1 \leq i \leq s} \|\tilde{A}_i - A_i^{(r)}\|_2 \leq O\left(\frac{\epsilon}{\sigma_r}\right)$$

where  $\epsilon^2 = O\left(N_S^{-\delta_s}\right)$  and  $\delta_s = \left(\log \frac{1}{p_{\max}}\right) \left(\log \frac{s}{p_{\max}}\right)^{-1}$ . Here  $\tilde{C}, \tilde{A}_i, \tilde{B}$  is the output of Algorithm 1 and  $(C^{(r)}, A_i^{(r)}, B^{(r)})$  are  $r$ -order balanced truncated model given in Eq. (16).

Theorem 2 indicates that finding an  $r$ -order balanced truncated model depends inversely on the  $r^{\text{th}}$  singular value of  $\Sigma$  in Eq. (15). Note that in Eq. (25) as  $N_S$  increase  $\epsilon$  decreases, i.e., the estimate  $\mathcal{H}^{(\tilde{N})}$  becomes better and indeed if  $\epsilon = 0 \implies \mathcal{H}^{(\tilde{N})} = \mathcal{H}^{(\infty)} \implies r = n$ .

#### IV. DISCUSSION

In this work we provide finite sample error guarantees for learning realizations of SLS when stability radius or order is unknown. Specifically, we construct a Hankel-like matrix of size  $\tilde{N}$ , chosen in a data dependent fashion. From this Hankel-like matrix we recover system parameters using a data dependent threshold rule in Eq. (25). Under stated assumptions, we obtain  $O(\sqrt{N^{-\delta_s}})$  error rates which are also the parametric estimation error rates and are known to be optimal for the case when  $s = 1$  (See for e.g.: [1]). Furthermore, from a computational perspective our algorithm is polynomial in the number of samples,  $N_S$ , because we are doing SVD on a matrix of dimension at most  $p s^{\tilde{N}} \times m s^{\tilde{N}}$  but  $\tilde{N}$  is logarithmic in  $N_S$  with high probability and as a result the matrix size is polynomial in  $N_S$ .

Due to the nature of the analysis we believe that this work can be easily extended to the case when  $\{\theta_t\}_{t=1}^{\infty}$  evolution is more complex, for e.g.: state dependent or a markov chain. Furthermore, we assumed in this paper the discrete switches are completely observable. However, in many cases the discrete state itself might be noisy or not observed. In such cases it important to predict the switch sequence and following that use the procedure described. This appears to be an interesting avenue for future work.

#### V. APPENDIX

**Proposition 6** (Bernstein's Inequality). *Let  $\{X_i\}_{i=1}^n$  be zero mean random variables. Suppose that  $|X_i| \leq M$  almost surely, for all  $i$ . Then, for all positive  $t$ ,*

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| > t\right) \leq \exp\left(-\frac{\frac{1}{2}t^2}{\sum_{j=1}^n \mathbb{E}[X_j^2] + \frac{1}{3}Mt}\right) \quad (26)$$

Recall from Eq. (12) that

$$\hat{p}_{1^k} = \frac{N_{1^k}}{N_S(N-k+1)} = \frac{1}{N_S} \sum_{i=1}^{N_S} \underbrace{\sum_{l=1}^{N-k+1} \mathbf{1}_{\{\mathbf{1}_{1^k} \text{ starts at } l\}}^{(i)}}}_{x_i}$$

where  $0 \leq x_i \leq 1$  and  $\{x_i\}_{i=1}^{N_S}$  are i.i.d random variables. Since  $\mathbb{E}[x_i] = p_{1^k}$ . We can use Bernstein's inequality on  $\sum_{i=1}^{N_S} (x_i - p_{1^k})$ . Note  $\mathbb{E}[(x_i - p_{1^k})^2] \leq p_{1^k}$ . Then by Bernstein's inequality we have

**Proposition 7.** *Fix  $\delta, N > 0$ . For all sequences  $1^k$  with  $k \leq N$  we have simultaneously with probability at least  $1 - \delta$*

$$\left| \sum_{i=1}^{N_S} x_i - p_{1^k} N_S \right| \leq \begin{cases} \alpha \log \frac{s_N}{\delta}, & \text{if } \alpha \log \frac{s_N}{\delta} > p_{1^k} N_S \\ \sqrt{p_{1^k} N_S \log \frac{s_N}{\delta}}, & \text{otherwise} \end{cases}$$

for some known absolute constant  $\alpha \geq 1$  and  $s_N = \frac{s^{N+1}-1}{s-1}$ .

**Proposition 8.** *Fix  $0 \leq \hat{N} \leq N_S$ , then with probability at least  $1 - \delta$  we have*

$$\sum_{1^k \in [s]^k} \left( \sqrt{p_{1^k}} - \sqrt{\hat{p}_{1^k}} \right)^2 \|CA_{1^k} B\|_F^2 \leq \frac{2\alpha\beta^2}{N_S} \log \frac{(s^{\hat{N}+1} - 1)}{(s-1)\delta}$$

where  $\alpha \geq 1$  is a known absolute constant.

*Proof.* Let  $s_0 = \alpha \log \left( \frac{s^{N+1}-1}{(s-1)\delta} \right)$ . Now we break the sum in two parts

$$\begin{aligned} & \sum_{1^k \in [s]^k} \left( \sqrt{p_{1^k}} - \sqrt{\hat{p}_{1^k}} \right)^2 \|CA_{1^k} B\|_F^2 \\ & \leq \underbrace{\sum_{p_{1^k} N_S \leq s_0} \left( \sqrt{p_{1^k}} - \sqrt{\hat{p}_{1^k}} \right)^2 \|CA_{1^k} B\|_F^2}_{(i)} \\ & \quad + \underbrace{\sum_{p_{1^k} N_S > s_0} \left( \sqrt{p_{1^k}} - \sqrt{\hat{p}_{1^k}} \right)^2 \|CA_{1^k} B\|_F^2}_{(ii)} \quad (27) \end{aligned}$$

For (i) combine  $(\sqrt{p_{1^k}} - \sqrt{\hat{p}_{1^k}})^2 \leq |p_{1^k} - \hat{p}_{1^k}|$  and use

Proposition 7 which gives (i)  $\leq \sum_{1^k \in [s]^k} \frac{s_0 \|CA_{1^k} B\|_F^2}{N_S}$ . For (ii), it follows from the second part in Proposition 7 that  $(\sqrt{p_{1^k}} - \sqrt{\hat{p}_{1^k}})^2 \leq p_{1^k} \left( \sqrt{1 + \sqrt{\frac{s_0}{p_{1^k} N_S}}} - 1 \right)^2 \leq \alpha \frac{s_0}{N_S}$ . Then

we get that (ii)  $\leq \sum_{1^k \in [s]^k} \frac{\alpha s_0 \|CA_{1^k} B\|_F^2}{N_S}$ . By assumption we have  $\sum_{1^k \in [s]^k} \|CA_{1^k} B\|_F^2 \leq \beta^2$ .  $\square$   $\square$

**Proposition 9.** *For  $N_{up}$  defined in Eq. (11) we can show with probability at least  $1 - \delta$  that*

$$N_{up} = O\left(\frac{\log N_S}{\log \frac{1}{p_{\max}}}\right)$$

where  $p_{\max} = \max_{1 \leq i \leq s} p_i$ . Here  $O(\cdot)$  also hides a dependence of  $\log \log \frac{1}{\delta}$ .

*Proof.* Due to a shortage of space we only sketch the proof here. Note from Proposition 7 that for all sequences of length  $N$ ,  $N_{1^N} \leq \alpha \log \frac{s^{N+1}-1}{(s-1)\delta}$  with high probability if

$$\alpha \left( m + \log \left( \frac{s^{N+1} - 1}{(s-1)\delta} \right) \right) \geq p_{1^N} N_S$$



since  $p_{i_1^N} \leq p_{\max}^N N_S$ , if we ensure  $p_{\max}^N N_S = \alpha \left( m + \log \left( \frac{s^{N+1}-1}{(s-1)\delta} \right) \right)$  we get our desired result for all sequences of length up to  $N$ .  $\square$   $\square$

**Proposition 10.** Let  $T_{\hat{N}}^2 = \|\hat{\mathcal{H}}^{(\hat{N})} - \mathcal{H}^{(\infty)}\|_F^2$ . Then for a large enough  $N_S$  there exists  $\hat{N}$  such that

$$T_{\hat{N}}^2 \leq \mathcal{E}^{\delta, \hat{N}}(N_S)$$

In fact,  $\hat{N} = O(\log N_S)$  where  $O(\cdot)$  hides system level dependence.

*Proof.* Define  $s_0 = \log \left( \frac{s^{N_{up}+1}-1}{\delta(s-1)} \right)$ . Since the SLS is mean-square stable and by our assumptions we have

$$\begin{aligned} \sum_{i+j \geq 2\hat{N}-1} p_{i_1, i_1^j} \|CA_{i_1, i_1^j} B\|_F^2 &= \sum_{k \geq 2\hat{N}-1} p_{i_1^k} (k+1) \|CA_{i_1^k} B\|_F^2 \\ &\leq \sum_{k \geq 2\hat{N}-1} \sum_{i_1^k \in [s]^k} (k+1) \|CA_{i_1^k} B\|_F^2 p_{\max}^k \end{aligned}$$

This gives us  $\sum_{k \geq 2\hat{N}-1} \sum_{i_1^k \in [s]^k} k \|CA_{i_1^k} B\|_F^2 p_{\max}^k \leq \sum_{k \geq 2\hat{N}-1} (k+1) p_{\max}^k \beta^2 \leq \frac{2\hat{N}^2 p_{\max}^{2\hat{N}-1} \beta^2}{1-p_{\max}}$ . Then clearly  $\hat{N}^*$  is less than  $\hat{N}$  that satisfies

$$\begin{aligned} \mathcal{E}^{\delta, \hat{N}}(N_S) &= \alpha^2 \hat{N}^2 \frac{\beta^2 (m + s_0)}{N_S} \left( \frac{s^{2\hat{N}+1} - 1}{s-1} \right) \\ &\geq \frac{2\hat{N}^2 p_{\max}^{2\hat{N}-1} \beta^2}{1-p_{\max}} \geq T_{\hat{N}}^2 \end{aligned} \quad (28)$$

The last inequality is satisfied for all  $\hat{N}^*$  such that  $\mathbf{I}_1^{\hat{N}}$  occurs often enough. Furthermore, from the proof of Proposition 9,  $\hat{N}^* < N_{up}$ , since  $N_{up}$  satisfies  $\alpha \left( m + \log \left( \frac{s^{N_{up}+1}-1}{(s-1)\delta} \right) \right) \approx p_{\max}^{2N_{up}-1} N_S$ . Then by solving the functional equation in Eq. (28) we get  $\hat{N}^* \leq \frac{1}{2} \left( \frac{\log N_S}{\log \frac{s}{p_{\max}}} \right)$ . This gives us the error

$$\mathcal{E}^{\delta, \hat{N}^*} \leq O(N_S^{-\delta_s}) \quad (29)$$

where  $\delta_s = \left( \log \frac{1}{p_{\max}} \right) \left( \log \frac{s}{p_{\max}} \right)^{-1}$  and  $O(\cdot)$  hides logarithmic factors in  $N_S$ . Note that  $\delta_1 = N_S^{-1}$  which is the correct rate for LTI systems.  $\square$   $\square$

**Proposition 11.** Let  $\hat{N}^* = \inf \{N | T_N \leq \mathcal{E}^{\delta, N}\}$ , then  $\hat{N} \leq \hat{N}^*$ .

*Proof.* We prove this by showing for  $l, h \geq \hat{N}^*$  satisfies  $\|\hat{\mathcal{H}}^{(l)} - \hat{\mathcal{H}}^{(h)}\|_F^2 \leq 9\mathcal{E}^{\delta, h}$ . To see this

$$\begin{aligned} \|\hat{\mathcal{H}}^{(l)} - \hat{\mathcal{H}}^{(h)}\|_F &\leq \|\hat{\mathcal{H}}^{(l)} - \mathcal{H}^{(l)}\|_F + \|\hat{\mathcal{H}}^{(h)} - \mathcal{H}^{(h)}\|_F \\ &\quad + \|\mathcal{H}^{(l)} - \mathcal{H}^{(h)}\|_F \end{aligned} \quad (30)$$

Since  $\|\hat{\mathcal{H}}^{(l)} - \mathcal{H}^{(l)}\|_F^2 \leq \mathcal{E}^{\delta, l} \leq \mathcal{E}^{\delta, h}$  and further more  $\|\mathcal{H}^{(l)} - \mathcal{H}^{(h)}\|_F^2 \leq \|\mathcal{H}^{(l)} - \mathcal{H}^{(\infty)}\|_F^2 \leq \mathcal{E}^{\delta, l}$ . Combining all of this we get  $\|\hat{\mathcal{H}}^{(l)} - \hat{\mathcal{H}}^{(h)}\|_F^2 \leq 3\mathcal{E}^{\delta, h}$  and this means that  $\hat{N} \leq \hat{N}^*$ .  $\square$

## REFERENCES

- [1] T. Sarkar, A. Rakhlin, and M. A. Dahleh, "Finite-time system identification for partially observed lti systems of unknown order," *arXiv preprint arXiv:1902.01848*, 2019.
- [2] S. R. Venkatesh and M. A. Dahleh, "On system identification of complex systems from finite data," *IEEE Transactions on Automatic Control*, vol. 46, no. 2, pp. 235–257, 2001.
- [3] J. Ezzine and A. H. Haddad, "Controllability and observability of hybrid systems," *International Journal of Control*, vol. 49, no. 6, pp. 2045–2055, 1989.
- [4] M. S. Branicky, "Introduction to hybrid systems," in *Handbook of networked and embedded control systems*. Springer, 2005, pp. 91–116.
- [5] Z. Sun, *Switched linear systems: control and design*. Springer Science & Business Media, 2006.
- [6] S. W. Linderman, A. C. Miller, R. P. Adams, D. M. Blei, L. Paninski, and M. J. Johnson, "Recurrent switching linear dynamical systems," *arXiv preprint arXiv:1610.08466*, 2016.
- [7] E. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "Bayesian nonparametric inference of switching dynamic linear models," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1569–1585, 2011.
- [8] R. L. Westra, M. P. Ralf, and L. Peeters, "Identification of piecewise linear models of complex dynamical systems," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 14863–14868, 2011.
- [9] R. L. Grossman and R. G. Larson, "An algebraic approach to hybrid systems."
- [10] M. Petreczky and J. H. van Schuppen, "Realization theory for linear hybrid systems," *IEEE Transactions on Automatic Control*, vol. 55, no. 10, pp. 2282–2297, 2010.
- [11] M. Petreczky and R. Vidal, "Realization theory for a class of stochastic bilinear systems," *IEEE Transactions on Automatic Control*, vol. 63, no. 1, pp. 69–84, Jan 2018.
- [12] L. Bako, G. Mercère, R. Vidal, and S. Lecoeuche, "Identification of switched linear state space models without minimum dwell time," *IFAC Proceedings Volumes*, vol. 42, no. 10, pp. 569–574, 2009.
- [13] L. Bako, "Identification of switched linear systems via sparse optimization," *Automatica*, vol. 47, no. 4, pp. 668–677, 2011.
- [14] N. Ozay, C. Lagoa, and M. Szaier, "Robust identification of switched affine systems via moments-based convex optimization," in *Proceedings of the 48th IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*. IEEE, 2009, pp. 4686–4691.
- [15] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (gpca)," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 12, pp. 1945–1959, 2005.
- [16] C. Feng, C. M. Lagoa, N. Ozay, and M. Szaier, "Hybrid system identification: An sdp approach," in *49th IEEE Conference on Decision and Control (CDC)*. IEEE, 2010, pp. 1546–1552.
- [17] F. Lauer and G. Bloch, "Hybrid system identification," in *Hybrid System Identification*. Springer, 2019, pp. 77–101.
- [18] G. Kotsalis, A. Megretski, and M. A. Dahleh, "Balanced truncation for a class of stochastic jump linear systems and model reduction for hidden markov models," *IEEE Transactions on Automatic Control*, vol. 53, no. 11, pp. 2543–2557, 2008.
- [19] A. Birouche, B. Mourllion, and M. Basset, "Model order-reduction for discrete-time switched linear systems," *International Journal of Systems Science*, vol. 43, no. 9, pp. 1753–1763, 2012.
- [20] M. Petreczky, R. Wisniewski, and J. Leth, "Balanced truncation for linear switched systems," *Nonlinear Analysis: Hybrid Systems*, vol. 10, pp. 4–20, 2013.
- [21] I. V. Gosea, M. Petreczky, and A. C. Antoulas, "Data-driven model order reduction of linear switched systems in the loewner framework," *SIAM Journal on Scientific Computing*, vol. 40, no. 2, pp. B572–B610, 2018.
- [22] R. Vershynin, "High-dimensional probability," 2019.
- [23] Q. Huang, R. Ge, S. Kakade, and M. Dahleh, "Minimal realization problem for hidden markov models," in *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2014, pp. 4–11.
- [24] H. Ito, S.-I. Amari, and K. Kobayashi, "Identifiability of hidden markov information sources and their minimum degrees of freedom," *IEEE transactions on information theory*, vol. 38, no. 2, pp. 324–333, 1992.

- [25] B. Balle, X. Carreras, F. M. Luque, and A. Quattoni, "Spectral learning of weighted automata," *Machine learning*, vol. 96, no. 1-2, pp. 33–63, 2014.
- [26] S. Tu, R. Boczar, A. Packard, and B. Recht, "Non-asymptotic analysis of robust control from coarse-grained identification," *arXiv preprint arXiv:1707.04791*, 2017.
- [27] K. Zhou, J. Doyle, and K. Glover, "Robust and optimal control," 1996.
- [28] P.-Å. Wedin, "Perturbation bounds in connection with singular value decomposition," *BIT Numerical Mathematics*, vol. 12, no. 1, pp. 99–111, 1972.