1. Extended introduction

System identification is an important area in control theory, and 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. In this work, we focus on estimating a linear time-invariant system. Given a system with order R, we parameterize it via the state-space representation

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

Here $x_t \in \mathbb{R}^R$ is the state, $u_t \in \mathbb{R}^p$ is the input, $y_t \in \mathbb{R}^m$ is output, $w_t \in \mathbb{R}^R$ and $z_t \in \mathbb{R}^m$ are the state and output noise, and $A \in \mathbb{R}^{R\times R}$, $B \in \mathbb{R}^{R\times p}$, $C \in \mathbb{R}^{m\times R}$, $D \in \mathbb{R}^{m\times p}$ are the system parameters. The system identification problem is to fit the system parameters based on system input u_t and measurements y_t. When $y_t = x_t$, we can directly observe the state at all times, otherwise we only get to see the input and output which provides partial state observations. For instance, for a single output system ($m = 1$), output carries much less information compared to the full state observation making the problem challenging.

A notable line of work provides statistical bounds for system identification using limited number of *full state observations* obtained from a single system trajectory Abbasi-Yadkori and Szepesvári (2011); Simchowitz et al. (2018); Sarkar and Rakhlin (2019). These results make use of random inputs and the core approach is essentially using the least-squares estimator and then adapting the self-normalized martingale bounds from Abbasi-Yadkori et al. (2011).

For hidden-state system of (1) Markov parameters uniquely identify the end-to-end behavior of the system, where the output is determined by input. The Markov parameters are essentially the impulse response terms and are given by the matrices D and $CA^iB \in \mathbb{R}^{T\times p}$ for $i = 0, 1, ...$ The impulse response is given by the operator

$$h = [D, CB, CAB, CA^2B, ...]^T. \tag{2}$$

Grabbing the first $2n-1$ elements of h, we define the Hankel map $\mathcal{H} : \mathbb{R}^{m\times(2n-1)p} \to \mathbb{R}^{mn\times pn}$

$$H := \mathcal{H}(h) = \begin{bmatrix} h_1 & h_2 & ... & h_n \\ h_2 & h_3 & ... & h_{n+1} \\ ... & & & \\ h_n & h_{n+1} & ... & h_{2n-1} \end{bmatrix} \tag{3}$$

If R is the system order and $n \geq R$, the Hankel matrix H is rank R regardless of n. Hence a practically interesting scenario is when the order R is not exactly known in advance and n 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. The notion of simplicity of a system by low-order condition (i.e., low-rank Hankel matrix) is assumed in a wide range of applications, including signal recovery of sum of complex exponentials Cai et al. (2016); Xu et al. (2018) shape from moments estimation in tomography and geophysical inversion Elad et al. (2004), video inpainting Ding et al. (2007), etc.

The traditional unregularized methods include Cadzow approach Cadzow (1988); Gillard (2010), matrix pencil method Sarkar and Pereira (1995), Ho-Kalman approach Ho and

Kálmán (1966) and the subspace method raised in Van Overschee and De Moor (2012), further modified as frequency domain subspace method in McKelvey et al. (1996) when the inputs are single frequency signals. After obtaining an (noisy) estimate of impulse response, the algorithms reduce the rank of Hankel matrix or the order of the system impulse response. It is known that the output of the nuclear norm regularized problems (with proper penalty parameter choice) is usually low rank (if the true matrix is low rank), and it lowers sample complexity as well. The regularized methods that directly modify from subspace method are Hansson et al. (2012); Verhaegen and Hansson (2016) and Smith (2014) (frequency space), whose algorithms run nuclear norm regularization on top of it. Liu et al. (2013); Fazel et al. (2013) propose a slightly different algorithms which regress low rank matrix of output Hankel, both adding a Hankel nuclear norm regularization. Grossmann et al. (2009) specifies the regime when not all output data is collected, and runs Hankel nuclear norm regularization. Ayazoglu and Sznaier (2012) proposes a fast algorithm on solving the regularization algorithm. All above regularization works emphasize on optimization algorithms and have no statistical bounds, and more recently Cai et al. (2016) theoretically proves that a low order SISO system from multi-trajectory input-outputs can be recovered by this approach. Blomberg et al. (2015) analyses a more generic regularization problem with less concrete bound for system identification specifically, and they proposed Hankel matrix nuclear norm regularization as an example. Blomberg (2016) gives a thorough analysis on Hankel nuclear norm regularization applied in system identification, including discussion on proper error metrics, role of rank/system order in formulating the problem, implementable algorithm and selection of tuning parameters.

Alternatively, least-squares can be used to recover the Markov parameters and reconstruct A, B, C, D from Hankel matrix via Ho-Kalman algorithm (Ho and Kálmán (1966)). To identify a stable system using single rollout measurements, Oymak and Ozay (2018) regresses Markov parameter matrix h and Sarkar et al. (2019) regresses Hankel matrix via least-squares. The latter provides optimal Hankel spectral norm error rates however results in suboptimal sample complexity. Related work by Tu et al. (2017) analyzes SISO systems with specific inputs and provides sharp rates in $\mathcal{H}_\infty$ norm. These works assume (roughly) known system order. For Ho-Kalman approach, one can truncate the small singular values to get a low-rank Hankel estimate. Sarkar et al. (2019) employs this strategy and find the system order by back testing. There are several interesting generalizations of least square approaches with non-asymptotic guarantees. Hazan et al. (2018) and Simchowitz et al. (2019) introduced filtering strategies on top of least square to identify stable systems, which are based on sampling in frequency domain. With the filter, the prediction model regresses future output on both previous input and output, and predicts a marginally stable system $1 - \rho(A) \to 0$ as well. They do not recover system parameters such as impulse response, thus we don't compare it with theoretical bounds of regularized or least square approaches, such as our work and Cai et al. (2016), Oymak and Ozay (2018), Sarkar et al. (2019) and Tu et al. (2017). We believe if one truncates the impulse response as FIR when fitting the model, the algorithm always requires a decay in impulse response. Tsiamis and Pappas (2019) gives non-asymptotic analysis for learning a Kalman filter system, which can also be applied to auto-regressive setting. As an extension, Dean et al. (2019) and Mania et al. (2019) applies system identification guarantee for further robust control, and Agarwal et al. (2019) do

online control and regret analysis in adversarial setting, whose algorithm directly learns the policy end-to-end.

This work studies the sample complexity and estimation errors for nuclear norm regularized estimators as comparison of least-squares. Oymak and Ozay (2018) and Sarkar et al. (2019) recover the system from single rollout/trajectory of input, where as 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 used in the problem formulation.

We motivate nuclear norm regularization by experiment with DaISy dataset. In Section **??**, we compare regularized estimator with the least square based algorithm in Oymak and Ozay (2018). We observe that, regularization outperforms least square method in (1) when the system order is unknown, it's easy to tune the weight of the regularizer than solving least square problem of different size many times; (2) more robust with respect to the size and condition number of the input matrix, ending up in smaller error when the sample size is small and no worse error when we have enough observations; (3) approximate low rank solution, i.e., clear gap of singular value of Hankel matrix, so that one can identify the true system order from the recovered system. We also compare Oymak and Ozay (2018) and Sarkar et al. (2019) by experiment, and see that the error of Oymak and Ozay (2018) is much smaller than Sarkar et al. (2019), indicating that we have compared with the right least square solver.

With this, we list our contributions below and contrast with recent work.

• **Nuclear Norm Regularization:** For multi-input/single-output (MISO) systems, we establish optimal sample complexity bounds for the nuclear norm regularized system identification problem, i.e., required sample size grows as $O(pR\log^2 n)$ (based on analysis of SISO system in Cai et al. (2016)). Notably, this bound grows logarithmically with n. We also establish error rates in Hankel spectral norm which are optimal when sample complexity $T \sim O(pR^2\log^2 n)$ and suboptimal in sample complexity at most by a factor of R, whereas the approach of Cai et al. (2016) only works for impulse response Frobenius norm and suffers suboptimality when translated into Hankel spectral norm.[1]

As the comparison of related work in regularization regime, Cai et al. (2016) requires multiple rollout measurements and identifies a SISO system with sample complexity $T \sim O(R\log^2 n)$. We get an easy extension in MISO by modifying the Gaussian width calculation and get $T \sim O(pR\log^2 n)$. In terms of error rate, Cai et al. (2016) analyses Frobenius norm error of impulse response, whereas we study the error rate of spectrum of Hankel matrix, which is $\sqrt{R}$ better than Frobenius norm bound.

The reason that we bound spectral norm error of Hankel matrix recovery is that, Hankel matrix is closely related to $\mathcal{H}_\infty$ norm of a linear system. More importantly, from system identification point of view, people often use Ho-Kalman procedure Ho and Kálmán (1966) to recover A, B, C matrix. Denote $\|\cdot\|, \|\cdot\|_*, \|\cdot\|_F$ as spectral norm, nuclear norm and Frobenius norm when applied on a matrix, and we quote the following theorem.

1. When $T < n$, we lose a factor R in sample complexity, i.e., the optimal sample complexity should be $O(\min\{n, pR\log^2 n\})$. This appears to be an artifact of the proof technique. We include a discussion in Sec **??**.

Theorem 1 *Oymak and Ozay (2018) Suppose $h \in \mathbb{R}^{2n-1}$ is the impulse response associated with linear dynamics*

$$y_t = Cx_t + z_t \quad \text{and} \quad x_{t+1} = Ax_t + Bu_t$$

Then, for some unitary T, Ho-Kalman procedure returns impulse response estimation $\hat{h}$ and system parameter estimation $\hat{A}, \hat{B}, \hat{C}$ such that

$$\|C - \hat{C}T\|_F, \ \|B - T^*\hat{C}\|_F, \ \|A - T^*\hat{A}T\|_F \lesssim \sqrt{n}\|\mathcal{H}(h) - \mathcal{H}(\hat{h})\|$$

The theorem suggests that, the error of recovering A, B, C matrix of the system is closely related to spectral norm error of Hankel matrix, instead of (weighted or not) Frobenius norm error of impulse response suggested in Cai et al. (2016).

• **Least-Squares Estimator:** It is fairly straightforward to show that least-squares estimator regresses Markov parameters h when $T \gtrsim np$ and $n \gtrsim R$ (c.f. Oymak and Ozay (2018)). However establishing the spectal error rates is more challenging. For multi-input/multi-output (MIMO) systems we establish optimal spectral error bound on the Hankel matrix. This bound improves over the result Oymak and Ozay (2018), which bounds Hankel spectral error by Frobenius norm and suffers a $O(\sqrt{n})$ suboptimality. Sarkar et al. (2019) and Tu et al. (2017) also provide optimal spectral norm error, however their sample complexities are suboptimal as they require minimum of $O(n^2)$ measurements rather than $O(n)$. In terms of error rate in Hankel spectral norm, our result is optimal compared to Oymak and Ozay (2018) and Sarkar et al. (2019)

• **Understanding the practical algorithmic performance:** An important contribution of this work is clarifying which algorithm really works in practice. Regularized algorithm requires nontrivial random matrix analysis and we are unable to get a better theoretical bound than correctly tuned least square. However the experiments show that regularized algorithm has empirical benefits in sample complexity, error, and smoothness of training/validation curve and also demonstrate that regularized algorithm is easier to tune (flexible to hyperparameter choice). Another experiment compares two least-squares approaches in Oymak and Ozay (2018) and Sarkar et al. (2019) and finds that the former performs substantially better.

Let the Hankel matrix be of size $n \times n$, the number of output observation is T. Denote the spectral norm of A as $\rho < 1$, the variance of i.i.d. input and output noise are σ_u and σ_z. A summary of the existing results is the following:

1. Oymak and Ozay (2018) requires the system to be strictly stable, and the Frobenius norm error of impulse response of length $2n-1$ is $O((\frac{\sigma_z}{\sigma_u} + \frac{\rho^n}{1-\rho})\sqrt{\frac{n}{T}})$ from $O(n \cdot \text{polylog}(n))$ output observations. The $O(\sqrt{\frac{n}{T}})$ dependence on n and T are optimal (as finite impulse case shown in Djehiche et al. (2019)). They did not propose a Hankel spectral norm error with the same rate. We proved that the Hankel spectral norm error is $O(\sqrt{n/T}\log(n))$.

2. Sarkar et al. (2019) proves that they can achieve a Hankel spectral norm error on the order of $O(\frac{1}{1-\rho}(1+\frac{\sigma_z}{\sigma_u})\sqrt{\frac{n}{T}})$, but requires $O(n^2)$ observations. The sample complexity is n times larger than Oymak and Ozay (2018), and in noiseless case there is a $O(\frac{1}{1-\rho}\sqrt{\frac{n}{T}})$ constant overhead, compared to $O(\frac{\rho^n}{1-\rho}\sqrt{\frac{n}{T}})$ in Oymak and Ozay (2018). This was reflected in both experiments of noiseless and noisy settings in Section **??**. They

also include steps of trying different Hankel size and choose the size/system order by validation.

3. Tu et al. (2017) gives SISO analysis for regressing from independent simulations of impulse response or single frequency inputs. Different from sample complexity defined in our context, they collect all outputs in one rollout. The bound does not distinguish between n and R. They prove that with $T = O(n)$ rollouts and $O(nT)$ output observations. Their algorithm can get $\mathcal{H}_\infty$ error bound $O(\sqrt{\log n/T})$.

Table 1 compare sample complexity and error bounds among related least square works. We compare four related works with sample complexity and error bound under Gaussian noise: our bound of Hankel nuclear norm regularized method, our bound of unregularized least square method, Oymak and Ozay (2018) and Sarkar et al. (2019). To make the fair comparison, we use the same SNR, which is defined as the average power of input divided by power of noise.[2] The takeaways are the following.

- With regularization, our sample complexity is better when the same error rate is same as best least square work.
- Our least square Hankel spectral norm error bound improves upon Oymak and Ozay (2018), and matches the $\sqrt{n}$ rate as in Sarkar et al. (2019).
- Oymak and Ozay (2018) beats Sarkar et al. (2019) with an additive $O(\sqrt{n/T})$ error. The better error is also revealed in our experiments.

Table 1: Related least square work. Let dimension of Hankel be n, system order be R and number of samples be T. Noise level $\sigma = 1/\sqrt{\mathbf{snr}}$. LS-IR and LS-Hankel stands for least square reguression on impulse response and Hankel matrix. All numbers are orderwise (up to constant 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 (11)	$\sigma\sqrt{n/T}$	$\sigma\sqrt{n/T}$	$(1+\sigma)\sqrt{n/T}$
Hankel Spectral Error	see (11)	$\sigma\sqrt{n/T}$	$\sigma n/\sqrt{T}$	$(1+\sigma)\sqrt{n/T}$

2. More experiments

First we generate synthetic data and compare the performance of Oymak and Ozay (2018) and Sarkar et al. (2019) in Figure 1. We can see that, due to the constant overhead $O(\frac{1}{1-\rho}\sqrt{n/T})$ in Sarkar et al. (2019) algorithm, the resulted error is larger than Oymak and

2. The distributions of input in different work are different. In Oymak and Ozay (2018) and Sarkar et al. (2019), they both use i.i.d. standard Gaussian as input, whereas in this paper and Cai et al. (2016), the input follows the form in (**??**). To make fair comparison among the works, we consider SNR in the error bound, where the power of input is defined as the average power of single inputs.

Ozay (2018). Figure 2 compares them in the setting when output noise exists and Oymak and Ozay (2018) has smaller error as well.

In this subsection, we check Theorem 11 via synthetic experiments, and compare with least square estimator. In the following experiment, we have a fixed strictly stable SISO linear system with order 9, the Hankel size n is initiated as 20 which exceeds the order. The input is multiple rollout, scaled i.i.d Gaussian, which means that we send in the input up to time $2n-1$, and observe the output at the end as an observation, and restart the system. The input satisfies that, after scaling by K^{-1}, $\mathbf{E}(\boldsymbol{U}^T\boldsymbol{U}) = I$, which is the assumption in Theorem 11. The observed output can be noiseless and noisy, and the numbers of observations are 30 (undetermined for least square) and 60 (determined for least square).

We tune the regularized model by training with different weight λ of regularization. To tune the least square model, there are two ways: (1) fix the size of Hankel matrix, and run Ho-Kalman algorithm with different rank truncation, or (2) change the size of Hankel matrix. We pick the model associated with the smallest validation error at the end, and run it on test set. The size of training, validation and test set is $1:3:6$.

2.1. Noiseless, enough observations (Fig 3 and 4)

When the output is noiseless and $T = 60$, we can see that both regularized and least square algorithms do well. When $\lambda \to 0$ in regularization or the size and rank tends to 20 in least square method, it almost perfectly fit the model. The singular values of the estimated Hankel is the same since it is perfect recovery.

2.2. Noisy, enough observations (Fig 5 and 6)

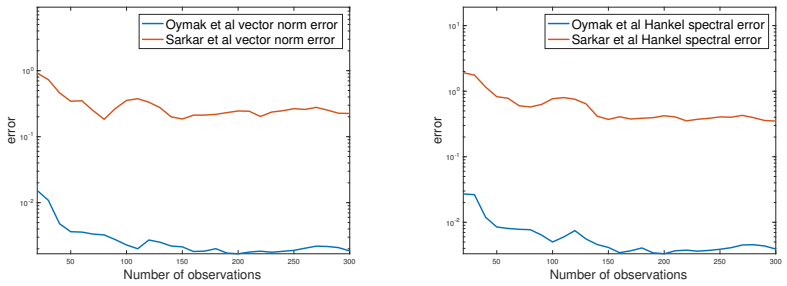
With enough data, when the output is noisy, both regularization and least square do the job well. In Figure 5, we can see that in terms of validation error, there is a best weight λ and Hankel size n, below and above which the validation error both grow. Then we can pick the optimizer associated with those weight, size or rank as our estimation of the system.

2.3. Noiseless, not enough observations (Fig 7 and 8)

Without enough data for least square, even if the output is noiseless, least square is underdetermined, even if we take the solution with the smallest 2 norm in impulse response, it suffers big error on validation and test set. However, the error of regularization remains small and as λ getting small, the error still tends 0. It indicates that, the solution with the least Hankel nuclear norm behaves better than least impulse Frobenius norm in low sample complexity case.

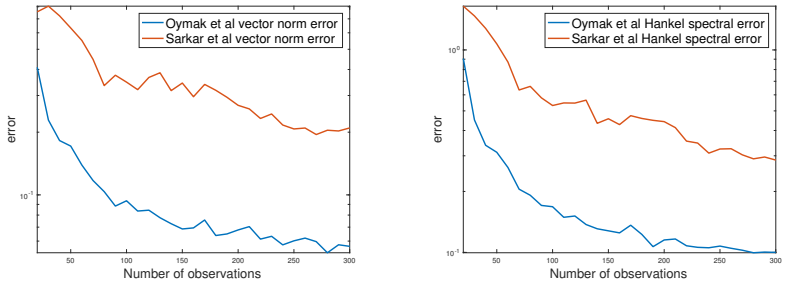
2.4. Noisy, not enough observations (Fig 9 and 10)

Finally not enough data and noisy. We can see that regularized algorithm is robust to noise, where as least square algorithms remain bad.



(a) (b)

Figure 1: Comparison of (a) impulse Frobenius norm (b) Hankel spectral norm error when output is noiseless between Oymak and Ozay (2018) and Sarkar et al. (2019) with synthetic data. System is randomly generated with order 9 and Hankel $H \in \mathbb{R}^{9\times 9}$. Single trajectory and input is i.i.d. Gaussian.



(a) (b)

Figure 2: Comparison of (a) impulse Frobenius norm (b) Hankel spectral norm error when output SNR is 10 between Oymak and Ozay (2018) and Sarkar et al. (2019) with synthetic data. System is randomly generated with order 9 and Hankel $H \in \mathbb{R}^{9\times 9}$. Single trajectory and input is i.i.d. Gaussian.

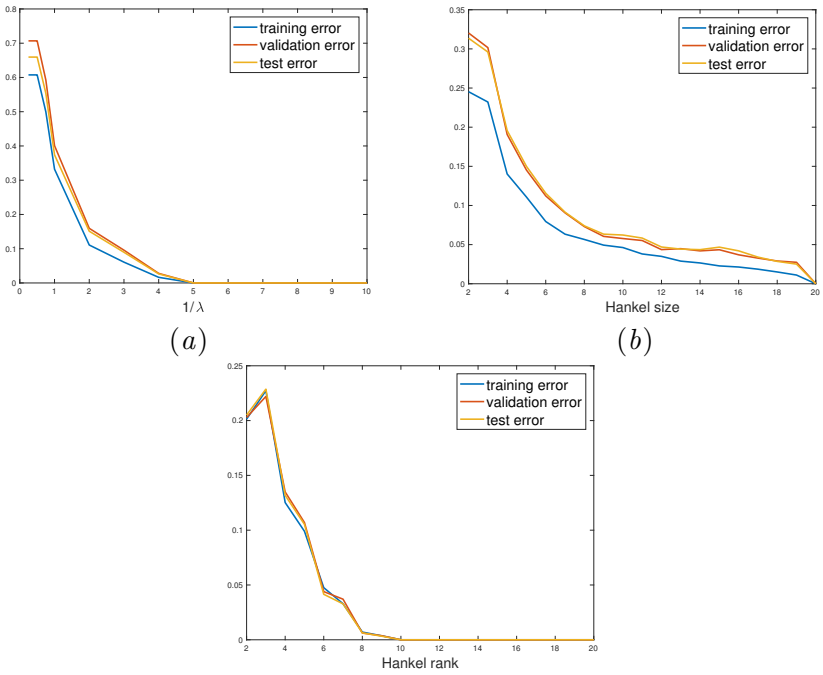


Figure 3: System estimation for synthetic data, noiseless, assuming $n = 20$. Training data size $= 60$. (a) Training and validation error of different λ, (b) Training and validation error of different Hankel size n. (c) Training and validation error of different Hankel rank with same size $n = 20$.

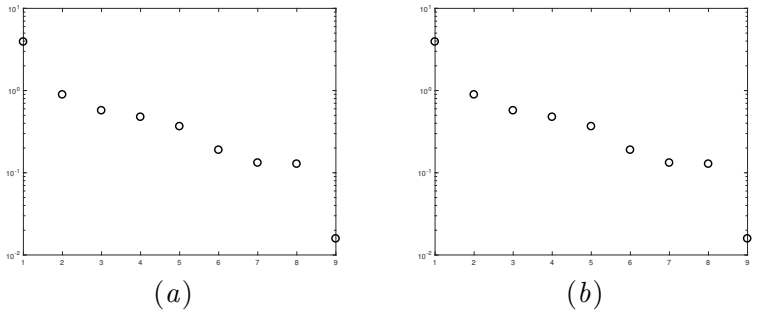


Figure 4: Synthetic, $SNR = 10$, training size is 60, singular value of (a) unregularized Hankel before rank truncation (b) regularized Hankel.

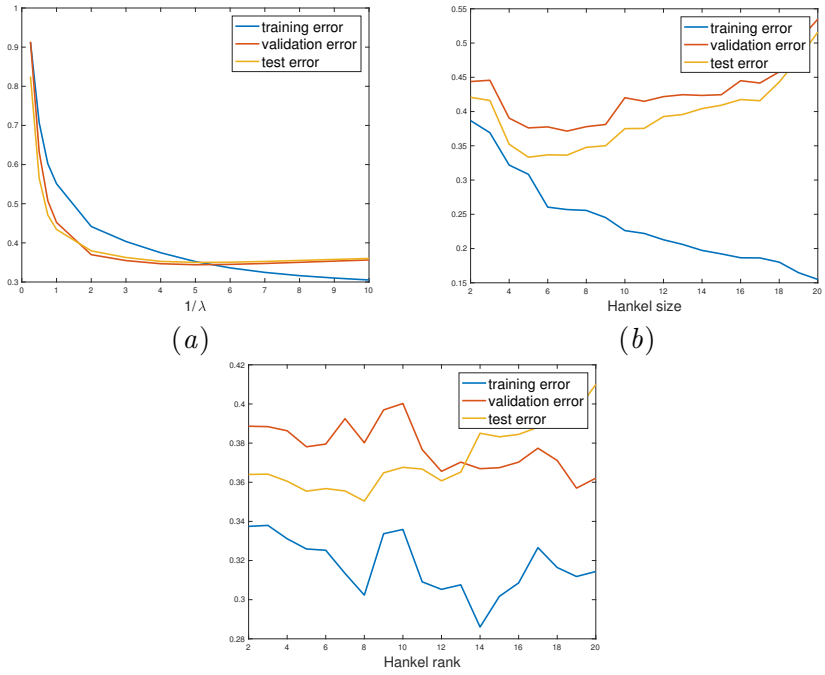


Figure 5: System estimation for synthetic data, SNR = 10, assuming $n = 20$. Training data size = 60. (a) Training and validation error of different λ, (b) Training and validation error of different Hankel size n. (c) Training and validation error of different Hankel rank with same size $n = 20$.

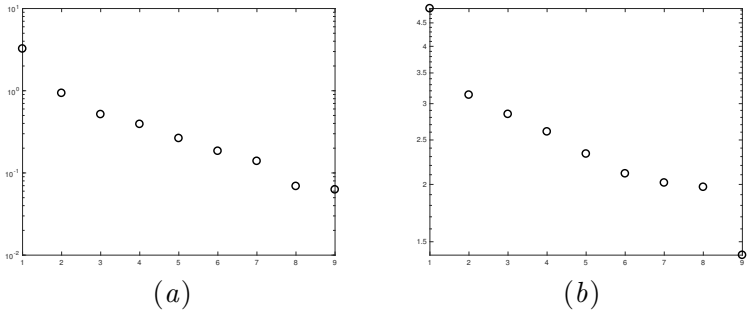


Figure 6: Synthetic, $SNR = 10$, training size is 60, singular value of (a) unregularized Hankel before rank truncation (b) regularized Hankel.

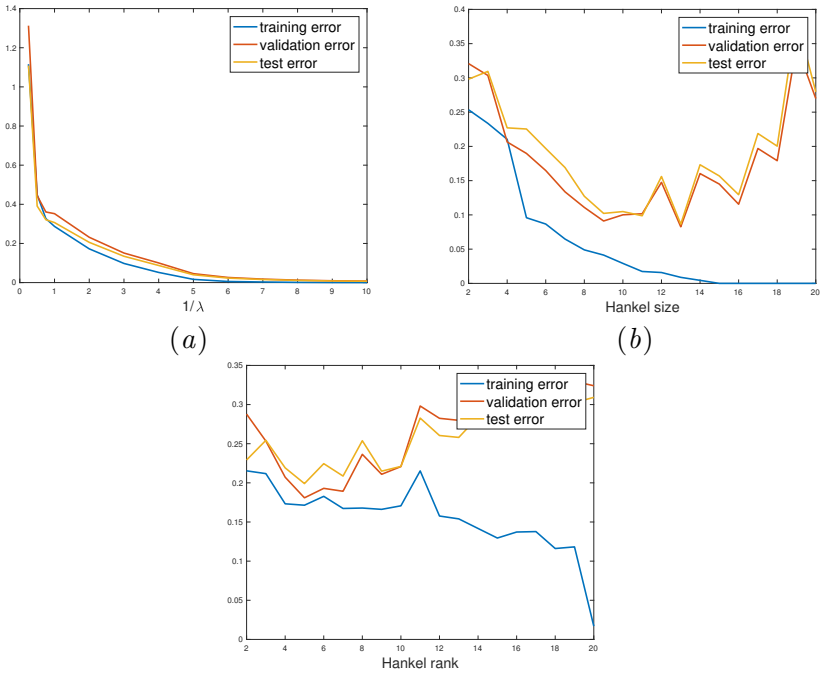


Figure 7: System estimation for synthetic data, noiseless, assuming $n = 20$. Training data size $= 30$. (a) Training and validation error of different λ, (b) Training and validation error of different Hankel size n. (c) Training and validation error of different Hankel rank with same size $n = 20$.

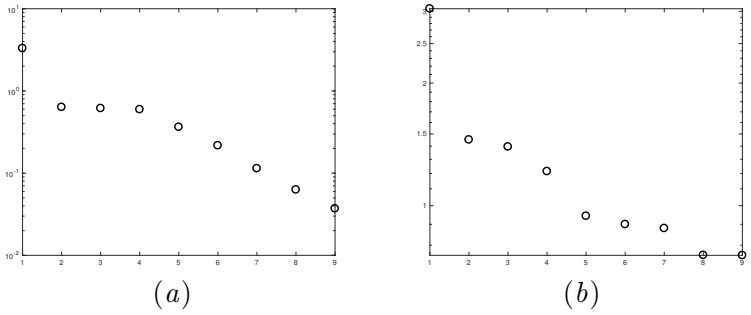


Figure 8: Synthetic, noiseless, training size is 30, singular value of (a) unregularized Hankel before rank truncation (b) regularized Hankel.

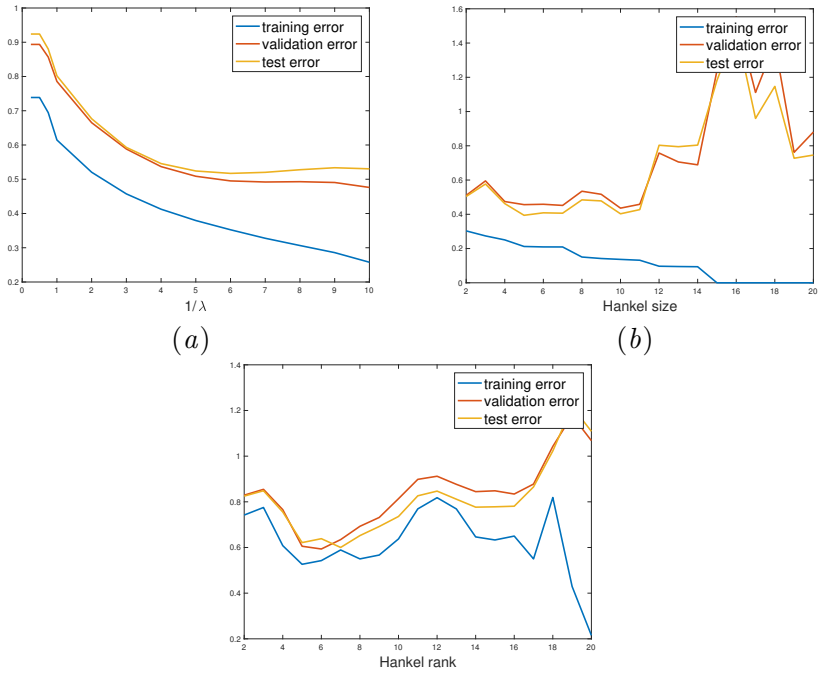


Figure 9: System estimation for synthetic data, SNR = 10, assuming $n = 20$. Training data size = 30. (a) Training and validation error of different λ, (b) Training and validation error of different Hankel size n. (c) Training and validation error of different Hankel rank with same size $n = 20$.

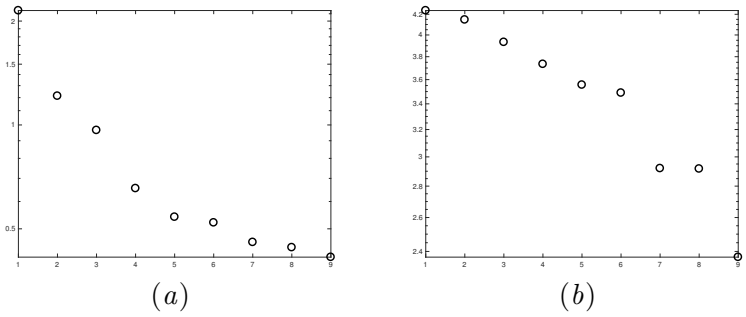


Figure 10: Synthetic, $SNR = 10$, training size is 30, singular value of (a) unregularized Hankel before rank truncation (b) regularized Hankel.

3. Proof of least square spectral norm error

Theorem 2 *Denote the discrete Fourier transform matrix by F. Denote $z_{(i)} \in \mathbb{R}^T, i = 1, ..., m$ as the noise that corresponds to each dimension of output. The solution $\hat{h}$ of*

$$\hat{h} := h + \boldsymbol{U}^\dagger z = \min_{h'} \frac{1}{2}\|\boldsymbol{U}h' - y\|_F^2. \tag{4}$$

obeys

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

Proof First we clarify the notation here. In regularization part, we only consider the MISO system, whereas we can prove the bound for MIMO system as well in least square. Here we assume the input is p dimension and output is m dimension, at each time. For the notation in (4), $\boldsymbol{U} \in \mathbb{R}^{T\times(2n-1)p}$, whose each row is the input in a time interval of length $2n-1$. The impulse response is $h \in \mathbb{R}^{(2n-1)p\times m}$ and output and noise are $y, z \in \mathbb{R}^{T\times m}$, where each column corresponds to one channel of the output. Each row of y is an output observation at a single time point. $z_{(i)} \in \mathbb{R}^T$ is a column of the noise, meaning one channel of the noise contaminating all observations at this channel.

(4) has close form solution and we have $\|\hat{h} - h\| = \|\boldsymbol{U}^\dagger z\| \leq \|z\|/\sigma_{\min}(\boldsymbol{U})$. To get the error bound in Hankel matrix, we can denote $\bar{z} = \boldsymbol{U}^\dagger z = (\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T z$, and

$$H_{\bar{z}} = \begin{bmatrix} \bar{z}_1 & \bar{z}_2 & ... & \bar{z}_{2n-1} \\ \bar{z}_2 & \bar{z}_3 & ... & \bar{z}_1 \\ ... & & & \\ \bar{z}_{2n-1} & \bar{z}_1 & ... & \bar{z}_{2n-2} \end{bmatrix}.$$

If $m = 1$, $\bar{z} \in \mathbb{R}^{(2n-1)p}$ is a vector (Krahmer et al., 2014, Section 4) proves that

$$H_{\bar{z}} = F^{-1}\text{diag}(F\bar{z})F.$$

So the spectral norm error is bounded by $\|\text{diag}(F\bar{z})\|_2 = \|F\bar{z}\|_\infty$.

If $m > 1$, all columns of z are independent, so $H_{\bar{z}}$ can be seen as concatenation of m independent noise matrices where each satisfies the previous derivation. ■

Theorem 3 *Denote the solution to (4) as $\hat{h}$. Let $\boldsymbol{U} \in \mathbb{R}^{T\times(2n-1)p}$ is multiple rollout input, where every entry is i.i.d. Gaussian random variable, y be the corresponding output and z is i.i.d. Gaussian matrix with each entry has mean 0 and variance σ_z, then the spectral norm error is $\|\mathcal{H}(\hat{h} - h)\| \lesssim \sigma_z\sqrt{\frac{mnp}{T}}\log(np)$.*

Proof We use Theorem 2. First let $m = 1$. The covariance of $F\bar{z} = F\boldsymbol{U}^\dagger z$ is $F(\boldsymbol{U}^T\boldsymbol{U})^{-1}F^T$. If $T = \tilde{\Omega}(n)$, it's proven Vershynin (2018) that $\frac{TI}{2} \preceq \boldsymbol{U}^T\boldsymbol{U} \preceq \frac{3TI}{2}$ then $\frac{n}{2T}I \preceq F(\boldsymbol{U}^T\boldsymbol{U})^{-1}F^T \preceq \frac{3n}{2T}I$. So $\|F\bar{z}\|_\infty$ should scale as $O(\sigma_z\sqrt{\frac{n}{T}}\log n)$. So $\|\mathcal{H}(\bar{z})\|_2 \leq \|H_{\bar{z}}\|_2 \leq \|F\bar{z}\|_\infty = O(\sigma_z\sqrt{\frac{n}{T}}\log n)$. If $m > 1$, then by concatenation we simply bound the specral norm by m times MISO case. When $m > 1$, with previous discussion of concatenation, and each submatrix to be concatenated has the same distribution, so the spectral norm error is at most $\sqrt{m}$ times larger. ■

4. Gaussian width of nuclear norm normal cone in MISO

We consider recovering a MISO system impulse response. We first calculate the minimum number of observations needed to recover the system regardless of noise rate, which is a simple extension from SISO case in Cai et al. (2016). This can be seen as the sample complexity requirement in noiseless case. For multi-rollout case, we only observe the output at time $2n-1$, we have

$$y_{2n-1} = \sum_{i=1}^{2n-2} CA^{2n-2-i}Bu_i + Du_{2n-1}. \tag{5}$$

Denote the impulse response by $h \in \mathbb{R}^{p(2n-1)}$, which is a block vector

$$h = \begin{bmatrix} h^{(1)} \\ h^{(2)} \\ \dots \\ h^{(2n-1)} \end{bmatrix}$$

where each block $h^{(i)} \in \mathbb{R}^p$. $\beta \in \mathbb{R}^{p(2n-1)}$ is a weighted version of h, with weights

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

and

$$x^{(i)} = K_i h^{(i)}$$

Define the reweighted Hankel map for the same h by

$$\mathcal{G}(\beta) = \begin{bmatrix} \beta^{(1)}/K_1 & \beta^{(2)}/K_2 & \beta^{(3)}/K_3 & \dots \\ \beta^{(2)}/K_2 & \beta^{(3)}/K_3 & \beta^{(4)}/K_2 & \dots \\ \dots & & & \end{bmatrix}^T \in \mathbb{R}^{n \times pn}$$

and $\mathcal{G}^*$ is the adjoint of $\mathcal{G}$. We define each rollout input $u_1, \dots, u_{2n-1}$ as independent Gaussian vectors with

$$u_i \sim \mathcal{N}(0, K_i^2 \mathbf{I})$$

Now let $\boldsymbol{U} \in \mathbb{R}^{T \times p(2n-1)}$, each entry is iid standard Gaussian. We consider the question

$$\begin{aligned} \min_{\beta'} \quad & \|\mathcal{G}(\beta')\|_* \\ \text{s.t.,} \quad & \|\boldsymbol{U}\beta' - y\|_2 \le \delta \end{aligned} \tag{6}$$

where the norm of overall (state and output) noise is bounded by δ.

Theorem 4 *Cai et al. (2016) Let $\hat{\beta}$ be the true impulse response. If $T = \Omega((\sqrt{pR}\log(n) + \epsilon)^2)$, C is some constant, the solution $\hat{\beta}$ to (6) satisfies $\|\hat{\beta} - \beta\|_2 \le 2\delta/\epsilon$ with probability*

$$1 - \exp\left(-\frac{1}{2}(\sqrt{T-1} - C(\sqrt{pR}\log(n) + \epsilon) - \epsilon)^2\right).$$

Let $\mathcal{I}(\beta)$ be the descent cone of $\|\mathcal{G}(\beta)\|_*$ at β, we have the following lemma:

Lemma 5 *Assume*

$$\min_{z\in\mathcal{I}(\beta)} \frac{\|\boldsymbol{U}z\|_2}{\|z\|_2} \geq \epsilon,$$

then $\|\hat{\beta}-\beta\|_2 \leq 2\delta/\epsilon$.

(Proof omitted) To prove Theorem 4, we only need lower bound LHS with Lemma 5. The following lemma gives the probability that LHS is lower bounded.

Lemma 6 *Define the Gaussian width*

$$w(S) := E_g(\sup_{\gamma\in S} \gamma^T g) \tag{7}$$

where g is standard Gaussian vector of size p. Let $\Phi = \mathcal{I}(\beta)\cap\mathbb{S}$ where $\mathbb{S}$ is unit sphere. We have

$$P(\min_{z\in\Phi} \|\boldsymbol{U}z\|_2 < \epsilon) \leq \exp\left(-\frac{1}{2}(\sqrt{T-1} - w(\Phi) - \epsilon)^2\right). \tag{8}$$

Now we need to study $w(\Phi)$.

Lemma 7 *(Cai et al. (2016) eq. (17)) Let $\mathcal{I}^*(\beta)$ be the dual cone of $\mathcal{I}(\beta)$, then*

$$w(\Phi) \leq E(\min_{\gamma\in\mathcal{I}^*(\beta)} \|g-\gamma\|). \tag{9}$$

Note that $\mathcal{I}^*(\beta)$ is just the cone of subgradient of $\mathcal{G}(\beta)$, so it can be written as

$$\mathcal{I}^*(\beta) = \{\mathcal{G}^*(V_1V_2^T + W) | V_1^T W = 0, WV_2 = 0, \|W\| \leq 1\}$$

where $\mathcal{G}(\beta) = V_1\Sigma V_2^T$ is the SVD of $\mathcal{G}(\beta)$[3]. So

$$\min_{\gamma\in\mathcal{I}^*(\hat{x})} \|g-\gamma\|_2 = \min_{\lambda,W} \|\lambda\mathcal{G}^*(V_1V_2^T+W) - g\|_2.$$

For RHS, we have

$$\begin{aligned}
\|\lambda\mathcal{G}^*(V_1V_2^T+W) - g\|_2 &= \|\lambda\mathcal{G}\mathcal{G}^*(V_1V_2^T+W) - \mathcal{G}(g)\|_F \\
&= \|\lambda(V_1V_2^T+W) - \mathcal{G}(g)\|_F + \|\lambda(I-\mathcal{G}\mathcal{G}^*)(V_1V_2^T+W)\|_F \\
&\leq \|\lambda(V_1V_2^T+W) - \mathcal{G}(g)\|_F.
\end{aligned}$$

Let $\mathcal{P}_W$ be projection operator onto subspace spanned by W, i.e.,

$$\{W | V_1^T W = 0, WV_2 = 0\}$$

3. For simplicity, we only write down real case. Complex case can be seen as a dimension increase by 2 times as in Cai et al. (2016).

and $\mathcal{P}_V$ be projection onto its orthogonal complement. Choose $\lambda = \|\mathcal{P}_W(\mathcal{G}(g))\|$ and $W = \mathcal{P}_W(\mathcal{G}(g))/\lambda$.

$$\begin{aligned}\|\lambda(V_1V_2^T + W) - \mathcal{G}(g)\|_F &= \|\mathcal{G}(g) - \mathcal{P}_W(\mathcal{G}(g)) - \|\mathcal{P}_W(\mathcal{G}(g))\|V_1V_2^T\|_F \\ &\le \|\mathcal{P}_V(\mathcal{G}(g)) - \|\mathcal{P}_W(\mathcal{G}(g))\|V_1V_2^T\|_F \\ &\le \|\mathcal{P}_V(\mathcal{G}(g))\|_F + \|\mathcal{P}_W(\mathcal{G}(g))\|\|V_1V_2^T\|_F \\ &= \|\mathcal{P}_V(\mathcal{G}(g))\|_F + \sqrt{R}\|\mathcal{P}_W(\mathcal{G}(g))\| \\ &= \|\mathcal{P}_V(\mathcal{G}(g))\|_F + \sqrt{R}\|\mathcal{G}(g)\|.\end{aligned}$$

Bound the first term by (note V_1 and V_2 span R dimensional space, so $V_1 \in \mathbb{R}^{n\times R}$ and $V_2 \in \mathbb{R}^{pn\times R}$)

$$\begin{aligned}\|\mathcal{P}_V(\mathcal{G}(g))\|_F &= \|V_1V_1^T\mathcal{G}(g) + (I - V_1V_1^T)\mathcal{G}(g)V_2V_2^T\|_F \\ &\le \|V_1V_1^T\mathcal{G}(g)\|_F + \|\mathcal{G}(g)V_2V_2^T\|_F \\ &\le 2\sqrt{R}\|\mathcal{G}(g)\|.\end{aligned}$$

we get

$$\begin{aligned}w(\Phi) &\le E(\min_{\lambda,W}\|\lambda\mathcal{G}^*(V_1V_2^T + W) - g\|_2) \\ &\le E(\|\lambda\mathcal{G}^*(V_1V_2^T + W) - g\|_2)\big|_{\lambda=\|\mathcal{P}_W(\mathcal{G}(g))\|, W=\mathcal{P}_W(\mathcal{G}(g))/\lambda} \\ &\le 3\sqrt{R}\|\mathcal{G}(g)\|.\end{aligned}$$

We know that, if $p = 1$, then $E\|\mathcal{G}(g)\| = O(\log(n))$. For general p, let

$$g^{(i)} = [g_1^{(i)}, ..., g_p^{(i)}]^T,$$

we rearrange the matrix as

$$\begin{aligned}\bar{\mathcal{G}}(g) &= \begin{bmatrix}\begin{bmatrix} g_1^{(1)} & g_1^{(2)}/\sqrt{2} & ... \\ g_1^{(2)}/\sqrt{2} & g_1^{(3)}/\sqrt{3} & ... \\ ... & & \end{bmatrix} & \begin{bmatrix} g_2^{(1)} & g_2^{(2)}/\sqrt{2} & ... \\ g_2^{(2)}/\sqrt{2} & g_2^{(3)}/\sqrt{3} & ... \\ ... & & \end{bmatrix} & ...\end{bmatrix} \\ &= [G_1, ..., G_p]\end{aligned}$$

where expectation of operator norm of each block is $\log(n)$. Then (note v below also has a block structure $[v^{(1)}; ...; v^{(n)}]$)

$$\begin{aligned}\|\bar{\mathcal{G}}(g)\| &= \max_{u,v}\frac{u^T\bar{\mathcal{G}}(g)v}{\|u\|\|v\|} \\ &= \max_{u,v^1,...,v^p}\sum_{i=1}^p\frac{u^TG_iv^{(i)}}{\|u\|\|v\|} \\ &\le \max_{v^1,...,v^p} O(\log(n))\frac{\sum_{i=1}^p\|v^{(i)}\|}{\sqrt{\sum_{i=1}^p\|v^{(i)}\|^2}} \\ &\le O(\sqrt{p}\log(n)).\end{aligned}$$

And $\|\bar{\mathcal{G}}(g)\| = \|\mathcal{G}(g)\|$. So we have $\|\mathcal{G}(g)\| = \sqrt{p}\log(n)$. So $w(\Phi) = C\sqrt{pR}\log(n)$. Get back to (8), we want the probability be smaller than 1, and we get

$$\sqrt{T-1} - C\sqrt{pR}\log n - \epsilon > 0$$

thus $T = O((\sqrt{pR}\log(n) + \epsilon)^2)$.

Before stepping to the proof of main theorem, we give a different version of Theorem 4. Theorem 4 in Cai et al. (2016) works for the any noise with bounded norm. Here we consider the iid Gaussian noise, and use the result in Oymak et al. (2013), we have the following theorem.

Theorem 8 *Let the system output $y = \boldsymbol{U}\beta + z$ where $\boldsymbol{U}$ entries are iid Gaussian $\mathcal{N}(0, 1/T)$, β is the true system parameter and $z \sim \mathcal{N}(0, \sigma_z^2)$. Then (6) recovers $\hat{\beta}$ with error $\|\hat{\beta} - \beta\|_2 \leq w(\Phi)\|z\|_2/\sqrt{T} \lesssim \sqrt{pR}\sigma_z \log n$ with high probability.*

Remark 9 *Since the power of $\boldsymbol{U}$ is n times of that of $\bar{\boldsymbol{U}}$ and the variance of $\boldsymbol{U}$ is $1/T$, $\sigma_z = \sqrt{n/T}\sigma$, we have $\|\hat{h} - h\|_2 \leq \|\hat{\beta} - \beta\|_2 \lesssim \sqrt{\frac{pnR}{T}}\sigma \log n$.*

5. Proof of main theorem

Theorem 10 *Consider problem*

$$\hat{h} = \arg\min_{h'} \quad \frac{1}{2}\|\bar{\boldsymbol{U}}h' - y\|_F^2 + \lambda\|\mathcal{H}(h')\|_*. \tag{10}$$

in the MISO (multi-input single-output) setting (m=1, p inputs), the system is order R, $\bar{\boldsymbol{U}} \in \mathbb{R}^{T\times(2n-1)p}$, each row consisting an input rollout $u^{(i)} \in \mathbb{R}^{(2n-1)p}$, and the scaled $\boldsymbol{U}$ has i.i.d Gaussian entries. Let $\boldsymbol{snr} = \mathbb{E}[\|u\|^2/n]/\mathbb{E}[\|z\|^2]$ and $\sigma = 1/\sqrt{\boldsymbol{snr}}$. Let $\lambda = \sigma\sqrt{\frac{pn}{T}}\log(n)$, (10) returns $\hat{h}$ such that

$$\|\hat{h} - h\|_2 \lesssim \|\mathcal{H}(\hat{h} - h)\| \lesssim \begin{cases} \sqrt{\frac{np}{\boldsymbol{snr}\times T}}\log(n) & \text{if} \quad T \gtrsim \min(R^2, n) \\ \sqrt{\frac{Rnp}{\boldsymbol{snr}\times T}}\log(n) & \text{if} \quad R \lesssim T \lesssim \min(R^2, n). \end{cases} \tag{11}$$

We will prove the first case of (11). The second case is a direct application of Theorem 8.

Theorem 11 *We study the problem*

$$\min_{\beta'} \quad \frac{1}{2}\|\boldsymbol{U}\hat{\beta}' - y\|^2 + \lambda\|\mathcal{G}(\hat{\beta}')\|_*, \tag{12}$$

in the MISO (multi-input single-output) setting (m=1, p inputs), 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., $\operatorname{rank}(\mathcal{G}(\beta)) = R$, and let $y = \boldsymbol{U}\beta + \xi$ be the measured output, where ξ is the measurement noise. Finally, denote the minimizer of (12) by $\hat{\beta}$. Define

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

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

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

Lemma 12 *Suppose $\xi \sim \mathcal{N}(0, \sigma_\xi I)$, $T \lesssim pR^2 \log^2 n$, and $\boldsymbol{U}$ has iid Gaussian entries with $\mathbf{E}(\boldsymbol{U}^\top \boldsymbol{U}) = 1$. Then, we have that $\mathbf{E}(\Gamma) < 0.5$, and $P(\Gamma < 0.5) \geq 1 - O(R \log n \sqrt{p/T})$. In this case $\|\mathcal{G}(\hat{\beta} - \beta)\| \lesssim \sigma_\xi \sqrt{p} \log n$.*

Remark 13 *To be consistent with the main theorem in the paper, we need to find the relation between σ_ξ and SNR, or σ. We do the following computation: (1) $\mathcal{G}(\hat{\beta} - \beta) = \mathcal{H}(\hat{h} - h)$, so we are bounding the Hankel spectral norm error here; (2) Each column of the input is unit norm, so each input is $\mathcal{N}(0, 1/T)$, and the average power of input is $1/T$; (3) Because of the scaling matrix K, the actual input of $\bar{\boldsymbol{U}}$ is n times the power of entries in $\boldsymbol{U}$. With all above discussion, we have $\sigma_\xi = \sigma\sqrt{n/T}$, which results in $\|\mathcal{G}(\hat{\beta} - \beta)\| \lesssim \sqrt{\frac{np}{T}} \sigma \log n$.*

Proof Now we bound $\|\mathcal{G}(\hat{\beta} - \beta)\|$ by partitioning it to $\|\mathcal{G}(I - \boldsymbol{U}^T\boldsymbol{U})(\hat{\beta} - \beta)\|$ and $\|\mathcal{G}(\boldsymbol{U}^T\boldsymbol{U}(\hat{\beta} - \beta))\|$. We have

$$\begin{aligned} \|\mathcal{G}(I - \boldsymbol{U}^T\boldsymbol{U})(\hat{\beta} - \beta)\| &= \|\mathcal{G}(I - \boldsymbol{U}^T\boldsymbol{U})\mathcal{G}^*\mathcal{G}(\hat{\beta} - \beta)\| \\ &\leq \|\mathcal{G}(I - \boldsymbol{U}^T\boldsymbol{U})\mathcal{G}^*\|_{2,\mathcal{G}\mathcal{J}(\beta)} \|\mathcal{G}(\hat{\beta} - \beta)\| \\ &= \Gamma \|\mathcal{G}(\hat{\beta} - \beta)\|. \end{aligned} \tag{13}$$

And then we also have

$$\begin{aligned} \|\mathcal{G}(\boldsymbol{U}^T\boldsymbol{U}(\hat{\beta} - \beta))\| &= \|\mathcal{G}\boldsymbol{U}^T(\boldsymbol{U}\hat{\beta} - y + \xi)\| \\ &\leq \|\mathcal{G}\boldsymbol{U}^T(\boldsymbol{U}\hat{\beta} - y)\| + \|\mathcal{G}(\boldsymbol{U}^T\xi)\|. \end{aligned}$$

Since $\hat{\beta}$ is the optimizer, we have

$$\boldsymbol{U}^T(\boldsymbol{U}\hat{\beta} - y) + \lambda\mathcal{G}^*(\hat{V}_1\hat{V}_2^T + \hat{W}) = 0,$$

where $\mathcal{G}(\hat{\beta}) = \hat{V}_1\hat{\Sigma}\hat{V}_2^T$ is the SVD of $\mathcal{G}(\hat{\beta})$, $\hat{W} \in \mathbb{R}^{n \times n}$ where $\hat{V}_1^T\hat{W} = 0, \hat{W}\hat{V}_2 = 0, \|\hat{W}\| \leq 1$. We have

$$\|\mathcal{G}(\boldsymbol{U}^T\boldsymbol{U}(\hat{\beta} - \beta))\| \leq \|\mathcal{G}(\boldsymbol{U}^T\xi)\| + \lambda. \tag{14}$$

Combining (13) and (14), we have

$$\begin{aligned} \|\mathcal{G}(\hat{\beta} - \beta)\| &\leq \|\mathcal{G}(I - \boldsymbol{U}^T\boldsymbol{U})(\hat{\beta} - \beta)\| + \|\mathcal{G}(\boldsymbol{U}^T\boldsymbol{U}(\hat{\beta} - \beta))\| \\ &\leq \Gamma\|\mathcal{G}(\hat{\beta} - \beta)\| + \|\mathcal{G}(\boldsymbol{U}^T\xi)\| + \lambda \end{aligned}$$

or equivalently,

$$\|\mathcal{G}(\hat{\beta} - \beta)\| \leq \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\| + \lambda}{1 - \Gamma}, \quad \Gamma = \|\mathcal{G}(I - \boldsymbol{U}^T\boldsymbol{U})\mathcal{G}^*\|_{2,\mathcal{G}\mathcal{J}(\beta)}.$$

Bounding Γ. Denote the SVD of $\mathcal{G}(\beta) = V_1 \Sigma V_2^T$. Denote projection operators $\mathcal{P}_V(M) = V_1 V_1^T M + M V_2 V_2^T - V_1 V_1^T M V_2 V_2^T$ and $\mathcal{P}_W(M) = M - \mathcal{P}_V(M)$. First we prove some side results for later use. From optimality of $\hat{\beta}$, we have

$$\begin{aligned}
& \frac{1}{2}\|y - \boldsymbol{U}\hat{\beta}\|^2 + \lambda\|\mathcal{G}\hat{\beta}\|_* \le \frac{1}{2}\|y - \boldsymbol{U}\beta\|^2 + \lambda\|\mathcal{G}\beta\|_* = \frac{1}{2}\|\xi\|^2 + \lambda\|\mathcal{G}\beta\|_* \\
\Rightarrow \quad & \frac{1}{2}\|\boldsymbol{U}\beta + \xi - \boldsymbol{U}\hat{\beta}\|^2 + \lambda\|\mathcal{G}\hat{\beta}\|_* \le \frac{1}{2}\|\xi\|^2 + \lambda\|\mathcal{G}\beta\|_* \\
\Rightarrow \quad & \frac{1}{2}\|\boldsymbol{U}(\beta - \hat{\beta})\|^2 + \xi^T\boldsymbol{U}(\beta - \hat{\beta}) + \lambda\|\mathcal{G}\hat{\beta}\|_* \le \lambda\|\mathcal{G}\beta\|_* \\
\Rightarrow \quad & \lambda\|\mathcal{G}\hat{\beta}\|_* \le \lambda\|\mathcal{G}\beta\|_* + \xi^T\boldsymbol{U}(\hat{\beta} - \beta) \\
\Rightarrow \quad & \|\mathcal{G}\hat{\beta}\|_* - \|\mathcal{G}\beta\|_* \le \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}\|\mathcal{G}(\hat{\beta} - \beta)\|_* \qquad (15)
\end{aligned}$$

(15) is an important result to note, and following that,

$$\begin{aligned}
& \|\mathcal{G}\hat{\beta}\|_* - \|\mathcal{G}\beta\|_* \le \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}\|\mathcal{G}(\hat{\beta} - \beta)\|_* \\
\Rightarrow \quad & \langle \mathcal{G}(\hat{\beta} - \beta), V_1 V_2^T + W\rangle \le \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}\|\mathcal{G}(\hat{\beta} - \beta)\|_* \\
\Rightarrow \quad & \|\mathcal{P}_W\mathcal{G}(\hat{\beta} - \beta)\|_* \le -\langle \mathcal{G}(\hat{\beta} - \beta), V_1 V_2^T\rangle + \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}\|\mathcal{G}(\hat{\beta} - \beta)\|_* \\
\Rightarrow \quad & \|\mathcal{P}_W\mathcal{G}(\hat{\beta} - \beta)\|_* \le \|\mathcal{P}_V\mathcal{G}(\hat{\beta} - \beta)\|_* + \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}(\|\mathcal{P}_V\mathcal{G}(\hat{\beta} - \beta)\|_* + \|\mathcal{P}_W\mathcal{G}(\hat{\beta} - \beta)\|_*) \\
\Rightarrow \quad & \|\mathcal{P}_W\mathcal{G}(\hat{\beta} - \beta)\|_* \le \frac{1 + \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}}{1 - \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}}\|\mathcal{P}_V\mathcal{G}(\hat{\beta} - \beta)\|_* \qquad (16)
\end{aligned}$$

Let $\boldsymbol{U}$ be iid Gaussian matrix with scaling $\boldsymbol{E}(\boldsymbol{U}^T\boldsymbol{U}) = I$. Here we need to study the Gaussian width of the normal cone $w(\mathcal{J}(\beta))$ of (12). Banerjee et al. (2014) proves that, if (15) is true, and $\lambda \ge 2\|\mathcal{G}(\boldsymbol{U}^T\xi)\|$, then the Gaussian width of this set (intersecting with unit ball) is less than 3 times of Gaussian width of $\{\hat{\beta} : \|\mathcal{G}(\hat{\beta})\|_* \le \|\mathcal{G}(\beta)\|_*\}$, which is $O(\sqrt{R}\log n)$ Cai et al. (2016).
A simple bound is that, let $\delta = \hat{\beta} - \beta$, Γ can be replaced by

$$\max \|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\|/\|\mathcal{G}(\delta)\|$$

subject to $\hat{\beta} \in \mathcal{J}(\beta)$. With (16), we have $\|\mathcal{P}_W\mathcal{G}(\delta)\|_* \le 3\|\mathcal{P}_V\mathcal{G}(\delta)\|_*$.
Denote $\sigma = \|\mathcal{G}(\delta)\|$, we know that $\sigma \ge \max\{\|\mathcal{P}_W\mathcal{G}(\delta)\|, \|\mathcal{P}_V\mathcal{G}(\delta)\|\}$ and $\|\mathcal{P}_V\mathcal{G}(\delta)\| \ge \|\mathcal{P}_V\mathcal{G}(\delta)\|_*/(2R)$. And simple algebra gives that

$$\max_{0<\sigma_i<\sigma, \sum_i \sigma = S} \sum_i \sigma_i^2 \le S\sigma.$$

So let σ_i be singular values of $\mathcal{P}_V\mathcal{G}(\delta)$ or $\mathcal{P}_W\mathcal{G}(\delta)$, and $S = \|\mathcal{P}_V\mathcal{G}(\delta)\|_*$ or $\|\mathcal{P}_W\mathcal{G}(\delta)\|_*$,

$$\frac{\sigma}{\|\mathcal{P}_V\mathcal{G}(\delta)\|_F} \geq \sqrt{\frac{\|\mathcal{P}_V\mathcal{G}(\delta)\|_*}{2R\|\mathcal{P}_V\mathcal{G}(\delta)\|_*}} \geq \sqrt{1/2R}$$
$$\frac{\sigma}{\|\mathcal{P}_W\mathcal{G}(\delta)\|_F} \geq \sqrt{\frac{\|\mathcal{P}_V\mathcal{G}(\delta)\|_*}{2R\|\mathcal{P}_W\mathcal{G}(\delta)\|_*}} \geq \sqrt{1/6R}$$

the second last inequality comes from (16). Thus if $\|(I - \boldsymbol{U}^T\boldsymbol{U})\delta\| = O(1/\sqrt{R})\|\delta\|$, in other words, $\|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\|_F = O(1/\sqrt{R})\|\mathcal{G}(\delta)\|_F$, whenever δ in normal cone, we have

$$\|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\| \leq \|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\|_F \leq O(1/\sqrt{R})\|\mathcal{G}(\delta)\|_F \leq \|\mathcal{G}(\delta)\| \tag{17}$$

so $\Gamma < 1$. To get this, we need $\sqrt{T}/w(\mathcal{J}(\beta)) = O(\sqrt{R})$ where $T = O(pR^2\log^2 n)$ (Vershynin, 2018, Thm 9.1.1), still not tight in R, but $O(\min\{n, R^2\log^2 n\})$ is as good as Oymak and Ozay (2018) and better than Sarkar et al. (2019), which are $O(n)$ and $O(n^2)$ correspondingly. (Vershynin, 2018, Thm 9.1.1) is a bound in expectation, but it naively turns into high probability bound since $\Gamma \geq 0$. ■

6. Bounding Γ, where do we lose?

The previous proof is not tight here.

$$\underbrace{\|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\| \leq \|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\|_F}_{\text{not tight}} \leq O(1/\sqrt{R})\|\mathcal{G}(\delta)\|_F \leq \|\mathcal{G}(\delta)\| \tag{18}$$

If we can show that, for all δ in the normal cone (thus independent of $\boldsymbol{U}$), $\|\mathcal{G}((I-\boldsymbol{U}^T\boldsymbol{U})\delta)\| = O(1/\sqrt{R})\|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\|_F$ for $\boldsymbol{U} \in \mathbb{R}^{O(R\log^2 n)\times n}$, then we can get the correct sample complexity. The difficulty is that, we do not know the distribution of $(I - \boldsymbol{U}^T\boldsymbol{U})\delta$. Let $M = I - \boldsymbol{U}^T\boldsymbol{U}$ and $g := M\delta$. Let $\tilde{g}$ be a Gaussian vector with same mean and covariance as g that will be studied later. We know that $g_i = \sum M_{ij}\delta_j$. Let $z_{ij} = U_{:,i}^T U_{:,j}$, u, v denote

standard Gaussian vectors of dimension T, we have (the last equation: $i \neq j$)

$$\begin{aligned}
E((1-z_{ii}^2)^2) &= E((1-\frac{1}{T}u^T u)^2) \\
&= 1 - \frac{2}{T}\sum_{i=1}^{T} E(u_i^2) + \frac{1}{T^2}(\sum_{i=1}^{T} E(u_i^4) + \sum_{i\neq j}^{T} E(u_i^2 u_j^2)) = \frac{2}{T}. \\
E(z_{ij}^2) &= E((\frac{1}{T}u^T v)^2) \\
&= \frac{1}{T^2}E(\sum u_i^2 v_i^2) = \frac{1}{T}. \\
E(g_i) &= 0, \\
E(g_i^2) &= E((\sum M_{ij}\delta_j)^2) \\
&= \delta_i^2 E((1-z_{ii}^2)^2) + \sum_{j\neq i}\delta_j^2 E(z_{ij}^2) + \sum_{j\neq k}\delta_j\delta_k E(M_{ij}M_{ik}) \\
&\leq \frac{1}{T}(\delta_i^2 + \|\delta\|^2). \\
E(g_i g_j) &= E((\sum M_{ik}\delta_k)(\sum M_{jl}\delta_l)) \\
&= \delta_i\delta_j E(M_{ij}M_{ji}) \\
&= \frac{1}{T}\delta_i\delta_j.
\end{aligned}$$

So

$$Cov(g) = \frac{1}{T}(\|\delta\|^2 I + \delta\delta^T).$$

The problem is that g is not Gaussian so even we know mean and variance it's still hard to deal with. Let's study Gaussian first. If $\tilde{g} = \tilde{g}_1 + \check{g}_2\delta$ where $\tilde{g}_1 \sim \mathcal{N}(0, \frac{\|\delta\|^2}{T}I)$ and $\check{g}_2 \sim \mathcal{N}(0, 1/T)$, then we have

$$\begin{aligned}
E(\|\mathcal{G}(\tilde{g})\|) &\leq E(\|\mathcal{G}(\tilde{g}_1)\|) + E(|\check{g}_2|\|\mathcal{G}(\delta)\|) \\
&\leq \frac{1}{\sqrt{T}}(\|\delta\|\frac{\log n}{\sqrt{n}} + \|\mathcal{G}(\delta)\|) \\
&\leq \frac{1}{\sqrt{T}}(\underbrace{\frac{\sqrt{R}\log n}{\sqrt{n}}}_{\text{proven in paper}} + 1)\|\mathcal{G}(\delta)\| \\
&\leq \frac{2}{\sqrt{T}}\|\mathcal{G}(\delta)\|.
\end{aligned}$$

If we have

$$P(\|\mathcal{G}(\tilde{g})\| > \alpha E(\|\mathcal{G}(\tilde{g})\|)) \leq \psi(\alpha),$$

then let $\alpha = \sqrt{T}/2$, we have

$$P(\|\mathcal{G}(\tilde{g})\| > E(\|\mathcal{G}(\delta)\|)) \leq \psi(\sqrt{T}/2)$$

We hope that $\psi(\alpha) = \exp(-O(\alpha^2))$ or $\log(\psi(\alpha)) = -O(\alpha^2)$. Then with a set of Gaussian width $\sqrt{R}\log n$, we use a union bound and have (if we ignore the difference between g and $\tilde{g}$)

$$P(\max_{\delta} \|\mathcal{G}(g)\| > \|\mathcal{G}(\delta)\|) \leq \psi(\sqrt{T}/2)\exp(O(R\log^2 n)) = \exp(O(R\log^2 n) + \log(\psi(\sqrt{T}/2))).$$

So if the derivation of a Gaussian vector can be applied to a non-Gaussian $g = (I - \boldsymbol{U}^T\boldsymbol{U})\delta$ with the same mean and variance, and $\|\mathcal{G}(g)\|$ is a subGaussian random variable, then we can get the tight bound.

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.

Naman Agarwal, Brian Bullins, Elad Hazan, Sham M Kakade, and Karan Singh. Online control with adversarial disturbances. *arXiv preprint arXiv:1902.08721*, 2019.

Mustafa Ayazoglu and Mario Sznaier. An algorithm for fast constrained nuclear norm minimization and applications to systems identification. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 3469–3475. IEEE, 2012.

Arindam Banerjee, Sheng Chen, Farideh Fazayeli, and Vidyashankar Sivakumar. Estimation with norm regularization. In *Advances in Neural Information Processing Systems*, pages 1556–1564, 2014.

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.

James A Cadzow. Signal enhancement-a composite property mapping algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(1):49–62, 1988.

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.

Sarah Dean, Stephen Tu, Nikolai Matni, and Benjamin Recht. Safely learning to control the constrained linear quadratic regulator. In *2019 American Control Conference (ACC)*, pages 5582–5588. IEEE, 2019.

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.

Jonathan Gillard. Cadzows basic algorithm, alternating projections and singular spectrum analysis. *Statistics and its Interface*, 3(3):335–343, 2010.

Cristian Grossmann, Colin N Jones, and Manfred Morari. System identification with missing data via nuclear norm regularization. In *2009 European Control Conference (ECC)*, pages 448–453. IEEE, 2009.

Anders Hansson, Zhang Liu, and Lieven Vandenberghe. Subspace system identification via weighted nuclear norm optimization. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 3439–3444. IEEE, 2012.

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.

Felix Krahmer, Shahar Mendelson, and Holger Rauhut. Suprema of chaos processes and the restricted isometry property. *Communications on Pure and Applied Mathematics*, 67(11): 1877–1904, 2014.

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

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalent control of lqr is efficient. *arXiv preprint arXiv:1902.07826*, 2019.

Tomas McKelvey, Hüseyin Akçay, and Lennart Ljung. Subspace-based multivariable system identification from frequency response data. *IEEE Transactions on Automatic Control*, 41 (7):960–979, 1996.

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

Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. Simple bounds for noisy linear inverse problems with exact side information. *arXiv preprint arXiv:1312.0641*, 2013.

Tapan K Sarkar and Odilon Pereira. Using the matrix pencil method to estimate the parameters of a sum of complex exponentials. *IEEE Antennas and Propagation Magazine*, 37(1):48–55, 1995.

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.

Roy S Smith. Frequency domain subspace identification using nuclear norm minimization and hankel matrix realizations. *IEEE Transactions on Automatic Control*, 59(11):2886–2896, 2014.

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.

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

Michel Verhaegen and Anders Hansson. N2sid: Nuclear norm subspace identification of innovation models. *Automatica*, 72:57–63, 2016.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Weiyu Xu, Jirong Yi, Soura Dasgupta, Jian-Feng Cai, Mathews Jacob, and Myung Cho. Sep] ration-free super-resolution from compressed measurements is possible: an orthonormal atomic norm minimization approach. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 76–80. IEEE, 2018.