# A New Approach to Learning Linear Dynamical Systems 

Ainesh Bakshi<br>ainesh@mit.edu<br>Massachusetts Institute of Technology<br>Cambridge, Massachusetts, USA

Allen Liu<br>cliu568@mit.edu<br>Massachusetts Institute of Technology<br>Cambridge, Massachusetts, USA

Ankur Moitra<br>moitra@mit.edu<br>Massachusetts Institute of Technology<br>Cambridge, Massachusetts, USA

Morris Yau<br>morrisy@mit.edu<br>Massachusetts Institute of Technology<br>Cambridge, Massachusetts, USA


#### Abstract

Linear dynamical systems are the foundational statistical model upon which control theory is built. Both the celebrated Kalman filter and the linear quadratic regulator require knowledge of the system dynamics to provide analytic guarantees. Naturally, learning the dynamics of a linear dynamical system from linear measurements has been intensively studied since Rudolph Kalman's pioneering work in the 1960's. Towards these ends, we provide the first polynomial time algorithm for learning a linear dynamical system from a polynomial length trajectory up to polynomial error in the system parameters under essentially minimal assumptions; observability, controllability, and marginal stability. Our algorithm is built on a method of moments estimator to directly estimate Markov parameters from which the dynamics can be extracted. Furthermore we provide statistical lower bounds when our observability and controllability assumptions are violated.


## CCS CONCEPTS

- Theory of computation $\rightarrow$ Machine learning theory.


## KEYWORDS

linear dynamical systems, marginal stability, method of moments, information-theoretically minimal, minimal assumptions, observability, controllability

ACM Reference Format:
Ainesh Bakshi, Allen Liu, Ankur Moitra, and Morris Yau. 2023. A New Approach to Learning Linear Dynamical Systems. In Proceedings of the 55th Annual ACM Symposium on Theory of Computing (STOC '23), June 20-23, 2023, Orlando, FL, USA. ACM, New York, NY, USA, 14 pages. https: //doi.org/10.1145/3564246.3585247

## 1 INTRODUCTION

Linear dynamical systems are the canonical model for time series data. At each time step $t$ there is an unknown hidden state $x_{t} \in \mathbb{R}^{n}$ and a known exogenous input $u_{t} \in \mathbb{R}^{p}$. The transition dynamics


This work is licensed under a Creative Commons Attribution 4.0 International License.

STOC '23, 7une 20-23, 2023, Orlando, FL, USA
© 2023 Copyright held by the owner/author(s)
ACM ISBN 978-1-4503-9913-5/23/06.
https://doi.org/10.1145/3564246.3585247
and observations $y_{t} \in \mathbb{R}^{m}$ are generated according to the following rules:

$$
\begin{aligned}
x_{t+1} & =A x_{t}+B u_{t}+w_{t}, \\
y_{t} & =C x_{t}+D u_{t}+z_{t},
\end{aligned}
$$

Here $A, B, C$ and $D$ are matrices of dimension $n \times n, n \times p, m \times$ $n$ and $m \times p$ respectively. Moreover $w_{t}$ and $z_{t}$ are independent random variables and are called the process and observation noise respectively. It is standard to assume that they, along with the inputs and the initial state $x_{0}$, are all Gaussian, though we will work in a more general setting.

Linear dynamical systems have wide-ranging applications in control theory [26], computer vision [18], speech recognition [48], econometrics [2], healthcare [45] and neuroscience [61]. They are the de facto model of choice due to their mathematical simplicity and because, when the parameters are known, making predictions about subsequent observations and making inferences about the unknown state are both algorithmically tractable. In fact these algorithms are simple, practical and statistically optimal.
But what happens when the parameters are unknown? The problem of estimating $A, B, C$ and $D$ from input-output sequences is called system identification and has been intensively studied since Rudolph Kalman's pioneering work in the 1960's [35]. There is a well-developed theory that furnishes asymptotic guarantees [1, 46]. And more recently, many researchers have sought finite-sample guarantees both in the fully observed setting where $C=I[13,20$, 58,64 ] and in the partially observed setting [27, 51, 59, 63, 68]. Our focus here will be on obtaining running time and sample complexity bounds that are polynomial in the appropriate parameters and work under the most general conditions.

### 1.1 Previous Work

In the fully observed setting, the maximum likelihood estimator can be computed by solving ordinary least squares. It is known to be statistically optimal and there are strong finite sample guarantees on its performance [13, 20, 58, 64]. The partially observed setting is significantly more challenging because the problem of computing the maximum likelihood estimator becomes nonconvex. The EM algorithm [24] is often used in practice but it can get stuck in bad local minima. Our main focus will be on algorithms for learning partially observed linear dynamical systems with provable guarantees. There is a vast literature on this and related prediction problems (see Section 2). But all existing algorithms need to make one or more of the following types of restrictive assumptions:
(1) Assumptions about the characteristic polynomial $q$ of $A$ or the phases of its roots. Hardt, Ma and Recht [27] assumed that the image of the complex unit disk under $q$ is contained in the cone of complex numbers whose real part is larger than the absolute value of its imaginary part. For example, this is satisfied if the $\ell_{1}$-norm of the coefficients of $q$ is at most $\sqrt{2} / 2$. Hazan et al. [28] studied the problem of predicting subsequent observations in a non-stochastic setting. Their bounds depend on the $\ell_{1}$-norm of the coefficients of a polynomial $p$ that vanishes on the phases of the eigenvalues of $A$. In particular, when there are few distinct roots or they are pairwise separated, the $\ell_{1}$-norm of the coefficients of $p$ can be much smaller than for $q$. This notion was further refined by Simchowitz et al. [63]. However it is not clear why one would expect these norm bounds to be small. In many settings, unless there is extreme cancellation, the coefficients of $q$ would in fact be exponentially large.
(2) Strict stability and mixing. Another popular assumption is called strict stability, which stipulates that the spectral radius $\rho(A)<1$. Often the transition matrix $A$ only satisfies $\rho(A) \leqslant 1$, which is called marginal stability. Consider a classic application in control theory, of tracking an object from radar measurements. The state of the object at some time step is its position, velocity and acceleration. The transition matrix is derived from Newton's laws and is upper triangular with ones along the diagonal, and so all of its eigenvalues are one. There are many other such examples, particularly in econometrics and coming from discretizations of ODEs. Algorithms that assume strict stability generally have bounds that depend on $1 /(1-\rho(A))$ [27, 62]. Essentially, strict stability requires that the distribution of the $y_{t}$ 's eventually converges and that there are no long-range correlations. So after about $1 /(1-\rho(A))$ steps we essentially get fresh independent samples. Yet in many applications longrange correlations are an essential feature of the problem. Moreover getting around strict stability has many qualitative parallels with learning in graphical models without correlation decay [6], and learning in Gaussian graphical models without the restricted eigenvalue condition [38].
(3) Restrictions on the dimension, etc. Some algorithms only work in the single-input single-output setting, i.e. when $m=$ $p=1$ [27]. Others have bounds that depend exponentially on the size of the largest Jordan block of $A$, or even treat the number of parameters of the linear dynamical system as a constant [63].
By now, there is a standard blueprint which works as follows: The first step is to estimate the Markov parameters, given by

$$
\left[\begin{array}{lllll}
D & C B & C A B & \cdots & C A^{s} B
\end{array}\right]
$$

The second step is to apply the Ho-Kalman algorithm [30], which uses the Markov parameters to compute estimates $\hat{A}, \hat{B}, \hat{C}$ and $\hat{D}$ that are close to the true parameters in the appropriate metric. Oymak and Ozay [51] gave the first effective stability bounds for the Ho-Kalman algorithm. Thus the main issue is: How do you estimate the Markov parameters? Essentially all previous works use some form of linear regression. The analysis is based on expressing the observation $y_{t}$ as a linear function of the previous inputs and some
noise terms. The Markov parameters can then be extracted from the regressor. The noise terms are a function of observation and process noise and also the quantity $A^{s} x_{t-s}$, which captures how the state at some previous time step affects the current state. When $A$ is strictly stable, this term decays exponentially. But when $A$ is only marginally stable, controlling this error presents many challenges. Our main question is:

> Are there efficient algorithms for learning high-dimensional linear dynamical systems whose running time and sample complexity are polynomial in
> the appropriate parameters, and whose assumptions are essentially optimal?

### 1.2 Our Assumptions

It is important to draw a sharp distinction between the assumptions featured in the previous subsection and the more standard assumptions from control theory. In 1960, Rudolph Kalman [36] introduced the concepts of observability and controllability. Since then, it has been understood that they ought to in some sense govern what sorts of linear dynamical systems can be learned. In this subsection, we will review these assumptions and their natural quantitative counterparts.

Observability and Controllability. Consider the observability matrix: for an integer $s$, let

$$
O_{s}=\left[\begin{array}{llll}
C^{\top} & (C A)^{\top} & \ldots & \left(C A^{s-1}\right)^{\top}
\end{array}\right]^{\top}
$$

A linear dynamical system is observable if for some $s$, the matrix $O_{s}$ has full column rank. Intuitively, this condition ensures that there is no portion of the state space that we cannot observe eventually.

Now consider the controllability matrix: for an integer $s$, let

$$
Q_{s}=\left[\begin{array}{llll}
B & A B & \ldots & A^{s-1} B
\end{array}\right]
$$

A linear dynamical system is controllable if the controllability matrix has full row rank. Intuitively this condition ensures that there is no portion of the state space that cannot be reached by the appropriate inputs. If either observability or controllability are violated, it is information-theoretically impossible to learn.

While the full rank conditions are enough to build an asymptotic theory, we will need natural quantitative counterparts to get finite sample guarantees. In particular we assume that $O_{s}$ and $Q_{s}$ have bounded condition number for some $s$. These assumptions are usually made in addition to the ones from the previous subsection, as they are needed in the stability bounds for the Ho-Kalman algorithm [51]. Furthermore we show that (see Theorem 1.4) they are information-theoretically necessary in order to learn a linear dynamical system from a polynomial length trajectory.

Finally, as is standard, we also assume that the system is nonexplosive, i.e. the eigenvalues of $A$ are bounded by 1 in magnitude. Note that assuming the eigenvalues are bounded is much weaker than assuming the singular values are bounded (e.g. consider the types of upper triangular matrices that arise in control theory, including $n$-dimensional integrators [57]).

Relaxed Control and Noise. In the literature, the standard assumption is that the initial state, the process and observation noise are all drawn from a Gaussian. But Gaussianity is not meant to literally be
true and it is often assumed for convenience. We show that we can dramatically relax this assumption to allow heavy-tailed distributions instead. In particular, for the control input, $u_{t}$, we only require the underlying distribution to have well-behaved fourth-moments:

Definition 1.1 ((4,2)-Hypercontactivity). A distribution $\mathcal{D}$ over $\mathbb{R}^{d}$ is (4,2)-hypercontractive if for all $v$,

$$
\underset{x \sim \mathcal{D}}{\mathbb{E}}\left[\langle x, v\rangle^{4}\right] \leqslant O(1) \underset{x \sim \mathcal{D}}{\mathbb{E}}\left[\langle x, v\rangle^{2}\right]^{2} .
$$

We note that several families of distributions are hypercontractive, including Gaussians, uniform distributions over the hypercube, sphere and other convex bodies, the Laplace, gamma, chi-squared, Wishart, Dirichlet and beta distributions, and in general, all logconcave distributions. Further, the set of hypercontractive distributions is closed under affine transformations, products and mixtures.

Finally, we only require that the distributions of the process noise, $w_{t}$, and observation noise, $z_{t}$, have bounded covariance. We state these assumptions formally in Section 4.

### 1.3 Our Results

Our approach is based on the method-of-moments rather than leastsquares regression. Our starting point is the following folklore observation: for any integers $t>j$,

$$
\underset{u_{t} \sim \mathcal{D}_{u}}{\mathbb{E}}\left[y_{t+j} u_{t}^{\top}\right]=\left\{\begin{array}{lc}
D & \text { if } j=0  \tag{1}\\
C A^{j-1} B & \text { otherwise }
\end{array}\right.
$$

However getting accurate estimates of the Markov parameters is a challenging task. For a fixed $j$, since the expectation of the estimator $y_{t+j} u_{t}^{\top}$ does not depend on $t$, a natural approach to estimate $\mathbb{E}\left[y_{t+j} u_{t}^{\top}\right]$ is to average over several control-observation pairs: $\widehat{C A^{j-1} B}=\frac{1}{T} \sum_{t \in[T]} y_{t+j} u_{t}^{\top}$ and hope that this estimator converges to its expectation. Unfortunately, this is just not true!

The first issue is that samples of the form $y_{t+j} u_{t}^{\top}$ are not independent for different values of $t$. The second issue is that, in the marginally stable setting, the variance of this statistic grows with $t$, even when the control and the noise are Gaussian (see the appendix of the full version for an example). Thus, directly using the empirical estimate can be highly inaccurate no matter how long our trajectory is. One of the key steps in our algorithm is to learn a transformation of the observations to a new time series $\left\{\hat{y}_{1}, \hat{y}_{2}, \ldots, \hat{y}_{T}\right\}$ such that $\mathbb{E}\left[\hat{y}_{t+j} u_{t}^{\top}\right]=C A^{j-1} B$ and the variance of our estimator is bounded. As a result, we obtain the following theorem:

Theorem 1.2 (Efficiently Learning a Linear Dynamical System, informal Theorem 6.4). Given $\epsilon>0$, a fixed polynomial length trajectory from a linear dynamical system satisfying mild nondegeneracy assumptions (see Subsection 1.2), there exists an algorithm that outputs estimates $\hat{A}, \hat{B}, \hat{C}, \hat{D}$ such that with probability at least $9 / 10$, there exists a similarity transform $U$ satisfying

$$
\left\|A-U^{-1} \hat{A} U\right\| \leqslant \epsilon,\left\|B-U^{-1} \hat{B}\right\| \leqslant \epsilon,\|C-\hat{C} U\| \leqslant \epsilon,\|D-\hat{D}\| \leqslant \epsilon .
$$

Further, the algorithm runs in time that is a fixed polynomial in all the parameters.
Remark 1.3. Note that it is only possible to recover the system parameters up to some global transformation $U$ since all such transformations lead to equivalent dynamics, see e.g. [51].

The main appeal of our algorithm is that it works in essentially the most general setting possible. In particular we show the following lower bound:

Theorem 1.4 (Sample Complexity Lower Bound for Ill-Conditioned Systems). [Informal, see Theorem 9.4] If for an LDS, the observability matrix $O_{s}$ has smallest singular value less than $\delta$ for all orders $s$, then any algorithm that uses less than $\sim 1 / \sqrt{\delta}$ length trajectories incurs constant error in estimating $A, B, C, D$ with constant probability. The same statement holds with the observability matrix $O_{s}$ replaced by the controllability matrix $Q_{s}$. In particular, if $\delta$ is exponentially small, then an exponential number of samples are required to learn the parameters.

It turns out that super-resolution [7, 17], namely the task of recovering a sparse signal from noisy low-frequency measurements, corresponds to a special case of learning linear dynamical systems. It is known that super-resolution exhibits a sharp phase transition, where the problem goes from having efficient algorithms with polynomial running time and sample complexity, to being information-theoretically impossible, unless the noise is exponentially small [49]. Thus there are some linear dynamical systems where it is impossible to learn the true parameters with bounded length trajectories. We refine this connection to show instancewise lower bounds for learning any linear dynamical system whose observability or controllability matrices are close to singular. Thus the assumptions our algorithm needs are qualitatively tight, and our results close the question of what linear dynamical systems can be efficiently learned.

## 2 RELATED WORK

Linear Time Invariant Systems: Identification, Prediction, Estimation. There is a long history of identifying linear dynamical systems from measurements, see [23] for extensive references. A focus of these works is on the "pre-filtering" approach to handling long range correlations in learning dynamical systems, see [15] [70] [66] [63]. Recently, there is a flurry of work on prediction and estimation for LDS's through the framework of no regret learning both in the fully observable setting [65] [58] [19] and the partially observed setting [62] [27]. For a variety of assumptions on the dynamics matrix, such as diagonalizability, there is work on learning marginally stable LDS's [28] [29]. Many works take a regression approach to estimating the markov parameters of the LDS for strictly stable systems see [44] [60] [21] [16]. In these settings it is possible to take advantage of the decay of the coefficients of the associated regressors. Marginal stability can be handled with multiple trajectories see [71] [67]. Closed loop system identification has also been studied; see [43] [42].

Somewhat related to our work is the problem of prediction without system identification in marginally stable LDS's [68] [25] but with an assumption on the exponential decay of the kalman filter coefficients. In [56], the exponential stability of the Kalman filter assumption is removed via a procedure that builds a succinct bank of filters for the prediction task and with an additional assumption on the dynamics having real eigenvalues. For a survey of the area see [69].

Relaxing distributional assumptions. In recent years, there has been a tremendous amount of work on designing algorithms that do not rely on strong distributional assumptions, such as Gaussianity, and only require much milder conditions. In particular, hypercontractivity of linear functions and low-degree polynomials have been identified as key analytic conditions that admit efficient algorithms for numerous problems in high-dimensional algorithmic statistics. In particular, souped up variants of hypercontractivity are used for heavy-tailed mean and covariance estimation [11, 31, 47], robust moment estimation [41], robust regression [5, 10, 34, 39, 52, 53, 72], robustly clustering mixture models [4, 14, 32, 40] and list-decodable learning [ $3,12,33,37,54,55$ ]. Algorithms with relaxed distributional assumptions were also recently given for online regression [9] and Kalman filtering [8].

## 3 TECHNICAL OVERVIEW

In this section, we describe our key algorithmic ideas and the corresponding technical challenges involved.

### 3.1 A Thought Experiment

Consider the setting where we already know the parameters, $A, B, C$ and $D$, of the underlying linear dynamical system. While there is nothing left to learn in this setting, we can still ask whether there exists a transformation of the observations $\left\{y_{t}\right\}_{t \in[0, T]}$ to a new time series $\left\{\hat{y}_{t}\right\}_{t \in[0, T]}$, such that the variance of the random variable $\hat{y}_{t+j} u_{t}$ is bounded.

It is indeed possible to do so by considering a simple linear transformation of the observations: let $\hat{y}_{t}=y_{t}-\sum_{j=1}^{n} c_{j} y_{t-j}$, where the $c_{j}$ 's are the coefficients of the characteristic polynomial of $A$. To see why this works, we recall that by the Cayley-Hamilton theorem (see Fact 5.4), the coefficients of the characteristic polynomial satisfy the following algebraic identity:

$$
\begin{equation*}
A^{n}-\sum_{j \in[n]} c_{j} A^{n-j}=0 \tag{2}
\end{equation*}
$$

Therefore, assuming (for the purposes of exposition) that $w_{t}$ 's and $z_{t}$ 's are bounded, and $D=0$, we have

$$
\begin{align*}
\hat{y}_{t} & =y_{t}-\sum_{j=1}^{n} c_{j} y_{t-j} \\
& =\underbrace{\sum_{i=1}^{n}\left(C A^{i-1} B-\sum_{j=1}^{i-1} c_{j} C A^{i-j-1} B\right) u_{t-i}}_{(a)}  \tag{3}\\
& +\underbrace{\sum_{i=n+1}^{t}\left(C A^{i-n-1}\left(A^{n}-\sum_{j=1}^{n} c_{j} A^{n-j}\right) B\right) u_{t-i}}_{(b)}
\end{align*}
$$

We note that term (a) above only has $n$ terms and does not grow as a function of $t$, and term (b) is in fact zero, since we can repeatedly apply the identity from Equation (2). A simple computation then implies that the estimator $\hat{y}_{t+j} u_{t}^{\top}$ satisfies $\mathbb{E}\left[\hat{y}_{t+j} u_{t}^{\top}\right]=C A^{j-1} B$ and has bounded variance. We replicate this thought experiment, by learning the coefficients that stabilize the variance of our estimator
from the observations directly. We dedicate the rest of the technical overview to describe how we accomplish this task.

### 3.2 Learning the Stabilizing Transform

For ease of exposition, we assume that $C$ is a $1 \times n$ matrix and therefore the resulting observations, $y_{t}$, are scalars. A natural approach is to then consider the following least-squares regression problem:

$$
\min _{c_{1}, c_{2}, \ldots c_{n}} \sum_{t \in[T]}\left(y_{t}-\sum_{j \in[n]} c_{j} y_{t-j}\right)^{2}
$$

We know that the coefficients of the characteristic polynomial are a feasible solution to this regression problem, and the resulting linear transformation of the observations results in an estimator with bounded variance. Such an approach also appears in [63] but they incur unspecified, potentially exponential dependencies on the system parameters due to the complexities of analyzing this regression problem directly. In particular, we do not have finegrained control over the solution returned by solving the regression problem, and apriori, the regression solution need not be close to the coefficients of the characteristic polynomial of $A$.

Convex Program. Instead, we take a more direct approach to stabilizing the variance and consider a different convex program, specifically designed to do so. In particular, we find a vector $\alpha=$ ( $\alpha_{1}, \alpha_{2}, \ldots \alpha_{s}$ ) such that the following constraint system is feasible:

$$
C_{\alpha}=\left\{\begin{array}{lr}
\forall j \in[s] & \left|\alpha_{j}\right|^{2} \leqslant P_{0}  \tag{4}\\
\forall i \in[T] & \left|y_{i+k}-\sum_{j \in[s]} \alpha_{j} \cdot y_{i-j}\right|^{2} \leqslant P_{1}
\end{array}\right\}
$$

where $s$ is the integer satisfying the observability and controlability assumptions from Definition 4.4 , and $P_{0}$ and $P_{1}$ are sufficiently large polynomials in the system parameters (see Algorithm 6.6 for details).

Intuitively, the first constraint posits that each coefficient, $\alpha_{j}$ is bounded in magnitude. This is necessary since the process and observation noise scale proportional to the coefficients in the linear transformation, and we cannot afford to pay exponentially in these quantities. The second constraint posits that the resulting observations themselves are bounded, and tries to enforce a universal bound on the variance of each $\hat{y}_{t}$ appearing in the estimator $\frac{1}{T} \sum_{t \in[T]} \hat{y}_{t+k} u_{t}^{\top}$.

Feasibility. Observe, in contrast to the characteristic polynomial, we are only taking a linear combination of the previous $s$ (potentially $\ll n$ ) observations. While this difference does not manifest itself when $C$ is $1 \times n$, it becomes crucial when $C$ is $m \times n$ for $m>1$ for obtaining guarantees that depend only on the observability and controllability matrix. Also, note that we are expressing $y_{t+k}$, rather than $y_{t}$, as a linear combination of $y_{t-1}, \ldots y_{y-s}$. This difference is also crucial in the construction and analysis of our estimator.

In order to establish feasibility of $C_{\alpha}$, we invoke the observability assumption: since $O_{s}$ has bounded condition number, there exists a vector $\alpha^{*}=\left(\alpha_{1}^{*}, \ldots, \alpha_{s}^{*}\right)$ such that each $\alpha$ is bounded and the following identity holds:

$$
C A^{k+s}-\alpha_{1} C A^{s-1}-\alpha_{2} C A^{s-2}-\ldots-\alpha_{s} C=0
$$

We then follow an argument similar to the one in Equation (3) to show that the magnitude of $\hat{y}_{t}$ 's is bounded (see Lemma 7.2 for details). We also note that the above program is convex, and admits an efficient separation oracle, and therefore, we can find a feasible $\alpha$ in polynomial time. Interestingly, the feasibility analysis only requires that the covariance of the control input, process noise and observation noise be bounded, and does not require strong assumptions such as sub-Gaussian tails.

The Anti-concentration Potential. Next, we show that any feasible solution to the constraint system actually yields a stabilized estimator. To accomplish this goal, we design a potential function that captures the variance of our estimator, and argue that if the potential is large, with high probability, some constraint in $C_{\alpha}$ must be violated. In particular, for any vector $\alpha$, and integer $l$, we consider the potential

$$
\mathcal{G}_{\alpha, l}=\sum_{i=0}^{l}\left\|F_{\alpha}(A) A^{i} B\right\|_{F}^{2}
$$

where

$$
F_{\alpha}(A)=C A^{k+s}-\alpha_{1} C A^{s-1}-\alpha_{2} C A^{s-2}-\ldots-\alpha_{s} C
$$

We observe that the terms appearing in this potential are the trace of $\mathbb{E}\left[\left(F_{\alpha} A^{i} B u_{t-i}\right)^{\top}\left(F_{\alpha} A^{i} B u_{t-i}\right)\right]$, which captures how large intermediate terms are, as a function of $\alpha$. In particular, if $\alpha=\alpha^{*}$, the trace would be 0 . We make this intuition precise in Lemma 7.4.

Next, we show that we can split up the terms appearing in $\hat{y}_{t+k}$ into three parts as follows:

$$
\hat{y}_{t+k}=X_{t}+V_{t}+W_{t}
$$

where $X_{t}$ is polynomially bounded in the system parameters, $V_{t}$ is a random variable such that the covariance matrix of $V_{t}$, denoted by $\Sigma_{V_{t}}$ satisfies $\operatorname{Tr}\left(\Sigma_{V_{t}}\right)=\mathcal{G}_{\alpha, l}$, and $W_{t}$ is a random variable that we do not have control over, and may potentially be unbounded. This presents obstacle since $W_{t}$ can wipe out the information contained in $V_{t}$, and $\mathcal{G}_{\alpha, l}$ may be large without violating any constraint in $C_{\alpha}$.

Here, we observe that such an event can be avoided precisely when the random variable $V_{t}$ is anti-concentrated, i.e. the probability that $V_{t}$ lands in a ball of small radius is small. Perhaps counterintuitively, we show that if the 4 -th moment of $V_{t}$ concentrates, then it already possesses the anti-concentration properties we require. We make this precise in Lemma 5.9, where we establish a Paley-Zigmund style inequality, showing that if a random variable is (4, 2)-hypercontractive (see Definition 5.7), then the probability it lands in any interval that is a constant fraction of its variance is bounded by a constant. Finally, we show that $V_{t}$ is $(4,2)$ hypercontractive if the control inputs are (4,2)-hypercontractive, and therefore, significantly relax the Gaussianity assumption.

To summarize, we show that if the potential is large, the magnitude of $V_{t}$ is large, and since $V_{t}$ is anti-concentrated, $W_{t}$ cannot wash away this information. Therefore, $\left|\hat{y}_{t+k}\right|$ must be large, for some $t$, which is a contradiction to the feasibility of $C_{\alpha}$.

Dependent Random Variables and Decoupling. We then establish that if the potential $\mathcal{G}_{\alpha}$ is small, for a fixed setting of $\alpha$, the resulting
estimator has bounded variance:

$$
\begin{equation*}
\mathbb{E}\left[\left\|\frac{1}{L} \sum_{t \in[L]} \hat{y}_{t+j} u_{t}^{\top}-C A^{j-1} B\right\|^{2}\right] \leqslant \frac{P_{1}}{L}\left(\|\alpha\|^{2}+G_{\alpha, L}\right) \tag{5}
\end{equation*}
$$

for some fixed polynomial $P_{1}$ in the system parameters. We treat $L$ as a sufficiently large polynomial in the system parameters and $1 / \varepsilon$ (where $\varepsilon$ is the desired accuracy).

This argument is fairly involved and heavily uses the independence of the $u_{t}, w_{t}$ and $z_{t}$ 's. We refer the reader to Lemma 7.8 for a complete proof. While the above inequality holds for a fixed setting of $\alpha$, we note that the $\alpha$ 's output by solving the constraint $C_{\alpha}$ themselves depend on the randomness in the control input and the noise non-trivially.

To overcome this issue, we decouple the $\alpha$ 's from $u_{t}$ 's and establish a symbolic matrix inequality, where the matrices only depend on the $u_{t}$ 's. Here, we treat the vector $v_{\alpha}=\left(1, \alpha_{1}, \ldots, \alpha_{s}\right)$ as a formal variable, and write the potential as a quadratic form in the vector $v_{\alpha}$ :

$$
\mathcal{G}_{\alpha, L}=v_{\alpha}^{\top} G_{L} v_{\alpha}
$$

where $G_{L}$ is a PSD matrix. We note that such a representation always exists and is unique since $\mathcal{G}_{\alpha, L}$ is a sum-of-squares in $\alpha$. Similarly, we observe that the variance we want to bound admits a similar decomposition: let $M_{j}$ be the PSD matrix such that

$$
\mathbb{E}\left[\left\|\frac{1}{L} \sum_{t \in[L]} \hat{y}_{t+j} u_{t}^{\top}-C A^{j-1} B\right\|^{2}\right]=v_{\alpha}^{\top} M_{j} v_{\alpha}
$$

Observe, the matrices $G_{L}$ and $M_{j}$ are independent of the formal variables $\alpha$, and in Corollary 7.10 we establish the following inequality:

$$
\begin{equation*}
\underset{u, w, z}{\mathbb{E}}\left[M_{j}\right] \leq \frac{P_{1}}{L}\left(I+G_{L}\right) \tag{6}
\end{equation*}
$$

Since the above inequality holds for all quadratic forms simultaneously, one natural way to proceed would be to consider an $\epsilon$-net over the $\alpha$ 's and union bound over each vector in the net satisfying the quadratic form in Equation (6). To execute this, we need a net that is fine enough to account for how much error we accumulate in a term of the from $\hat{y}_{t+j} u_{t}^{\top}$. Note, the largest terms we need to account for are roughly of the form $\alpha_{j} A^{L}$, which naively requires a $1 /\|A\|^{L}$-net. Unfortunately, this net is too fine and we cannot afford to union bound over all the vectors in this net because in some sense we only have $L$ samples.

Bounded Eigenvalues to Smaller Nets. To address the issue above, we show that for any $n \times n$ matrix $A$ with complex entries, if the eigenvalues of $A$ are bounded by 1 in magnitude, the operator norm of $A^{L}$ in fact grows as $L^{n}$, instead of exponentially in $L$ (see Lemma 5.6 for a precise statement). Here, we crucially note that we only assume the eigenvalues, instead of the singular values (which would make this statement trivial but would rule out several important families of linear dynamical systems), are bounded. With this insight, we can then union bound over all vectors $\alpha$ in a $1 / L^{n_{-}}$ net, and establish equation (5) for the $\alpha$ output by solving the constraint system $C_{\alpha}$. To conclude, we have shown that the variance of our estimator is bounded, and therefore, we can estimate each block of the Markov Parameter matrix.

### 3.3 Lower Bound for Ill-Conditioned LDS's

In light of our results on learning LDS's, a natural question is whether the assumptions on observability and controllability are necessary. We exhibit information-theoretic lower bounds on the sample complexity of learning a linear dynamical systems with Gaussian noise and Gaussian inputs, when the observability or controllability matrices are exponentially ill conditioned. For this section, we discuss the case when the observability matrix is ill conditioned. The case when the controllabiltiy matrix is ill conditioned follows essentially the same argument.
Definition 3.1. We say that an $\operatorname{LDS} \mathcal{L}(A, B, C, D)$ is $(\delta, v)$-unobservable if $v$ is a unit vector such that for all integers $s \geqslant 0$,

$$
\left\|C A^{s} v\right\| \leqslant \delta
$$

Now the key to proving an information-theoretic lower bound is the observation that when the input and noise distributions are Gaussian, the system measurements and control inputs

$$
\left(u_{0}, \ldots, u_{T}, y_{0}, \ldots, y_{T}\right)
$$

are a Gaussian process. Therefore, the joint distribution is uniquely determined by its covariance matrix. On the other hand, we can explicitly compute the covariance matrix in terms of the system parameters $A, B, C, D$. While there are several terms in the expression, the main point is that essentially all of the terms look like $C A^{j} B$. Now, when the system is $(\delta, v)$-unobservable, we can replace $B$ with $B+v u^{\top}$ for an arbitrary vector $u \in \mathbb{R}^{p}$ while only changing expressions of the form $C A^{j} B$ by a little bit.

Overall, we can show that the pair of LDS's $\mathcal{L}=\mathcal{L}(A, B, C, D)$ and $\mathcal{L}^{\prime}=\mathcal{L}\left(A, B+v u^{\top}, C, D\right)$ are statistically close up to time $T \ll$ $1 / \delta$. This means that no algorithm using $\ll 1 / \delta$ length trajectories can distinguish the two systems and since their parameters are not close to equivalent up to similarity (for generic choices of parameters), the algorithm must incur large error.

## 4 FORMAL SETUP

In this section, we formally state the linear dynamical system model, and our assumptions.

Model 4.1 (Linear Dynamical System). Let $A \in \mathbb{C}^{n \times n}, B \in$ $\mathbb{C}^{n \times p}, C \in \mathbb{C}^{m \times n}$, and $D \in \mathbb{C}^{m \times p}$ be complex valued matrices. Let $\mathcal{D}_{0}, \mathcal{D}_{u}, \mathcal{D}_{w}, \mathcal{D}_{z}$ be distributions with mean zero. Then, a Linear Dynamical System, $\mathcal{L}(A, B, C, D)$, is defined as follows:

$$
\begin{aligned}
x_{t+1} & =A x_{t}+B u_{t}+w_{t} \\
y_{t} & =C x_{t}+D u_{t}+z_{t}
\end{aligned}
$$

where $x_{0} \sim \mathcal{D}_{0}$, and for all $t \in \mathbb{N}$, $u_{t} \sim \mathcal{D}_{u}, w_{t} \sim \mathcal{D}_{w}$ and $z_{t} \sim \mathcal{D}_{z}$.
We see only the sequence of observations $y_{1}, y_{2}, \ldots, y_{T}$ up to some time $T$ and our goal is to learn the parameters of the system $A, B, C, D$. We need some assumptions about the parameters and also the input and noise distributions which we discuss below (as otherwise the system may be degenerate and it may be informationtheoretically impossible to learn, see Section 9).

### 4.1 Assumptions on the System Parameters

We begin by ensuring that the linear dynamical system at hand is not degenerate. This notion can be made precise by considering the observability matrix:

Definition 4.2 (Observability Matrix). For an integer $s$, define the matrix $O_{s} \in \mathbb{R}^{s m \times n}$ as

$$
O_{s}=\left[\begin{array}{c}
C \\
C A \\
\vdots \\
C A^{s-1}
\end{array}\right]
$$

A LDS is observable if for some $s$, the matrix $O_{s}$ has full columnrank. Similarly, we need to ensure that the control input is not degenerate, and only acts in a subspace that is not spanned by $A$. This is made precise by considering the controllability matrix:
Definition 4.3 (Controllability Matrix). For an integer $s$, define the matrix $Q_{s} \in \mathbb{R}^{s p \times n}$ as

$$
Q_{s}=\left[\begin{array}{llll}
B & A B & \ldots & A^{s-1} B
\end{array}\right]
$$

A LDS is controllable is the controllability matrix has full rowrank. We note that we assume a quantitative strengthening of these two assumptions to $O_{s}$ and $Q_{s}$ having bounded condition number (and this is necessary, recall Theorem 1.4).
Definition 4.4 (Well-Behaved Linear Dynamical System). We say a linear dynamical system $\mathcal{L}(A, B, C, D)$ is well-behaved if the following assumptions hold:
(1) Non-trivial Controller. The matrix $B$ satisfies $\|B\| \geqslant 1$.
(2) Non-trivial Measurement. The matrix $C$ satisfies $\|C\| \geqslant 1$.
(3) Non-exposive System. All eigenvalues of $A$ have magnitude at most 1.
(4) Bounded Condition Number. $O_{s}$ has full column-rank, $Q_{s}$ has full row-rank and for some integer $s$, and parameter $\kappa \geqslant 1$,

$$
\begin{aligned}
& \sigma_{\max }\left(O_{2 s}\right) / \sigma_{\min }\left(O_{s}\right) \leqslant \kappa \\
& \sigma_{\max }\left(Q_{2 s}\right) / \sigma_{\min }\left(Q_{s}\right) \leqslant \kappa
\end{aligned}
$$

Remark 4.5. Crucially, the above assumption is on the eigenvalues and not the singular values of $A$, which would be a far stronger assumption, as discussed in the introduction.
Remark 4.6. While the bounded condition number assumption is standard in the literature $[22,50]$, in Section 9 , we show that a polynomial bound on the condition number is necessary in the sample complexity, even information-theoretically. As a consequence, it is impossible to learn an exponentially ill-conditioned system with polynomially bounded observations. Finally, note that up to polynomial factors, it suffices to have a bound on $\sigma_{\max }\left(O_{t}\right) / \sigma_{\min }\left(O_{s}\right)$ as long as $t / s>1+c$ for some positive constant $c$ (see Claim 5.16). For simplicity, we wrote the above condition for $t=2 s$.

### 4.2 Assumptions on the Distribution of the Control and Noise

We consider the following assumptions over the control input, system and process noise distributions:

Definition 4.7 (Distributional Assumptions). For all $t \in[T]$, we assume that $u_{t} \sim \mathcal{D}_{u}, w_{t} \sim \mathcal{D}_{w}$ and $z_{t} \sim \mathcal{D}_{z}$ are each sampled independently from the corresponding distributions. Additionally, $x_{0} \sim \mathcal{D}_{0}$. Then,

- Mean Zero: $\mathcal{D}_{u}, \mathcal{D}_{w}, \mathcal{D}_{z}$ and $\mathcal{D}_{0}$ are all mean 0 distribution.
- Isotropic and Hypercontractive Control: The covariance of $\mathcal{D}_{u}, \Sigma_{\mathcal{D}_{u}}=I$, and $\mathcal{D}_{u}$ is (4, 2, K)-hypercontractive for a fixed constant $K \geqslant 3$ (see Definition 5.7).
- Bounded Variance Noise: For $\sigma_{w}, \sigma_{z} \geqslant 1$, the covariances of $\mathcal{D}_{w}$ and $\mathcal{D}_{z}$ satisfy $\Sigma_{\mathcal{D}_{w}} \leq \sigma_{w} I$ and $\Sigma_{\mathcal{D}_{z}} \leq \sigma_{z} I$.
- Starting Point: The distribution $\mathcal{D}_{0}$ has covariance $\Sigma_{\mathcal{D}_{0}} \leq$ $\sigma_{0} I$

Remark 4.8. We don't actually need that the $u_{i}, w_{i}, z_{i}$ are all drawn from the same distribution across different time-steps. We only need that they are independent. In other words, all of our results still hold if we allow for there to be different distributions $\mathcal{D}_{u, t}, \mathcal{D}_{w, t}, \mathcal{D}_{z, t}$ at each time-step that all satisfy the above assumptions.

## 5 PRELIMINARIES

We begin with some notation and basic facts from linear algebra and probability. For a matrix $A \in \mathbb{C}^{n \times n}$, we use $\|A\|=\|A\|_{\text {op }}=$ $\max _{\|u\|=1}\|A u\|_{2}$ and $\|A\|_{F}=\sqrt{\sum_{i, j \in[n]}\left|A_{i, j}\right|^{2}}$. We use the notation $A^{\top}$ to denote the transpose when $A$ only has real entries. Further, for $A \in \mathbb{C}^{n \times m}$ such that $n \geqslant m$, let $\operatorname{SVD}(A)=U \Sigma V^{\top}$ denote the singular value decomposition of $A$, where $U \in \mathbb{C}^{n \times m}$ and $V^{\top} \in$ $\mathbb{C}^{m \times m}$ are unitary matrices (see Definition 5.1 , and $\Sigma$ is a diagonal matrix, with the singular values denoted by $\sigma_{1} \geqslant \sigma_{2} \geqslant \ldots \sigma_{m} \geqslant 0$.

### 5.1 Linear Algebra Background

Definition 5.1 (Unitary Matrices). Given a symmetric matrix $U \in$ $\mathbb{C}^{n \times n}$ we say $U$ is a unitary matrix if $U^{\top} U=U U^{\top}=I$.

Fact 5.2 (Operator Norm of Unitary Matrices). If $Q \in \mathbb{C}^{n \times n}$ is a unitary matrix, $\|Q\|_{o p}=1$.

Fact 5.3 (Sub-Multiplicativity of Operator Norms). Given matrices $A \in \mathbb{C}^{n \times d}, B \in \mathbb{C}^{d \times m},\|A B\|_{o p} \leqslant\|A\|_{o p} \cdot\|B\|_{o p}$.

Fact 5.4 (Cayley-Hamilton Theorem). Given a square matrix $A \in \mathbb{C}^{n \times n}$, the characteristic polynomial of $A$ is defined as $p_{A}(\lambda)=$ $\operatorname{det}(\lambda I-A)=\lambda^{n}+c_{n-1} \lambda^{n-1}+\ldots c_{1} \lambda+c_{0}$, where the coefficients, $c_{i}$, are scalar. Then, consider the matrix valued polynomial $p_{A}(A)=$ $A^{n}+c_{n-1} A^{n-1}+\ldots c_{1} A+c_{0}$ I. The Cayley-Hamilton theorem states $p_{A}(A)=0$.

Fact 5.5. For any matrix $A \in \mathbb{C}^{n \times n}$, there is a unitary matrix $Q$ such that $Q^{-1} A Q$ is upper triangular.

Proof. A must have some eigenvector, say $v$. Also normalize $v$ so that it is a unit vector. Let $Q_{0}$ be a matrix whose first column is $v$ and whose columns form an orthonormal basis of $\mathbb{C}^{n}$. Then $A^{\prime}=Q_{0}^{-1} A Q_{0}$ has all entries in the first column equal to 0 except possibly the first entry. Now it suffices to compute a unitary matrix in $\mathbb{C}^{(n-1) \times(n-1)}$ that transforms the $(n-1) \times(n-1)$ submatrix of $A^{\prime}$ (excluding the first row and column) into an upper triangular matrix but this can be done by induction.

We also establish the following key lemma to upper bound the operator norm of matrix polynomials, when the underlying matrix has bounded eigenvalues.

Lemma 5.6 (Operator Norm of a matrix with bounded Eigenvalues ). Let $A \in \mathbb{C}^{n \times n}$ be a matrix and assume that all eigenvalues
of A have magnitude at most 1 . Then for any integer $L$,

$$
\left\|A^{L}\right\| \leqslant n \cdot(2(1+\|A\|) L)^{n} .
$$

Proof. By Fact 5.5 and Fact 5.2, without loss of generality we can assume that $A$ is upper triangular. Then, all of its diagonal entries are eigenvalues so all of its diagonal entries have magnitude at most 1. Now we bound the magnitude of all entries of $A^{L}$. Consider the entry indexed by $i, j$. Clearly $i \leqslant j$ or the corresponding entry is 0 . Next, by definition of $\left(A^{L}\right)_{i, j}$,

$$
\begin{align*}
\left|\left(A^{L}\right)_{i j}\right| & =\sum_{\substack{i_{1}, \ldots, i_{L} \\
i \leqslant i_{1} \leqslant \ldots \leqslant i_{L-1} \leqslant j}} A_{i i_{1}} \cdots A_{i_{L-1} j} \\
& \leqslant \sum_{\substack{i_{1}, \ldots, i_{L} \\
i \leqslant i_{1} \leqslant \ldots \leqslant i_{L-1} \leqslant j}}\left|A_{i i_{1}}\right| \cdots\left|A_{i_{L-1} j}\right|  \tag{7}\\
& \leqslant\binom{ L-1+n}{n}(1+\|A\|)^{n} \\
& \leqslant(2(1+\|A\|) L)^{n} .
\end{align*}
$$

where the second inequality counts the number of paths and uses that each path can contain at most $n$ entries strictly above the diagonal.

### 5.2 Probability Background

Next, we recall the definition of $\ell_{4 \rightarrow 2}$-hypercontractivity for distributions.

Definition 5.7 (Hypercontractivity). We say a distribution $\mathcal{D}$ on $\mathbb{R}^{n}$ is (4, 2, K)-hypercontractive if for any vector $v \in \mathbb{R}^{n}$, we have $\mathbb{E}_{x \sim \mathcal{D}}\left[\langle v, x\rangle^{4}\right] \leqslant K \cdot \mathbb{E}_{x \sim \mathcal{D}}\left[\langle v, x\rangle^{2}\right]^{2}$.

We note that hypercontractivity is a very mild assumption on the concentration behavior of 4-th moments of a distribution, and several well-studied families of distributions, including all subGaussian, sub-Exponential and log-concave distributions satisfy this assumption.

Lemma 5.8 (Linear Transform of a Hypercontractive Distribution). Let $\left\{u_{i}\right\}_{i \in[t]}$ be $t$ iid samples from a distribution $\mathcal{D}$ that is (4, 2, K)-hypercontractive (see Definition 5.7). Then, for any matrices $M_{1}, \ldots, M_{t} \in \mathbb{R}^{m \times p}$, the random variable $\sum_{i \in[t]} M_{i} u_{i}+\cdots+M_{t} u_{t}$ is ( $4,2, K$ )-hypercontractive.

Proof. See full version.
Next, we obtain a weak anti-concentration bound via a Paley Zygmund like inequality:

Lemma 5.9 (Weak Anti-Concentration via Hypercontracтivity). Let $z$ be a real-valued random variable such that $\mathbb{E}[z]=$ $0, \mathbb{E}\left[z^{2}\right] \geqslant 1, \mathbb{E}\left[z^{4}\right] \leqslant K$ for some constant $K \geqslant 3$. Then for any real number $\beta$,

$$
\operatorname{Pr}[|z-\beta| \leqslant 0.1] \leqslant 1-\frac{1}{10 K} .
$$

Proof. Clearly we must have $K \geqslant 1$. Assume for the sake of contradiction that the desired inequality is false. Without loss of generality we have $\beta \geqslant 0$. First consider the case where $\beta \geqslant 0.3$.

Let $p$ be the probability that $z \leqslant 0.2$. We must have $p \leqslant 1 /(10 K)$. Furthermore, since $\mathbb{E}[z]=0$, we must have

$$
\begin{aligned}
0=\mathbb{E}[z]=p \mathbb{E}[z \mid z & \leqslant 0.2]+(1-p) \mathbb{E}[z \mid z>0.2] \\
& \geqslant p \mathbb{E}[z \mid z \leqslant 0.2]+0.2(1-p)
\end{aligned}
$$

which rearranges as

$$
\mathbb{E}[z \mid z \leqslant 0.2] \leqslant \frac{-0.2(1-p)}{p}
$$

Thus, by Jensen's inequality (since $z^{4}$ is convex), this implies that

$$
\mathbb{E}\left[z^{4}\right] \geqslant p \mathbb{E}\left[z^{4} \mid z \leqslant 0.2\right] \geqslant p\left(\frac{0.2(1-p)}{p}\right)^{4}>K
$$

which is a contradiction. Now it remains to consider the case where $\beta \leqslant 0.3$. Then let $q$ be the probability that $|z-\beta| \leqslant 0.1$. We have

$$
\begin{aligned}
\mathbb{E}\left[z^{2}\right] & =q \mathbb{E}\left[z^{2}| | z-\beta \mid \leqslant 0.1\right]+(1-q) \mathbb{E}\left[z^{2}| | z-\beta \mid>0.1\right] \\
& \leqslant q(\beta+0.1)^{2}+(1-q) \mathbb{E}\left[z^{2}| | z-\beta \mid>0.1\right]
\end{aligned}
$$

Thus, since $\mathbb{E}\left[z^{2}\right] \geqslant 1$, we must have

$$
\mathbb{E}\left[z^{2}| | z-\beta \mid>0.1\right] \geqslant \frac{0.8}{1-q}
$$

Thus, by convexity, we must have

$$
\mathbb{E}\left[z^{4}\right] \geqslant(1-q) \cdot\left(\frac{0.8}{1-q}\right)^{2}=\frac{0.64}{1-q}
$$

and combining with the fact that $\mathbb{E}\left[z^{4}\right] \leqslant K$, we deduce that $q \leqslant$ $1-\frac{1}{10 K}$ and we are done.

### 5.3 Linear Dynamical Systems Background

Next, we establish some basic definitions and identities that we utilize throughout. We begin with the definition of the Markov parameters of a LDS.

Definition 5.10 (Markov Parameters). Given a linear dynamical system, $\mathcal{L}(A, B, C, D)$, and an integer $T \geqslant 1$, the Markov Parameter matrix $G \in \mathbb{R}^{m \times T p}$ is defined as the following block matrix:

$$
G=\left[\begin{array}{lllll}
D & C B & C A B & \ldots & C A^{T-2} B
\end{array}\right] .
$$

It will be important to consider linear combinations of the observations $y_{t}$. In particular, for different integers $k$, we will consider linear combinations of the form $\hat{y}_{t+k}=y_{t+k}-\sum_{j=1}^{s} \alpha_{j} y_{t-j}$ where $\alpha_{1}, \ldots, \alpha_{s} \in \mathbb{R}^{m \times m}$ are $m \times m$ matrices. To ease notation, it will be useful to consider the following matrix polynomial.

Definition 5.11 (Matrix Polynomial). For $\alpha=\left(\alpha_{1}, \ldots, \alpha_{s}\right)$ where $\alpha_{1}, \ldots, \alpha_{s} \in \mathbb{R}^{m \times m}$ are matrices, define the matrix polynomial $F_{\alpha, k}: \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{m \times n}$

$$
F_{\alpha, k}(X)=C X^{s+k}-\alpha_{1} C X^{s-1}-\cdots-\alpha_{s} C .
$$

Also for $i=0,1, \ldots, k+s$, let

$$
F_{\alpha, k}^{(i)}(X)= \begin{cases}C X^{i} & 0 \leqslant i \leqslant k \\ C X^{i}-\sum_{j=1}^{i-k} \alpha_{j} C X^{i-k-j} & k+1 \leqslant i \leqslant k+s\end{cases}
$$

We now have the following identities.

Fact 5.12 (Algebraic Identities for LDS’s). Let $\mathcal{L}(A, B, C, D)$ be a Linear Dynamical System (see Definition 4.1). Then, for any $t \in \mathbb{N}$,

$$
y_{t}=\sum_{i=1}^{t}\left(C A^{i-1} B u_{t-i}+C A^{i-1} w_{t-i}\right)+C A^{t} x_{0}+D u_{t}+z_{t}
$$

Further, given $k \in \mathbb{N}$ and $\alpha_{1}, \alpha_{2}, \ldots, \alpha_{s} \in \mathbb{R}^{m \times m}$, let $\hat{y}_{t+k}=y_{t+k}-$ $\sum_{j=1}^{s} \alpha_{j} y_{t-j}$. Then,

$$
\begin{aligned}
\widehat{y}_{t+k}= & \left(z_{t+k}-\sum_{j=1}^{s} \alpha_{j} z_{t-j}\right)+\left(D u_{t+k}-\sum_{j=1}^{s} \alpha_{j} D u_{t-j}\right) \\
& +\left(F_{\alpha, k}(A) A^{t-s}\right) x_{0}+\sum_{i=1}^{k+s} F_{\alpha, k}^{(i-1)}(A)\left(B u_{t+k-i}+w_{t+k-i}\right) \\
& +\sum_{i=k+s+1}^{t+k} F_{\alpha, k}(A) A^{i-(k+s+1)}\left(B u_{t+k-i}+w_{t+k-i}\right)
\end{aligned}
$$

Proof. The proof is direct computation (see full version).
Next, we observe that the cross-covariance between the control input and the observation is an unbiased estimator of the Markov parameters.

Fact 5.13 (Cross-Covariance of Control and Observation). For any $t, k \in \mathbb{N}$, and any $0 \leqslant j \leqslant k$, we have

$$
\mathbb{E}\left[\hat{y}_{t+j} u_{t}^{\top}\right]=\left\{\begin{array}{l}
D \text { if } j=0  \tag{8}\\
C A^{j-1} B \text { otherwise }
\end{array}\right.
$$

Proof. We use the formula in Fact 5.12 and the independence of the $u_{i}, w_{i}, z_{i}$. When $j=0$, we immediately have

$$
\mathbb{E}\left[\hat{y}_{t} u_{t}^{\top}\right]=\mathbb{E}\left[D u_{t} u_{t}^{\top}\right]=D
$$

When $j>0$, we have

$$
\mathbb{E}\left[\hat{y}_{t+j} u_{t}^{\top}\right]=\mathbb{E}\left[F_{\alpha, k}^{(j-1)}(A) B u_{t} u_{t}^{\top}\right]=F_{\alpha, k}^{(j-1)}(A) B
$$

Next, since $j \leqslant k$, by definition we have $F_{\alpha, k}^{(j-1)}(A)=C A^{j-1}$ and we are done.

In light of the above, we make the following definition.
Definition 5.14. For an integer $j \geqslant 0$, we define the matrix $X_{j}$ as

$$
X_{j}=\left\{\begin{array}{l}
D \text { if } \mathrm{j}=0 \\
C A^{j-1} B \text { otherwise }
\end{array} .\right.
$$

Of course, we have $X_{j}=\mathbb{E}\left[\hat{y}_{t+j} u_{t}^{\top}\right]$ by the previous fact.
We also require the following straight-forward consequences of the bounded condition number assumption on the observability and controlability matrices from Section 4:

Claim 5.15. Given an integer $s$, let $O_{s}, Q_{s}$ be the observability and controlability matrices from definition 4.2 and 4.3. Then, $\sigma_{\min }\left(O_{s}\right) \leqslant$ $\sqrt{s}\|C\|$ and $\sigma_{\min }\left(Q_{s}\right) \leqslant \sqrt{s}\|B\|$.

Proof. Let $v$ be an eigenvector of $A$ and say $A v=\lambda v$. Note that $|\lambda| \leqslant 1$. Thus,

$$
\left\|O_{s} v\right\| \leqslant \sqrt{\|C v\|^{2}\left(1+\lambda^{2}+\lambda^{4}+\cdots+\lambda^{2 s-2}\right)} \leqslant \sqrt{s}\|C\|\|v\| .
$$

A similar argument works for $Q_{s}$.

Claim 5.16. Given integers $t>s$, let $\sigma_{\max }\left(O_{t}\right) / \sigma_{\min }\left(O_{s}\right) \leqslant \kappa$. Then for any integer $k>0$,

$$
\left\|A^{k}\right\|_{F} \leqslant(\sqrt{n} \kappa)^{k /(t-s)} .
$$

The same holds with $O$ replaced with $Q$.
Proof. Note that $\left\|O_{s} A^{t-s}\right\|_{F} \geqslant \sigma_{\min }\left(O_{s}\right)\left\|A^{t-s}\right\|_{F}$. On the other hand,

$$
\left\|O_{s} A^{t-s}\right\|_{F} \leqslant\left\|O_{t}\right\|_{F} \leqslant \sqrt{n}\left\|O_{t}\right\| \leqslant \sqrt{n} \kappa \sigma_{\min }\left(O_{s}\right) .
$$

Thus, $\left\|A^{t-s}\right\|_{F} \leqslant \sqrt{n} \kappa$ and this immediately implies the desired inequality.

## 6 ALGORITHM

Our algorithm follows the outline described in Section 3. We first learn a stabilizing transformation that transforms $y_{t} \rightarrow \hat{y}_{t}$ (see Algorithm 6.6). Note that in the general case, when $C$ is $m \times n$ (as opposed to $1 \times n$ as described in Section 3), the coefficients $\alpha_{1}, \ldots, \alpha_{s}$ are $m \times m$ matrices. Once we have these matrices, we empirically estimate the Markov parameters (recall Fact 5.13). Finally, once we obtain estimates for the Markov parameters, we extract the system parameters via the Ho-Kalman algorithm.

Before we describe our algorithm formally, we introduce a few definitions and notational simplifications. Throughout this section, we will use $\varepsilon$ to denote the desired accuracy and $\delta$ to be a parameter for the failure probability. We will also make the simplification that $x_{0}=0$. This is allowed because we can absorb the distribution of $x_{0}$ into the distribution for $w_{0}$ i.e. $\mathcal{D}_{w} \leftarrow A \mathcal{D}_{0}+\mathcal{D}_{w}$ which is still polynomially bounded in the system parameters. As mentioned in Remark 4.8, we do not need the $w_{i}$ to have identical distributions, just that they are independent and bounded variance.

## Algorithm 6.1 (Learning a Linear Dynamical System).

Input: $T$ observations $\left\{y_{1}, y_{2}, \ldots, y_{T}\right\}$ and the corresponding control inputs $\left\{u_{1}, u_{2}, \ldots, u_{T}\right\}$ generated from a Linear Dynamical System $\mathcal{L}(A, B, C, D)$ satisfying the assumptions in Section 4.
Input: $s \in \mathbb{N}$ such that the observability and controllability matrices satisfy the condition number bounds in Definition 4.4, accuracy parameter $0<\epsilon<1$, and failure probability parameter $0<\delta<1$.
Operation:
(1) Stabilizing the System: Run Algorithm 6.6 to obtain coefficient matrices $\alpha_{1}, \ldots, \alpha_{s}$.
(2) Estimating the Markov Parameters: Set $k=10 \mathrm{~s}$. For all $j=0,1, \ldots, k$
(a) For all $t \in[T]$ with $t>k+s$, compute $\hat{y}_{t}=$ $y_{t}-\sum_{i=1}^{s} \alpha_{i} y_{t-k-i}$.
(b) Compute $\hat{X}_{j}=\frac{1}{T-k-s} \sum_{t=k+s+1}^{T} \hat{y}_{t} u_{t-j}^{\top}$.
(3) Robust Ho-Kalman: Run Ho-Kalman on

$$
\hat{G}=\left[\hat{X}_{0}, \hat{X}_{1}, \ldots, \hat{X}_{2 s}\right]
$$

to obtain estimates $\hat{A}, \hat{B}, \hat{C}, \hat{D}$
Output: $\hat{A}, \hat{B}, \hat{C}, \hat{D}$ satisfying guarantees of Theorem 6.4

In the definition below, we condense all of the system parameters into a set $\mathcal{S}$ so that we don't need to list out the full set of parameters in future computations. Recall Definition 4.7 where $K$ is the hypercontractivity parameter for the input distribution (and will be defined as such throughout this section) and $\sigma_{w}, \sigma_{z}$ are upper bounds on the variances of the noise distributions.

Definition 6.2. Let $\mathcal{S}$ denote the set of parameters

$$
\left\{\|A\|,\|B\|,\|C\|,\|D\|, m, n, p, s, \kappa, K, 1 / \delta, \sigma_{w}, \sigma_{z}\right\}
$$

We will write $\operatorname{poly}(\mathcal{S})$ for a quantity that depends polynomially on these parameters.

Note that we will be more explicit about dependencies on the accuracy $\varepsilon$ so it is not included in the definition of $\mathcal{S}$.
Now we define a constraint system for a convex program that is at the core of our algorithm. Let $\varepsilon$ be the desired accuracy, and $k=10 s$. Throughout this section, we will treat $k$ as fixed. Choose sufficiently large polynomials $P_{0}, P_{1}, P_{2}$ in terms of the parameters in $\mathcal{S}$ such that $P_{0} \ll P_{1} \ll P_{2}$ and let $L=P_{2} \log ^{2}(1 / \varepsilon) / \varepsilon^{2}$.

Definition 6.3. We define the constraint system $C_{\alpha}$ for matrices of coefficients $\alpha_{1}, \ldots, \alpha_{s} \in \mathbb{R}^{m \times m}$ as follows. We define $\hat{y}_{t+k}=y_{t+k}-$ $\alpha_{1} y_{t-1}-\cdots-\alpha_{s} y_{t-s}$. We also enforce the following constraints
(1) $\left\|\alpha_{i}\right\|_{F} \leqslant P_{0}$ for all $i=1,2, \ldots, s$
(2) For all $i \in\left[100 \mathrm{snm}^{2} K \log L\right]$, we have

$$
\left\|\hat{y}_{i L+k}\right\| \leqslant P_{1} \log (1 / \varepsilon)
$$

Our main theorem is that Algorithm 6.1 runs with polynomial time and sample complexity and outputs estimates $\hat{A}, \hat{B}, \hat{C}, \hat{D}$ that are $\varepsilon$-close to the true system parameters up to a global similarity transformation (which is always necessary).

Theorem 6.4 (Learning a Linear Dynamical System). Given $0<\epsilon, \delta<1$, an integer $s$, and trajectory length

$$
T=\Omega\left(\operatorname{poly}(\mathcal{S}) \cdot \frac{\log ^{3}\left(\frac{1}{\epsilon}\right)}{\epsilon^{4}}\right),
$$

and the corresponding observations and inputs $\left\{y_{i}, u_{i}\right\}_{i \in[T]}$, from a linear dynamical system $\mathcal{L}(A, B, C, D)$, satisfying the assumptions in Section 4, Algorithm 6.1 outputs estimates $\hat{A}, \hat{B}, \hat{C}, \hat{D}$ such that with probability at least $1-\delta$, there exists a similarity tranform $U$ satisfying

$$
\max \left(\left\|A-U^{-1} \hat{A} U\right\|,\|C-\hat{C} U\|,\left\|B-U^{-1} \hat{B}\right\|,\|D-\hat{D}\|\right) \leqslant \epsilon
$$

Further, $\operatorname{Algorithm} 6.1$ runs in $\operatorname{poly}\left(\mathcal{S}, \frac{1}{\epsilon}\right)$ time.
The overall structure of our approach is to first estimate the Markov parameters (see Definition 5.10). We then recover the estimates of the system parameters by setting up a Generalized eigenvalue problem and using the Ho-Kalman Algorithm (see for example Theorem 5.3 in [51]).

The key techincal theorem we obtain for learning each block matrix in the Markov Parameter matrix is as follows:

Theorem 6.5 (Learning the Markov Parameters). Given $0<$ $\epsilon, \delta<1$, an integer $s$, and trajectory length

$$
T=\Omega\left(\operatorname{poly}(\mathcal{S}) \cdot \frac{\log ^{3}\left(\frac{1}{\epsilon}\right)}{\epsilon^{2}}\right)
$$

with the corresponding observations and inputs $\left\{y_{i}, u_{i}\right\}_{i \in[T]}$, from a linear dynamical system $\mathcal{L}(A, B, C, D)$, satisfying the assumptions in Section 4, Algorithm 6.6 outputs $\alpha_{1}, \alpha_{2}, \ldots, \alpha_{s}$ in time $\operatorname{poly}(\mathcal{S}, 1 / \epsilon)$ such that with probability at least $1-\delta$, for all $0 \leqslant j \leqslant k$

$$
\left\|\frac{1}{T-k-s} \sum_{t=k+s+1}^{T}\left(y_{t}-\sum_{i \in[s]} \alpha_{i} y_{t-k-i}\right) u_{t-j}^{\top}-X_{j}\right\|_{F}^{2} \leqslant \epsilon^{2}
$$

where recall $X_{j}$ is defined in Definition 5.14.
Once we have proven Theorem 6.5, we can combine it with a result from [51] for estimating the system parameters from the Markov Parameters using Ho-Kalman. See Section 8 for details. The main technical work of this paper is in proving Theorem 6.5, which we focus on in Section 7.

## Algorithm 6.6 (Stabilizing the System).

Input: $T$ observations $\left\{y_{1}, y_{2}, \ldots, y_{T}\right\}$ from a single trajectory and the corresponding inputs $\left\{u_{1}, u_{2}, \ldots, u_{T}\right\}$ generated from a Linear Dynamical System $\mathcal{L}(A, B, C, D)$ satisfying the assumptions in Section 4
Input: $s \in \mathbb{N}$ such that the observability and controllability matrices satisfy the condition number bounds in Definition 4.4, accuracy parameter $0<\epsilon<1$, and failure probability parameter $0<\delta<1$.
Operation:
(1) Let $P_{0}, P_{1}, P_{2}$ be sufficiently large polynomials in the parameters $m, n, p, s, k, \kappa$ and $K$ such that $P_{0} \ll P_{1} \ll$ $P_{2}$. Let $L=P_{2} \log ^{2}(1 / \epsilon) / \epsilon^{2}$. Let $S=\{1,2, \ldots\}$.
(2) Solve the following system in the matrix variables $\alpha_{1}, \alpha_{2}, \ldots \alpha_{s} \in \mathbb{R}^{m \times m}:$
$C_{\alpha}=\left\{\begin{array}{l}\forall j \in[s] \\ \left\|\alpha_{j}\right\|_{F}^{2} \leqslant P_{0} \\ \forall i \in\left[100 s_{n m}^{2} K \log (L)\right] \\ \left\|y_{i L+k}-\sum_{j \in[s]} \alpha_{j} \cdot y_{i L-j}\right\|_{2}^{2} \leqslant P_{1} \log (1 / \epsilon)\end{array}\right\}$
Output: Matrices $\alpha_{1}, \alpha_{2}, \ldots \alpha_{s}$ obtained above.

## 7 ANALYSIS OF ALGORITHM 6.6

In this section, we analyze Algorithm 6.6 and prove Theorem 6.5. Since we treat $k$ as fixed throughout this section, we will write $F_{\alpha}(A)$ for $F_{\alpha, k}(A)$ (recall Definition 5.11). First, we need the following basic observation.

Claim 7.1 (Uniform bounds on the control and noise). Let $P_{0}, P_{1}, P_{2}$ be sufficiently large polynomials in the parameters in $\mathcal{S}$ such that $P_{0} \ll P_{1} \ll P_{2}$. Let $L=P_{2} \log ^{2}(1 / \epsilon) / \epsilon^{2}$. With probability $1-0.1 \delta$, the following events all hold:
(1) For all $i \in\left[100 \mathrm{snm}^{2} K \log L\right]$ and integers $-10(k+s) \leqslant c \leqslant$ $10(k+s)$, we have

$$
\left\|u_{i L+c}\right\|,\left\|w_{i L+c}\right\|,\left\|z_{i L+c}\right\| \leqslant P_{0} \log (1 / \varepsilon)
$$

(2) For all $t \leqslant P_{2}^{2} \log ^{3}(1 / \varepsilon) / \varepsilon^{2}$, we have

$$
\left\|u_{t}\right\|,\left\|w_{t}\right\|,\left\|z_{t}\right\| \leqslant L P_{0}
$$

Proof. Note that from the covariance bounds on $u_{t}, w_{t}, z_{t}$ (and using Markov's inequality), we have that for a fixed index $t$ and any parameter $\alpha$,

$$
\begin{aligned}
& \operatorname{Pr}\left[\left\|u_{t}\right\| \geqslant \alpha\right] \leqslant \frac{n}{\alpha^{2}} \\
& \operatorname{Pr}\left[\left\|w_{t}\right\| \geqslant \alpha\right] \leqslant \frac{n \sigma_{w}^{2}}{\alpha^{2}} \\
& \operatorname{Pr}\left[\left\|z_{t}\right\| \geqslant \alpha\right] \leqslant \frac{n \sigma_{z}^{2}}{\alpha^{2}}
\end{aligned}
$$

Now to prove the first statement, note that we only need to union bound over

$$
O\left(100(k+s) s n m^{2} K \log L\right)=\operatorname{poly}(\mathcal{S}) \log (1 / \varepsilon)
$$

variables so as long as we choose $P_{0}$ sufficiently large we get that the statement holds with at least $1-0.05 \delta$ probability. The proof of the second statement is similar except we union bound over more variables.

We begin by establishing the feasibility of the constraint system, $\mathcal{C}_{\alpha}$, as defined in Algorithm 6.6:

Lemma 7.2 (Feasibility of the Constraint System). Assume we are given $0<\epsilon, \delta<1$, and $T=\Omega\left(\operatorname{poly}(\mathcal{S}) \cdot \frac{\log ^{3}\left(\frac{1}{\epsilon}\right)}{\epsilon^{2}}\right)$ observations $\left\{y_{i}, u_{i}\right\}_{i \in[T]}$ from a single trajectory of a linear dynamical system $\mathcal{L}(A, B, C, D)$, satisfying the assumptions in Section 4. Then, as long as the events in Claim 7.1 hold, the constraint system $C_{\alpha}$ is feasible.

Proof. See full version.
Next, we argue that any $\alpha$ that is feasible for $\mathcal{C}_{\alpha}$ must actually be useful for stabilizing the system. To do this, we introduce the following potential.

Definition 7.3 (Anti-Concentration Potential). For an integer $l$ and coefficients $\alpha=\left(\alpha_{1}, \ldots, \alpha_{s}\right)$, define the function

$$
G_{\alpha, l}=\sum_{i=0}^{l}\left\|F_{\alpha}(A) A^{i} B\right\|_{F}^{2}
$$

$G_{\alpha, L}$ is a potential measuring the variance of $\hat{y}$. We will show that with high probability over the randomness of the $u_{t}, w_{t}, z_{t}$, any $\alpha$ that is feasible for $\mathcal{C}_{\alpha}$ must have $G_{\alpha, l}$ be small.

First, we express $G_{\alpha, l}$ as the variance of a random variable that naturally arises when computing $\hat{y}_{t}$ (using the formula in Fact 5.12).

Lemma 7.4 (Potential captures Variance). Given $l<t \in \mathbb{N}$, consider the random variable

$$
\gamma_{t, l}=\sum_{i=0}^{l} F_{\alpha}(A) A^{i} B u_{t-i}
$$

Then we have

$$
\operatorname{Tr}\left[\mathbb{E}\left[\gamma_{t, l} \gamma_{t, l}^{\top}\right]\right]=G_{\alpha, l}
$$

Proof. Since the $u_{i}$ are independent $\mathbb{E}\left[u_{i} u_{j}\right]=0$ for $i \neq j$ and $\mathbb{E}\left[u_{i} u_{i}^{\top}\right]=I$ for all $i$, we have

$$
\begin{aligned}
\mathbb{E}\left[\gamma_{t, l} \gamma_{t, l}^{\top}\right] & =\mathbb{E}\left[\left(\sum_{i=0}^{l} F_{\alpha}(A) A^{i} B u_{t-i}\right)\left(\sum_{i=0}^{l} F_{\alpha}(A) A^{i} B u_{t-i}\right)^{\top}\right] \\
& =\mathbb{E}\left[\sum_{i=0}^{l}\left(F_{\alpha}(A) A^{i} B u_{t-i}\right)\left(F_{\alpha}(A) A^{i} B u_{t-i}\right)^{\top}\right] \\
& =\sum_{i=0}^{l}\left(( F _ { \alpha } ( A ) A ^ { i } B ) \left(\left(F_{\alpha}(A) A^{i} B\right)^{\top} .\right.\right.
\end{aligned}
$$

Now taking the trace of both sides, and using the linearity of trace,

$$
\begin{aligned}
\operatorname{Tr}\left[\mathbb{E}\left[\gamma_{t, l} \gamma_{t, l}^{\top}\right]\right] & =\sum_{i=0}^{l} \operatorname{Tr}\left[\left(\left(F_{\alpha}(A) A^{i} B\right)\left(\left(F_{\alpha}(A) A^{i} B\right)^{\top}\right]\right.\right. \\
& =\sum_{i=0}^{l}\left\|F_{\alpha}(A) A^{i} B\right\|_{F}^{2}=G_{\alpha, l}
\end{aligned}
$$

as desired.
Next, we show that if the potential $G_{\alpha, L}$ is large, with high probability, there must be a violated constraint in $C_{\alpha}$.

Lemma 7.5. Let $\alpha=\left(\alpha_{1}, \ldots, \alpha_{S}\right)$ be a fixed sequence with $\left\|\alpha_{i}\right\|_{F} \leqslant$ $P_{0}$ and $G_{\alpha, L}>m\left(100 P_{1} \log (1 / \varepsilon)\right)^{2}$. Now, consider a trajectory of length $T$, and the corresponding observations and inputs, $\left\{y_{i}, u_{i}\right\}_{i \in[T]}$, from the $\operatorname{LDS} \mathcal{L}(A, B, C, D)$ after fixing this sequence and let $\zeta$ be the event that Lemma 7.1 holds. Conditioned on $\zeta$, with $1-(1 / L)^{10 \text { snm }^{2}}$ probability, there exists some integer $1 \leqslant i \leqslant 10^{2}$ snm $^{2} K \log L$ such that

$$
\left\|\hat{y}_{i L+k}\right\| \geqslant 2 P_{1} \log (1 / \varepsilon) .
$$

Proof. See full version.
Next, we relate the potential function $G_{\alpha, L}$ to the variance of our actual estimators for the Markov parameters. First, we will need to define a similar-looking potential and relate it to $G_{\alpha, L}$.

Definition 7.6. For an integer $l$ and matrix-valued coefficients $\alpha=\left(\alpha_{1}, \ldots, \alpha_{s}\right)$, define the function

$$
H_{\alpha, l}=\sum_{i=0}^{l}\left\|F_{\alpha}(A) A^{i}\right\|_{F}^{2} .
$$

Intuitively, this potential simply does not include the matrix $B$ and thus is polynomially related to $G_{\alpha, l}$, since the condition number of the controllability matrix is bounded. We make this precise as follows:

Lemma 7.7 (Potential without B). For a fixed sequences of matrices $\alpha=\left(\alpha_{1}, \alpha_{2}, \ldots, \alpha_{s}\right)$, we have

$$
H_{\alpha, l} \leqslant \kappa^{2} s G_{\alpha, l+s} .
$$

Proof. See full version.

Next, for any fixed sequence of $\alpha$ 's we show that the variance of our estimator can be bounded in terms of the norm of the $\alpha_{i}$ 's and the two potentials we defined above. In particular,

Lemma 7.8 (Potential to Variance Bound). For any fixed $\alpha_{1}, \ldots, \alpha_{s} \in \mathbb{R}^{m \times m}$ and any $0 \leqslant j \leqslant k$, we have

$$
\begin{array}{r}
\mathbb{E}\left[\left\|\frac{1}{L} \sum_{t=s+1}^{s+L} \hat{y}_{t+k} u_{t+k-j}^{\top}-X_{j}\right\|_{F}^{2}\right] \\
\leqslant \frac{P_{1}}{L}\left(1+\left\|\alpha_{1}\right\|_{F}^{2}+\cdots+\left\|\alpha_{s}\right\|_{F}^{2}+G_{\alpha, L}+H_{\alpha, L}\right) .
\end{array}
$$

where recall $X_{j}$ is defined as in Definition 5.14 and the expectation is over the draws of $u_{t}, w_{t}, z_{t}$.

Proof. See full version.
Unfortunately, we cannot use Lemma 7.8 directly because the choice of $\alpha=\left(\alpha_{1}, \ldots, \alpha_{s}\right)$ that we compute using the program $C_{\alpha}$ already depends on the realizations of the $u_{t}, w_{t}, z_{t}$ meaning that there is no more fresh randomness. In order to use the randomness over $u_{t}, w_{t}$ and $z_{t}$ to bound the variance of our estimator, we need decouple the $\alpha$ 's out of the variance expression.
To circumvent this, we will derive a symbolic inequality from Lemma 7.8 that holds simultaneously for all choices of $\alpha$ and thus can be applied even if $\alpha$ depends on the realizations of $u_{t}, w_{t}, z_{t}$.

Note that $G_{\alpha, l}$ and $H_{\alpha, l}$ are both quadratic expressions in the $\alpha_{i}$. It will be useful to extract out the matrix of coefficients, which we do in the following definition.

Definition 7.9 (Coefficient Matrix of a quadratic polynomial). Define $v_{\alpha}=\left(1, \alpha_{1}, \ldots, \alpha_{s}\right)$ where we view the $\alpha_{i}$ as formal variables and flatten each of the matrices $\alpha_{i}$ into a vector and concatenate them so that $v_{\alpha}$ has length $s m^{2}+1$. Let $G_{L}$ (respectively $H_{L}$ ) be the unique symmetric $\left(s m^{2}+1\right) \times\left(s m^{2}+1\right)$ matrix such that

$$
v_{\alpha}^{T} G_{L} v_{\alpha}=G_{\alpha, L} .
$$

Note that matrices defined above are unique because we force them to be symmetric so the coefficients of the monomials in the $\alpha_{i}$ uniquely determine the entries of the matrices. Also, the entries of $G_{l}, H_{l}$ are purely functions of the system parameters $A, B, C, D$. Now we can prove a symbolic version of Lemma 7.8. Intuitively, this symbolic version is a way to decouple the vector valued random variables $u_{t}, w_{t}$ and $z_{t}$ from the matrix random variables $\alpha$, and then use properties of the input distribution on the expressions that are independent of $\alpha$.

Corollary 7.10 (Symbolic Matrix Inequality). Consider the vector of formal variables $v_{\alpha}=\left(1, \alpha_{1}, \ldots, \alpha_{s}\right)$. For an integer $j$ with $0 \leqslant j \leqslant k$, define $M_{j}$ to be the unique symmetric matrix such that

$$
\left\|\frac{1}{L} \sum_{t=s+1}^{s+L} \hat{y}_{t+k} u_{t+k-j}^{\top}-X_{j}\right\|_{F}^{2}=v_{\alpha}^{\top} M_{j} v_{\alpha},
$$

where recall $X_{j}$ is defined in Definition 5.14. Then we have

$$
\mathbb{E}\left[M_{j}\right] \leq \frac{P_{1}}{L}\left(I+G_{L}+H_{L}\right),
$$

where the expectation is over the randomness of the realizations of the $u_{i}, w_{i}, z_{i}$.

Proof. We simply use the fact that Lemma 7.8 holds over all choices of $\alpha=\left(\alpha_{1}, \ldots, \alpha_{s}\right)$. See full version for details.

Now we can complete the analysis of our algorithm for learning the Markov parameters.

Proof of Theorem 6.5. With probability at least $1-0.1 \delta$, the event in Claim 7.1 holds and we condition on it. Now we solve the program in Definition 6.3. By Lemma 7.2, it is feasible. Let $\tilde{\alpha}=\left(\tilde{\alpha}_{1}, \ldots, \tilde{\alpha}_{s}\right)$ be a feasible solution.

Now construct a $\gamma$-net, denoted by $\mathcal{T}$, over the matrices $\left(\alpha_{1}, \ldots, \alpha_{s}\right)$ in Frobenius norm with $\gamma=1 / L^{4 n}$. Then, observe for any $\left(\alpha_{1}, \ldots, \alpha_{s}\right)$ with $\left\|\alpha_{i}\right\| \leqslant P_{0}$, there exists $\left(\alpha_{1}^{\prime}, \ldots, \alpha_{s}^{\prime}\right) \in \mathcal{T}$ such that

$$
\sqrt{\left\|\alpha_{1}-\alpha_{1}^{\prime}\right\|_{F}^{2}+\cdots+\left\|\alpha_{s}-\alpha_{S}^{\prime}\right\|_{F}^{2}} \leqslant \frac{1}{L^{4 n}} .
$$

It is clear that such a net exists with $|\mathcal{T}| \leqslant L^{4 n s m^{2}}$. Now, we can union bound over all the events such that with probability at least $1-0.2 \delta$, simultaneously, for all $\alpha$, for all $i \in\left[100 \mathrm{snm}^{2} K \log (L)\right]$, it follows from Lemma 7.5 that $\left\|\hat{y}_{i L+k}\right\|^{2} \leqslant 2 P_{1} \log (1 / \epsilon)$.

Next, we show that the solution $\hat{\alpha}$ we obtain must satisfy

$$
\begin{equation*}
G_{\hat{\alpha}, L} \leqslant m\left(200 P_{1} \log (1 / \varepsilon)\right)^{2} \tag{9}
\end{equation*}
$$

To see this, assume for the sake of contradiction that the above doesn't hold. Now round our solution $\hat{\alpha}$ to the nearest $\alpha^{\prime}=\left(\alpha_{1}^{\prime}, \ldots, \alpha_{s}^{\prime}\right)$ in the net. By Lemma 5.6, as long as $P_{2}$ (and recall $\left.L=P_{2} \log ^{2}(1 / \varepsilon) / \varepsilon^{2}\right)$ is chosen sufficiently large, we have

$$
G_{\alpha^{\prime}, L} \geqslant m\left(100 P_{1} \log (1 / \varepsilon)\right)^{2}
$$

However, Lemma 7.5 implies that there is some integer $1 \leqslant i \leqslant$ $10^{2}$ snm $^{2} K \log L$ such that

$$
\left\|y_{i L+k}-\alpha_{1}^{\prime} y_{i L-1}-\cdots-\alpha_{s}^{\prime} y_{i L-s}\right\| \geqslant 2 P_{1} \log (1 / \varepsilon)
$$

However, using the assumptions in Lemma 7.1, the formula for $y_{t}$ in Fact 5.12 and the bounds in Lemma 5.6, and the properties of the net, the above implies that

$$
\left\|y_{i L+k}-\hat{\alpha}_{1} y_{i L-1}-\cdots-\hat{\alpha}_{s} y_{i L-s}\right\| \geqslant P_{1} \log (1 / \varepsilon)
$$

which contradicts the fact that $\left(\hat{\alpha}_{1}, \ldots, \hat{\alpha}_{S}\right)$ is a feasible solution. Thus, we actually must have (9). Now by Lemma 7.7, we have $H_{\alpha, L-s} \leqslant \kappa^{2} \operatorname{sm}\left(200 P_{1} \log (1 / \varepsilon)\right)^{2}$. Now let $L^{\prime}=L-s$ and set

$$
\hat{X}_{j}=\frac{1}{L^{\prime}} \sum_{t=s+1}^{s+L^{\prime}} \hat{y}_{t+k} u_{t+k-j}^{\top}
$$

for all $0 \leqslant j \leqslant k$. Let $M_{j}$ be defined as in Corollary 7.10 (with $L$ replaced by $L^{\prime}$ ). Let

$$
\widetilde{M_{j}}=\frac{P_{1}}{L^{\prime}}\left(I+G_{L^{\prime}}+H_{L^{\prime}}\right) .
$$

Note that $\widetilde{M}_{j}$ is clearly PSD. Also $M_{j}$ is always PSD regardless of the realizations of the $u_{i}, w_{i}, z_{i}$, so by Markov's inequality and Corollary 7.10, we have that with $1-0.1 \delta / k$ probability

$$
\begin{aligned}
\operatorname{Tr}\left({\widetilde{M_{j}}}^{-1 / 2} M_{j}{\widetilde{M_{j}}}^{-1 / 2}\right) & \leqslant \frac{10 k}{\delta} \mathbb{E}\left[\operatorname{Tr}\left({\widetilde{M_{j}}}^{-1 / 2} M_{j}{\widetilde{M_{j}}}^{-1 / 2}\right)\right] \\
& \leqslant \frac{10 k}{\delta}\left(s m^{2}+1\right) .
\end{aligned}
$$

This means that with $1-0.1 \delta / k$ probability,

$$
\widetilde{M}_{j}^{-1 / 2} M_{j} \widetilde{M}_{j}^{-1 / 2} \leq \frac{10 k\left(s m^{2}+1\right)}{\delta}
$$

which then implies

$$
M_{j} \leq \frac{10 k\left(s m^{2}+1\right)}{\delta} \widetilde{M_{j}} .
$$

Assuming that this happens, we have that

$$
\begin{align*}
& \left\|\hat{X}_{j}-X_{j}\right\|_{F}^{2}=v_{\alpha}^{\top} M_{j} v_{\alpha} \\
& \leqslant \frac{10 k\left(s m^{2}+1\right)}{\delta} \frac{P_{1}}{L^{\prime}}\left(1+\left\|\alpha_{1}\right\|_{F}^{2}+\cdots+\left\|\alpha_{s}\right\|_{F}^{2}+G_{\alpha, L^{\prime}}+H_{\alpha, L^{\prime}}\right) \\
& \leqslant \varepsilon^{2} \tag{10}
\end{align*}
$$

where the last inequality uses that $L=P_{2}(\log (1 / \varepsilon))^{2} / \varepsilon^{2}$ and $P_{2}$ is chosen sufficiently large. Finally, union bounding the above over all choices of $j=0,1, \ldots k$ completes the proof for the guarantees of the estimator.

Finally, note that our algorithm runs in polynomial time in all the parameters because the convex program $C_{\alpha}$ admits an efficient separation oracle. To see this, note that there are only polynomially many constraints in $C_{\alpha}$ and each one is either a linear constraint or an ellipsoid constraint both of which admit an efficient separation oracle.

## 8 FROM MARKOV PARAMETERS TO SYSTEM PARAMETERS

Note that Theorem 6.5 guarantees that we can get good estimates for the markov parameters. To complete the proof of our full learning result, Theorem 6.4, we apply the Ho-Kalman algorithm black box to extract the system matrices $\{A, B, C, D\}$ where $\{A, B, C\}$ are recovered up to a similarity transformation. Recall that linear dynamical systems are specified only up to similarity transformation (see [51] for a discussion on this point).

The following lemma from [51] establishes error guarantees for the Ho-Kalman algorithm given operator norm bounds on estimating $G$.

Lemma 8.1 ([51]). For observability and controllability matrices that are rank $n$, the Ho-Kalman algorithm applied to $\hat{G}$ produces estimates $\hat{A}, \hat{B}$, and $\hat{C}$ such that there exists similarity transform $T \in$ $\mathbb{R}^{n \times n}$ such that

$$
\max \left\{\|C-\hat{C} T\|_{F},\left\|B-T^{-1} \hat{B}\right\|_{F}\right\} \leqslant 5 \sqrt{n\|G-\hat{G}\|}
$$

and

$$
\left\|A-T^{-1} \hat{A} T\right\|_{F} \leqslant \frac{\sqrt{n\|G-\hat{G}\|\|H\|}}{\sigma_{\text {min }}^{3 / 2}\left(H^{-}\right)}
$$

and

$$
\|D-\hat{D}\|_{F} \leqslant \sqrt{n}\|G-\hat{G}\|
$$

where in the above

$$
G=\left[D, C B, C A B, \ldots, C A^{2 s-1} B\right]
$$

A straightforward application of this lemma allows us to complete the proof of Theorem 6.4.

Proof of Theorem 6.4. By Theorem 6.5, in Algorithm 6.1, the input $\hat{G}$ to the Ho-Kalman algorithm satisfies

$$
\|G-\hat{G}\|_{F} \leqslant \sqrt{2 s+1} \varepsilon
$$

with probability at least $1-\delta$. Now we apply Lemma 8.1. The two things we need to do are upper bound $\|H\|$ and lower bound $\sigma_{\min }\left(H^{-}\right)$. We have an upper bound on $\|H\| \leqslant \sigma_{\max }\left(O_{s}\right) \sigma_{\max }\left(Q_{s}\right) \leqslant$ $\kappa^{2} s\|B\|\|C\|$ where we use Claim 5.15. We also have $\sigma_{\min }\left(H^{-}\right) \geqslant$ $\sigma_{\min }\left(O_{s}\right) \sigma_{\min }\left(Q_{s}\right) \geqslant\|B\|\|C\| \geqslant 1$. Therefore we conclude that there is a similarity transform $T$ such that

$$
\begin{array}{r}
\max \left\{\left\|A-T^{-1} \hat{A} T\right\|_{F},\|B-\hat{B} T\|_{F},\left\|C-T^{-1} \hat{C}\right\|_{F}\right\} \\
\leqslant \operatorname{poly}(n, \kappa, s,\|B\|,\|C\|) \sqrt{\epsilon} .
\end{array}
$$

Redefining $\varepsilon$ appropriately immediately gives the desired result.

## 9 SAMPLE COMPLEXITY LOWER BOUND FOR ILL-CONDITIONAL LDS

In this section, we prove a lower bound, that when the observability or controllability matrix of an LDS is close to singular, then it is information-theoretically impossible to learn. We consider the case where the distributions $\mathcal{D}_{u}=\mathcal{D}_{0}=N(0, I)$ and $\mathcal{D}_{w}=N\left(0, \Sigma_{w}\right)$ where $\Sigma_{w}$ will be set later. For simplicity, we also set $x_{0}=0$ and also $D=0$.

Definition 9.1. We say that an $\operatorname{LDS} \mathcal{L}(A, B, C, D)$ is $(\delta, v)$-unobservable if $v$ is a unit vector such that for all integers $s \geqslant 0$,

$$
\left\|C A^{s} v\right\| \leqslant \delta
$$

Note that the above condition depends only on $A, C$ so we will sometimes talk about a pair of matrices $A, C$ being $(\delta, v)$-unobservable.
Definition 9.2. We say that an $\operatorname{LDS} \mathcal{L}(A, B, C, D)$ is $(\delta, v)$-uncontrollable if $v$ is a unit vector such that for all integers $s \geqslant 0$,

$$
\left\|\left(A^{s} B\right)^{\top} v\right\| \leqslant \delta
$$

Note that the above condition depends only on $A, B$ so we will sometimes talk about a pair of matrices $A, B$ being $(\delta, v)$-unobservable.

For our formal lower bound, we need a minor assumption that $A$ is generic, in particular, we need that it is not too close to a multiple of the identity plus a rank-1 perturbation. Essentially all matrices satisfy this assumption as long as $n \geqslant 3$.
Definition 9.3. We say a matrix $A \in \mathbb{R}^{n \times n}$ and vector $v$ are $c$ generic if $\|A\| \geqslant c$ and there are unit vectors $u, w$ such that

$$
\begin{aligned}
\langle u, v\rangle & =0 \\
\langle u, w\rangle & =0
\end{aligned}
$$

$$
\langle u, A w\rangle \geqslant c\|A\| .
$$

Theorem 9.4. Let $A, C$ be matrices that are ( $\delta, v$ )-unobservable for some $0<\delta<0.1$ and unit vector $v$. Assume that ( $A, v$ ) are cgeneric for some constant $c$. Then any algorithm that is given an LDS $\mathcal{L}(A, B, C, D)$ that uses at most

$$
o\left(\frac{1}{\sqrt{\delta(m+p)}}\right)
$$

samples has probability at least 0.4 of outputting $\hat{A}, \hat{B}, \hat{C}, \hat{D}$ such that there is no invertible matrix $U$ with

$$
\left\|A-U^{-1} \hat{A} U\right\|_{F},\left\|B-U^{-1} \hat{B}\right\|_{F},\|C-\hat{C} U\|_{F} \leqslant 0.1 c^{2} .
$$

Similarly, the same holds if $A, B$ are matrices that are $(\delta, v)$ uncontrollable.

Proof. See full version.

## ACKNOWLEDGMENTS

Ainesh Bakshi was supported by Ankur Moitra's ONR grant. Allen Liu was supported by an NSF Graduate Research Fellowship and a Fannie and John Hertz Foundation Fellowship. Ankur Moitra was supported by an ONR grant and a David and Lucile Packard Fellowship.

## REFERENCES

[1] Karl Johan Åström and Peter Eykhoff. 1971. System identification-a survey. Automatica 7, 2 (1971), 123-162.
[2] Michael Athans. 1974. The importance of Kalman filtering methods for economic systems. In Annals of Economic and Social Measurement, Volume 3, number 1. NBER, 49-64.
[3] Ainesh Bakshi and Pravesh Kothari. 2020. List-Decodable Subspace Recovery via Sum-of-Squares. arXiv preprint arXiv:2002.05139 (2020).
[4] Ainesh Bakshi and Pravesh Kothari. 2020. Outlier-Robust Clustering of NonSpherical Mixtures. arXiv preprint arXiv:2005.02970 (2020).
[5] Ainesh Bakshi and Adarsh Prasad. 2021. Robust linear regression: Optimal rates in polynomial time. In Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing. 102-115.
[6] Guy Bresler. 2015. Efficiently learning Ising models on arbitrary graphs. In Proceedings of the forty-seventh annual ACM symposium on Theory of computing. 771-782.
[7] Emmanuel J Candès and Carlos Fernandez-Granda. 2014. Towards a mathematical theory of super-resolution. Communications on pure and applied Mathematics 67, 6 (2014), 906-956.
[8] Sitan Chen, Frederic Koehler, Ankur Moitra, and Morris Yau. 2022. Kalman filtering with adversarial corruptions. In Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing. 832-845.
[9] Sitan Chen, Frederic Koehler, Ankur Moitra, and Morris Yau. 2022. Online and distribution-free robustness: Regression and contextual bandits with huber contamination. In 2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS). IEEE, 684-695.
[10] Yeshwanth Cherapanamjeri, Efe Aras, Nilesh Tripuraneni, Michael I Jordan, Nicolas Flammarion, and Peter L Bartlett. 2020. Optimal robust linear regression in nearly linear time. arXiv preprint arXiv:2007.08137 (2020).
[11] Yeshwanth Cherapanamjeri, Samuel B Hopkins, Tarun Kathuria, Prasad Raghavendra, and Nilesh Tripuraneni. 2020. Algorithms for heavy-tailed statistics: Regression, covariance estimation, and beyond. In Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing. 601-609.
[12] Yeshwanth Cherapanamjeri, Sidhanth Mohanty, and Morris Yau. 2020. List Decodable Mean Estimation in Nearly Linear Time. arXiv preprint arXiv:2005.09796 (2020).
[13] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. 2020. On the sample complexity of the linear quadratic regulator. Foundations of Computational Mathematics 20, 4 (2020), 633-679.
[14] Ilias Diakonikolas, Samuel B Hopkins, Daniel Kane, and Sushrut Karmalkar. 2020. Robustly Learning any Clusterable Mixture of Gaussians. arXiv preprint arXiv:2005.06417 (2020).
[15] Feng Ding. 2013. Two-stage least squares based iterative estimation algorithm for CARARMA system modeling. Applied Mathematical Modelling 37, 7 (2013), 4798-4808. https://doi.org/10.1016/j.apm.2012.10.014
[16] Boualem Djehiche and Othmane Mazhar. 2022. Efficient learning of hidden state LTI state space models of unknown order. https://doi.org/10.48550/ARXIV.2202. 01625
[17] David L Donoho. 1992. Superresolution via sparsity constraints. SIAM journal on mathematical analysis 23, 5 (1992), 1309-1331.
[18] Gianfranco Doretto, Alessandro Chiuso, Ying Nian Wu, and Stefano Soatto. 2003. Dynamic textures. International Journal of Computer Vision 51, 2 (2003), 91-109.
[19] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. 2017. Finite Time Identification in Unstable Linear Systems. CoRR abs/1710.01852 (2017). arXiv:1710.01852 http://arxiv.org/abs/1710.01852
[20] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. 2018. Finite time identification in unstable linear systems. Automatica 96 (2018), 342-353.
[21] Salar Fattahi. 2020. Learning Partially Observed Linear Dynamical Systems from Logarithmic Number of Samples. CoRR abs/2010.04015 (2020). arXiv:2010.04015 https://arxiv.org/abs/2010.04015
[22] Erwin Enrique Fetzer and PM Anderson. 1975. Observability in the state estimation of power systems. IEEE transactions on power Apparatus and Systems 94, 6 (1975), 1981-1988.
[23] Miguel Galrinho. 2016. Least Squares Methods for System Identification of Structured Models.
[24] Zoubin Ghahramani and Geoffrey E Hinton. 1996. Parameter estimation for linear dynamical systems. (1996).
[25] Udaya Ghai, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. 2020. NoRegret Prediction in Marginally Stable Systems. CoRR abs/2002.02064 (2020). arXiv:2002.02064 https://arxiv.org/abs/2002.02064
[26] Mohinder S Grewal and Angus P Andrews. 2010. Applications of Kalman filtering in aerospace 1960 to the present [historical perspectives]. IEEE Control Systems Magazine 30, 3 (2010), 69-78.
[27] Moritz Hardt, Tengyu Ma, and Benjamin Recht. 2018. Gradient Descent Learns Linear Dynamical Systems. Journal of Machine Learning Research 19 (2018), 1-44.
[28] Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. 2018. Spectral filtering for general linear dynamical systems. Advances in Neural Information Processing Systems 31 (2018).
[29] Elad Hazan, Karan Singh, and Cyril Zhang. 2017. Learning Linear Dynamical Systems via Spectral Filtering. CoRR abs/1711.00946 (2017). arXiv:1711.00946 http://arxiv.org/abs/1711.00946
[30] BL HO and Rudolf E Kálmán. 1966. Effective construction of linear state-variable models from input/output functions. at-Automatisierungstechnik 14, 1-12 (1966), 545-548.
[31] Samuel B Hopkins. 2018. Sub-Gaussian Mean Estimation in Polynomial Time. arXiv preprint arXiv:1809.07425 (2018).
[32] Samuel B Hopkins and Jerry Li. 2018. Mixture models, robustness, and sum of squares proofs. In Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing. 1021-1034.
[33] Misha Ivkov and Pravesh K Kothari. 2022. List-decodable covariance estimation. In Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing. 1276-1283.
[34] Arun Jambulapati, Jerry Li, Tselil Schramm, and Kevin Tian. 2021. Robust regression revisited: Acceleration and improved estimation rates. Advances in Neural Information Processing Systems 34 (2021), 4475-4488.
[35] Rudolph Emil Kalman. 1960. A new approach to linear filtering and prediction problems. (1960).
[36] Rudolf E Kalman. 1960. On the general theory of control systems. In Proceedings First International Conference on Automatic Control, Moscow, USSR. 481-492.
[37] Sushrut Karmalkar, Adam Klivans, and Pravesh Kothari. 2019. List-decodable linear regression. In Advances in Neural Information Processing Systems. 74237432.
[38] Jonathan Kelner, Frederic Koehler, Raghu Meka, and Ankur Moitra. 2020. Learning some popular gaussian graphical models without condition number bounds. Advances in Neural Information Processing Systems 33 (2020), 10986-10998.
[39] Adam Klivans, Pravesh K Kothari, and Raghu Meka. 2018. Efficient Algorithms for Outlier-Robust Regression. arXiv preprint arXiv:1803.03241 (2018).
[40] Pravesh K Kothari, Jacob Steinhardt, and David Steurer. 2018. Robust moment estimation and improved clustering via sum of squares. In Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing. ACM, 1035-1046.
[41] Pravesh K Kothari and David Steurer. 2017. Outlier-robust moment-estimation via sum-of-squares. arXiv preprint arXiv:1711.11581 (2017).
[42] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. 2020. Regret Bound of Adaptive Control in Linear Quadratic Gaussian (LQG) Systems. CoRR abs/2003.05999 (2020). arXiv:2003.05999 https://arxiv.org/abs/ 2003.05999
[43] Bruce Lee and Andrew Lamperski. 2019. Non-asymptotic Closed-Loop System Identification using Autoregressive Processes and Hankel Model Reduction. https://doi.org/10.48550/ARXIV.1909.02192
[44] Holden Lee. 2020. Improved rates for identification of partially observed linear dynamical systems. CoRR abs/2011.10006 (2020). arXiv:2011.10006 https://arxiv. org/abs/2011.10006
[45] Qiao Li, Roger G Mark, and Gari D Clifford. 2007. Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter. Physiological measurement 29, 1 (2007), 15.
[46] Lennart Ljung. 1998. System identification. In Signal analysis and prediction. Springer, 163-173.
[47] Gábor Lugosi and Shahar Mendelson. 2019. Sub-Gaussian estimators of the mean of a random vector. The annals of statistics 47, 2 (2019), 783-794.
[48] Bertrand Mesot and David Barber. 2007. Switching linear dynamical systems for noise robust speech recognition. IEEE Transactions on Audio, Speech, and Language Processing 15, 6 (2007), 1850-1858.
[49] Ankur Moitra. 2015. Super-resolution, extremal functions and the condition number of Vandermonde matrices. In Proceedings of the forty-seventh annual

ACM symposium on Theory of computing. 821-830.
[50] PC Müller and HI Weber. 1972. Analysis and optimization of certain qualities of controllability and observability for linear dynamical systems. Automatica 8,3 (1972), 237-246.
[51] Samet Oymak and Necmiye Ozay. 2019. Non-asymptotic identification of lti systems from a single trajectory. In 2019 American control conference (ACC). IEEE, 5655-5661.
[52] Ankit Pensia, Varun Jog, and Po-Ling Loh. 2020. Robust regression with covariate filtering: Heavy tails and adversarial contamination. arXiv preprint arXiv:2009.12976 (2020).
[53] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. 2020. Robust estimation via robust gradient estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82, 3 (2020), 601-627. https://doi.org/10.1111/rssb. 12364 arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssb. 12364
[54] Prasad Raghavendra and Morris Yau. 2020. List decodable learning via sum of squares. In Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms. SIAM, 161-180.
[55] Prasad Raghavendra and Morris Yau. 2020. List decodable subspace recovery. In Conference on Learning Theory. PMLR, 3206-3226.
[56] Paria Rashidinejad, Jiantao Jiao, and Stuart Russell. 2020. SLIP: Learning to predict in unknown dynamical systems with long-term memory. In Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 5716-5728. https://proceedings. neurips.cc/paper/2020/file/3e91970f771a2c473ae36b60d1146068-Paper.pdf
[57] Derek Rowell. 2002. State-space representation of LTI systems. URL: http://web. mit. edu/2.14/www/Handouts/StateSpace. pdf (2002), 1-18.
[58] Tuhin Sarkar and Alexander Rakhlin. 2019. Near optimal finite time identification of arbitrary linear dynamical systems. In International Conference on Machine Learning. PMLR, 5610-5618.
[59] Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. 2019. Nonparametric finite time LTI system identification. arXiv preprint arXiv:1902.01848 (2019).
[60] Tuhin Sarkar, Alexander Rakhlin, and Munther A. Dahleh. 2022. Finite Time LTI System Identification. 7. Mach. Learn. Res. 22, 1, Article 26 (jul 2022), 61 pages.
[61] Steven J Schiff. 2009. Kalman meets neuron: the emerging intersection of control theory with neuroscience. In 2009 annual international conference of the IEEE engineering in medicine and biology society. IEEE, 3318-3321.
[62] Parikshit Shah, Badri Narayan Bhaskar, Gongguo Tang, and Benjamin Recht. 2012. Linear system identification via atomic norm regularization. In 2012 IEEE 51st IEEE conference on decision and control (CDC). IEEE, 6265-6270.
[63] Max Simchowitz, Ross Boczar, and Benjamin Recht. 2019. Learning linear dynamical systems with semi-parametric least squares. In Conference on Learning Theory. PMLR, 2714-2802.
[64] Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. 2018. Learning without mixing: Towards a sharp analysis of linear system identification. In Conference On Learning Theory. PMLR, 439-473.
[65] Max Simchowitz, Horia Mania, Stephen Tu, Michael I. Jordan, and Benjamin Recht. 2018. Learning Without Mixing: Towards A Sharp Analysis of Linear System Identification. CoRR abs/1802.08334 (2018). arXiv:1802.08334 http: //arxiv.org/abs/1802.08334
[66] W. Spinelli, L. Piroddi, and M. Lovera. 2005. On the role of prefiltering in nonlinear system identification. IEEE Trans. Automat. Control 50, 10 (2005), 1597-1602. https://doi.org/10.1109/TAC.2005.856655
[67] Yue Sun, Samet Oymak, and Maryam Fazel. 2022. System Identification via Nuclear Norm Regularization. https://doi.org/10.48550/ARXIV.2203.16673
[68] Anastasios Tsiamis and George J Pappas. 2019. Finite sample analysis of stochastic system identification. In 2019 IEEE 58th Conference on Decision and Control (CDC). IEEE, 3648-3654.
[69] Anastasios Tsiamis, Ingvar Ziemann, Nikolai Matni, and George J. Pappas. 2022. Statistical Learning Theory for Control: A Finite Sample Perspective. https: //doi.org/10.48550/ARXIV.2209.05423
[70] Yong Zhang. 2011. Unbiased identification of a class of multi-input singleoutput systems with correlated disturbances using bias compensation methods. Mathematical and Computer Modelling 53, 9 (2011), 1810-1819. https: //doi.org/10.1016/j.mcm.2010.12.059
[71] Yang Zheng and Na Li. 2021. Non-Asymptotic Identification of Linear Dynamical Systems Using Multiple Trajectories. IEEE Control Systems Letters 5, 5 (nov 2021), 1693-1698. https://doi.org/10.1109/lcsys.2020.3042924
[72] Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. 2020. Robust estimation via generalized quasi-gradients. arXiv preprint arXiv:2005.14073 (2020).

Received 2022-11-07; accepted 2023-02-06

