Outlier-Robust Linear System Identification Under Heavy-Tailed Noise

Vinay Kanakeri and **Aritra Mitra** VKANAKE, AMITRA2@**ncsu.edu**
Department of Electrical and Computer Engineering, North Carolina State University

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

Abstract

We consider the problem of estimating the state transition matrix of a linear time-invariant (LTI) system, given access to multiple independent trajectories sampled from the system. Several recent papers have conducted a non-asymptotic analysis of this problem, relying crucially on the assumption that the process noise is either Gaussian or sub-Gaussian, i.e., "light-tailed". In sharp contrast, we work under a significantly weaker noise model, assuming nothing more than the existence of the fourth moment of the noise distribution. For this setting, we provide the first set of results demonstrating that one can obtain sample-complexity bounds for linear system identification that are nearly of the same order as under sub-Gaussian noise. To achieve such results, we develop a novel robust system identification algorithm that relies on constructing multiple weakly-concentrated estimators, and then boosting their performance using suitable tools from high-dimensional robust statistics. Interestingly, our analysis reveals how the kurtosis of the noise distribution, a measure of heavy-tailedness, affects the number of trajectories needed to achieve desired estimation error bounds. Finally, we show that our algorithm and analysis technique can be easily extended to account for scenarios where an adversary can arbitrarily corrupt a small fraction of the collected trajectory data. Our work takes the first steps towards building a robust statistical learning theory for control under non-ideal assumptions on the data-generating process.

Keywords: Linear system identification; finite-time bounds; heavy-tailed noise; robust statistics.

1. Introduction

Given the empirical success of reinforcement learning in various complex tasks spanning video games to robotics, there has been a recent growth of interest in understanding the performance of feedback control systems when the model of the system is *unknown* (Hu et al., 2023; Tsiamis et al., 2023). To mitigate uncertainty in the model, one natural strategy is to first use data generated by the system to learn the system parameters - a task known as *system identification.* Subsequently, using the learned system model, one can appeal to either certainty-equivalent or robust control. For such a data-driven approach to yield the desired stability and performance guarantees, it is essential to quantify how data sampled from the system can help reduce the uncertainty in the underlying dynamics. This is particularly important for reliable operation in safety-critical applications (e.g., self-driving cars) when one integrates data-driven approaches into the feedback control loop. In this context, a growing body of work has drawn upon tools from learning theory and high-dimensional statistics to characterize the number of samples needed to accurately estimate the system parameters, given access to noisy data. While the results in this space have provided a fine-grained understanding of how the nature of the dynamical system (stable vs. unstable) shapes the sample-complexity bounds for system identification, all such results have been derived under somewhat idealistic assumptions on the data-generating process. In particular, the process noise exciting the dynamics is assumed to be either Gaussian or sub-Gaussian, i.e., “light-tailed", an assumption that may not hold for real-world environments. *Under less favorable circumstances when the noise is heavy-tailed or even adversarial,*

can we still provide a finite-sample analysis of system identification? If so, is there any hope of recovering similar estimation error bounds as achievable under sub-Gaussian noise? In this paper, we provide the first rigorous examination of the above questions for linear system identification.

More precisely, we consider a linear time-invariant (LTI) system: $x_{t+1} = Ax_t + w_t$, where $x_t \in \mathbb{R}^d$ and $w_t \in \mathbb{R}^d$ are the state of the system and process noise at time t, respectively, and $A \in \mathbb{R}^{d\times d}$ is the unknown state transition matrix. Given access to N independent trajectories sampled from this system, the goal is to construct an estimate $\hat{A}$ of A, and characterize the corresponding sample-complexity bounds within the probably approximately correct (PAC) framework. In other words, we wish to precisely quantify the number of trajectories needed to achieve a prescribed estimation accuracy ε with a confidence level specified by failure probability δ. The main contribution of this work is to offer the first PAC bounds for this setting under the assumption that the noise process $\{w_t\}$ admits a finite fourth moment, *and nothing more*. Interestingly, we show that with a suitably designed robust estimator of A, one can (almost) recover bounds known in the literature under the significantly stronger assumption of sub-Gaussian noise.

Related Work. Linear system identification is a fundamental problem in control theory that finds applications in time-series forecasting, finance, and reinforcement learning. Classical treatments of this problem primarily focus on asymptotic results (Lai and Wei, 1983; Ljung et al., 1987). Our interest, however, is in a more recent strand of literature that aims to provide a finer understanding by deriving non-asymptotic bounds on the amount of data that needs to be collected from the system to meet desired performance guarantees. To our knowledge, the first results of this kind were obtained by Rantzer (2018) for scalar LTI systems. In follow-up work, the results were extended to vector (potentially unstable) LTI systems by Dean et al. (2020) for the multi-trajectory setting, i.e., when multiple independent trajectories are available as data to the learner. The multi-trajectory setting we consider here is akin to that in Dean et al. (2020) and has also appeared in various other works (Zheng and Li, 2020; Xing et al., 2020; Xin et al., 2022). When data is collected from a single trajectory, the analysis becomes much more challenging since such data is no longer independent and identically distributed (i.i.d.), but rather temporally correlated. For stable and marginally stable LTI systems, finite-sample results for the single-trajectory case were derived by Simchowitz et al. (2018), Oymak and Ozay (2019), and Jedra and Proutiere (2022). When the state transition matrix can contain unstable modes, results under single-trajectory data were obtained by Sarkar and Rakhlin (2019).

Variations of the basic linear system identification problem involve scenarios where the system state is not fully observed, and the system is only excited by noise; see the work of Tsiamis and Pappas (2019) in this regard. Other variants include the problem of sparse system identification, considered by Fattahi et al. (2019) and Sun et al. (2020). For a detailed discussion of the latest results on system identification, we refer the reader to the excellent tutorials by Tsiamis et al. (2023) and Ziemann et al. (2023). Despite the wealth of literature that has emerged on the topic in recent years, the analysis in each of the papers mentioned above hinges crucially on leveraging concentration bounds for Gaussian or sub-Gaussian noise distributions. One notable exception is the work of Faradonbeh et al. (2018), where the authors consider a noise model weaker than ones admitting sub-Gaussian tails. Nonetheless, the sub-Weibull noise model in Faradonbeh et al. (2018) ensures the existence of all finite moments of the noise distribution. This leads to the following question:

Can we derive finite sample bounds for linear system identification under heavy-tailed noise distributions that admit no more than the fourth moment?

The recent survey paper by Tsiamis et al. (2023) identifies this as an open question. We provide an answer in the affirmative via the following contributions.

• **Problem Formulation.** Our study is motivated by an interesting observation made in Tsiamis et al. (2023). The authors note that for a heavy-tailed noise model where, for instance, $\mathbb{E}[\|w_t\|^4] < \infty$ but $\mathbb{E}[\|w_t\|^p] = \infty$ for some finite $p > 4$, while the ordinary least squares (OLS) estimator might still be optimal in expectation under i.i.d. data, it is no longer optimal w.r.t. its dependence on the failure probability δ (Oliveira, 2016). In particular, it fails to achieve the optimal logarithmic dependence of $\log(1/\delta)$ for all distributions within the aforementioned heavy-tailed noise class. In the context of heavy-tailed linear system identification, we examine for the first time whether such an optimal $\log(1/\delta)$ dependence can be reinstated.

• **Novel Algorithm.** In practice, it may not be possible to ascertain ahead of time whether the noise is sub-Gaussian or heavy-tailed. As such, we would ideally like to have a system identification algorithm that is *agnostic* to the nature of the noise and yields similar guarantees under both light- and heavy-tailed distributions. As discussed earlier, the OLS estimator fails in this regard. This motivates us to develop a novel algorithm titled `Robust-SysID` in Section 3. Our main idea is to first construct multiple OLS estimators of A by suitably partitioning the collected trajectories into buckets. To "boost" the performance of such weakly concentrated estimators, we employ the notion of a geometric median w.r.t. the Frobenius norm. While similar ideas have been pursued for robust mean estimation (Minsker, 2015), we show how they can be also employed for system identification.

• **Matching Sub-Gaussian Rates under Heavy-tailed Noise.** Existing analyses for system identification exploit various concentration tools for sub-Gaussian and sub-exponential distributions. Unfortunately, our noise model precludes the use of such tools, necessitating a new proof technique. To illustrate some of the key ideas that show up in our analysis, we consider a scalar setting in Section 4. Our main result for the scalar case, namely Theorem 1, recovers the *exact same error bound as under sub-Gaussian noise*; in particular, we are able to achieve the desired $\log(1/\delta)$ dependence on the failure probability. Unlike the analogous sub-Gaussian result, however, the number of trajectories needs to scale with the *kurtosis* of the noise distribution, i.e., the ratio of the fourth moment to the square of the variance. While this requirement captures the effect of heavy-tailed noise, whether it is fundamental is an open question. The extension to the vector setting requires much more work to control the smallest eigenvalue of the empirical covariance matrix. Our main result for this case, namely Theorem 4, once again nearly recovers the same error bound as reported in Dean et al. (2020) under sub-Gaussian noise, up to an extra multiplicative $O(d)$ term. *To our knowledge, these are the first results to demonstrate that one can (almost) match sub-Gaussian error rates under heavy-tailed noise for system identification.*

• **Robustness to Outlier Trajectories.** Finally, in Theorem 8, we show that our algorithm and analysis technique can be seamlessly extended to account for the scenario where an adversary can arbitrarily corrupt a small fraction η of the trajectories. Our result in this context is consistent with those for robust mean estimation with adversarial outliers (Lugosi and Mendelson, 2021).

Overall, by drawing on ideas from robust statistics, we take the first steps towards building the foundations of data-driven control under non-ideal (yet more realistic) assumptions on the data-generating process. While we focus on system identification in this paper, we anticipate that our ideas will find broader applicability to more complex feedback control problems under uncertainty.

Notation. Given a positive integer $n \in \mathbb{N}$, we define the shorthand $[n] \triangleq \{1, 2, \ldots, n\}$. For a vector $w \in \mathbb{R}^d$, we will use $w^\top$ to denote its transpose, and $w(i)$ to represent its i-th component. Unless otherwise specified, $\|\cdot\|$ will be used to denote the Euclidean norm for vectors and spectral norm for matrices. Given a matrix $M \in \mathbb{R}^{d\times d}$, we will use $\|M\|_F$ to denote its Frobenius norm. Finally, we will use $c, C, c_1, c_2, \ldots$ to represent universal constants that may change from one line to another.

2. Problem Formulation

Consider an uncontrolled discrete-time linear time-invariant (LTI) system of the following form:

$$x_{t+1} = Ax_t + w_t, \tag{1}$$

where $x_t \in \mathbb{R}^d$ and $w_t \in \mathbb{R}^d$ are the state of the system and process noise at time t, respectively, and $A \in \mathbb{R}^{d\times d}$ is the a priori *unknown* state transition matrix. Without loss of generality, we assume that $x_0 = 0$. We further assume that the noise sequence $\{w_t\}$ is a zero-mean, independent and identically distributed (i.i.d.) stochastic process satisfying the following second- and fourth-moment bounds:

$$\mathbb{E}[w_t w_t^\top] = \sigma^2 I_d, \quad \mathbb{E}[(w_t(i))^4] = \tilde{\sigma}^4, \forall i \in [d], \forall t \geq 0. \tag{2}$$

Data Collection. Suppose we have access to N independent trajectories of the system (1), each of length T. Such trajectories can be generated by rolling out the dynamics for T time-steps, and then resetting the system to the zero initial condition after each rollout. Let us use $\mathcal{D}^{(i)}$ to denote the trajectory data $\{x_t^{(i)}\}_{1\leq t\leq T+1}$ collected during the i-th rollout, where $i \in [N]$. Using the collective data set $\mathcal{D} = \bigcup_{i\in[N]} \mathcal{D}^{(i)}$, the goal of a learner is to obtain an estimate of the system matrix A. Formally, our problem of interest can now be stated as follows.

Problem 1 *Consider the system in (1) and the noise model in (2). Fix an accuracy parameter $\varepsilon > 0$ and a failure probability $\delta \in (0, 1)$. Given the data set $\mathcal{D}$, construct an estimator $\hat{A}$ of A, and characterize its sample-complexity $N_S(\varepsilon, \delta, C_A, C_w)$, such that with probability at least $1 - \delta$, we have $\|\hat{A} - A\| \leq \varepsilon$, provided $N \geq N_S$. Here, C_A and C_w are constants that depend on the system matrix A, and the noise parameters $\sigma, \tilde{\sigma}$, respectively.*

Several comments are now in order regarding our problem formulation.

1. The key departure of our problem setting from existing finite-time results on linear system identification stems from the generality of the assumptions we make on the noise process $\{w_t\}$. In particular, existing work on this topic has either assumed "light-tailed" Gaussian or sub-Gaussian noise. The only notable exception we are aware of in this regard is the work of Faradonbeh et al. (2018), where the authors consider a noise process with sub-Weibull distribution. Although more general than sub-Gaussian distributions, all finite moments of a sub-Weibull distribution exist, as shown by Vladimirova et al. (2020). In sharp contrast, the assumptions we make on the noise process in (2) require nothing more than the existence of the fourth moment of the noise distribution.

2. To build intuition regarding our results, let us consider the well-studied setting where the noise process is Gaussian with variance σ^2. Given N independent trajectories, the ordinary least squares (OLS) estimator yields the following guarantee in this scenario (Dean et al., 2020):

$$\|\hat{A} - A\| \leq c_1 \sqrt{\frac{d\log(1/\delta)}{\lambda_{\min}(G_T)N}} \text{ holds with probability at least } 1 - \delta, \tag{3}$$

provided $N \geq c_2 d\log(1/\delta)$. Here, $c_1, c_2 > 0$ are suitable universal constants, and $G_T := \sum_{t=0}^{T-1} A^t(A^\top)^t$. Our **goal** is to understand whether, and to what extent, similar guarantees can be recovered under the significantly more general noise model we consider in this paper. In particular, we ask: *Is it possible to retain the mild logarithmic dependence on the failure probability δ in (3)*? This is particularly relevant when one seeks high-probability guarantees.

3. To focus on answering the above question, we consider a system model with no inputs. Nonetheless, under the standard assumption of controllability, our techniques can be easily extended to account for a somewhat more general system of the form $x_{t+1} = Ax_t + Bu_t + w_t$, where B is an unknown input matrix, and u_t is the control input at time t. In a similar spirit, to isolate the effect of heavy-tailed noise, here, we do not pursue other natural extensions pertaining to partial observability, measurement noise, single-trajectory data, and nonlinear dynamics. While each of these generalizations are certainly interesting avenues for future work, they are orthogonal to the subject of this paper.

In the sequel, we will show that it is indeed possible to recover bounds of the form in (3). Arriving at such bounds will, however, require an algorithmic technique different from the standard OLS approach. Furthermore, as we will elaborate later in the paper, we cannot appeal to the existing proof techniques for linear system identification that rely heavily on concentration properties of light-tailed distributions. This is all to say that the "simple" model in (1) is interesting in its own right.

3. Robust System Identification Algorithm

In this section, we will develop our proposed algorithm called `Robust-SysID` that enables system identification in the face of heavy-tailed noise. Later, in Section 6, we will see that a minor tweak to this algorithm suffices to accommodate the presence of arbitrarily corrupted adversarial data. In other words, *we will establish that `Robust-SysID` is not only robust to heavy-tailed noise, but also to adversarial outliers.* Our algorithm has three main components that we outline below.

Step 1: Bucketing. In the first step, we partition the N data sets into K buckets denoted by $\mathcal{B}_1, \ldots, \mathcal{B}_K$, such that each bucket contains M independent trajectories; here, for simplicity, we have assumed that $N = MK$. The choice of K is crucial for our final bounds and will be outlined later.

Step 2: Local Estimation per Bucket. In the second step, we use the trajectories within each bucket to construct an OLS estimator per bucket. To make this idea precise, fix a bucket $j \in [K]$. We use the last two samples of each trajectory within $\mathcal{B}_j$ to construct the OLS estimator $\hat{A}_j$ for bucket j:

$$\hat{A}_j = \operatorname*{argmin}_{\theta \in \mathbb{R}^{d \times d}} \sum_{i \in \mathcal{B}_j} \|x_{T+1}^{(i)} - \theta x_T^{(i)}\|^2. \tag{4}$$

Step 3: Boosting. In the last step, we fuse the "weak" estimates obtained from each bucket to create a more powerful estimator for A. Specifically, we leverage the notion of a geometric median for matrices to construct $\hat{A}$ as follows:

$$\hat{A} = \texttt{Med}(\hat{A}_1, \ldots, \hat{A}_K) := \operatorname*{argmin}_{\theta \in \mathbb{R}^{d \times d}} \sum_{j \in [K]} \|\theta - \hat{A}_j\|_F. \tag{5}$$

In the above step, the geometric median $\hat{A}$ is computed with respect to the Frobenius norm since it induces an inner-product space on the space of all matrices in $\mathbb{R}^{d \times d}$. In turn, this guarantees the existence of $\hat{A}$ as defined in (5) (Minsker, 2015). At a high level, we note that our algorithmic strategy is inspired by the popular "median of means" device from robust statistics. While such ideas have been explored in the past for robust mean estimation, we employ them here for the first time in the context of linear system identification. In the subsequent sections, we will discuss the performance guarantees of `Robust-SysID`, and sketch out the main steps in the analysis, while highlighting the challenges that arise in the way.

4. Warm Up: The Scalar Case

We start by analyzing a scalar version of the system in (1) since it captures much of the challenges posed by heavy-tailed noise. Therefore, analyzing `Robust-SysID` for this case provides us with valuable insights into the nature of the bounds to be expected in the more challenging vector setting. Our main result on the performance of `Robust-SysID` for the scalar case is stated in the following theorem, where we define a to be the scalar counterpart of A from (1), and $g_T = \sum_{t=0}^{T-1} a^{2t}$.

Theorem 1 *Consider the scalar version of the system in (1) and the noise assumptions in (2). With probability at least $1-\delta$, the following bound holds for the output $\hat{a}$ of `Robust-SysID`:*

$$|\hat{a} - a| \leq C\sqrt{\frac{\log(1/\delta)}{N g_T}}, \text{ provided } K = \lceil c_1 \log(1/\delta) \rceil, M \geq c_2(\tilde{\sigma}^4/\sigma^4), \text{ and } N = MK. \quad (6)$$

Before providing a proof sketch of the above result, some remarks are in order.

Discussion. We note that the error bound in (6) matches the one obtained by the standard OLS estimator under Gaussian noise, as indicated in (3). This reveals the robustness of our algorithm to a general heavy-tailed noise process. Theorem 1 also specifies the design parameters of `Robust-SysID`, namely the number of buckets K, and the number of samples per bucket M. Since $N = MK$, we note from (6) that the number of trajectories depends on the kurtosis of the noise process - a dependence not observed under Gaussian noise. Although we are uncertain whether a dependence on the kurtosis is inevitable, it is, however, meaningful as it captures the heaviness of the tail.

Proof Sketch for Theorem 1. In what follows, we sketch the main ideas in the proof of Theorem 1. Due to space constraints, detailed proofs of all our results are made available in Kanakeri and Mitra (2024). The main hurdle in our analysis is that we can no longer leverage concentration bounds for sub-Gaussian and sub-exponential distributions that have appeared in prior works. Nonetheless, we start by deriving bounds for the OLS estimator from each bucket. Accordingly, the OLS estimator $\hat{a}_j$ for the jth bucket can be expressed as

$$\hat{a}_j = a + \frac{\sum_{i \in \mathcal{B}_j} x_T^{(i)} w_T^{(i)}}{\sum_{i \in \mathcal{B}_j} \left(x_T^{(i)}\right)^2}. \quad (7)$$

We bound the numerator and the denominator of the error term separately and then combine them by applying an union bound. In this regard, the following lemmas provide key results.

Lemma 2 ***(Scalar numerator upper bound)*** *Fix a bucket $j \in [K]$. With probability at least $1 - p/2$,*

$$\left| \sum_{i \in \mathcal{B}_j} x_T^{(i)} w_T^{(i)} \right| \leq c\sigma^2 \sqrt{g_T M/p}.$$

To prove the above result, one can start by noting that due to the independence of trajectories, $\mathrm{Var}(\sum_{i \in \mathcal{B}_j} x_T^{(i)} w_T^{(i)}) = M \,\mathrm{Var}(x_T^{(1)} w_T^{(1)})$, where $M = |\mathcal{B}_j|$, and $\mathrm{Var}(Z)$ is used to represent the variance of a real-valued random variable Z. Next, observe that for each individual term, $\mathrm{Var}(x_T^{(1)} w_T^{(1)}) = \mathbb{E}[(x_T^{(1)})^2]\mathbb{E}[(w_T^{(1)})^2] = \sigma^2 g_T \times \sigma^2 = \sigma^4 g_T$; here, we exploited the fact that $x_T^{(1)}$ and $w_T^{(1)}$ are independent, and $\mathbb{E}[x_T^{(1)}] = \mathbb{E}[w_T^{(1)}] = 0$. The rest follows from a straightforward application of Chebyshev's inequality. Next, the following lemma provides a lower bound on $\sum_{i \in \mathcal{B}_j} \left(x_T^{(i)}\right)^2$, by exploiting the existence of the fourth moment of $w_T^{(i)}$.

Lemma 3 ***(Scalar denominator lower bound)*** *Fix a bucket $j \in [K]$. With probability at least $1 - p/2$,*

$$\sum_{i \in \mathcal{B}_j} \left(x_T^{(i)}\right)^2 \geq \sigma^2 g_T M/2, \;\; \textit{provided} \;\; M \geq (c/p)(\tilde{\sigma}^4/\sigma^4).$$

Proof Due to the i.i.d nature of the trajectories, we have $\mathrm{Var}\left(\sum_{i \in \mathcal{B}_j} \left(x_T^{(i)}\right)^2\right) = M \,\mathrm{Var}\left(\left(x_T^{(1)}\right)^2\right)$. For clarity of notation, let us drop the superscript in the rest of the proof. Since $\mathrm{Var}\left(x_T^2\right) \leq \mathbb{E}\left[x_T^4\right]$, it suffices to bound the fourth moment of x_T. Under the zero initial condition, observe that $x_T = \sum_{t=0}^{T-1} a^t n_t$, where we have defined $n_t \triangleq w_{T-(t+1)}$ for brevity. The fourth moment of x_T can be expressed as follows:

$$\begin{aligned}
\mathbb{E}[x_T^4] &= \mathbb{E}\left[\left(\sum_{t=0}^{T-1} a^t n_t\right)^2 \left(\sum_{s=0}^{T-1} a^s n_s\right)^2\right] \\
&= \mathbb{E}\left[\left(\underbrace{\sum_{t=0}^{T-1} a^{2t} n_t^2}_{T_1} + \underbrace{\sum_{t' \neq t=0}^{T-1} a^{t'} a^t n_t n_{t'}}_{T_2}\right)\left(\underbrace{\sum_{s=0}^{T-1} a^{2s} n_s^2}_{T_3} + \underbrace{\sum_{s' \neq s=0}^{T-1} a^{s'} a^s n_s n_{s'}}_{T_4}\right)\right].
\end{aligned}$$

In the above, as a result of the noise process being i.i.d. with zero mean, only the terms that contribute either a fourth power or a product of squares survive the expectation, causing the cross terms $T_1 \times T_4$ and $T_2 \times T_3$ to vanish. Furthermore, in $T_2 \times T_4$, only terms of the form $n_t^2 n_{t'}^2$ survive, yielding

$$\mathbb{E}[x_T^4] = \sum_{t=0}^{T-1} a^{4t}\,\mathbb{E}[n_t^4] + 3\sum_{t' \neq t=0}^{T-1} a^{2(t+t')}\,\mathbb{E}[n_t^2 n_{t'}^2] = \sum_{t=0}^{T-1} a^{4t}\tilde{\sigma}^4 + 3\sum_{t' \neq t=0}^{T-1} a^{2(t+t')}\sigma^4 \overset{(a)}{\leq} 3(g_T)^2\tilde{\sigma}^4.$$

In the above steps, (a) follows from the definition of g_T in Theorem 1, and $\tilde{\sigma}^4 \geq \sigma^4$ due to Jensen's inequality. Now applying Chebyshev's bound with the above result, we have for any $t > 0$:

$$\mathbb{P}\left(\left|\sum_{i \in \mathcal{B}_j} \left(x_T^{(i)}\right)^2 - \mathbb{E}\left[\sum_{i \in \mathcal{B}_j} \left(x_T^{(i)}\right)^2\right]\right| \geq t\right) \leq \frac{3M(g_T)^2\tilde{\sigma}^4}{t^2}.$$

Setting the R.H.S. of the above inequality to $p/2$, we get $t = g_T\tilde{\sigma}^2\sqrt{6M/p}$. Notice that $\mathbb{E}\left[\sum_{i \in \mathcal{B}_j} \left(x_T^{(i)}\right)^2\right] = Mg_T\sigma^2$, giving us the following with probability at least $1 - p/2$:

$$\sum_{i \in \mathcal{B}_j} \left(x_T^{(i)}\right)^2 \geq g_T\left(\sigma^2 M - \tilde{\sigma}^2\sqrt{6M/p}\right).$$

In the above display, setting the R.H.S $\geq g_T\sigma^2 M/2$ and solving for M completes the proof. ■

Combining the results from Lemma 2 and Lemma 3, and using an union bound, we have that when $M \geq (c_1/p)(\tilde{\sigma}^4/\sigma^4)$, the following holds with probability at least $1 - p$:

$$|\hat{a}_j - a| \leq c_2\sqrt{\frac{1}{pMg_T}}. \tag{8}$$

Note that the failure probability p appears polynomially (and not logarithmically) in the above bound.

The Role of Boosting. In (8), set $p = 1/4$, and let $\varepsilon = c_2(pMg_T)^{-1/2}$. In the scalar case, note that $\hat{a}$ in (5) is simply the standard (scalar) median of $\{\hat{a}_1, \ldots, \hat{a}_K\}$. By the property of the median, observe that the "bad" event $\{|\hat{a} - a| > \varepsilon\}$ implies $\{\sum_{j\in[K]} Y_j \geq K/2\}$, where Y_j is an indicator random variable of the event $\{|\hat{a}_j - a| > \varepsilon\}$. Using the fact that each of the Y_j's are i.i.d. random variables in $\{0, 1\}$ satisfying $\mathbb{E}[Y_j] \leq p = 1/4$, we can use Hoeffding's inequality to infer that

$$\mathbb{P}\left(\{|\hat{a} - a| > \varepsilon\}\right) \leq \exp(-K/8) \leq \delta,$$

when $K = \lceil 8\log(1/\delta)\rceil$. Using this expression for K in $\varepsilon = c_2(pMg_T)^{-1/2}$ and noting that $M = N/K$, we arrive at the bound in (6). This completes the proof sketch for Theorem 1. In simple words, for the median estimate $\hat{a}$ to deviate from a beyond our desired error tolerance ε, at least half of the OLS estimates from the buckets must also deviate by ε. Although the failure probability of each one of such (independent) events is at most $1/4$, asking $K/2$ of such events to occur *simultaneously* diminishes the overall failure probability, thereby "boosting" the quality of $\hat{a}$. With this intuition in mind, we now proceed to analyze the vector case in the following section.

5. The Vector Case

In this section, we demonstrate how `Robust-SysID` addresses Problem 1. We also discuss some of the unique challenges posed by the vector setting compared to the scalar case from the previous section. The following theorem captures our main result.

Theorem 4 *(**Main Result**) Consider the system in (1) and the noise assumptions in (2). With probability at least $1 - \delta$, the following bound holds for the output $\hat{A}$ of `Robust-SysID`:*

$$\|\hat{A} - A\| \leq Cd^{3/2}\sqrt{\frac{\log(1/\delta)}{N\lambda_{\min}(G_T)}}, \text{ when } K = \lceil c_1\log(1/\delta)\rceil, M \geq c_2 d^2 C_A C_w,\ N = MK, \quad (9)$$

$$\text{where } C_A \triangleq \left(\frac{\sum_{t=0}^{T-1}\|A^t\|^2}{\lambda_{\min}(G_T)}\right)^2, \text{ and } C_w \triangleq \frac{\tilde{\sigma}^4}{\sigma^4}. \quad (10)$$

The detailed proof of the above result is available in Kanakeri and Mitra (2024). Before sketching the main ingredients in the analysis, we discuss the implications of Theorem 4.

Discussion. Comparing (3) and (9), we note that `Robust-SysID` recovers the logarithmic dependence on the failure probability. To our knowledge, this is the first result to provide such a guarantee for the general heavy-tailed noise model considered in this work. That said, we note from (9) that our error bound, and the requirement on the number of trajectories, both suffer from an extra multiplicative factor of $O(d)$ relative to the Gaussian benchmark in (3). Furthermore, unlike the scalar case in (6), the requirement on the number of trajectories in the vector case exhibits an additional dependency on the system matrix A via the constant C_A in (10).

Proof Sketch for Theorem 4. Analogous to the scalar case, the proof of Theorem 4 first involves deriving bounds for the OLS estimators of each bucket. To simplify our analysis, we whiten the vector $x_T^{(i)}$ and define $z_T^{(i)} = \Sigma_x^{-1/2}x_T^{(i)}$, where $\Sigma_x = \mathbb{E}[x_T^{(i)}(x_T^{(i)})^\top]$. Under this definition, it suffices to individually bound $\left\|\sum_{i\in\mathcal{B}_j} w_T^{(i)}(z_T^{(i)})^\top\right\|$ and $\lambda_{\min}\left(\sum_{i\in\mathcal{B}_j} z_T^{(i)}(z_T^{(i)})^\top\right)$ as shown in Dean et al. (2020), Matni and Tu (2019). The following lemmas provide key results in this regard.

Lemma 5 *(**Vector numerator upper bound**) Fix a bucket $j \in [K]$. With probability at least $1 - p/2$,*

$$\left\| \sum_{i \in \mathcal{B}_j} w_T^{(i)} \left(z_T^{(i)}\right)^\top \right\| \leq c_1 d \sqrt{\sigma^2 M/p}.$$

Challenges in Analysis. Let us discuss some of the challenges that arise in the proof of the above result by outlining potential proof strategies, and their limitations for our setting. In Dean et al. (2020), the authors derive a similar result under Gaussian noise by exploiting variational properties of the spectral norm along with covering arguments. This was made possible due to the availability of sub-Gaussian and sub-exponential tail bounds with logarithmic dependence on the error probability, which, in turn, help in controlling certain covering numbers. Clearly, without the logarithmic factor, as is the case with heavy-tailed noise, using such an approach would lead to a prohibitive exponential dependence on the dimension d due to the covering number. An alternative strategy to bound the norm of $\sum_{i \in \mathcal{B}_j} w_T^{(i)} \left(z_T^{(i)}\right)^\top$ is to bound each scalar entry of this matrix by invoking the analysis from Section 4. However, this approach fails to provide the bound in Lemma 5 as it would involve union bounding over d^2 elements, leading to an additional dimension factor.

Our methods. In light of the above discussion, we identify two different approaches. In the first approach, we define a new variance statistic for a random square matrix X as $\mathrm{var}(X) \triangleq \mathbb{E}[\|X - \mathbb{E}[X]\|_F^2]$. Such a definition using the Frobenius norm leverages independence in the sense that $\mathrm{var}(\sum_i X_i) = \sum_i \mathrm{var}(X_i)$ for independent matrices $\{X_i\}$. Since $\mathrm{var}(X)$ as defined above is a scalar, one can use the standard Markov's inequality in this case. The second approach exploits a matrix version of Markov's inequality proposed by Ahlswede and Winter (2001). It turns out that both approaches lead to exactly the same bounds in Lemma 5 and Lemma 6. Our next result controls the smallest eigenvalue of the (whitened) empirical covariance matrix.

Lemma 6 *(**Vector denominator lower bound**) For each bucket j, with probability at least $1 - p/2$,*

$$\lambda_{\min} \left(\sum_{i \in \mathcal{B}_j} z_T^{(i)} \left(z_T^{(i)}\right)^\top \right) \geq M/2, \quad \textit{provided} \quad M \geq c(d^2/p) C_A C_w.$$

The key step in the proof of the above result involves bounding the trace of the matrix $\mathbb{E}[(x_T x_T^\top)^2]$; here, we have dropped the superscript (i) for clarity. By exploiting the i.i.d. and zero-mean properties of the noise process $\{w_t\}$, the next result helps considerably simplify the expression for $\mathbb{E}[(x_T x_T^\top)^2]$.

Lemma 7 *Define $n_t \triangleq w_{T-(t+1)}$. Given the system in (1) and the noise assumptions in (2), we have*

$$\begin{aligned}\mathbb{E}\left[(x_T x_T^\top)^2\right] = &\sum_{t=0}^{T-1} \mathbb{E}\left[(A^t n_t n_t^\top (A^t)^\top)^2\right] + 2 \sum_{s \neq t = 0}^{T-1} \mathbb{E}\left[A^t n_t n_t^\top (A^t)^\top A^s n_s n_s^\top (A^s)^\top\right] \\ &+ \sum_{s \neq t = 0}^{T-1} \mathbb{E}\left[A^t n_t n_s^\top (A^s)^\top A^s n_s n_t^\top (A^t)^\top\right].\end{aligned}$$

We use the above result in tandem with various trace inequalities to establish Lemma 6. Combining the results from Lemma 5 and Lemma 6 immediately provides a guarantee for the OLS estimate from each bucket, much like in (8). For boosting, we employ an argument similar to the scalar case, along with properties of the geometric median from Minsker (2015).

Avenues for improvement. We discuss the sources of the extra $O(d)$ factor (relative to the sub-Gaussian noise case) in our error-bound of (9). To invoke the results from Minsker (2015) for the geometric median, we need to work with the Frobenius norm. As such, we use the inequality $\|\hat{A}_j - A\|_F \leq \sqrt{d}\|\hat{A}_j - A\|$, costing us an extra $\sqrt{d}$ factor in the boosting step. This could be avoided if it were possible to provide guarantees for robust matrix aggregation directly w.r.t. the spectral norm. The other $\sqrt{d}$ factor comes from Lemma 5 for which we used variants of Markov's inequality. One might hope that a more powerful concentration tool can lead to a tighter bound. Perhaps the most relevant result in this context is provided by Theorem 5.48 in Vershynin (2010), which concerns bounding the expected value of the spectral norm of a matrix with heavy-tailed rows. Applied to our setting, we obtain with probability at least $1-p$, $\left\|\sum_{i\in\mathcal{B}_j} w_T^{(i)} \left(z_T^{(i)}\right)^\top\right\| \leq c_1\sqrt{d\sigma^2 M/p} + c_2\sqrt{m\log(d)/p}$, where $m = \mathbb{E}\left[\max_{k\in[d]}\left\|\sum_{i\in\mathcal{B}_j} w_T^{(i)}(k) z_T^{(i)}\right\|^2\right]$. Notice that if we could commute the $\max$ and $\mathbb{E}[\cdot]$ operators in the definition of m, we would be able to shave off a $\sqrt{d}$ factor from the bound of Lemma 5. However, such an operation is not valid in general. On the other hand, if we upper-bound the $\max$ by summing over $k \in [d]$, we end up with the same bound as in Lemma 5. Although tighter concentration bounds are available for the maxima of sub-Gaussian random variables, we are unaware of analogous bounds under our noise assumptions. Thus, it remains an open problem to ascertain whether our current bounds can be further improved.

6. System Identification under Adversarial Corruptions

In this section, we show that our prior developments concerning `Robust-SysID` can be extended to account for adversarial corruption in conjunction with heavy-tailed noise. To make this precise, we consider the *strong-contamination* attack model from the robust statistics literature (Lugosi and Mendelson, 2021), where an adversary can *arbitrarily* corrupt a small fraction $\eta \in [0, 1/2)$ of the data. In our context, we allow the adversary to contaminate up to ηN number of trajectories in the data set $\mathcal{D}$. We have the following result for this setting.

Theorem 8 ***(Robustness to adversarial corruptions)*** *Consider the strong-contamination model described above. With probability at least $1-\delta$, the following bound holds for the output $\hat{A}$ of* `Robust-SysID` *when $\eta < 0.5/(c_1 d^2 C_A C_w)$, $K \geq \lceil c_2\log(1/\delta) + c_3\eta N\rceil$, and $M \geq c_4 d^2 C_A C_w$:*

$$\|\hat{A} - A\| \leq C d^{3/2}\left(\sqrt{\frac{\log(1/\delta)}{N\lambda_{\min}(G_T)}} + \sqrt{\frac{\eta}{\lambda_{\min}(G_T)}}\right). \tag{11}$$

Comparing the above result with the case without adversarial corruptions from Theorem 4, notice that the error bound in (11) recovers the bound in (9) up to an additive $O(\sqrt{\eta})$ term; this is consistent with analogous results for robust mean estimation in Lugosi and Mendelson (2021). Therefore, Theorem 8 shows that `Robust-SysID` can effectively counter adversarial corruption by carefully designing the number of buckets, and leveraging the inherent robustness of the geometric median. To gain intuition, consider the worst-case scenario where the adversary corrupts ηN buckets by corrupting exactly one trajectory in each such bucket. To keep the median well-concentrated in this case, we need the number of *uncorrupted* buckets to be in the order of $\log(1/\delta)$ as shown in (9), which can be ensured with $O(\eta N)$ extra buckets. This explains the requirement on K in Theorem 8, which, in turn, imposes bounds on N and η. Interestingly, similar constraints on N and η are *not* required for robust mean estimation (Lugosi and Mendelson, 2021). This difference can be attributed to the fact that, unlike mean estimation, sys-ID requires a minimum number of trajectories M per bucket to ensure that the empirical covariance matrix in each uncorrupted bucket is well-behaved.

References

R. Ahlswede and A. Winter. Strong converse for identification via quantum channels, 2001. URL https://arxiv.org/abs/quant-ph/0012127.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018.

Salar Fattahi, Nikolai Matni, and Somayeh Sojoudi. Learning sparse dynamical systems from a single sample trajectory. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 2682–2689. IEEE, 2019.

Bin Hu, Kaiqing Zhang, Na Li, Mehran Mesbahi, Maryam Fazel, and Tamer Başar. Toward a theoretical foundation of policy optimization for learning control policies. *Annual Review of Control, Robotics, and Autonomous Systems*, 6(1):123–158, 2023.

Yassir Jedra and Alexandre Proutiere. Finite-time identification of linear systems: Fundamental limits and optimal algorithms. *IEEE Transactions on Automatic Control*, 68(5):2805–2820, 2022.

Vinay Kanakeri and Aritra Mitra. Outlier-robust linear system identification with heavy-tailed noise. *arXiv preprint*, 2024.

TL Lai and CZ Wei. Asymptotic properties of general autoregressive models and strong consistency of least-squares estimates of their parameters. *Journal of multivariate analysis*, 13(1):1–23, 1983.

Lennart Ljung et al. Theory for the user. *System identification*, 1987.

Gabor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. 2021.

Nikolai Matni and Stephen Tu. A tutorial on concentration bounds for system identification, 2019. URL https://arxiv.org/abs/1906.11395.

Stanislav Minsker. Geometric median and robust estimation in banach spaces. 2015.

Roberto Imbuzeiro Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166:1175–1194, 2016.

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

Anders Rantzer. Concentration bounds for single parameter adaptive control. In *2018 Annual American Control Conference (ACC)*, pages 1862–1866. IEEE, 2018.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pages 5610–5618. PMLR, 2019.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

Yue Sun, Samet Oymak, and Maryam Fazel. Finite sample system identification: Optimal rates and the role of regularization. In *Learning for dynamics and control*, pages 16–25. PMLR, 2020.

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

Anastasios Tsiamis, Ingvar Ziemann, Nikolai Matni, and George J Pappas. Statistical learning theory for control: A finite-sample perspective. *IEEE Control Systems Magazine*, 43(6):67–97, 2023.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9 (1):e318, 2020.

Lei Xin, George Chiu, and Shreyas Sundaram. Learning the dynamics of autonomous linear systems from multiple trajectories. In *2022 American Control Conference (ACC)*, pages 3955–3960. IEEE, 2022.

Yu Xing, Ben Gravell, Xingkang He, Karl Henrik Johansson, and Tyler Summers. Linear system identification under multiplicative noise from multiple trajectory data. In *2020 American Control Conference (ACC)*, pages 5157–5261. IEEE, 2020.

Yang Zheng and Na Li. Non-asymptotic identification of linear dynamical systems using multiple trajectories. *IEEE Control Systems Letters*, 5(5):1693–1698, 2020.

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