Nonconvex Linear System Identification with Minimal State Representation

Uday Kiran Reddy Tadipatri UKREDDY@SEAS.UPENN.EDU
University of Pennsylvania, USA

Benjamin D. Haeffele HAEFFELE@SEAS.UPENN.EDU
University of Pennsylvania, USA

Joshua Agterberg JAGT@ILLINOIS.EDU
University of Illinois Urbana-Champaign, USA

Ingvar Ziemann INGVARZ@SEAS.UPENN.EDU
University of Pennsylvania, USA

René Vidal VIDALR@SEAS.UPENN.EDU
University of Pennsylvania, USA

Editors: N. Ozay, L. Balzano, D. Panagou, A. Abate

Abstract

Low-order linear System IDentification (SysID) addresses the challenge of estimating the parameters of a linear dynamical system from finite samples of observations and control inputs with minimal state representation. Traditional approaches often utilize Hankel-rank minimization, which relies on convex relaxations that can require numerous, costly singular value decompositions (SVDs) to optimize. In this work, we propose two nonconvex reformulations to tackle low-order SysID (i) Burer-Monterio (BM) factorization of the Hankel matrix for efficient nuclear norm minimization, and (ii) optimizing directly over system parameters for real, diagonalizable systems with an atomic norm style decomposition. These reformulations circumvent the need for repeated heavy SVD computations, significantly improving computational efficiency. Moreover, we prove that optimizing directly over the system parameters yields lower statistical error rates, and lower sample complexities that do not scale linearly with trajectory length like in Hankel-nuclear norm minimization. Additionally, while our proposed formulations are nonconvex, we provide theoretical guarantees of achieving global optimality in polynomial time. Finally, we demonstrate algorithms that solve these nonconvex programs and validate our theoretical claims on synthetic data.

Keywords: System Identification, Hankel rank minimization, Sample Complexity.

1. Introduction

We consider the linear time-invariant system (LS) with time $t \in \mathbb{N}$, $\mathbf{x}_t \in \mathbb{R}^{n_x}$ as hidden states, $\mathbf{u}_t \in \mathbb{R}^{n_u}$ as control inputs, $\mathbf{y}_t \in \mathbb{R}^{n_y}$ as outputs, $\zeta_t \in \mathbb{R}^{n_y}$ as output noise, and system parameters $(A \in \mathbb{R}^{n_x \times n_x}, B \in \mathbb{R}^{n_x \times n_u}, C \in \mathbb{R}^{n_y \times n_x}, D \in \mathbb{R}^{n_y \times n_u})$ described as

$$\begin{aligned} \mathbf{x}_{t+1} &= A\mathbf{x}_t + B\mathbf{u}_t; \mathbf{x}_1 = 0, \\ \mathbf{y}_t &= C\mathbf{x}_t + D\mathbf{u}_t + \zeta_t. \end{aligned} \tag{LS}$$

Linear *System IDentification* (SysID) aims to estimate these parameters $(A,\ B,\ C,\ D)$ using finite rollouts with finite length trajectories, $\{(\mathbf{u}_t^i, \mathbf{y}_t^i)_{t=1}^{2L+2}\}_{i=1}^N$. However, the true system order or

state dimension n_x is typically unknown. It is of practical interest to find systems that are *minimal order* as such models provide faster response times, simplified control designs, and improved robustness, while mitigating overfitting during the learning phase. This task can be formulated as a dimension minimization problem under an L_2 norm constraint (see Program (P0)). However, solving this problem is known to be NP-HARD (Vandenberghe and Boyd, 1996).

In classical linear system theory, the order of the system is determined by the rank of the Hankel matrix constructed from the Markov parameters $(CB, CAB, \ldots, CA^{2L+1}B)$ (Sontag, 2013, Corollary 6.5.7). Convex relaxations, such as minimizing the Hankel nuclear norm instead of the rank, have been proposed to make this problem computationally tractable (Fazel et al., 2003). However, this presents two key challenges: (i) **Computational complexity**: The number of parameters scales linearly with the trajectory length L, making nuclear norm minimization increasingly difficult. (ii) **Statistical inefficiency**: The high dimensionality leads to worse statistical error rates and larger sample complexity requirements (Oymak and Ozay, 2019).

In the literature on matrix sensing, nuclear norm minimization is tackled via nonconvex reparameterization using its variational form with Frobenius norms (Burer and Monteiro, 2003). This reformulation enables the use of efficient first-order optimization methods, such as gradient descent and Nesterov acceleration (Nesterov, 1983), which converge in $\mathcal{O}(1/T)$, and $\mathcal{O}(1/T^2)$, respectively, where T is the number of iterations. In contrast, direct nuclear norm minimization often requires computing proximal operators (Parikh et al., 2014) that depend on SVD, making it less computationally efficient. In this work, we propose two reformulations, the first of which is Hankel nuclear norm minimization for SysID (see Program (P1)) using the BM factorization (Burer and Monteiro, 2003). Despite being nonconvex, this formulation eliminates the need for frequent SVD computations, significantly enhancing computational efficiency, and can still be provably globally optimized in polynomial time.

While this first reformulation offers clear advantages in terms of optimization efficiency, the statistical error rates and sample complexities remain comparable to those of Hankel nuclear norm minimization. In particular, the statistical error rates scale with trajectory length, despite the number of system parameters remaining constant. Our second reformulation (see Program (P2)) proposes a method of SysID of a real, diagonalizable system where we perform optimization directly over the system parameters space (A, B, C, D) with a structured regularization similar to an atomic norm decomposition. Although this reformulation is also nonconvex, we show that it can be efficiently optimized to global optimality. Furthermore, this approach achieves improved statistical error rates and reduced sample complexity compared to existing Hankel nuclear norm minimization heuristics.

Contributions. Our key contributions are as follows:

1. **Nonconvex Relaxations for low-order SysID.** We propose two nonconvex reformulations (i) for general linear systems via BM re-parametrization (P1) of the Hankel matrix, (ii) by directly optimizing over system parameters (P2), restricted to real, diagonalizable systems. These nonconvex problems are first of their kind and serve as relaxations to the NP-HARD problem (P0).

2. **Global Optimality Guarantees.** We provide guarantees of global optimality for the two nonconvex reformulations when they are solved using first-order optimization procedures. Furthermore, we affirm that these nonconvex problems can be solved in polynomial time. The second reformulation overcomes the quadratic dependency on trajectory length in the computational efficiency of existing methods.

3. **Statistical Error Rates and Sample Complexities.** We provide statistical error rates and sample complexities for both the reformulations while the first reformulation scales poorly compared to the second due to its linear dependency on trajectory length.
4. **Numerical evidence.** We propose algorithms to tackle these nonconvex programs, provide numerical simulations that corroborate our theoretical insights. Under a fixed compute budget, our algorithms outperform existing methods.

These results highlight that for real, diagonalizable systems the reformulation (P2) achieves both superior statistical efficiency and reduced sample complexity compared to (P1), Sun et al. (2022), and Lee (2022).

Organization. First, in §2 we introduce the problem and mathematical formulations. Next, we move onto §3 where we present optimality certificates for the nonconvex formulations. Then, in §4 we present statistical error rates and sample complexities for each of the formulations. Later, we move to §5 that presents numerical experimentation and comparison with existing methods. Finally, we conclude and discuss the future work in §6. Further related works and proofs for all the mathematical statements can be found in Appendix of our arXiv version.

Notation. For a matrix A, $[A]_{i_1:i_2,j_1:j_2}$ denotes its submatrix from rows i_1–i_2 and columns j_1–j_2; $[[A_{i,j}]_{i\in\mathcal{I}}]_{j\in\mathcal{J}}$ stacks blocks $A_{i,j}$ row- and column-wise over $\mathcal{I}$ and $\mathcal{J}$. The superscript $\dagger$ denotes the adjoint for operators and the Moore–Penrose pseudoinverse for matrices. The inequality $f(x) \gtrsim g(x)$ or $f(x) \geq \mathcal{O}(g(x))$ means there exists a positive constant c such that $f(x) \geq cg(x)$ for all x. For some $\alpha > 0$, $(\ln x)^\alpha$ is denoted as $\mathsf{polylog}(x)$, and $\tilde{\mathcal{O}}(x)$ denotes $\mathcal{O}(x \cdot \mathsf{polylog}(x))$.

2. Problem Formulation

Given a noisy linear time-invariant system (LS), we are interested in special approximation of outputs $\{\mathbf{y}_t\}_{t=1}^{2L+2}$ for given inputs $\{\mathbf{u}_t\}_{t=1}^{2L+2}$ that has low-order, or smaller state dimension, n_x. Formally, define impulse response [1] $G(A,B,C,D) \in \mathbb{R}^{2(L+1)n_y \times 2(L+1)n_u}$ where $i,j \in [2(L+1)]$, and

$$[G(A,B,C,D)]_{i,j} = \begin{cases} 0 & i < j, \\ D & i = j, \\ CA^{i-j-1}B & i > j. \end{cases} \in \mathbb{R}^{n_y \times n_u}. \tag{1}$$

Our goal is to solve the optimization problem (P0),

$$\min_{n_x,A,B,C,D} n_x, \quad \text{such that } \frac{1}{4N(L+1)} \sum_{i=1}^{N} \|\mathsf{vec}(Y_i) - G(A,B,C,D)\mathsf{vec}(U_i)\|_F^2 \leq \epsilon, \quad \text{(P0)}$$

where $Y_i := \begin{bmatrix} \mathbf{y}_1^i & \mathbf{y}_2^i & \cdots & \mathbf{y}_{2(L+1)}^i \end{bmatrix} \in \mathbb{R}^{n_y \times 2(L+1)}$, $U_i := \begin{bmatrix} \mathbf{u}_1^i & \mathbf{u}_2^i & \cdots & \mathbf{u}_{2(L+1)}^i \end{bmatrix} \mathbb{R}^{n_y \times 2(L+1)}$, $\mathsf{vec} : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\cdot n\times 1}$ is column-wise stacking operator. It is well known that solving (P0) in general is NP-HARD (Vandenberghe and Boyd, 1996). Numerous convex relaxations of (P0) have been proposed to tackle this issue under certain conditions (beyond the scope of this work), such as minimizing the nuclear norm of the Hankel matrix (Fazel et al., 2003; Recht et al., 2008, 2010). To place this in context, define the Hankel operator $\mathcal{H} : \mathbb{R}^{n_y\times(2L+1)n_u} \to \mathbb{R}^{(L+1)n_y\times(L+1)n_u}$ for a block sequence $\{K_t\}_{t\in[2L+1]} \subset \mathbb{R}^{n_y\times n_u}$ as

1. Due to the trivial estimation of D we often omit its dependency in the impulse response and denote it as $G(A,B,C)$.

$$\forall i,j \in [L+1], \left[\mathcal{H}\left(\begin{bmatrix} K_1 & K_2 & \dots & K_{2L+1}\end{bmatrix}\right)\right]_{1+(i-1)\cdot n_y : i\cdot n_y, 1+(j-1)\cdot n_u : j\cdot n_u} = K_{i+j-1}. \quad (2)$$

In Sun et al. (2022) the authors proposed to solve the reformulated problem

$$\min_{H,D} \frac{1}{2N}\sum_{i=1}^{N} \|\mathbf{y}^i_{2L+1} - \begin{bmatrix} H & D\end{bmatrix} \mathsf{vec}(U_i)\|_F^2 + \lambda \|\mathcal{H}(H)\|_*. \quad \text{(P1')}$$

After obtaining the Markov parameters H they perform the Ho-Kalman procedure (Ho and Kalman, 1966) to obtain the system parameters (A, B, C, D). The matrix H has $2(L+1)n_y n_u$ number of unique parameters to estimate. The time complexity of SVD on the matrix $\mathcal{H}(H)$ is $\mathcal{O}(L^2 n_y n_x)$ (Xu and Qiao, 2008), so the quadratic dependence on L it causes a computational bottleneck when L is large. Therefore, we propose adopting a BM type re-parameterization (Burer and Monteiro, 2003) in two ways: (i) by learning two low-rank matrices whose product forms a Hankel matrix, and (ii) by directly estimating the system parameters (A, B, C, D). First, recall the adjoint of the Hankel operator $\mathcal{H}^\dagger : \mathbb{R}^{(L+1)n_y \times (L+1)n_u} \to \mathbb{R}^{n_y \times (2L+1)n_u}$ as

$$\mathcal{H}^\dagger\left(\left[[K_{i+j-1}]_{i\in[L+1]}\right]_{j\in[L+1]}\right) = \begin{bmatrix} K_1 & K_2 & \dots & K_{2L+1}\end{bmatrix} \in \mathbb{R}^{n_y \times (2L+1)n_u}, \quad (3)$$

and note that $\mathcal{H} \circ \mathcal{H}^\dagger = \mathsf{id}$, and $\mathcal{H}^\dagger \circ \mathcal{H} = \mathsf{id}$, where id is the identity operator.

First formulation. We define the optimization program (P1) as

$$\min_{n_x,V,Z,D} \frac{1}{2N}\sum_{i=1}^{N} \|\mathbf{y}^i_{2L+1} - \begin{bmatrix} \mathcal{H}^\dagger(VZ^T) & D\end{bmatrix} \mathsf{vec}(U_i)\|_F^2 + \frac{\lambda}{2}\left[\|V\|_F^2 + \|Z\|_F^2\right], \quad \text{(P1)}$$

where $V \in \mathbb{R}^{(L+1)n_y \times n_x}$, and $Z \in \mathbb{R}^{(L+1)n_u \times n_x}$.

Program (P1) redefines the nuclear norm minimization via the sum of squared Frobenius norms $\frac{1}{2}\left[\|V\|_F^2 + \|Z\|_F^2\right]$. In contrast to the nuclear norm regularization, the Frobenius norm minimization has time complexity $\mathcal{O}(L(n_y + n_u))$, which is only linear in L. However, in both Programs (P1') and (P1), the number of parameters scales with the trajectory length. For (P1'), it has been shown that the sample complexity is $NL \gtrsim \tilde{\mathcal{O}}(L^2 n_u(n_y + n_u))$ (Sun et al., 2022). In contrast, we will see in Theorem 2 (P1) achieves a smaller sample complexity of $NL \gtrsim \tilde{\mathcal{O}}(L n_u(n_y + n_u))$. Nevertheless, there remains room for improvement in the sample complexity, given that we have access to $\mathcal{O}(NL)$ data points. We consider another reformulation that ameliorates this suboptimal trajectory dependence. Unfortunately, the theoretical guarantees necessitate that the systems be real, diagonalizable, which imposes limitations on the class of realizable systems (Fernández-Cara et al., 2015). However, previous studies have shown that estimation of non-diagonalizable systems is highly challenging (Tu et al., 2024) which is an avenue for future work.

Second formulation. For real, diagonalizable systems we optimize over the system parameters and dimension directly via a regularization that can induce low-order structure. We achieve this by considering the program

$$\min_{n_x,\mathbf{a},B,C,D} \frac{1}{4N(L+1)}\sum_{i=1}^{N} \|\mathsf{vec}(Y_i) - G(\mathsf{diag}(\mathbf{a}), B, C, D)\mathsf{vec}(U_i)\|_F^2 + \lambda\Theta(\mathbf{a}, B, C, D), \quad \text{(P2)}$$

where $\mathbf{a} \in \mathbb{R}^{n_x}$, and $\Theta(\mathbf{a}, B, C, D)$ is a regularizer that we will define subsequently in §3. Since we are dealing with real, diagonalizable systems the program (P2) optimizes only on the spectrum

of state-transition matrix A, and eigen matrix is assumed to be absorbed in B, and C. Although Program (P2) is nonconvex, specific choices of $\Theta(\mathbf{a}, B, C, D)$ allow us to provide a certificate of optimality when employing first-order optimization methods. In this work, we provide certificates for global optimality and analyze statistical recovery errors for the programs (P1), and (P2) for a trivial feed-through matrix, $D = 0$. The bulk of the work (in §3) relies in re-writing the programs (P1), and (P2) as sums of slightly generalized positively homogenous functions for which global optimality guarantees are well studied in Haeffele and Vidal (2020). In §4 we instantiate the approach of Ziemann and Tu (2022) to provide tight (up-to log factors) error rates and sample complexities.

3. Optimality certificates of nonconvex programs

In this section, we present certificates for the global optimality to each of the formulations (P1), and (P2), whose proofs can be found in §A. First, we state Proposition 1 that establishes optimality of any stationary points of (P1).

Proposition 1 *Let (U_i, Y_i) be N roll-outs of the system (LS). Consider the estimator via factors $\hat{V} \in \mathbb{R}^{(L+1)n_y \times n_x}$, $\hat{Z} \in \mathbb{R}^{(L+1)n_u \times n_x}$. Define $U_i' := [U_i]_{1:2L+1}$, and*

$$M := \frac{1}{N\lambda} \sum_{i=1}^{N} \left(\mathbf{y}_{2L+2}^{(i)} - \mathcal{H}^{\dagger}(\hat{V}\hat{Z}^T)\mathsf{vec}(U_i') \right) \mathsf{vec}(U_i')^T; \; \mathsf{Polar}^{(\mathrm{P1})} := \|\mathcal{H}(M)\|_2. \tag{4}$$

Suppose $\hat{V}, \hat{Z}$ are any stationary points of Program (P1). If $\mathsf{Polar}^{(\mathrm{P1})} = 1$, then $\hat{V}$ and $\hat{Z}$ are globally optimal. Otherwise, the objective can be reduced by augmenting the top-singular vectors $(\mathbf{v}^*, \mathbf{z}^*)$ of $\mathcal{H}(M)$ to $\begin{bmatrix} \hat{V} & \tau^* \mathbf{v}^* \end{bmatrix} \in \mathbb{R}^{(L+1)n_y \times (n_x+1)}$ and $\begin{bmatrix} \hat{Z} & \tau^* \mathbf{z}^* \end{bmatrix} \in \mathbb{R}^{(L+1)n_u \times (n_x+1)}$, for some scaling factor τ^*.

Corollary 1 *Under the conditions of Proposition 1, if $(\hat{n}_x, \hat{V}, \hat{Z})$ are the globally optimal points of program (P1), then the system has the order $\hat{n}_x$ and system parameters take the form*

$$\hat{A} = \left(\left[\hat{V} \right]_{1:n_y,:} \right)^{\dagger} \left[\hat{V} \right]_{1+n_y:2n_y,:}, \hat{B} = \left[\hat{Z}^T \right]_{:,1:n_u}, \hat{C} = \left[\hat{V} \right]_{1:n_y,:}. \tag{5}$$

Remarks: From Proposition 1 we can utilize any first-order algorithms such as gradient descent, Polyak's momentum method (Polyak, 1964), or Nesterov accelerated method (Nesterov, 1983) to approximately reach stationary points in polynomial time, we assume exact convergence for technical convenience. Then from the Equation 4 we need to verify the condition $\mathsf{Polar}^{(\mathrm{P1})} = 1$. For which we need to compute the Hankel norm that is computationally expensive requiring $\mathcal{O}(Ln_yn_u)$ iterations (Cariow and Gliszczyński, 2012). Moreover, $\mathsf{Polar}^{(\mathrm{P1})}$ can never be strictly less than 1, see Haeffele and Vidal (2017) for further discussion. However, for program (P2) we will see that such optimality check is faster. To estimate the system parameters we can perform a pseudo inverse on the learned factors, V, Z as presented in Corollary 1. By the variational nuclear norm re-paramterization, we have eliminated the need for separate Ho-Kalman procedure that was needed in many Sub-space recovery algorithms like N4SID (Van Overschee and De Moor, 1994).

Before we state our next results we define $\gamma(a) := \sum_{t=0}^{L} a^{2t}$, $P(a) \in \mathbb{R}^{2(L+1)\times 2(L+1)}$ where

$$[P(a)]_{i,j} = \begin{cases} a^{i-j-1} \text{ if } i > j \\ 0 \text{ otherwise} \end{cases} \in \mathbb{R}, \text{ and } \Theta(\mathbf{a}, B, C, D) := \sum_{j=1}^{n_x} \gamma(a_j) \|\mathbf{b}_j\|_2 \|\mathbf{c}_j\|_2. \tag{6}$$

Regularization $\Theta(\mathbf{a}, B, C, D)$ resembles the atomic norm type norm considered in matrix factorization problems (Bach, 2013; Haeffele and Vidal, 2017) whose optimality guarantees are well-studied. We next present optimality guarantees for applying this atomic-norm-type regularization directly to system parameters in the context of low-order SysID.

Theorem 1 *Let (U_i, Y_i) be N roll-outs from the system (LS) with $\mathbf{x}_0 = 0$. Consider the estimator of system parameters via $\left(\mathit{diag}\,(\mathbf{a}),\ \left[\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{n_x}\right]^T,\ \left[\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_{n_x}\right]\right)$, of order n_x where $\mathbf{a} \in \mathbb{R}^{n_x}$, $\mathbf{b}_\cdot \in \mathbb{R}^{n_u}$, $\mathbf{c}_\cdot \in \mathbb{R}^{n_y}$. Define $U_i' := [U_i]_{1:2L+1}$, and*

$$M(a) := \frac{1}{2N(L+1)\lambda} \sum_{i=1}^{N} \left[Y_i - \sum_{j=1}^{n_x} \mathbf{c}_j \mathbf{b}_j^T U_i' P^T(a_j) \right] \frac{P(a)}{\gamma(a)} U_i'^T,\ \mathit{Polar}^{(\mathrm{P2})} := \sup_{a \in \mathbb{R}} \|M(a)\|_2. \quad (7)$$

Let $\{a_j, \mathbf{c}_j, \mathbf{b}_j\}_{j=1}^{n_x}$ be any stationary points of program (P2). If $\mathit{Polar}^{(\mathrm{P2})} = 1$, then $\{a_j, \mathbf{c}_j, \mathbf{b}_j\}$ are global optimal points. Otherwise, we can reduce the objective by the parameters $\{a_j, \mathbf{c}_j, \mathbf{b}_j\}_{j=1}^{n_x} \cup \{\tau_1 a^*, \tau \mathbf{c}^*, \tau \mathbf{b}^*\}$ for some scalings, τ_1, τ, where a^* is the supremizer of κ and $\mathbf{c}^*, \mathbf{b}^*$ are the top-singular vectors the matrix $M(\mathbf{a}^*)$.

Remarks: Similar to the discussion in Proposition 1, we can use any first-order optimization methods to approximately reach stationary points in polynomial time. Once we reach stationary points, this leaves us to check the condition $\mathsf{Polar}^{(\mathrm{P2})} = 1$, this requires us to maximize the rational matrix polynomial, this can be done through line search methods in one-dimension such as golden-section or Fibonacci search in polynomial time (Ben-Tal and Nemirovski, 2001).

Proof Strategy: Our proofs for Proposition 1, and Theorem 1 rely using the general framework for global optimality by Haeffele and Vidal (2017), who studied global optimality guarantees for the objective of the form,

$$\min_{r, \{W_j\}_{j=1}^r} \ell\left(Y, \sum_{j=1}^{r} \phi(W_j; X)\right) + \lambda \sum_{j=1}^{r} \theta(W_j).$$

where $\phi(\cdot)$, and $\theta(\cdot)$ are positively homogeneous with the same degree.

For program (P1), (P2) we can re-write the predictions and regularization as

$$\Phi_{n_x}{}^{(\mathrm{P1})}(V, Z; U) := \sum_{j=1}^{n_x} \mathcal{H}^\dagger(\mathbf{v}_j \mathbf{z}_j^T)\mathsf{vec}(U'); \Theta_{n_x}{}^{(\mathrm{P1})}(V, Z) := \frac{1}{2} \sum_{j=1}^{n_x} \|\mathbf{v}_j\|_2^2 + \|\mathbf{z}_j\|_2^2, \quad (8)$$

$$\Phi_{n_x}{}^{(\mathrm{P2})}(\mathbf{a}, B, C; U) := \sum_{j=1}^{n_x} \left[P(a_j) \otimes \mathbf{c}_j \mathbf{b}_j^T\right] \mathsf{vec}(U'); \Theta_{n_x}{}^{(\mathrm{P2})}(\mathbf{a}, B, C) := \sum_{j=1}^{n_x} \gamma(a_j) \|\mathbf{b}_j\|_2 \|\mathbf{c}_j\|_2. \quad (9)$$

The pair $\left(\Phi_{n_x}{}^{(\mathrm{P1})}(\cdot), \Theta_{n_x}{}^{(\mathrm{P1})}(\cdot)\right)$ satisfy homogeneity property which enables use the results of Haeffele and Vidal (2017). However, the pair $\left(\Phi_{n_x}{}^{(\mathrm{P2})}(\mathbf{a}, B, C; U), \Theta_{n_x}{}^{(\mathrm{P2})}(\mathbf{a}, B, C)\right)$ are positively homogeneous with respective to only B and C but not $\mathbf{a}$. Nevertheless, this technical challenge was resolved in Tadipatri et al. (2025) by relaxing to a weaker positive homogeneity property.

4. Statistical error rates and Sample complexities for nonconvex programs

In this section, we present statistical error rates and sample complexities for each of the formulations (P1), and (P2). The high-level proof strategy is same for Theorem 2, and 3 which is discussed near the end of this section for full proofs see §B. To begin, we state few assumptions that are required for statistical recovery of the system parameters. We impose tail conditions on the inputs and noise.

Assumption 1 (Data Model) *The control inputs, $\mathbf{u}_t \in \mathbb{R}^{n_u}$ are drawn independently from sub-Gaussian distribution with a proxy variance σ_U^2/n_u and zero mean, i.e., for any unit vector $\mathbf{v} \in \mathbb{R}^{n_u}$ and $\forall \lambda \geq 0$ $\mathbb{E}\left[\exp(\lambda\langle \mathbf{u}_t, \mathbf{v}\rangle)\right] \leq \exp\left(\lambda^2\sigma_U^2/2n_u\right)$. Let the covariance of $\mathbf{u}_t$ be $\Sigma_U \succeq 0$, and define block matrix $\tilde{\Sigma}_U = I_{2(L+1)} \otimes \Sigma_U$. The outputs are governed by the system (LS) with order n_x^*, and $\zeta_t|\mathbf{u}_{1:t}$ is conditionally independent sub-Gaussian distribution with a proxy variance σ^2/n_y. Denote the impulse response for this linear system as G^*.*

Our next assumption ensures compactness of the parameter class.

Assumption 2 (Bounded parameters for (P1)) *The learned parameters lie in the parametric class defined by*

$$\mathcal{F}_\theta^{(\mathrm{P1})} := \{(V, Z) : n_x \in \mathbb{N}, \|\mathbf{v}_j\|_2 \leq B_v, \|\mathbf{z}_j\|_2 \leq B_z\}. \tag{10}$$

Before we state Theorem 2 we define optimal regularizer

$$\Omega^{(\mathrm{P1})}(\hat{G}) := \inf_{n_x, V, Z \in \mathcal{F}_\theta^{(\mathrm{P1})}} \frac{1}{2}\left[\|V\|_F^2 + \|Z\|_F^2\right], \text{ s.t. } \mathcal{H}^\dagger(VZ^T) = \hat{G}, \tag{11}$$

and let $E_t \in \mathbb{R}^{n_y \times 2(L+1)n_y}$ such that $[E_t]_{1+(t-1)n_y:tn_y} = I_{n_y}$, and rest of them to be zero. Now define condition number as

$$\mathsf{cond}_{\mathcal{F}_\theta^P} := \sup_{\theta \in \mathcal{F}_\theta^P} \sup_{t \in [2(L+1)]} \frac{\sqrt{\mathbb{E}\left[\|E_t \Phi_{n_x}^P(\theta; U)\|_F^4\right]}}{\mathbb{E}\left[\|E_t \Phi_{n_x}^P(\theta; U)\|_F^2\right]}; P \in \{(\mathrm{P1}), (\mathrm{P2})\}. \tag{12}$$

Note that $\Phi_{n_x}^P(\theta; U)$ (see Equations (8), and (9)) is linear in U. Therefore, if U follows a Gaussian distribution, the condition number $\mathsf{cond}_{\mathcal{F}_\theta^P}$ evaluates to 3. For sub-Gaussian distributions, this can be bounded using Proposition 6.1 from Ziemann et al. (2023).

Theorem 2 *Let (U_i, Y_i) be N i.i.d roll-outs following the Assumptions 1. Fix a $\delta \in (0, 1]$. Suppose the regularization parameter is such that $\lambda \leq \tilde{\mathcal{O}}\left(\frac{n_x(n_y+n_u)}{N} + \frac{\ln(1/\delta)}{NL}\right)$. For any global optimal points $(n_x, \hat{V}, \hat{Z})$ of program (P1) satisfying Assumption 2. If $N/\ln(NL) \gtrsim \mathsf{cond}^2_{\mathcal{F}_\theta^{(\mathrm{P1})}} \times [Ln_x(n_y + n_u) + \ln(1/\delta)]$, then w.p at-least $1 - \delta$ we have that*

$$\|(\mathcal{H}^\dagger(\hat{V}\hat{Z}^T) - G^*)\tilde{\Sigma}_U^{1/2}\|_F^2 \leq \tilde{\mathcal{O}}\left(\left(\sigma^2 + \Omega^{(\mathrm{P1})}(G^*)\right)\left[\frac{n_x(n_y+n_u)}{N} + \frac{\ln(1/\delta)}{NL}\right]\right). \tag{13}$$

Corollary 2 *Under the conditions of Theorem 2, the sample complexity for the recovery of G^* through program (P2) is $NL \gtrsim (Ln_x(n_y + n_u) + \ln(1/\delta)) \cdot \mathsf{polylog}(NL)$.*

Remarks. For a fixed failure rate in Equation (13) we recover near tight statistical error rates, naïvely we have $\mathsf{recovery\ error} \lesssim \sqrt{\#\ \mathsf{parameters}/\#\ \mathsf{samples}}$. This is optimal up-to a logarithmic factor in comparison to least-squares error. From Corollary 2 we infer that the sample complexity grows nearly linearly with the trajectory length, i.e., $NL \geq \tilde{\mathcal{O}}(Ln_x(n_y+n_u))$. This recovers the program (P1') studied in (Sun et al., 2022). In practice, dependency on the trajectory length is undesirable because each trajectory in itself provides data points albeit dependent ones.

However, this limitation does not apply to formulation (P2). We next present Theorem 3 which provide error rates and sample complexities for program (P2), effectively addressing the aforementioned issue. To set the stage, we begin by introducing a compactness assumption on the system parameters, similar to Assumption 2.

Assumption 3 (Bounded parameters for (P2)) *The learned parameters lie in the parametric class defined by*

$$\mathcal{F}_\theta^{(\mathrm{P2})} := \{(\mathbf{a}, B, C) : n_x \in \mathbb{N}, |a_j| \leq B_a, \|\mathbf{b}_j\|_2 \leq B_b, \|\mathbf{c}_j\|_2 \leq B_c\}\,. \tag{14}$$

Furthermore for all $(\mathbf{a}', B', C'), (\mathbf{a}, B, C) \in \mathcal{F}_\theta^{(\mathrm{P2})}$ *and some constant* k_a *dependent only on* B_a *it holds true that*

$$\|G(\mathsf{diag}(\mathbf{a}'), B', C') - G(\mathsf{diag}(\mathbf{a}), B, C)\|_2 \leq k_a \left[\|B' - B\|_F + \|C' - C\|_F\right]. \tag{15}$$

Next define optimal regularizer similar to Equation (11)

$$\Omega^{(\mathrm{P2})}(\hat{G}) := \inf_{n_x, \mathbf{a}, B, C \in \mathcal{F}_\theta^{(\mathrm{P2})}} \sum_{j=1}^{n_x} \gamma(a_j)\|\mathbf{b}_j\|_2\|\mathbf{c}_j\|_2, \text{ s.t. } G(\mathsf{diag}(\mathbf{a}), B, C) = \hat{G}. \tag{16}$$

Now we present the statistical recovery guarantee of program (P2).

Theorem 3 *Let* (U_i, Y_i) *be* N *i.i.d roll-outs following the Assumptions 1. Fix a* $\delta \in (0, 1]$. *Suppose the regularization parameter is such that* $\lambda \leq \tilde{\mathcal{O}}\left(\frac{n_x(n_y+n_u)+\ln(1/\delta)}{NL}\right)$. *For any global optimal points* $(n_x, \{\hat{a}_j\}, \hat{B}, \hat{C})$ *of program (P2) satisfies Assumption 3. If* $N/\ln(NL) \gtrsim \mathsf{cond}^2_{\mathcal{F}_\theta^{(\mathrm{P2})}} \times [n_x(n_y+n_u)+\ln(1/\delta)]$, *then w.p at-least* $1-\delta$ *we have that*

$$\|(G(\mathsf{diag}(\{\hat{a}_j\}), \hat{B}, \hat{C}) - G^*)\tilde{\Sigma}_U^{1/2}\|_F^2 \leq \tilde{\mathcal{O}}\left(\left(\sigma^2 + \Omega^{(\mathrm{P2})}(G^*)\right)\left[\frac{n_x(n_y+n_u)+\ln(1/\delta)}{NL}\right]\right). \tag{17}$$

Corollary 3 *Under the conditions of Theorem 3, the following statements holds true,*

1. *If* A^* *is not real, diagonalizable then the upper bound of Equation 17 evaluates to* ∞.
2. *Otherwise, the sample complexity for the recovery of* G^* *with program (P2) is* $NL \gtrsim [n_x(n_y+n_u)+\ln(1/\delta)] \cdot \mathsf{polylog}(NL)$.

Remarks: Statement 1 of Corollary 3 establishes that if the underlying system is not real, diagonalizable then the statistical error obtained in Theorem 3 becomes trivial. Suppose that underlying system was real, diagonalizable then we would require a total samples, $NL \gtrsim \tilde{\mathcal{O}}\left(n_x(n_y+n_x)\right)$. By naïve counting argument we have $n_x(n_y+n_u)$ parameters to estimate, therefore, we require atleast those many samples to estimate all the parameters. This fact is indeed reflected in Statement 2. In comparison to Program (P1) the statistical error rate and sample complexity are L-folds tighter.

Comparison with existing works. Table 1 summarizes the comparison between our bounds and those from existing key works. We observe that program (P1) achieves statistical error rates and sample complexity comparable to those of Lee (2022). However, unlike our approach, the method in Lee (2022) assumes knowledge of the true system order, which we do not require. For real, diagonalizable systems, program (P2) additionally removes the dependency on trajectory length present in program (P1) and prior methods. In terms of both statistical error rate and sample complexity, we observe at least an L-fold improvement.

Methods for low-order SysID	**(Error rate)^2**	**Sample complexity**, $N_{\mathsf{tol}} \gtrsim$
Hankel nuclear norm minimization (Sun et al., 2022)	$L^2 n_x(n_y+n_u)/N_{\mathsf{tol}}$	$L^2 n_x(n_y+n_u)$
SVD truncation (Lee, 2022) (n_x^* *is required*)	$Ln_x(n_y+n_u)/N_{\mathsf{tol}}$	$Ln_x(n_y+n_u)$
BM re-parametrization (P1) (Theorem 2)	$Ln_x(n_y+n_u)/N_{\mathsf{tol}}$	$Ln_x(n_y+n_u)$
System Parameters (P2) (Theorem 3)‡	$n_x(n_y+n_u)/N_{\mathsf{tol}}$	$n_x(n_y+n_u)$

Table 1: Summary of error rates and sample complexities for low-order SysID (up to log factors). Here, $N_{\mathbf{tol}} = N \cdot L$, and ‡ indicates restriction to real, diagonalizable systems.

Proof Strategy. Although Programs (P1) and (P2) are nonconvex in the parameter space, the input-output map remains linear. Thus, we can directly apply the time-dependent excess risk bounds from Theorem 6.1 of Ziemann and Tu (2022), under the additional condition that the impulse response is Lipschitz continuous with respect to the parameters and the parameter space is compact. These conditions are ensured by Assumptions 2 and 3.

5. Numerical Simulations

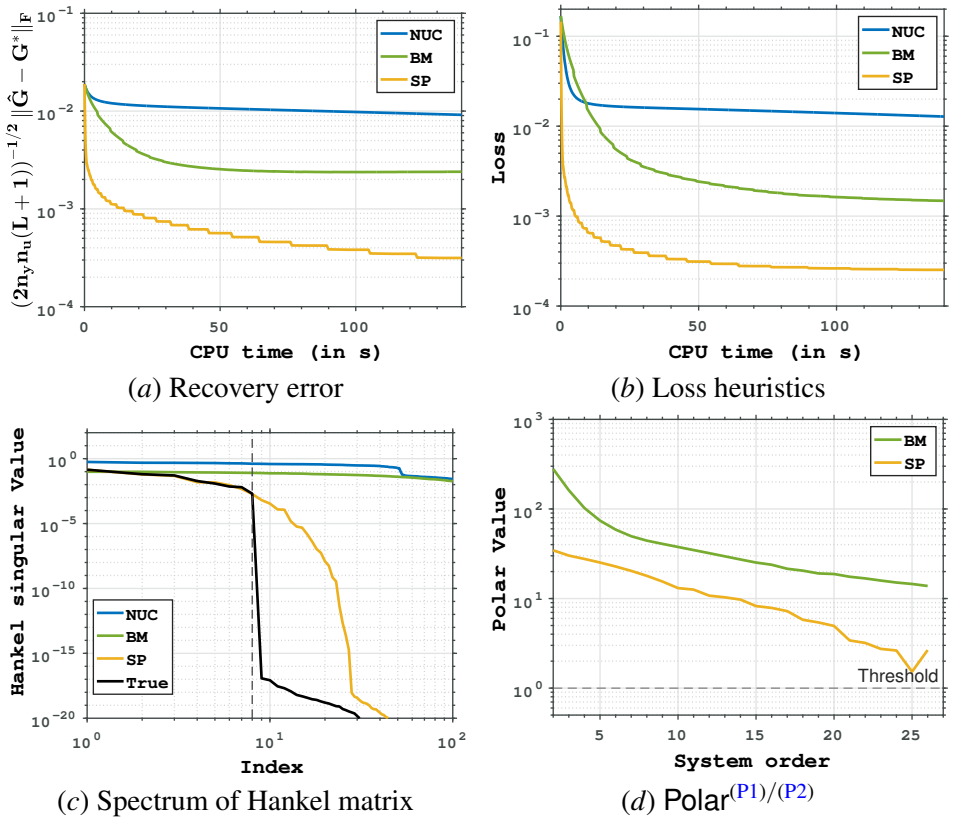
In this section, we compare low-order SysID using programs (P1'), (P1), and (P2) through numerical simulations, evaluating time complexity, sample complexity, and trajectory complexity.

Data Generation. We simulate linear trajectories with system order $n_x^* = 5$ and dimensions $n_u = n_y = 8$. The system matrix A is symmetric with Normal entries, while B and C are Normal matrices. Control inputs are sampled from $\mathcal{N}(0, I_{n_u}/n_u)$, and outputs are corrupted with Gaussian noise of variance 0.01. We generate $N = 500$ rollouts, and trajectory length $2(L+1) = 102$.

Algorithmic Implementation. We use accelerated proximal gradient descent (D.1)(Becker et al., 2011) to solve convex program (P1'). For programs (P1), and (P2) we unroll Proposition 1, and Theorem 1 into algorithms (D.2), (D.3) that uses Polyak's gradient descent (Polyak, 1964). For each of the algorithm we fixed regularization parameter, $\lambda = 0.001$ and choose the best learning rate, and momentum rate. For fair comparison, recovery error, $\|\hat{G} - G^*\|_F/\sqrt{2n_y n_u(L+1)}$ was evaluated against CPU runtime rather than iterations of the algorithms. All algorithms are initialized such that they share identical impulse response to ensure consistent starting points.

Ablation Studies. We evaluate recovery loss across varying sample sizes, trajectory lengths, and algorithm performance on real- and non-real-diagonalizable systems. From the results in Figures 1, D.1, and D.2, we conclude the following:

- *Computational Efficiency.* As shown in Figure 1(*a*), program (P2) outperforms program (P1) and program (P1') in recovering the impulse response within a fixed CPU time budget. Figure 1(*b*)



(*a*) Recovery error (*b*) Loss heuristics

(*c*) Spectrum of Hankel matrix (*d*) Polar$^{(P1)/(P2)}$

Figure 1: Performance metrics for a real, diagnosable system. NUC, BM, and SP corresponds to programs (P1'), (P1), and (P2), respectively. Dashed line in 1(*c*) represents $n_x = 5$.

highlights faster loss reduction for program (P2). Noise often necessitates higher-order approximations for program (P1) and (D.3) to achieve global optima, as shown in Figure 1(*c*) and 1(*d*).

- *Statistical Efficiency.* For a fixed sample size or trajectory length (Figure D.1), program (P2) consistently outperforms the others as discussed earlier in the Table 1.

6. Conclusions

In this work, we focus on performing system identification with minimal state representation. We propose two nonconvex reformulations: (i) a BM-style reparameterization of Hankel nuclear norm minimization, and (ii) direct optimization over the system parameters. Despite their nonconvex nature, we design algorithms with theoretical guarantees of global convergence using first-order optimization methods. Additionally, we derive near tight statistical error rates and sample complexity bounds. Our findings suggest that directly estimating system parameters outperforms alternative approaches in both optimization efficiency and statistical recovery. However, our analysis is currently restricted to systems that are real, diagonalizable. Extending this framework to handle non-diagonalizable systems is an avenue for future work.

Acknowledgments

UKRT gratefully acknowledges Hancheng Min for his valuable feedback.

References

Karl Johan Aström and Peter Eykhoff. System identification—a survey. *Automatica*, 7(2):123–162, 1971.

Francis Bach. Convex relaxations of structured matrix factorizations, September 2013. arXiv:1309.3117 [cs, math].

Stephen R. Becker, Emmanuel J. Candès, and Michael C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3): 165–218, September 2011. ISSN 1867-2957. doi: 10.1007/s12532-011-0029-5.

Aharon Ben-Tal and Arkadi Nemirovski. Lectures on modern convex optimization: Analysis. *Algorithms, and Engineering Applications*, 2, 2001.

Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical programming*, 95(2):329–357, 2003.

Marco C Campi, Simone Garatti, and Federico A Ramponi. Non-convex scenario optimization with application to system identification. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 4023–4028. IEEE, 2015.

Aleksandr Cariow and Marek Gliszczyński. Fast algorithms to compute matrix-vector products for toeplitz and hankel matrices. *Electrical Review*, 88(8):166–171, 2012.

Alessandro Chiuso, Tianshi Chen, Lennart Ljung, and Gianluigi Pillonetto. Regularization strategies for nonparametric system identification. In *52nd IEEE conference on decision and control*, pages 6013–6018. IEEE, 2013.

Bart De Schutter. Minimal state-space realization in linear system theory: an overview. *Journal of computational and applied mathematics*, 121(1-2):331–354, 2000.

M. Fazel, H. Hindi, and S.P. Boyd. Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices. In *Proceedings of the 2003 American Control Conference, 2003.*, volume 3, pages 2156–2162 vol.3, 2003. doi: 10.1109/ACC.2003.1243393.

Maryam Fazel, Haitham Hindi, and Stephen P Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference.(Cat. No. 01CH37148)*, volume 6, pages 4734–4739. IEEE, 2001.

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, January 2013. ISSN 0895-4798, 1095-7162. doi: 10.1137/110853996.

Enrique Fernández-Cara, Manuel González-Burgos, and Luz de Teresa. Controllability of linear and semilinear non-diagonalizable parabolic systems. *ESAIM: Control, Optimisation and Calculus of Variations*, 21(4):1178–1204, 2015.

Paris Giampouras, Rene Vidal, Athanasios Rontogiannis, and Benjamin Haeffele. A novel variational form of the Schatten-p quasi-norm. In *Advances in Neural Information Processing Systems*, volume 33, pages 21453–21463. Curran Associates, Inc., 2020.

Benjamin D. Haeffele and Rene Vidal. Global optimality in neural network training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Benjamin D. Haeffele and Rene Vidal. Structured Low-Rank Matrix Factorization: Global Optimality, Algorithms, and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(6):1468–1482, June 2020. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2019.2900306.

B.L. Ho and R.E. Kalman. Die konstruktion von linearen modeilen in der darstellung durch zustandsvariable aus den beziehungen für ein- und ausgangsgrößen. *at - Automatisierungstechnik*, 14(1-12):545–548, 1966. doi: doi:10.1524/auto.1966.14.112.545.

Arya Honarpisheh, Rajiv Singh, Jared Miller, and Mario Sznaier. Identification of low order systems in a loewner framework. *IFAC-PapersOnLine*, 58(15):199–204, 2024.

Holden Lee. Improved rates for prediction and identification of partially observed linear dynamical systems. In *International Conference on Algorithmic Learning Theory*, pages 668–698. PMLR, 2022.

Ljung Lennart and Peter E Caines. Asymptotic normality of prediction error estimators for approximate system models. *Stochastics*, 3(1-4):29–46, 1980.

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

Ivan Markovsky. Application of low-rank approximation for nonlinear system identification. In *2017 25th Mediterranean Conference on Control and Automation (MED)*, pages 12–16. IEEE, 2017.

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

Karthik Mohan and Maryam Fazel. Reweighted nuclear norm minimization with application to system identification. In *Proceedings of the 2010 American Control Conference*, pages 2953–2959. IEEE, 2010.

Yurii Nesterov. A method for solving the convex programming problem with convergence rate o (1/k2). In *Dokl akad nauk Sssr*, volume 269, page 543, 1983.

Samet Oymak and Necmiye Ozay. Non-asymptotic identification of lti systems from a single trajectory. In *2019 American control conference (ACC)*, pages 5655–5661. IEEE, 2019.

Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014.

Gianluigi Pillonetto, Tianshi Chen, Alessandro Chiuso, Giuseppe De Nicolao, and Lennart Ljung. Regularized linear system identification using atomic, nuclear and kernel-based norms: The role of the stability constraint. *Automatica*, 69:137–149, 2016.

Gianluigi Pillonetto, Aleksandr Aravkin, Daniel Gedon, Lennart Ljung, Antônio H Ribeiro, and Thomas B Schön. Deep networks for system identification: a survey. *Automatica*, 171:111907, 2025.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

S Joe Qin. An overview of subspace identification. *Computers & chemical engineering*, 30(10-12): 1502–1513, 2006.

Benjamin Recht, Weiyu Xu, and Babak Hassibi. Necessary and sufficient conditions for success of the nuclear norm heuristic for rank minimization. In *2008 47th IEEE Conference on Decision and Control*, pages 3065–3070, 2008. doi: 10.1109/CDC.2008.4739332.

Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. *SIAM Review*, 52(3):471–501, January 2010. ISSN 0036-1445, 1095-7200. doi: 10.1137/070697835.

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

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. PMLR, 2018.

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.

Eduardo D Sontag. *Mathematical control theory: deterministic finite dimensional systems*, volume 6. Springer Science & Business Media, 2013.

Yue Sun, Samet Oymak, and Maryam Fazel. Finite sample identification of low-order lti systems via nuclear norm regularization. *IEEE Open Journal of Control Systems*, 1:237–254, 2022. doi: 10.1109/OJCSYS.2022.3200015.

Uday Kiran Reddy Tadipatri, Benjamin David Haeffele, Joshua Agterberg, and Rene Vidal. A convex relaxation approach to generalization analysis for parallel positively homogeneous networks. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Anastasios Tsiamis and George J Pappas. Finite sample analysis of stochastic system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3648–3654. IEEE, 2019.

Stephen Tu, Roy Frostig, and Mahdi Soltanolkotabi. Learning from many trajectories. *Journal of Machine Learning Research*, 25(216):1–109, 2024.

Peter Van Overschee and Bart De Moor. N4sid: Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica*, 30(1):75–93, 1994.

Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, 1996.

Wei Xu and Sanzheng Qiao. A fast symmetric svd algorithm for square hankel matrices. *Linear Algebra and its Applications*, 428(2-3):550–563, 2008.

L. A. Zadeh. From circuit theory to system theory. *Proceedings of the IRE*, 50(5):856–865, 1962. doi: 10.1109/JRPROC.1962.288302.

Ingvar Ziemann and Stephen Tu. Learning with little mixing. *Advances in Neural Information Processing Systems*, 35:4626–4637, 2022.

Ingvar Ziemann, Anastasios Tsiamis, Bruce Lee, Yassir Jedra, Nikolai Matni, and George J. Pappas. A tutorial on the non-asymptotic theory of system identification. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 8921–8939, 2023. doi: 10.1109/CDC49753.2023.10383489.

Ingvar Ziemann, Stephen Tu, George J. Pappas, and Nikolai Matni. Sharp rates in dependent learning theory: Avoiding sample size deflation for the square loss. In *Forty-first International Conference on Machine Learning*, 2024.