Finite Sample System Identification: Improved Rates and the Role of Regularization

Yue Sun YUESUN@UW.EDU
University of Washington, Seattle, WA 98105, US

Samet Oymak OYMAK@ECE.UCR.EDU
University of California, Riverside, CA 92507, US

Maryam Fazel MFAZEL@UW.EDU
University of Washington, Seattle, WA 98105, US

Editors: A. Bayen, A. Jadbabaie, G. J. Pappas, P. Parrilo, B. Recht, C. Tomlin, M.Zeilinger

Abstract

This paper studies low-order linear system identification via regularized regression. The nuclear norm of the system's Hankel matrix is added as a regularizer to the least-squares cost function due to the following advantages. The regularized problem is (1) often easier to tune, (2) has lower sample complexity, and (3) returns a Hankel matrix with a clear singular value gap, which robustly recovers a low-order linear system from noisy output observations. Recently, the performance of the unregularized least-squares formulations have been studied statistically in terms of finite sample complexity and recovery errors; however, no results are known for the regularized approach. In this work, we provide a novel statistical analysis of the regularized algorithm. Our analysis leads to new bounds on estimating the impulse response and the Hankel matrix associated with the system while using smaller number of observations than the least-squares estimator.

1. Introduction

System identification is an important topic in control theory. An accurate estimation of system dynamics is the basis of the associated control or policy decision problems in tasks varying from linear-quadratic control to deep reinforcement learning. Consider a linear time-invariant system of order R with the state-space representation

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t, \\ y_t &= Cx_t + Du_t + z_t, \end{aligned} \tag{1}$$

where $x_t \in \mathbb{R}^R$ is the state, $u_t \in \mathbb{R}^p$ is the input, $y_t \in \mathbb{R}^m$ is the output, $z_t \in \mathbb{R}^m$ is the output noise, $A \in \mathbb{R}^{R\times R}$, $B \in \mathbb{R}^{R\times p}$, $C \in \mathbb{R}^{m\times R}$, $D \in \mathbb{R}^{m\times p}$ are the system parameters, and x_0 is the initial state (in this paper, we assume $x_0 = 0$). The system identification problem is finding the system parameters, given input and output observations. When $C = I$, we directly observe the state, otherwise we may obtain only partial state information. A notable line of work derives statistical bounds for system identification with limited *state* observations from a single output trajectory with a random input Abbasi-Yadkori and Szepesvári (2011); Simchowitz et al. (2018); Sarkar and Rakhlin (2019). The core approach is using the least-squares estimator and then adapting the self-normalized martingale bounds from Abbasi-Yadkori et al. (2011). These works require observing the state.

For the hidden-state system in (1), the impulse response sequence $h_0 = D$, $h_t = CA^{t-1}B \in \mathbb{R}^{m\times p}$ for $t = 1, 2, \ldots$ (also known as the Markov parameters) uniquely identifies the end-to-end behavior of the system. The impulse response of the system has infinite length, and we let $h = [D, CB, CAB, CA^2B, \ldots, CA^{2n-3}B]^\top$ denote its first $2n-1$ entries. We also define the Hankel map $\mathcal{H} : \mathbb{R}^{m\times(2n-1)p} \to \mathbb{R}^{mn\times pn}$ as

$$H := \mathcal{H}(h) = \begin{bmatrix} h_1 & h_2 & \ldots & h_n \\ h_2 & h_3 & \ldots & h_{n+1} \\ \ldots & & & \\ h_n & h_{n+1} & \ldots & h_{2n-1} \end{bmatrix}. \tag{2}$$

If R is the system order and $n \geq R$, the Hankel matrix H is of rank R regardless of n. A practically relevant scenario is when the order R is not known in advance or may be misspecified. Specifically, we will assume that R is small, and explore the use of nuclear norm regularization to find a low-rank Hankel matrix.

Least-squares can be used to recover the Markov parameters and reconstruct A, B, C, D from the Hankel matrix via the Ho-Kalman algorithm (Ho and Kálmán (1966)). To identify a stable system from a single trajectory, Oymak and Ozay (2018) estimates the Markov parameter matrix h and Sarkar et al. (2019) estimates the Hankel matrix via least-squares. The latter provides optimal Hankel spectral norm error rates, however has suboptimal sample complexity (see the table in Section 2). While Oymak and Ozay (2018); Sarkar et al. (2019) use random input, (Tu et al., 2017, Thm 1.1, 1.2) use impulse and single frequency signal respectively as input, both recovering system Markov parameters. These works assume (roughly) known system order, or traverse the Hankel size n to fit the system order. When the system order is unknown and assumed low, Cai et al. (2016) studies least-squares with Hankel nuclear norm regularization, showing that the number of observations is proportional to the underlying system order and is insensitive to the problem dimension.

In this paper, we study the sample complexity and estimation errors for least-squares and nuclear norm regularized estimators. Oymak and Ozay (2018) and Sarkar et al. (2019) recover the system from single rollout/trajectory of input, whereas our work, Tu et al. (2017) and Cai et al. (2016) require multiple rollouts. To ensure a standardized comparison, we define *sample complexity* to be the number of equations (equality constraints in variables h_t) used in the problem formulation. With this, we explore the following performance metrics of learning the system from a finite set of T measurements.

- **Sample complexity:** The minimum sample size T for recovering system parameters with zero error when the noise $z = 0$. This quantity is lower bounded by the system order which is the degrees of freedom.
- **Impulse Response (IR) Estimation Error:** The Frobenius norm error for the IR $\|\hat{h} - h\|_F$. Knowing the impulse response enables accurate prediction of the system output.
- **Hankel Estimation Error:** The spectral norm error of the Hankel matrix $\|\mathcal{H}(\hat{h} - h)\|$. This performance metric is particularly important for system identification as described below.

The Hankel spectral norm error is a critical quantity to control for several reasons. First, the Hankel spectral error connects to the $\mathcal{H}_\infty$ estimation of the system via classical arguments Sanchez-Pena and Sznaier (1998). Secondly, bounding this error allows for robustly finding balanced realizations of the system; for example, the error in reconstructing A, B, C, D via the Ho-Kalman procedure is bounded by the Hankel spectral error. Finally, it is beneficial in model selection, as a small spectral error helps

distinguish the true singular values of the system from the spurious ones. Indeed, as illustrated in the experiments, the Hankel singular value gap of the solution of the regularized algorithm is more visible compared to least-squares, which helps in identifying the order of the system with a parameter λ that is easy to tune in (5) as explored in section 4.

Contributions. Below, we list our contributions, and contrast our results with the existing work.

• **Nuclear norm regularization** (Sec 3.1): For multi-input/single-output (MISO) systems, we establish sample complexity bounds for the nuclear norm regularized system identification problem, showing the required sample size grows as $O(pR \log^2 n)$, which is linear in the system order R. This result build directly on Cai et al. (2016) which analyzed the recovery a sum-of-exponentials signal using Hankel nuclear norm (which is equivalent to SISO system identification).

Our work also establishes statistically consistent error rates on the IR and Hankel spectral errors (i.e., the estimates to the ground-truth system parameters with growing sample size). This is in contrast to the error bounds of Cai et al. (2016). Our rates are at least as good as least-squares rates; however, they apply in the small sample size regime[1] $T \lesssim pR^2 \log^2 n$[2].

• **Least-squares estimator** (Sec 3.2): It is fairly straightforward to show that least-squares estimator for the impulse response h has a guaranteed error bound when $T \gtrsim np$ (c.f. Oymak and Ozay (2018)). However the bound of Oymak and Ozay (2018) is loose when it comes to Hankel spectral error. For multi-input/multi-output (MIMO) systems, we establish the *optimal spectral error bound* on the Hankel matrix. Sarkar et al. (2019) and Tu et al. (2017) also provide similar bounds, however their sample complexities are suboptimal as they require $O(n^2)$ measurements rather than $O(n)$.

• **Relating IR and Hankel errors:** Note that one can upper/lower bound the Hankel error in terms of IR error using the fact that rows of the Hankel matrix are subsets of the IR sequence. Specifically, we always have the inequality

$$\|\hat{h} - h\|_F/\sqrt{2} \leq \|\mathcal{H}(\hat{h} - h)\| \leq \sqrt{n}\|\hat{h} - h\|_F. \tag{3}$$

Observe that there is a factor of $\sqrt{n}$ difference between the left and right-hand side inequalities. One contribution of this work is that, perhaps surprisingly, we show that the left-hand side inequality is typically the tighter one and we have $\|\hat{h} - h\|_F \sim \|\mathcal{H}(\hat{h} - h)\|$.

• **Experimental performance** (Sec 4): Finally, we numerically explore the regularized and unregularized algorithms for system identification from single-trajectory data. Our synthetic and real-data experiments (on a low-order example from the DaiSy De Moor et al. (1997) datasets) suggest that the regularized algorithm has empirical benefits in sample complexity, error, and Hankel spectral gap, and demonstrate that the regularized algorithm is less sensitive to the choice of the tuning parameter than the least squares algorithm is to the Hankel matrix size. Another experiment compares the two least-squares approaches in Oymak and Ozay (2018) and Sarkar et al. (2019), showing that the former (which estimates the impulse response) performs substantially better than the latter (which estimates the Hankel matrix). This highlights the role of proper parameterization in system identification.

Further literature review. Nuclear norm regularization has been shown to recover an unstructured low-rank matrix in a sample-efficient way in many settings (e.g., Recht et al. (2010); Candes and Plan (2010)). Far less is known about recovery when the matrix has linear structure such Hankel, block-Hankel, or Toeplitz structure that arise frequently in signal processing and control Ding et al.

1. $a \gtrsim b$ and $a \lesssim b$ stand for "there exist a constant c (that does not depend on other parameters) such that $a \geq cb$ or $a \leq cb$".
2. We also get slightly weaker results in the regime $pR^2 \log^2 n \gtrsim T \gtrsim pR \log^2 n$.

(2007); Elad et al. (2004). For system identification, while the use of nuclear norm is common (e.g., Wahlberg and Rojas (2013); Blomberg et al. (2015); Blomberg (2016); Markovsky (2012)) and some efficient algorithms have been proposed to solve the regularized regression problem Liu et al. (2013); Fazel et al. (2013), few statistical recovery guarantees are available.

There are several interesting generalizations of least squares with non-asymptotic guarantees for different goals. Hazan et al. (2018) and Simchowitz et al. (2019) introduced filtering strategies on top of least squares. The filters in Hazan et al. (2018) is the top eigenvectors of a deterministic matrix, used for output prediction in stable systems. Simchowitz et al. (2019) uses filters in frequency domain to recover the system parameters of a stable system, Tsiamis and Pappas (2019) gives a non-asymptotic analysis for learning a Kalman filter system, which can also be applied to an auto-regressive setting.

2. Problem Setup and Algorithms

Let $\|\cdot\|, \|\cdot\|_*, \|\cdot\|_F$ denote the spectral norm, nuclear norm and Frobenius norm respectively. Throughout, we work with the first $2n-1$ terms of the impulse response denoted by h. The system is excited by input u in the time interval $[0, t]$ and output y is measured at time t, i.e.,

$$y_t = \sum_{i=1}^{t} h_{t+1-i} u_i + z_t. \tag{4}$$

We start by describing data acquisition models. In the setting that we refer to as "multi-rollout" (Figure 1(a)), for each input signal $u^{(i)}$ we take only one output measurement y_t at time $t = 2n-1$ and then the system is restarted with a new input (for example, in a chemical system experiment, or more generally in cases where measuring the output is expensive). Here the *sample complexity* is T, the number of inputs. Many papers in the literature (e.g., Oymak and Ozay (2018) and Sarkar et al. (2019)) use the "single rollout" model (Figure 1(b)) where we apply an input signal from time 1 to $T + 2n - 1$ without restart, and collect all output from time $2n-1$; we use this model in the numerical experiments in section 4. We consider two estimators in this paper: the *nuclear norm*

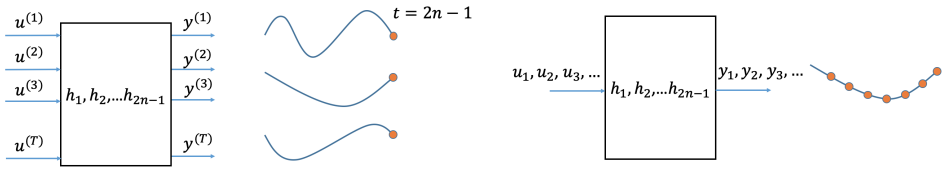

Figure 1: Two data aqcuisition models: (a) Multi-rollout (left), and (b) single rollout (right).

regularized estimator and the *least square estimator*. The nuclear norm regularized estimator is

$$\hat{h} = \arg\min_{h'} \quad \frac{1}{2}\|\bar{\boldsymbol{U}} h' - y\|_F^2 + \lambda \|\mathcal{H}(h')\|_*, \tag{5}$$

which reduces to the (unregularized) least-squares estimator when $\lambda = 0$.

We would like to bound the various error metrics mentioned earlier in terms of the true system order R, the dimension of impulse response $n \gg R$, and signal to noise ratio (SNR) defined as

snr $= \mathbb{E}[\|u\|^2/n]/\,\mathbb{E}[\|z\|^2]$. The following table provides a summary and comparison of these bounds. In the table, the Hankel matrix is $n \times n$, the system order is R, and the number of samples is T, and $\sigma = 1/\sqrt{\textbf{snr}}$ denotes the noise level. LS-IR and LS-Hankel stands for least square regression on the impulse response and on the Hankel matrix. All bounds are order-wise and hide log factors.

Paper	This work	This work	Oymak and Ozay (2018)	Sarkar et al. (2019)
Sample complexity	R	n	n	n^2
Method	Nuc-norm	LS-IR	LS-IR	LS-Hankel
IR error	see (7)	$\sigma\sqrt{n/T}$	$\sigma\sqrt{n/T}$	$(1+\sigma)\sqrt{n/T}$
Hankel spectral error	see (7)	$\sigma\sqrt{n/T}$	$\sigma n/\sqrt{T}$	$(1+\sigma)\sqrt{n/T}$

We consider a multiple rollout setup where we measure the system dynamics with T separate rollouts. For each rollout, we drive the system with an input sequence $u^{(i)} \in \mathbb{R}^{(2n-1)p}$ and measure the system output at time $2n-1$. Note that the output at time $2n-1$ is simply $h^\top u$. Define $\bar{\boldsymbol{U}} \in \mathbb{R}^{T\times(2n-1)p}$ where each row is a rollout of inputs, and let $y \in \mathbb{R}^{T\times m}$ denote the corresponding observed outputs. We consider the nuclear norm regularized problem (5). Note that the $\mathcal{H}$ operator does not preserve the Euclidean norm, so Cai et al. (2016) proposes using a normalized operator $\mathcal{G}$, where they first define the weights

$$K_j = \begin{cases} \sqrt{j}, & 1 \le j \le n \\ \sqrt{2n-j}, & n < j \le 2n-1 \end{cases}$$

and let $K = \mathrm{diag}(K_j \mathbf{I}_{p\times p}) \in \mathbb{R}^{(2n-1)p\times(2n-1)p}$, and define the mapping $\mathcal{G}(h) = \mathcal{H}(K^{-1}h)$. In other words, if $\beta = Kh$ then $\mathcal{G}(\beta) = \mathcal{H}(h)$. Define $\mathcal{G}^* : \mathbb{R}^{mn\times np} \to \mathbb{R}^{m\times(2n-1)p}$ as the adjoint of $\mathcal{G}$, where $[\mathcal{G}^*(M)]_i = \sum_{j+k-1=i} M_{(j)(k)}/K_i$ if we denote the j,k-th block of M (defined in (2)) by $M_{(j)(k)}$. Using this change of variable and letting $\boldsymbol{U} = \bar{\boldsymbol{U}}K^{-1}$, problem (5) can be written as

$$\hat{\beta} = \arg\min_{\beta'} \quad \frac{1}{2}\|\boldsymbol{U}\beta' - y\|_F^2 + \lambda\|\mathcal{G}(\beta')\|_*. \tag{6}$$

3. Main results

3.1. Hankel Nuclear Norm Regularization

To promote a low-rank Hankel matrix, we add nuclear norm regularization in our objective and solve the regularized regression problem. Here we give a finite sample analysis for the recovery of the Hankel matrix and the impulse response found via this approach. We consider a random input matrix $\bar{\boldsymbol{U}}$ and observe the corresponding noisy output vector y as in (4). We then regress y and $\bar{\boldsymbol{U}}$ such that $y = \bar{\boldsymbol{U}}h + z$ where z is the noise vector.

Theorem 1 [3] *Consider the problem (5) in the MISO (multi-input single-output) setting (m=1, p inputs). Suppose the system is order R, $\bar{\boldsymbol{U}} \in \mathbb{R}^{T\times(2n-1)p}$, each row consists of an input rollout $u^{(i)} \in \mathbb{R}^{(2n-1)p}$, and the scaled $\boldsymbol{U} = \bar{\boldsymbol{U}}K^{-1}$ has i.i.d Gaussian entries. Let* $\textbf{snr} = \mathbb{E}[\|u\|^2/n]/\,\mathbb{E}[\|z\|^2]$ *and $\sigma = 1/\sqrt{\textbf{snr}}$. Let $\lambda = \sigma\sqrt{\frac{pn}{T}}\log(n)$. Then, the problem (5) returns $\hat{h}$ such that*

$$\frac{\|\hat{h}-h\|_2}{\sqrt{2}} \le \|\mathcal{H}(\hat{h}-h)\| \lesssim \begin{cases} \sqrt{\frac{np}{\textbf{snr}\times T}}\log(n) & \text{if} \quad T \gtrsim \min(R^2, n) \\ \sqrt{\frac{Rnp}{\textbf{snr}\times T}}\log(n) & \text{if} \quad R \lesssim T \lesssim \min(R^2, n). \end{cases} \tag{7}$$

3. Due to the limitation of space, we refer readers to Sun et al. (2020) for the proofs.

Theorem 1 jointly bounds the impulse response and Hankel spectral errors of the system under mild conditions. We highlight the improvements that our bounds provide: (1) When the system is low order, the sample complexity T is logarithmic in n and improves upon the $O(n)$ bound of the least-squares algorithm. (2) The error rate with respect to the system parameters n, R, T is same as Oymak and Ozay (2018), Sarkar et al. (2019) and Tu et al. (2017) (e.g. compare to Theorem 4).

The regularized method also has the intrinsic advantage that it does not require knowledge of the rank or the singular values of the Hankel matrix beforehand. Numerical experiments on real data in Section 4 demonstrate the performance and robustness of the regularized method.

The theorem above follows by combining statistical analysis with a more general deterministic result (Theorem 2). We will state this result in terms of a restricted singular value (RSV) condition. While RSV is a common condition in sparse estimation literature, our analysis requires introducing a spectral norm variation of RSV. Given a matrix M spectral RSV over a set S is defined as follows:

$$\|M\|_S = \max_{v \in S, v \neq 0} \|\mathcal{G}(Mv)\| / \|\mathcal{G}(v)\|.$$

Theorem 2 *Consider the problem (6) in the MISO setting, where $\boldsymbol{U} \in \mathbb{R}^{T \times (2n-1)p}$. Let β denote the (weighted) impulse response of the true system which has order R, i.e., $\mathrm{rank}(\mathcal{G}(\beta)) = R$, and let $y = \boldsymbol{U}\beta + \xi$ be the measured output, where ξ is the measurement noise. Finally, denote the minimizer of (6) by $\hat{\beta}$. Define*

$$\mathcal{J}(\beta) := \left\{ v \;\middle|\; \langle v, \partial(\frac{1}{2}\|\boldsymbol{U}\beta - y\|_2^2 + \lambda \|\mathcal{G}(\beta)\|_*) \rangle \leq 0 \right\}, \quad \Gamma := \|I - \boldsymbol{U}^\top \boldsymbol{U}\|_{\mathcal{J}(\beta)},$$

where $\mathcal{J}(\beta)$ is the normal cone at β, and Γ is the spectral RSV. If $\Gamma < 1$, $\hat{\beta}$ satisfies

$$\|\mathcal{G}(\hat{\beta} - \beta)\| \leq \frac{\|\mathcal{G}(\boldsymbol{U}^\top \xi)\| + \lambda}{1 - \Gamma}.$$

This theorem determines the generic conditions on the measurements $\boldsymbol{U}$ to ensure successful system identification. As future work, it would be desirable to extend our results to a wider range of measurement models.

3.2. Least-Squares Bounds

Next we consider the least-squares estimator given measurements $y = \bar{\boldsymbol{U}}h + z$. We consider the MIMO setup where $y \in \mathbb{R}^{T \times m}$ and $h \in \mathbb{R}^{(2n-1)p \times m}$. This is obtained by setting $\lambda = 0$ in (5) hence the estimator is given via the pseudo-inverse

$$\hat{h} := h + \bar{\boldsymbol{U}}^\dagger z = \min_{h'} \frac{1}{2}\|\bar{\boldsymbol{U}}h' - y\|_F^2. \tag{8}$$

The following theorem characterizes the spectral norm bound in terms of discrete Fourier transform.

Theorem 3 *Denote the discrete Fourier transform matrix by F. Denote $z_{(i)} \in \mathbb{R}^T, i = 1, ..., m$ as the noise that corresponds to the i'th coordinate of the output. The solution $\hat{h}$ of (8) obeys*

$$\|\hat{h} - h\|_F \leq \|z\|_F / \sigma_{\min}(\bar{\boldsymbol{U}})$$
$$\|\mathcal{H}(\hat{h} - h)\| \leq \left\| \left[\|F\bar{\boldsymbol{U}}^\dagger z_{(1)}\|_\infty, ..., \|F\bar{\boldsymbol{U}}^\dagger z_{(m)}\|_\infty \right] \right\|.$$

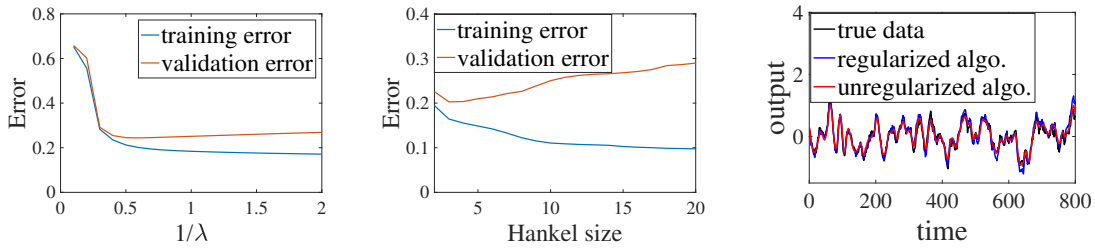

Figure 2: System identification for CD player arm data. Training data size = 200 and validation data size = 600. The first two figures are the training/validation errors of varying λ in regularized algorithm ($n = 10$), and training/validation errors of varying Hankel size n in unregularized algorithm. The last figure is the output trajectory of the true system and the recovered systems (best validation chosen for each).

The next theorem bounds the error when inputs and noise are randomly generated.

Theorem 4 *Denote the solution to (8) as $\hat{h}$. Let $\bar{\boldsymbol{U}} \in \mathbb{R}^{T\times(2n-1)p}$ be input matrix obtained from multiple rollouts, with i.i.d. standard normal entries, $y \in \mathbb{R}^{T\times m}$ be the corresponding outputs and $z \in \mathbb{R}^{T\times m}$ be the noise matrix with i.i.d. $\mathcal{N}(0, \sigma_z^2)$ entries. Then the spectral norm error obeys* $\|\mathcal{H}(\hat{h} - h)\| \lesssim \sigma_z \sqrt{\frac{mnp}{T}} \log(np)$.

This theorem improves the spectral norm bound compared to Oymak and Ozay (2018) which naively bounds the spectral norm in terms of IR error using the right-hand side of (3). Instead, we show that spectral error is same as the IR error up to a log factor (when there is only output noise). Our bound also loses a log factor compared with Sarkar et al. (2019) however is applicable with much fewer samples ($O(n)$ vs $O(n^2)$). We remark that $O(\sigma_z\sqrt{np/T})$ is a tight lower bound for $\|\mathcal{H}(h - \hat{h})\|$ as well as $\|h - \hat{h}\|$ (Oymak and Ozay (2018); Arias-Castro et al. (2012); Djehiche et al. (2019)).

The proofs of the theorems above are provided in Sun et al. (2020). As a proof sketch, we first use the fact that the spectral norm of a circulant matrix is the infinity norm of its Fourier transform. To conclude with Theorem 4, we develop probabilistic bounds on the spectrum of the Hankel error matrix which is circulant.

4. Experiments

Out experiment uses the DaISy dataset De Moor et al. (1997), where a known input signal (not random) is applied and the resulting noisy output trajectory is measured. Consider the following input matrix $\bar{\boldsymbol{U}}$ and the corresponding measurements y

$$\bar{\boldsymbol{U}} = \begin{bmatrix} u_{2n-1}^\top & u_{2n-2}^\top & \dots & u_1^\top \\ u_{2n}^\top & u_{2n-1}^\top & \dots & u_2^\top \\ \dots & & & \\ u_{2n+T-2}^\top & u_{2n+T-3}^\top & \dots & u_T^\top \end{bmatrix} \quad \text{and} \quad y = [y_{2n}, \dots, y_{2n+T-1}]^\top. \tag{9}$$

We consider the optimization problem (5) using single trajectory data. While the input model doesn't satisfy the assumptions of Theorem 2, experiments will demonstrate the advantage of regularization in terms of sample complexity, singular value gap and ease of tuning.

In our experiments, we find that with fixed Hankel size, empirically Oymak and Ozay (2018) performs better than Sarkar et al. (2019), hence we compare our method with the approach of Oymak

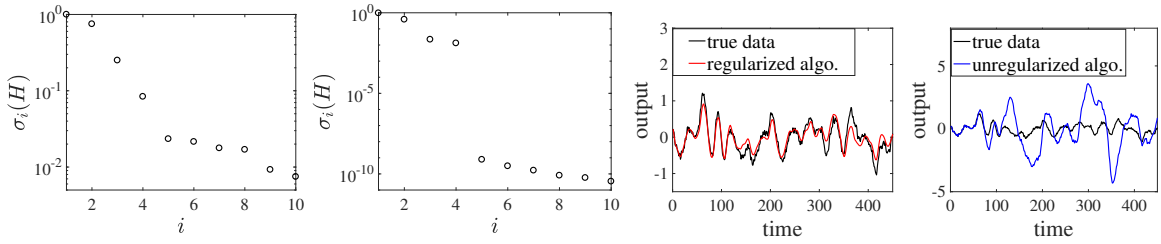

Figure 3: The left two figures: CD player arm data, singular values of the *unregularized* and *regularized* Hankel. The right two figures: Recovery by *regularized* and *unregularized* algorithms when Hankel matrix is 10×10. Training size is 50 and validation size is 400.

and Ozay (2018). When necessary, to select the system order in Oymak and Ozay (2018), we simply keep running the estimated system after time $2n + T - 1$, compare predictions with the true outputs, and choose the order with the smallest validation error.

With enough data for unregularized version, the algorithms perform well in both cases. The first two figures in Figure 2 show the training and validation error. The tuning parameters are weight λ and Hankel size n for the regularized and unregularized problems respectively. This step is to find the best system order by choosing the tuning parameters with the smallest validation error. The third figure in Figure 2 plots the training and validation sequence from dataset and two algorithms. We see that with sufficient sample size, the system can be recovered well. However, the validation error of regularized algorithm is more flat and λ is easier to tune compared to n.

The first two figures in Fig. 3 show that the Hankel spectrums of the two algorithms have a notable difference: The system recovered by the regularized algorithm is low-order and has larger singular value gap. The last two figures in Fig. 3 show the advantage of regularization with much better validation performance. As expected from our theory, the difference is most visible in small sample size (this experiment uses 50 training samples). When the number of observations T is small, regularization still returns a solution close to the true system while least-squares cannot recover the system properly. Further experiments on synthetic systems are provided in (Sun et al., 2020, Sec 2).

5. Future directions

This paper established new sample complexity and estimation error bounds for system identification. We showed that nuclear norm penalization works well with small sample size regardless of the mis-specification in the problem (i.e. fitting impulse response with a much larger length rather than the true order). For least-squares we provide the first guarantee that is optimal in sample complexity and the Hankel spectral norm error. These results can be refined in several directions. In the proof of Theorem 2, we use a weighted version of the Hankel operator. We expect that directly computing the Gaussian width of the original Hankel operator will lead to improvements. It would also be interesting to extend the results to account for single trajectory analysis or process noise. In both cases, an accurate analysis of the regularized problem would lead to new algorithmic insights.

Acknowledgements

Y.S. and M.F. are supported in part by the grants NSF CCF-1740551, NSF DMS-1839371, and DARPA Lagrange FA8650-18-2-7836. S.O. is supported by the NSF grant CNS-1932254.

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26, 2011.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011.

Ery Arias-Castro, Emmanuel J Candes, and Mark A Davenport. On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory*, 59(1):472–481, 2012.

Niclas Blomberg. *On nuclear norm minimization in system identification*. PhD thesis, KTH Royal Institute of Technology, 2016.

Niclas Blomberg, Cristian R Rojas, and Bo Wahlberg. Approximate regularization paths for nuclear norm minimization using singular value bounds–with implementation and extended appendix. *arXiv preprint arXiv:1504.05208*, 2015.

Jian-Feng Cai, Xiaobo Qu, Weiyu Xu, and Gui-Bo Ye. Robust recovery of complex exponential signals from random gaussian projections via low rank hankel matrix reconstruction. *Applied and computational harmonic analysis*, 41(2):470–490, 2016.

Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6): 925–936, 2010.

Bart De Moor, Peter De Gersem, Bart De Schutter, and Wouter Favoreel. Daisy: A database for identification of systems. *JOURNAL A*, 38:4–5, 1997.

Tao Ding, Mario Sznaier, and Octavia I Camps. A rank minimization approach to video inpainting. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

Boualem Djehiche, Othmane Mazhar, and Cristian R Rojas. Finite impulse response models: A non-asymptotic analysis of the least squares estimator. *arXiv preprint arXiv:1911.12794*, 2019.

Michael Elad, Peyman Milanfar, and Gene H Golub. Shape from moments-an estimation theory perspective. *IEEE Transactions on Signal Processing*, 52(7):1814–1829, 2004.

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.

Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Spectral filtering for general linear dynamical systems. In *Advances in Neural Information Processing Systems*, pages 4634–4643, 2018.

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

Zhang Liu, Anders Hansson, and Lieven Vandenberghe. Nuclear norm system identification with missing inputs and outputs. *Systems & Control Letters*, 62(8):605–612, 2013.

Ivan Markovsky. How effective is the nuclear norm heuristic in solving data approximation problems? *IFAC Proceedings Volumes*, 45(16):316–321, 2012.

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

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Ricardo S Sanchez-Pena and Mario Sznaier. *Robust systems theory and applications*. Wiley-Interscience, 1998.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. *arXiv preprint arXiv:1812.01251*, 2019.

Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. Finite-time system identification for partially observed lti systems of unknown order. *arXiv preprint arXiv:1902.01848*, 2019.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473, 2018.

Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning linear dynamical systems with semi-parametric least squares. *arXiv preprint arXiv:1902.00768*, 2019.

Yue Sun, Samet Oymak, and Maryam Fazel. Finite sample system identification: Optimal rates and the role of regularization. https://github.com/sunyue93/Finite-Sample-System-Identification/blob/master/supp.pdf, 2020.

Anastasios Tsiamis and George J Pappas. Finite sample analysis of stochastic system identification. *arXiv preprint arXiv:1903.09122*, 2019.

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.

Bo Wahlberg and Cristian Rojas. Matrix rank optimization problems in system identification via nuclear norm mimization. In *Tutorial) at the European Control Conference*, 2013.