Sample Complexity Bounds for Linear System Identification from a Finite Set

Nicolas Chatzikiriakos and Andrea Iannelli

***Abstract*— This paper considers a finite sample perspective on the problem of identifying an LTI system from a finite set of possible systems using trajectory data. To this end, we use the maximum likelihood estimator to identify the true system and provide an upper bound for its sample complexity. Crucially, the derived bound does not rely on a potentially restrictive stability assumption. Additionally, we leverage tools from information theory to provide a lower bound to the sample complexity that holds independently of the used estimator. The derived sample complexity bounds are analyzed analytically and numerically.**

***Index Terms*— Linear System Identification, Maximum Likelihood Estimation, Finite Sample Analysis**

I. Introduction

In this work, we consider the problem of identifying an linear time-invariant (LTI) system from a finite set of system models using data from a finite and noisy trajectory. To this end, we use tools from statistical learning theory and information theory to derive high probability upper and lower bounds of the sample complexity of identifying the true system using the maximum likelihood estimator (MLE).

The problem of identifying the true system from a given finite set of systems naturally appears in many applications [1]. In particular, switched systems are prevalent across different domains such as mechanical systems or the automotive industry [2]. Identifying the active mode of a switched system is also a relevant problem for control, as shown, e.g., by several works in the adaptive control literature [3]. Recently, renewed interest has emerged in this literature for considering uncertainty in the form of a finite set of models [4]. An additional application domain motivating our interest are ecological and evolutionary models where often the correct hypothesis out of a finite hypothesis class needs to be determined from observations [5]. Choosing an element from a finite set by leveraging information coming from measured data is also paradigmatic of many decision-making problems, e.g., in the bandits literature [6]. A notable example is best-arm identification, a pure exploration problem that has been successfully applied, e. g., for clinical trials [7]. In the fixed confidence setting [8], the goal is to identify the best arm among a selection of arms with a desired confidence. In the dynamical systems setting, this translates into the high-probability identification of the true system from a finite set of systems. A key difference, to the setup considered in this work is that best-arm identification assumes static arms and uses the control input to achieve the goal as fast as possible In this work, we assume Gaussian control inputs however the system dynamic introduces correlation we need to address.

Related Works: While in the asymptotic regime, the statistic analysis of system identification has a long history [9], recently novel tools in statistical learning theory [10], [11] sparked an increasing interest in non-asymptotic results, analyzing error bounds of estimated models, especially the ordinary least squares estimator (OLS), from a finite-sample perspective. While first works [12], [13] relied on i.i.d data the now predominant part of the literature analyzes the case of trajectory data. Hereby, [14] was the first work to provide a finite sample analysis for fully observed marginally stable systems. The analysis was extended to unstable systems [15] uncovering a statistical inconsistency of the OLS under certain conditions. Results providing lower bounds on the sample complexity [16] further enabled a fundamental analysis and understanding of difficulties in identifying linear systems from trajectory data [17]. The finite sample analysis of the identification of switched systems using OLS has been considered in [18], with known switching sequences and unknown system matrices. For a comprehensive overview of non-asymptotic system identification, we refer to [19].

Contribution: To the best of our knowledge, we provide the first sample complexity analysis for the problem of identifying an unknown LTI system from a finite set of models. To this end, we derive an instance specific sample complexity upper bound for the MLE. Hereby, the tools used to analyze the unconstrained OLS when identifying an unknown system from a continuous set can not be applied, since they rely on the closed-form solution of the OLS. In contrast, our proof relies on showing high-probability concentration of the cost of the MLE for the true system and anti-concentration for all other systems. A further advantage of the proposed approach is that, unlike most finite sample results for the continuous set case, the results in this work do not rely on any stability assumptions. In addition, we provide an instance specific sample complexity lower bound for δ-stable algorithms. An analytical and numerical analysis of the bounds shows which factors influence the hardness of identification.

Notation: The unit sphere in $\mathbb{R}^n$ is denoted by $\mathbb{S}^{n-1}$. Given a matrix M we denote its Frobenius norm by $\|M\|_{\mathrm{F}}$. Given a vector $x \in \mathbb{R}^n$ and a matrix $P \succ 0$ we define $\|x\|_P^2 = x^\top P x$.

This work is funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 – 390740016. We acknowledge the support by the Stuttgart Center for Simulation Science (SimTech).

The authors are with the University of Stuttgart, Institute for Systems Theory and Automatic Control, 70550 Stuttgart, Germany. (e-mail: {nicolas.chatzikiriakos, andrea.iannelli}@ist.uni-stuttgart.de)

In the special case of $P = I$ we omit the subscript. We denote matrix blocks that can be inferred from symmetry by $\star$, i. e., we write $\Lambda^\top \Sigma \Lambda = [\star]\Sigma\Lambda$. We denote the canonical basis in $\mathbb{R}^n$ with $e_1, \dots, e_n$. Given some $z \in \mathbb{R}$ we write $\lfloor z \rfloor$ to denote the floor-function. We use the shorthand $\mathbb{P}_\theta[\cdot]$ ($\mathbb{E}_\theta[\cdot]$) when referring to the probability (expectation) given that the system generating the data is given by θ.

II. Preliminaries

A. Problem setup

Consider the linear time-invariant discrete-time system

$$x_{t+1} = A_* x_t + B_* u_t + w_t, \tag{1}$$

where $x_t \in \mathbb{R}^{n_x}$ is the state of the system, $u_t \in \mathbb{R}^{n_u}$ is the control input and $w_t \in \mathbb{R}^{n_x}$ is unknown process noise. We seek to identify the unknown system matrices $\theta_* = (A_*, B_*)$ from a single trajectory $\{x_t\}_{t=1}^T$, $\{u_t\}_{t=1}^{T-1}$ of length T. We assume the data is collected as follows.

Assumption 1: The process noise and control input are i. i. d. zero-mean Gaussian with known covariance matrices $\Sigma_w, \Sigma_u \succ 0$, i. e., $w_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_w)$ and $u_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_u)$.

While Assumption 1 is standard in non-asymptotic identification literature [14], [19], extensions to other noise classes are an interesting topic for future work. Further, we assume to know a finite set of possible systems which contains the true system, i. e., $\theta_* \in \mathcal{S} := \{\theta_0, \dots, \theta_N\}$, where $\theta_i = (A_i, B_i)$. We assume $(A_*, B_*) = (A_0, B_0)$ to simplify the notation.

B. Maximum likelihood estimation

To identify the true system (A_*, B_*) from data we resort to the MLE, which is asymptotically optimal and efficient [20]. Here, we analyze its non-asymptotic behavior, by means of statistical learning theory tools. To this end, observe that the probability of collecting the data $d_t := \{x_{t+1}, x_t, u_t\}$ from system i is given by

$$\mathbb{P}_{\theta_i}(d_t) = \frac{1}{\sqrt{(2\pi)^{n_x}|\Sigma_w|}} e^{-\frac{1}{2}[\star]\Sigma_w^{-1}(x_{t+1} - A_i x_t - B_i u_t)}.$$

Based on this observation, we define the cost

$$\ell_{\theta_i}(x_t, u_t) = \|x_{t+1} - A_i x_t - B_i u_t\|^2_{\Sigma_w^{-1}}, \tag{2}$$

which is proportional to the negative log-likelihood of observing the data $\{x_{t+1}, x_t, u_t\}$ from system i. Thus, the MLE minimizes the empirical risk $\hat{L}(\theta) = \frac{1}{T}\sum_{t=1}^T \ell_\theta(x_t, u_t)$, i. e.,

$$\hat{\theta}_T \in \arg\min_{\theta \in \mathcal{S}} \hat{L}(\theta). \tag{3}$$

For our analysis we adopt the notion of sample complexity (see, e. g., [19]) and aim at deriving an instance specific sample complexity upper bound for the MLE (3). That is, given a chosen failure probability $\delta \in (0, 1)$, we provide guarantees of the form

$$\mathbb{P}[\hat{\theta}_T = \theta_*] \geq 1 - \delta, \quad \text{if } T \geq T_{\text{ub}}, \tag{4}$$

where $T_{\text{ub}} = T_{\text{ub}}(\delta, \mathcal{A}, \mathcal{S}, \theta_*)$ is an upper bound of the problem specific sample complexity of an analyzed estimation algorithm $\mathcal{A}$. Additionally, we provide an instance-specific sample complexity lower bound that holds independently of the used estimator, i. e., we show that any

$$\mathbb{P}\left[\hat{\theta}_T = \theta_*\right] \leq 1 - \delta \quad \text{if } T \leq T_{\text{lb}}, \tag{5}$$

where $T_{\text{lb}} = T_{\text{lb}}(\delta, \mathcal{S}, \theta_*)$ is a lower bound of the problem specific sample complexity for any estimation algorithm yielding the estimate $\hat{\theta}_T$.

III. Finite sample identification

A. A sample complexity upper bound

Plugging the dynamics (1) into the cost (2) yields

$$\ell_{\theta_i}(x_t, u_t) = \|w_t + \Delta A_i x_t + \Delta B_i u_t\|^2_{\Sigma_w^{-1}},$$

where we defined $\Delta A_i := A_* - A_i$ and $\Delta B_i := B_* - B_i$. Further, we define

$$z_t^i := \Sigma_w^{-1/2}(w_t + \Delta A_i x_t + \Delta B_i u_t), \tag{6}$$

where $\Sigma_w^{-1} = \left(\Sigma_w^{-1/2}\right)^\top \Sigma_w^{-1/2}$. With this, the empirical risk reads $\hat{L}(\theta_i) = \frac{1}{T}\sum_{t=1}^T \|z_t^i\|^2$. Since the sum of Gaussian random variables is Gaussian we have that for all $t \in [1, T]$

$$x_t \sim \mathcal{N}\Bigg(A^t x_0, \underbrace{\sum_{\tau=0}^{t} [\star]\Sigma_u (A_*^\tau B_*)^\top + [\star]\Sigma_w A_*^{\tau\top}}_{:=\Sigma_{x_t}}\Bigg). \tag{7}$$

With this, we see that z_t^i is Gaussian distributed as well, i. e.,

$$z_t^i \sim \mathcal{N}\left(\Sigma_w^{-1/2}\Delta A_i A^t x_0, \Sigma_{z_t}^i\right), \tag{8}$$

where

$$\Sigma_{z_t}^i := I + \Sigma_w^{-1/2}\left([\star]\Sigma_u \Delta B_i^\top + [\star]\Sigma_{x_t}\Delta A_i^\top\right)\Sigma_w^{-1/2\top}. \tag{9}$$

It is important to note that the sequence $(z_t^i)_{t\geq 1}$ is highly correlated due to the underlying LTI system generating the data. To handle the correlation, we leverage the block martingale small-ball condition introduced in [14].

Definition 1 (Block Martingale Small-Ball [14]): Let $(\zeta_t)_{t\geq 1}$ be a $\{\mathcal{F}_t\}_{t\geq 1}$-adapted random process taking values in $\mathbb{R}$. We say $(\zeta_t)_{t\geq 1}$ satisfies the (k, ν, p)-block martingale small-ball (BMSB) condition if, for any $j \geq 0$

$$\frac{1}{k}\sum_{i=1}^k \mathbb{P}[|\zeta_{j+i}| \geq \nu | \mathcal{F}_j] \geq p \quad \text{a.s.}$$

Given a process $(z_t)_{t\geq 1}$ taking values in $\mathbb{R}^{n_z}$, we say that it satisfies the $(k, \Gamma_{\text{sb}}, p)$-BMSB condition for $\Gamma_{\text{sb}} \succ 0$ if, for any fixed $v \in \mathbb{S}^{n_x - 1}$ the process $\zeta_t = \langle v, z_t \rangle$ satisfies $(k, \sqrt{v^\top \Gamma_{\text{sb}} v}, p)$-BMSB.

The BMSB condition establishes a level of anti-concentration along a sequence. We can show that the sequence $(z_t^i)_{t=1}^T$ satisfies the BMSB condition.

Proposition 1: Let Assumption 1 hold, let z_t^i be defined according to (6) and define

$$\mathcal{F}_t := (w_0, \dots, w_t, x_0, \dots, x_t, u_0, \dots, u_t).$$

Then the $\{\mathcal{F}_t\}_{t=1}^T$-adapted random process $(z_t^i)_{t=1}^T$ satisfies the $\left(k, \Sigma_{z_{k/2}}^i, 3/20\right)$-BMSB condition for all $k \in [1, T]$.

The proof of Proposition 1 builds on arguments used in [14] and can be found in Appendix B. Before we state the main theorem of this section we define

$$\Delta\Lambda_i^u(t) \coloneqq \Delta A_i \sum_{s=0}^{t} A_*^s B_* \Sigma_u^{1/2} \tag{10a}$$

$$\Delta\Lambda_i^w(t) \coloneqq \Delta A_i \sum_{s=0}^{t} A_*^s \Sigma_w^{1/2} \tag{10b}$$

describing the excitations due to the control input (10a) and noise (10b) projected on the differences between A_* and A_i.

Theorem 1: Let $\{x_t\}_{t=1}^T$, $\{u_t\}_{t=1}^T$ be data collected from system (1) according to Assumption 1. Fix a failure probability $\delta \in (0,1)$. If there exists $k \in [1,T]$ such that

$$\lfloor T/k \rfloor \geq 320/3 \log(2n_x N/\delta) \tag{11a}$$

and for all $i \in [1,N]$

$$\sqrt{n_x} + n_x \leq \frac{9k\lfloor T/k \rfloor}{3200T} \Bigg(\|\Sigma_w^{-1/2} \Delta\Lambda_i^w(k/2)\|_{\mathrm{F}}^2 + n_x + \|\Sigma_w^{-1/2} \Delta B_i \Sigma_u^{1/2}\|_{\mathrm{F}}^2 + \|\Sigma_w^{-1/2} \Delta\Lambda_i^u(k/2)\|_{\mathrm{F}}^2 \Bigg), \tag{11b}$$

then the MLE (3) yields the true system θ_* with probability at least $1-\delta$, i. e., $\mathbb{P}[\hat{\theta}_T = \theta_*] \geq 1-\delta$.

Proof: The proof is articulated in three steps. First, we use the BMSB condition to show anti-concentration of $\frac{1}{T}\sum_{t=1}^T \|z_t^i\|$. Secondly, we show that $\frac{1}{T}\sum_{t=1}^T \|w_t\|$ concentrates. Finally, we combine the previous two results to obtain that $\hat{L}(\theta_*) < \hat{L}(\theta_i) \quad \forall \theta_i \neq \theta_*$ with high probability.

a) Lower bounding the empirical risk of $\theta \neq \theta_*$: Recall that Proposition 1 shows that the process $(z_t^i)_{t=1}^T$ satisfies the $\left(k, \Sigma_{z_{k/2}}^i, 3/20\right)$-BMSB condition. Note that $e_\ell \in \mathbb{S}^{n_x-1}\, \forall l \in [1,n_x]$. Thus, for some fixed $\ell \in [1,n_x]$, it follows from the BMSB condition that

$$\frac{1}{k}\sum_{t=1}^{k} \mathbb{P}\left[|[z_{s+t}^i]_\ell| \geq \sqrt{{e_\ell}^\top \Sigma_{z_{k/2}}^i e_\ell} \right] \geq \frac{3}{20},$$

where $[\cdot]_\ell$ extracts the ℓ-th element from a vector. By applying Corollary 1 (Appendix C), which is a tighter version of [14, Proposition 2.5], we obtain the anti-concentration result

$$\mathbb{P}\left[\sum_{t=1}^{T} [z_t^i]_\ell^2 \leq \frac{9k\lfloor T/k \rfloor}{3200} {e_\ell}^\top \Sigma_{z_{k/2}}^i e_\ell \right] \leq e^{-\frac{3}{320}\lfloor T/k \rfloor}. \tag{12}$$

In the following we impose $\exp(-3/320 \lfloor T/k \rfloor) \leq \frac{\delta}{2n_x N}$, which requires the burn-in time condition

$$\lfloor T/k \rfloor \geq 320/3 \log(2n_x N/\delta), \tag{13}$$

which is (11a). We consider the events

$$\mathcal{E}_\ell \coloneqq \sum_{t=1}^{T} [z_t^i]_\ell^2 \geq 9k/320 \lfloor T/k \rfloor {e_\ell}^\top \Sigma_{z_{k/2}}^i e_\ell,$$

as well as their union $\mathcal{E} \coloneqq \bigcup_{\ell=1}^{n_x} \mathcal{E}_\ell$. It follows that

$$\mathcal{E} \implies \sum_{\ell=1}^{n_x}\sum_{t=1}^{T} [z_t^i]_\ell^2 \geq \sum_{\ell=1}^{n_x} 9k/3200 \lfloor T/k \rfloor {e_\ell}^\top \Sigma_{z_{k/2}}^i e_\ell.$$

Because of (13) we have $\mathbb{P}[\mathcal{E}_\ell] \geq 1 - \frac{\delta}{2n_x N}$. Hence, using union bound arguments, we obtain

$$\begin{aligned}
\mathbb{P}\left[\sum_{t=1}^{T} \|z_t^i\|^2 \geq \sum_{\ell=1}^{n_x} 9k/3200 \lfloor T/k \rfloor {e_\ell}^\top \Sigma_{z_{k/2}}^i e_\ell \right] &= \mathbb{P}\left[\sum_{\ell=1}^{n_x}\sum_{t=1}^{T} [z_t^i]_\ell^2 \geq \sum_{\ell=1}^{n_x} 9k/3200 \lfloor T/k \rfloor {e_\ell}^\top \Sigma_{z_{k/2}}^i e_\ell \right] \\
&\geq \mathbb{P}\left[\mathcal{E}\right] \geq \prod_{l=1}^{n_x} \left(1 - \frac{\delta}{2n_x N}\right) \geq 1 - \frac{\delta}{2N}.
\end{aligned}$$

Finally, observe $\sum_{\ell=1}^{n_x} {e_\ell}^\top \Sigma_{z_{k/2}}^i e_\ell = \mathrm{Tr}\left(\Sigma_{z_{k/2}}^i\right)$ to obtain

$$\mathbb{P}\left[\sum_{t=1}^{T} \|z_t^i\|^2 \geq 9k/3200 \lfloor T/k \rfloor \,\mathrm{Tr}\left(\Sigma_{z_{k/2}}^i\right) \right] \geq 1 - \frac{\delta}{2N}. \tag{14}$$

b) Upper bounding the empirical risk of θ_0: Note that $\Delta A_0 = \Delta B_0 = 0$. Set $\zeta_t^\top = w_t^\top \Sigma_w^{-\frac{1}{2}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I)$ to obtain

$$\hat{L}(\theta_0) = \frac{1}{T}\sum_{t=1}^{T} \|w_t\|_{\Sigma_w^{-1}} = \frac{1}{T}\sum_{t=1}^{T} \zeta_t^\top \zeta_t = \frac{1}{T}\sum_{t=1}^{n_x T} \xi_t^2,$$

where $\xi_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Clearly, ξ_k^2 is sub-exponential with parameters $(4,4)$. Since the sum of sub-exponential random variables X_i with parameters (ν_i^2, α_i) is sub-exponential with parameters $(\sum_i \nu_i^2, \max_i(\alpha_i))$ [11], we have $\sum_{t=1}^{T\cdot n_x} \xi_t^2 \sim \mathrm{subExpo}\left(4Tn_x, 4\right)$. Using [11, Proposition 2.9] with $t = \sqrt{n_x}T$ after minor reformulations we obtain

$$\mathbb{P}\left[\frac{1}{T}\sum_{k=1}^{n_x T} \xi_k^2 \geq \sqrt{n_x} + n_x \right] \leq \exp(-T/8) \leq \frac{\delta}{2}, \tag{15}$$

where the last inequality is satisfied if

$$T \geq 8 \log(2/\delta). \tag{16}$$

c) Leveraging concentration and anti-concentration: First, note that the burn-in time condition (13) implies the burn-in time condition (16). Hence, from (15) we have

$$\mathbb{P}[\mathcal{E}_{\theta_0}] \coloneqq \mathbb{P}\left[\hat{L}(\theta_0) \leq \sqrt{n_x} + n_x \right] \geq 1 - \frac{\delta}{2}.$$

Because of (14) for each θ_i, $i \in [1,N]$ it holds that

$$\mathbb{P}[\mathcal{E}_{\theta_i}] \coloneqq \mathbb{P}\left[\hat{L}(\theta_i) \geq \frac{9k\lfloor T/k \rfloor}{3200T} \,\mathrm{Tr}\left(\Sigma_{z_{k/2}}^i\right) \right] \geq 1 - \frac{\delta}{2N}.$$

If $\sqrt{n_x} + n_x < \frac{9k\lfloor T/k \rfloor}{3200T} \,\mathrm{Tr}(\Sigma_{z_{k/2}}^i)$ holds $\forall i \in [1,N]$ we have

$$\mathbb{P}\left[\hat{\theta} = \theta_* \right] \geq \mathbb{P}\left[\bigcup_{i=0}^{N} \mathcal{E}_{\theta_i} \right] = \left(1 - \tfrac{\delta}{2}\right) \prod_{i=1}^{N} \left(1 - \tfrac{\delta}{2N}\right) \geq 1 - \delta,$$

where the second inequality follows from union-bound arguments. From here we can conclude the proof, by using linearity of the trace, as well as $\mathrm{Tr}(AA^\top) = \|A\|_{\mathrm{F}}^2$ to obtain

$$\begin{aligned}
\mathrm{Tr}\left(\Sigma_{z_{k/2}}^i\right) = {} & \|\Sigma_w^{-1/2} \Delta\Lambda_i^w(k/2)\|_{\mathrm{F}}^2 + n_x \\
& + \|\Sigma_w^{-1/2} \Delta B_i \Sigma_u^{1/2}\|_{\mathrm{F}}^2 + \|\Sigma_w^{-1/2} \Delta\Lambda_i^u(k/2)\|_{\mathrm{F}}^2.
\end{aligned}$$

■

Remark 1: Theorem 1 imposes no stability assumptions on the system (1). This is in contrast with non-asymptotic results for the unconstrained OLS, where a statistical inconsistency has been shown for certain classes of unstable systems [15]. Note that any T satisfying the conditions in Theorem 1 is an upper bound to the sample complexity as defined in (4). Further, the cardinality $N+1$ of the set $\mathcal{S}$ enters Theorem 1 in (11a) and (11b). Whereas the former shows the dependence $\mathcal{O}(\log N)$, the influence of N on (11b) is more subtle and depends on the particular systems being added as N increases. Note also, that Theorem 1 depends on the true systems matrices, which are unknown in practice. While data-dependent results might prove more useful from a practical perspective, the value of Theorem 1 lies in understanding the fundamental difficulty of the learning problem. To investigate the conservatism of Theorem 1 analytically, we now derive finite-sample identification lower bounds.

B. A sample complexity lower bound

In this section, we provide a sample complexity lower bound for the class of δ-stable estimation algorithms, which we define as follows.

Definition 2 (δ-stable algorithms): Consider the setup described in Section II-A. An algorithm is called δ-stable, if for all $\delta \in (0,1)$, and any θ_* and $\mathcal{S}$ there exists a finite time $\bar{T}$ s.t. for all $t \geq \bar{T}$ we have $\mathbb{P}_{\theta_*}(\hat{\theta}_t = \theta_*)$.

The notion of δ-stable algorithms excludes algorithms that yield the same estimate independently of the data observed and is inspired by (ε,δ)-locally-stable algorithms in [16].

Theorem 2: Let Assumption 1 hold. Then, for any δ-stable algorithm, all $\delta \in (0,1)$ and all $i \in [1,N]$ it holds that

$$\bar{T}\left\|\Sigma_w^{-\frac{1}{2}}\Delta B_i \Sigma_u^{1/2}\right\|_{\mathrm{F}}^2 + \sum_{s=0}^{\bar{T}-1}\left\|\Sigma_w^{-1/2}\Delta\Lambda_i^u(s)\right\|_{\mathrm{F}}^2 + \left\|\Sigma_w^{-1/2}\Delta\Lambda_i^w(s)\right\|_{\mathrm{F}}^2 \geq 2\log\left(\tfrac{1}{2.4\delta}\right). \tag{17}$$

The proof of Theorem 2 is inspired by [16] and can be found in Appendix A. Importantly, Theorem 2 can be analyzed to see which factors contribute to identifying the true system.

C. Analysis of the sample complexity bounds

Having established a sample complexity upper bound for the MLE (3) as well as an estimation algorithm independent lower bound, we can now analyze and compare Theorems 1 and 2. To this end, we will focus on three key factors that influence the identification of the true system.

a) Excitation and noise level: For simplicity consider the case where $\Sigma_w = \sigma_w^2 I$ and $\Sigma_u = \sigma_u^2 I$. In this case, according to both the upper bound (Theorem 1) and the lower bound (Theorem 2) increasing the ratio σ_u/σ_w allows for either a decrease in the number of samples T or in the failure probability δ.

b) Excitation directions: Observe that condition (11b) of the upper bound and condition (17) of the lower bound qualitatively both depend on the same three terms and can be interpreted as a signal-to-noise ratio (SNR) condition. To this end, recall that $\Delta\Lambda_i^B(t)$, $\Delta B_i \Sigma_u^{\frac{1}{2}}$ and $\Delta\Lambda_i^I(t)$ can be interpreted as the excitation of the system projected to the difference between systems θ_* and θ_i. Weighting this measure of relevant excitation with the covariance of the noise yields an effective SNR. Importantly, the directions of the excitation matter. That is, using the control input to excite the system where ΔA_i is large results in a smaller burn-in time or failure probability. Further, if the noise affects some states more than others this will also affect the lower and upper bounds.

c) Sample efficiency: Assume Σ_w and Σ_u are fixed. Then, decreasing the failure probability δ requires a larger T to satisfy the lower bound (17), i. e., the failure probability can be decreased when more samples are available. Regarding the upper bound derived in Theorem 1, the burn-in time condition (11a) shows a similar coupling between T and δ. On the other hand, the SNR-condition (11b) does only exhibit a weak dependence on T through the parameter k. An increase in T allows for a larger k which in turn enters (11b) through $\Delta\Lambda_i^B(k/2)$ and $\Delta\Lambda_i^I(k/2)$. The dependence of these quantities on k depends heavily on the stability properties of the system. As expected, the more stable A_*, the harder it is to satisfy (11b). To conclude, it has to be said that, even though the upper bound (Theorem 1) and lower bound (Theorem 2) qualitatively depend on similar quantities, there still is a substantial gap between the two. This is largely due to the leading constants in condition (11a) and (11b), which appear due to the BMSB condition needed because of the correlation in the data.

IV. Numerical Example

In the following, we investigate the results and observations of the previous sections using a numerical example.[1] To do so we consider the set $\mathcal{S} = \{(A_0,B),(A_1,B),(A_2,B)\}$, with

$$A_i = \begin{bmatrix} a_i & 0.1 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & b_i \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix},$$

where $a_0 = a_2 = 0.2$, $a_1 = 0.1$, $b_0 = b_1 = 0.5$ and $b_2 = 0.6$. We choose a small cardinality of $\mathcal{S}$ and low state space dimension since this setup suffices to make a number of key observations while maintaining clarity of exposition. Note first, that each $\theta_i \in \mathcal{S}$ has a weak coupling between x_1 and u making it hard to excite the first mode of the system. This is a structure known to make identification hard using OLS [17] and also plays a key role here. To show the influence of the directions of excitation on the identification of the true system, we conduct three numerical experiments:

Exp. 1: $\Sigma_u = \mathrm{diag}(10, 0.1)$, $\Sigma_w = 0.1I$
Exp. 2: $\Sigma_u = \mathrm{diag}(0.1, 10)$, $\Sigma_w = 0.1I$
Exp. 3: $\Sigma_u = \mathrm{diag}(10, 0.1)$, $\Sigma_w = \mathrm{diag}(10, 0.1, 0.001)$.

Table I shows the percentage of the estimates (within 1000 trials) for varying T for estimation using MLE as presented in this work as well as using OLS and projecting on the closest system in spectral norm. The results in Table I allow us to numerically show the observations made in Section III-C. Firstly, the true positive rate increases as T increases. Secondly, the different directions of excitation in

[1]The Python code for the numerical example can be accessed at: `https://github.com/col-tasas/2024-bounds-finite-set-ID`

TABLE I

ESTIMATION PERCENTAGES FOR DIFFERENT NUMBER OF SAMPLES. LOWER BOUND IS SATISFIED WITH $\boldsymbol{\delta = 0.05}$ FOR $\boldsymbol{T \geq 192}$ (EXP. 1), $\boldsymbol{T \geq 400}$ (EXP. 2) AND $\boldsymbol{T \geq 404}$ (EXP 3). UPPER BOUND IS NOT SATISFIED FOR ANY OF THE DISPLAYED EXPERIMENTS AND SAMPLE SIZES.

T	$\mathbb{P}_{\mathrm{MLE}}[\theta_0]$ ($\mathbb{P}_{\mathrm{OLS}}[\theta_0]$) in %			$\mathbb{P}_{\mathrm{MLE}}[\theta_1]$ ($\mathbb{P}_{\mathrm{OLS}}[\theta_1]$) in %			$\mathbb{P}_{\mathrm{MLE}}[\theta_2]$ ($\mathbb{P}_{\mathrm{OLS}}[\theta_2]$) in %		
	Exp 1	Exp 2	Exp 3	Exp 1	Exp 2	Exp 3	Exp 1	Exp 2	Exp 3
250	80.2 (72.3)	77.7 (71.9)	78.3 (73.9)	10.9 (16.1)	22.3 (27.0)	21.7 (23.4)	8.9 (11.6)	0.0 (1.1)	0.0 (2.7)
500	92.8 (86.8)	87.6 (84.2)	87.7 (84.7)	4.7 (7.9)	12.4 (15.3)	12.3 (12.7)	2.5 (5.3)	0.0 (0.5)	0.0 (2.6)
750	96.6 (92.9)	91.2 (87.8)	90.7 (88.4)	2.5 (3.8)	8.8 (12.1)	9.3 (9.7)	0.9 (3.3)	0.0 (0.1)	0.0 (1.9)
1000	98.5 (96.9)	95.1 (93.9)	94.8 (92.9)	1.3 (2.3)	4.9 (6.1)	5.2 (5.2)	0.2 (0.8)	0.0 (0.0)	0.0 (1.9)
1250	99.4 (99.0)	98.1 (97.3)	95.4 (93.7)	0.5 (0.7)	1.9 (2.7)	4.6 (4.7)	0.1 (0.3)	0.0 (0.0)	0.0 (1.6)

the experiments play an important role. Clearly, Exp. 1 allows for the fastest identification of the true system, both in the numerical simulation and in the lower bound. When considering Exp. 2 and Exp. 3 we can observe that θ_2 can be ruled out very quickly because it differs from the true system in the x_3 directions which has a very high SNR (either through large excitation or low noise). Since the excitation of x_1 is very low in Exp. 2 and the noise affecting x_1 is very large in Exp. 3 distinguishing between θ_0 and θ_1 takes longer than in Exp. 1. Again this can be observed both numerically and in the lower bound, even though the inputs only differ in their directionality, not their size. Interestingly, both the lower bound and the numerical results also indicate that Exp. 3 poses the hardest identification problem out of the three. Numerically it can be seen that MLE consistently outperforms OLS for our setup, reinforcing the interest in its statistical analysis. Note that the system considered is strongly damped and hence based on the discussions in Section III-C the number of samples only has a weak influence on the upper bound. Reducing the conservatism in Theorem 1 especially in this strictly stable regime remains an important open problem.

V. CONCLUSION

In this paper we provided upper and lower bounds for the sample complexity of identifying an LTI system from a finite set of candidates in absence of stability assumptions. These are the first finite sample guarantees for this setting that, albeit relevant, was not studied before. Future work includes reducing the conservatism in the upper bound, considering additional noise classes and relaxing the assumption that $\theta_* \notin \mathcal{S}$. Finally, the fact that no stability is required here (Remark 1), suggests to better understand whether this depends on the choice of the estimator made here or the finite hypothesis class.

REFERENCES

[1] C. Choirat and R. Seri, "Estimation in discrete parameter models," *Statistical Science*, vol. 27, no. 2, May 2012.

[2] D. Liberzon and A. Morse, "Basic problems in stability and design of switched systems," *IEEE Control Systems Magazine*, vol. 19, no. 5, pp. 59–70, 1999.

[3] K. Narendra and J. Balakrishnan, "Adaptive control using multiple models," *IEEE Transactions on Automatic Control*, vol. 42, no. 2, pp. 171–187, 1997.

[4] A. Rantzer, "Minimax adaptive control for a finite set of linear systems," in *Proceedings of the 3rd Conf. on Learning for Dynamics and Control*, 2021, pp. 893–904.

[5] J. Johnson and K. Omland, "Model selection in ecology and evolution," *Trends in Ecology & Evolution*, vol. 19, no. 2, pp. 101–108, 2004.

[6] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[7] S. Villar, J. Bowden, and J. Wason, "Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges," *Statistical Science*, vol. 30, no. 2, May 2015.

[8] K. Jamieson and R. Nowak, "Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting," in *48th Annu. Conf. on Information Sciences and Systems (CISS)*. J, 2014.

[9] L. Ljung, *System Identification: Theory for the User*. Prentice Hall PTR, 1999.

[10] Y. Abbasi-Yadkori, D. P., and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[11] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge university press, 2019, vol. 48.

[12] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *Foundations of Computational Mathematics*, vol. 20, no. 4, pp. 633–679, 2020.

[13] N. Matni and S. Tu, "A tutorial on concentration bounds for system identification," in *IEEE 58th Conf. on Decision and Control (CDC)*, 2019.

[14] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," in *Conf. On Learning Theory*. PMLR, 2018, pp. 439–473.

[15] T. Sarkar and A. Rakhlin, "Near optimal finite time identification of arbitrary linear dynamical systems," in *International Conf. on Machine Learning*. PMLR, 2019, pp. 5610–5618.

[16] Y. Jedra and A. Proutiere, "Sample complexity lower bounds for linear system identification," in *IEEE 58th Conf. on Decision and Control (CDC)*, 2019.

[17] A. Tsiamis and G. J. Pappas, "Linear systems can be hard to learn," in *60th IEEE Conf. on Decision and Control (CDC)*. IEEE, 2021.

[18] T. Sarkar, A. Rakhlin, and M. Dahleh, "Nonparametric system identification of stochastic switched linear systems," in *IEEE 58th Conf. on Decision and Control (CDC)*, 2019.

[19] A. Tsiamis, I. Ziemann, N. Matni, and G. J. Pappas, "Statistical learning theory for control: A finite-sample perspective," *IEEE Control Systems*, vol. 43, no. 6, pp. 67–97, 2023.

[20] I. J. Myung, "Tutorial on maximum likelihood estimation," *Journal of Mathematical Psychology*, vol. 47, no. 1, pp. 90–100, Feb. 2003.

[21] A. Garivier, P. Ménard, and G. Stoltz, "Explore first, exploit next: The true shape of regret in bandit problems," *Mathematics of Operations Research*, vol. 44, no. 2, pp. 377–399, May 2019.

APPENDIX

A. Proof of Theorem 2

Define the data observed up to time t as $\mathcal{D}_t := \{x_1, u_1, \dots x_t, u_t\}$ and the probability of the observing $\mathcal{D}_t$ under system θ as $\mathbb{P}_\theta(\mathcal{D}_t)$. Then, we define the log-likelihood ratio of the first t observations under θ_* and some $\theta_i \in \mathcal{S} \setminus \{\theta_*\}$ as $L_t = \log\left(\frac{\mathbb{P}_{\theta_*}(\mathcal{D}_t)}{\mathbb{P}_{\theta_i}(\mathcal{D}_t)}\right)$. Following the change of measurement argument in [16], we use the generalized data processing inequality [21, Lemma 1] to obtain

$$\begin{aligned}\mathbb{E}_{\theta_*}[L_t] &= \mathrm{KL}(\mathbb{P}_{\theta_*}(\mathcal{D}_t) \| \mathbb{P}_{\theta_i}(\mathcal{D}_t)) \\ &\geq \sup_{\mathcal{E} \in \mathcal{F}_t} \mathrm{kl}(\mathbb{P}_{\theta_*}(\mathcal{E}) \| \mathbb{P}_{\theta_i}(\mathcal{E})),\end{aligned}$$

where $\mathrm{kl}(x\|y)$ is the KL-divergence of two Bernoulli distributions of means x and y, respectively. Since we analyze δ-stable algorithms we define the event $\mathcal{E} \coloneqq \{\hat{\theta}_t = \theta_*\}$ s.t. consequently $\mathbb{P}_{\theta_*}(\mathcal{E}) \geq 1-\delta$ and $\mathbb{P}_{\theta_i}(\mathcal{E}) \leq \delta$ and hence

$$\mathrm{kl}(\mathbb{P}_{\theta_*}(\mathcal{E})\|\mathbb{P}_{\theta_i}(\mathcal{E})) \geq (2\delta-1)\log\left(\frac{1-\delta}{\delta}\right) \geq \log(1/2.4\delta).$$

Further, we follow [16, Section IV.A] to obtain

$$\begin{aligned}\mathbb{E}_{\theta_*}[L_t] &= \frac{1}{2}\mathbb{E}_{\theta_*}\left[\sum_{s=0}^{t-1}[\star][\star]\Sigma_w^{-1}\begin{bmatrix}\Delta A_i & \Delta B_i\end{bmatrix}\begin{bmatrix}x_s\\u_s\end{bmatrix}\right] \\ &= \frac{1}{2}\operatorname{Tr}\left([\star]\Sigma_w^{-1}\begin{bmatrix}\Delta A_i & \Delta B_i\end{bmatrix}\sum_{s=0}^{t-1}\mathbb{E}_{\theta_*}\left[\begin{bmatrix}x_s\\u_s\end{bmatrix}[\star]\right]\right),\end{aligned}$$

where in the last step we used the fact $\mathbb{E}\left[X^\top A X\right] = \operatorname{Tr}(A\mathbb{E}\left[XX^\top\right])$. Note that up to this point, we have not used that $u \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\Sigma_u)$. By this assumption, we can write $\mathbb{E}_{\theta_*}\left[[\star]\begin{bmatrix}x_s & u_s\end{bmatrix}\right] = \operatorname{diag}(\Sigma_{x_t},\Sigma_u)$, where Σ_{x_t} is the t-step controllability Gramian defined in (7). Using $\operatorname{Tr}(ABC) = \operatorname{Tr}(BCA)$ with $A = \begin{bmatrix}\Delta A_i & \Delta B_i\end{bmatrix}^\top \Sigma_w^{-\frac{1}{2}}$, $B = A^\top$ and $C = \sum_{s=0}^{t-1}\operatorname{diag}(\Sigma_{x_t},\Sigma_u)$ we finally obtain

$$\mathbb{E}_{\theta_*}[L_t] = \operatorname{Tr}\left([\star]\left(\sum_{s=0}^{t-1}[\star]\Sigma_u\Delta B_i^\top + [\star]\Sigma_{x_s}\Delta A_i^\top\right)\Sigma_w^{-\frac{1}{2}}\right)$$

and hence for any δ-stable algorithm

$$\begin{aligned}\operatorname{Tr}\left(\Sigma_w^{-\frac{1}{2}}\left(\sum_{s=0}^{\bar{T}-1}[\star]\Sigma_u\Delta B_i^\top + [\star]\Sigma_{x_s}\Delta A_i^\top\right)\Sigma_w^{-\frac{1}{2}}\right) \\ \geq 2\log(\frac{1}{2.4\delta}).\end{aligned}$$

Finally, we similarly as in the proof of Theorem 2 obtain

$$\begin{aligned}&\operatorname{Tr}\left(\Sigma_w^{-\frac{1}{2}}\left(\sum_{s=0}^{\bar{T}-1}\Delta B_i\Sigma_u\Delta B_i{}^\top + \Delta A_i\Sigma_{x_s}\Delta A_i^\top\right)\Sigma_w^{-\frac{1}{2}}\right) \\ &= \bar{T}\left\|\Sigma_w^{-\frac{1}{2}}\Delta B_i\Sigma_u^{\frac{1}{2}}\right\|_{\mathrm{F}}^2 + \sum_{s=0}^{\bar{T}-1}\sum_{k=0}^{s-1}\left\|\Sigma_w^{-\frac{1}{2}}\Delta A_i A^k B\Sigma_u^{\frac{1}{2}}\right\|_{\mathrm{F}}^2 \\ &\qquad + \left\|\Sigma_w^{-\frac{1}{2}}\Delta A_i A^k\Sigma_w^{\frac{1}{2}}\right\|_{\mathrm{F}}^2.\end{aligned}$$

B. Proof of Proposition 1

Recall the definition of the random variable z_t^i (6) and its distribution (8). For some $v \in \mathbb{S}^{n_x-1}$ we have $\langle v, z_{s+t}^i\rangle|\mathcal{F}_s \sim \mathcal{N}\left(\langle v, \Delta A_i A^t x_s\rangle, v^\top\Sigma_{z_t}^i v\right)$. Now consider some $k' \leq t$ s.t.

$$\begin{aligned}&\mathbb{P}\left[|\langle v, z_{s+t}^i\rangle| \geq \sqrt{v^\top\Sigma_{z_{k'}}^i v}|\mathcal{F}_s\right] \\ &\qquad = \mathbb{P}\left[|\langle v, z_t^i\rangle| \geq \sqrt{v^\top\Sigma_{z_{k'}}^i v}\right] \\ &\qquad \geq \mathbb{P}\left[|\langle v, z_t^i\rangle| \geq \sqrt{v^\top\Sigma_{z_t}^i v}\right], \qquad (18)\end{aligned}$$

where the first step follows since the distributions are equal and the inequality follows from $\Sigma_{z_t}^i \succeq \Sigma_{z_{k'}}^i$ for $k' \leq t$. Defining $\zeta_t^i = \langle v, z_t^i - \Delta A_i A^t x_0\rangle \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, v^\top\Sigma_{z_t}^i v)$ yields

$$\begin{aligned}&\mathbb{P}\left[|\langle v, z_t^i\rangle| \geq \sqrt{v^\top\Sigma_{z_t}^i v}\right] \\ &\qquad = \mathbb{P}\left[|\zeta_t^i + \langle v, \Delta A_i A^t x_0\rangle| \geq \sqrt{v^\top\Sigma_{z_t}^i v}\right] \\ &\qquad \geq \mathbb{P}\left[|\zeta_t^i| \geq \sqrt{v^\top\Sigma_{z_t}^i v}\right] \geq \frac{3}{10},\end{aligned}$$

where the last inequality follows from the fact that for any $\xi \sim \mathcal{N}(0,\sigma^2)$ we have $\mathbb{P}[|\xi| \geq \sigma] \geq 3/10$. It follows that

$$\begin{aligned}&\frac{1}{k}\sum_{t=1}^{k}\mathbb{P}\left[|\langle v, z_{s+t}^i\rangle| \geq \sqrt{v^\top\Sigma_{z_{k'}}^i v}|\mathcal{F}_s\right] \\ &= \frac{1}{k}\sum_{t=1}^{k}\mathbb{P}\left[|\langle v, z_t^i\rangle| \geq \sqrt{v^\top\Sigma_{z_{k'}}^i v}\right] \\ &\geq \frac{1}{k}\sum_{t=k'}^{k}\mathbb{P}\left[|\langle v, z_t^i\rangle| \geq \sqrt{w^\top\Sigma_{z_t}^i w}\right] \geq \frac{3}{10}\frac{k-k'+1}{k}.\end{aligned}$$

The result then follows by choosing $k' = \frac{1}{2}k$.

C. Using the BMSB condition to show anti-concentration

If a sequence $(z_t)_{t\geq 0}$ satisfies the BMSB condition, anti-concentration can be shown. In the following, we provide a milder version of [14, Proposition 2.5], in which the probability bound scales with p instead of p^2.

Corollary 1: Suppose that $(z_1,\ldots,z_T) \in \mathbb{R}^T$ satisfies the (k,ν,p)-BMSB condition. Then

$$\mathbb{P}\left[\sum_{t=1}^{T} z_t^2 \leq \frac{\nu^2 p^2}{8}k\lfloor T/k\rfloor\right] \leq \exp\left(-\frac{\lfloor T/k\rfloor p}{16}\right).$$

Proof: The proof of Corollary 1 builds on the original work, thus we only provide a sketch here. Set $S = \lfloor T/k\rfloor$ and

$$B_j = \mathbb{I}\left[\sum_{t=1}^{k} z_{jk+t}^2 \geq \frac{\nu^2 pk}{2}\right] \quad \forall j \in [0, S-1].$$

Using the same arguments as in the original proof we obtain

$$\mathbb{P}\left[\sum_{t=1}^{T} z_t^2 \leq \frac{\nu^2 p^2}{8}kS\right] \leq \inf_{\lambda\leq 0} e^{-\lambda S\frac{p}{4}}\mathbb{E}\left[e^{\lambda\sum_{j=0}^{S-1}B_j}\right]. \quad (19)$$

Inserting the optimizer $\lambda_* = \log(\frac{1-\frac{p}{2}}{2-\frac{p}{2}})$ of (19), yields

$$\begin{aligned}\mathbb{P}\left[\sum_{t=1}^{T} z_t^2 \leq \frac{\nu^2 p^2}{8}kS\right] &\leq \left(\left(\frac{2-p}{4-p}\right)^{-p/4}\left(1-\frac{p}{4-p}\right)\right)^S \\ &\leq \left(2^{p/4}\left(1-\frac{p}{4-p}\right)\right)^S \leq e^{-\frac{S}{4}p\left(\frac{p}{4}+1-\log(2)\right)} \leq e^{-\frac{S}{16}p},\end{aligned}$$

where the last step follows from $p > 0$ and $1-\log(2) > 1/4$. ∎