Data-Driven Sparse System Identification

Salar Fattahi and Somayeh Sojoudi

***Abstract*—In this paper, we study the system identification problem for sparse linear time-invariant systems. We propose a sparsity promoting Lasso-type estimator to identify the dynamics of the system with only a limited number of input-state data samples. Using contemporary results on high-dimensional statistics, we prove that $\Omega(k_{\max}\log(m+n))$ data samples are enough to reliably estimate the system dynamics, where n and m are the number of states and inputs, respectively, and $k_{\max}$ is the maximum number of nonzero elements in the rows of input and state matrices. The number of samples in the developed estimator is significantly smaller than the dimension of the problem for sparse systems, and yet it offers a small estimation error entry-wise. Furthermore, we show that, unlike the recently celebrated least-squares estimators for system identification problems, the method developed in this work is capable of *exact recovery* of the underlying sparsity structure of the system with the aforementioned number of data samples. Extensive case studies on synthetically generated systems and physical mass-spring networks are offered to demonstrate the effectiveness of the proposed method.**

I. Introduction

With their ever-growing size and complexity, real-world dynamical systems are hard to model. Today's systems are complex and large, often with a massive number of unknown parameters which render them doomed to the so-called *curse of dimensionality*. Therefore, system operators should rely on simple and tractable estimation methods to identify the dynamics of the system via a limited number of recorded input-output interactions, and then design control policies to ensure the desired behavior of the entire system. The area of *system identification* is created to address this problem [1].

Despite the long history in control theory, most of the results on system identification deal with asymptotic behavior of the proposed estimation methods [1]–[4]. Although these results shed light on the theoretical consistency of these methodologies, they are not applicable to the finite time/sample settings. In many applications, the dynamics of the system should be estimated under the *large dimension-small sample size* regime, where the dimension of the states and inputs of the system is overwhelmingly large compared to the number of available input-output data. Under such circumstances, the classical approaches for checking the asymptotic consistency of estimators face major breakdowns. Simple examples of such failures can be easily found in high-dimensional statistics. For instance, given a number of independent and identically distributed (i.i.d.) samples with Gaussian distribution, none of the eigenvectors and eigenvalues of the sample covariance matrix are consistent estimators of their true counterparts if the sample size and the dimension of variables grow at the same rate [5]. As another example, it is well-known that the least-squares estimators, which are widely used in system identification problems, seize to exist when the sample size is smaller than the dimension of the system [6].

On the other hand, with the unprecedented interest in data-driven control approaches, such as model-free reinforcement learning, robust, and adaptive control [7]–[9], a question arises as to what the minimum number of input-output data samples should be to guarantee a small error in the estimated model. Answering this question has been the subject of many recent studies on the sample complexity of the system identification problem [10]–[14]. Most of these results are tailored to a specific type of dynamics, depend on the stability of the open-loop system, or do not exploit the *a priori* information on the structure of the system.

In this work, the objective is to employ modern results on high-dimensional statistics to reduce the sample complexity of one of the most fundamental problems in control theory, namely the linear time-invariant (LTI) systems with perfect state measurements. This type of dynamical system forms the basis of many classical control problems, such as Linear Quadratic Regulator and Linear Quadratic Gaussian problems. Our results are built upon the fact that, in many practical large-scale systems, the states and inputs exhibit sparse interactions with one another, which in turn translates into a sparse representation of the state space equations of the system. It will be shown that the sparsity of the dynamics enables us to develop an estimator that is guaranteed to achieve infinitesimal estimation error with small number of samples. In particular, we introduce a Lasso-type estimator, i.e. a least-squares estimator augmented by an ℓ_1 regularizer, and show that, with an appropriate scaling of the regularization coefficient, its sample complexity is only $\Omega(k_{\max}\log(n+m))$, where $k_{\max}$ is maximum number of nonzero elements in the columns of input and state (or system) matrices, and n and m are the dimensions of the states and the inputs, respectively. This is a significant improvement over the recently derived sample complexity $\Omega(n+m)$ for the least-squares estimator, in the case where $k_{\max}$ is much smaller than $m+n$, i.e., when the system is sparse.

In this work, we derive upper bounds on the elementwise error of the proposed estimator. In particular, we show that the elementwise error decreases at the rate $O(\sqrt{\log(n+m)/d})$, where d is the number of available sample trajectories. Although this error is not comparable to the operator norm error bound $O(\sqrt{n+m/d})$ for the least-squares estimator introduced in [13], we show tremendous improvements in the

Email: fattahi@berkeley.edu and sojoudi@berkeley.edu.

Salar Fattahi is with the Department of Industrial Engineering and Operations Research, University of California, Berkeley. Somayeh Sojoudi is with the Departments of Electrical Engineering and Computer Sciences and Mechanical Engineering as well as the Tsinghua-Berkeley Shenzhen Institute, University of California, Berkeley. This work was supported by the ONR grant N00014-17-1-2933, DARPA grant D16AP00002, and AFOSR grant FA9550-17-1-0163.

accuracy of the developed estimator through different case studies. Another advantage of the proposed Lasso estimator over its least-squares analog is its *exact recovery* property. More specifically, we show that while the least-squares estimator is unable to identify the sparsity pattern of the input and state matrices for any finite number of samples, the proposed estimator is guaranteed to recover the true sparsity pattern of these matrices with only $\Omega(k_{\max}\log(n+m))$ samples. It is worthwhile to mention that this work generalizes the results in [12], where the authors use a similar Lasso estimator to learn the dynamics of a particular type of systems. However, [12] assumes that the system is autonomous (input-free) and inherently stable, both of which will be relaxed in this work.

Notations: For a matrix M, the symbols $\|M\|_F$, $\|M\|_2$, $\|M\|_1$, and $\|M\|_\infty$ denote its Frobenius, operator, ℓ_1/ℓ_1, and ℓ_∞/ℓ_∞ norms, respectively. Furthermore, $\kappa(M)$ refers to its 2-norm condition number, i.e., the ratio of its maximum and minimum singular values. Given integer sets I and J, the notation M_{IJ} refers to the submatrix of M whose rows and columns are indexed by I and J, respectively. Given the sequences $f_1(n)$ and $f_2(n)$, the notations $f_1(n) = O(f_2(n))$ and $f_1(n) = \Omega(f_2(n))$ imply that there exist $c_1 < \infty$ and $c_2 > 0$ such that $f_1(n) \le c_1 f_2(n)$ and $f_1(n) \ge c_2 f_2(n)$, respectively. Finally, $f_1(n) = o(f_2(n))$ is used to show that $f_1(n)/f_2(n) \to 0$ as $n \to \infty$. A zero-mean Gaussian distribution with covariance Σ is shown as $N(0,\Sigma)$. Given a function $f(x)$, the expression $\arg\min f(x)$ refers to its minimizer. For a set $\mathcal{I}$, the symbol $|\mathcal{I}|$ denotes its cardinality.

II. Problem Formulation

Consider the LTI system

$$x[t+1] = Ax[t] + Bu[t] + w[t] \tag{1a}$$

where t is the time step, $A \in \mathbb{R}^{n\times n}$ is the state matrix, and $B \in \mathbb{R}^{n\times m}$ is the input matrix. Furthermore, $x[t] \in \mathbb{R}^n$, $u[t] \in \mathbb{R}^m$, and $w[t] \in \mathbb{R}^n$ are the state, input, and disturbance vectors at time t, respectively. The dimension of the system is defined as $m+n$. It is assumed that the input disturbance vectors are identically distributed and independent with distribution $N(0,\Sigma_w)$ across different times. In this work, we assume that the matrices A and B are sparse and the goal is to estimate them based on a limited number of *sample trajectories*, i.e. a sequence $\{(x^{(i)}[\tau], u^{(i)}[\tau])\}_{\tau=0}^{T}$ with $i = 1,2,...,d$, where d is the number of available sample trajectories. The i^{th} sample trajectory $\{(x^{(i)}[\tau], u^{(i)}[\tau])\}_{\tau=0}^{T}$ is obtained by running the system from $t=0$ to $t=T$ and collecting the input and state vectors. The sparsity assumption on A and B is practical in many applications because of two reasons:

- In order to model many large-scale real-world systems accurately, one needs to consider an overwhelmingly large number of internal states. However, it is often the case that the interactions between different states and inputs obey a sparse structure, which translates into a sparsity pattern in state and input matrices. An important example of these types of problems is multi-agent systems, where the agents (subsystems) interact with one another via a sparse communication network.
- More generally, it is well-known that there may not be a unique representation of the state-space model of the system. For example, one can entirely change the state and input matrices in (1) via linear/nonlinear transformations to arrive at a different, but equally accurate state-space equation. Furthermore, the dynamics of the system may have a sufficiently accurate sparse approximation. The question of interest is whether it is possible to design a data-driven method in order to estimate the *sparsest* state-space representation of the system? Answering this question is crucial, specially in the context of a *distributed control* problem where the goal is to design a decentralized controller whose structure respects the dynamics of the system [15]–[17].

Given the sample trajectories $\{(x^{(i)}[\tau], u^{(i)}[\tau])\}_{\tau=0}^{T}$ for $i = 1,2,...,d$, one can obtain an estimate of (A,B) by solving the following least-squares optimization problem:

$$\min_{A,B} \sum_{i=1}^{d}\sum_{t=0}^{T-1} \left\| x^{(i)}[t+1] - \left(Ax^{(i)}[t] + Bu^{(i)}[t]\right)\right\|_2^2 \tag{2}$$

In order to describe the behavior of the least-squares estimator, define

$$Y^{(i)} = \begin{bmatrix} x^{(i)}[1]^T \\ \vdots \\ x^{(i)}[T]^T \end{bmatrix}, \quad X^{(i)} = \begin{bmatrix} x^{(i)}[0]^T & u^{(i)}[0]^T \\ \vdots & \vdots \\ x^{(i)}[T-1]^T & u^{(i)}[T-1]^T \end{bmatrix},$$
$$W^{(i)} = \begin{bmatrix} w^{(i)}[0]^T \\ \vdots \\ w^{(i)}[T-1]^T \end{bmatrix}. \tag{3}$$

for every sample trajectory $i = 1,2,...,d$. Furthermore, let Y, X, and W be defined as vertical concatenations of $Y^{(i)}$, $X^{(i)}$, and $W^{(i)}$ for $i = 1,2,...,d$, respectively. Finally, denote $\Theta = [A \;\; B]^T$ as the unknown system parameter and Θ^* as its true value. Based on these definitions, it follows from (1) that

$$Y = X\cdot\Theta + W \tag{4}$$

The system identification problem is then reduced to estimating Θ based on the *observation matrix* Y and the *design matrix* X. Consider the following least-squares estimator:

$$\Theta_{\text{ls}} = \arg\min_{\Theta} \|Y - X\Theta\|_F^2 \tag{5}$$

One can easily verify the equivalence of (2) and (5). The optimal solution of 5 can be written as

$$\Theta_{\text{ls}} = (X^TX)^{-1}X^TY = \Theta^* + (X^TX)^{-1}X^TW \tag{6}$$

Notice that Θ_{ls} is well-defined and unqiue if and only if X^TX is invertible, which necessitates $d \ge n+m$. The estimation error is then defined as

$$E = \Theta_{\text{ls}} - \Theta^* = (X^TX)^{-1}X^TW \tag{7}$$

Thus, one needs to study the behavior of $(X^TX)^{-1}X^TW$ in order to control the estimation error of the least-squares estimator. However, since the state of the system at time t is affected by random input disturbances at times $0,1,...t-1$, X and W are correlated, which renders (7) hard to analyze. In

order to circumvent this issue, [18] simplifies the estimator and considers only the state of the system at time T in $Y^{(i)}$. By ignoring the first $T-1$ rows in $Y^{(i)}$, $X^{(i)}$, and $W^{(i)}$, one can ensure that the random matrix $(X^TX)^{-1}X^T$ is independent of W. Therefore, it is assumed in the sequel that

$$Y = \begin{bmatrix} x^{(1)}[T]^T \\ \vdots \\ x^{(d)}[T]^T \end{bmatrix}, \qquad X = \begin{bmatrix} x^{(1)}[T-1]^T & u^{(1)}[T-1]^T \\ \vdots & \vdots \\ x^{(d)}[T-1]^T & u^{(d)}[T-1]^T \end{bmatrix},$$
$$W = \begin{bmatrix} w^{(1)}[T-1]^T \\ \vdots \\ w^{(d)}[T-1]^T \end{bmatrix} \tag{8}$$

With this simplification, [18] shows that, with input vectors $u^{(i)}[t]$ chosen randomly from $N(0,\Sigma_u)$ for every $t = 1,2,...,T-1$ and $i = 1,2,...,d$, the least-squares estimator requires at least $d = \Omega(m+n)$ sample trajectories to guarantee $\|E\|_2 = \mathcal{O}\left(\sqrt{\frac{m+n}{d}}\right)$ with high probability. In what follows, a regularized estimator will be introduced that exploits the underlying sparsity structure of the system dynamics to significantly reduce the number of sample trajectories for accurate estimation of the parameters.

III. Main Results

Consider the following variant of (5):

$$\Theta_{\text{lasso}} = \arg\min_{\Theta} \frac{1}{2d}\|X - Y\Theta\|_F^2 + \lambda_d\|\Theta\|_1 \tag{9}$$

where the estimator is now regularized by the sparsity-promoting ℓ_1/ℓ_1 penalty term. Under the sparsity assumption on (A,B), we will show that the non-asymptotic statistical properties of Θ_{lasso} significantly outperform those of Θ_{ls}. In particular, the primary objective is to prove that $\|\Theta_{\text{lasso}} - \Theta^*\|_\infty$ decreases at the rate $\mathcal{O}(\sqrt{\log(n+m)/d})$ with high probability with an appropriate scaling of the regularization coefficient and the sparsity assumption on Θ^*. The second goal is to show that for a large class of regularization coefficients, only $\Omega(k_{\max}\log(n+m))$ sample trajectories are needed to guarantee the uniqueness of the solution and a small estimation error of the Lasso estimator. Here, $k_{\max}$ is the maximum number of nonzero elements in the columns of $[A \;\; B]^T$. Comparing this number with the required lower bound $\Omega(n+m)$ on the number of sample trajectories for the least-squares estimator, we conclude that the proposed method needs significantly less number of samples when A and B are sparse. The third objective is to prove that this method is able to find the correct sparsity structure of A and B with high probability. In contrast, it will be shown that the solution of the least-squares estimator is fully dense for any finite number of sample trajectories and, hence, it cannot correctly extract the sparsity structures of A and B. We will showcase the superior performance of the Lasso estimator in both sparsity identification and estimation accuracy in simulations.

Before presenting the main result of this work, note that

$$\begin{aligned} x^{(i)}[T-1] = {} & A^{T-2}Bu^{(i)}[0] + A^{T-3}Bu^{(i)}[1] + \cdots + Bu^{(i)}[T-2] \\ & + A^{T-2}w^{(i)}[0] + A^{T-3}w^{(i)}[1] + \cdots + w^{(i)}[T-2] \end{aligned} \tag{10}$$

where, without loss of generality, the initial state is assumed to be zero for every sample trajectory. The results can be readily extended to the case where the initial state is an unknown random vector with Gaussian distribution. Recalling that $u^{(i)}[t]$ and $w^{(i)}[t]$ are i.i.d samples of $N(0,\Sigma_u I)$ and $N(0,\Sigma_w I)$, respectively, (10) and (8) imply that

$$X_{i,:}^T \sim N\left(0,\tilde{\Sigma}\right) \tag{11}$$

where $X_{i,:}$ is the i^{th} row of X and

$$\tilde{\Sigma} = \begin{bmatrix} F\Sigma_u F^T + G\Sigma_w G^T & 0 \\ 0 & \Sigma_u \end{bmatrix} \tag{12a}$$
$$F = [A^{T-2}B \;\; A^{T-3}B \cdots B] \tag{12b}$$
$$G = [A^{T-2} \;\; A^{T-3} \cdots I] \tag{12c}$$

Define $\mathcal{A}_j$ as the index set of the rows corresponding to the nonzero elements in the j^{th} column of Θ^*. The complement of $\mathcal{A}_j$ is denoted by $\mathcal{A}_j^c$. The following assumption plays a key role in deriving the main result of this paper:

Assumption 1 (Mutual Incoherence Property)**.** *There exists a number $\gamma \in (0,1]$ such that*

$$\max_{1\le j\le n}\left\{\max_{i\in\mathcal{A}_j^c}\left\{\left\|\tilde{\Sigma}_{i\mathcal{A}_j}(\tilde{\Sigma}_{\mathcal{A}_j\mathcal{A}_j})^{-1}\right\|_1\right\}\right\} \le 1-\gamma \tag{13}$$

The mutual incoherence property is a commonly known assumption for the exact recovery of unknown parameters in compressive sensing and classical Lasso regression problems [19]–[22]. This assumption entails that the effect of those submatrices of $\tilde{\Sigma}$ corresponding to zero (unimportant) elements of Θ on the remaining entries of $\tilde{\Sigma}$ should not be large. Roughly speaking, this condition guarantees that the unknown parameters are *recoverable* in the noiseless scenario, i.e. when $W = 0$. If the recovery cannot be guaranteed in the noise-free setting, then there is little hope for the Lasso estimator to recover the true structure of A and B when the system is subject to noise.

Assumption 2. *For every $1 \le j \le n$, k_j has the order $\Omega\left((n+m)^{\epsilon_j}\right)$ for some $\epsilon_j \in (0,1]$.*

Note that this assumption adds a very mild restriction: the growth rate of the number of nonzero elements in each column of Θ^* should be at least a polynomial function of the dimension of the system. As it will be shown later, this assumption is required to guarantee the consistency of the Lasso estimator in the high-dimensional setting. Next, we define some notations to streamline the presentation of the main theorem. $\Lambda_{\min}$ is used to denote the minimum eigenvalue of $\tilde{\Sigma}$. Furthermore, define

$$k_{\max} = \max_{1\le j\le n}|\mathcal{A}_j|,\; q = \max_{1\le j\le n}\left\{\max_{i\in\mathcal{A}_j}\left\|(\tilde{\Sigma}^{-1}_{\mathcal{A}_j\mathcal{A}_j})_{i\mathcal{A}_j}\right\|_1\right\},\; \epsilon_{\min} = \min_{1\le j\le n}\epsilon_j \tag{14}$$

We assume that $\Lambda_{\min}$ and q are uniformly bounded away from 0 and $+\infty$.

Theorem 1. *Given arbitrary constants $c_1, c_2 > 2$ and $c_3 >$*

$\frac{1}{2\epsilon_{\min}}+1$, *suppose that* d *and* λ_d *satisfy the inequalities*

$$\lambda_d \geq \sqrt{\frac{32c_1\sigma_w\sigma_G}{\gamma^2}\cdot\frac{1+\log(n+m)}{d}} \tag{15}$$

$$d \geq \frac{72c_2}{\gamma^2}\cdot\kappa(\tilde{\Sigma})\cdot k_{\max}\cdot(1+\log(n+m)) \tag{16}$$

Then, with probability of at least

$$1-\frac{K_1}{(n+m)^{c_1-2}}-\frac{K_2}{(n+m)^{c_2-2}}-\frac{K_3}{(n+m)^{2\epsilon_{\min}(c_3-1)-1}}\to 1 \tag{17}$$

for some universal constants $K_1, K_2, K_3 > 0$*, the following statements hold:*

- *Statement 1:* Θ_{lasso} *is unique and in addition,*

$$\|\Theta_{lasso}-\Theta^*\|_\infty \leq g \tag{18}$$

where

$$g=\sqrt{\frac{36c_3\sigma_w\log(k_{\max})}{\Lambda_{\min}d}}+\lambda_d\left(\frac{8k_{\max}}{\Lambda_{\min}\sqrt{d}}+q\right) \tag{19}$$

- *Statement 2: The support of* Θ_{lasso} *is a subset of the support of* Θ^**. Furthermore, the supports of* Θ_{lasso} *and* Θ^* *are the same if* $\theta_{\min} > g$*, where* $\theta_{\min} = \min_{1\leq j\leq n}\{\min_{t\in\mathcal{A}_j}|\Theta_{tj}|\}$.

Proof. See Appendix. □

Theorem 1 states that the support of Θ_{lasso} will not contain any *false positives* with a relatively small number of sample trajectories. Furthermore, it shows that if the decrease rate of the smallest nonzero element in Θ is slow enough, Θ_{lasso} will also exclude *false negatives*.

Another important observation can be made from Theorem 1. Notice that the minimum number of required sample trajectories depends on $\kappa(\tilde{\Sigma})$. Upon assuming that Σ_u and Σ_w are identity matrices, $\kappa(\tilde{\Sigma})$ is reduced to $\kappa(FF^T+GG^T)$. F and G are commonly known as finite time *controllability matrices* for the input and disturbance noise, respectively. Roughly speaking, $\kappa(FF^T+GG^T)$ quantifies the ratio between the eigenvalues of the easiest- and hardest-to-identify modes of the system. Therefore, Theorem 1 implies that only a small number of samples is required to accurately identify the dynamics of the system if all of its modes are easily excitable. The dependency of the estimation error on the modes of the system is also reflected in the non-asymptotic error bound of the least-squares estimator in [18]. This is completely in line with the conventional results on the identifiability of dynamical systems: independent of the proposed method, it is significantly harder to identify the parameters of the system accurately if it possesses nearly-hidden modes. The connection between the identifiability of the system and the number of required sample trajectories to guarantee a small estimation error will be elaborated through different case studies in Section IV.

As mentioned before, Theorem 1 implies that, with an appropriate scaling of the regularization coefficient, $d = \Omega(k_{\max}\log(n+m))$ suffices to have a unique solution for the Lasso estimation problem and to guarantee its small elementwise error with high probability. The following corollary shows that, with a factor of increase in the number of sample trajectories, the error rate of Lasso estimator becomes significantly smaller in the sparse setting.

Corollary 1. *Suppose that* $k_{\max} = o(d^{1/2})$*. Then,* $d = \Omega(k_{\max}^2)$ *is enough to obtain*

$$\|\Theta_{lasso}-\Theta^*\|_\infty = O\left(\sqrt{\frac{\log(n+m)}{d}}\right) \tag{20}$$

with probability of at least (17)*.*

Proof. According to (15), we have

$$\lambda_d = O\left(\sqrt{\frac{\log(n+m)}{d}}\right) \tag{21}$$

Using $\sqrt{d}=\Omega(k_{\max})$, it yields that

$$\lambda_d\cdot\frac{k_{\max}}{\sqrt{d}} = O(\frac{k_{\max}\sqrt{\log(n+m)}}{d}) = O\left(\sqrt{\frac{\log(n+m)}{d}}\right) \tag{22}$$

Combining this with (18) and noting that $k_{\max} = O(n+m)$ concludes the proof. □

Corollary 1 introduces settings in which our theoretical bounds perform the best. In particular, when $\Omega((n+m)^{\epsilon_{\min}}) = k_{\min} \leq k_{\max} = O((n+m)^{\epsilon_{\max}})$ and $0 < \epsilon_{\min} \leq \epsilon_{\max} \ll 1/2$, the Lasso estimator requires a significantly smaller number of samples than $\Omega(m+n)$ to guarantee infinitesimal elementwise estimation error. However, notice that Theorem 1 is more general than this corollary and proves the elementwise consistency of the Lasso estimator even in the absence of such assumptions.

While it is shown that the proposed estimator can recover the correct sparsity structure of Θ with a relatively small number of sample trajectories, we prove in the next theorem that the least-squares estimator defined as (7) does not extract the correct sparsity structure of Θ for any finite number of sample trajectories.

Theorem 2. *If* A *and* B *are not fully dense matrices,* Θ_{ls} *does not recover the support of* Θ^* *for any finite number of sample trajectories with probability one.*

Proof. Define $R = ((X^TX)^{-1}X^T)^T$ and note that R and W are independent random variables due to the construction of X. Now, suppose that $\Theta^*_{ij} = 0$. We show that with probability zero, it holds that $E_{ij} = |(\Theta_{\text{ls}})_{ij} - \Theta^*_{ij}| = 0$ and hence $(\Theta_{\text{ls}})_{ij} = 0$. Note that $E_{ij} = R^T_{:,i}W_{:,j}$. If $R_{:,i} \neq 0$, then E_{ij} is a linear combination (with at least one nonzero coefficient) of identically distributed normal random variables with mean zero and variance $(\Sigma_w)_{jj}$. Since $R_{:,i}$ and $W_{:,j}$ are independent, we have $E_{ij} = 0$ with probability zero. Now, assume that $R_{:,i} = 0$. This means that the i^{th} row of R^T is a zero vector. This, in turn, implies that the i^{th} row of R^TX is zero. However, $R^TX = (X^TX)^{-1}X^TX = I$, which is a contradiction. This completes the proof. □

Remark 1. *In this work, we considered those LTI systems that are subject to input disturbance with Gaussian distribution. In many areas within control theory, including robust and*

model predictive control, it is more practical to assume that the input disturbances belong to a bounded set [23], [24]. Moreover, note that this work is based on random generation of inputs from a user-defined Gaussian distribution. Again, this may not be possible since in many applications, the input is constrained to be bounded in order to guarantee the well-being of the entire system. However, the general methodology of this paper is valid in the bounded disturbance-bounded input regime. This is due to the fact that many (although not all) non-asymptotic and high-dimensional properties of estimators involving Gaussian random variables can be carried over to the problems with bounded-support random variables; they benefit from the light-tailed properties of sub-Gaussian probability distributions. Due to the space limitations, these extensions are not included in this work. Similarly, the extension to noisy measurements is left out.

IV. Numerical Results

In this section, we illustrate the performance of the Lasso estimator and compare it with its least-square counterpart. We consider two case studies on synthetically generated systems and physical mass-spring networks. The simulations are run on a laptop computer with an Intel Core i7 quad-core 2.50 GHz CPU and 16GB RAM. The reported results are for a serial implementation in MATLAB R2017b and the function `lasso` is used to solve 9. Define the mismatch error as the total number of false positives and false negatives in the sparsity pattern of the estimator. Moreover, define *relative number of sample trajectories* (RST) as the number of sample trajectories normalized by the dimension of the system, and *relative mismatch error* (RME) as the mismatch error normalized by total number of elements in Θ.

A. Synthetically Generated Systems

Given the numbers n and w, and for each instance of the problem, the state and input matrices are constructed as follows: The diagonal elements of $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ are set to 1 (the dimensions of the inputs and states are chosen to be equal); the elements of the first w upper and lower diagonals of A and B are set to 0.3 or -0.3 with equal probability. Furthermore, at each row of A, another w elements are randomly chosen from the elements not belonging to the first w upper and lower diagonals and set to 0.3 or -0.3 with equal probability. The structure of $[A \quad B]$ is visualized in Figure 1 for $w = 2, 3, 4$. Based on this procedure, the number of nonzero elements at each row of $[A \quad B]$ varies between $3w+2$ and $5w+2$. Assuming that $\Sigma_u = I$ and $\Sigma_w = 0.5I$, the mutual incoherence property is satisfied for all constructed instances. To verify the developed theoretical results, λ_d is set to

$$\sqrt{\frac{2(1+\log(n+m))}{d}} \tag{23}$$

Note that this choice of λ_d does not require any additional fine-tuning and is at most a constant factor away from the lower bound introduced in (15).

In the first set of experiments, we consider the mismatch error of Θ_{lasso} with respect to the number of sample trajectories and for different system dimensions. The length of the time horizon T is set to 3. The results are illustrated in Figure 2a for $n+m$ equal to 200, 600, 1200, and 2000. In all of these test cases, w is chosen in such a way that the number of nonzero elements in each column of Θ is between $(n+m)^{0.3}$ and $(n+m)^{0.4}$. It can be observed that, as the dimension of the system increases, a higher number of sample trajectories is required to have a small mismatch error in the Lasso estimator. Conversely, the required value of RST to achieve a small RME reduces as the dimension of the system grows. More precisely, RST should be at least 1.80, 1.13, 0.37, and 0.20 to guarantee RME $\leq 0.1\%$, when $m+n$ is equal to 200, 600, 1200, and 2000, respectively.

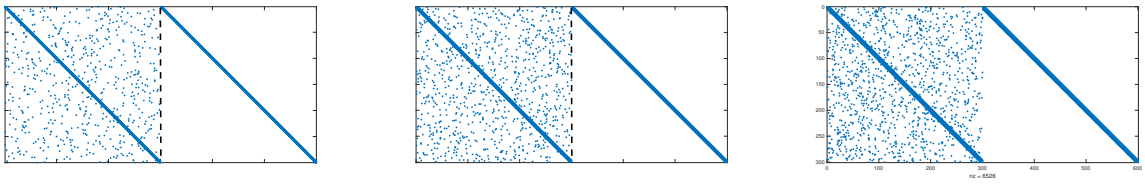
In the next set of experiments, we consider the mismatch error for different time horizons $T = 3, 4, ..., 7$, when fixing $m+n$ and w to 600 and 2, respectively. As mentioned before, large values of T tend to inflate the easily identifiable modes of the system and suppress the nearly hidden ones, thereby making it hard to obtain an accurate estimation of the parameters. It is pointed out that $\kappa(FF^T+GG^T)$ is a good indicator of the gap between these modes. This relationship is clearly reflected in Figures 2b and 2c. As can be observed in Figure 2b, 330 sample trajectories are enough to guarantee RME $\leq 0.1\%$ for $T = 3$. However, for $T = 7$, RME cannot be reduced below 0.42% with 1000 sample trajectories. To further elaborate on this dependency, Figure 2c is used to illustrate the value of $\kappa(FF^T+GG^T)$ with respect to T in a log-log scale. One can easily verify that $\kappa(FF^T+GG^T)$ associated with $T = 7$ is 485 times greater than this parameter for $T = 3$.

Finally, we study the Lasso estimator for different per-column number of nonzero elements in Θ and compare its accuracy to the least-squares estimator. Fixing $T = 3$ and $m+n = 600$, Figure 3a depicts the mismatch error of the Lasso estimator when the maximum number of nonzero elements at each column of Θ ranges from 7 (corresponding to $w = 1$) to 27 (corresponding to $w = 5$). Not surprisingly, the required number of samples to achieve a small mismatch error increases as the number of nonzero elements in each column of Θ grows. On the other hand, the least-squares estimator is fully dense in all of these experiments, regardless of the number of sample trajectories. To have a better comparison between the two estimators, we consider the 2-norm of the estimation errors normalized by the 2-norm of Θ^*, for different number of nonzero elements in each column of Θ^*. As it is evident in Figure 3b, the Lasso estimator significantly outperforms the least-squares estimator for any number of sample trajectories. Furthermore, the least-squares estimator is not defined for $d < 600$.

B. Mass-Spring Systems

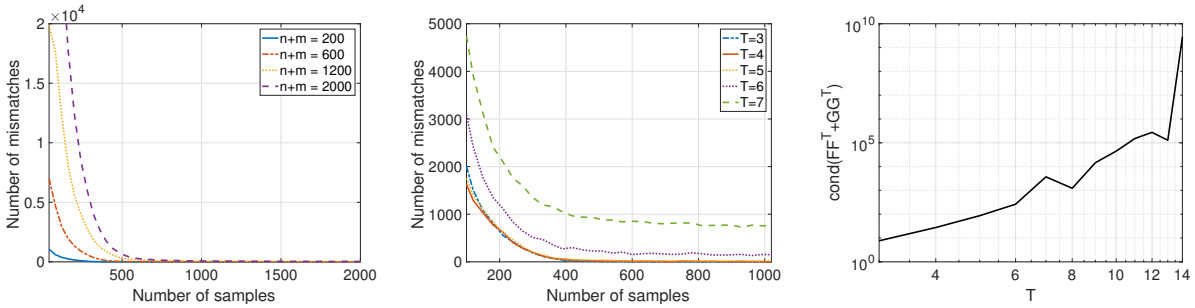
In this case study, we conduct simulations on mass-spring networks with different sizes to elucidate the performance of the Lasso estimator on physical systems. Consider N identical masses connected via a path of springs. Figure 4 exemplifies this system for $N = 2$. The state-space equation of this system in continuous domain can be written as

$$\dot{x}_c(t) = A_c x(t) + B_c u(t) \tag{24}$$



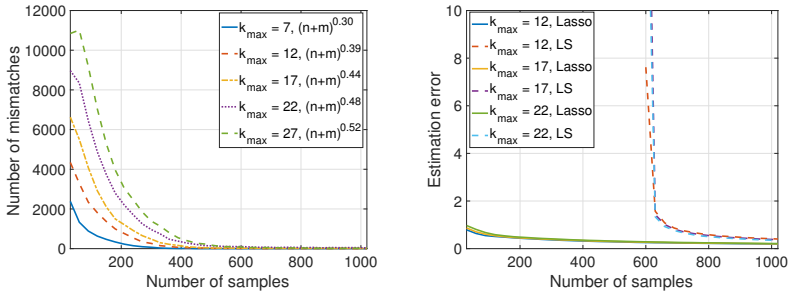
(a) $w = 2$ (b) $w = 3$ (c) $w = 4$

Fig. 1: The sparsity structure of the matrix $[A \;\; B]$ for $w = 2, 3, 4$.



(a) Mismatch error for different d (b) Mismatch error for different T (c) Condition number for different T

Fig. 2: (a) The mismatch error with respect to the number of sample trajectories for different system dimensions, (b) the mismatch error with respect to the number of sample trajectories for different time horizons, (c) the condition number of $FF^T + GG^T$ with respect to the time horizon.



(a) Mismatch error for different $k_{\max}$ (b) Estimation error for different $k_{\max}$

Fig. 3: (a) The mismatch error with respect to the number of sample trajectories for different per-column number of nonzero elements in Θ^*, (b) the normalized estimation error for Lasso and least-square (abbreviated as LS) estimators with respect to the number of sample trajectories.

where $A_c \in \mathbb{R}^{2N\times 2N}$, $B_c \in \mathbb{R}^{2N\times N}$, and $x_c(t)$ consists of two parts: the first N elements correspond to the location of the masses while the second N elements capture their velocity. With the unit masses and spring constants, we have

$$A_c = \begin{bmatrix} 0 & I \\ S & 0 \end{bmatrix}, \quad B_c = \begin{bmatrix} 0 \\ I \end{bmatrix} \tag{25}$$

where $S \in \mathbb{R}^{n\times n}$ is a tridiagonal matrix whose diagonal elements are set to -2 and its first upper and lower diagonal elements are set to 1 [25]. This continuous system is discretized using the forward Euler method with sampling time of 0.2s. Similar to the previous case, Σ_u and Σ_w are set to I and $0.5I$, respectively. Furthermore, T is set to 3 and λ_d is chosen as 23. The mutual incoherence property holds in all of these simulations. Notice that RST is equal to $3d/N$ in this case study, since $n = 2N$ and $m = N$. Figure 5a depicts the required RST to achieve RME $\leq 0.1\%$ for different number of masses. If $N = 30$, the minimum number of sample trajectories to guarantee a small mismatch error is 5 times the dimension of the system, whereas this number is dropped to 0.02 for $N = 500$. Furthermore, Figure 5b shows the normalized estimation errors of the Lasso and least-squares estimators for fixed $N = 300$ and $T = 3$. Similar to the previous case study, the proposed estimator results in significantly smaller estimation errors in medium- and large-sampling regime, while it remains as the only viable estimator when the number of available sample trajectories is smaller than 900.

V. Conclusion

We consider the problem of identifying the parameters of linear time-invariant (LTI) systems. In many real-world

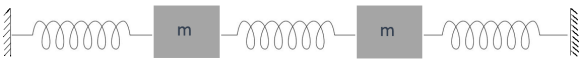


Fig. 4: Mass-spring system with two masses

problems, the state-space equation describing the evolution of the system admits a sparse representation, due to localized or internally limited interactions of states and inputs. In this work, we leverage this property and introduce a Lasso-type estimator to identify the parameters of the system. Using modern high-dimensional statistics, we derive sharp non-asymptotic bounds on the minimum number of input-state data samples to guarantee a small elementwise estimation error. In particular, upon defining n and m as the respective numbers of states and inputs, and $k_{\max}$ as the maximum number of nonzero elements in the rows of input and state matrices, we prove that $\Omega(k_{\max}\log(n+m))$ data samples suffice to ensure infinitesimal elementwise estimation error and exact recovery of the sparsity structure of the system. Through different case studies on synthetically generated systems and mass-spring networks, we demonstrate substantial improvements in the accuracy of the proposed estimator, compared to its well-known least-squares counterpart.

REFERENCES

[1] L. Ljung, "System identification," in *Signal analysis and prediction*. Springer, 1998, pp. 163–173.

[2] L. Ljung, "Convergence analysis of parametric identification methods," *IEEE transactions on automatic control*, vol. 23, no. 5, pp. 770–783, 1978.

[3] R. Pintelon and J. Schoukens, *System identification: a frequency domain approach*. John Wiley & Sons, 2012.

[4] E.-W. Bai, "Non-parametric nonlinear system identification: An asymptotic minimum mean squared error estimator," *IEEE Transactions on automatic control*, vol. 55, no. 7, pp. 1615–1626, 2010.

[5] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Annals of statistics*, pp. 295–327, 2001.

[6] D. R. Cox and D. V. Hinkley, *Theoretical statistics*. CRC Press, 1979.

[7] S. Ross and J. A. Bagnell, "Agnostic system identification for model-based reinforcement learning," *arXiv preprint arXiv:1203.1007*, 2012.

[8] S. Sadraddini and C. Belta, "Formal guarantees in data-driven model identification and control synthesis," in *21st ACM International Conference on Hybrid Systems: Computation and Control*. ACM, 2018.

[9] Z. Hou and S. Jin, "Data-driven model-free adaptive control for a class of mimo nonlinear discrete-time systems," *IEEE Transactions on Neural Networks*, vol. 22, no. 12, pp. 2173–2188, 2011.

[10] E. Weyer, R. C. Williamson, and I. M. Mareels, "Finite sample properties of linear model identification," *IEEE Transactions on Automatic Control*, vol. 44, no. 7, pp. 1370–1383, 1999.

[11] E. Weyer, "Finite sample properties of system identification of arx models under mixing conditions," *Automatica*, vol. 36, no. 9, pp. 1291–1299, 2000.

[12] J. Pereira, M. Ibrahimi, and A. Montanari, "Learning networks of stochastic differential equations," in *Advances in Neural Information Processing Systems*, 2010, pp. 172–180.

[13] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *arXiv preprint arXiv:1710.01688*, 2017.

[14] S. Tu, R. Boczar, A. Packard, and B. Recht, "Non-asymptotic analysis of robust control from coarse-grained identification," *arXiv preprint arXiv:1707.04791*, 2017.

[15] G. Darivianakis, S. Fattahi, J. Lygeros, and J. Lavaei, "High-performance cooperative distributed model predictive control for linear systems," *to appear in American Control Conference*, 2018.

[16] Y.-S. Wang, N. Matni, and J. C. Doyle, "Localized lqr optimal control," in *IEEE 53rd Conference on Decision and Control*, 2014, pp. 1661–1668.

[17] G. Fazelnia, R. Madani, A. Kalbat, and J. Lavaei, "Convex relaxation for optimal distributed control problems," *IEEE Transactions on Automatic Control*, vol. 62, no. 1, pp. 206–221, 2017.

[18] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *arXiv preprint arXiv:1710.01688*, 2017.

[19] P. Zhao and B. Yu, "On model selection consistency of lasso," *Journal of Machine learning research*, vol. 7, no. Nov, pp. 2541–2563, 2006.

[20] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *The annals of statistics*, pp. 1436–1462, 2006.

[21] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution," *Communications on pure and applied mathematics*, vol. 59, no. 6, pp. 797–829, 2006.

[22] E. Candes and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse problems*, vol. 23, no. 3, p. 969, 2007.

[23] D. Q. Mayne, M. M. Seron, and S. Raković, "Robust model predictive control of constrained linear systems with bounded disturbances," *Automatica*, vol. 41, no. 2, pp. 219–224, 2005.

[24] G. E. Dullerud and F. Paganini, *A course in robust control theory: a convex approach*. Springer Science & Business Media, 2013, vol. 36.

[25] F. Lin, M. Fardad, and M. R. Jovanović, "Design of optimal sparse feedback gains via the alternating direction method of multipliers," *IEEE Transactions on Automatic Control*, vol. 58, no. 9, pp. 2426–2431, 2013.

[26] S. N. Negahban and M. J. Wainwright, "Simultaneous support recovery in high dimensions: Benefits and perils of block ℓ_1/ℓ_∞-regularization," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3841–3863, 2011.

[27] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso)," *IEEE transactions on information theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

APPENDIX

In this section, we present the proof of Theorem 1. First, note that (4) can be reformulated as the set of linear equations

$$Y_{:,j} = X\Theta_{:,j} + W_{:,j} \quad \forall\, j = 1, ..., n \tag{26}$$

where $Y_{:,j}$, $\Theta_{:,j}$, and $W_{:,j}$ are the j^{th} column of Y, Θ, and W, respectively. Based on this definition, consider the following set of Lasso regression subproblems:

$$\Theta^j_{\text{lasso}} = \arg\min_{\Theta_{:,j}\in\mathbb{R}^d} \frac{1}{2d}\|Y_{:,j} - X\Theta_{:,j}\|_2^2 + \lambda_d\|\Theta_{:,j}\|_1 \tag{27}$$

Furthermore, define

$$\gamma_j = 1 - \max_{i\in\mathcal{A}_j^c}\{\|\tilde{\Sigma}_{i\mathcal{A}_j}(\tilde{\Sigma}_{\mathcal{A}_j\mathcal{A}_j})^{-1}\|_1\} \tag{28}$$

for every $j = 1, 2, ..., n$. Note that we have $0 < \gamma_j \leq 1$ due to Assumption 1. The following important lemma holds for the newly defined subproblems (27).

Lemma 1. *Given arbitrary constants $c_1, c_2, c_3 > 1$ and for every $1 \leq j \leq n$, suppose that d and λ_d satisfy*

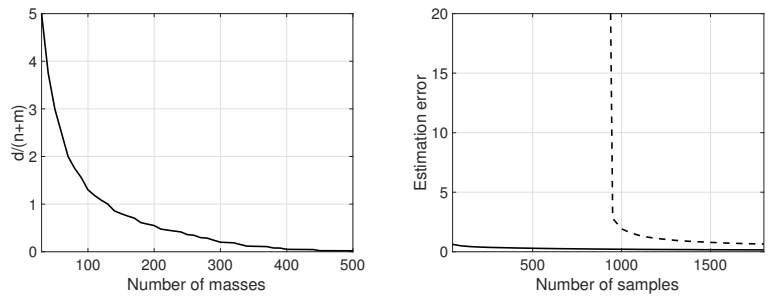
$$\lambda_d \geq \sqrt{\frac{32c_1\cdot(\Sigma_w)_{jj}\cdot\sigma_G}{\gamma_j^2}\cdot\frac{1+\log(n+m)}{d}} \tag{29}$$

$$d \geq \frac{72c_2}{\gamma_j^2}\cdot\kappa(\tilde{\Sigma})\cdot k_j\cdot(1+\log(n+m)) \tag{30}$$

Then, with probability at least

$$\begin{aligned} &1 - 3\exp\big(-(c_1-1)(1+\log(n+m))\big) \\ &\quad - 2\exp\big(-(c_2-1)(1+\log(n+m))\big) \\ &\quad - 2\exp\big(-2(c_3-1)(\log(k_j))\big) - 6\exp\left(-\frac{k_j}{2}\right) \to 1 \end{aligned} \tag{31}$$

the following statements hold:



(a) Relative mismatch error (b) Normalized estimation error

Fig. 5: (a) The minimum RST to guarantee RME $\leq 0.1\%$ with respect to the number of masses, (b) the normalized estimation error for Lasso and least-square estimators with respect to the number of sample trajectories.

1. *Θ^j_{lasso} is unique and in addition,*

$$\|\Theta^j_{lasso}-\Theta_{:,j}\|_\infty \leq \underbrace{\sqrt{\frac{36c_3(\Sigma_w)_{jj}\log(k_j)}{\Lambda_{\min}d}}+\lambda_d\left(\frac{8k_j}{\Lambda_{\min}\sqrt{d}}+q\right)}_{g_j} \tag{32}$$

2. *The support of Θ^j_{lasso} is a subset of the support of $\Theta_{:,j}$. Furthermore, the supports of Θ^j_{lasso} and $\Theta_{:,j}$ are identical if $\theta^j_{\min} > g_j$, where $\theta^j_{\min} = \min_{t\in\mathcal{A}_j}|\Theta_{tj}|$.*

Proof. The proof is based a the primal-dual witness approach introduced in [26], [27]. The details are omitted due to the space restrictions. □

Proof of Theorem 1: First, we present the sketch of the proof in a few steps:

1. We decompose the multivariate lasso regression problem (9) into n disjoint lasso regression subproblems defined in (27).
2. For each of these Lasso regression subproblems, we consider the event that Statements 1 and 2 of Theorem 1 hold.
3. We consider the intersection of these n events and show that they hold simultaneously with a probability that is lower bounded by (17).

Step 1: (9) can be rewritten as follows:

$$\Theta_{\text{lasso}} = \arg\min_{\Theta}\sum_{j=1}^{n}\left(\frac{1}{2d}\|X_{:,j}-Y\Theta_{:,j}\|_2^2+\lambda\|\Theta_{:,j}\|_1\right) \tag{33}$$

The above optimization problem can be naturally decomposed into n disjoint lasso regression subproblems in the form of (27).

Step 2: Assume that (30) and (29) hold for every $1 \leq j \leq n$. Upon defining $\mathcal{T}_j$ as the event that the first and second statements of Lemma 1 hold, one can write

$$\begin{aligned}\mathbb{P}(\mathcal{T}_j) \geq\ & 1-3\exp\big(-(c_1-1)(1+\log(n+m))\big)\\ & -2\exp\big(-(c_2-1)(1+\log(n+m))\big)\\ & -2\exp\big(-2(c_3-1)(\log(k_j))\big)-6\exp\left(-\frac{k_j}{2}\right)\end{aligned} \tag{34}$$

For every $1 \leq j \leq n$.

Step 3: Consider the event $\mathcal{T} = \mathcal{T}_1 \cap \mathcal{T}_2 \cap \cdots \cap \mathcal{T}_n$. Based on the Frechet inequality, we have

$$\mathbb{P}(\mathcal{T}_1)+\mathbb{P}(\mathcal{T}_2)+\cdots+\mathbb{P}(\mathcal{T}_n)-(n-1) \leq \mathbb{P}(\mathcal{T}) \tag{35}$$

Combining (34) and (35) yields that

$$\begin{aligned}\mathbb{P}(\mathcal{T}) \geq\ & 1-\underbrace{2\sum_{j=1}^{n}\exp\big(-2(c_3-1)(\log(k_j))\big)}_{(a)}\\ & -\underbrace{3n\exp\big(-(c_1-1)(1+\log(n+m))\big)}_{(b)}\\ & -\underbrace{2n\exp\big(-(c_2-1)(1+\log(n+m))\big)}_{(c)}\\ & -\underbrace{6n\exp\left(-\frac{k_j}{2}\right)}_{(d)}\end{aligned} \tag{36}$$

Denote $\mathcal{N}$ as the event that Statements 1 and 2 of the theorem hold. Note that the lower bounds introduced in (16) and (15) for d and λ_d guarantee the validity of (30) and (29) for every $1 \leq j \leq n$. Furthermore, based on (18) and (32), we have $g \geq g_j$ for every $1 \leq j \leq n$. This implies that $\mathbb{P}(\mathcal{T}) \leq \mathbb{P}(\mathcal{N})$. It remains to show that the right hand side in (36) is lower bounded by (17). We have

$$(a) \leq 2\cdot\frac{n}{(n+m)^{2\epsilon_{\min}(c_3-1)}} \tag{37a}$$

$$(b) \leq 3\exp(-(c_1-1))\frac{n}{(n+m)^{c_1-1}} \tag{37b}$$

$$(c) \leq 2\exp(-(c_2-1))\frac{n}{(n+m)^{c_2-1}} \tag{37c}$$

where (a) is due to the assumption $k_{\min} = \Omega((n+m)^{\epsilon_{\min}})$ and (b) and (c) follow from a simple calculation. In light of Assumption 2, one can write $k_{\max} = \Omega((n+m)^{\epsilon_{\max}})$ for some $\epsilon_{\max} > 0$. This yields that (d) is dominated by the right hand sides of (37a), (37b), and (37c). The proof is completed by noting that $c_1, c_2 > 2$ and $c_3 > \frac{1}{2\epsilon_{\min}} + 1$ are enough to ensure that (a), (b), (c), and (d) converge to zero as the dimension of the system grows. □