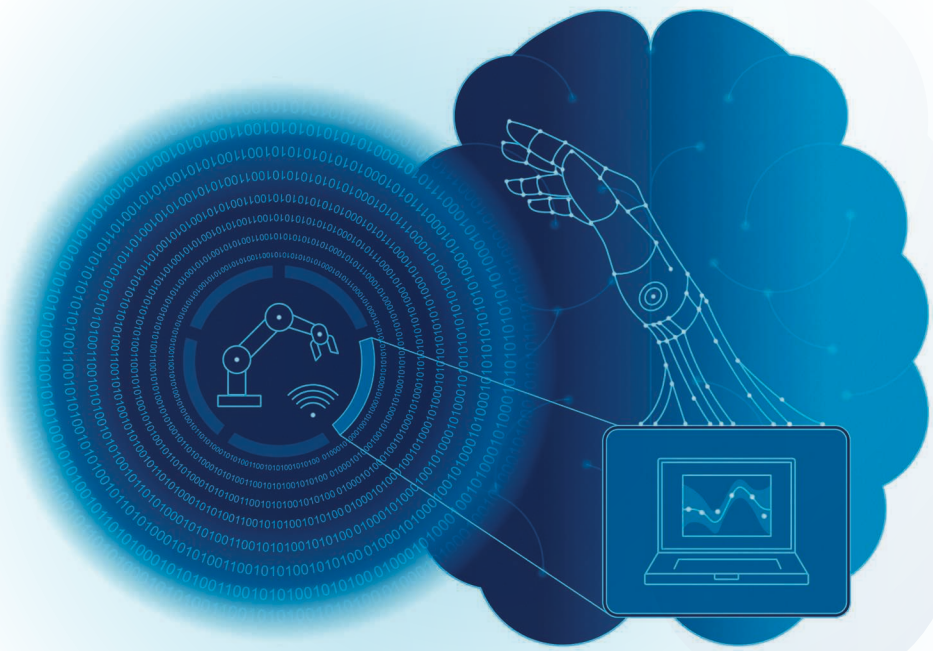
Statistical Learning Theory for Control

A FINITE-SAMPLE PERSPECTIVE



ANASTASIOS TSIAMIS, INGVAR ZIEMANN, NIKOLAI MATNI, and GEORGE J. PAPPAS

Learning algorithms have become an integral component to modern engineering solutions. Examples range from self-driving cars and recommender systems to finance and even critical infrastructure, many of which are typically under the purview of control theory. While these algorithms have already shown tremendous promise in certain applications [1], there are considerable challenges, in particular, with respect to guaranteeing safety and gauging fundamental limits of operation. Thus, as we integrate tools from machine learning into our systems, we also require an integrated theoretical understanding of how they operate in the presence of dynamic and system-theoretic phenomena. Over the past few years, intense efforts toward this goal—an integrated theoretical understanding of learning, dynamics, and control—have been made. While much work remains to be done, a relatively clear and complete picture has begun to emerge for (fully observed) linear dynamical systems. These systems already allow for reasoning about concrete failure modes, thus helping to indicate a path forward. Moreover, while simple at a glance, these systems can be challenging to analyze. Recently, a host of methods from learning theory and high-dimensional statistics, not typically in the control-theoretic toolbox, have been introduced to our community.

Digital Object Identifier 10.1109/MCS.2023.3310345
Date of current version: 14 November 2023

This tutorial survey serves as an introduction to these results for learning in the context of unknown linear dynamical systems (see "Summary"). We review the current state of the art and emphasize which tools are needed to arrive at these results. Our focus is on characterizing the sample efficiency and fundamental limits of learning algorithms. Along the way, we also delineate a number of open problems. More concretely, this article is structured as follows. We begin by revisiting recent advances in the finite-sample analysis of system identification. Next, we discuss how these finite-sample bounds can be used downstream to give guaranteed performance for learning-based offline control. The final technical section discusses the more challenging online control setting. Finally, in light of the material discussed, we outline a number of future directions.

FINITE-SAMPLE ANALYSIS OF SYSTEM IDENTIFICATION

In linear system identification, the goal is to recover the model of an *unknown* system of the form

$$\begin{aligned} x_{t+1} &= A_\star x_t + B_\star u_t + w_t \\ y_t &= C_\star x_t + v_t \end{aligned} \tag{1}$$

where $x_t \in \mathbb{R}^{d_x}$ represents the state, $y_t \in \mathbb{R}^{d_y}$ represents the observations, $u_t \in \mathbb{R}^{d_u}$ is the control signal, and $w_t \in \mathbb{R}^{d_x}$, $v_t \in \mathbb{R}^{d_y}$ are the process and measurement noises, respectively. The question we answer in this section is, *How many samples are needed to guarantee that the system identification error is small*? We make this question more formal by introducing the notion of *sample complexity*. Prior to doing so, we establish the statistical learning framing of the problem. While many of the results presented in the following sections can be extended to more general noise models, we keep the exposition simple by focusing on Gaussian noise models. In particular, we assume that both the process noise w_t and measurement noise v_t are independent identically distributed (i.i.d.) zero-mean Gaussians with covariance matrices Σ_w and Σ_v, respectively, and that these process are all mutually independent of one another. Similarly, we let the initial state x_0 be a zero-mean Gaussian, with covariance Γ_0, and independent of the process and measurement noise. We denote the covariance of the state x_t at time t by $\Gamma_t \triangleq \mathbf{E} x_t x_t^\top$. Here and in the sequel, the state parameters $(A_\star, B_\star, C_\star) \in \mathbb{R}^{d_x \times (d_x + d_u + d_y)}$ are unknown. The goal of the system identification problem is to recover the a priori unknown model of system (1) from finite input–output samples $\{(y_i, u_i)\}_{i=1}^{N_{\text{tot}}}$, where N_{tot} is the total number of samples. Thus, this is an *offline learning* problem. The data can come from a single trajectory of length T (that is, $N_{\text{tot}} = T$) or from N_{traj} multiple independent trajectories with horizon T; that is, $N_{\text{tot}} = TN_{\text{traj}}$. While the learning task is to recover the state-space parameters $\theta_\star \triangleq (A_\star, B_\star, C_\star, \Sigma_w, \Sigma_v)$ of (1) using these data, the state-space representation of system (1) is, in general, not unique. Hence, we instead seek to recover one such representation or a function $f(\theta_\star)$ of the underlying true parameters $\theta_\star$. To streamline the exposition, we focus on the single-trajectory case $N_{\text{tot}} = T$. A more refined analysis can be used when samples are drawn from multiple trajectories to yield similar conclusions [2] but under weaker stability-type assumptions. Let the identification algorithm $\mathcal{A}$ be a (measurable) function that takes as an input the horizon T and the data $\{(y_0, u_0), (y_1, u_1), \ldots, (y_T, u_T)\}$ and returns an estimate $\hat{f}_T$ of the desired system quantity $f(\theta_\star)$. In some settings, the algorithm $\mathcal{A}$ may also encompass an exploration policy, that is, the choice of control inputs u_t used during the data collection phase. The goal of the exploration policy is to excite the system in a way that maximizes the "richness" of the data, that is, how much information the data carry about the underlying system. Formally, we define an exploration policy π to be a sequence of (measurable) functions $\pi = \{\pi_t\}_{t=0}^{\infty}$, where every function π_t maps previous output–input values $y_0, \ldots, y_t$, $u_0, \ldots, u_{t-1}$ and potentially an auxiliary randomization signal to the new input u_t. This definition encompasses both closed- and open-loop policies—in the latter case, the exploration policy is a function only of the auxiliary randomness. We can now define the notion of sample complexity. Let $\mathbf{P}_{\theta_\star, \pi}$ denote the probability distribution of the input–output data for the system (1) defined by parameters $\theta_\star$ evolving under the exploration policy π.

Summary

This tutorial survey provides an overview of recent advances in statistical learning theory relevant to control and system identification featuring nonasymptotic methods. While there has been substantial progress across all areas of control, the theory is most well developed when it comes to linear system identification and learning for the linear quadratic regulator, which are the focus of this article. From a theoretical perspective, much of the work underlying these advances has been in adapting tools from modern high-dimensional statistics and learning theory. While highly relevant to control theorists interested in integrating tools from machine learning, the foundational material has not always been easily accessible. To remedy this, we provide a self-contained presentation of the relevant material, outlining all the key ideas and offering an overview of the technical machinery that underpins recent results. We also present a number of open problems and future directions.

Sample Complexity

Fix a class $\mathscr{C}$ of systems of the form (1) and a norm $\|\cdot\|$. Let $f(\theta_\star)$ be the system quantity to be identified. Fix an identification algorithm $\mathcal{A}$ with an exploration policy π. Pick an accuracy parameter ε and a failure probability $\delta \in (0,1)$. Let $\hat{f}_T$ be the system identification output under the algorithm $\mathcal{A}$. Then, the sample complexity N_c of learning f given the class $\mathscr{C}$, the algorithm $\mathcal{A}$, and the policy π is the minimum $N_c = N_c(\varepsilon, \delta, C, \mathcal{A}, \pi)$ such that

$$\sup_{\theta_\star \in \mathscr{C}} \mathbf{P}_{\theta_\star, \pi}\left(\left\| f(\theta_\star) - \hat{f}_T \right\| \geq \varepsilon\right) \leq \delta \quad \text{if } T \geq N_c(\varepsilon, \delta, \mathscr{C}, \mathcal{A}, \pi). \tag{2}$$

We say that a class of systems $\mathscr{C}$ is learnable if there exist an algorithm $\mathcal{A}$ and a policy π such that for any $\varepsilon > 0, \delta \in (0,1)$, the sample complexity N_c is finite.

In the case of multiple trajectories, we can replace T with N_{tot} in the preceding definition. We can also define algorithm-independent and/or policy-independent sample complexity by considering the minimum N_c over all possible algorithms/policies. By choosing $\mathscr{C}$ to be a neighborhood around some system $\theta_\star$, we can also define local instance-specific sample complexities; see, for example, [5]. Note that for the sample complexity to be nontrivial, the algorithm should perform well across all possible $\theta_\star \in \mathscr{C}$, which is what the supremum over $\mathscr{C}$ achieves in (2). Otherwise, we can construct trivial algorithms that overfit to a specific system and fail to identify any other system in the class. Note that one often encounters ranges of T and δ for which the sample complexity dependency on ε behaves poorly. Typically, this is due to transient phenomena. For instance, in a d-dimensional linear regression problem, the design matrix can be nearly singular if we have too few measurements (for example, if fewer than d independent measurements are available). Informally, for a fixed δ, one typically refers to the smallest sample size T such that there exists a finite (or meaningful) sample complexity at accuracy ε as the *burn-in time*. The burn-in for linear system identification is given in (9).

From Asymptotics to Finite-Sample Guarantees

Before we proceed, let us take a step back and briefly discuss the historical development of system identification from a mathematical methods perspective. Clearly, the statistical analysis of system identification algorithms has a long history [6]. Until recently, this line of work has emphasized providing guarantees for system identification algorithms in the *asymptotic regime* [7], [8], [9], [10], [11], in which the number of collected samples tends to infinity. The main focus of asymptotic analysis has been to establish consistency, that is, the convergence of the estimated system parameters to the ground truth (as modeled). Typically, this is achieved if certain *persistency of excitation* (*PE*) conditions hold [12]. Asymptotic tools can also go beyond consistency and provide convergence rates. Standard tools for characterizing such rates are the law of the iterated logarithm (LIL) and the central limit theorem (CLT); see [14] for a detailed exposition of both techniques. Nevertheless, even the more advanced techniques (that is, the LIL and the CLT) hold only as the number of samples tend to infinity.

Toward a Finite-Sample Analysis

Early work on the nonasymptotic analysis of system identification appeared in the 1990s [14], [15], [16], [17], [18] and 2000s [19], [20]. The setting of [14] and [15] focuses on worst-case noise, which is different from the statistical setting considered in this article. In [16], approximate expressions for the finite-time identification error variance are given. We cannot derive sample complexity guarantees directly from [16]; the expressions therein are not directly computable in our setting (they require exact computation of expectations), and they do not characterize the finite-sample distribution of the identification error and how it depends on the number of samples. The statistical learning setting was first studied in [18], [19], and [20], where guarantees are typically given for the prediction error of the learned model. Moreover, the guarantees rely heavily on having a mixing, that is, a stable, process. As we soon see, in many settings, mixing is not required, and in fact, faster mixing systems can be harder to learn—at least when it comes to parameter recovery [4]. Following papers by Abbasi-Yadkori and Szepesvári [21] and Dean et al. [22], there has been a resurgence of interest in using finite-data tools for system identification and controls. This is partially motivated by recent advances in high-dimensional probability [3] and statistics [23], which provide us with new powerful tools and allow us to bypass asymptotic reasoning.

Why Do We Need Finite-Sample Guarantees?

In principle, our view is that both asymptotic and nonasymptotic methods are useful for both control and learning theorists to have in their toolbox. On the one hand, a careful asymptotic analysis can provide sharp bounds and give a clear picture of some key quantities involved in the problem at hand. However, in reality, all data are finite, and asymptotic bounds are heuristics, albeit often sharp if the sample size is large enough. On the other hand, nonasymptotic analysis is often more appropriate to carefully delineate notions, such as transient phenomena (for example, burn-in times) and failure probabilities; see "What Do Finite-Sample Methods Bring?" We gain a more detailed qualitative characterization of learning difficulty, often at the expense of sharpness in the asymptotic regime. For instance, the question, How many samples do we need to stabilize an unknown linear system with a certainty-equivalent (CE) linear quadratic regulator (LQR) controller? is necessarily answered using finite-sample methods. Being able to combine these sometimes distinct styles of analysis gives us a richer understanding of the dynamic phenomena under consideration.

Many datasets are high dimensional, with the number of explanatory variables not necessarily being small in proportion to the number of samples collected; for example, the state dimension d_x might be of the same order as T. In this case, asymptotic bounds with fixed dimension d_x are not always meaningful, while finite-sample guarantees still hold. Examples from systems theory for when this may be relevant include large networked systems and autoregressions of unknown order. An insightful discussion of this matter from a statistics perspective is held by Wainwright [23, Ch. 1]. From the perspective of a control theorist, obtaining sample complexity bounds as a function of system-theoretic parameters, for example, the system dimension, controllability Gramian, and stability radius, could be very useful. Finite-sample bounds can be qualitatively informative about learning difficulty and what can go wrong with it. That is, they provide us with tools to answer questions like, Which systems are hard to learn? How does the controllability structure affect learnability? Which algorithms are optimal? Naturally, some of these questions can also be answered using asymptotic tools. Nonetheless, we believe that a finite-sample approach offers a new perspective, enabling us to even pose new questions; see, for instance, the open problems in the following. Learning

What Do Finite-Sample Methods Bring?

Consider an *unknown* scalar system

$$x_{t+1} = a_\cdot x_t + w_t \tag{S1}$$

where $|a_\cdot| < 1$, w_t is independent identically distributed and mean-zero Gaussian with variance one. Assume that our goal is to recover the unknown scalar $a_\cdot$ from single-trajectory data $(x_0, \ldots, x_T)$. One of the simplest algorithms is to minimize the squared prediction errors

$$\hat{a}_T = \operatorname*{argmin}_a \sum_{t=1}^{T} (x_t - a x_{t-1})^2.$$

Given the stochastic nature of the data, the least-squares estimate $\hat{a}_T$ will fluctuate around the "true" value $a_\cdot$. Both asymptotic and nonasymptotic methods aim to characterize the statistical variability of the error $\hat{a}_T - a_\cdot$. One of the most powerful asymptotic tools is establishing asymptotic normality, that is, a time series version of the central limit theorem (CLT). For this particular scalar system, Mann and Wald [S1] proved that as the number of samples approaches infinity $T \to \infty$, the estimation error is asymptotically normal:

$$\sqrt{T}\,(\hat{a}_T - a_\cdot) \Rightarrow \mathcal{N}(0, 1 - a_\cdot^2)$$

where $\Rightarrow$ denotes convergence in the distribution and $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution, with mean μ and variance σ^2. This result can give us the exact distribution of the estimation error in the asymptotic regime. However, being an asymptotic result, it holds only approximately under finite samples; the approximation error cannot be ignored. Hence, some questions remain unanswered: What is the distribution of the error under finite samples? What is the transient behavior? We can partially answer these questions by applying the nonasymptotic tools reviewed in this survey. In particular, by following the arguments in the "Sample Complexity Upper Bounds" section [see (9) and (17)], we can establish a finite-sample tail bound of the form

$$\mathbf{P}(|\hat{a}_T - a_\cdot| \geq \varepsilon) \leq \delta$$

for a large enough sample size

$$T \geq \max\left\{T_{\text{burn-in}}, c\,\frac{1 - a_\cdot^2}{\varepsilon^2} \log \frac{1}{\delta}\right\}$$

where ε controls the accuracy of identification and δ controls the confidence. The constant c is a so-called universal constant; that is, it takes just a numerical value and is independent of system parameters, confidence, and accuracy. The burn-in time $T_{\text{burn-in}}$ captures the complexity of transient phenomena, for example, the minimum time until we achieve persistency of excitation (excitation of all modes of the system). It typically depends on the desired confidence and the size of the system, that is, its state dimension d_x. For the simple scalar system (S1), we can take $T_{\text{burn-in}} = c' \log 1/\delta$, where c' is another universal constant. For nonscalar systems, we can multiply the preceding burn-in time with the state dimension d_x. While we did not fully characterize the finite-sample distribution of the estimation error, we managed to characterize the tail probabilities. For example, we have a $\log 1/\delta$ term in the required number of samples, which is sharp. This was not possible before by applying only asymptotic tools. For example, CLT approximation results, such as the Berry–Esseen bound, provide a conservative characterization of tail probabilities; see [3, Ch 2.1] for a detailed explanation. We can generalize finite-sample bounds to the case $a_\cdot \geq 1$ (not presented in this sidebar) when the system does not converge to a steady-state distribution. We can also generalize the bounds to the case vector-valued systems of high dimensions $d_x > 1$, as presented later. In fact, we can even allow the state dimension d_x to increase with the number of samples T, which is not covered by the CLT. A downside of finite-sample bounds is that we lose sharpness in the asymptotic regime. In particular, the universal constants c, c' (see [4] for exact expressions) are typically large numerical values, much larger than the ones that we would obtain from a heuristic application of the CLT.

REFERENCE

[S1] H. B. Mann and A. Wald, "On the statistical treatment of linear stochastic difference equations," *Econometrica, J. Econometric Soc.*, vol. 11, no. 3/4, pp. 173–220, Jul./Oct. 1943, doi: 10.2307/1905674.

control systems under finite samples is also interesting from the perspective of a machine learning theorist. While the setting of learning under finite, independent, or weakly dependent (mixing) data has been studied extensively, new challenges arise in control systems, where the data are not only dependent but also affected by control inputs. Some questions that are of interest are, When is learning under dependent data as easy as learning under independent data? Is mixing required? What is the tradeoff between exploration and exploitation? Finally, a goal of this survey is to establish a common language among control theorists, learning theorists, and statisticians. Machine learning theory has, in principle, been nonasymptotic from the outset, and modern statistics has very much moved in this direction. Meanwhile, the classical literature of system identification and adaptive control relies, more often than not, on asymptotic tools. A common language facilitates an exchange of ideas that is likely to benefit all three fields. Besides, machine learning, statistics, and control theory share common research agendas and often seek to tackle the same problems.

Asymptotic Notation

In this article, we sometimes use the asymptotic notation O, Θ, Ω to simplify the presentation. This does not imply that our statements are asymptotic. For example, the statement $f(T) = O(g(T))$ $[f(T) = \Omega(g(T))]$ can be replaced by statements of the form "there exists universal positive constant $c > 0$ such that $f(T) \le cg(T)$ $[f(T) \ge cg(T)]$, for $T \ge T_{\text{burn-in}}$," where a universal constant takes just a numerical value and is independent of system and algorithmic parameters. Exact finite-time expressions for $g(T)$, c, $T_{\text{burn-in}}$ are given either here, for example, see (18), or in the respective articles. The statement $f(T) = \Theta(g(T))$ is equivalent to $f(T) = O(g(T))$, $f(T) = \Omega(g(T))$ holding simultaneously. Finally, the $\tilde{O}$ notation ignores polylogarithmic terms; for example, $f(T) = \tilde{O}(g(T))$ is equivalent to f(T) = O(g(T)poly (logT)), where "poly" denotes some arbitrary polynomial function of fixed degree.

Fully Observed Systems

Let us now return to the technical task at hand: to provide a finite-sample analysis of system identification. Recall that we focus on the single-trajectory case $N_{\text{tot}} = T$. We start by analyzing the simplest system identification problem, namely, the case of fully observed systems with $C_\star = I$ and $\Sigma_v = 0$, yielding direct state measurements $y_t = x_t$, $t \le T$. We focus only on the identification of $A_\star, B_\star$, but the same techniques could be applied for the estimation of the covariance Σ_w. For this reason, abusing the notation introduced in the preceding, we denote $\theta_\star = (A_\star, B_\star)$, $f(\theta) = \theta$. Given the data $\{(x_0, u_0), \ldots, (x_T, u_T)\}$, a natural way to obtain an estimate of the system matrices is to employ the least-squares algorithm

$$\hat{\theta}_T \triangleq (\hat{A}_T, \hat{B}_T) \in \operatorname{argmin}_{A,B} \sum_{t=0}^{T-1} \left\| x_{t+1} - Ax_t - Bu_t \right\|_2^2. \tag{4}$$

After some algebraic manipulations, we can verify that

$$\hat{\theta}_T - \theta_\star = \left(\sum_{t=0}^{T-1} w_t [x_t^\top \;\; u_t^\top] \right) \left(\sum_{t=0}^{T-1} \begin{bmatrix} x_t \\ u_t \end{bmatrix} [x_t^\top \;\; u_t^\top] \right)^{-1} \tag{5}$$

provided that the matrix inverse on the right-hand side of (5) exists. We characterize the sample complexity of the least-squares estimator (5) by establishing bounds on the operator norm $\| \hat{\theta}_T - \theta_\star \|_{\text{op}}$. It is possible to provide similar guarantees for the Frobenius norm, but the dimensional factors differ slightly. The techniques presented in the following can be applied to open-loop nonexplosive systems when all the eigenvalues of matrix $A_\star$ are inside or on the unit circle; that is, $\rho(A_\star) \le 1$, where $\rho(A_\star)$ denotes the spectral radius. We also assume that the open-loop inputs are i.i.d. zero-mean Gaussians with $\mathbf{E} u_t u_t^\top = \sigma_u^2 I$ for some $\sigma_u > 0$. We discuss generalizations later on. To simplify the exposition, we also assume that the noise is full rank; that is, $\Sigma_w \succ 0$. This implies that the noise directly excites all system states directly, making PE easier to establish. We can also obtain PE for indirectly excited systems as long as the controllability structure of the system is well-defined [24]. Finally, we assume that the system starts from the fixed initial condition $x_0 = 0$, and hence, the initial state covariance is $\Gamma_0 = 0$. The following terms will be useful in the analysis of the least-squares algorithm:

$$S_T \triangleq \sum_{t=0}^{T-1} w_t [x_t^\top \;\; u_t^\top], \; V_T \triangleq \sum_{t=0}^{T-1} \begin{bmatrix} x_t \\ u_t \end{bmatrix} [x_t^\top \;\; u_t^\top]. \tag{6}$$

Using the preceding notation, we can break the least-squares error into two separate terms:

$$\| \hat{\theta}_T - \theta_\star \|_{\text{op}} \le \underbrace{\left\| S_T V_T^{-1/2} \right\|_{\text{op}}}_{\text{Self-normalized term}} \underbrace{\left\| V_T^{-1/2} \right\|_{\text{op}}}_{\text{PE term}}$$

where $V_T^{-1/2}$ denotes a symmetric positive definite matrix such that $V_T^{-1/2} V_T^{-1/2} = V_T^{-1}$. To obtain sample complexity bounds for the least-squares algorithm, we need to analyze both terms. The self-normalized term captures the contribution of the noise to the least-squares error. The PE term captures PE, that is, the richness of the data. The richer the data, the larger the magnitude of the eigenvalues of the Gram matrix V_N, leading to a smaller identification error.

Persistency of Excitation (PE)

If the collected trajectory data are rich enough, that is, if all modes of the system are excited, then the Gram matrix V_T defined in (6) is both invertible and well-conditioned. In particular, if $\lambda_{\min}(V_T)$ grows unbounded with T, we say that PE holds. Moreover, the smallest eigenvalue of V_T captures the direction of the system that is the most difficult to excite. Recall that $\Gamma_t = \mathbf{E} x_t x_t^\top$ is the covariance of the state. Under i.i.d. white inputs, we can compute

$$\Gamma_t = \sum_{k=0}^{t-1} A^k (\sigma_u^2 BB^\top + \Sigma_w)(A^\top)^k, \Gamma_0 = 0. \tag{7}$$

Since the state is driven by both exogenous inputs and noise, both factors appear in the state covariance. By the definition of the Gram matrix V_T,

$$\mathbf{E}V_T = \begin{bmatrix} \sum_{t=0}^{T} \Gamma_t & 0 \\ 0 & \sigma_u^2 TI \end{bmatrix}.$$

Note that Γ_t is increasing in the positive semidefinite cone since $\Gamma_0 = 0$. It is easy to show that the expected Gram matrix $\mathbf{E}V_T$ is invertible and well-conditioned; that is, its eigenvalues increase with time T. For example, we can choose a $\tau > 0$ such that $\Gamma_\tau \succ 0$. Then, by monotonicity, $\sum_{t=0}^{T} \Gamma_t \succeq (T-\tau)\Gamma_\tau$. The main technical difficulty is to control the difference between the Gram matrix and its expectation $\| V_T - \mathbf{E}V_T \|$. Such a task might be possible in the case of strictly stable systems $\rho(A_\cdot) < 1$ by using concentration inequalities and mixing arguments. However, this approach gives sample complexity bounds that explode as $\rho(A_\cdot)$ approaches one: two-sided concentration necessitates stability. Instead, we appeal to *small-ball* techniques [25]. Rather than bounding the difference between V_T and its expectation, we seek only to obtain a one-sided lower bound. The name small-ball refers to the fact that the distribution of $\lambda_{\min}(V_T)/T$ is not concentrated in a neighborhood of the origin—it exhibits anticoncentration. Define the extended covariance matrix

$$\tilde{\Gamma}_t \triangleq \mathbf{E}\begin{bmatrix} x_t \\ u_t \end{bmatrix}\begin{bmatrix} x_t^\top & u_t^\top \end{bmatrix} = \begin{bmatrix} \Gamma_t & 0 \\ 0 & \sigma_u^2 I \end{bmatrix}.$$

Choose a time index $\tau > 0$. Invoking the small-ball methods described in "Persistency of Excitation and Small-Ball Bounds," it is possible to show that with a probability of at least $1-\delta$,

$$V_T \} \succeq c\tau \left\lfloor \frac{T}{\tau} \right\rfloor \tilde{\Gamma}_{\lfloor \tau/2 \rfloor} \tag{8}$$

where c is a universal constant, provided that we have a large enough number of samples:

$$T \geq \tau O\left((d_{\mathsf{x}} + d_{\mathsf{u}}) \log \frac{d_{\mathsf{x}} + d_{\mathsf{u}}}{\delta} + \log \frac{\det \tilde{\Gamma}_T}{\det \tilde{\Gamma}_{\lfloor \tau/2 \rfloor}} \right). \tag{9}$$

Persistency of Excitation and Small-Ball Bounds

Let $z_t \in \mathbb{R}^{d_z}$, $t \geq 0$ be a stochastic process adapted to a filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. Let the Gram matrix be

$$V_T = \sum_{t=0}^{T} z_t z_t'.$$

The process z_t is persistently exciting with a probability of at least $1-\delta$ if there exist c, $T_0(\delta) > 0$ such that

$$\mathbf{P}(V_T \succeq cTI) \geq 1-\delta$$

for all $T \geq T_0(\delta)$. To prove persistency of excitation (PE), we need only to establish one-sided lower bounds of the form

$$\mathbf{P}(\lambda_{\min}(V_T) \geq cT) \geq 1-\delta.$$

In other words, we need to show that the least singular value of the Gram matrix does not concentrate in a small ball around the origin. We now discuss a sufficient condition first presented in [4], based on the small-ball method [25]. An alternative approach via exponential inequalities can be found in [S2].

BLOCK MARTINGALE SMALL-BALL CONDITION

Before establishing PE for the whole vector z_t, we first study the projected processes $\xi^\top z_t$, where $\xi \in \mathbb{R}^{d_z}$ is a unit vector. The process z_t satisfies the block martingale small-ball condition with parameters $(k, \Gamma_{\mathrm{lb}}, p)$ if for every unit $\xi \in \mathbb{R}^{d_z}$ and every $t \geq 0$,

$$\frac{1}{k}\sum_{i=1}^{k} \mathbf{P}\left(\left| \xi^\top z_{t+i} \right|^2 \geq \xi^\top \Gamma_{\mathrm{lb}} \xi \middle| \mathcal{F}_t \right) \geq p \text{ almost surely.} \tag{S2}$$

The preceding condition states that, conditioned on t, the block average probability of being away from the origin is nonzero. The average probability is taken over blocks of size k. The geometry of the lower bound is captured by the matrix Γ_{lb}. Let condition (S2) hold. Then, it follows that z_t is persistently exciting, with the lower bound depending on the parameter Γ_{lb},

$$\mathbf{P}\left(V_T \succeq \frac{p^2}{16} k \lfloor T/k \rfloor \Gamma_{\mathrm{lb}} \right) \geq 1-\delta \tag{S3}$$

as long as we have a large enough number of samples

$$T \geq T_0 = \frac{10k}{p}\left(\log \frac{1}{\delta} + 2d_z \log \frac{10}{p} + \operatorname{logdet}(\Gamma_{\mathrm{ub}} \Gamma_{\mathrm{lb}}^{-1}) \right)$$

with $\Gamma_{\mathrm{ub}} = (d_z/\delta) \max_{t \leq T} \{ \mathbf{E} z_t z_t' \}$. Informally, the term Γ_{ub} is an upper bound of V_T/T, while the term Γ_{lb} is a lower bound of V_T/T. Hence the burn-in time N_0 depends logarithmically on the condition number of V_T. The proof of the result can be found in [4] and [26].

LINEAR SYSTEMS

In the case of fully observed linear systems, we can select $z_t = [x_t^\top \; u_t^\top]^\top$ to be the vector of the stacked state and input. Under white noise inputs, it can be shown [4] that the process z_t satisfies the $(k, \tilde{\Gamma}_{\lfloor k/2 \rfloor}, 3/20)$ block martingale small-ball condition, where

$$\tilde{\Gamma}_t \triangleq \begin{bmatrix} \Gamma_t & 0 \\ 0 & \sigma_u^2 I \end{bmatrix}.$$

REFERENCE

[S2] I. Ziemann, "A note on the smallest eigenvalue of the empirical covariance of causal Gaussian processes," 2022, *arXiv:2212.09508*.

The right-hand side of the preceding equation increases with T; fortunately, under the assumption that the system is nonexplosive $\rho(A) \leq 1$, it increases at most logarithmically with T. Hence, condition (9) will be satisfied for nonexplosive systems for a large enough T. The minimum time such that condition (9) is satisfied is also known as the *burn-in time*. The time index τ gives us some control of the size of the lower bound $\tilde{\Gamma}_{\lfloor \tau/2 \rfloor}$. Recall that the sequence Γ_t is increasing in the positive semidefinite cone. Hence, choosing a larger time index τ allows us to guarantee a stronger lower bound $\tilde{\Gamma}_{\lfloor \tau/2 \rfloor}$. On the other hand, the required burn-in time increases linearly with τ.

Self-Normalized Term

We begin with two observations about the self-normalized term

$$S_T V_T^{-1/2} = \left(\sum_{t=0}^{T-1} w_t \begin{bmatrix} x_t \\ u_t \end{bmatrix}^T \right) \left(\sum_{t=0}^{T-1} \begin{bmatrix} x_t \\ u_t \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix}^T \right)^{-1/2}.$$

First, note that the process noise w_t is independent of x_t, u_t for all $t \leq T$; that is, the sum S_T has a martingale structure. Second, as its name suggests, the term is self-normalized: if the covariates x_t, u_t are large for some t, then any increase in S_T will be compensated by an increase in $V_T^{-1/2}$. For this reason, $S_T V_T^{-1/2}$ is called a *self-normalized martingale*. Such terms have been studied previously in statistics in the asymptotic regime [7]. Here, we are interested in establishing finite-sample bounds. We invoke the results of Abbasi-Yadkori et al. [27]; see "Self-Normalized Martingales" for more details. Let V be a symmetric positive definite matrix (to be decided later), and set $\bar{V}_t = V_t + V$. The extra term V guarantees the positive definiteness of matrix $\bar{V}_t$. Then,

$$\left\| S_T \bar{V}_T^{-1/2} \right\|_{\text{op}}^2 \leq 8 \left\| \Sigma_w \right\|_{\text{op}} \log\left(\frac{\det(\bar{V}_T)^{1/2}}{\det(V)^{1/2}} \frac{5^{d_x}}{\delta} \right). \tag{12}$$

Crucially, self-normalization implies that the preceding term increases slowly (at most, logarithmically) with the

Self-Normalized Martingales

An object that arises often in standard least-squares analyses is the so-called self-normalized martingale. Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration, and let $z_t \in \mathbb{R}^{d_z}$, for some $d_z > 0$, be a stochastic process such that z_t is $\mathcal{F}_{t-1}$ measurable. Let $\eta_t \in \mathbb{R}^{d_\eta}$, $d_\eta > 0$ be a martingale difference sequence with respect to $\mathcal{F}_t$, that is, η_t is integrable, and $\mathcal{F}_t$ measurable, with $\mathbf{E}(\eta_t | \mathcal{F}_{t-1}) = 0$. Then, a self-normalized martingale $M_k \in \mathbb{R}^{d_\eta \times d_z}$ is defined as

$$M_k = \left(\sum_{t=0}^{k} \eta_t z_t^\top \right) \left(V + \sum_{t=0}^{k} z_t z_t^\top \right)^{-1/2}$$

where V is an arbitrary symmetric positive definite matrix of appropriate dimensions.

BOUNDS FOR SCALAR PROCESSES

Assume that $\eta_t \in \mathbb{R}$ is a scalar process. Under some regularity conditions on the tail of η_t, we can establish finite-sample bounds on the magnitude of M_k. Let the process η_t be conditionally K *sub-Gaussian* for some $K > 0$:

$$\mathbf{E}(e^{\lambda \eta_t} | \mathcal{F}_{t-1}) \leq e^{\frac{K^2 \lambda^2}{2}}, \quad \text{for all } \lambda \in \mathbb{R}.$$

The preceding condition requires that the tails of η_t decay at least as quickly as a Gaussian distribution. Now, we can invoke [27, Th. 1]. Letting

$$\bar{V}_k = V + \left(\sum_{t=0}^{k} z_t z_t^\top \right)$$

we then have the following finite-sample bound. Pick a failure probability $\delta \in (0, 1)$. Then, with a probability of at least $1 - \delta$,

$$\left\| M_k \right\|_2^2 \leq 2K^2 \log\left(\frac{\det(\bar{V}_k)^{1/2}}{\det(V)^{1/2}} \frac{1}{\delta} \right). \tag{S4}$$

EXTENSION TO VECTOR PROCESSES

Assume now that the process η_t is *vectored valued*, with $d_\eta > 1$, and conditionally K sub-Gaussian (that is, for any unit vector $v \in \mathbb{R}^{d_\eta}$, $\|v\|_2 = 1$, the projected process $v^\top \eta_t$ is conditionally K sub-Gaussian). The bound (S4) does not apply directly since it relies on the process η_t being scalar. Nevertheless, by appealing to *covering techniques* [3], it is straightforward to generalize this argument to vector processes. The idea is to apply (S4) to projections $v^\top \eta_t$ of η_t onto several directions v of the unit sphere. In particular, we discretize the unit sphere by considering points v_i, $i = 1, \ldots, N_\varepsilon$ such that the points are an ε net; that is, they cover the whole sphere with ε balls around them. Then, by taking a union bound over all points v_i, we obtain with a probability of at least $1 - \delta$,

$$\left\| M_k \right\|_{\text{op}}^2 \leq 2(1-\varepsilon)^{-2} K^2 \log\left(\frac{\det(\bar{V}_k)^{1/2}}{\det(V)^{1/2}} \frac{N_\varepsilon}{\delta} \right) \tag{S5}$$

where the number of points is at most

$$N_\varepsilon \leq \left(1 + \frac{2}{\varepsilon}\right)^{d_\eta}.$$

The term $(1-\varepsilon)^{-2}$ comes from the discretization error and decreases as the discretization becomes finer. However, as the discretization becomes finer, the number of points N_ε increases. A typical choice is $\varepsilon = 1/2$. The preceding guarantees are with respect to the operator norm. We could also obtain guarantees for the Frobenius norm by applying (S4) to $e_i^\top v_t$, where e_i, $i = 1, \ldots, d_\eta$ are the canonical vectors of $\mathbb{R}^{d_\eta}$. In this case, with a probability of at least $1 - \delta$,

$$\left\| M_k \right\|_F^2 \leq 2 d_\eta K^2 \log\left(\frac{\det(\bar{V}_k)^{1/2}}{\det(V)^{1/2}} \frac{d_\eta}{\delta} \right). \tag{S6}$$

norm of $\bar{V}_T$. If the data are generated by a stable system, this dependency can be further reduced [at the cost of inflating lower-order complexity terms by the inverse of the stability margin $1-\rho(A_\star)$]; see [28, Sec. 5.2]. To apply (12), we need to carefully select V. Moreover, to obtain data-independent sample complexity guarantees, we require a data-independent upper bound of $\bar{V}_T$. For the former, we choose $V = c\tau\lfloor T/\tau\rfloor\tilde{\Gamma}_{\lfloor\tau/2\rfloor}$. When lower bound (8) on V_T holds, then

$$\left\| S_T V_T^{-1/2} \right\|_{\text{op}}^2 \le 2 \left\| S_T \bar{V}_T^{-1/2} \right\|_{\text{op}}^2.$$

For the latter, we may appeal to the matrix version of Markov's inequality (due to Ahlswede and Winter [29, Th. 12]):

$$\mathbf{P}\left(V_T \not\preceq \frac{d_{\text{x}}+d_{\text{u}}}{\delta} T\tilde{\Gamma}_T \right) \le \delta$$

where $\{V_T \not\preceq ((d_{\text{x}}+d_{\text{u}})/\delta)T\tilde{\Gamma}_T\}$ is the complement of $\{V_T((d_x+d_u)/\delta)T\tilde{\Gamma}_T\}$. We could improve the preceding upper bound by applying the Hanson–Wright inequality instead of Markov's inequality. In this case, we would get logarithmic dependence on the confidence $\log 1/\delta$ instead of linear $1/\delta$. The improvement would be minor since $\bar{V}_T$ (and a factor $1/\delta$) already appears inside a logarithm in (12).

Sample Complexity Upper Bounds

Combining the previous bounds, we finally obtain instance-specific sample complexity upper bounds. For the least-squares estimator (5),

$$\mathbf{P}\left(\left\| \theta_\star - \hat{\theta} \right\|_{\text{op}} \ge \varepsilon \right) \le \delta \tag{16}$$

if the burn-in time condition (9) is satisfied along with

$$T \ge c' \frac{\left\| \Sigma_w \right\|_{\text{op}}}{\varepsilon^2 \lambda_{\min}(\Gamma_{\lfloor\tau/2\rfloor})} \left((d_{\text{x}}+d_{\text{u}}) \log \frac{d_{\text{x}}+d_{\text{u}}}{\delta} + \log \frac{\det \tilde{\Gamma}_T}{\det \tilde{\Gamma}_{\lfloor\tau/2\rfloor}} \right) \tag{17}$$

where c' is a universal constant. Once again, the right-hand side of inequality (17) increases at most logarithmically with the estimation horizon T for nonexplosive systems ($\rho(A) \le 1$), and hence will be satisfied for a large enough T. In fact, the rate defined in (17) is nearly optimal in the sense that it nearly matches the linear regression rate achieved *when all the samples are drawn independently*. See Figure 1 for an illustration.

To simplify the presentation, assume for now that we have strict stability $\rho(A) < 1$. In this case, the burn-in condition (9) and sample complexity bound (17) can be combined and rewritten as

$$T \ge c'' \max\left\{ \tau, \frac{1}{\varepsilon^2 \text{SNR}_\tau} \right\} (d_{\text{x}}+d_{\text{u}}) \log \frac{d_{\text{x}}+d_{\text{u}}}{\delta}$$

where c'' is another universal constant, and

$$\text{SNR}_\tau = \frac{\lambda_{\min}(\tilde{\Gamma}_{\lfloor\tau/2\rfloor})}{\left\| \Sigma_w \right\|_{\text{op}}}$$

captures the signal-to-noise ratio (SNR) of the system. The larger the SNR, the larger the excitation of the state compared to the magnitude of the noise. If the system has eigenvalues on the unit circle [$\rho(A) = 1$], then the expression looks similar but with some additional logarithmic terms; for simplicity, we omit this discussion here.

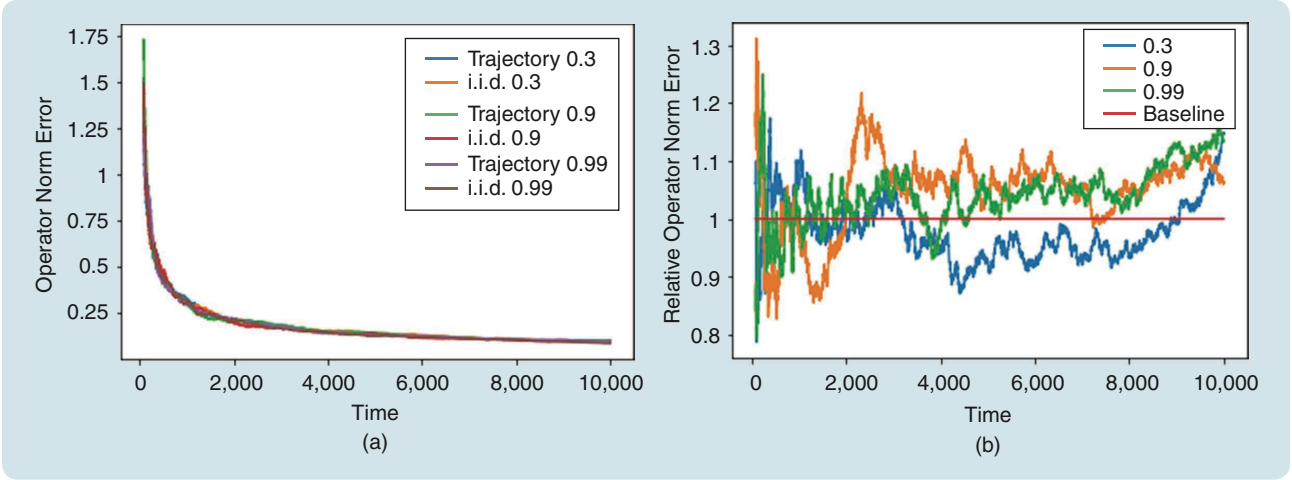


FIGURE 1 The essence of the learning-without-mixing phenomenon [5]: dependence does not necessarily impede the rate of convergence. (a) We plot the operator norm error of least-squares identification for $\rho(A_\cdot) \in \{0.3, 0, 9, 0.99\}$, $\lambda_{\min}(A_\cdot) \approx 0$, and $d_{\text{x}} = 25$. Lines marked "Trajectory" are sampled from a linear dynamical system $x_{t+1} = A_\cdot x_t + w_t$, whereas lines marked "i.i.d." are drawn from an independent baseline motivated by [2]. These i.i.d. lines correspond to a linear regression model $y_t = A_\cdot x_t + w_t$ in which the x_t are drawn i.i.d. from $\mathcal{N}(0, \text{dlyap}(A_\cdot, I_{d_{\text{x}}}))$. (b) Even as the correlation length $1/(1-\rho(A_\cdot))$ increases, the relative performance of the dynamic model to the independent baseline oscillates around one.

Ignoring logarithmic terms, the sample complexity grows as fast as $1/\varepsilon^2$, as we require more accuracy. Alternatively, the identification error decays as fast as $\tilde{O}(1/\sqrt{T})$ with the number of samples T. It also increases linearly with the dimension of the unknowns $d_x + d_u$. Intuitively, matrices $A_\star, B_\star$ have $d_x^2 + d_x d_u$ unknown entries. Every state measurement has d_x entries. Hence, we need at least $d_x + d_u$ state samples to match the number of unknowns in $A_\star, B_\star$. The sample complexity is also inversely proportional to the SNR. Finally, it depends logarithmically on δ, as (heuristically) predicted by the CLT. It is worth mentioning that the SNR depends heavily on the controllability structure of the system. In particular, under white noise inputs, the state covariance matrix Γ_k is actually the controllability Gramian of the pair $(A, [\sigma_u^2 B \;\; \Sigma_w^{1/2}])$. In this setting, controllability is equivalent to the excitability of the system. When the noise is isotropic (or nonsingular), the noise covariance Σ_w has full rank. Then, we can confirm that $\Gamma_1 \succeq \Sigma_w \succ 0$, which implies that the state is directly excited. It is, thus, sufficient to select $\tau = 2$ in the burn-in time condition (9) and sample complexity bound (17). When the noise is rank deficient, the state can be only indirectly excited; we can still achieve PE if there exists a $\tau > 0$ such that $\tilde{\Gamma}_{\lfloor \tau/2 \rfloor}$ is nonzero. In particular, we can select $\lfloor \tau/2 \rfloor$ to be equal to the controllability index of the system [24], that is, the smallest possible $\kappa > 0$ such that $\Gamma_\kappa \succ 0$. The preceding sample complexity upper bound is instance specific; that is, it holds for a specific system $(A_\star, B_\star, \Sigma_w)$. To obtain class-specific sample complexity upper bounds for some class $\mathscr{C}$, we need to impose global bounds on the norms of all $(A_\star, B_\star, \Sigma_w) \in \mathscr{C}$ as well as a global bound on $\lambda_{\max}^{-1}(\Gamma_\tau)$, for some $\tau > 0$; see, for example, [24].

Confidence Ellipsoids

Sample complexity guarantees are qualitative and data independent. That is, they provide intuition about how the number of required samples depends on various control-theoretic parameters, such as the dimension of the system and SNR. These guarantees depend directly on the quantities of the unknown system being estimated—see (9) and (17)—limiting their practical applicability. Another limitation is that the operator norm $\|\theta_\star - \hat{\theta}\|_{\text{op}}$ picks up the direction of the largest error. As a result, a guarantee, as in (16) and (17), provides confidence balls, which can be conservative in certain directions of the state space. In practice, it might be more useful to provide data-dependent confidence ellipsoids. Toward this end, we can still apply the tools for self-normalized martingales presented in "Self-Normalized Martingales." Let V be symmetric positive definite, and define $\bar{V}_t = V_t + V$. Using the properties of the least-squares estimator,

$$\left\| (\theta_\star - \hat{\theta}) \bar{V}_T^{1/2} \right\|_{\text{op}}^2 \le \left\| S_T \bar{V}_T^{-1/2} \right\|_{\text{op}}^2 \left\| \bar{V}_T^{1/2} V_T^{-1/2} \right\|_{\text{op}}^2.$$

Define the ellipsoid radius to be

$$r(\delta) \triangleq 8 \|\Sigma_w\|_{\text{op}} \log\left(\frac{\det(\bar{V}_T)^{1/2}}{\det(V)^{1/2}} \frac{5^{d_x}}{\delta} \right) \left\| \bar{V}_T^{1/2} V_T^{-1/2} \right\|_{\text{op}}^2.$$

Invoking (12), we obtain

$$\mathbf{P}\left(\left\| (\theta_\star - \hat{\theta}) \bar{V}_T^{1/2} \right\|_{\text{op}}^2 \le r(\delta) \right) \ge 1 - \delta. \tag{18}$$

Interestingly, the ellipsoid adapts to the informativity of the data, as captured by $\bar{V}_T$. If some mode of the system is well excited in V_T, the respective parameter error will be small. With the exception of $\|\Sigma_w\|_{\text{op}}$, all other quantities can be computed directly from data. In practice, one could replace $\|\Sigma_w\|_{\text{op}}$ by an upper bound or compute an empirical covariance from data. Although this quantity provides sharper confidence ellipsoids, it does not reveal directly how the identification error depends on the number of samples; that is, it does not reveal the statistical rate of estimating $\theta_\star$. Other data-dependent methods for establishing confidence ellipsoids can be found in [22], [26], and [30].

Sample Complexity Lower Bounds

The upper bounds on the sample complexity of the system identification of the previous section are valid only for the least-squares estimator (5). One may naturally ask whether we can do better with a different algorithm; that is, are the sample requirements of the least-squares algorithm a fundamental limitation, or are they suboptimal? One way to answer these questions is by establishing minimax lower bounds. The main technical workhorse underpinning such lower bounds is information-theoretic inequalities. As we show next, the least-squares identification algorithm analyzed in the preceding is nearly optimal in the case of fully observed systems. To prove this, it is sufficient to construct system instances that are difficult to identify *for all possible identification algorithms*. By invoking information-theoretic inequalities, we can show that any algorithm requires at least as many samples as the least-squares algorithm. We establish lower bounds for systems without exogenous inputs, but the same results also apply to systems with white noise exogenous inputs. For simplicity, we focus on the former case. Since there is no control input to implement an exploration policy, we denote this setting by $\pi = \varnothing$. Note that the case of more general exploration policies is an active front of research and is also discussed later on. Fix a spectral radius ρ, and define the class of scaled orthogonal systems

$$\mathcal{O}_\rho = \{A_\star \in \mathbb{R}^{d_x \times d_x} : A_\star = \rho O, O^\top O = I\}.$$

Let $N_c = N_c(\varepsilon, \delta, \mathcal{O}_\rho, \mathcal{A}, \varnothing)$ denote the best possible sample complexity for learning over the class of scaled

orthogonal systems. In [4], it is shown that for any identification algorithm $\mathcal{A}$,

$$N_c = \Omega\Big(\frac{d_x + \log 1/\delta}{\varepsilon^2 \mathrm{SNR}_{N_c}}\Big).$$

The result follows from a standard application of information-theoretic lower bounds; see "Birgé's Inequality" for more details. This shows that the rate $1/\varepsilon^2$, the dimension factor d_x, and the confidence $\log 1/\delta$ are fundamental, implying that the least-squares algorithm is nearly optimal. The preceding result holds for the specific subclass $\mathcal{O}_\rho$ of autonomous scaled orthogonal systems. It is also possible to obtain stronger instance-specific lower bounds, namely, lower bounds that hold locally around any fixed system. In particular, let $\theta_\cdot$ be an unknown system, and consider a ball $\mathcal{B}(\theta_\cdot, 3\varepsilon)$ of radius 3ε around $\theta_\cdot$. Let $\mathrm{N_c} = \mathrm{N_c}(\varepsilon, \delta, \mathcal{B}(\theta_\cdot, 3\varepsilon), \mathrm{A}, \varnothing)$ denote the minimum number of samples for identifying the local class $\mathcal{B}(\theta_\cdot, 3\varepsilon)$. In [31], it is shown that for any identification algorithm $\mathcal{A}$, failure probability $\delta \in (0,1)$, and accuracy $\varepsilon \in (0,\infty)$, it holds true that

$$N_c = \Omega\Big(\frac{d_x + \log 1/\delta}{\varepsilon^2 \mathrm{SNR}_{N_c}}\Big).$$

The proof is also based on Birgé's inequality. Terms capturing the SNR appear in both the upper and lower bounds. However, there is a gap between the upper and lower bounds. The former depend on $\lambda_{\min}^{-1}(\Gamma_\tau)$ for some small enough τ, while the latter depend on $\lambda_{\min}^{-1}(\Gamma_T)$, where T is the number of samples collected. Note that we cannot increase τ too much since it affects the burn-in time condition (9). In the case of stable systems $\rho(A_\cdot) < 1$, this gap can be closed at the expense of a burn-in time that depends on the mixing time $1/(1-\rho(A_\cdot))$ of the system [28]. The gap can also be made small, that is, $\tau = \Theta(T)$, in the case of diagonalizable marginally stable systems with $\rho(A_\cdot) = 1$ [2]. In the case of systems with white noise control inputs, the same analysis can be applied. In the case of general exploration policies, the landscape is more complex since both the policy π and the identification algorithm $\mathcal{A}$ affect the sample complexity. Let $N_c = N_c(\varepsilon, \delta, \mathcal{B}(\theta_\cdot, 3\varepsilon), \mathcal{A}, \pi)$ be the local sample complexity defined as before, where now the policy π can also be varied. Following the result of [5], we obtain the lower bound condition

$$N_c = \Omega\Big(\frac{\log 1/\delta}{\varepsilon^2 \mathrm{SNR}^{\star}_{N_c}}\Big)$$

Birgé's Inequality

Birgé's inequality is a sharper version of Fano's inequality, a classical tool from information theory [S3]. It can be used to establish lower bounds in multiple testing problems. Before we state the inequality, recall the definition of Kullback–Leibler (KL) divergence between two probability distributions $(\mathbf{P}, \mathbf{Q})$,

$$D(\mathbf{Q}\|\mathbf{P}) \triangleq \mathbf{E}_\mathbf{Q}\Big(\log \frac{d\mathbf{Q}}{d\mathbf{P}}\Big)$$

where we assume that $\mathbf{Q}$ is absolutely continuous with respect to $\mathbf{P}$ and $d\mathbf{Q}/d\mathbf{P}$ denotes the density of $\mathbf{Q}$ with respect to $\mathbf{P}$. Now, let $\mathbf{P}_0, \ldots, \mathbf{P}_n$ be probability distributions over some measurable space $(\Omega, \mathcal{F})$ such that $\mathbf{P}_i$, $i = 1, \ldots, n$ are absolutely continuous with respect to $\mathbf{P}_0$. These probability distributions represent, for instance, different hypotheses in a multiple-hypothesis testing scenario. Let $E_0, \ldots, E_n \in \mathcal{F}$ be disjoint events. For instance, $\mathbf{P}_i(E_i)$ might represent the probability of making a correct guess. Birgé's inequality states that a necessary condition for the minimum success probability to be lower bounded as

$$\min_{i=0,\ldots,n} \mathbf{P}_i(E_i) \triangleq 1 - \delta \geq \frac{1}{n+1} \tag{S7}$$

is that the average pairwise KL divergence between $\mathbf{P}_i$ and $\mathbf{P}_0$ satisfies the lower bound

$$\frac{1}{n}\sum_{i=1}^{n} D(\mathbf{P}_i\|\mathbf{P}_0) \geq h(1-\delta, \delta/n) \tag{S8}$$

where $h(p,q) = p\log p/q + (1-p)\log(1-p)/(1-q)$. The preceding condition states that making a correct guess with high probability is possible only if the distributions $\mathbf{P}_1, \ldots \mathbf{P}_n$ are sufficiently distinguishable from $\mathbf{P}_0$. Note that condition (S7) is permutation invariant; that is, it is independent of the ordering of the probability distributions. Hence, Birgé's inequality (S8) should also hold if we swap $\mathbf{P}_0$ with any $\mathbf{P}_j$, $j \leq n$. Hence, $\mathbf{P}_0, \ldots, \mathbf{P}_n$ should be mutually distinguishable.

SYSTEM IDENTIFICATION

Let $\mathscr{C} = \{\theta_0, \ldots, \theta_n\}$ be a class of systems that are 2ε separated; that is, $\|\theta_i - \theta_j\| > 2\varepsilon$. Let $\mathbf{P}_i$ be the probability distribution of the data $\{(y_0, u_0), \ldots, (y_T, u_T)\}$ when the underlying system is θ_i. Let $\hat{\theta}$ be the output of any identification algorithm. Since the systems are separated, the events $E_i \triangleq \{\|\theta_i - \hat{\theta}\| \leq \varepsilon\}$ will be disjoint. If some algorithm performs well with high probability across all systems, then (S7) holds, which (in turn) implies that (S8) holds. To obtain the tightest lower bounds possible, we aim to construct sets of 2ε-separated systems that nonetheless lead to data distributions with a small KL divergence. In other words, the separation should not be too large so that the distributions are as indistinguishable as possible.

REFERENCE

[S3] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, U.K.: Oxford Univ. Press, 2013.

where the exploration policy π is chosen to optimize the SNR term:

$$\mathrm{SNR}^{\star}_{N_c} = \max_{\pi} \frac{1}{N_c} \sum_{t=1}^{N_c} \mathbf{E}\begin{bmatrix} x_t \\ u_t \end{bmatrix}\begin{bmatrix} x_t^\top & u_t^\top \end{bmatrix}.$$

To avoid arbitrarily large exploration inputs, we limit the control input energy

$$\mathbf{E}\|\mathrm{u}_t\|_2^2 \leq \sigma_{\mathrm{u}}^2$$

for some $\sigma_u > 0$. Otherwise, we trivially obtain $\mathrm{SNR}^{\star} = \infty$. Finding the optimal exploration policy is not a simple problem and requires knowledge of the system dynamics. In [32], it is shown that the preceding lower bound can be achieved asymptotically (as $\delta \to 0$) by following an active exploration policy based on sinusoidal signals.

Summary and Generalizations

In Table 1, we summarize some of the main results for the sample complexity of identifying fully observed systems. For compactness, we denote $d = d_{\mathrm{x}} + d_{\mathrm{u}}$. Only results for open-loop nonexplosive systems [$\rho(A_\star) \leq 1$] are shown. If a stabilizing feedback gain K_0 is somehow known beforehand, the results can immediately be extended to the case of closed-loop stable systems [$\rho(A_\star - B_\star K_0) < 1$] under the stabilizing feedback law $u_t = K_0 x_t + \eta_t$. The case of open-loop unstable systems with $\rho(A_\star) > 1$ is analyzed in [33] and [35], where it is shown that under a regularity condition on the eigenvalues of $A_\star$, the error of learning explosive systems decays exponentially quickly with the number of samples. In [33], it is further shown that the error of learning systems with all eigenvalues on the unit circle decays at least as fast as $\tilde{O}(1/T)$, as opposed to the $\tilde{O}(1/\sqrt{T})$ error we get for strictly stable systems. The preceding rates agree with previous asymptotic results [7]. As discussed in the presentation of the lower bounds, the least-squares algorithm is nearly optimal in the case of white noise excitation. In the case of nonexplosive systems $\rho(A_\star) = 1$, there is a gap between the upper and lower bounds. The gap can be closed in the case of stable systems $\rho(A_\star) < 1$ [28]. This can be achieved by exploiting the Hanson–Wright inequality (see "The Hanson–Wright Inequality" for more details) instead of small-ball techniques. However, the downside of using Hanson–Wright is that the burn-in time depends on the mixing time of the system $1/(1-\rho(A_\star))$. As the system approaches instability $\rho(A_\star) \to 1$, the finite-sample guarantees degrade rapidly due to the burn-in time going to infinity. A benefit of small-ball techniques is that they hold even in the regime $\rho(A_\star) = 1$.

The Excitation Policy

In the presentation of sample complexity upper bounds, we considered only white noise input signals. Although white noise input signals can guarantee PE and lead to parameter recovery, they constitute a suboptimal exploration policy. It is a passive form of exploration that does not adapt online to the gathered information. Instead, in [32], an active exploration policy is employed based on sinusoidal inputs, leading to sharper sample complexity guarantees. In fact, in the regime where the failure probability goes to zero,

TABLE 1 **Sample complexities of fully observed system identification. Define $d = d_{\mathrm{x}} + d_{\mathrm{u}}$. The total number of nonzero elements is denoted by d_{s}. By $\mathrm{snr}^{\star}$, we denote the SNR under the best possible active exploration policy. For [36], we show only the result for $\rho(A_\star) \leq 1$. The sample complexities are given in terms of $N_{\mathrm{tot}} = N_{\mathrm{traj}} T$, that is, the total number of samples, where T is the horizon and N_{traj} is the number of trajectories. For single-trajectory data, $N_{\mathrm{tot}} = T$. All bounds are nonasymptotic, and we use only the big-O notation to simplify the presentation of the bounds.**

Paper	Trajectory	Stability	Actuation	Upper Bound	Burn-In Time	Lower Bound
[23]	Multiple	Any	White noise	$\tilde{O}\left(T\frac{d\log 1/\delta}{\varepsilon^2 \mathrm{snr}_T}\right)$	$T\tilde{O}(d+\log 1/\delta)$	—
[5]	Single	$\rho(A_\star) \leq 1$	White noise	$\tilde{O}\left(\frac{d\log d/\delta}{\varepsilon^2 \mathrm{snr}_\tau}\right)$	$\tilde{O}(\tau d \log d/\delta)$	$\Omega\left(\frac{d+\log 1/\delta}{\varepsilon^2 \mathrm{snr}_T}\right)$
[36]	Single	Any	White noise	$\tilde{O}\left(\frac{d\log d/\delta}{\varepsilon^2 \mathrm{snr}_1}\right)$	$\tilde{O}(d\log d/\delta)$	—
[6]	Single	Any	Active	—	—	$\Omega\left(\frac{\log 1/\delta}{\varepsilon^2 \mathrm{snr}_T^{\star}}\right)$
[30]	Single	$\rho(A_\star) < 1$	White noise	$\tilde{O}\left(\frac{d+\log 1/d}{\varepsilon^2 \mathrm{snr}_T}\right)$	$\tilde{O}\left(\frac{d+\log 1/d}{(1-\rho(A_\star))^2}\right)$	—
[34]	Single	$\rho(A_\star) < 1$	Active	$\tilde{O}\left(\frac{d+\log 1/\delta}{\varepsilon^2 \mathrm{snr}_\tau^{\star}}\right)$	$\mathrm{poly}\left(\frac{1}{1-\rho(A_\star)}\right)\tilde{O}(d+\log 1/\delta)$	$\Omega\left(\frac{\log 1/\delta}{\varepsilon^2 \mathrm{snr}_\infty^{\star}}\right)$
[37]	Single	$\rho(A_\star) < 1$	White noise	$\tilde{O}\left(\frac{d_{\mathrm{s}}\log d/\delta}{\varepsilon^2 \mathrm{snr}_\infty (1-\rho(A_\star))}\right)$	$\tilde{O}\left(\frac{d_{\mathrm{s}}^2\log d/\delta}{(1-\rho(A_\star))^4}\right)$	—
[25]	Single	$\rho(A_\star) \leq 1$	Any	$\tilde{O}\left(\exp(d)\frac{\log 1/\delta}{\varepsilon^2}\right)$	$\tilde{O}(d\log d/\delta)$	$\Omega\left(\exp(d)\frac{\log 1/\delta}{\varepsilon^2}\right)$

$\delta \to 0$, the proposed active exploration policy together with the least-squares identification algorithm are nearly optimal and achieve the minimax lower bound.

Systems With Sparse Structure

Another interesting problem is sparse system identification, where there might be an underlying sparse structure in the matrices $(A_\cdot, B_\cdot)$. In [34], it is shown that under an ℓ_1-regularization penalty and certain mutual incoherence conditions, the sample complexity of correctly identifying the nonzero elements of $(A_\cdot, B_\cdot)$ scales with d_{s}^2, that is, the number of nonzero elements squared, instead of the problem's dimensions $d_{\mathrm{x}} + d_{\mathrm{u}}$. Hence, if the nonzero elements are fewer than the dimension of the problem, we suffer from a smaller sample complexity. It is an open problem whether the square exponent of term d_{s}^2 can be improved. Moreover, it is an open question whether the results of [34] can be extended to open-loop nonexplosive systems $\rho(A_\cdot) = 1$.

Data From Multiple Trajectories

So far, we have focused on single-trajectory data. In practice, we might have access to data generated by several trajectories. In [2] and [22], learning from multiple independent trajectories is studied, where $N_{\mathrm{tot}} = N_{\mathrm{traj}} T$ is the total number of samples, T is the trajectory length, and N_{traj} is the number of trajectories. In [22], many samples are discarded (all but the last two) to turn system identification into an i.i.d. regression problem. As a result, there is an $O(T)$ extra sample overhead. These limitations are addressed by [2], where single-trajectory and multiple-trajectory learning are treated in a unified way; the parameter recovery guarantees are different and given in expectation, and hence, we did not include them in Table 1. An interesting conclusion in [2] is that in the "many" trajectories regime [for example, $N_{\mathrm{traj}} = \Omega(d)$], learning is more efficient than in the "few" trajectories regime [for example, $N_{\mathrm{traj}} = o(d)$]. Hence, it might be more beneficial to increase the number of trajectories N_{traj} rather than the horizon T while keeping the total number of samples constant.

Systems That Are Hard to Learn

All previous results rely on the process noise being full rank with positive definite covariance $\Sigma_w \succ 0$. In this case, all modes of the system are directly excited by the process noise, making learning easier, as the system SNR is always lower bounded by the condition number of the noise; that is, $\mathrm{SNR}_t \geq \left(\|\Sigma_w\|_{\mathrm{op}}\right) / \left(\lambda_{\min}(\Sigma_w)\right)$. As a result, in this case, system identification exhibits sample complexity, which scales polynomially with the system dimension d. If we take away this structural assumption and allow degenerate noise, the sample complexity can increase dramatically. In [24], it is shown that there exist nontrivial classes of systems for which the sample complexity scales exponentially with the dimension d. Such classes include underactuated systems, for example, systems with an integrator/network structure. Such systems are structurally hard to control/excite and, thus, difficult to identify. Under an additional robust controllability requirement, it is shown in [24] that the sample complexity of identifying underactuated systems cannot be worse than exponential with the dimension d. In fact, it cannot be worse than exponential in the so-called controllability index, which quantifies the degree of the underactuation of a system.

The Noise Model

We can obtain finite-sample guarantees if the process noise sequence is a martingale difference sequence [36], thus relaxing the i.i.d. requirement. Still, the methods presented here are quite fragile to the martingale difference noise assumption, which essentially amounts to a strong realizability assumption, implying, in some sense, that the model class contains the true model. In certain situations with colored noise, it is still possible to reduce the problem to a white noise problem, allowing us to invoke the self-normalized martingale inequality, for instance, by fitting a filter of sufficient length [37]. However, in full generality, sharply dealing with colored noise in the nonasymptotic regime is very challenging. If one seeks to go beyond sub-Gaussian tails, the situation becomes even more subtle. In a heavy-tailed noise model (with, for instance, $\mathbf{E}\|w_t\|^4 < \infty$,

The Hanson–Wright Inequality

In many situations of interest (for example, when analyzing Gram matrices), we need to work with quadratic functions of random variables. The Hanson–Wright inequality [3] is a standard tool for analyzing the concentration of such quadratic forms when the underlying random variables are sub-Gaussian. Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with independent mean-zero K-sub-Gaussian coordinates satisfying

$$\mathbf{E}e^{tX_i} \leq e^{\frac{K^2 t^2}{2}}, \quad i = 1, \ldots, n.$$

Let $M \in \mathbb{R}^{n \times n}$ be a matrix. Then, there exists a universal constant c such that for every $s \geq 0$,

$$\mathbf{P}\left(\left|X^\top MX - \mathbf{E}X^\top MX\right| \geq K^2 s\right) \leq 2e^{-c\min\left\{\frac{s^2}{\|M\|_{\mathrm{F}}^2}, \frac{s^2}{\|M\|_{\mathrm{op}}}\right\}}.$$

Hanson–Wright has been used as an alternative method for establishing persistency of excitation in the case of the identification of fully observed stable systems [28]. Contrary to small-ball methods, the Hanson–Wright inequality is a two-sided result, which is a stronger requirement. Hence, it can be conservative in the case of unstable or marginally stable systems. The Hanson–Wright inequality has also been utilized for proving isometry for Hankel matrices when the elements of the Hankel matrix are independent identically distributed.

but $\mathbf{E}\|w_t\|^p = \infty$ for some finite $p > 4$), the least-squares estimator is still optimal in expectation for most problems (at least for i.i.d. data [38]). However, it is no longer optimal in deviation—not even for i.i.d. data—meaning that it does not uniformly in δ attain the optimal $\log(1/\delta)$ failure probability [39]. Still, for i.i.d. data, this optimal dependency can be obtained by an alternative estimator (obtained by minimizing the so-called Huber loss; see [40, Sec. 6.4]). We do not know of any results that sharply characterize the failure probability in heavy-tailed linear system identification.

Partially Observed Systems

We now consider the more general case of partially observed systems with $C_\star \neq I$ and $\Sigma_v \neq 0$. Partial observability makes system identification harder, as we do not have direct access to state measurements. In the case where we do not know anything about the system, identifying the "true" state-space parameters is impossible, as the state-space representation is no longer unique, as the input–output map from inputs u to measured outputs y remains the same under similarity transformations. That is, for any invertible matrix Ξ, the systems

$$\theta_\star = (A_\star, B_\star, C_\star, \Sigma_w, \Sigma_v)$$
$$\theta'_\star = (\Xi^{-1}A_\star\Xi, \Xi^{-1}B_\star, C_\star\Xi, \Xi^{-1}\Sigma_w\Xi^{-\top}, \Sigma_v)$$

are equivalent from an input–output point of view. Another source of ambiguity is that the noise model is also nonunique [41]. Consider the system

$$\begin{aligned} \hat{x}_{k+1} &= A_\star\hat{x}_k + B_\star u_k + L_\star e_k \\ y_k &= C_\star\hat{x}_k + e_k \end{aligned} \tag{21}$$

where $L_\star$ is the steady-state Kalman filter gain

$$L_\star = A_\star S_\star C_\star^\top (C_\star S_\star C_\star^\top + \Sigma_v)^{-1}$$
$$S_\star = A_\star S_\star A_\star^\top + \Sigma_w - A_\star S_\star C_\star^*(C_\star S_\star C_\star^\top + \Sigma_v)^{-1}C_\star S_\star A_\star^\top.$$

The innovation error is defined as

$$e_k \triangleq y_k - C_\star\hat{x}_k.$$

The innovation process is i.i.d. zero-mean Gaussian with covariance $\Sigma_e \triangleq C_\star S_\star C_\star^\top + \Sigma_v$ [42]. System (21) is called the (steady-state) Kalman filter form or innovations form of system (1). Under the assumption that the system is initialized under its stationary distribution (that is, $\Gamma_0 = S_\star$), system (1) and its innovation form (21) are statistically equivalent from an input–output perspective in that they generate outputs with identical statistics. It has been common practice in the system identification literature [43] to work with the representation (21) instead of the original system (1). One reason is that the innovation noise is always output measurable, as opposed to the process/measurement noise. Another reason is that under certain observability conditions, the closed-loop map $A_\star - L_\star C_\star$ is stable; that is, $\rho(A_\star - L_\star C_\star) < 1$. We present techniques that can be applied to open-loop nonexplosive systems that satisfy $\rho(A_\star) \leq 1$. Again, assume that the open-loop inputs are white noise zero-mean Gaussian i.i.d., with $\mathbf{E}u_t u_t^\top = \sigma_u^2 I$ for some $\sigma_u > 0$. Also, assume that $(A_\star, C_\star)$ is detectable, $(A_\star, \Sigma_w^{1/2})$ is stabilizable, and Σ_v is invertible so that the innovation form (21) is well-defined and $\rho(A_\star - L_\star C_\star) < 1$. To simplify the analysis, assume that the Kalman filter starts from its steady state $\Gamma_0 = S_\star$, $\mathbf{E}x_0 = 0$. The latter is a weak assumption; due to the stability of the Kalman filter, we converge to the steady state exponentially fast. Most identification methods follow the prediction error approach [6] or the subspace method [41]. The prediction error approach is typically nonconvex and directly searches over the system parameters $\theta_\star$ by minimizing a prediction error cost. In the subspace approach, Hankel matrices of the system are estimated first, based on a convex regression problem. Then, realization is performed, typically based on singular value decomposition (SVD). Here, we focus on the subspace/realization approach. Recent work on the analysis of the prediction error method can be found in [44].

Regression Step

The first step is to establish a regression between future outputs and past inputs and outputs. Let $p > 0$ be a past horizon. By unrolling the innovation form (21), at any time step $k > 0$, we can express y_k as a function of p past outputs and inputs,

$$y_k = \underbrace{C_\star\mathcal{K}_p}_{G_p} Z_k + \underbrace{C_\star(A_\star - L_\star C_\star)^p\hat{x}_{k-p}}_{\text{bias}} + e_k \tag{22}$$

where Z_k is the vector of all the regressors stacked,

$$Z_k = \begin{bmatrix} y_{k-1}^\top & u_{k-1}^\top & \cdots & y_{k-p}^\top & u_{k-p}^\top \end{bmatrix}^\top$$

and $\mathcal{K}_p$ is an extended controllability matrix,

$$\mathcal{K}_p \triangleq \begin{bmatrix} \begin{bmatrix} B_\star & L_\star \end{bmatrix} & \cdots & (A_\star - L_\star C_\star)^{p-1}\begin{bmatrix} B_\star & L_\star \end{bmatrix} \end{bmatrix}.$$

Equation (22) shows that there is a linear relation between future outputs and past inputs/outputs, which is determined by matrix $G_p = C_\star\mathcal{K}_p$. We have a linear regression problem that is similar to the one encountered in the fully observed case since the innovation process e_t is i.i.d. and the regressors Z_k are independent of e_k at time k. The main differences are that 1) there exists a bias error term and 2) the unknown matrix G_p has a special structure. We can deal with the bias by increasing the past horizon p; the bias term goes to zero exponentially fast due to the stability of the Kalman filter. Note that (22) is also utilized by prediction error methods. In the prediction error approach, we optimize over the original state-space parameters (for example, A, B, and C), hence preserving the special structure of G_p. Here, following the subspace approach, we do not optimize over the original system parameters. Instead, we optimize directly over the higher-dimensional representation G_p by treating it as an unknown without structure. This leads to a convex least-squares problem:

$$\hat{G}_{p,T} \in \operatorname{argmin}_G \sum_{t=p}^{T} \|y_t - GZ_t\|_2^2. \tag{23}$$

In machine learning, this lifting to higher dimensions is referred to as improper learning [45]. After some algebraic manipulations, we can verify that

$$\hat{G}_{p,T} - G_p = \left(\sum_{t=p}^{T} e_t Z_t^\top\right)\left(\sum_{t=p}^{T} Z_t Z_t^\top\right)^{-1} + \text{bias}$$

where the bias term includes factors $(A_\cdot - L_\cdot C_\cdot)^p$ that decay exponentially with the past horizon p. The analysis now proceeds in a similar way as in the case of fully observed systems. We break the least-squares error into two terms, a self-normalized term and a term capturing PE:

$$\left\| \hat{G}_{p,T} - G_p \right\|_{\text{op}} \le \left\| S_{p,T} V_{p,T}^{-1/2} \right\|_{\text{op}} \left\| V_{p,T}^{-1/2} \right\|_{\text{op}}$$

where S_T and V_T are analogously defined as

$$S_{p,T} = \sum_{t=p}^{T} e_t Z_t^\top, V_{p,T} = \sum_{t=p}^{T} Z_t Z_t^\top.$$

For the self-normalized term, we exploit the techniques for self-normalized martingales. For the second term, we need to show PE. One way is to use again the small-ball techniques discussed in the fully observed case. An alternative way is establishing isometry for Hankel matrices (see "Isometry for Hankel Matrices"). Using the tools listed in the preceding, we can obtain sample complexity upper bounds for recovering the matrix G_p. Let $\Gamma_{Z,k} = \mathbf{E} Z_k Z_k^\top$ be the covariance of the regressors. For example, in the case of no inputs $B_\cdot = 0$, Tsiamis and Pappas [37] show that under the least-squares algorithm defined in the preceding,

$$\mathbf{P}\left(\left\| \mathrm{G_p} - \hat{\mathrm{G}}_{\mathrm{p,T}} \right\|_{\text{op}} \ge \varepsilon\right) \le \delta$$

if we select $p = \Omega(\log T)$ and

$$T \ge c \frac{p}{\varepsilon^2 \mathrm{SNR}_p} d_{\mathrm{y}} \log\left(\frac{p d_{\mathrm{y}}}{\delta} \frac{\left\| \Gamma_{Z,T} \right\|_{\text{op}}}{\lambda_{\min}(\Gamma_{Z,p})}\right)$$

where c is a universal constant and the SNR is defined as

$$\mathrm{SNR}_k = \frac{\left\| \Sigma_e \right\|_{\text{op}}}{\lambda_{\min}(\Gamma_{Z,p})}.$$

When we have inputs $B_\cdot \neq 0$, we can obtain a similar result by repeating the same arguments as in [37] and replacing d_{y} with $d_{\mathrm{y}} + d_{\mathrm{u}}$. Once again, we recover a rate of $\tilde{O}(1/\varepsilon^2)$. Equivalently, the error scales as $\tilde{O}(1/\sqrt{T})$. The main caveat is that we need to select p to increase logarithmically with the horizon T to mitigate the bias term. Ignoring ε, the SNR, and other system-theoretic parameters, the sample complexity upper bound scales with $p(d_{\mathrm{y}} + d_{\mathrm{u}})$; that is, it depends linearly on the size of the past horizon p. This upper bound suggests that there is a tradeoff between reducing the bias term (a large p) and reducing the sample complexity (a small p), as also discussed in prior work [10]. This dependence on the past horizon p arises because we ignore the structure of G_p and treat it as an unknown matrix. In this case, G_p has $p(d_{\mathrm{y}} + d_{\mathrm{u}})d_{\mathrm{y}}$ unknown entries. Since every measurement y_k contributes with d_{y} components, a sample complexity of $O(p(d_{\mathrm{y}} + d_{\mathrm{u}}))$ suffices. However, it might be the case that this sample complexity is suboptimal since the true number of unknowns in $\theta_\cdot$ is of the order of $d_{\mathrm{x}}^2 + d_{\mathrm{x}}(d_{\mathrm{y}} + d_{\mathrm{u}})$. It seems that by lifting the problem to higher dimensions in (23), we suffer from larger sample complexity.

Realization

Let us introduce the notations $A_{\mathrm{cl},\cdot} \triangleq A_\cdot - L_\cdot C_\cdot$ and $\tilde{B}_\cdot \triangleq [B_\cdot \; L_\cdot]$. For this section, assume for simplicity that system $(C_\cdot, A_{\mathrm{cl},\cdot}, \tilde{B}_\cdot)$ is minimal; that is, $(C_\cdot, A_{\mathrm{cl},\cdot})$ is observable,

Isometry for Hankel Matrices

Let $\eta_0, \ldots, \eta_{N-1}$ be a sequence of independent identically distributed zero-mean isotropic Gaussian variables in $\mathbb{R}^{d_\eta}$ [that is, $\eta_t \sim \mathcal{N}(0, I_{d_\eta})$], and consider the following Hankel matrix:

$$H_{L,N} \triangleq \begin{bmatrix} \eta_0 & \eta_1 & \cdots & \eta_{N-L-1} \\ \vdots & & & \\ \eta_L & \eta_{L+1} & \cdots & \eta_{N-1} \end{bmatrix}.$$

Such matrices arise in the analysis of system identification algorithms that use information of the past L steps for prediction. For example, η_t could be the input process u_t and/or the (normalized) innovations e_t. A crucial problem is determining whether the matrices $H_{L,N}$ are persistently exciting. One solution is to exploit the small-ball approach, as reviewed in "Persistency of Excitation and Small-Ball Bounds." Here, we review an alternative way to answer this question, which leads to a stronger two-sided result [S4], [46]. Fix a failure probability $\delta \le 1/2$. Then, there exists a universal constant c such that if

$$N \ge c L d_\eta \log \frac{L d_\eta}{\delta}$$

then with a probability of at least $1 - \delta$,

$$\frac{N}{2} I_{L d_\eta} \preceq H_{L,N} H_{L,N}^\top \preceq \frac{3N}{2} I_{L d_\eta}.$$

The result is adapted from [46, Th. A.2]. The proof is based on the Hanson–Wright inequality along with Fourier domain techniques. Similar results appear in [47] and [48] but require a slightly larger burn-in time.

REFERENCE

[S4] B. Djehiche, O. Mazhar, and C. R. Rojas, "Finite impulse response models: A non-asymptotic analysis of the least squares estimator," *Bernoulli*, vol. 27, no. 2, pp. 976–1000, May 2021, doi: 10.3150/20-BEJ1262.

and $(A_{\mathrm{cl},\star}, \tilde{B}_\star)$ is controllable. Under this notation, matrix G_p contains the Markov parameters $C_\star A_{\mathrm{cl},\star}^k \tilde{B}_\star$, $k \leq p-1$ of system $(C_\star, A_{\mathrm{cl},\star}, \tilde{B}_\star)$, allowing for the use of standard realization techniques to extract $(C_\star, A_{\mathrm{cl},\star}, \tilde{B}_\star)$ from the Markov parameters. A standard such approach is the Ho–Kalman realization technique. If we assume that we know the true Markov parameters G_p, then we can construct the following Hankel matrix:

$$\mathcal{H}_{k_1,p} \triangleq \begin{bmatrix} C_\star \tilde{B}_\star & C_\star A_{\mathrm{cl},\star} \tilde{B}_\star & \cdots & C_\star A_{\mathrm{cl},\star}^{p-1-k_1} \tilde{B}_\star \\ C_\star A_{\mathrm{cl},\star} \tilde{B}_\star & C_\star A_{\mathrm{cl},\star}^2 \tilde{B}_\star & \cdots & C_\star A_{\mathrm{cl},\star}^{p-k_1} \tilde{B}_\star \\ \vdots & & \ddots & \vdots \\ C_\star A_{\mathrm{cl},\star}^{k_1} \tilde{B}_\star & C_\star A_{\mathrm{cl},\star}^{k_1+1} \tilde{B}_\star & \cdots & C_\star A_{\mathrm{cl},\star}^{p-1} \tilde{B}_\star \end{bmatrix}.$$

The Hankel matrix has rank d_{x} since it can be written as the outer product of a controllability matrix and an observability matrix:

$$\mathcal{H}_{k_1,p} = \underbrace{\begin{bmatrix} C_\star \\ C_\star A_{\mathrm{cl},\star} \\ \vdots \\ C_\star A_{\mathrm{cl},\star}^{k_1} \end{bmatrix}}_{\mathcal{O}_{k_1}} \underbrace{\begin{bmatrix} \tilde{B}_\star & A_{\mathrm{cl},\star} \tilde{B}_\star & \cdots & A_{\mathrm{cl},\star}^{p-1-k_1} \tilde{B}_\star \end{bmatrix}}_{\mathcal{C}_{p-1-k_1}}.$$

To make sure that the Hankel matrix is of rank d_{x}, it is sufficient to select $k_1, p-1-k_1 \geq d_{\mathrm{x}}$. In the setting where we know the true Markov parameters, a simple SVD suffices to recover the observability and controllability matrices up to a similarity transformation. In particular, letting the singular decomposition be written as

$$\mathcal{H}_{k_1,p} = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^\top \\ V_2^\top \end{bmatrix}$$

we can select a *balanced realization* $\mathcal{O}_{k_1} = \mathrm{U}_1 \Sigma_1^{1/2}, \mathcal{C}_{p-1-k_1} = \Sigma_1^{1/2} \mathrm{V}_1^\top$. Then, from the observability/controllability matrices, it is easy to recover $(C_\star, A_{\mathrm{cl},\star}, \tilde{B}_\star)$ up to a similarity transformation; see, for example, [48]. However, in practice, we have access only to noisy Markov parameter estimates $\hat{G}_{p,N}$, obtained, for example, via the least-squares identification step described previously. In this case, the corresponding Hankel matrix $\hat{\mathcal{H}}_{k_1,p}$ will also be noisy and no longer have rank d_{x}; instead, it will, in general, have a higher rank. In this case, a low-rank approximation step is crucial for recovering the correct observability and controllability matrices. Assume that we know the true order d_{x} of the system. Then, we can perform SVD truncation, that is, choose the singular vectors corresponding to the d_{x} largest singular values. If the SVD of the noisy Hankel matrix is

$$\hat{\mathcal{H}}_{k_1,p} = \begin{bmatrix} \hat{U}_1 & \hat{U}_2 \end{bmatrix} \begin{bmatrix} \hat{\Sigma}_1 & 0 \\ 0 & \hat{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \hat{V}_1^\top \\ \hat{V}_2^\top \end{bmatrix}$$

then one solution is to keep the d_{x} largest singular values, that is, select $\hat{\mathcal{O}}_{k_1,T} = \hat{U}_1 \hat{\Sigma}_1^{1/2}, \hat{\mathcal{C}}_{p-1-k_1,T} = \hat{\Sigma}_1^{1/2} \hat{V}_1^\top$. To capture the error between the true and estimated observability/controllability matrices, we appeal to SVD perturbation results; more details can be found in [49] and [50, Th. 5.14]. Essentially, these results state that, for some similarity transformation T, the error $\left\| \mathcal{O}_{k_1} - \hat{\mathcal{O}}_{k_1,T} \right\|_{\mathrm{op}}$ (similarly for the controllability matrix) scales with the Markov parameter error $\left\| G_p - \hat{G}_{p,T} \right\|_{\mathrm{op}}$ as long as a robustness condition is satisfied. Ignoring dependencies on k_1, p, the robustness condition is typically of the form

$$\left\| G_p - \hat{G}_{p,T} \right\|_{\mathrm{op}} \leq O(\sigma_{d_{\mathrm{x}}}(\mathcal{H}_{k_1,p})). \tag{24}$$

That is, the Markov parameter estimation error should be smaller than the smallest singular value of the true Hankel matrix $\mathcal{H}_{k_1,p}$. Such a condition is a fundamental limitation of the SVD procedure; it guarantees that the d_{x} singular vectors of the Hankel matrix $\mathcal{H}_{k_1,p}$ are approximated continuously, while the extra singular vectors in $\hat{\mathcal{H}}_{k_1,p}$, which come from the noise and contribute to full rank, are rejected. While in the asymptotic regime such a condition is satisfied asymptotically, in the finite-sample regime, it imposes a high sample complexity, as the smallest singular value of the Hankel matrix can be very small in practice. It is an interesting open problem to look at different realization approaches or model reduction techniques so that we avoid this restrictive robustness condition.

Open Problem 1: Comparison of Subspace Algorithms

Most results in the finite-sample regime analyze the performance of the Ho–Kalman method (or similar variants) [37], [47], [48], [51]. However, in the subspace identification literature, this realization approach is rarely used. Popular subspace identification algorithms (for example, Multivariable Output Error State Space [52] and Numerical Algorithms for Subspace State Space System Identification [41]) premultiply and/or postmultiply the Hankel matrix, with appropriate weighting matrices, before performing the SVD step; see, for example, [43, Sec. 3]. Several asymptotic properties of such algorithmic variations have been studied before [53]. However, it is an open problem to compare such algorithms using finite-sample methods. In particular, under finite samples, a robustness condition like (24) should be satisfied for the SVD step to be well-behaved. Different methods lead to different robustness conditions, affecting finite-sample performance. Such robustness conditions did not appear before in asymptotic analyses, for example, [54], since as the number of samples goes to infinity, the SVD error decays continuously.

Overview and Limitations

An overview of prior work can be found in Table 2. Up to now, we have studied the identification of Markov

parameters of both the deterministic part, that is, $(C_\cdot, A_{\mathrm{cl},\cdot}, \tilde{B}_\cdot)$, and the stochastic part of the system, that is, $(C_\cdot, A_\cdot, L_\cdot)$. Prior work has also studied the identification of exclusively the deterministic part [46], [47], [48], [55], [56], [57], that is, the Markov parameters of $(C_\cdot, A_\cdot, B_\cdot)$, where only past inputs are used as regressors. By using only inputs, these results hold only for stable systems $\rho(A_\cdot) < 1$ unless we use multiple trajectories [58]. In [59], it is shown that the identification of nonexplosive systems $\rho(A_\cdot) = 1$ is possible if we also use past outputs as regressors and include a prefiltering step in the system identification algorithm, that is, learn an autoregressive filter first before estimating the Markov parameters. The identification of the stochastic part [that is, the Markov parameters of $(C_\cdot, A_\cdot, L_\cdot)$] is investigated in [37]. A nonparametric approach is considered in [17].

The Excitation Policy

Most of the aforementioned works rely on white noise open-loop excitation to achieve parameter recovery. Closed-loop identification under finite samples has been analyzed in [51] and [60], where the closed-loop controller is a linear dynamic feedback law, potentially driven by white noise [51]. The problem of experiment design (that is, finding good excitation policies in the finite-sample regime) remains quite open. Still, it was studied in the classical system identification literature using asymptotic tools [6].

The Noise Model

In the case of non-Gaussian noise, the system (1) and its Kalman form (21) have similar second moments. However, they are no longer statistically equivalent, and the innovation process is no longer i.i.d. Gaussian. For this reason, some of the techniques presented in the preceding might not be applicable. We also point out that in the case of i.i.d. sub-Gaussian noise, the results of [47], [55], and [59] still hold but recover only the deterministic part of the system.

System Order

The realization procedure that we presented previously requires the order of the system d_x to be known. The identification of systems under an unknown model order is studied in [47], [56], and [57]. In [47], an approximate order, which does not necessarily converge to the true one, is obtained by truncating the estimated Hankel matrices at a desired level of accuracy. In [56] and [57], the problem of learning low-rank Hankel matrices via nuclear norm regularization is studied.

Lower Bounds

Lower bounds have been studied before in the classical literature [6, Ch. 7]. In the case of a known system order, we can characterize the best possible parameter estimation variance among all estimators by invoking the Cramér–Rao inequality [61], a variant of Van Trees' inequality that is studied in the following. One difference from Birgé's inequality is that the Cramér–Rao inequality characterizes the expected error (variance), while Birgé's inequality characterizes tail probabilities providing information about the confidence level δ. Unlike fully observed systems, existing lower bounds for partially observed systems do not have transparent expressions in terms of system-theoretic properties, such as the system dimension and controllability Gramians; see, for example, the derivation of Cramér–Rao bounds in [62]. This is mainly due to the nonlinearity of the input-to-output map with respect to the state-space parameters. Another issue is the nonuniqueness of state-space representations.

Open Problems in the Partially Observed Setting

Under the assumption that the model order is known and under certain conditions on the inputs, the asymptotic optimality of several algorithms has been established. In particular, it has been shown that the prediction error method is equivalent to the maximum-likelihood method [6, Ch. 9], while some subspace identification algorithms

TABLE 2 **The system identification of partially observed systems.**

Paper	Trajectory	Stability	System Part	Order d_x	Actuation	Noise
[49], [52]	Single	$\rho(A_\cdot) < 1$	Deterministic	Known	Open loop	Gaussian
[63]	Single	$\rho(A_\cdot) \leq 1$	Deterministic	Known	Open loop	Sub-Gaussian
[40]	Single	$\rho(A_\cdot) \leq 1$	Stochastic	Known	—	Gaussian
[51]	Single	$\rho(A_\cdot) < 1$	Deterministic	Unknown	Open loop	Sub-Gaussian
[61]	Single	$\rho(A_\cdot) < 1$	Deterministic	Unknown	Open loop	Gaussian
[59]	Single	$\rho(A_\cdot) < 1$	Deterministic	Known	Open loop	Sub-Gaussian
[55], [64]	Single	Closed loop	Both	Known	Closed loop	Gaussian
[62]	Multiple	Any	Deterministic	Known	Open loop	Gaussian
[60]	Multiple	Any	Deterministic	Unknown	Open loop	Gaussian

asymptotically match the maximum-likelihood method under white noise excitation [53], [63]. Obtaining a finite-sample analog is an open problem.

Open Problem 2: Optimal Sample Complexity
What is the optimal sample complexity in the case of partial observability? In the case of a known system order, can we match the asymptotic performance of maximum likelihood by a nonasymptotic analysis? What if the order is unknown? How do system-theoretic parameters affect complexity?

An open question is whether the optimal sample complexity should depend on the past horizon p. As discussed in the "Regression Step" section, this might not be the case since the number of unknowns in $\theta_\cdot$ is independent of the horizon p. Some progress in this regard has already been made: in [46], it is shown that in the absence of process noise, the sample complexity depends only logarithmically on the past horizon p while retaining the $1/\varepsilon^2$ complexity rate. This is achieved by exploiting repeated entries in Hankel matrices, which are computed at different scales, that is, for different horizons p. In the case of process noise, the complexity bound in [46] still scales linearly with p. In [55], the sample complexity is shown to be logarithmic with p at the expense of a worse $1/\varepsilon^4$ complexity rate. This is achieved by adding an ℓ_1-regularization penalty on G_p in the regression step. To conclude, another open problem is the identification of open-loop (explosively) unstable systems, $(\rho(A_\cdot) > 1)$, in the case of single-trajectory data. While this problem is resolved in the case of fully observed systems, (under certain regularity conditions) it is still open in the case of partial observability.

Open Problem 3
Existing results for partially observable systems rely on stability $\rho(A_\cdot) \leq 1$. What, if any, are the necessary conditions for conducting open-loop unstable identification based on a single trajectory of data?

One of the main technical difficulties in the case of unstable systems is dealing with the bias term in (22). If the state is increasing exponentially fast with time k, the bias term might not decay fast enough with p. In the case of nonexplosive systems, two-step procedures (for example, performing a prefiltering step [59] or estimating components of the marginally stable subspace first [64]) guarantee learnability. It is an open question whether a two-step procedure would work for (explosively) unstable systems.

OFFLINE CONTROL

In the previous section, we studied the system identification of unknown systems under a finite number of samples. Although system identification is a problem of independent interest, our ultimate goal is to control the underlying unknown system. In this section, we connect the previous results with controlling unknown systems in a model-based framework. We also review some model-free methods. We focus on offline learning architectures, where we design the controller once after collecting the data. This setup is very similar to the setting of episodic reinforcement learning (RL), which has received renewed interest recently due to its success in settings such as games [1], [65]. However, most existing analyses focus on finite state and input (action) spaces. Since learning methods are becoming increasingly ubiquitous even for complex continuous control tasks [66], the gap between theory and practice has become considerable. The LQR and the linear quadratic Gaussian (LQG) problems offer a theoretically tractable path forward to reason about RL for continuous control tasks. By leveraging the theoretically tractable natures of the LQR and LQG, we obtain baselines and are able to quantify the performance of learning algorithms in terms of natural control-theoretic parameters. Perhaps most importantly, given the safety-critical nature of many applications [67], we are able to quantify what makes learning hard and when it necessarily fails. To make this concrete, suppose a learner (control engineer) knows that the system has dynamics of the form

$$x_{t+1} = A_\cdot x_t + B_\cdot u_t + w_t \tag{25}$$

where, as in the previous section, x_t and $w_t \in \mathbb{R}^{d_x}$ are the state and process noise, respectively, and $u_t \in \mathbb{R}^{d_u}$ is the control input. The dynamics matrices are $A_\cdot \in \mathbb{R}^{d_x \times d_x}$ and $B_\cdot \in \mathbb{R}^{d_x \times d_u}$. In the learning task, the parameters $(A_\cdot, B_\cdot)$ are unknown to the learner. All that is known is that $(A_\cdot, B_\cdot) \in \Theta$, where Θ is some subset of parameters, typically those corresponding to stabilizable systems. In the offline setting, the learner is given access to N_{traj} sampled trajectories of length T (a total of $N_{\text{tot}} = N_{\text{traj}} T$ samples) from the system (25) and is tasked to output a policy π that renders the following cost as small as possible:

$$\bar{V}(\theta; K) \triangleq \limsup_{T \to \infty} \mathbf{E}_\theta^K \left[\frac{1}{T} \sum_{t=0}^{T-1} (x_t^\top Q x_t + u_t^\top R u_t) \right] \tag{26}$$

where expectation $\mathbf{E}_\theta^K$ is taken with respect to dynamics $\theta = (A, B)$ under the feedback law $u_t = K x_t$. In this case, it is of course known that the optimal controller is a constant state feedback law of the form $u_t = K(A_\cdot, B_\cdot) x_t = K_\cdot x_t$, where the controller gain $K(A, B)$ is specified in terms of the solution $P = P(A, B)$ to a discrete-time algebraic Riccati equation:

$$P = Q + A^\top P A - A^\top P B (B^\top P B + R)^{-1} B^\top P A \tag{27}$$

$$K = -(B^\top P B + R)^{-1} B^\top P A. \tag{28}$$

Model-Based Methods

A classical approach to designing the optimal LQR controller for an unknown system (25), which we revisit from a finite-data perspective, is to perform system identification followed by a control design step. In RL terminology, this approach is referred to as a model-based approach because we explicitly parameterize and learn the transition dynamics, which are then used to compute a policy. In particular, suppose that we have obtained estimates $(\hat{A}, \hat{B})$ of $\theta_\star = (A_\star, B_\star)$, and these estimates are guaranteed to be ε accurate; that is, $\max\left\{\left\|A_\star - \hat{A}\right\|_{\mathrm{op}}, \left\|B_\star - \hat{B}\right\|_{\mathrm{op}}\right\} \leq \varepsilon$. Such estimates can be acquired and guaranteed to satisfy the desired accuracy level (with high probability) by leveraging the results of the discussion in the "Sample Complexity Upper Bounds" section. Based on the system estimates, we can either apply CE control or design a robust controller using the error information ε.

Certainty Equivalence

The CE approach is to simply use the estimates $(\hat{A}, \hat{B})$ as if they were the ground truth and play the controller $\hat{K} = K(\hat{A}, \hat{B})$. This setting is analyzed in Mania et al. [68, Th. 2]. They demonstrate that the controller $\hat{K} = K(\hat{A}, \hat{B})$ enjoys the suboptimality guarantee

$$\bar{V}(\theta_\star; \hat{K}) - \bar{V}(\theta_\star; K_\star) \leq \mathrm{poly}_{\theta_\star}\varepsilon^2 \tag{29}$$

where $\mathrm{poly}_{\theta_\star}$ denotes a quantity polynomial in system quantities, such as $\|P_\star\|_{\mathrm{op}}$, and the spectral radius of the optimal closed-loop dynamics $A_\star + B_\star K_\star$—one can view the term $\mathrm{poly}_{\theta_\star}$ as capturing the fact that systems with well-conditioned closed-loop behavior [a small $\|P_\star\|_{\mathrm{op}}$, $\rho(A_\star + B_\star K_\star)$] are easier to learn to control. Similar guarantees can also be provided for the partially observed LQG setting, in which the entire linear dynamic controller is estimated from data [68, Th. 3]. It is important to recognize, however, that guarantee (29) comes with the caveat that the accuracy ε needs to be small enough so that the controller $\hat{K}$ can be shown to be stabilizing for the instance $\theta_\star = (A_\star, B_\star)$. Mania et al. [68] provide sufficient conditions on the accuracy ε in terms of system parameters by leveraging Riccati equation perturbation theory (see "Riccati Equation Perturbation Theory"). The dependence on ε in inequality (29) is optimal, and it can be shown that *for almost every* experiment consisting of input state data $\{(x_0, u_0), \ldots, (u_{N_{\mathrm{tot}}-1}, x_{N_{\mathrm{tot}}})\}$, the least-squares estimator described previously (in combination with CE control) is optimal [69, Th. 2.1] in that up to universal constants, there exists no better strategy. In fact, it is later shown that the CE approach is also the best-known strategy in the more challenging online control setting. Combining guarantee (29) with the sample complexity upper bounds of the previous section, we can obtain end-to-end guarantees for the offline learning of the optimal LQR controller. In particular, we obtain that the suboptimality gap decreases at least as fast as $\tilde{O}(1/N_{\mathrm{tot}})$. However, as stated earlier, this result assumes that the number of samples is large enough that the CE controller $\hat{K}$ is stabilizing for the original system, which may require a large burn-in time.

Robust Control Methods

While the CE controller is optimal when the model error ε is very small, there are nevertheless many cases of interest where only a coarse model is available and where the model error is too large to guarantee that the CE controller is stabilizing [22]. In such settings, an alternative is to design a robust controller that stabilizes all possible systems consistent with the model estimates and error bounds. In [70], the problem of robust control from coarse system identification was studied in the nonasymptotic regime. In [22], a robust control scheme based on system-level synthesis (SLS) [71] is introduced that uses finite-sample model error information. The aforementioned robust control designs are safer than the CE controller in general. However, the cost of this robustness is that the resulting controller suboptimality guarantees are worse. Contrary to (29), the suboptimality guarantees are of the order of

$$\bar{V}(\theta_\star; \hat{K}) - \bar{V}(\theta_\star; K_\star) \leq \mathrm{poly}_{\theta_\star}\varepsilon \tag{30}$$

where $\hat{K}$ is the robust controller. It is unknown whether this suboptimality is inherent or an artifact of the analysis. SLS controllers can also be deployed in the case of state/input constraints [72] as well as partially observed systems [73]. An alternative input–output parameterization framework was adapted in [74] to deal with uncertain partially observed systems.

Model-Free Methods

Model-free methods, in which (essentially) no structural information about the problem is used to derive a learning-based

Riccati Equation Perturbation Theory

To provide a guarantee of the form (29) for the certainty-equivalent approach, we need to guarantee that small errors in the estimates $\max\left\{\left\|A_\star - \hat{A}\right\|_{\mathrm{op}}, \left\|B_\star - \hat{B}\right\|_{\mathrm{op}}\right\} \leq \varepsilon$ translate to small errors in Riccati equation quantities (27)–(28). Key to achieving such guarantees is an operator-theoretic proof strategy, due to [S5]. Roughly, the idea is to construct a map Φ, of which the error $P_\star - \hat{P}$ is the unique fixed point over a set of elements with a small norm. A more detailed account can be found in [68, Sec. 4.1]. Also note that [36, Sec. 3] has recently developed an alternative ordinary differential equation approach, which gives tighter bounds in terms of system-theoretic parameters.

REFERENCE

[S5] M. M. Konstantinov, P. H. Petkov, and N. D. Christov, "Perturbation analysis of the discrete Riccati equation," *Kybernetika*, vol. 29, no. 1, pp. 18–29, 1993.

policy, are very popular in the RL literature. The most basic class of such methods are policy gradient methods, which we discuss next in the context of the LQR problem.

Policy Gradient Methods

Policy gradient methods work exactly as their name advertises: they run (stochastic) gradient descent on a controller parameterization with respect to the cost (26). To make this concrete, let us (for simplicity) first discuss the state feedback setting in which $C_\star = I_{d_\mathbf{x}}$ and $v_t = 0$. In light of the form (27)–(28) of the optimal policy, it appears reasonable to parameterize the cost (26) by linear controllers of the form $u_t = Kx_t$ and run our descent steps on matrices $K \in \mathbb{R}^{d_\mathbf{u} \times d_\mathbf{x}}$.

Do Exact Gradients Converge?

Assume for the moment that we have oracle access to *exact gradients*, and we are able to run (nonstochastic) gradient descent on the cost function (26):

$$K_{j+1} = K_j - \nabla_K V_T(\theta; K)\big|_{K=K_j}.$$

It is not obvious that such an algorithm will work, as even in this simplified setting, there are two potential obstacles to convergence: 1) the cost function (26) is nonconvex in K, and 2) the cost function (26) is not globally smooth—in fact, it is not even finite for those K that do not stabilize the system (25). Thankfully, the LQR objective (26) satisfies "weaker versions" of convexity and smoothness, which are entirely sufficient (see "Linear Quadratic Regulator, Polyak–Łojasiewicz, and Approximate Smoothness"). These weaker conditions were first established by Fazel et al. [75], who showed that if initialized with a stabilizing controller K_0, after only $O(\log 1/\varepsilon)$ iterations, (nonstochastic) gradient descent outputs a controller $\tilde{K}$ satisfying

$$\bar{V}(\theta, \tilde{K}) - \min_K \bar{V}(\theta, K) \le \varepsilon. \tag{31}$$

It should be noted that the authors of [75] consider a slightly different cost function than the cost considered here. Namely, they consider the infinite-horizon case with $w_t = 0$, and only the initial condition x_0 is allowed to be random. However, the infinite-horizon and ergodic average cost functions are almost identical (as functions of K), and it is straightforward to verify that the convergence guarantee mentioned in the preceding remains true with only minor modifications to problem-specific constants when applied to the ergodic average cost (26). Having established that the exact gradient method converges, Fazel et al. [75] also showed that a method based on zero-order gradient estimates also converges. However, their results apply only to the noiseless setting with a random initial condition. By contrast, [76] analyzes a noisy finite-horizon setting and shows that such methods still provably converge. Note that the assumption of an initial stabilizing controller mentioned in the preceding can be removed with a more sophisticated gradient strategy [77]. We refer the reader to the recent survey [78] for a more comprehensive overview of policy gradient methods.

Fundamental Limits and Model Based Versus Model Free

Given the optimality of the CE controller in the offline LQR setting, it is natural to wonder whether similar guarantees are achievable by model-free methods based on policy gradients. To this end, Tu and Recht [79] study a simplified version of LQR (26) in which $R = 0$ and the optimal solution is of the form $K_\star = -B_\star^\dagger A_\star$. In this simplified scenario, they compute asymptotically exact expressions for the risk of CE and a stochastic policy gradient method (REINFORCE) and show that

Linear Quadratic Regulator, Polyak–Łojasiewicz, and Approximate Smoothness

While the linear quadratic regulator (LQR) objective is not convex, the objective (26) satisfies the so-called Polyak–Łojasiewicz (PL) condition. Namely, Fazel et al. [75, Lemma 3] show that as long as the tuple $(A, \sqrt{\Sigma_w})$ is controllable, the following PL condition holds:

$$\bar{V}(\theta, K) - \min_K \bar{V}(\theta, K) \le \lambda \left\| \nabla_K \bar{V}(\theta, K) \right\|_F^2 \tag{S9}$$

for some problem-specific constant $\lambda > 0$. PL conditions, such as inequality (S9), are known to be sufficient alternatives to (strong) convexity in the optimization literature [S6], [S7]. In particular, condition (S9) enforces that any stationary point is a global minimizer, as is the case for convex functions. An alternative perspective on the condition (S9) is offered in [S8], in which it is shown to be a consequence of the existence of a convex reparameterization for the LQR objective. Similarly, even though the objective (26) is not globally smooth, it is sufficiently regular in that

$$\bar{V}(\theta, K) - \bar{V}(\theta, K_\star) = \langle \nabla_K \bar{V}(\theta, K), K - K_\star \rangle_F + O(\|K - K_\star\|_F^2)$$

in a neighborhood of the optimal policy $K_\star$. In combination, these properties can be used to verify that if gradient descent is initialized with a stabilizing controller, its updates remain stable and converge to the global optimum at the rate (31).

REFERENCES

[S6] P. Boris Teodorovich, "Gradient methods for minimizing functionals," *Zh. Vychisl. Mat. Mat. Fiz.*, vol. 3, no. 4, pp. 643–653, 1963.

[S7] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Cham, Switzerland: Springer-Verlag, 2016, pp. 795–811.

[S8] Y. Sun and M. Fazel, "Learning optimal controllers by policy gradient: Global optimality via convex parameterization," in *Proc. 60th IEEE Conf. Decis. Control (CDC)*, Piscataway, NJ, USA: IEEE Press, 2021, pp. 4576–4581, doi: 10.1109/CDC45484.2021.9682821.

that there is a polynomial gap in the problem dimension in their respective sample complexities (with CE outperforming REINFORCE). The fundamental limits of policy gradient methods are further investigated and related to various system-theoretic quantities in [80]. It is still an open problem to explore whether the result of Tu and Recht [79] can be extended to more general systems/gradient-based methods.

ONLINE CONTROL

Having discussed episodic RL tasks through the lens of control, we now turn our attention to the more technically challenging setting of online adaptive control. We rely on the notion of *regret* to quantify the performance of an online algorithm. Just as in offline control, suppose the system has dynamics are of the form

$$\begin{aligned} x_{t+1} &= A_\star x_t + B_\star u_t + w_t \\ y_t &= C_\star x_t + v_t \end{aligned} \tag{33}$$

where $x_t, w_t \in \mathbb{R}^{d_{\mathsf{x}}}$, $u_t \in \mathbb{R}^{d_{\mathsf{u}}}$, $y_t, v_t \in \mathbb{R}^{d_{\mathsf{y}}}$, and $\mathrm{A}_\star \in \mathbb{R}^{\mathrm{d_x}\times\mathrm{d_x}}$, $\mathrm{B}_\star \in \mathbb{R}^{\mathrm{d_x}\times\mathrm{d_u}}$, and $C_\star \in \mathbb{R}^{d_{\mathsf{y}}\times d_{\mathsf{x}}}$. However, in contrast to the offline control setting, the learner now interacts iteratively with only a single trajectory ($N_{\mathrm{traj}} = 1, T = N_{\mathrm{tot}}$) from the system (33). The parameters of $(A_\star, B_\star, C_\star)$ are, as before, unknown to the learner. For simplicity, assume that $\{w_t\}$ and $\{v_t\}$ are mutually independent i.i.d. sequences of mean-zero sub-Gaussian random variables, with covariance matrices Σ_w and Σ_v, respectively. Most of the current literature focuses on the LQR setting, where $C = I_{d_{\mathsf{x}}}$ and $v_t = 0$. Relatively less is known about regret minimization for the partially observed setting (in which case, the noise sequences are Gaussian). In either setting, the goal in the adaptive LQR and LQG problems is to regulate the system (33) by using a policy π so as to render the following cost functional as small as possible:

$$V_T^\pi(\theta) \triangleq \mathbf{E}_\theta^\pi\left[x_T^\top Q_T x_T + \sum_{t=0}^{T-1} x_t^\top Q x_t + u_t^\top R u_t\right] \tag{34}$$

where $\mathbf{E}_\theta^\pi$ stands for expectation with respect to dynamics $\theta = (A, B, C)$ under policy π and (Q, Q_T, R) are positive definite weighting matrices. The difficulty of the task arises from the fact that the parameter θ is assumed to be a priori unknown, and hence, the optimal cost $V_T^\star(\theta_\star) \triangleq \inf_\pi V_T^\pi(\theta)$ cannot be realized. Instead, one seeks to design a policy (algorithm) π with small regret.

Regret

The regret of an algorithm measures the cumulative suboptimality accrued over the entire time horizon as compared to the optimal policy:

$$\mathcal{R}_T^\pi(\theta) \triangleq x_T^\top Q_T x_T + \sum_{t=0}^{T-1} x_t^\top Q x_t + u_t^\top R u_t - V_T^\star(\theta) \tag{35}$$

where the law of $\{x_t, u_t\}_{t=0}^T$ is specified by (θ, π). Alternatively, one may be interested in the expected regret:

$$\mathbf{E}\mathcal{R}_T^\pi(\theta) = V_T^\pi(\theta) - V_T^\star(\theta). \tag{36}$$

Note that the regret is a random quantity, whereas the expected regret is not; however, in either case, the interpretation is that one seeks to design a policy that has small cumulative suboptimality as compared to the optimal policy $\pi_\star(x) = K_\star x$, which can be computed via Riccati equations (38)–(39). Abstracting slightly, the regret of an algorithm can be thought of as the rate of convergence of an adaptive algorithm [see (40)]. Moreover, it quantifies the dual nature of control [81], [82] (in RL terminology, the exploration–exploitation tradeoff). We see in the sequel that for an algorithm to have low regret, it necessarily must generate sufficiently rich data. At a high level, by relating (35) [or (36)] to quantities of interest (such as the time horizon T, dimensional factors, and system-theoretic quantities), we gain an understanding of the statistical properties of adaptation and under which circumstances adaptation—if only in an idealized environment—is easy or hard. Also note that in the formulation (35)–(36), we compete with a policy that has good average case performance (LQR) but does not necessarily take into account robust or stability margins. While certainly important, in this survey, we do not cover robustness aspects of adaptive methods but, rather, emphasize their statistical analysis.

State Feedback Systems

For state feedback systems ($C_\star = I_{d_{\mathsf{x}}}, v_t = 0$), it has been shown by Simchowitz and Foster [36] that certainty equivalence with naive exploration (additive Gaussian noise injected into the control input) attains, with probability $1 - \delta$,

$$R_T^\pi(\theta_\star) \le c_{\mathsf{sys}} \sqrt{d_{\mathsf{x}} d_{\mathsf{u}}^2 T \log(1/\delta)}$$

for a system-dependent constant $c_{\mathsf{sys}} > 0$ and provided that T is sufficiently large (polynomial in dimension and system-dependent quantities). Their result refined an earlier result of [68] and essentially settled the question of what the optimal dependence on system dimensions and the time horizon is. A recent result due to Jedra and Proutiere [83] also shows that, up to logarithmic factors, the same rate can be attained in expectation $\mathbf{E}R_T^\pi(\theta_\star) = \tilde{O}(\sqrt{d_{\mathsf{x}} d_{\mathsf{u}}^2 T})$. Simchowitz and Foster [36] also provide a matching lower bound with $\sup_{\theta \in B(\theta_\star, \varepsilon)} \mathbf{E}R_T^\pi(\theta) = \Omega(\sqrt{d_{\mathsf{x}} d_{\mathsf{u}}^2 T})$. However, characterizing the optimal dependence on the system parameters $(A_\star, B_\star)$ is still open. For instance, there is a polynomial gap between the best-known upper bounds [36] and the best-known lower bounds [84] in regard to the dependence on $P_\star = P(A_\star, B_\star)$ [recall (27)]. A summary of the state of the art for both state feedback and partially observed systems is given in Table 3.

Certainty Equivalence

The key algorithmic idea to solve the regret minimization problem for the LQR is again CE. The idea dates back to the late 1950s [81], [82], [91] and was first analyzed in the context of adaptive control of linear models by Åström and

Wittenmark [92], in 1973. Initially, the emphasis was solely on asymptotic average cost optimality, corresponding to sublinear regret, $\mathcal{R}_T^\pi = o(T)$, in our formulation. Regret minimization was introduced to the adaptive control literature roughly a decade later by Lai [93]. Online CE LQR control takes continuously updated parameter estimates $(\hat{A}, \hat{B}, \hat{C})$ of $(A., B., C.)$ as inputs and then solves the dynamic programming problem for these estimates as if they were the ground truth. For the LQR, the dynamic programming solution has a closed-form solution in terms of the (discrete algebraic) Riccati recursion (38)–(39), which can be solved efficiently by numerical schemes. The resulting controller is then used to regulate the system. To see why the CE strategy is successful in the LQR, we note the following elementary relation between the expected regret and the Riccati recursion [84]:

$$\mathbf{E}R_T^\pi(\theta) = \sum_{t=0}^{T-1} \mathbf{E}_\theta^\pi\left[(u_t - K_t x_t)^\top (B^\top P_{t+1} B + R)(u_t - K_t x_t)\right] \quad (37)$$

where $\theta = (A, B)$, $P_t = P_t(\theta)$ and $K_t = K_t(\theta)$ are given by

$$P_{t-1} = Q + A^\top P_t A - A^\top P_t B (B^\top P_t B + R)^{-1} B^\top P_t A \quad (38)$$

$$K_t = -(B^\top P_t B + R)^{-1} B^\top P_t A \quad (39)$$

and where the terminal condition is $P_T = Q_T$. We further denote the steady-state versions of the recursion (38)–(39) by $P(A, B)$ and $K(A, B)$. It will be convenient to denote $P. \triangleq P(A., B.)$ and $K. \triangleq K(A., B.)$. Equation (37) follows from the "completing-the-square" proof of LQR optimality; see [94, Th. 11.2]. Crucially, for naive exploration policies of the form $\pi : u_t = \hat{K}_t x_t + \eta_t$ (with $\{\eta_t\}$ a mean-zero sequence of exploratory noise, independent of all other randomness), equation (37) becomes

$$\mathbf{E}R_T^\pi(\theta) = \mathbf{E}_\theta^\pi \sum_{t=0}^{T-1} \eta_t^\top \eta_t + \sum_{t=0}^{T-1} \mathbf{E}_\theta^\pi\left[x_t^\top (K_t - \hat{K}_t)^\top (B^\top P_{t+1} B + R)(K_t - \hat{K}_t) x_t\right]. \quad (40)$$

Equation (40) shows that the expected regret of a CE policy is a quadratic form in the estimation error $\hat{K}_t - K_t$. Moreover, by a stability argument, it suffices to use the steady-state versions of the Riccati recursion (38)–(39). This suggests that the CE strategy with $\hat{K}_t = K(\hat{A}, \hat{B})$ can be shown to be successful, provided that one shows that the

1) estimates $(\hat{A}, \hat{B})$ are consistent estimators of the true dynamics
2) map $(A, B) \mapsto K_t(A, B)$ is sufficiently smooth in the parameters (A, B)
3) policy π is stabilizing in that the state process x_t does not become too large.

Analogous reasoning is applicable in the high-probability regret setting, but it becomes a little more involved (see [36, Lemma 5.2]). Before we proceed, one remark is in order: (40) suggests that $\mathbf{E}\mathcal{R}_T^\pi(\theta) = O(\log T)$ should be possible. Namely, we noted in the finite-sample analysis of system identification that identification errors generally decline as $O(1/\sqrt{t})$, where t is the number of samples collected so far. Since the suboptimality bound (29) is quadratic in the identification error, the square errors decline as $O(1/t)$, and the regret induced will scale as the sum of $1/t, t = 0, \ldots, T-1$, which is of order $\log T$. We soon ask, Why do we need exploration? and see that logarithmic regret is not possible in general, for reasons of closed-loop identifiability.

Why Do We Need Exploration?

In the sketch of the CE approach presented in the preceding, we mentioned that one typically requires a perturbation η_t of the input u_t. The most common exploration strategy, known as ε-greedy exploration, uses simple additive perturbations to the control policy, yielding inputs of the form $u_t = K_t x_t + \eta_t$, as previously. More intricate exploration strategies are possible, as described in "Optimism and Thompson Sampling." To understand why such perturbations are necessary, consider again the least-squares algorithm (4). Recall that the error of the estimator $\hat{\theta}_s = (\hat{A}_s, \hat{B}_s)$ satisfies the following equation:

$$\hat{\theta}_s - \theta. = \left(\sum_{t=0}^{s-1} w_t [x_t^\top \; u_t^\top]\right)\left(\sum_{t=0}^{s-1} \begin{bmatrix} x_t \\ u_t \end{bmatrix} [x_t^\top \; u_t^\top]\right)^{-1} \quad (41)$$

TABLE 3 A summary of the results: regret minimization in adaptive control (the state of the art is in blue).

Paper	Setting	Method	Upper Bound	Lower Bound
[22]	SF: (A, B) unknown	Optimism	$\tilde{O}(\sqrt{T})$ but intractable	
[93]	SF: (A, B) unknown	CE	$\tilde{O}(T^{2/3})$	
[94]		CE		
[72]	SF: (A, B) unknown	CE	$\tilde{O}(\sqrt{T})$	
[95]		Optimism		
[39]	SF: (A, B) unknown	CE	$O(\sqrt{d_x d_u^2 T})$	$\Omega(\sqrt{d_x d_u^2 T})$
[96]	SF: A unknown	CE	$\tilde{O}(\log T)$	
	b Scalar unknown	CE		$\Omega(\sqrt{T})$
[97]	PO: (A, B, C) unknown	Gradient	$\tilde{O}(\sqrt{T})$	
[92]	SF: (A, B) unknown			$\Omega(\sqrt{d_x d_u^2 T})$
	PO: (A, B, C) unknown			$\Omega(\sqrt{T})$
[98]	SF: (A, B) unknown	CE	$O(\exp(\kappa) \times \sqrt{d_x d_u^2 T})$	$\Omega\left(\sqrt{\frac{1}{d_x} 2^\kappa T}\right)$

SF: state feedback; PO: partially observed.

provided that the matrix inverse on the right-hand side of (41) exists. As mentioned in the preceding, as long as the covariates do not grow more than polynomially with the time horizon, it can be shown, using the theory of self-normalized martingales, that the rate of convergence of $\hat{\theta}_s - \theta_\star$ is dictated by the smallest eigenvalue of the covariates matrix

$$\|\hat{\theta}_s - \theta_\star\|_{\text{op}} = \tilde{O}\left[\lambda_{\min}^{-1}\left(\sqrt{\left(\sum_{t=0}^{s-1}\begin{bmatrix}x_t\\u_t\end{bmatrix}\begin{bmatrix}x_t^\top & u_t^\top\end{bmatrix}\right)}\right)\right]. \tag{42}$$

Suppose, for the moment, $u_t \approx K_\star x_t$ in (42). In this case, the matrix

$$\sum_{t=0}^{s-1}\begin{bmatrix}x_t\\u_t\end{bmatrix}\begin{bmatrix}x_t^\top & u_t^\top\end{bmatrix} \approx \sum_{t=0}^{s-1}\begin{bmatrix}I_{d_x}\\K_\star\end{bmatrix}x_t x_t^\top \begin{bmatrix}I_{d_x} & K_\star^\top\end{bmatrix} \tag{43}$$

is nearly singular. To see this, note that $[I_{d_x} \;\; K_\star^\top]^\top$ is a tall matrix—the outer product of tall matrices is singular. Thus, the error (42) diverges if the policy is too close to the optimal policy $K_\star$; that is, the true parameters $A_\star$ and $B_\star$ are not identifiable under the optimal closed-loop policy $K_\star$. In fact, this lack of identifiability is true under any policy of the form $u_t = Kx_t$. Alternatively, the need for exploration can be seen by noting that for every perturbation $\Delta \in \mathbb{R}^{d_x \times d_u}$ and $(A(\Delta), B(\Delta))$ of the form $A(\Delta) = A_\star - s\Delta K_\star$, $B(\Delta) = B + s\Delta$ $(s \in \mathbb{R})$, the closed-loop systems $A_\star + B_\star K_\star$ and $A(\Delta) + B(\Delta)K_\star$ are identical: $A_\star + B_\star K_\star = A(\Delta) + B(\Delta)K_\star$ for all such Δ, s. Thus, from observing trajectories generated by the two systems

$$x_{t+1} = (A_\star + B_\star K_\star)x_t + w_t$$
$$x_{t+1} = (A(\Delta) + B(\Delta)K_\star)x_t + w_t$$

it is impossible to distinguish between them. The reasoning in the preceding indicates that to obtain estimates that converge sufficiently quickly to the true parameters $(A_\star, B_\star)$, exciting inputs that lead to exploration away from the optimal policy $K_\star$ are necessary.

Do We Actually Need to Identify the True Parameters $(A_\star, B_\star)$?

The answer to this question is in the affirmative. To see this, we recall from [36, Lemma 2.1] that

$$\frac{d}{ds}K(A_\star - s\Delta K_\star, B_\star + s\Delta)\Big|_{s=0} = -(B_\star^\top P_\star B_\star + R)^{-1}\Delta^\top P_\star (A_\star + B_\star K_\star). \tag{44}$$

As long as $(A_\star + B_\star K_\star)$ in the matrix on the right-hand side of (44) is nonzero, this implies that there exists a confusing parameter variation (which is not closed-loop distinguishable) that has a different optimal policy. Hence, one necessarily must identify the true parameters $A_\star$ and B_Δ in the adaptive control problem.

A Historical Tangent on Identifiability

Closed-loop identifiability issues are well known in the system identification literature [95], [96], [97]. Indeed, in the LQR setting, Polderman [97] gives an elegant geometric argument showing that the true parameters need to be identified. It is also interesting to note that, precisely because the minimum variance controller $(Q = I, R = 0)$ is closed-loop identifiable [95] (in contrast to the more general LQR controller), logarithmic regret can be achieved in this setting [93]. Reiterating the point in the preceding: the reason for the necessity of the "exploratory signals" η_t in (40) is precisely a lack of closed-loop identifiability.

Returning to the estimation guarantee (42), note that an i.i.d. sequence η_t of rescaled isotropic noise of magnitude (standard deviation) $t^{-\alpha}$ is sufficient to guarantee parameter recovery at the rate $\|\hat{\theta}_t - \theta_\star\|_{\text{op}} = \tilde{O}(t^{\alpha-1/2})$. In this case, smoothness (combined with a naive Taylor expansion) suggests that $\|K(\hat{A}_t, \hat{B}_t) - K_\star\|_{\text{op}} = \tilde{O}(t^{\alpha-1/2})$. Balancing the two terms in (40) demonstrates that $\alpha = 1/4$ leads to $R_T = \tilde{O}(\sqrt{T})$, which is optimal. While the reasoning in the

Optimism and Thompson Sampling

Alternative exploration strategies include optimism and Thompson sampling. Indeed, the first complete treatment of regret minimization in the linear quadratic regulator, due to Abbasi-Yadkori and Szepesvári [21], relies on the principle of *optimism in the face of uncertainty* (*OFU*). Just as in the certainty-equivalent (CE) approach discussed in the main text, OFU is based on constructing parameter estimates $(\hat{A}, \hat{B})$. However, OFU also maintains a (tuned) confidence interval for these estimates. The adaptive control law is then obtained by selecting the most *optimistic* parameter and CE control law—those resulting in the lowest estimated cost—in this confidence interval. The original algorithm of [27] was not computationally tractable, but this was later remedied by [S9]. A related method, *Thompson sampling*, is studied in [S10] and [S11]. Even though these strategies, in principle, are more sophisticated, to date, the tightest bounds have been proved for the simple input perturbation approach described in the main text [36].

REFERENCES

[S9] M. Abeille and A. Lazaric, "Efficient optimistic exploration in linear-quadratic regulators via Lagrangian relaxation," *in Proc. 37th Int. Conf. Mach. Learn.*, PMLR, 2020, pp. 23–31.

[S10] Y. Ouyang, M. Gagrani, and R. Jain, "Control of unknown linear systems with Thompson sampling," *in Proc. 55th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Piscataway, NJ, USA: IEEE Press, 2017, pp. 1198–1205, doi: 10.1109/ALLERTON.2017.8262873.

[S11] M. Abeille and A. Lazaric, "Improved regret bounds for Thompson sampling in linear quadratic control problems," *in Proc. 35th Int. Conf. Mach. Learn.*, PMLR, 2018, pp. 1–9.

preceding about the necessity of the perturbations η_t is entirely heuristic, it can be made formal and will be discussed further in the following section.

Regret Lower Bounds

We now argue that the scaling $R_T^\pi = \Theta(\sqrt{d_{\mathrm{x}} d_{\mathrm{u}}^2 T})$ is optimal for state feedback systems by finding matching lower bounds. The modern approach to lower bounds, or fundamental performance limits, for sequential decision-making problems seeks to characterize local minimax lower bounds. Such bounds quantify statements of the form "there exists no algorithm that uniformly outperforms a certain fundamental limit across a small (local) neighborhood of problem parameters." For the regret minimization problem, such lower bounds typically take the form

$$\sup_{\theta \in B(\theta_\cdot, \varepsilon)} \mathbf{E} R_T^\pi(\theta) \geq f(\theta_\cdot, \varepsilon, T) \tag{45}$$

for some $\varepsilon > 0$, some function f, and every (causal) policy π. The lower bound (45) states that the worst-case expected regret over a neighborhood of the true parameter is lower bounded by some function of the instance parameter $\theta_\cdot$ and the horizon T. The appearance of $\sup_{\theta \in B(\theta_\cdot, \varepsilon)}$ in inequality (45) is not restrictive—while such lower bounds are "worst-case," one can typically allow for $\varepsilon \to 0$. In other words, such lower bounds are applicable to all algorithms that are, in some sense, robust to infinitesimal perturbations in the model parameter $\theta_\cdot$, a rather mild criterion. Put differently, a lower bound of the form (45) for vanishing $\varepsilon \to 0$ states that there exists no algorithm that uniformly outperforms the lower bound in an infinitesimal neighborhood.

Regret Lower Bounds via Reduction to Bayesian Estimation

To arrive at a local minimax lower bound (45), suppose, for simplicity, that $QT = P$ so that (37) becomes

$$\begin{aligned} &\sup_{\theta \in B(\theta_\cdot, \varepsilon)} \mathbf{E} R_T^\pi(\theta) \\ &= \sum_{t=0}^{T-1} \mathbf{E}_\theta^\pi \left[(u_t - K(\theta) x_t)^\top (B^\top(\theta) P(\theta) B(\theta) + R)(u_t - K(\theta) x_t) \right] \\ &\geq \lambda_\varepsilon \sup_{\theta \in B(\theta_\cdot, \varepsilon)} \sum_{t=0}^{T-1} \mathbf{E}_\theta^\pi \left\| u_t - K(\theta) x_t \right\|_2^2 \end{aligned} \tag{46}$$

where $\lambda_\varepsilon = \min_{\theta \in B(\theta_\cdot, \varepsilon)} \lambda_{\min}(B^\top(\theta) P(\theta) B(\theta) + R) \geq \lambda_{\min}(R) > 0$. The next step is crucial: we relax the supremum in inequality (46) by a introducing a prior λ over $\theta \in B(\theta_\cdot, \varepsilon)$. The exact choice of λ is not particularly interesting, and its influence on the final bound can be made to vanish. By weak duality, we have for any such λ that

$$\sup_{\theta \in B(\theta_\cdot, \varepsilon)} \mathbf{E} R_T^\pi(\theta) s \gtrsim \sum_{t=0}^{T-1} \mathbf{E}_{\theta \sim \lambda} \mathbf{E}_\theta^\pi \left\| u_t - K(\theta) x_t \right\|_2^2. \tag{47}$$

The key insight is now that the quantity $\inf_{u_t} \mathbf{E}_{\theta \sim \lambda} \mathbf{E}_\theta^\pi \left\| u_t - K(\theta) x_t \right\|_2^2$ is simply the minimum mean-square error for estimating the random variable $K(\theta) x_t$, where θ is drawn according to the prior distribution λ. Although it does require rather a few intermediate steps [84, Th. 4.1], one can, in principle, lower bound the right-hand side of inequality (47) by using estimation-theoretic lower bounds, such as the Bayesian Cramér–Rao inequality [61], namely, Van Trees' inequality The leading term in such lower bounds is the inverse of the Fisher information:

$$\mathtt{I}_p(\theta) = \mathbf{E}\left[\sum_{t=0}^{T-1} \left(\begin{bmatrix} x_t \\ u_t \end{bmatrix} \begin{bmatrix} x_t^\top & u_t^\top \end{bmatrix} \right) \otimes \sum_w^{-1} \middle| \, \theta \right]. \tag{48}$$

Heuristically, as $\varepsilon \to 0$, for two problem-dependent constants $c(\theta)$, $c'(\theta)$,

$$\sup_{\theta \in B(\theta_\cdot, \varepsilon)} \mathbf{E} R_T^\pi(\theta) \geq T \times c(\theta_\cdot) \lambda_{\min}(\mathbf{E}_{\theta_\cdot}^\pi \mathtt{I}_p(\theta_\cdot) + c'(\theta_\cdot))^{-1}. \tag{49}$$

The reason the constant $c(\theta_\cdot)$ is nonzero is a consequence of the derivative calculation (44). This expression concludes that the Jacobian terms discussed in "Van Trees' Inequality and Fisher Information" are invertible. Further, it is instructive to note that the expression inside the conditional expectation in (48) is proportional to the leading term in the estimation error (41) related to the recovery of the parameter $\theta = (A, B)$. As argued in the preceding,

Van Trees' Inequality and Fisher Information

Van Trees' inequality is a mean-square-error lower bound for Bayesian estimation problems. Suppose the learner seeks to estimate a smooth function $\psi(\theta)$ of a parameter θ. The learner is given access to a sample Z, which is drawn conditionally from a density $p(z \,|\, \theta)$ and has access to a prior $\lambda(\theta)$. To state Van Trees' inequality, define the Fisher information as

$$\mathtt{I}_p(\theta) \triangleq \int [\nabla_\theta \log p(z \,|\, \theta)] [\nabla_\theta \log p(z \,|\, \theta)]^\top p(z \,|\, \theta) \, dz$$

and the prior information as

$$\mathtt{J}(\lambda) \triangleq \int [\nabla_\theta \log \lambda(\theta)] [\nabla_\theta \log \lambda(\theta)]^\top \lambda(\theta) \, d\theta.$$

Under a few relatively mild regularity conditions, Van Trees' inequality states that any estimate using Z satisfies the lower bound

$$\mathbf{E}[(\hat{\psi} - \psi(\theta))(\hat{\psi} - \psi(\theta))^\top] \geq \mathbf{E} \nabla_\theta \psi(\theta) [\mathbf{E} \mathtt{I}_p(\theta) + \mathtt{J}(\lambda)]^{-1} \mathbf{E} [\nabla_\theta \psi(\theta)]^\top$$

where $\mathbf{E}$ denotes expectation with respect to $p(y, \theta) = p(y \,|\, \theta) \lambda(\theta)$. For these purposes, note that the Fisher information $\mathtt{I}_p(\theta)$ for $Z = \{x_t, u_t\}_{t=0}^{T-1}$, with $x_{t+1} = Ax_t + Bu_t + w_t$ and $\theta = \mathrm{vec}(A, B)$, is equal to

$$\mathtt{I}_p(\theta) = \mathbf{E}\left[\sum_{t=0}^{T-1} \left(\begin{bmatrix} x_t \\ u_t \end{bmatrix} \begin{bmatrix} x_t^\top & u_t^\top \end{bmatrix} \right) \otimes \Sigma_w^{-1} \middle| \, \theta \right].$$

following (43), the optimal policy $u_t = Kx_t$ renders the matrix (48) singular, and so one needs to deviate from this policy to consistently estimate the parameter $\theta = (A, B)$. In fact, it can be shown that the expected regret is an upper bound for the Fisher information (48):

$$\lambda_{\min}(\mathbf{E}_\theta^\pi \mathtt{I}_p(\theta)) \leq c''(\theta)\mathbf{E}R_T^\pi(\theta) \tag{50}$$

for a third problem-dependent constant $c''(\theta)$; see [84, Lemma 3.6]. This offers a slight change of perspective: the expected regret (36) acts as a constraint on the set of possible experiment designs available to the learner. This idea has also been explored from the perspective of regret *upper bounds* in [98]. Balancing the upper and lower bounds on the Fisher information in terms of the regret, as in the heuristic inequalities (49)–(50), yields that the optimal scaling must be $\sqrt{T}$. In particular, any policy attaining expected regret on the order of magnitude $O(\sqrt{T})$ generates a dataset in which the smallest eigenvalue of the Fisher information is $O(\sqrt{T})$. Hence, identification of the parameter $\theta_\star = (A_\star, B_\star)$ can occur no faster than at the rate $O(1/\sqrt{T})$ for a regret-optimal policy, by which we deduce that the optimal rate is $\Omega(\sqrt{T})$. To obtain the correct dimensional dependence in the lower bound $\Omega(\sqrt{d_{\mathtt{x}} d_{\mathtt{u}}^2 T})$, this argument needs to be slightly refined. Namely, we note that it, in fact, is not just the smallest eigenvalue of $\mathtt{I}_p(\theta)$ that is zero for laws of the form $u_t = Kx_t$ but all the smallest $d_{\mathtt{x}} d_{\mathtt{u}}$-many eigenvalues. To see this, note that the entire linear manifold $\{(A, B) : A + BK_\star = A_\star + B_\star K_\star\}$ corresponds to parameters lacking PE in a closed loop. As mentioned in the preceding, the optimal dimensional scaling of regret for feedback systems has been settled by [36]. However, there is currently a gap in our understanding of the best possible scaling of the regret in terms of key system-theoretic quantities. In particular, tight bounds for the scaling in terms of the solution $P_\star$ to the steady-state Riccati equation are unavailable; the best known upper bound is due to [36, Th. 2] and is of order $\sqrt{\|P_\star\|_{\mathrm{op}}^{11}}$, whereas the best known lower bound is of order $\sigma_{\min}(P_\star)$ [84, Corollaries 4.2 and 4.3]. Note that ascertaining the exact optimal dependence of the regret on $P_\star$ and other system-theoretic quantities in the LQR remains an open problem.

Quantity $P_\star$ Can Be Exponential in the Dimension

We saw in the preceding that if one regards system-theoretic parameters as "dimension-less," the optimal dimensional dependency for the state feedback regret minimization scenario is polynomial in $d_{\mathtt{x}}$ and $d_{\mathtt{u}}$. We now see that these system-theoretic quantities can be rather significant. To this end, consider the following system, which consists of two independent subsystems:

$$A = \left[\begin{array}{c|ccc|cc} 0 & 0 & 0 & & 0 & 0 \\ \hline 0 & 1 & 1 & & 0 & 0 \\ & & & \ddots & & \\ 0 & 0 & 0 & & 1 & 1 \\ 0 & 0 & 0 & & 0 & 1, \end{array}\right] B = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & 0 \\ \vdots & \\ 0 & 1 \end{array}\right]. \tag{51}$$

The first subsystem (A_1, B_1), corresponding to the top and leftmost part of the arrays in (51), is just a simple memoryless system. The second subsystem (A_2, B_2) is an integrator of order $d_{\mathtt{x}} - 1$. The system (51) is decoupled but is very sensitive to misspecification in the coupling, due to the integrator component's potential for error amplification. Moreover, the solution $P_\star(A_2, B_2)$ is on the order $2^{d_{\mathtt{x}}}$ [90, Lemma 9]. Using this, one can construct a local minimax regret lower bound for the instance (A, B) [system (51)] with scaling:

$$\sup_{\theta \in B((A,B),\varepsilon)} \mathbf{E}R_T^\pi(\theta) = \Omega(2^{d_{\mathtt{x}}}\sqrt{T}).$$

A more general statement is given in [90, Th. 3]. While the particular system (51) has exponential complexity in the state dimension $d_{\mathtt{x}}$, it establishes a more general phenomenon: the controllability index κ—the number of steps it takes to reset a noise-free system to the origin—can be used to characterize the local minimax regret, and this dependence is exponential (see also Table 3). The preceding discussion leads to two conclusions:

1) Learning to control can be hard; exponential complexity in the dimension can arise, for example, as simply as integrators.
2) To appreciate this hardness, we need to understand the role of control-theoretic quantities, such as $P_\star$.

Partially Observed Systems

While our current understanding of the state feedback setting is relatively complete, less is known when the learner has access only to a measured output and not the actual system state. In the state feedback setting, we know that the correct scaling with time is $\sqrt{T}$, that the dimensional dependence is $\sqrt{d_{\mathtt{x}} d_{\mathtt{u}}^2}$, and that the key system-theoretic quantity is $P_\star$. In contrast, in the partially observed setting, we currently know only that the correct scaling with the time horizon is $\sqrt{T}$. Determining the correct instance-specific scaling, and which quantities are key to this, is an open problem. Moreover, no current approach can handle the general LQG cost structure (34) but instead applies to the criterion

$$\tilde{V}_T^\pi(\theta) \triangleq \mathbf{E}_\theta^\pi\left[\sum_{t=0}^{T-1} y_t^\top Q y_t + u_t^\top R u_t\right].$$

With these caveats in mind, we now sketch an elegant approach due to [89] and based on the classical Youla parameterization [99], [100], leading to $\tilde{O}(\sqrt{T})$ regret for partially observed systems.

Disturbance Feedback Control

Unrolling the dynamics (33), it is straightforward to verify that

$$y_t = e_t + \sum_{s=0}^{t-1} C_\star A_\star^{t-s-1} w_s + \sum_{s=0}^{t-1} C_\star A_\star^{t-s-1} B_\star u_s \tag{52}$$

for some error signal e_t decaying exponentially fast to zero for stable systems. The approach as sketched here requires $\rho(A_\star) < 1$ but can be extended to open-loop unstable

systems [89, Appendix C]. The representation (52) suggests that there are two separate components to the input–output dynamics. The first component,

$$y_t^{\mathsf{nat}} = e_t + \sum_{s=0}^{t-1} C_\star A_\star^{t-s-1} w_s \tag{53}$$

is referred to as "nature's y" and is a counterfactual object representing the evolution of the output in the absence of controller inputs. The second component is simply the discrete convolution of the inputs $u_{0:t-1}$ with the system Markov parameters $G_\star^{0:t-1}$, where $G_\star(s) = C_\star A_\star^s B_\star$. Hence, $y_t = y_t^{\mathsf{nat}} + G_\star^{0:t-1} * u_{0:t-1}$. With these preliminaries established, for a sequence of matrices $\{M_s\}_{s=0}^{m-1}$ [89], define disturbance response controllers (DRCs) of order m as controllers of the form

$$u_t = \sum_{s=0}^{m-1} M_s y_{t-s}^{\mathsf{nat}}. \tag{54}$$

Notice that since $y_t^{\mathsf{nat}} = y_t - \Sigma_{s=0}^{t-1} C_\star A_\star^{t-s-1} B_\star u_s$, these are admissible causal controllers by construction—had the dynamics $(A_\star, B_\star, C_\star)$ been known, we would have been able to execute controllers of the form (54). It can be shown that controllers of the form (54) can approximate linear dynamic controllers, such as the separation principle solution to the LQG (a Kalman filter with an LQR controller).

Regret Bounds for Partially Observed Systems

The following algorithm combines the convex Youla-like parameterization (54) with modern online convex optimization [101]. In particular, Simchowitz et al. [89] propose an algorithm in which they

1) inject exploratory noise for a period of length proportional to $\sqrt{T}$
2) use this dataset to estimate the Markov parameters M
3) for the remainder of the horizon, compute estimates of nature's y (53) using the estimated Markov parameters
4) use the estimated nature's y to run online (projected) gradient descent on the parameters M_s of the disturbance feedback controller.

Simchowitz et al. [89] show that for a properly tuned order m of DRCs, the approach outlined in the preceding yields $\tilde{O}(\sqrt{T})$ regret. While, in this setting, there is no general lower bound to date, the authors of [84] have shown that $\Omega(\sqrt{T})$ regret is unavoidable in the worst case by considering instances with a large input dimension.

Logarithmic Regret?

It is also interesting to note that for an alternative notion of regret, in which the learner competes with the best *persistently exciting* policy instead of the optimal policy, [102] has shown that logarithmic regret is possible in the partially observed setting. Note, however, that the optimal LQG policy might not necessarily be persistently exciting. Indeed, known lower bounds show that it is not persistently exciting in 1) the state feedback setting [see (43)] and 2) the partially observed setting for certain large-input-dimension systems. Thus, it is an open problem to characterize the relation between the regret definition (35) and the one defined in [102]. Note that a related situation arises in the state feedback setting if the learner is given access to the precise value of $B_\star$. In this case, it suffices to identify the matrix $A_\star$, which is identifiable in a closed loop given knowledge of $B_\star$. Cassel et al. [88] show that this observation leads to logarithmic regret—against the optimal controller—if $B_\star$ is known a priori. A related problem where logarithmic regret is possible is that of adaptive Kalman filtering or online prediction [45], [103], [104], [105]. The objective is to predict future observations y_k online based on the past $y_{k-1}, u_{k-1}, \ldots, y_0, u_0$. Since the only goal is prediction, the cost of control does not enter the objective. Interestingly, for this problem, it is possible to attain logarithmic regret [103], [104], [105]. Hence, we can learn the Kalman filter online with a smaller regret than that achievable in online LQR control. In light of our discussion, this is hopefully no longer surprising. In the LQR problem, we need to inject additional exploratory signals into the system, which also affects the cost of control. In the prediction problem, exploration is "free," as the cost of control does not affect prediction performance. In fact, we can predict even without PE [104]; informally, if the covariates lie on a certain subspace, so will their future versions.

Open Problem 4

Provide matching upper and lower bounds on either the regret (35) or the expected regret (36). In the partially observed setting, we currently do not even know the correct dimensional dependence (or what the correct notion of the dimension is, although it is to be suspected that this is related to the order of the system and the input and output dimensions d_{u} and d_{y}). To resolve this problem, it is required to find a function f such that for a universal constant $c_1 > 0$ independent of all problem parameters,

$$\mathcal{R}_T^{\pi}(A_\star, B_\star) \le c_1 f(A_\star, B_\star, C_\star, Q, R, \Sigma_{\mathsf{W}}, \Sigma_V, T)$$

for some specific algorithm π and T sufficiently large with high probability (or in expectation). A resolution will also provide a matching lower bound, which for some $\varepsilon = o_T(1)$ and some constant $c_2 > 0$ depending only on ε, establishes that

$$\sup_{A,B \in B((A_\star, B_\star), \varepsilon)} \mathcal{R}_T^{\pi}(A, B) \ge c_2 f(A_\star, B_\star, C_\star, Q, R, \Sigma_{\mathsf{W}}, \Sigma_V, T)$$

for all algorithms π and T sufficiently large with at least constant probability (or in expectation). A partial resolution applying only to state feedback systems, thus determining the correct dependence on system-theoretic quantities, is also of interest.

SUMMARY AND DISCUSSION

We have provided a tutorial survey of recent advances in statistical learning for control. One of the key takeaway messages is that we now have a relatively complete picture of the learning problem in fully observed linear dynamical systems, both in terms of system identification (as summarized in Table 1) and in terms of regret minimization (as summarized in Table 3). We have also provided an overview and listed a number of open problems with respect to partially observed extensions of the previously mentioned results. As exciting as the developments over the past few years in this field have been, there is still much work to be done. With this mind, we now outline some future directions we believe are important for the field to consider as next steps.

Future Directions

Control-Oriented Identification

In finite-sample analysis of system identification, we studied methods of obtaining high-probability bounds on the parameter estimation error of the form

$$\left\| \hat{A}_T - A_\star \right\|_{\mathsf{op}} \le \varepsilon$$

where $\hat{A}_T$ is the output of the least-squares algorithm (4). Similar bounds can be obtained for the other state parameters as well. As discussed in the "Confidence Ellipsoids" section, the operator norm picks up the worst-case direction, which is the most difficult to identify. As shown in [24], the sample complexity of identifying the worst-case direction can grow very large for certain systems. However, a question that arises is whether this worst-case direction affects control: *Does the bottleneck of identification, that is, the worst direction, affect control design? Do we always need to identify everything?* Consider, for example, the following system:

$$A_1 = \begin{bmatrix} 0 & \alpha & 0 \\ 0 & 0 & \beta \\ 0 & 0 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \Sigma_w = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}^\top$$

where only α and β are unknown. Let the control objective be stabilization by state feedback, that is, finding a feedback gain K such that the closed-loop system $A + BK$ is asymptotically stable. The only way to excite $x_{k,2}$ is via $x_{k,3}$; the coupling coefficient β determines the degree of excitation. Note that as the coupling β goes to zero, the excitation of $x_{k,2}$ becomes smaller. As a result, if β is very small, it is very difficult to identify the parameter α, and the complexity of the system identification increases with β^{-1}. However, it is trivial to stabilize the system, even without knowledge of α, for example, with $K = 0$. In this particular example, the worst direction of the identification error is not relevant for stabilization. Hence, the complexity of stabilization should be independent of β^{-1}. On the other hand, consider system

$$A_2 = \begin{bmatrix} 1 & \alpha & 0 \\ 0 & 0 & \beta \\ 0 & 0 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \Sigma_w = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}^\top$$

where now the first state has marginally stable dynamics. Unfortunately for this pathological example, it is, in fact, necessary to identify α to stabilize the system (this example is adapted from [90]) suffering from complexity that scales with β^{-1}. In particular, we cannot stabilize the system unless we identify the sign of α, showing that for some systems, the worst direction of the identification error matters. The preceding example shows a system for which stabilization depends on an *identification bottleneck*. However, it seems that the constructed systems are artificial or pathological. It is an open problem to characterize the conditions under which we can avoid such corner cases. Similar questions have been previously studied in the context of control-oriented identification or identification for control [106]. In many situations of practical interest, we need to identify only the part of the model that matters for a specific closed-loop objective. In this case, it is reasonable to tune the identification toward the objective for which the model is to be used, that is, to ensure that the model error is "orthogonal" to the control objective. This is particularly important in the case of agnostic learning, that is, when there is no "true model" and the model class can only approximate the system, which is typically the case in practice.

Learning With Structure and Regularization

In many practical situations, certain structural properties of the system to be identified and controlled are known a priori. For instance, when trying to learn a networked system, the engineer might have prior knowledge that interconnections between states are relatively sparse. Other examples of relevant structural priors include low-order (as captured by the rank of a system Hankel matrix) or physical properties, such as passivity and dissipativity.

Sparsity

In the case of a linear dynamical system, sparsity amounts to the matrix $A_\star$ in the dynamics $x_{t+1} = A_\star x_t + w_t$ having many zero entries; that is, $A_\star$ will be sparse and have only $s \ll d_x^2$ nonzero entries. Many modern networked systems have the property that they are large-scale but not maximally connected, leading to a high-dimensional state vector with a sparse $A_\star$. There are many other examples that fall into this category, including snake-like robots, which can be modeled by an integrator-like structure:

$$A_{\mathsf{snake}} = \begin{bmatrix} a_{11} & a_{12} & 0 & 0 & & 0 \\ 0 & a_{22} & a_{23} & 0 & & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \end{bmatrix}.$$

The matrix A_{snake} has only $s = 2d_{\mathsf{x}} \ll d_{\mathsf{x}}^2$ many nonzero entries, and so one is justified to hope for a polynomial speedup in the sample complexity of the system identification as compared to the standard minimax rate achieved by the least-squares estimator.

In such high-dimensional situations, running linear regression, which suffers a minimax rate of convergence proportional to d_{x}^2 in the Frobenius norm (proportional to d_{x} in the operator norm), is not sample efficient or might not even be tractable. To alleviate this issue, Fattahi et al. [34] analyze the least absolute shrinkage and selection operator (LASSO) estimator as applied to system identification. Recall that the ℓ^1 norm of a vector $v = (v_1, \ldots, v_{\mathsf{d}}) \in \mathbb{R}^{\mathsf{d}}$ takes the form $\|v\|_{\ell^1} = \Sigma_{i=1}^{\mathsf{d}} |v_i|$. The LASSO penalizes the least-squares solution by this norm by using a fixed regularization parameter $\lambda > 0$ and takes the form

$$\hat{A} \in \operatorname*{argmin}_{A \in \mathbb{R}^{d_{\mathsf{x}}}} \left\{ \frac{1}{T} \sum_{t=0}^{T-1} \| x_{t+1} + Ax_t \|_2^2 + \lambda \| \operatorname{vec}(A) \|_{\ell^1} \right\}. \tag{55}$$

It is by now well known that ℓ^1 regularization promotes sparse least-squares solutions [107], [108]. The authors of [34] show that the LASSO also avoids polynomial dependence on the state dimension for linear dynamical systems. Unfortunately, the rate in [34] degrades with the stability of the system—precisely that which we sought to avoid in our discussion of the finite-sample analysis of system identification by leveraging the PE and small-ball bounds. Moreover, by instantiating recent results in [109], it can be shown that the minimax rate (in the Frobenius norm) over the class of s-sparse linear dynamical systems is no more than $\tilde{O}(\sqrt{s\sigma_w^2 / \lambda_{\min}(\Gamma_T)})$, where Γ_T is as in (7) (with $B_\cdot = 0$). Unfortunately, instantiating [109] does not yield an effective algorithm and reduces to running $\binom{d_{\mathsf{x}}^2}{s} = O(d_{\mathsf{x}} \exp(2s))$ separate regressions, each one over an s-dimensional submanifold. This quickly becomes intractable, even for rather moderate cases of the degree of sparsity s.

Open Problem 5

Studying the tension between dependence on mixing time (stability) and computational intractability is an exciting direction for future work. Can we refine existing analyses of the LASSO (or provide some other polynomial algorithm) to match minimax rates, or is there a fundamental computational barrier introduced by sparsity? Resolving this issue may well require the development of new tools since existing analyses of the LASSO in the i.i.d. setting invariably depend on the condition number of the covariates matrix [23], [110], which, for a linear dynamic system, is proportional to the mixing time (degree of stability), leading to suboptimal rates.

Low-Order Models

Sparsity, as discussed in the preceding, is also relevant when estimating input–output models of unknown order. For example, consider the following model:

$$y_{t+1} = \sum_{j=0}^{t} A_j y_{t-j} + \sum_{j=0}^{t} B_j u_{t-j} + w_t, \quad y_j = 0 \text{ for } j \le 0. \tag{56}$$

In this scenario, there is no nontrivial upper bound on the lag order available to the engineer, and it may be as large as the entire horizon T. Converting the process (56) into state-space form and running least squares is not tractable: recall that the minimax rate of convergence depends on the ratio of the number of unknown parameters and the number of samples (in this case, given by the horizon T). Without further assumption, this ratio is constant in the worst-case for model (56). However, if there is hope that the true model is of low order so that many of the $\{A_j, B_j\}$ are zero, a variation of the LASSO (55) may also be appropriate for model selection in this scenario.

Low-Rank Models

A more sophisticated notion of model order than discussed in the preceding section is that of the Hankel matrix rank (McMillan degree). Let $h_\cdot = [C_\cdot B_\cdot \; C_\cdot A_\cdot B_\cdot \; C_\cdot A_\cdot^2 B_\cdot \; \ldots]$ denote the impulse response (matrix) associated to the tuple $(A_\cdot, B_\cdot, C_\cdot)$, and notice that model (1) can be written as

$$y_t = h * u_{0:t-1} + \eta_t$$

where $*$ denotes discrete convolution and $\{\eta_t\}$ is some (not necessarily i.i.d.) noise sequence. Denote by $\mathcal{H}$ the Hankel (linear) operator mapping impulse responses to Hankel matrices. The nuclear norm of a matrix $M \in \mathbb{R}^{\mathsf{d} \times \mathsf{d}}$ is $\|M\|_* = \Sigma_{i=1}^{\mathsf{d}} \sigma_i(M)$. This norm plays a similar role to the ℓ^1 norm but promotes low-rank solutions rather than sparse solutions [108]. Since the rank of the Hankel matrix $\mathcal{H}(h_\cdot)$ coincides with the McMillan degree of the system (1), it is natural to consider the following nuclear norm-regularized problem (see, for example, [56]):

$$\hat{h} \in \operatorname{argmin}_h \left\{ \frac{1}{T} \sum_{t=0}^{T-1} \| y_t + h * u_{t-1:0} \|_2^2 + \lambda \| \mathcal{H}(h) \|_* \right\}. \tag{57}$$

As of the writing of this article, no finite-sample analysis exists for the nuclear norm-regularized estimator (57).

Learning for Nonlinear Identification and Control

While the vast majority of the literature on statistical learning for identification and control has been on linear systems, most real systems are not linear. Learning in linear dynamical systems escapes many nonlinear phenomena and does not capture one of the most fundamental issues in modern machine learning: distribution shift. For linear models, parameter recovery is always possible as long as

the average covariance matrix of the covariates is sufficiently nondegenerate (invertible) and the rate of the parameter recovery is (asymptotically) completely described by the second-order statistics of the process under investigation. Put differently, all equilibrium points of a linear system are (dynamically) equivalent. This stands in stark contrast to more general nonlinear systems in which, in the worst case, learning the behavior around one equilibrium point gives no information about the behavior of the system in other regions of the state space. Moreover, recent advances in learning and estimation for nonlinear dynamics bypass these issues of distribution shift by either considering models that behave almost linearly [111], [112], [113], [114], [115] or by sidestepping the issue entirely and considering only a prediction error associated with the invariant measure of the system [109], [116]. For statistical learning to be truly informative for downstream control applications, a more integrated understanding of learnability, nonlinear dynamic phenomena, and control-theoretic notions, such as incremental stability or contraction, are needed [117], [118], [119].

Realizability and Approximation

Existing work on learning in dynamical systems makes strong realizability assumptions. For instance, it is often assumed that the true model is generated by a linear dynamical system of the form (1) driven by i.i.d. mean-zero (or martingale difference) noise. Even if one considers more complicated nonlinear models, such additive mean-zero noise models completely sidestep bias or misspecification challenges. This is significant since ignoring this issue might mean that existing analyses are overly optimistic. The work in [120] shows that in the worst case, misspecification in a simple linear regression model leads to a deflated sample complexity by a factor linear in the mixing time of the covariates process. This stands in stark contrast to the results in [4], in which linear regression over a well-specified model class is analyzed completely without reference to mixing. While the fundamental limits in [120] may seem discouraging at first, they are worst-case and may be avoidable by introducing further regularity assumptions. As a first step, one could analyze the sample complexity of recovering the best linear approximation to an almost linear autoregression, for example, adding a small nonlinearity, or considering a generalized linear model with a nearly isometric link function.

Structured Nonlinear Identification

A host of new opportunities present themselves in structural nonlinear identification as compared to the linear setting. While sparse and low-rank structures are certainly of interest and applicable to learning in nonlinear dynamical systems, there are other exciting (and arguably more fundamentally system-theoretic) alternatives. For instance, one might ask how properties such as passivity or dissipativity affect the minimax rate of estimation and whether there are efficient algorithms that might take advantage of this. More concretely, one might be interested in the 1D autoregression $x_{t+1} = f_\star(x_t) + w_t$ and seek to identify $f_\star$ under the physically motivated hypothesis that $f_\star$ is the negative gradient of an unknown convex potential. Taking advantage of structure may also be inherently more important in nonlinear identification since, otherwise, the curse of dimensionality is quick to present itself. For instance, in the model

$$y_t = f_\star(x_t) + w_t$$

running regression over the hypothesis class $\mathscr{F} = \{f : \mathbb{R}^{d_x} \to [0, 1] \subset \mathbb{R} \text{ and } f \text{ is } k \text{ smooth}\}$ incurs a minimax rate that degrades *exponentially* with a large d_x.

ACKNOWLEDGMENT

Ingvar Ziemann is supported by a Swedish Research Council International Postdoc Grant. The work of Nikolai Matni is supported, in part, by National Science Foundation (NSF) Award CPS-2038873, NSF CAREER Award ECCS-2045834, and a Google Research Scholar Award. The authors are grateful to three anonymous reviewers for excellent feedback. Anastasios Tsiamis and Ingvar Ziemann contributed equally to this work.

AUTHOR INFORMATION

Anastasios Tsiamis (atsiamis@control.ee.ethz.ch) received the diploma degree in electrical and computer engineering from the National Technical University of Athens, Athens, Greece, in 2014 and the Ph.D. degree in electrical and systems engineering from the University of Pennsylvania, Philadelphia, PA, USA, in 2022. He is currently a postdoctoral scholar with the Department of Information Technology and Electrical Engineering, ETH Zürich, 8092 Zürich, Switzerland. His research interests include statistical learning for control, risk-aware control and optimization, and networked control systems. He was a finalist for the International Federation of Automatic Control (IFAC) Young Author Prize at the 2017 IFAC World Congress and a finalist for the Best Student Paper Award at the 2019 American Control Conference. He was also a coauthor of a paper that won the 2022 IEEE Conference on Decision and Control Best Student Paper Award. He is a Member of IEEE.

Ingvar Ziemann received the Ph.D. degree in 2022 from the Division of Decision and Control Systems, KTH Royal Institute of Technology, Stockholm, Sweden, under the supervision of Henrik Sandberg. His research is centered on using statistical and information-theoretic tools to study learning-enabled control methods, with a current interest in studying how learning algorithms generalize in the context of dynamical systems. Prior to starting his Ph.D., he obtained two sets of M.S. and B.S. degrees in mathematics (Stockholm University/KTH) and in economics and

finance (Stockholm School of Economics). He was the recipient of a Swedish Research Council International Postdoc Grant, the 2022 IEEE Conference on Decision and Control Best Student Paper Award, and the 2017 Stockholm Mathematics Center Excellent Master Thesis Award. He is a Student Member of IEEE.

Nikolai Matni is an assistant professor in the Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA, where he is also a member of the Department of Computer and Information Sciences (by courtesy); the General Robotics, Automation, Sensing, and Perception Lab; the Penn Research in Embedded Computing and Integrated Systems Engineering Center; and the Applied Mathematics and Computational Science graduate group. He has held positions as a visiting faculty researcher at Google Brain Robotics, New York City, NY, USA; as a postdoctoral scholar in electrical engineering and computer science at the University of California, Berkeley, Berkely, CA, USA; and as a postdoctoral scholar of computing and mathematical sciences at the California Institute of Technology (Caltech), Pasadena, CA, USA. He received the Ph.D. degree in control and dynamical systems from Caltech in 2016. He also holds the B.A.Sc. degree and M.A.Sc. degree in electrical engineering from the University of British Columbia, Vancouver, BC, Canada. His research interests broadly encompass the use of learning, optimization, and control in the design and analysis of autonomous systems. He was a recipient of the National Science Foundation CAREER Award (2021), a Google Research Scholar Award (2021), the 2021 IEEE Control Systems Society George S. Axelby Award, and the 2013 IEEE Conference on Decision and Control (CDC) Best Student Paper Award. He was also a coauthor of papers that won the 2022 CDC Best Student Paper Award and the 2017 American Control Conference Best Student Paper Award. He is a Member of IEEE.

George J. Pappas received the Ph.D. degree in electrical engineering and computer sciences from the University of California, Berkeley, Berkeley, CA, USA, in 1998. He is currently the Joseph Moore Professor in and the chair of the Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA. He also holds a secondary appointment with the Department of Computer and Information Sciences and the Department of Mechanical Engineering and Applied Mechanics. He is a member of the General Robotics, Automation, Sensing, and Perception Lab and the Penn Research in Embedded Computing and Integrated Systems Engineering Center. He was previously the deputy dean for research with the School of Engineering and Applied Science. His research interests include control theory and, in particular, hybrid systems, embedded systems, cyberphysical systems, and hierarchical and distributed control systems, with applications to unmanned aerial vehicles, distributed robotics, green buildings, and biomolecular networks. He was a recipient of various awards, such as the Antonio Ruberti Young Researcher Prize, the IEEE Control Systems Society George S. Axelby Award, the O. Hugo Schuck Best Paper Award, the George H. Heilmeier Award, the National Science Foundation Presidential Early Career Award for Scientists and Engineers, and numerous best student papers awards. He is a Fellow of IEEE.

REFERENCES

[1] D. Silver et al., "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016, doi: 10.1038/nature16961.

[2] S. Tu, R. Frostig, and M. Soltanolkotabi, "Learning from many trajectories," 2022, *arXiv:2203.17193*.

[3] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, vol. 47. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[4] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," 2018, *arXiv:1802.08334*.

[5] Y. Jedra and A. Proutiere, "Sample complexity lower bounds for linear system identification," in *Proc. IEEE 58th Conf. Decis. Control (CDC)*, Piscataway, NJ, USA: IEEE Press, 2019, pp. 2676–2681, doi: 10.1109/CDC40024.2019.9029303.

[6] L. Ljung, *System Identification: Theory for the User*. Upper Saddle River, NJ, USA: Prentice-Hall, 1999.

[7] T. L. Lai and C. Z. Wei, "Asymptotic properties of general autoregressive models and strong consistency of least-squares estimates of their parameters," *J. Multivariate Anal.*, vol. 13, no. 1, pp. 1–23, Mar. 1983, doi: 10.1016/0047-259X(83)90002-7.

[8] L. Ljung and B. Wahlberg, "Asymptotic properties of the least-squares method for estimating transfer functions and disturbance spectra," *Adv. Appl. Probability*, vol. 24, no. 2, pp. 412–440, 1992, doi: 10.2307/1427698.

[9] M. Deistler, K. Peternell, and W. Scherrer, "Consistency and relative efficiency of subspace methods," *Automatica*, vol. 31, no. 12, pp. 1865–1875, Dec. 1995, doi: 10.1016/0005-1098(95)00089-6.

[10] D. Bauer, M. Deistler, and W. Scherrer, "Consistency and asymptotic normality of some subspace algorithms for systems without observed inputs," *Automatica*, vol. 35, no. 7, pp. 1243–1254, Jul. 1999, doi: 10.1016/S0005-1098(99)00031-X.

[11] A. Chiuso and G. Picci, "The asymptotic variance of subspace estimates," *J. Econometrics*, vol. 118, nos. 1–2, pp. 257–291, Jan./Feb. 2004, doi: 10.1016/S0304-4076(03)00143-X.

[12] E.-W. Bai and S. S. Sastry, "Persistency of excitation, sufficient richness and parameter convergence in discrete time adaptive control," *Syst. Control Lett.*, vol. 6, no. 3, pp. 153–163, Aug. 1985, doi: 10.1016/0167-6911(85)90035-0.

[13] E. J. Hannan and M. Deistler, *The Statistical Theory of Linear Systems*. Philadelphia, PA, USA: SIAM, 2012.

[14] M. A. Dahleh, T. V. Theodosopoulos, and J. N. Tsitsiklis, "The sample complexity of worst-case identification of FIR linear systems," in *Proc. 32nd IEEE Conf. Decis. Control*, 1993, pp. 2082–2086, doi: 10.1109/CDC.1993.325566.

[15] K. Poolla and A. Tikku, "On the time complexity of worst-case system identification," *IEEE Trans. Autom. Control*, vol. 39, no. 5, pp. 944–950, May 1994, doi: 10.1109/9.284870.

[16] L. Guo and L. Ljung, "Performance analysis of general tracking algorithms," *IEEE Trans. Autom. Control*, vol. 40, no. 8, pp. 1388–1402, Aug. 1995, doi: 10.1109/9.402230.

[17] A. Goldenshluger, "Nonparametric estimation of transfer functions: Rates of convergence and adaptation," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 644–658, Mar. 1998, doi: 10.1109/18.661510.

[18] E. Weyer, R. C. Williamson, and I. M. Mareels, "Finite sample properties of linear model identification," *IEEE Trans. Autom. Control*, vol. 44, no. 7, pp. 1370–1383, Jul. 1999, doi: 10.1109/9.774109.

[19] M. C. Campi and E. Weyer, "Finite sample properties of system identification methods," *IEEE Trans. Autom. Control*, vol. 47, no. 8, pp. 1329–1334, Aug. 2002, doi: 10.1109/TAC.2002.800750.

[20] M. Vidyasagar and R. L. Karandikar, "A learning theory approach to system identification and stochastic adaptive control," *J. Process Control*, vol. 18, nos. 3–4, pp. 421–430, Mar. 2008, doi: 10.1016/j.jprocont.2007.10.009.

[21] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems," in *Proc. 24th Annu. Conf. Learn. Theory*, 2011, pp. 1–26.
[22] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *Found. Comput. Math.*, vol. 20, no. 4, pp. 633–679, Aug. 2020, doi: 10.1007/s10208-019-09426-y.
[23] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, vol. 48. Cambridge, U.K.: Cambridge Univ. Press, 2019.
[24] A. Tsiamis and G. J. Pappas, "Linear systems can be hard to learn," 2021, *arXiv:2104.01120*.
[25] S. Mendelson, "Learning without concentration," in *Proc. 27th Conf. Learn. Theory*, PMLR, 2014, pp. 25–39.
[26] N. Matni and S. Tu, "A tutorial on concentration bounds for system identification," in *Proc. IEEE 58th Conf. Decis. Control (CDC)*, Piscataway, NJ, USA: IEEE Press, 2019, pp. 3741–3749, doi: 10.1109/CDC40024.2019.9029621.
[27] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 2312–2320.
[28] Y. Jedra and A. Proutiere, "Finite-time identification of stable linear systems optimality of the least-squares estimator," in *Proc. 59th IEEE Conf. Decis. Control (CDC)*, Piscataway, NJ, USA: IEEE Press, 2020, pp. 996–1001, doi: 10.1109/CDC42340.2020.9304362.
[29] R. Ahlswede and A. Winter, "Strong converse for identification via quantum channels," *IEEE Trans. Inf. Theory*, vol. 48, no. 3, pp. 569–579, Mar. 2002, doi: 10.1109/18.985947.
[30] A. Carè, B. C. Csáji, M. C. Campi, and E. Weyer, "Finite-sample system identification: An overview and a new correlation method," *IEEE Contr. Syst. Lett.*, vol. 2, no. 1, pp. 61–66, Jan. 2018, doi: 10.1109/LCSYS.2017.2720969.
[31] Y. Jedra and A. Proutiere, "Finite-time identification of linear systems: Fundamental limits and optimal algorithms," *IEEE Trans. Autom. Control*, vol. 68, no. 5, pp. 2805–2820, May 2022, doi: 10.1109/TAC.2022.3221705.
[32] A. Wagenmaker and K. Jamieson, "Active learning for identification of linear dynamical systems," in *Proc. 33rd Conf. Learn. Theory*, PMLR, 2020, pp. 3487–3582.
[33] T. Sarkar and A. Rakhlin, "Near optimal finite time identification of arbitrary linear dynamical systems," in *Proc. 36th Int. Conf. Mach. Learn.*, PMLR, 2019, pp. 5610–5618.
[34] S. Fattahi, N. Matni, and S. Sojoudi, "Learning sparse dynamical systems from a single sample trajectory," 2019, *arXiv:1904.09396*.
[35] M. K. Shirani Faradonbeh, A. Tewari, and G. Michailidis, "Finite time identification in unstable linear systems," *Automatica*, vol. 96, pp. 342–353, Oct. 2018, doi: 10.1016/j.automatica.2018.07.008.
[36] M. Simchowitz and D. J. Foster, "Naive exploration is optimal for online LQR," 2020, *arXiv:2001.09576*.
[37] A. Tsiamis and G. J. Pappas, "Finite sample analysis of stochastic system identification," in *Proc. IEEE 58th Conf. Decis. Control (CDC)*, 2019, pp. 3648–3654, doi: 10.1109/CDC40024.2019.9029499.
[38] J. Mourtada, "Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices," *Ann. Statist.*, vol. 50, no. 4, pp. 2157–2178, Aug. 2022, doi: 10.1214/22-AOS2181.
[39] R. I. Oliveira, "The lower tail of random quadratic forms with applications to ordinary least squares," *Probability Theory Related Fields*, vol. 166, nos. 3–4, pp. 1175–1194, Dec. 2016, doi: 10.1007/s00440-016-0738-9.
[40] S. Mendelson, "Learning without concentration for general loss functions," *Probability Theory Related Fields*, vol. 171, nos. 1–2, pp. 459–502, Jun. 2018, doi: 10.1007/s00440-017-0784-y.
[41] P. Van Overschee and B. De Moor, *Subspace Identification for Linear Systems: Theory–Implementation–Applications*. New York, NY, USA: Springer Science and Business Media, 2012.
[42] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. New York, NY, USA: Dover, 2005.
[43] S. Joe Qin, "An overview of subspace identification," *Comput. Chem. Eng.*, vol. 30, nos. 10–12, pp. 1502–1513, 2006, doi: 10.1016/j.compchemeng.2006.05.045.
[44] M. Hardt, T. Ma, and B. Recht, "Gradient descent learns linear dynamical systems," *J. Mach. Learn. Res.*, vol. 19, no. 29, pp. 1–44, 2018.
[45] M. Kozdoba, J. Marecek, T. Tchrakian, and S. Mannor, "On-line learning of linear dynamical systems: Exponential forgetting in Kalman filters," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, pp. 4098–4105, 2019, doi: 10.1609/aaai.v33i01.33014098.
[46] H. Lee, "Improved rates for prediction and identification of partially observed linear dynamical systems," in *Proc. Int. Conf. Algorithmic Learn. Theory*, PMLR, 2022, pp. 668–698.
[47] T. Sarkar, A. Rakhlin, and M. A. Dahleh, "Finite time LTI system identification," *J. Mach. Learn. Res.*, vol. 22, no. 1, pp. 1186–1246, 2021.
[48] S. Oymak and N. Ozay, "Revisiting Ho–Kalman-based system identification: Robustness and finite-sample analysis," *IEEE Trans. Autom. Control*, vol. 67, no. 4, pp. 1914–1928, Apr. 2022, doi: 10.1109/TAC.2021.3083651.
[49] P.-Å. Wedin, "Perturbation bounds in connection with singular value decomposition," *BIT Numer. Math.*, vol. 12, no. 1, pp. 99–111, Mar. 1972, doi: 10.1007/BF01932678.
[50] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 964–973.
[51] B. Lee and A. Lamperski, "Non-asymptotic closed-loop system identification using autoregressive processes and hankel model reduction," in *Proc. IEEE 59th Conf. Decis. Control (CDC)*, 2020, pp. 3419–3424, doi: 10.1109/CDC42340.2020.9304468.
[52] M. Verhaegen and V. Verdult, *Filtering and System Identification: A Least Squares Approach*. Cambridge, U.K.: Cambridge Univ. Press, 2007.
[53] D. Bauer, "Asymptotic properties of subspace estimators," *Automatica*, vol. 41, no. 3, pp. 359–376, Mar. 2005, doi: 10.1016/j.automatica.2004.11.012.
[54] D. Bauer, M. Deistler, and W. Scherrer, "On the impact of weighting matrices in subspace algorithms," *IFAC Proc. Volumes*, vol. 33, no. 15, pp. 97–102, Jun. 2000, doi: 10.1016/S1474-6670(17)39733-1.
[55] S. Fattahi, "Learning partially observed linear dynamical systems from logarithmic number of samples," in *Proc. 33rd Learn. Dyn. Control*, PMLR, 2021, pp. 60–72.
[56] Y. Sun, S. Oymak, and M. Fazel, "System identification via nuclear norm regularization," 2022, *arXiv:2203.16673*.
[57] B. Djehiche and O. Mazhar, "Efficient learning of hidden state LTI state space models of unknown order," 2022, *arXiv:2202.01625*.
[58] Y. Zheng and N. Li, "Non-asymptotic identification of linear dynamical systems using multiple trajectories," *IEEE Contr. Syst. Lett.*, vol. 5, no. 5, pp. 1693–1698, Nov. 2021, doi: 10.1109/LCSYS.2020.3042924.
[59] M. Simchowitz, R. Boczar, and B. Recht, "Learning linear dynamical systems with semi-parametric least squares," in *Proc. Conf. Learn. Theory*, PMLR, 2019, pp. 2714–2802.
[60] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, "Adaptive control and regret minimization in linear quadratic Gaussian (LQG) setting," in *Proc. Amer. Control Conf. (ACC)*, Piscataway, NJ, USA: IEEE Press, 2021, pp. 2517–2522, doi: 10.23919/ACC50511.2021.9483309.
[61] R. D. Gill and B. Y. Levit, "Applications of the van trees inequality: A Bayesian Cramér-Rao bound," *Bernoulli*, vol. 1, no. 1/2, pp. 59–79, Mar./Jun. 1995, doi: 10.2307/3318681.
[62] T. Söderström, "On computing the Cramer-Rao bound and covariance matrices for PEM estimates in linear state space models," *IFAC Proc. Volumes*, vol. 39, no. 1, pp. 600–605, 2006, doi: 10.3182/20060329-3-AU-2901.00092.
[63] D. Bauer, "Comparing the CCA subspace method to pseudo maximum likelihood methods in the case of no exogenous inputs," *J. Time Series Anal.*, vol. 26, no. 5, pp. 631–668, Sep. 2005, doi: 10.1111/j.1467-9892.2005.00441.x.
[64] D. Bauer and M. Wagner, "Estimating cointegrated systems using subspace algorithms," *J. Econometrics*, vol. 111, no. 1, pp. 47–84, Nov. 2002, doi: 10.1016/S0304-4076(02)00119-7.
[65] O. Vinyals et al., "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, Nov. 2019, doi: 10.1038/s41586-019-1724-z.
[66] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.
[67] J. Garc Ia and F. Fernández, "A comprehensive survey on safe reinforcement learning," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 1437–1480, 2015.
[68] H. Mania, S. Tu, and B. Recht, "Certainty equivalence is efficient for linear quadratic control," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 10,154–10,164.
[69] A. J. Wagenmaker, M. Simchowitz, and K. Jamieson, "Task-optimal exploration in linear dynamical systems," in *Proc. 38th Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 10,641–10,652.
[70] S. Tu, R. Boczar, A. Packard, and B. Recht, "Non-asymptotic analysis of robust control from coarse-grained identification," 2017, *arXiv:1707.04791*.
[71] J. Anderson, J. C. Doyle, S. H. Low, and N. Matni, "System level synthesis," *Annu. Rev. Control*, vol. 47, pp. 364–393, 2019, doi: 10.1016/j.arcontrol.2019.03.006.
[72] S. Dean, S. Tu, N. Matni, and B. Recht, "Safely learning to control the constrained linear quadratic regulator," in *Proc. Amer. Control Conf.*

(ACC), Piscataway, NJ, USA: IEEE Press, 2019, pp. 5582–5588, doi: 10.23919/ACC.2019.8814865.
[73] R. Boczar, N. Matni, and B. Recht, "Finite-data performance guarantees for the output-feedback control of an unknown system," in *Proc. IEEE Conf. Decis. Control (CDC)*, Piscataway, NJ, USA: IEEE Press, 2018, pp. 2994–2999, doi: 10.1109/CDC.2018.8618658.
[74] L. Furieri, B. Guo, A. Martin, and G. Ferrari-Trecate, "Near-optimal design of safe output-feedback controllers from noisy data," *IEEE Trans. Autom. Control*, vol. 68, no. 5, pp. 2699–2714, May 2023, doi: 10.1109/TAC.2022.3180692.
[75] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *Proc. 35th Int. Conf. Mach. Learn.*, PMLR, 2018, pp. 1467–1476.
[76] B. Hambly, R. Xu, and H. Yang, "Policy gradient methods for the noisy linear quadratic regulator over a finite horizon," *SIAM J. Control Optim.*, vol. 59, no. 5, pp. 3359–3391, 2021, doi: 10.1137/20M1382386.
[77] J. Perdomo, J. Umenberger, and M. Simchowitz, "Stabilizing dynamical systems via policy gradient methods," in *Proc. 34th Adv. Neural Inf. Process. Syst.*, 2021, pp. 29,274–29,286.
[78] B. Hu, K. Zhang, N. Li, M. Mesbahi, M. Fazel, and T. Başar, "Towards a theoretical foundation of policy optimization for learning control policies," 2022, *arXiv:2210.04810*.
[79] S. Tu and B. Recht, "The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint," in *Proc. 32nd Conf. Learn. Theory*, PMLR, 2019, pp. 3036–3083.
[80] I. Ziemann, A. Tsiamis, H. Sandberg, and N. Matni, "How are policy gradient methods affected by the limits of control?" in *Proc. IEEE 61st Conf. Decis. Control (CDC)*, Cancun, Mexico, 2022, pp. 5992–5999, doi: 10.1109/CDC51059.2022.9992612.
[81] A. A. Feldbaum, "Dual control theory. I," *Avtomat. Telemekh.*, vol. 21, no. 9, pp. 1240–1249, 1960.
[82] A. A. Feldbaum, "Dual control theory. II," *Avtomat. Telemekh.*, vol. 21, no. 11, pp. 1453–1464, 1960.
[83] Y. Jedra and A. Proutiere, "Minimal expected regret in linear quadratic control," in *Proc. Int. Conf. Artif. Intell. Statist.*, PMLR, 2022, pp. 10,234–10,321.
[84] I. Ziemann and H. Sandberg, "Regret lower bounds for learning linear quadratic gaussian systems," 2022, *arXiv:2201.01680*.
[85] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "Regret bounds for robust adaptive control of the linear quadratic regulator," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4188–4197.
[86] M. K. Shirani Faradonbeh, A. Tewari, and G. Michailidis, "Input perturbations for adaptive control and learning," *Automatica*, vol. 117, Jul. 2020, Art. no. 108950, doi: 10.1016/j.automatica.2020.108950.
[87] A. Cohen, T. Koren, and Y. Mansour, "Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 1300–1309.
[88] A. Cassel, A. Cohen, and T. Koren, "Logarithmic regret for learning linear quadratic regulators efficiently," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1328–1337.
[89] M. Simchowitz, K. Singh, and E. Hazan, "Improper learning for non-stochastic control," in *Proc. 33rd Conf. Learn. Theory*, PMLR, 2020, pp. 3320–3436.
[90] A. Tsiamis, I. Ziemann, M. Morari, N. Matni, and G. J. Pappas, "Learning to control linear systems can be hard," in *Proc. 35th Conf. Learn. Theory*, 2022, pp. 3820–3857.
[91] H. A. Simon, "Dynamic programming under uncertainty with a quadratic criterion function," *Econometrica, J. Econometric Soc.*, vol. 24, no. 1, pp. 74–81, Jan. 1956, doi: 10.2307/1905261.
[92] K. J. Åström and B. Wittenmark, "On self tuning regulators," *Automatica*, vol. 9, no. 2, pp. 185–199, Mar. 1973, doi: 10.1016/0005-1098(73)90073-3.
[93] T. L. Lai, "Asymptotically efficient adaptive control in stochastic regression models," *Adv. Appl. Math.*, vol. 7, no. 1, pp. 23–45, 1986, doi: 10.1016/0196-8858(86)90004-7.
[94] T. Söderström, *Discrete-Time Stochastic Systems: Estimation and Control*. London, U.K.: Springer Science and Business Media, 2002.
[95] W. Lin, P. R. Kumar, and T. I. Seidman, "Will the self-tuning approach work for general cost criteria?" *Syst. Control Lett.*, vol. 6, no. 2, pp. 77–85, Jul. 1985, doi: 10.1016/0167-6911(85)90001-5.
[96] M. Gevers and L. Ljung, "Optimal experiment designs with respect to the intended model application," *Automatica*, vol. 22, no. 5, pp. 543–554, Sep. 1986, doi: 10.1016/0005-1098(86)90064-6.
[97] J. Willem Polderman, "On the necessity of identifying the true parameter in adaptive LQ control," *Syst. Control Lett.*, vol. 8, no. 2, pp. 87–91, Dec. 1986, doi: 10.1016/0167-6911(86)90065-4.
[98] K. Colin, M. Ferizbegovic, and H. Hjalmarsson, "Regret minimization for linear quadratic adaptive controllers using fisher feedback exploration," *IEEE Contr. Syst. Lett.*, vol. 6, pp. 2870–2875, Jun. 2022, doi: 10.1109/LCSYS.2022.3179668.
[99] D. Youla, H. Jabr, and J. Bongiorno, "Modern Wiener-Hopf design of optimal controllers–Part II: The multivariable case," *IEEE Trans. Autom. Control*, vol. 21, no. 3, pp. 319–338, Jun. 1976, doi: 10.1109/TAC.1976.1101223.
[100] G. Zames, "Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses," *IEEE Trans. Autom. Control*, vol. 26, no. 2, pp. 301–320, Apr. 1981, doi: 10.1109/TAC.1981.1102603.
[101] O. Anava, E. Hazan, and S. Mannor, "Online learning for adversaries with memory: Price of past mistakes," in *Proc. 28th Adv. Neural Inf. Process. Syst.*, 2015.
[102] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, "Logarithmic regret bound in partially observable linear dynamical systems," in *Proc. 33rd Adv. Neural Inf. Process. Syst.*, 2020, pp. 20,876–20,888.
[103] A. Tsiamis and G. Pappas, "Online learning of the Kalman filter with logarithmic regret," *IEEE Trans. Autom. Control*, vol. 68, no. 5, pp. 2774–2789, May 2023, doi: 10.1109/TAC.2022.3207670.
[104] U. Ghai, H. Lee, K. Singh, C. Zhang, and Y. Zhang, "No-regret prediction in marginally stable systems," in *Proc. 33rd Conf. Learn. Theory*, PMLR, 2020, pp. 1714–1757.
[105] P. Rashidinejad, J. Jiao, and S. Russell, "Slip: Learning to predict in unknown dynamical systems with long-term memory," 2020, *arXiv:2010.05899*.
[106] M. Gevers, "Identification for control: From the early achievements to the revival of experiment design," *Eur. J. Control*, vol. 11, nos. 4–5, pp. 335–352, 2005, doi: 10.3166/ejc.11.335-352.
[107] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Statist.*, vol. 37, no. 4, pp. 1705–1732, Aug. 2009, doi: 10.1214/08-AOS620.
[108] S. Negahban, B. Yu, M. J. Wainwright, and P. Ravikumar, "A unified framework for high-dimensional analysis of *m*-estimators with decomposable regularizers," in *Proc. 22nd Adv. Neural Inf. Process. Syst.*, 2009.
[109] I. Ziemann and S. Tu, "Learning with little mixing," 2022, *arXiv:2206.08269*.
[110] G. Lecué and S. Mendelson, "Regularization and the small-ball method I: Sparse recovery," *Ann. Statist.*, vol. 46, no. 2, pp. 611–641, 2018, doi: 10.1214/17-AOS1562.
[111] Y. Sattar and S. Oymak, "Non-asymptotic and accurate learning of nonlinear dynamical systems," *J. Mach. Learn. Res.*, vol. 23, no. 1, pp. 6248–6296, Jan. 2022.
[112] H. Mania, M. I. Jordan, and B. Recht, "Active learning for nonlinear system identification with guarantees," *J. Mach. Learn. Res.*, vol. 23, no. 1, pp. 1433–1462, Jan. 2022.
[113] D. Foster, T. Sarkar, and A. Rakhlin, "Learning nonlinear dynamical systems from a single trajectory," in *Proc. Learn. Dyn. Control*, PMLR, 2020, pp. 851–861.
[114] Y. Sattar, Z. Du, D. A. Tarzanagh, L. Balzano, N. Ozay, and S. Oymak, "Identification and adaptive control of Markov jump systems: Sample complexity and regret bounds," 2021, *arXiv:2111.07018*.
[115] S. Kowshik, D. Nagaraj, P. Jain, and P. Netrapalli, "Near-optimal offline and streaming algorithms for learning non-linear dynamical systems," in *Proc. 34th Adv. Neural Inf. Process. Syst.*, 2021, pp. 8518–8531.
[116] I. M. Ziemann, H. Sandberg, and N. Matni, "Single trajectory nonparametric learning of nonlinear dynamics," in *Proc. 35th Conf. Learn. Theory*, PMLR, 2022, pp. 3333–3364.
[117] S. Tu, A. Robey, T. Zhang, and N. Matni, "On the sample complexity of stability constrained imitation learning," in *Proc. 4th Learn. Dyn. Control Conf.*, PMLR, 2022, pp. 180–191.
[118] D. Pfrommer, T. T. Zhang, S. Tu, and N. Matni, "TaSIL: Taylor series imitation learning," 2022, *arXiv:2205.14812*.
[119] H. Tsukamoto, S.-J. Chung, and J.-J. E. Slotine, "Contraction theory for nonlinear stability analysis and learning-based control: A tutorial overview," *Annu. Rev. Control*, vol. 52, pp. 135–169, 2021, doi: 10.1016/j.arcontrol.2021.10.001.
[120] D. Nagaraj, X. Wu, G. Bresler, P. Jain, and P. Netrapalli, "Least squares regression with Markovian data: Fundamental limits and algorithms," in *Proc. 33rd Adv. Neural Inf. Process. Syst.*, 2020, pp. 16,666–16,676.