Finite Sample Identification of Low-Order LTI Systems via Nuclear Norm Regularization

YUE SUN (Student Member, IEEE), SAMET OYMAK (Member, IEEE), AND MARYAM FAZEL (Senior Member, IEEE)

1Department of Electrical and Computer Engineering, University of Washington, Seattle, WA 98195 USA
2Department of Electrical and Computer Engineering, University of California, Riverside, CA 92507 USA

Corresponding Author: MARYAM FAZEL (e-mail: mfazel@uw.edu).

This work was supported in part by the U.S. National Science Foundation under Grants DMS 2023166, CCF 2007036, CCF 2212261, and 1932254, and in part by ARO under Grant W911NF2110312.

ABSTRACT This paper studies the problem of identifying low-order linear time-invariant systems via Hankel nuclear norm (HNN) regularization. This regularization encourages the Hankel matrix to be low-rank, which corresponds to the dynamical system being of low order. We provide novel statistical analysis for this regularization, and contrast it with the unregularized ordinary least-squares (OLS) estimator. Our analysis leads to new finite-sample error bounds on estimating the impulse response and the Hankel matrix associated with the linear system using HNN regularization. We design a suitable input excitation, and show that we can recover the system using a number of observations that scales optimally with the true system order and achieves strong statistical estimation rates. Complementing these, we also demonstrate that the input design indeed matters by proving that intuitive choices, such as i.i.d. Gaussian input, lead to sub-optimal sample complexity. To better understand the benefits of regularization, we also revisit the OLS estimator. Besides refining existing bounds, we experimentally identify when HNN regularization improves over OLS: (1) For low-order systems with slow impulse-response decay, OLS method performs poorly in terms of sample complexity, (2) the Hankel matrix returned by regularization has a more clear singular value gap that makes determining the system order easier, (3) HNN regularization is less sensitive to hyperparameter choice. To choose the regularization parameter, we also outline a simple joint train-validation procedure.

INDEX TERMS Identification for control, information theory for control, linear systems.

I. INTRODUCTION

System identification is a central topic in control theory; accurate estimation of system dynamics is the basis of control or policy decisions in tasks varying from linear-quadratic regulator to model predictive control. A Linear time-invariant (LTI) dynamical system is described as

$$x_{t+1} = Ax_t + Bu_t,$$
$$y_t = Cx_t + Du_t + z_t,$$

where $x_t$, $u_t$, $y_t$, $z_t$ are the state, input, output, and noise vectors respectively, and the initial state is denoted by $x_0$. Suppose the true system order is $R$, i.e., $R$ is the smallest state dimension in (1). In practice, the true order $R$ is unknown so it is customary to pick a user-specified order $n \geq R$ and fit a larger system with state dimension $n$. For such a problem, the total number of parameters in $A$, $B$, $C$, $D$ matrices grow linearly in $n$. However, due to underlying low-order dynamics, the degrees of freedom of the system dynamics is proportional to $R$, regardless of $n$. This discussion naturally motivates regularization strategies that promote low-order systems to achieve system identification with near-optimal sample size. To this aim, we will use nuclear norm regularization for the Hankel matrix associated with the system. It is well known that the rank of a Hankel matrix corresponds to the system order $R$, and that nuclear norm regularization promotes low-rank solutions in matrix regression problems. In light of these, our main contribution is a statistical analysis of the Hankel nuclear norm (HNN) regularization approach. We will shed light on the fundamental limits of HNN, and contrast it with the...
Without Regularization, Sample Complexity Scales With the
User-Specified Order $n$. With HNN Regularization, We Establish
upper/lower Bounds That Overcome $O(n)$ Dependence and Also Reveal
the Importance of Input Design.

| Hankel nuclear norm | shaped input | i.i.d. input |
|---------------------|-------------|-------------|
| Ordinary least squares | $\lesssim \log^2(n)$ | $\gtrsim n^{1/3}$ |

Figure 1. (a) Shaped input (where variance of $u(t)$ changes over time): recovery is guaranteed when $T \approx R$; (b) i.i.d input (fixed variance): recovery fails with high probability when $T \lesssim n^{1/3}$. See Theorem 2.

ordinary least squares (OLS) estimator for low-order system identification.

Table 2 covers the core notation used in the paper, most of which are introduced in Section II.

We note that a shorter version of this paper was published as [1]. Compared to this work, we make several contributions by providing (1) a negative result for suboptimality of sample complexity with i.i.d. input (Section III Part II-B), (2) a discussion on model selection and how to tune the regularization parameter to establish estimation bounds for the train-validation procedure (Section V), (3) numerical simulations that demonstrate the advantage of HNN over least squares (Section VI Part VI-A).

A. CONTRIBUTIONS

Our main contribution is establishing data-driven guarantees for Hankel nuclear norm (HNN) regularization and shedding light on the benefit of regularization through a comparison to the ordinary least-squares (OLS) estimator. The core technical challenge in this work is the analysis of HNN-based system identification which requires novel statistical machinery compared to the prior finite-sample analysis which mostly focus on OLS. Table 1 summarizes our findings at a high-level: We not only show HNN can overcome the $O(n)$ sample size dependence but its performance strongly depends on the input design illustrated in Fig. 1. More specifically, we make the following contributions.

- **Sample complexity of Hankel nuclear norm** (Section III): For multi-input/single-output (MISO) systems ($p$ input channels), we establish near-optimal sample complexity bounds for the Hankel-regularized system identification and show that sample size grows as $O(pR\log^2 n)$ where $R$ is the true system order and $n$ is the user-specified order which is equal to the number of blocks in the Hankel matrix. This result utilizes an input-shaping strategy (rather than i.i.d. excitation, see Fig. 1(a)) and builds on [2] who studied the recovery of a sum-of-exponentials signal.

  In light of [3], systems with low-rank Hankel can be viewed as a simultaneously structured signal with Hankel and low-rank structures. Through degrees-of-freedom counting one can deduce that (A) enforcing only low-rankness without Hankel structure would require $O(nR)$ samples and (B) enforcing only Hankel structure without low-rankness would require $O(n)$ samples. Thus, $O(R\log^2(n))$ is a best-of-both worlds bound. This near-optimal bound is especially interesting since simultaneously structured signals can be provably hard to estimate with few measurements [3].

  Interestingly, in Section III-B, we also show that our input-shaping strategy is indeed critical for establishing logarithmic sample complexity dependence in $n$. Specifically, we prove that if the inputs are i.i.d. standard normal (Fig. 1(b)), the minimum number of observations to recover the impulse response grows polynomially as $T \gtrsim n^{1/3}$.

- **Improved statistical rates for HNN and OLS (Section IV):** We establish noise robustness of HNN regularization by upper bounding the estimation errors of impulse response (IR) of the system and the associated Hankel matrix. Unlike related results by [2], our bounds exhibit the optimal $O(1/\sqrt{T})$ decay with growing sample size $T$. These findings also reveal the competitiveness of HNN compared to OLS for noisy systems.

  For OLS, we also establish a near-optimal spectral error rate for the Hankel matrix estimation when $T \gtrsim np$. Our error rate improves over that of [4] and our sample complexity improves over [5] and [6] which require $O(n^2)$ samples rather than $O(n)$. This refinement is accomplished by tighter analysis that relate the IR and Hankel errors (see Thm. 3).

- **Guarantees on model-selection (Section V):** HNN method requires a proper choice of the regularization parameter $\lambda$. In practice, the optimal choice is data dependent and one usually estimates $\lambda$ via trial and error based on the validation error. We provide a complete procedure for model selection (training & validation phases), and establish statistical guarantees.

- **Contrasting HNN regularization and OLS (Section VI):** Finally, we assess the benefits of regularization via numerical experiments on system identification focusing on data collected from a single-trajectory. Following Table 1, we consider synthetic data and focus on low-order systems with slow impulse-response decay. The slow-decay is intended to exacerbate the FIR approximation error arising from truncating the impulse-response at $O(n)$ terms. In this setting, OLS as well as [5] are shown to perform poorly due to $O(n)$ dependence. In contrast, HNN regularization is able to avoid the truncation error as it allows for fitting a long impulse-response with few data (thanks to $\log^2(n)$ dependence).

  Our real-data experiments (on a low-order example from the DaSy datasets [7]) show that the regularized algorithm has empirical benefits in sample complexity, estimation error, and Hankel spectral gap, and demonstrate that the regularized algorithm is less sensitive to the choice of the tuning
Finally, we remark that, while this work focuses on the estimation of impulse response and the associated Hankel matrix, the results can be extended to identify the state-space matrices \((A, B, C, D)\). This can be accomplished via Ho-Kalman type post-processing algorithms which can estimate state-space matrices (up to a similarity transformation) given a noisy Hankel matrix \([8]\). Our primary interest is system identification and not control. However, it is also well-known that stronger estimation guarantees translate to stronger model-based controllers (e.g. certainty equivalent control). Upon reviewer suggestion, we further elaborate on benefits of our results to \(H_\infty\) or adaptive control.

### B. RELATED WORKS ON SYSTEM IDENTIFICATION

System identification has a rich and dynamic literature. We emphasize that our work focuses on regression methods (OLS and HNN) to estimate the impulse response of the system. These methods can be augmented with preprocessing/postprocessing techniques to improve estimation as well as identifying low-order state-space realizations. We gather the related works into following categories.

**Classical methods and asymptotic guarantees:** The traditional unregularized methods include Cadzow approach \([9], [10]\), matrix pencil method \([11]\), Ho-Kalman approach \([12]\) and the subspace method raised in \([13], [14], [15]\), further modified as frequency domain subspace method in \([16]\) when the inputs are single frequency (sine/cosine) signals. Much of the prior literature on these are asymptotic guarantees rather than finite sample analysis.

**Finite sample identification guarantees:** This literature is closer to our work and prior works can be categorized into identification with and without having direct access to state observations. State observation corresponds to setting \(C, D = I, 0\) in the state-space \((1)\). A notable line of work derives statistical bounds in this setting from a single output trajectory (defined in Fig. 2) by least squares \([17], [18], [19], [20]\) by estimating the \(A, B\) matrices. \([19]\) assumes that the system is marginally stable whereas \([20]\) removes the assumptions on the spectral radius of \(A\).

Closer to us, for the partially observed setting \((1)\), recent works show that least-squares can be used to recover the Markov parameters and reconstruct \(A, B, C, D\) from the Hankel matrix via least-squares. While \([4], [5]\) use random input, \([6]\), Thm 1.1, 1.2 uses impulse and single frequency signal respectively as input. These works assume known system order, or traverse the Hankel size \(n\) to fit the system order. There are more recent extensions on identifying unstable systems \([21], [22], [23], [24]\) propose equipping least squares with frequency domain filtering to assist system identification.

Representing the system as auto-regressive or ARX models: Besides fitting impulse response and system matrices, one can also predict the output via past outputs as well. Recent works develop methods and finite sample analysis to...
identify auto-regressive model [25]. There is also a rich system identification literature on ARX models [26], iterative ARX models [27], and related kernel methods [28], [29] to best explain the dynamics of a system. These works focus on either empirical/algorithmic performance or asymptotically correct theoretical guarantees, rather than finite sample error bounds, which is the focus of our work.

Results on regularized system identification: With only access to the input/output data, [30], [31] propose onvex optimization formulations to establish a confidence set corresponding to the estimated impulse response without making assumptions on the impulse response structure (e.g., low order, fast decay). However, these results do not say much about the required sample size under which the resulting confidence sets are guaranteed to be small. In contrast, our work is about determining the minimum sample size to ensure parameters truly lie on a small ball around the estimate.

There has also been growing interest in kernel-based regularization methods for system identification [28], [29], [32], [33], [34]. Section 7.5 of [32] provides non-asymptotic guarantees on system identification through stable spline estimators which is closer to our work. This is a kernel method that essentially solves regularized least squares and is an alternative to nuclear norm regularization. Specifically, they use a regularizer of the form $\text{Reg}(h) = h^\top M h$ where $M$ is a positive definite matrix that depends on the spectral radius or the a-priori distribution of the impulse response. The guarantees for spline estimators are heavily dependent on stability of the system and require a small spectral radius $\rho$ (see Theorem 7.11 of [32]). This is unlike our bounds where the sample size depends only on the true order of the system regardless of stability $\rho$. We also highlight that [34] provides a further survey on kernel/spline methods and extensively compares kernel, atomic norm, and nuclear norm approaches for system identification. In Section VI, we find that nuclear norm regularization outperforms least-squares methods for systems with slow impulse response decay. This finding is consistent with the guarantees on spline estimators which require small spectral radius and fast decay.

Hankel nuclear norm regularization: These works focus on recovering a low order system, in other words, a system with low-rank Hankel matrix. Low-rank Hankel matrices arise in a range of applications [2], [35], [36], [37]. It is known that nuclear norm regularization is used to find a low rank matrix [38], [39], [40], and [41] uses it for recovering a low rank Hankel matrix. Traditional methods such as subspace methods can synergistically be applied together with regularization to further boost performance [27], [42], [43], [44], [45], [46], [47], [48], [49], [50]. The prior works on HNN regularization emphasize efficient optimization algorithms and do not provide statistical bounds. More recently, Ref. [2] theoretically shows that a low order SISO system from multiple trajectories of input-output pairs can be recovered via HNN approach. However, they provide pessimistic noise robustness guarantees which are agnostic to sample size (unlike our optimal $O(1/\sqrt{T})$ rates). Ref. [50] provides a thorough analysis of Hankel nuclear norm regularization applied to system identification, including discussion on proper error metrics, role of rank/system order in formulating the problem, and selection of tuning parameters. Besides nuclear norm regularization, the singular value shrinkage method can also be applied for low rank matrix recovery [51], [52], which is anticipated to have similar benefits for finite sample system identification.

C. ROADMAP

The rest of the paper is organized as follows. Next section introduces the technical setup. Sections III Part III-A proposes our results on nuclear norm regularization. Section III Part III-B discusses the role of the input distribution and establishes lower bounds. Section IV provides our results on least-squares estimator. Section V discusses model selection algorithms. Finally Section VI presents the numerical experiments.

II. PROBLEM SETUP AND ALGORITHMS

Consider a linear time-invariant system of order $R$ with the minimal state-space representation

$$x_{t+1} = Ax_t + Bu_t,$$

$$y_t = Cx_t + Du_t + z_t,$$  \hspace{1cm} (2)

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. We denote $\rho(A)$ as the spectral radius of $A$. Generally with the same input and output, the dimension of the hidden state $x_t$ can be any number no less than $R$. In this paper we are interested in the minimum dimensional representation (i.e., minimal realization).

The goal of system identification is to find the system parameters, such as $A, B, C, D$ matrices or impulse response, given input and output observations. If $(C, D) = (I, 0)$, we directly observe the state.

The state evolves as $x_{t+1} = Ax_t + \eta_t$ where $\eta_t$ is the white noise that provides excitation to states [19], [20].

When we do not directly observe the state $x_t$ (also known as hidden-state), one has only access to $u_t$ and $y_t$ and lack the full information on $x_t$. We recover the impulse response (also known as the Markov parameters) sequence $h_0 = D, h_t = CA^{-1}B \in \mathbb{R}^{m \times p}$ for $t = 1, 2, \ldots$ that uniquely identifies the end-to-end behavior of the system. The impulse response can have infinite length, and we let $h = [D, CB, CAB, CA^2B, \ldots, CA^{2n-3}B]^\top$ denote its first $2n - 1$ entries, which can be later placed into an $n \times n$ block Hankel matrix. Without knowing the system order, we consider recovering the first $n$ terms of $h$ where $n$ is larger than system order $R$. To this end, let us also define the block Hankel map $\mathcal{H} : \mathbb{R}^{m \times (2n-1)p} \to \mathbb{R}^{mn \times mn}$ as

$$H := \mathcal{H}(h) = \begin{bmatrix} h_1 & h_2 & \ldots & h_n \\ h_2 & h_3 & \ldots & h_{n+1} \\ \vdots \\ h_n & h_{n+1} & \ldots & h_{2n-1} \end{bmatrix}. \hspace{1cm} (3)$$
If \( n \geq R \), the Hankel matrix \( H \) is of rank \( R \) regardless of \( n \) [53], Sec. 5.5]. Specifically, we will assume that \( R \) is small, so the Hankel matrix is low rank. Throughout, we estimate the first \( 2n-1 \) terms of the impulse response denoted by \( h \). The system is excited by an input \( u \) over the time interval \([0, t]\) and the output \( y \) is measured at time \( t \), i.e.,

\[
y_t = \sum_{i=1}^{j} h_{t+1-i} u_i + z_t.
\]  

We assumed \( x_0 = 0 \). If the system has spectral radius \( \rho \ll 1 \), then the effect of \( x_0 \) at the time \( N \) scales as \( \|x_0\|\rho^{N} \) which is a small number, thus does not harm our result. If the system is close to marginally stable, a thorough discussion is in Section VI-A.

**Performance criteria for system identification:** To explain our contributions, we introduce common performance metrics. To ensure a standardized comparison, we define *sample complexity* to be the number of equations (equality constraints in variables \( h_i \)) used in the problem formulation, which is same as the number of observed outputs (see Fig. 2 and Section II). With this, we explore the following performance metrics for learning the system from \( T \) output measurements.

- **Sample complexity**: The minimum sample size \( T \) for recovering system parameters with noise robustness. This quantity is lower bounded by the system order \( R \) which can be seen as the “degrees of freedom” of the system.
- **Impulse Response (IR) Estimation Error**: The Frobenius norm error \( \|\hat{h} - h\|_F \) for the IR. A good estimate of IR enables the accurate prediction of the system output.
- **Hankel Estimation Error**: The spectral norm error \( \|\mathcal{H}(\hat{h} - h)\| \) of the Hankel matrix. This metric is particularly important for system identification as described below.

The Hankel spectral norm error is a critical quantity for several reasons. First, the Hankel spectral norm error connects to the \( \mathcal{H}_\infty \) estimation of the system [54]. Secondly, bounding this error allows for robustly finding balanced realizations of the system; for example, the error in reconstructing state-space matrices \((A, B, C, D)\) 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 caused by estimation error. 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 aids in identifying the true order of the system as explored in Section VI.

**Algorithms:** Hankel-regularization & OLS. In our analysis, we consider a multiple rollout\(^1\) setup where we measure the system dynamics with \( T \) separate rollouts. For each rollout, the input sequence is\(^2\)

\[
u^{(i)} = [u^{(i)}_{2n-1}, \ldots, u^{(i)}_0] \in \mathbb{R}^{2n-1}p
\]

and we measure the system output at time \( 2n-1 \). Note that the \( j \)-th output at time \( 2n-1 \) is simply \( h^T u^{(i)} \).

Define \( \hat{U} \in \mathbb{R}^{T \times (2n-1)p} \) where the \( j \)-th row is \( u^{(i)} \). Let \( y \in \mathbb{R}^{T \times m} \) denote the corresponding observed outputs. Hankel-regularization refers to the nuclear norm regularized problem (HNN).

\[
\hat{h} = \underset{h'}{\arg \min} \quad \frac{1}{2} \| \hat{U} h' - y \|^2_F + \lambda \| \mathcal{H}(h') \|_*
\]  

(HNN)

Finally, setting \( \lambda = 0 \), we obtain the special case of ordinary least-squares (OLS).

**Data acquisition:** Generally there are several rounds (\( i \)th round is denoted with super script \((i)\) in Fig. 2) of inputs sent into the system, and the output can be collected or neglected at arbitrary time. In the setting that we refer to as “multi-rollout” (Fig. 2(b)), for each input signal \( u^{(i)} \) we take only one output measurement \( y_{ti} \), at time \( t = 2n-1 \) and then the system is restarted with a new input. Here the sample complexity is \( T \), the number of output measurements as well as the round of inputs. Recent papers (e.g., [4] and [5]) use the “single rollout” model (Fig. 2(c)) where we apply an input signal from time \( 1 \) to \( 2n-2 \) without restart, and collect all output from time \( 2n-1 \) to \( 2n-2 \), in total \( T \) output measurements; we use this model in the numerical experiments in Section VI.

We consider two estimators in this paper: the *nuclear norm regularized estimator* and the *least squares estimator* defined later.

We will bound the various error metrics mentioned earlier in terms of the sample complexity \( T \), the true system order \( R \), the dimension of impulse response \( n \gg R \), and signal to noise ratio (SNR) defined as \( \text{SNR} = \mathbb{E}[\|u\|_2^2/n]/\mathbb{E}[\|z\|_2^2] \). Table 3 provides a summary and comparison of these bounds. All bounds are order-wise and hide constants and log factors. We can see that, with nuclear norm regularization, our paper matches the least squares impulse response and Hankel spectral error bound while sample complexity can be as small as \( \mathcal{O}(R^2) \), and we can recover the impulse response with guaranteed suboptimal error when sample complexity is \( \mathcal{O}(R) \). Our least square error bound matches the best error bounds among [4] and [5], which is proven optimal for least squares.

Next, we discuss the design of the input signal and introduce input shaping matrix.

**Input shaping:** Note that the \( \mathcal{H} \) operator does not preserve the Euclidean norm, so [2] proposes using a normalized operator \( \mathcal{G} \), where they first define the weights

\[
K_j = \left\{ \begin{array}{ll} \sqrt{j}, & 1 \leq j \leq n, \\ \sqrt{2n-j}, & n < j \leq 2n-1. \end{array} \right.
\]  

(5)

and let \( K \in \mathbb{R}^{(2n-1)p \times (2n-1)p} \) be a block diagonal matrix where the \( j \)-th diagonal block of size \( p \times p \) is equal to \( K_j I_{p \times p} \).

---

\(^1\)We remark that in engineering each rollout usually amounts to an “experiment” on the system.

\(^2\)Note that, the time order of input if flipped. Since \( y \) is a convolution of \( u \) and \( h \), to keep the time order of \( h \) and write the proper linear map, we have to flip \( u \).
TABLE 3 Comparison of Recovery Error of Impulse Response. The Hankel Matrix is $n \times n$, the System Order is $R$, and the Number of Samples is $T$, and $\sigma = 1/\sqrt{\text{SNR}}$ Denotes the Noise Level. LS-IR and LS-Hankel Stands for Least Squares Regression on the Impulse Response and on the Hankel Matrix

| Paper | This work | [4] | [5] |
|-------|-----------|-----|-----|
| Sample complexity | $\hat{R}$ | $n$ | $n$ | $n^2$ |
| Method | Nuc-norm | LS-IR | LS-IR | LS-Hankel |
| Impulse response error | see (7) | $\sigma/\sqrt{n/T}$ | $\sigma/\sqrt{n/T}$ | $(1 + \sigma)/\sqrt{n/T}$ |
| Hankel spectral error | see (7) | $\sigma/\sqrt{n/T}$ | $\sigma/\sqrt{n/T}$ | $(1 + \sigma)/\sqrt{n/T}$ |

In other words,

$$K = \begin{bmatrix} K_1 I & 0 & 0 & \ldots & 0 \\ 0 & K_2 I & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & K_{2n-1} I \end{bmatrix}$$

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 $G^* : \mathbb{R}^{mn \times np} \rightarrow \mathbb{R}^{mn \times (2n-1)p}$ as the adjoint of $G$, where $[G^*(M)]_{i,j} = \sum_{k=1}^{n} M_{(i,j)(k)}/K_i$ if we denote the $j, k$-th block of $M$ (defined in (3)) by $M_{(i,j)(k)}$. Using this change of variable and letting $U = UK^{-1}$, problem (HNN) can be written as

$$\hat{\beta} = \arg \min_{\beta'} \frac{1}{2} \left\| U \beta' - y \right\|_F^2 + \lambda \left\| G(\beta') \right\|_s.$$  

III. HANKEL NUCLEAR NORM REGULARIZATION

To promote a low-rank Hankel matrix, we add nuclear norm regularization to the quadratic-loss 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 $U$ and observe the corresponding noisy output vector $y$ as in (4).

A. STATISTICAL GUARANTEES FOR HANKEL-REGULARIZATION

The theorem below shows that Hankel-regularization achieves near-optimal sample complexity with similarly strong estimation error rates that decay as $1/\sqrt{T}$. For simplicity, we assume initial state to be $x_0 = 0$. For stable systems, it is known that the impact of initial state decays exponentially in sampling time $2n - 1$.

**Theorem 1:** Consider the problem (HNN) in the MISO (multi-input single-output) setting $(m=1, p$ input channels). Suppose the system order is $R$, $U \in \mathbb{R}^{T \times (2n-1)p}$, each row consists of an input rollout $u^{(i)} \in \mathbb{R}^{(2n-1)p}$, and the scaled input $U = UK^{-1}$ has i.i.d Gaussian entries. Let $\text{snr} = \mathbb{E}[\|u\|^2/n]/\mathbb{E}[\|e\|^2]$ and $\sigma = 1/\sqrt{\text{snr}}$. For some constant $C$, let $\lambda = C\sigma \sqrt{\frac{mn}{T} \log(n)}$. Then, with probability $1 - O(R \log(n)/\sqrt{pT})$ for the first case and $1 - \exp(O(\sqrt{T} - O(\sqrt{pR\log(n)})))$ for the second case, solving (HNN) returns $\hat{h}$ such that

$$\frac{\|\hat{h} - h\|_2}{\sqrt{2}} \leq \|\mathcal{H}(\hat{h})\|$$

$$\leq \begin{cases} \sqrt{\frac{np}{\text{snr} \times T}} \log(n) & \text{if } T \gg \bar{R}, \\ \sqrt{\frac{Rnp}{\text{snr} \times T}} \log(n) & \text{if } pR\log(n) \leq T \leq \bar{R} \end{cases}$$

(7)

where $\bar{R} = \min(p\sigma^2 \log^2(n), n)$.

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: 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. The number of samples for recovering an order-$R$ system is $O(R \log(n))$, where the SISO case is also proven in [2], and we follow a similar proof. The error rate with respect to the system parameters $n, R, T$ is same as [4], [5], [6] (e.g. compare to Thm. 3). We can choose $\lambda$ from a range of a constant multiplicative factors, like $\lambda \in [C\sqrt{\frac{mn}{T}} \log(n), 2C\sigma \sqrt{\frac{mn}{T}} \log(n)]$ for a positive constant $C$. When $\lambda$ is in this range, there is always a sample complexity $T$ such that (7) holds. The flexibility of $\lambda$ makes tuning $\lambda$ for Algorithm 1 easier.

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. Tuning method Algorithm 1 and numerical experiments on real data in Section VI demonstrate the performance and robustness of the regularized method.

B. SAMPLE COMPLEXITY LOWER BOUNDS FOR IID INPUTS AND THE IMPORTANCE OF INPUT SHAPE

Theorem 1 uses shaped inputs whereas, in practice, one might expect that i.i.d. input sequence (without shaping matrix $K$) should be sufficient for recovering the impulse response with near-optimal sample size. For instance, [4] proves an optimal sample complexity bound for system identification via least-squares and i.i.d. standard normal inputs. Naturally, we ask: Does Hankel-regularization enjoy similar performance guarantees with i.i.d. inputs? Do we really need input design?

The following theorem proves that for a special system with order $r = 1$, the sample complexity of the problem under i.i.d. input is no less than $n^{1/3}$, compared to $O(\log n)$ under $\lambda$.

3In this theorem, we make the dependence on $\log(n)$ explicit and only hide constants.
the shaped input setting. This is accomplished by carefully lower bounding the Gaussian width induced by the Hankel-regularization. Gaussian width directly corresponds to the square-root of the sample complexity of the problem required for recovery in high probability [55], Thm. 1). Thus, Theorem 2 shows that the sample complexity with i.i.d. input can indeed be provably larger than with shaped input.

**Theorem 2:** Suppose the impulse response $h$ of the system is $h = 1$, $orall t \geq 1$, which is order 1. Consider the tangent cone associated with Hankel-regularization (normalized to unit $\ell_2$-norm) defined as $\{ x | \| H(h + x) \|_\ast \leq \| H(h) \|_\ast \}$. The Gaussian width of this set is lower bounded by $\frac{1}{4} n^{1/6}$ when $n \geq 2$.

This implies that, in the noiseless setting, the sample complexity to recover the impulse response is $T \gtrsim n^{1/3}$, which is larger than log $n$ dependence with shaped input. This result is rather counter-intuitive since i.i.d. inputs are often optimal for structured parameter estimation tasks (e.g. compressed sensing, low-rank matrix estimation). In contrast, our result shows the provable benefit of input shaping.

### IV. REFINED BOUNDS FOR THE LEAST-SQUARES ESTIMATOR

To better understand the Hankel-regularization bound of Theorem 1, one can contrast it with the performance of unregularized least-squares. Interestingly, the existing least-squares bounds are not tight when it comes to Hankel spectral norm. In this section, we revisit the least-squares estimator, tighten existing bounds, and contrast the result with our Theorem 1.

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 (HNN), hence the estimator is given via the pseudo-inverse

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

(8)

Using the fact that rows of the Hankel matrix are subsets of the IR sequence, we always have the inequality

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

(9)

Observe that there is a factor of $\sqrt{n}$ gap between the left-hand and right-hand side inequalities. We show that the left-hand side is typically the tighter one, thus $\| \hat{h} - h \|_F \sim \| H(\hat{h} - h) \|$. The next theorem formalizes this bound when inputs and noise are randomly generated.

**Theorem 3:** Denote the solution to (8) as $\hat{h}$. Let $U \in \mathbb{R}^{T \times (2n-1)p}$ 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

$$
\| H(\hat{h} - h) \| \leq \sigma_z \sqrt{\frac{mp}{T}} \log(np) \text{ when } T \gtrsim mnp.
$$

This theorem improves the spectral norm bound compared to [4], which naively bounds the spectral norm in terms of IR error using the right-hand side of (9). Instead, we show that spectral error is same as the IR error up to a log factor (when there is only output noise). We remark that $O(\sigma_z \sqrt{np/T})$ is a tight lower bound for $\| H(\hat{h} - h) \|$ as well as $\| h - \hat{h} \|$ [4, [56]. The proof of the theorem above is in Appendix A5. Note that, we apply the i.i.d input here for least squares which is different from regularized algorithm. It works with at least $O(n)$ samples, whereas the sample complexity for the regularized algorithm in Theorem 1 is $O(R)$.

### V. MODEL SELECTION FOR HANKEL-REGULARIZED SYSTEM ID

In Thm. 1, we established the recovery error for system’s impulse response for a particular parameter choice $\lambda$, which depends on the noise level $\sigma$. In practice, we do not know the noise level and the optimal $\lambda$ choice is data-dependent. Thus, given a set of parameter candidates $\Lambda \subset \mathbb{R}^+$, one can evaluate $\lambda \in \Lambda$ and minimize the validation error to perform model selection.

Covering model selection – selecting the model with the correct $\lambda$, the end-to-end algorithm will contain training-validation-selection procedures:

1. **Training:** For every $\lambda \in \Lambda$, we run (HNN) on training samples containing $T$ rollouts, and record the solution $\hat{h}_\lambda$.
2. **Validation:** The validation dataset contain $T_{val}$ rollouts. We collect all $\hat{h}_\lambda$ and check their validation error.
3. **Selection:** Choose the $\hat{h}_\lambda$ associated with smallest validation error.

In Appendix G, Algorithm 1 summarizes our training and validation procedure where $|\Lambda|$ denotes the cardinality of $\Lambda$. The theorem below states the performance guarantee for this algorithm.

**Theorem 4:** Consider the setting of Theorem 1. Sample $T$ i.i.d. training rollouts ($U$, $y$) and $T_{val}$ i.i.d. validation rollouts ($U_{val}$, $y_{val}$). Set $\lambda^* = C_0 \sqrt{\frac{R}{np}} \log(n)$ which is the choice in Thm. 1. Fix failure probability $P \in (0, 1)$. Suppose that:

(a) There is a candidate $\hat{\lambda} \in \Lambda$ obeying $\lambda^*/2 \leq \hat{\lambda} \leq 2\lambda^*$.

(b) Validation set obeys $T_{val} \gtrsim T \log^2(|\Lambda|/P) \log^3(n) \sqrt{R}$. 

Set $R = \min(R^2, n)$. With probability at least $1 - P$, Algorithm 1 achieves an estimation error equivalent to (7):

$$
\| H(\hat{h} - h) \| \lesssim \begin{cases} 
\sqrt{\frac{np}{\text{snr} \times T}} \log(n), & \text{if } T \gtrsim R, \\
\sqrt{R^{np} \text{snr} \times T} \log(n), & \text{if } R \lesssim T \lesssim R.
\end{cases}
$$

(10)

In a nutshell, this result shows that as long as candidate set contains a reasonable hyperparameter (Condition (a)), using few validation data (Condition (b)), one can compete with the hindsight parameter choice of Theorem 1. $T_{val}$ mildly depends on $T$ and only scales poly-logarithmically in $|\Lambda|$ which is consistent with the model selection literature [57].

In Algorithm 1, we fix the size of the Hankel matrix (which is usually large/overparameterized) and tune $\lambda > 0$. In contrast to this, for OLS and the least-squares procedure of [5], model selection is accomplished by varying the size of the Hankel matrix. The next section provides experiments and contrasts these methods and provides insights on how
regularization can improve over least-squares for certain class of dynamical systems.

Comparison of sample complexity of regularized algorithm and unregularized least-squares: model selection with data being requested online. Algorithm 1 uses static data for training and validation, which means that, the total $T + T_{val}$ samples are given and fixed, and we split the data and run Algorithm 1. We denote the total sample complexity $T_{tot} = T + T_{val}$. To be fully efficient in sample complexity, we can start from $T_{tot} = 0$, keep requesting new samples, which means increasing $T_{tot}$, and run Algorithm 1 for each $T_{tot}$. When the validation error is small enough (which happens when $T_{tot} \geq R$), we know the algorithm recovers a impulse response estimation with the error in Theorem 4 and we can terminate the algorithm.

We compare it with the model selection algorithm in [5] for least squares estimator, and we find that it does not terminate until $T_{tot} \geq n$. For least squares, the parameter to be tuned is the dimension of the variable, i.e., we vary the length of estimated impulse response. We call the tuning variable $n_t$ and it is upper bounded by $n$. We keep increasing $T_{tot}$ and train by varying $n_t \in [1, T_{tot}/2]$ (so the least squares problems are overdetermined). The output $y$ is collected at time $2n_t - 1$. We consider two impulse responses truncated at length $n$: $h^1 = I_n$ (order = 1) and $h^{n_1} = [I_{n_1} ; 0_{n-n_1}]$ (order = $n_1$). As long as $T_{tot} < n_1$, one cannot differentiate $h^1$ and $h^{n_1}$, because $y$ is collected at time $T_{tot}$ and the $T_{tot} + 1$-th to the $n$-th terms of $h^1, h^{n_1}$ do not contribute to $y$. Even if the system is order 1, one does not know it and cannot terminate the algorithm. Thus the tuning algorithm in [5] requires $T_{tot} \geq n$. This does not happen with regularization, because we always collect $y$ at time $n$ in Algorithm 1, but not at time $2n_t - 1$, thus the algorithm always detects the difference between $h^1, h^{n_1}$ after time $n_1$.

VI. EXPERIMENTS AND INSIGHTS

A. EXPERIMENTS WITH SYNTHETIC DATA

When does regularization beat least-squares? Low-order slow-decay systems (Fig. 3): We use an experiment with synthetic data to answer this question. So far, we showed that for fixed Hankel size, nuclear norm regularization requires less data than unregularized least-squares especially when the Hankel size is set to be large (Table 3). However, for least squares, one can choose to use a smaller Hankel size with $n \approx R$, so that we solve a problem of small dimension compared to $n \gg R$. We ask if there is a scenario in which fine-tuned nuclear norm regularization strictly outperforms fine-tuned least-squares.

In what follows, we discuss a single trajectory scenario. An advantage of the Hankel-regularization is that, it addresses scenarios which can benefit from large $n$ while both the sample complexity $T$ and the system order $R$ are small. Given choice of $n$, in a single trajectory setting, least-squares suffers an error of order $\rho(A)^{2n}(1 - \rho(A)^n)^{-1}$ (Thm 3.1 of [4]). This error arises from FIR truncation of impulse response (to $n$ terms) and occurs for both regularized and unregularized algorithms. In essence, due to FIR truncation, the problem effectively incurs an output noise of order $\rho(A)^{2n}(1 - \rho(A)^n)^{-1}$. Thus, if the system decays slowly, i.e., $\rho(A) \approx 1$, we will suffer from significant truncation error.

As an example, suppose sample size is $T = 40$ and $\rho(A) = 0.98$. If the problem is kept over-determined (i.e. $n < 40$), then $n$ will not be large enough to make the truncation error $\rho(A)^{2n}(1 - \rho(A)^n)^{-1} \approx 0.36$ vanishingly small. In contrast, Hankel-regularization can intuitively recover such slowly-decaying systems by choosing a large Hankel size $n$ as its sample complexity is (mostly) independent of $n$ (as in the setting of Theorem 1). This motivates us to compare the performance on recovering systems with low-order slow-decay. In Fig. 3, we set up an order-1 system with a pole at 0.98 and generate single rollout data with size 40. We tune $\lambda$ when applying regularized algorithm with $n = 45$ (as it is safe to choose a large Hankel dimension), whereas in unregularized method, Hankel size cannot be larger than 20 ($n \times n$ Hankel has $2n - 1$ parameters and we need least squares to remain overdetermined). With these in mind, in the first two figures we can see that the best validation error of regularization algorithm is 0.44, which is significantly smaller than the unregularized validation error 0.73. In the third figure, we use regularized least squares with $n = 20$, which also causes large truncation error (due to large $\rho(A)$) compared to the initial choice of $n = 45$ (the first figure). In this case, the best
validation error is 0.56 which is again noticeably worse than the 0.44 error in the first figure.

When the number of variables $2n - 1$ is larger than $T$, the system identification problem is overparameterized and there can be infinitely many impulse responses that achieve zero squared loss on the training dataset. This happens in the first figure when $\lambda \to 0$, and in the second figure when $n$ is large. In this case, regularized algorithm chooses the solution with the smallest Hankel nuclear norm and the least squares chooses the one with smallest $\ell_2$ norm. The first figure has smaller validation error when $1/\lambda$ tends to infinity. We verified that, when regularization weight is 0, the setup in the first figure achieves validation error 0.58, which is better than least squares for $n = 45$ with validation error 1.60. So among the solutions that overfits the training dataset, the one with small Hankel nuclear norm has better generalization performance when the true system is low order.

**B. EXPERIMENTS WITH DAISY DATASET**

Our experiment uses the DaISy dataset [7], where a known input signal (not random) is applied and the resulting noisy output trajectory is measured. Using the input and output matrices

$$
U = \begin{bmatrix}
u_{2n-1}^T & u_{2n-2}^T & \ldots & u_1^T \\
u_{2n}^T & u_{2n-1}^T & \ldots & u_2^T \\
\vdots & \vdots & \ddots & \vdots \\
u_{2n+T-2}^T & u_{2n+T-3}^T & \ldots & u_T^T
\end{bmatrix},
$$

$$
y = [y_{2n-1}, \ldots, y_{2n+T-2}].
$$

we solve the optimization problem (HNN) using single trajectory data.

While the input model is single instead of multiple roll-out, experiments will demonstrate the advantage of Hankel-regularization over least-squares in terms of sample complexity, singular value gap and ease of tuning.

**Large data regime:** Both Hankel and least-squares algorithms work well (Fig. 4): The first two figures in Fig. 4 show the training and validation errors of Hankel-regularized and unregularized methods with hyperparameters $\lambda$ and $n$ respectively. We then choose the best system by tuning the hyperparameters to achieve the smallest validation error. The third figure in Fig. 4 plots the training and validation output sequence of the dataset for these algorithms. We see that with sufficient sample size, the system is recovered well. However, the validation error is more flat as a function of $1/\lambda$ (first figure) whereas it is sensitive to the choice of $n$ (second figure), thus $\lambda$ is easier to tune compared to $n$. If the hyperparameter is very sensitive, it (intuitively) indicates that the algorithm is less robust because small variations in the problem variables can seriously hurt the performance. In the machine learning literature, this is also known as sharpness [58].

**Small data regime:** Hankel-regularization succeeds while least-squares may fail due to overfitting (Fig. 5): The first two figures in Fig. 5 show that the Hankel spectrums of the two algorithms have a notable difference: The system recovered by Hankel-regularization is low-order and has larger singular value gap. The last two figures in Fig. 5 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, Hankel-regularization still returns a solution close to the true system while least-squares cannot recover the system properly.
We also remark that nuclear norm regularized least squares has been empirically compared to subspace-based methods such as the Matlab implementation N4sid, for example see [45] which solves the Hankel nuclear-norm regularized least squares problem with interior-point methods, and [47] which solves the same problem with first-order optimization algorithms. Both papers make observations about the empirical benefits of the nuclear norm regularization versus standard subspace system ID for single-rollout experiments using Daisy benchmark data [7].

Identifying a linear approximation of a nonlinear system with few data samples (Fig. 6): Finally, we show that Hankel-regularization can identify a stable nonlinear system via its linearized approximation as well. We consider the inverted pendulum as the experimental environment. First we use a linearized controller to stabilize the system around the equilibrium, and apply single rollout input to the closed-loop system, which is i.i.d. random input of dimension 1. The dimension of the state is 4, and we observe the output of dimension 1, which is the displacement of the system. We then use the Hankel-regularization and least-squares to estimate the closed-loop system with a linear system model and predict the trajectory using the estimated impulse response. We use $T = 16$ observations for training, and set the dimension to $n = 45$. Fig. 6 shows the singular values and estimated trajectory of these two methods. Despite the nonlinearity of the ground-truth system, the regularized algorithm finds a linear model with order 6 and the predicted output has small error, while the correct order is not visible in the singular value spectrum of the unregularized least-squares.

VII. 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 misspecification of the problem (i.e. fitting impulse response with a much larger length rather than the true order). For least-squares, we provided 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 1, we used a weighted Hankel operator. We expect that directly computing the Gaussian width of the original Hankel operator will also lead to improvements over least-squares. We also hope to extend the results to account for single trajectory analysis and process noise. In both cases, an accurate analysis of the regularized problem will likely lead to new algorithmic insights.

APPENDIX
A. SAMPLE COMPLEXITY FOR MISO AND MIMO PROBLEMS
This section establishes sample complexity bounds for MISO and MIMO systems. The technical argument builds on [2] and extends their results from from SISO case to MISO. We consider recovering a MISO system impulse response. The system is given in (2)), with output size $m = 1$ and the system is order $R$. For multi-rollout case, we only observe the output at time $2n - 1$, and let $u_{2n} = 0$, we have

$$y_{2n-1} = \sum_{i=1}^{2n-2} CA^{n-2-i} Bu_i + Du_{2n-1}. \quad (13)$$

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

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

where each block $h^{(i)} \in \mathbb{R}^{p}$, $\beta \in \mathbb{R}^{p(n-2n-1)}$ is a weighted version of $h$, with weights $\beta^{(i)} = Kh^{(i)}$ and $\beta = \begin{bmatrix} \beta^{(1)\top} & \beta^{(2)\top} & \cdots & \beta^{(2n-1)\top} \end{bmatrix}^\top$. Define the reweighted Hankel map for the same $h$ by

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

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

$$u_i \sim \mathcal{N} (0, K_i^2 I) \quad (14)$$

Now let $U \in \mathbb{R}^{T \times p(n-2n-1)}$, each entry is iid standard Gaussian. Let $y \in \mathbb{R}^T$ be the concatenation of outputs

$$y = \begin{bmatrix} y_1^\top \\ y_2^\top \\ \cdots \\ y_T^\top \end{bmatrix}$$

where $y_i \in \mathbb{R}^m$ is defined in (13). We consider the question

$$\min_{\beta} \left\| \mathcal{G}(\beta') \right\|_* \quad \text{s.t.,} \quad \left\| U\beta' - y \right\|_2 \leq \delta \quad (15)$$

where the norm of overall (state and output) noise is bounded by $\delta$. We will present the following theorem, which generalizes the result of [2] from SISO case to MISO case.
Theorem 5: Let \( \beta \) be the true impulse response. If \( T = \Omega((\sqrt{pR} \log n + \epsilon)^2) \) is the number of output observations, \( C \) is some constant, the solution \( \hat{\beta} \) to (15) satisfies \( \| \beta - \hat{\beta} \|_2 \leq 2\delta/\epsilon \) with probability

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

When the system output is \( y = U\beta + z \) and \( z \) is i.i.d. Gaussian noise with variance \( \sigma^2 \), we have that \( \| \beta - \hat{\beta} \|_2 \leq (\sqrt{pR} + \epsilon)\sigma \log n \) with probability \((\delta/\epsilon)^3\) (Thm 1)

\[
1 - 6\exp\left(-\frac{1}{2} \left( \sqrt{T - 1} - C \left( \sqrt{pR} \log n + \epsilon \right) - \epsilon \right)^2 \right).
\]

This theorem says that when the input dimension is \( p \), the sample complexity is \( O(\sqrt{pR}\log n) \). The proof strongly depends on the following lemma \cite{2}, \cite{61}:

Lemma 1: Define the Gaussian width

\[
w(S) := E_g \left( \sup_{\gamma \in S} \gamma^T g \right)
\]

where \( g \) is standard Gaussian vector of size \( p \). The Gaussian width of the normal cone of (6) and (15) are different up to a constant \cite{62}. Let \( \Phi = T(\beta) \cap S \) where \( S \) is unit sphere. We have

\[
P\left( \min_{\beta \in \Phi} \| U\gamma \|_2 < \epsilon \right) \leq \exp\left(-\frac{1}{2} \left( \sqrt{T - 1} - w(\Phi) - \epsilon \right)^2 \right).
\]

We will present the proof in Appendix B.

MIMO: For MIMO case, we say output size is \( m \). We take each channel of output as a system of at most order \( R \), and solve \( m \) problems

\[
P_i : \min_{\beta_i} \| G_i(\beta_i) \|_*
\]

s.t. \( u_i \cdot [x_i, \ldots, x_m] \leq \delta \), \( y_i \in \mathbb{R}^T \) is the \( i \)th output.

and for each problem we have failure probability equal to (17), which means the total failure probability is

\[
m \exp\left(-\frac{1}{2} \left( \sqrt{T - 1} - w(\Phi) - \epsilon \right)^2 \right)
\]

so we need \( T = O((\sqrt{pR} \log n + \log(m) + \epsilon)^2) \). Let the solution to those optimization problems be \( [x_1^*, \ldots, x_m^*] \), and the true impulse response be \( [\hat{x}_1, \ldots, \hat{x}_m] \), then \( \| [x_1^*, \ldots, x_m^*] - [\hat{x}_1, \ldots, \hat{x}_m] \|_F \leq \sqrt{m\delta}/\epsilon \) with probability

\[
1 - \exp\left(-\frac{1}{2} \left( \sqrt{T - 1} - w(\Phi) - \epsilon \right)^2 \right).
\]

We propose another way of MIMO system identification. For each rollout of input data, the output is \( m \) dimensional, but we take 1 channel of output from the observation and throw away other \( m - 1 \) output. And we uniformly pick among channels and get \( T \) observations for each channel, and in total \( mT \) observations/input rollouts. In this case, when the sample complexity is \( m\sqrt{pR} \log n \) (\( m \) times of before), we can recover the impulse response with Frobenius norm \( \sqrt{m\delta}/\epsilon \) with probability

\[
1 - \exp\left(-\frac{1}{2} \left( \sqrt{T/m - 1} - w(\Phi) - \epsilon \right)^2 \right)
\]

B. GAUSSIAN WIDTH OF REGULARIZATION PROBLEM WITH MISO AND MIMO SYSTEM

The following sequence of theorems/lemmas are similar to [2], and we present them here for the integrity of the proof.

Theorem 6: Let \( \beta \) be the true impulse response. If \( T = \Omega((\sqrt{pR} \log(n) + \epsilon)^2) \) is the number of output observations, \( C \) is some constant, the solution \( \hat{\beta} \) to (15) satisfies \( \| \beta - \hat{\beta} \|_2 \leq 2\delta/\epsilon \) with probability

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

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

Lemma 2: Assume

\[
\min_{\gamma \in \mathcal{I}(\beta)} \| U\gamma \|_2 \geq \epsilon,
\]

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

This is proven in [2], Lemma 1. To prove Theorem 6, we only need lower bound the left hand side with Lemma 2. The following lemma gives the probability that the left hand side is lower bounded.

Lemma 3: Define the Gaussian width

\[
w(S) := E_g \left( \sup_{\gamma \in S} \gamma^T g \right)
\]

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

\[
P\left( \min_{\gamma \in \Phi} \| U\gamma \|_2 < \epsilon \right) \leq \exp\left(-\frac{1}{2} \left( \sqrt{T - 1} - w(\Phi) - \epsilon \right)^2 \right).
\]

(17)

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

Lemma 4: (\cite{2}, Eq.(17)) Let \( \mathcal{I}^*(\beta) \) be the dual cone of \( \mathcal{I}(\beta) \), then

\[
w(\Phi) \leq E \left( \min_{\gamma \in \mathcal{I}^*(\beta)} \| g - \gamma \|_2 \right).
\]

(18)

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

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

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

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

For right hand side, we have

\[
\| \lambda G^*(V_1V_2^T + W) - g \|_2 = \| \lambda G^* G(V_1V_2^T + W) - G(g) \|_F.
\]
\[ \begin{align*}
&= \| \lambda (V_1 V_2^T + W) - G(g) \|_F \\
&+ \| \lambda (I - G^*) (V_1 V_2^T + W) \|_F \\
&\leq \| \lambda (V_1 V_2^T + W) - G(g) \|_F.
\end{align*} \]

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

\[ \{ W | V_2^T W = 0, W V_2 = 0 \} \]

and \( \mathcal{P}_V \) be projection onto its orthogonal complement.

We have

\[ w(\Phi) \leq E \left( \min_{\lambda, W} \| \lambda G^* (V_1 V_2^T + W) - g \|_2 \right). \]

We will upper bound it by the specific choice \( \lambda = \| \mathcal{P}_W (g(g)) \|, W = \mathcal{P}_W (g(g))/\lambda. \)

\[ \begin{align*}
&= \| \lambda (V_1 V_2^T + W) - G(g) \|_F \\
&= \| G(g) - \mathcal{P}_W (g(g)) - \mathcal{P}_W (g(g)) \| V_1 V_2^T \|_F \\
&\leq \| \mathcal{P}_V (g(g)) - \mathcal{P}_W (g(g)) \| V_1 V_2^T \|_F \\
&\leq \| \mathcal{P}_V (g(g)) \|_F + \| \mathcal{P}_W (g(g)) \| V_1 V_2^T \|_F \\
&= \| \mathcal{P}_V (g(g)) \|_F + \sqrt{\mathcal{R}} \| \mathcal{P}_W (g(g)) \|.
\end{align*} \]

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 \( V_2 \in \mathbb{R}^{m \times R} \))

\[ \| \mathcal{P}_V (g(g)) \|_F = \| V_1 V_1^T g(g) + (I - V_1 V_1^T) (g(g) V_2 V_2^T) \|_F \\
\leq \| V_1 V_1^T g(g) \|_F + \| g(g) V_2 V_2^T \|_F \\
\leq 2 \sqrt{\mathcal{R}} \| g(g) \|.
\]

So we get

\[ w(\Phi) \leq 3 \sqrt{\mathcal{R}} \| g(g) \|. \]

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

\[ g^{(i)} = \left[ g^{(i)}_1, \ldots, g^{(i)}_p \right]^T, \]

we rearrange the matrix as

\[ g(g) = [G_1, \ldots, G_p], \]

where

\[ G_i = \left[ g^{(i)}_1, g^{(i)}_2/\sqrt{2}, g^{(i)}_3/\sqrt{3} \ldots \right. \]

and the expectation of operator norm of each block is \( \log(n) \).

Then (note \( v \) below also has a block structure \([v^{(1)}, \ldots, v^{(n)}])\)

\[ \| \tilde{G}(g) \| = \max_{u, v} \frac{u^T \tilde{G}(g) v}{\| u \| \| v \|} = \max_{u, v^{(1)}, \ldots, v^{(n)}} \sum_{i=1}^p \frac{u^T G_i v^{(i)}}{\| u \| \| v \|} \]

\[ \leq \max_{v^{(1)}, \ldots, v^{(p)}} \frac{\sum_{i=1}^p \| v^{(i)} \|}{\| u \| \| v \|} \]

\[ \leq \max_{v^{(1)}, \ldots, v^{(p)}} \| g^{(i)} \| \leq \| g^{(i)} \| \leq \| g^{(i)} \| \]

\[ \| \tilde{G}(g) \| = \| g(g) \|. \]

So we have \( \| g(g) \| = \sqrt{\mathcal{R}} \log(n) \).

\[ w(\Phi) = C \sqrt{\mathcal{R}} \log(n). \]

Back to (17), we want the probability be smaller than 1, and we get

\[ \sqrt{T - 1} = C \sqrt{\mathcal{R}} \log n - \epsilon > 0 \]

thus \( T = O \left( (\sqrt{\mathcal{R}} \log(n) + \epsilon)^2 \right). \)

At the end, we provide a different version of Theorem 6. Theorem 6 in [2] works for adversarial noise with bounded norm. Here we consider the i.i.d. Gaussian noise, and use the result in [59] to establish the following theorem.

**Theorem 7:** Let the system output \( y = U \beta + \xi \) where the entries of \( U \) are i.i.d. Gaussian \( N(0, 1/T \beta) \), \( \beta \) is the true system parameter and \( \xi \sim N(0, \sigma^2) \). Then (15) recovers \( \beta \) with error \( \| \beta - \beta^* \|_2 \leq w(\Phi)(\epsilon/\sqrt{T}) \lesssim \sqrt{\mathcal{R}} \sigma \log(n) \) with high probability.

**Remark 1:** Since the power of \( U \) is \( n \) times that of \( \tilde{U} \) and the variance of \( U \) is \( 1/T \), \( \sigma \xi = \sqrt{n/T} \sigma \), we have \( \| h - h^* \|_2 \leq \| \beta - \beta^* \|_2 \lesssim \sqrt{\mathcal{R}} \sigma \log(n) \).

**C. PROOF OF REGULARIZATION ALGORITHM’S SPECTRAL NORM ERROR (THM. 1)**

We will prove the first case of (7). The second case is a direct application of [2].

**Theorem 8:** We study the problem

\[ \min_{\beta} \frac{1}{2} \| U \beta - y \|^2 + \lambda \| \tilde{G}(\beta) \|_k, \]

in the MISO setting \((m=1, \ p \text{ inputs})\), where \( U \in \mathbb{R}^{T \times (2n-1)p} \).

Let \( \beta \) denote the (weighted) impulse response of the true system which has order \( R \), i.e., \( \text{rank}(g(g)) = R \), and let \( y = U \beta + \xi \) be the measured output, where \( \xi \) is the measurement noise. Finally, denote the minimizer of (19) by \( \beta^* \). Define

\[ J(\beta) : = \left\{ v \mid \left\{ \frac{1}{2} \| U^T \beta - y \|^2 + \lambda \| G(\beta) \|_k \right\} \leq 0 \right\}, \]

\[ : = \left\{ I - U^T U \right\} J(\beta), \]

\( J(\beta) \) is the normal cone at \( \beta \), and \( \Gamma \) is the spectral RSV. If \( \Gamma < 1 \), \( \tilde{\beta} \) satisfies

\[ \| G(\tilde{\beta}) \| \leq \frac{\| U^T \tilde{\xi} \| + \lambda}{1 - \Gamma}. \]

**Lemma 5:** Suppose \( \xi \sim N(0, \sigma \xi) \), \( T \lesssim p \mathcal{R}^2 \log^2(n) \), and \( U \) has iid Gaussian entries with \( E(U U^T) = 1 \). Then, we have that \( E(\Gamma) < 0.5 \), and \( P(\Gamma < 0.5) \geq 1 - O(\log n \sqrt{p}/T) \). In this case \( \| G(\tilde{\beta}) \| \lesssim \sigma \xi \sqrt{p} \log(n) \).

**Remark 2:** To be consistent with the main theorem in the paper, we need to find the relation between \( \sigma \xi \) and SNR, or \( \sigma \).

We do the following computation: (1) \( G(\tilde{\beta}) = \mathcal{H}(h - h) \),

so we are bounding the Hankel spectral norm error here;
(2) Each column of the input is unit norm, so each input is \(N(0, 1/T)\), and the average power of input is \(1/T\); (3) Because of the scaling matrix \(K\), the actual input of \(\hat{U}\) is \(n\) times the power of entries in \(U\). With all above discussion, we have
\[
\sigma_\xi = \sigma \sqrt{n/T},
\]
which results in \(\|\mathcal{G}(\hat{\beta} - \beta)\| \lesssim \sqrt{np\sigma \log n.}
\]

Proof: Now we bound \(\|\mathcal{G}(\hat{\beta} - \beta)\|\) by partitioning it to \(\|\mathcal{G}(I - U^\top U)(\hat{\beta} - \beta)\|\) and \(\|\mathcal{G}(U^\top U(\hat{\beta} - \beta))\|\). We have
\[
\begin{align*}
\|\mathcal{G}(I - U^\top U)(\hat{\beta} - \beta)\| & = \|\mathcal{G}(I - U^\top U)\hat{\beta}^* + \mathcal{G}(\hat{\beta} - \beta)\| \\
& \leq \|\mathcal{G}(I - U^\top U)\hat{\beta}^*\|_2,2,\text{G,\beta(\beta)} \|\mathcal{G}(\hat{\beta} - \beta)\| \\
& = \Gamma \|\mathcal{G}(\hat{\beta} - \beta)\|. \tag{20}
\end{align*}
\]

And then we also have
\[
\|\mathcal{G}(U^\top U(\hat{\beta} - \beta))\| \leq \|\mathcal{G}(U^\top \xi)\| + \lambda. \tag{21}
\]

Combining (20) and (21), we have
\[
\begin{align*}
\|\mathcal{G}(\hat{\beta} - \beta)\| & \leq \|\mathcal{G}(I - U^\top U)(\hat{\beta} - \beta)\| + \|\mathcal{G}(U^\top U(\hat{\beta} - \beta))\| \\
& \leq \Gamma \|\mathcal{G}(\hat{\beta} - \beta)\| + \|\mathcal{G}(U^\top \xi)\| + \lambda.
\end{align*}
\]

or equivalently,
\[
\|\mathcal{G}(\hat{\beta} - \beta)\| \leq \frac{\|\mathcal{G}(U^\top \xi)\| + 2\lambda}{\lambda - 1}, \quad \text{and} \quad \Gamma = \|\mathcal{G}(I - U^\top U)\hat{\beta}^*\|_2,2,\text{G,\beta(\beta)}.
\]

Bounding \(\Gamma\): Denote the SVD of \(\mathcal{G}(\beta) = V_1\Sigma V_2^\top\). Denote projection operators \(P_\mathcal{V}(M) = V_1V_2^\top M + MV_2V_1^\top - V_1V_2^\top MV_2V_2^\top\) and \(P_W(M) = M - P_V(M)\). First we prove a result for later use.
\[
\begin{align*}
\frac{1}{2}\|y - U\beta\|^2 + \lambda \|\mathcal{G}\beta\|_* & \leq \frac{1}{2}\|y - U\hat{\beta}\|^2 + \|\mathcal{G}\beta\|_* \\
\Rightarrow \quad \frac{1}{2}\|y - U\beta\|^2 + \lambda \|\mathcal{G}\beta\|_* & \leq \frac{1}{2}\|\xi\|^2 + \lambda \|\mathcal{G}\beta\|_* \\
\Rightarrow \quad \frac{1}{2}\|U\beta + \xi - U\hat{\beta}\|^2 + \lambda \|\mathcal{G}\beta\|_* & \leq \frac{1}{2}\|\xi\|^2 + \lambda \|\mathcal{G}\beta\|_* \\
\Rightarrow \quad \frac{1}{2}\|U(\beta - \hat{\beta})\|^2 + \|\xi + U(\beta - \hat{\beta})\|^2 + \lambda \|\mathcal{G}\beta\|_* & \leq \frac{1}{2}\|\xi\|^2 + \lambda \|\mathcal{G}\beta\|_* \\
\Rightarrow \quad \lambda \|\mathcal{G}\beta\|_* & \leq \lambda \|\mathcal{G}\beta\|_* + \|\xi + U(\beta - \hat{\beta})\|^2 \\
\Rightarrow \quad \|\mathcal{G}\beta\|_* - \|\mathcal{G}\beta\|_* & \leq \frac{\|\mathcal{G}(U^\top \xi)\|}{\lambda} \|\mathcal{G}(\hat{\beta} - \beta)\|_*(\lambda) \tag{22}
\end{align*}
\]

(22) is an important result to note, and following that,
\[
\|\mathcal{G}(\hat{\beta} - \beta)\|_* \leq \frac{\|\mathcal{G}(U^\top \xi)\|}{\lambda} \|\mathcal{G}(\hat{\beta} - \beta)\|_*
\]

Let \(U\) be iid Gaussian matrix with scaling \(E(U^\top U) = I\). Here we need to study the Gaussian width of the normal cone \(w(\mathcal{J}(\beta))\) of (19). [61] proves that, if (22) is true, and \(\lambda \geq 2\|\mathcal{G}(U^\top \xi)\|\), then the Gaussian width of this set (intersecting with unit ball) is less than 3 times of Gaussian width of \(\|\beta : \|\mathcal{G}(\beta)\| \leq \|\mathcal{G}(\beta)\|_*\), which is \(O(\sqrt{R\log n})\) [2]. A simple bound is that, let \(\delta = \hat{\beta} - \beta\), \(\Gamma\) can be replaced by max \(\|\mathcal{G}(I - U^\top U)\delta\|/\|\mathcal{G}(\delta)\|\) subject to \(\beta \in \mathcal{J}(\beta)\). With (23), we have \(\|\mathcal{P}_W(\beta)\|_* \leq 3\|\mathcal{P}_W(\beta)\|_*\).

Denote \(\sigma = \|\mathcal{G}(\delta)\|, \) we know that
\[
\sigma \geq \max(\|\mathcal{P}_W(\mathcal{G}(\delta))\|, \|\mathcal{P}_V(\mathcal{G}(\delta))\|).
\]

Consequently simple algebra gives that
\[
\|\mathcal{P}_V(\mathcal{G}(\delta))\| \geq \|\mathcal{P}_W(\mathcal{G}(\delta))\|_*/(2\sigma).
\]

the second last inequality comes from (23). Thus if \(\|\mathcal{G}(I - U^\top U)\|_F = O(1/\sqrt{R})\|\delta\|_F\), in other words, \(\|\mathcal{G}(I - U^\top U)\|_F = O(1/\sqrt{R})\|\mathcal{G}(\delta)\|_F\), whenever \(\delta\) in normal cone, we have
\[
\|\mathcal{G}(I - U^\top U)\|_F \leq O(1/\sqrt{R})\|\mathcal{G}(\delta)\|_F \leq \|\mathcal{G}(\delta)\|_F. \tag{24}
\]

so \(\Gamma < 1\). To get this, we need \(\sqrt{T}/w(\mathcal{J}(\beta)) \geq \sqrt{R}\) where \(T \geq pR\log^2 n\) [62], Thm 9.11.1, still not tight in \(R\), but \(O(\log(n, R^2\log^2 n))\) as good as [4] and better than [5], which are \(O(n)\) and \(O(n^2)\) correspondingly. [62], Thm 9.11.1 is a bound in expectation, but it naively turns into high probability bound since \(\Gamma \geq 0\).
D. PROOF OF SUBOPTIMAL RECOVERY GUARANTEE WITH IID INPUT (THM. 2)

We consider the Gaussian width \( w(\Phi) \) defined in this specific case.

Let \( \mathcal{I}^n(h) = \{ (H^*(V + W)) | 1^T W = 0, W 1 = 0, \|W\| \leq 1 \} \), and \( V = \frac{1}{n} \mathbf{1}^T \). We have \( w(\Phi) = E(\min_{1^T W} \| \lambda H^*(V + W) - g \|_2) \). In the instance, \( V = \frac{1}{n} \mathbf{1}^T \) and we take \( W \) such that \( \|W\| \leq 1 \) and \( W 1 = W^T 1 = 0 \).

First, we note that

\[
E \left( \min_{1^T W} \| \lambda H^*(V + W) - g \|_2 \right)
= \frac{1}{2} \left( E \left( \min_{1^T W} \| \lambda H^*(V + W) - g \|_2 | 1^T g \leq 0 \right) + E \left( \min_{1^T W} \| \lambda H^*(V + W) - g \|_2 | 1^T g > 0 \right) \right)
\geq \frac{1}{2} \min_{1^T W} \| \lambda H^*(V + W) - g \|_2 \geq \frac{1}{2} \| \lambda H^*(V + W) - g \|_2 \geq 0.
\]

(25)

Proof strategy: On the case when \( 1^T g \leq 0 \). Denote \( z = \lambda H^*(V + W) - g \), and the vector \( z_{i:k} \) is the first \( k \) entries of \( z \). Then we prove that (1) \( \lambda \leq \| z \|_2 / \sqrt{n} \), (2) \( \| z_{1:k} \|_2 \geq \lambda / 2 \). Then we have

\[
\| z \|_2 \geq \| z_{1:k} \|_2 \geq \lambda / 2 \geq \lambda / \sqrt{n} \geq \lambda / \sqrt{n}.
\]

which suggests \( \| z \|_2 \geq n^{1/6} \).

Lemma 6: Let \( g \) be a standard Gaussian vector of size \( 2n - 1 \) conditioned on \( 1^T g \leq 0 \). Denote \( z = \lambda H^*(V + W) - g \), where \( V = \frac{1}{n} \mathbf{1}^T \), and \( W 1 = W^T 1 = 0 \). Then we have that \( \lambda \leq \| z \|_2 / \sqrt{n} \).

We observe that \( 1^T H^*(X) \) is the summation of every entry in \( X \) for any matrix \( X \). Thus \( 1^T H^*(W) = 0 \) since \( W 1 = 0 \). Conditioned on \( 1^T g \leq 0 \), we have

\[
1^T (\lambda H^*(V + W) - g) \geq \lambda 1^T H^*(V) = \lambda n.
\]

So that \( \| \lambda H^*(V + W) - g \|_2 \geq \lambda \sqrt{n} \). Then \( \| z \|_2 / \sqrt{n} \geq \lambda \), we have proven the first point.

Lemma 7: Let \( g \) be a standard Gaussian vector of size \( 2n - 1 \) conditioned on \( 1^T g \leq 0 \). Denote \( z = \lambda H^*(V + W) - g \), where \( V = \frac{1}{n} \mathbf{1}^T \), and \( W 1 = W^T 1 = 0 \). Let the vector \( z_{i:k} \) is the first \( k \) entries of \( z \). Then we have that \( \| z_{1:k} \|_2 \geq \lambda / \sqrt{2} \).

If \( \| z \|_2 \leq \sqrt{n} \), we observe \( z_{1:1} [\sqrt{n} / \| z \|_2] \), where \( [\sqrt{n} / \| z \|_2] \) is the integer part of \( \sqrt{n} / \| z \|_2 \), and \( z_{1:1} [\sqrt{n} / \| z \|_2] \) is the subset of \( z \) containing the entries from index 1 to \( \sqrt{n} / \| z \|_2 \). When \( i \leq \sqrt{n} / \| z \|_2 \), the \( i \)-th entry of \( H^*(V + W) \), denoted as \( (H^*(V + W))_i \), is summation of 2 terms in \( V \) and \( W \). Since these two matrices have bounded spectral norm 1, then every entry of \( V \)

\[\text{is } 1/n \text{ and every entry of } W \text{ is no larger than } 1. \text{ So}
\]

\[z_i = \lambda (H^*(V + W))_i - g_i = \pm (1 + 1/n) i \lambda - g_i \]

\(\epsilon \leq \pm (1 + 1/n) \| z \|_2 \). We denote \( \begin{bmatrix} 1, \ldots, \sqrt{n} / \| z \|_2 \end{bmatrix} \) as \( \mathbf{r} \) the vector whose \( i \)th entry is \( \sqrt{i} / \| z \|_2 \), and get

\[
\| z_{1:1} [\sqrt{n} / \| z \|_2] \|_2 \geq \sqrt{1/2} \| z \|_2 \geq (1 + 1/n) \| z \|_2 \]

\[
\geq \sqrt{1/2} \| z \|_2 \geq (1 + 1/n)^{1/2} \sqrt{1/2} \| z \|_2 \geq (1 + 1/n)^{1/2} \sqrt{1/2} \| z \|_2 \geq (1 + 1/n)^{1/2} \sqrt{1/2} .
\]

Note that the first term is larger than the second term, so we have \( \| z_{1:1} [\sqrt{n} / \| z \|_2] \|_2 \geq C_1 \| z \|_2 \) for some constant \( C_1 \). Note this is the norm of a part of \( z \), which is smaller than the norm of \( z \), so we have \( C_1 \| z \|_2 \geq \| z \|_2 \). So that \( \| z \|_2 \geq (1 + 1/n)^{1/2} \) and we have bounded the quantity (25).

E. PROOF OF LEAST SQUARES’ SPECTRAL NORM ERROR (THM. 3)

Theorem 9: Denote the discrete Fourier transform matrix by \( F \). Denote \( \delta(t) \in \mathbb{R}^T, i = 1, \ldots, m \) as the noise that corresponds to each dimension of output. The solution \( \hat{h} \) of

\[
\hat{h} := h + \mathbf{U}^T z = \min_{h'} \| \mathbf{U} h' - y \|_F^2.
\]

obeys \( \| \hat{h} - h \|_F^2 \leq \| \mathbf{U} \|_F / \sigma_{\min}(\mathbf{U}) \), and \( \| \mathbf{H}(\hat{h} - h) \| \leq \| \mathbf{F} \mathbf{U} \|_F \| z_{1:1} \|_\infty \). 

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 (26), \( \mathbf{U} \in \mathbb{R}^{r \times (2n-1)p} \), each row is the input in a time interval of length \( 2n - 1 \). The impulse response is \( h \in \mathbb{R}^{2n-1p} \) and output and noise are \( y, z \in \mathbb{R}^{2n-1p} \), where each column corresponds to one channel of the output. Each row of \( y \) is an output observation at a single time point. Each column of \( z \) is a column of the noise, meaning one channel of the noise contaminating all observations at this channel.

(26) has close form solution and we have \( \| \hat{h} - h \| = \| \mathbf{U}^T z \| \leq \| z \|_F / \sigma_{\min}(\mathbf{U}) \). To get the error bound in Hankel matrix, we can denote \( \tilde{z} := \mathbf{U}^T z \) is \( (\mathbf{U}^T \mathbf{U})^{-1} \mathbf{U}^T z \), and

\[
H_{\mathbf{z}} := \begin{bmatrix}
\tilde{z}_1 & \tilde{z}_2 & \ldots & \tilde{z}_{2n-1} \\
\tilde{z}_2 & \tilde{z}_3 & \ldots & \tilde{z}_1 \\
\vdots & \vdots & \ddots & \vdots \\
\tilde{z}_{2n-2} & \tilde{z}_{2n-1} & \ldots & \tilde{z}_1
\end{bmatrix}.
\]

If \( m = 1, \tilde{z} \in \mathbb{R}^{2n-1p} \) is a vector [63], Sec. 4) proves that \( H_{\mathbf{z}} = F^{-1} diag(F z) F \). So the spectral norm error is bounded by \( \| \mathbf{U} \|_F \| z \|_2 = \| F \tilde{z} \|_\infty \).
If \( m > 1 \), all columns of \( z \) are independent, so \( H_z \) can be seen as concatenation of \( m \) independent noise matrices where each satisfies the previous derivation.

Next we prove Thm. 3.

**Proof:** We use Theorem 9. First let \( m = 1 \). The covariance of \( F \xi \) is \( F(U^T \bar{U})^{-1} F^T \). If \( T \gg n \), it’s proven in [62] that \( \frac{1}{n} I \preceq \bar{U}^T \bar{U} \preceq \frac{2n}{2m+1} I \). Then \( \frac{n}{2} I \preceq F(U^T \bar{U})^{-1} F^T \preceq \frac{2n}{2m+1} I \). So \( \|F \xi\|_\infty \) should scale as \( O(\sigma_F \sqrt{T \log n}) \), and then \( \|H(z)\|_2 \leq \|H_2\|_2 \leq \|F \xi\|_\infty = O(\sigma_F \sqrt{T \log n}) \).

If \( m > 1 \), then by concatenation we simply bound the spectral 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.

**F. PROOF OF END TO END BOUND OF SYSTEM IDENTIFICATION AND MODEL SELECTION (THM. 4)**

We select \( \delta > 0 \) such that \( T_{val} \geq \frac{1}{\delta^2} \log \frac{n_1}{p} \) and denote \( a_1 = 1 - \frac{\delta}{\sqrt{2}} \), \( a_2 = 1 + \frac{\delta}{\sqrt{2}} \). Then we have \( a_2/a_1 = 1 + \delta \). Let \( T_0 = \max\{1, T/T_{val}(R \log^2 n)\} \). We will show that

\[
\frac{\|\hat{h} - h\|_2}{\sqrt{2}} \leq \|H(\hat{h} - h)\| \\
\leq \left\{ \begin{array}{l}
(1 + T_0^{1/4}) \frac{\alpha_1}{a_1} \frac{\sqrt{n p}}{\sqrt{\text{SNR}}} \log(n) \text{, if } T \geq \min(R^2, n) \\
(1 + T_0^{1/4}) \frac{\alpha_1}{a_1} \frac{\sqrt{R n p}}{\sqrt{\text{SNR}}} \log(n) \text{, if } R \leq T \leq \min(R^2, n).
\end{array} \right.
\]

(27)

Note that we will need \( T_0^{1/4} \delta \leq 1 \) from our choice of \( T_{val} \) in the theorem, so the bound is sufficient for the theorem. This will be used later to calculate \( \delta \) in (30).

We use the change of variable as in (6). We learn the parameter \( \beta \) with different \( \lambda \), and get different estimations \( \hat{\beta} \) which is a function of \( \lambda \). To be more explicit, let \( \hat{\beta}(\lambda) \) be the estimator associated with a certain regularization parameter \( \lambda \). Among all the estimators, we choose the solution with the smallest validation error, which is denoted as

\[
\hat{\beta} = \arg\min_{\beta(\lambda)} \|U_{val}(\hat{\beta}(\lambda) - y_{val}\|_2^2
\]

Denote the noise in validation data as \( \xi_{val} \). We have that

\[
\|U_{val}(\hat{\beta} - y_{val}\|_2^2 = \|U_{val}(\hat{\beta} - \beta) - \xi_{val}\|_2^2 \equiv \|U_{val}(\hat{\beta} - \beta)\|_2^2 + \|\xi_{val}\|_2^2 - 2 \xi_{val}^T U_{val}(\hat{\beta} - \beta).
\]

(28)

In this formulation, \( \|\xi_{val}\|_2^2 \) in (28) is regarded as fixed among all validation instances, and we study the other two terms. Since \( U_{val} \) is normalized that each entry is i.i.d. \( N(0, 1/T_{val}) \), we have \( E\|U_{val}(\hat{\beta} - \beta)\|_2^2 = \|\beta - \beta\|_2^2 \).

The quantity \( \xi_{val}^T U_{val}(\hat{\beta} - \beta) \) is zero mean and we know that \( U_{val}(\hat{\beta} - \beta) \sim N(0, \frac{\|\beta - \beta\|_2^2}{T_{val}} I) \). Thus the variance of \( \xi_{val}^T U_{val}(\hat{\beta} - \beta) \) is bounded by \( O(\sigma_{\xi_{val}}^2 \|\beta - \beta\|_2^2/T_{val}) \) (the distribution of the inner product is sub-exponential). We know that

\[
\|\hat{\beta} - \beta\|_2 \approx \sqrt{\frac{R \log^2 n}{T}} \|\xi\|_2 = \sqrt{\frac{R \log^2 n}{T}} \sqrt{T_{val} \sigma_{\xi_{val}}^2}.
\]

**Case 1:** If \( T_{val} \geq \frac{T}{R \log^2 (\frac{n_1}{p})} \), we have that \( \|\hat{\beta} - \beta\|_2 \geq \sigma_{\xi_{val}} \).

Suppose the number of validated parameters \( \lambda \) is \(|\lambda|\) and we need to choose the size of validation data. With different validation data size \( T_{val} \), the variance of \( U_{val}(\hat{\beta} - \beta)\|_2^2 \) decreases with rate \( 1/T_{val} \).

We fix factors \( a_1, a_2 \), such that with high probability, for all choices of \( \beta \), \( \|U_{val}(\hat{\beta} - \beta)\|_2^2 - 2 \xi_{val}^T U_{val}(\hat{\beta} - \beta) \) is in the set \( (a_1 \|\beta - \beta\|_2^2, a_2 \|\beta - \beta\|_2^2) \). We know that: the terms \( \|U_{val}(\hat{\beta} - \beta)\|_2^2 \) and \( 2 \xi_{val}^T U_{val}(\hat{\beta} - \beta) \) are subexponential; The mean of \( \|U_{val}(\hat{\beta} - \beta)\|_2^2 \) is \( \|\beta - \beta\|_2^2 \) and the variance is \( O(\|\beta - \beta\|_2^4/T_{val}) \); The mean of \( 2 \xi_{val}^T U_{val}(\hat{\beta} - \beta) \) is 0 and the variance is \( O(\|\beta - \beta\|_2^2/T_{val}) \) (Note that \( \|\beta - \beta\|_2 \geq \sigma_{\xi_{val}} \) in this case).

By Bernstein bound [64], Prop. 5.16], we know that the probability that the quantity of (29) is not between \( (a_1, a_2) \cdot \|\beta - \beta\|_2^2 \) is at most \( \exp(-\min(a_i - 1)^2 T_{val}) \) where \( (a_i - 1)^2 \approx \delta^2 \).

Hence there exists a constant \( c \) such that for every choice of \( \lambda \),

\[
\Pr \left( \|U_{val}(\hat{\beta} - \beta)\|_2^2 - 2 \xi_{val}^T U_{val}(\hat{\beta} - \beta) \notin (a_1, a_2) \cdot \|\beta - \beta\|_2^2 \right) < \exp(-c \delta^2 T_{val}).
\]

(29)

We choose \( P \) to be an upper bound on the probability that any of the event in (29) happens. If all \(|\lambda|\) validations corresponding to \( \lambda \) succeed, then we use the union bound on (29) and solve for \(|\lambda|\) \( \exp(-c \delta^2 T_{val}) < P \). Thus we set \( T_{val} = \max \left\{ \frac{T}{R \log^2 (\frac{n_1}{p})}, \frac{T_{0}}{c^2} \log \left( \frac{\sqrt{\text{SNR}}}{\sqrt{n_1}} \right) \right\} \), so that (10) holds with probability \( 1 - P \).

**Case 2:** If \( T_{val} \leq \frac{T}{R \log^2 (\frac{n_1}{p})} \), then we denote \( T_0 = T/(T_{val} R \log^2 n) \), with similar derivation as above, we know that the mean of \( \|U_{val}(\hat{\beta} - \beta)\|_2^2 \) is \( \|\beta - \beta\|_2^2 \) and the variance is \( O(\|\beta - \beta\|_4^4/T_{val}) \); The mean of \( 2 \xi_{val}^T U_{val}(\hat{\beta} - \beta) \) is 0 and the variance is \( O(T_0 \|\beta - \beta\|_4^2/T_{val}) \). Thus, similar to (29),

\[
\Pr \left( \|U_{val}(\hat{\beta} - \beta)\|_2^2 - 2 \xi_{val}^T U_{val}(\hat{\beta} - \beta) \notin (a_1, a_2) \cdot \sqrt{T_0} \|\beta - \beta\|_2 \right) < \exp(-c \delta^2 T_{val}) .
\]

The following steps are same as the first case, and the error is multiplied by \( T_0^{1/4} \) compared to the first case.

At the end, we will need to argue about the lower bound for \( T_{val} \). We used two inequalities in the proof above:

\[
T_{val} \geq \frac{1}{\delta^2} \log \left( \frac{|\lambda|}{P} \right), T_{0}^{1/4} \delta \leq 1.
\]
Algorithm 1 Sequential System Identification & Model Selection

Require: \( T \) training rollouts: Input features \( \bar{U} \), outputs \( y_{T_{val}} \) validation rollouts: Input features \( \bar{U}_{val} \), outputs \( y_{val} \)
Hyperparameters: Hankel dimension \( n \), candidate set \( \Lambda \).

% Train a model for each \( \lambda \in \Lambda \)
for \( \lambda \in \Lambda \) do
    Training phase: Solve (HNN). Record the solution \( \hat{h}_{\lambda} \).
end for
% Return the minimum validation error
Model selection: \( \hat{h} = \arg \min_{\hat{h}_{\lambda}, \lambda \in \Lambda} \| \bar{U}_{val} \hat{h}_{\lambda} - y_{val} \|_2^2 \).

They are equivalent to

\[
T_{val} \gtrsim \frac{1}{\delta^2} \log \left( \frac{|\Lambda|}{P} \right), \quad T_{val} \gtrsim \frac{\delta^4 T}{R \log^2(n)}. \tag{30}
\]

Setting the right hand side to be equal, we have

\[
\delta^2 = \left( T^{-1} \log \left( \frac{|\Lambda|}{P} \right) R \log^2(n) \right)^{1/3}.
\]

Plugging it into any lower bound for \( T_{val} \) in (30), we get the bound in the main theorem.

G. EXPERIMENTS WITH SYNTHETIC DATA

We present Algorithm 1 for model selection which was omitted from the main text.

In this subsection, we check Theorem 1 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 true order. The input is multiple rollout, shaped Gaussian exactly as in Theorem 1, 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 shaping by \( K^{-1} \), \( \mathbf{E}(U^T U) = I \). The number of observations is 30 (undetermined for least square). We tune the regularized model by training with different weight \( \lambda \) of regularization, and tune the least square model by changing the size of Hankel matrix and observe the validation/test error.

In Fig. 7, we show the case with insufficient and noisy data. Since the number of samples is undetermined for least squares, even if we take the solution with the smallest \( \ell_2 \) norm in impulse response, it suffers large error on validation and test set. We can see that the regularized algorithm is robust to noise, while least square algorithm remains bad. It indicates that, the solution with the least Hankel nuclear norm behaves better than least Frobenius norm solution in low sample complexity case which verifies Theorem 1.

REFERENCES

[1] Y. Sun, S. Oymak, and M. Fazel, “Finite sample system identification: Optimal rates and the role of regularization,” in Proc. 2nd Conf. Learn. Dyn. Control, 2020, pp. 16–25.
[2] J.-F. Cai, X. Qu, W. Xu, and G.-B. Ye, “Robust recovery of complex exponential signals from random gaussian projections via low rank hankel matrix reconstruction,” Appl. Comput. Harmon. Anal., vol. 41, no. 2, pp. 470–490, 2016.
[3] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi, “Simultaneously structured models with application to sparse and low-rank matrices,” IEEE Trans. Inf. Theory, vol. 61, no. 5, pp. 2886–2908, May 2015.
[4] S. Oymak and N. Ozay, “Non-asymptotic identification of LTI systems from a single trajectory,” in Proc. Amer. Control Conf., 2019, pp. 5655–5661.
[5] T. Sarkar, A. Rahklin, and M. A. Dahleh, “Nonparametric finite time LTI system identification,” 2019, arXiv:1902.01848.
[6] S. Tu, R. Boczar, A. Packard, and B. Recht, “Non-asymptotic analysis of robust control from coarse-grained identification,” 2017, arXiv:1707.04791.
[7] B. De Moor, P. De Gersem, B. De Schutter, and W. Favoreel, “DAISY: A database for identification of systems,” J. A., vol. 38, pp. 4–5, 1997.
[8] S. Oymak and N. Ozay, “Revisiting ho-kalman based system identification: Robustness and finite-sample analysis,” IEEE Trans. Autom. Control, vol. 67, no. 4, pp. 1914–1928, Apr. 2022.
[9] J. A. Cadzow, “Signal enhancement—a composite property mapping algorithm,” IEEE Trans. Acoust., Speech, Signal Process., vol. 36, no. 1, pp. 49–62, Jan. 1988.
[10] J. Gillard, “Cadzow’s basic algorithm, alternating projections and singular spectrum analysis,” Statist. Interface, vol. 3, no. 3, pp. 335–343, 2010.
[11] T. K. Sarkar and O. Pereira, “Using the matrix pencil method to estimate the parameters of a sum of complex exponentials,” IEEE Antennas Propag. Mag., vol. 37, no. 1, pp. 48–55, Feb. 1995.
[12] B. Ho and R. E. Kälmän, “Effective construction of linear state-variable models from input/output functions,” Automatica, vol. 14, no. 1–2, pp. 545–548, 1966.
[13] L. Ljung, System Identification: Theory for the User, 2nd. Englewood Cliffs, NJ, USA: PTR Prentice Hall, pp. 1–14, 1999.
[14] P. Van Overschee and B. De Moor, “A unifying theorem for three subspace system identification algorithms,” Automatica, vol. 31, no. 12, pp. 1853–1864, Dec. 1995.
[15] P. Van Overschee and B. De Moor, Subspace identification for linear systems: Theory–Implementation–Applications. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2012.
[16] T. McKelvey, H. Akçay, and L. Ljung, “Subspace-based multivariable system identification from frequency response data,” IEEE Trans. Autom. Control, vol. 41, no. 7, pp. 960–979, Jul. 1996.
[17] Y. Abbassi-Yadkori, D. Pál, and C. Szepesvári, “Online least squares estimation with self-normalized processes: An application to bandit problems,” 2011, arXiv:1102.2670.
[18] Y. Abbasi-Yadkori and C. Szepesvári, “Regret bounds for the adaptive control of linear quadratic systems,” in Proc. 24th Annua. Conf. Learn. Theory, 2011, pp. 1–26.

[19] M. Simchowitz, H. Mannor, S. Tu, M. I. Jordan, and B. Recht, “Learning without mixing: Towards a sharp analysis of linear system identification,” in Proc. Conf. Learn. Theory, 2018, pp. 439–473.

[20] T. Sarkar and A. Rakhlík, “Near optimal finite time identification of arbitrary linear dynamical systems,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 5610–5618.

[21] S. Fattahi, “Learning partially observed linear dynamical systems from logarithmic number of samples,” in Proc. Learn. Dyn. Control, 2021, pp. 60–72.

[22] Y. Zheng and N. Li, “Non-asymptotic identification of linear dynamical systems using multiple trajectories,” IEEE Control Syst. Lett., vol. 5, no. 5, pp. 1693–1698, Nov. 2021.

[23] E. Hazan, H. Lee, K. Singh, C. Zhang, and Y. Zhang, “Spectral filtering for general linear dynamical systems,” in Proc. Adv. Neural Inf. Process. Syst, 2018, pp. 4634–4643.

[24] M. Simchowitz, R. Boczar, and B. Recht, “Learning linear dynamical systems with semi-parametric least squares,” in Proc. Learn. Theory, 2019, pp. 2714–2802.

[25] A. Tsiatis and G. J. Pappas, “Finite sample analysis of stochastic system identification,” in Proc. IEEE 58th Conf. Decis. Control, 2019, pp. 3648–3654.

[26] L. Jńg, T. Chen, and B. Mu, “A shift in paradigm for system identification,” Int. J. Control, vol. 93, no. 2, pp. 173–180, 2020.

[27] A. Chiuseo, “Regularization and bayesian learning in dynamical systems: Past, present and future,” Annu. Rev. Control, vol. 41, pp. 24–38, 2016.

[28] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Jńg, “Kernel methods in system identification, machine learning and function estimation: A survey,” Automatica, vol. 50, no. 3, pp. 657–682, 2014.

[29] G. Prando, A. Chiuseo, and G. Pillonetto, “Maximum entropy vector kernels for mimo system identification,” Automatica, vol. 79, pp. 326–339, 2017.

[30] M. C. Campi and E. Weyer, “Guaranteed non-asymptotic confidence regions in system identification,” Automatica, vol. 41, no. 10, pp. 1751–1764, 2005.

[31] M. C. Campi, G. Calafiore, and S. Garatti, “Interval predictor models: Identification and reliability,” Automatica, vol. 45, no. 2, pp. 382–392, 2009.

[32] G. Pillonetto, T. Chen, A. Chiuseo, G. De Nicolao, and L. Jńg, Regularized System Identification: Learning Dynamic Models from Data. Berlin, Germany: Springer, 2022.

[33] G. Pillonetto and A. Scampichio, “Sample complexity and minimax properties of exponentially stable regularized estimators,” IEEE Trans. Autom. Control, vol. 67, no. 5, pp. 2330–2342, May 2022.

[34] G. Pillonetto, T. Chen, A. Chiuseo, G. De Nicolao, and L. Jńg, “Regularized linear system identification using atomic, nuclear and kernel-based norms: The role of the stability constraint,” Automatica, vol. 69, pp. 137–149, 2016.

[35] W. Xu, J. Yi, S. Dasgupta, J.-F. Cai, M. Jacob, and M. Cho, “Sep} ration-free super-resolution from compressed measurements is possible: An orthonormal atomic norm minimization approach,” in Proc. IEEE Int. Symp. Inf. Theory, 2018, pp. 76–80.

[36] M. Elad, P. Milanfar, and G. H. Golub, “Shape from moments—an estimation theory perspective,” IEEE Trans. Signal Process., vol. 52, no. 7, pp. 1814–1829, Jul. 2004.

[37] T. Ding, M. Sznaier, and O. I. Camps, “A rank minimization approach to video inpainting,” in Proc. IEEE 11th Int. Conf. Comput. Vis., 2007, pp. 1–8.

[38] B. Recht, M. Fazel, and P. A. Parrilo, “ Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization,” SIAM Rev., vol. 52, no. 3, pp. 471–501, 2010.

[39] E. J. Candes and Y. Plan, “Matrix completion with noise,” Proc. IEEE, vol. 98, no. 6, pp. 925–936, Jun. 2010.

[40] M. Fazel, H. Hindi, and S. P. Boyd, “A rank minimization heuristic with application to minimum order system approximation,” in Proc. Amer. Control Conf., 2001, vol. 6, pp. 4734–4739.

[41] M. Fazel, “Matrix rank minimization with applications,” Ph.D. dissertation, Stanford Univ., Stanford, CA, USA, 2002.

[42] A. Hansson, Z. Liu, and L. Vandenberghe, “Subspace system identification via weighted nuclear norm optimization,” in Proc. IEEE 51st IEEE Conf. Decis. Control, 2012, pp. 3439–3444.
SAMET OYMAK (Member, IEEE) received the M.S. and Ph.D. degrees from the California Institute of Technology, Pasadena, CA, USA. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of California, Riverside (UCR), CA. Before joining UCR, he spent time at Google and in the financial industry, and prior to that he was a Fellow with the Simons Institute, Berkeley, CA, and a Postdoctoral Scholar with UC Berkeley. His research interests include the mathematical foundations of data science and machine learning by using tools from optimization and statistics, mathematical optimization, reinforcement learning, deep learning theory, and high-dimensional problems. He was the recipient of the Wilts Prize for the Best Thesis in electrical engineering at the California Institute of Technology.

MARYAM FAZEL (Senior Member, IEEE) received the B.S. degree from the Sharif University of Technology, Tehran, Iran, and the M.S. and Ph.D. degrees from Stanford University, Stanford, CA, USA. She is currently the Moorthy Family Professor of Electrical and Computer Engineering with the University of Washington (UW), Seattle, WA, USA, with adjunct appointments in computer science and engineering, mathematics, and statistics. Before joining UW, she was a Postdoctoral Scholar with Caltech, Pasadena, CA. She was the recipient of the NSF Career Award, UWEE Outstanding Teaching Award, and the UAI conference Best Student Paper Award with her student. She is currently the Director of the Institute for Foundations of Data Science, a multi-university NSF TRIPODS Institute. Prof. Fazel is also on the Editorial board of the MOS-SIAM Book Series on Optimization, and is an Associate Editor for the SIAM Journal on Mathematics of Data Science.