2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

The Fundamental Limitations of Learning Linear-Quadratic Regulators

Bruce D. Lee[1], Ingvar Ziemann[1], Anastasios Tsiamis[2], Henrik Sandberg[3], and Nikolai Matni[1]

[1]Department of Electrical and Systems Engineering, University of Pennsylvania
[2]Automatic Control Laboratory, ETH Zurich
[3]Division of Decision and Control Systems, KTH Royal Institute of Technology

***Abstract*— We present a local minimax lower bound on the excess cost of designing a linear-quadratic controller from offline data. The bound is valid for any offline exploration policy that consists of a stabilizing controller and an energy bounded exploratory input. The derivation leverages a relaxation of the minimax estimation problem to Bayesian estimation, and an application of van Trees inequality. We show that the bound aligns with system-theoretic intuition. In particular, we demonstrate that the lower bound increases when the optimal control objective value increases. We also show that the lower bound increases when the system is poorly excitable, as characterized by the spectrum of the controllability gramian of the system mapping the noise to the state and the $\mathcal{H}_\infty$ norm of the system mapping the input to the state. We further show that for some classes of systems, the lower bound may be exponential in the state dimension, demonstrating exponential sample complexity for learning the linear-quadratic regulator.**

I. Introduction

Reinforcement Learning (RL) has demonstrated success in a variety of domains, including robotics [1] and games [2]. However, it is known to be very data intensive, making it challenging to apply to complex control tasks. This has motivated efforts by both the machine learning and control communities to understand the statistical hardness of RL in analytically tractable settings, such as the tabular setting [3] and the linear-quadratic control setting [4]. Such studies provide insights into the fundamental limitations of RL, and the efficiency of particular algorithms.

There are two common problems of interest for understanding the statistical hardness of RL from the perspective of learning a linear-quadratic regulator (LQR): online LQR, and offline LQR. Online LQR models an interactive problem in which the learning agent attempts to minimize a regret-based objective, while simultaneously learning the dynamics [5]. Offline LQR models a two-step pipeline, where data from the system is collected, and then used to design a controller [6]. Guarantees in the online setting are in the form of regret bounds, whereas the offline setting focuses on Probably Approximately Correct (PAC) guarantees. The high data requirements of RL often render offline approaches the only feasible option for physical systems [7]. Despite this fact, recent years have seen greater efforts to provide lower bounds for the online LQR problem [8], [9]. Meanwhile, lower bounds that illustrate the hardness of learning in terms of interpretable system-theoretic parameters are conspicuously absent for the offline setting. Motivated by this fact, we derive lower bounds for designing a linear-quadratic controller from offline data.

Notation: The Euclidean norm of a vector x is denoted by $\|x\|$. The quadratic norm of a vector x with respect to a matrix P is denoted $\|x\|_P = \sqrt{x^\top P x}$. For a matrix A, the spectral norm is denoted $\|A\|$ and the Frobenius norm is denoted $\|A\|_F$. The spectral radius of a square matrix A is denoted $\rho(A)$. A symmetric, positive semidefinite matrix $A = A^\top$ is denoted $A \succeq 0$, and a symmetric, positive definite matrix is denoted $A \succ 0$. Similarly, $A \succeq B$ denotes that $A - B$ is positive semidefinite. The eigenvalues of a symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$ are denoted $\lambda_1(A), \dots, \lambda_n(A)$, and are sorted in non-ascending order. We also denote $\lambda_1(A) = \lambda_{\max}(A)$, and $\lambda_n(A) = \lambda_{\min}(A)$. For a matrix A, the vectorization operator $\mathsf{vec}\, A$ maps A to a column vector by stacking the columns of A. The kronecker product of A with B is denoted $A \otimes B$. Expectation and probability with respect to all the randomness of the underlying probability space are denoted $\mathbf{E}$ and $\mathbf{P}$, respectively. Conditional expectation and probability given the random variable X are denoted by $\mathbf{E}[\cdot|X]$ and $\mathbf{P}[\cdot|X]$. For an event $\mathcal{G}$, $\mathbf{1}_\mathcal{G}$ denotes the indicator function for $\mathcal{G}$. For a matrix $A \in \mathbb{R}^{d_\mathsf{x}\times d_\mathsf{x}}$ and a symmetric matrix $Q \in \mathbb{R}^{d_\mathsf{x}\times d_\mathsf{x}}$, we denote the solution P to the discrete Lyapunov equation, $A^\top P A - P + Q = 0$, by $\mathtt{dlyap}(A, Q)$. If we also have $B \in \mathbb{R}^{d_\mathsf{u}\times d_\mathsf{u}}$ and $R \in \mathbb{R}^{d_\mathsf{u}\times d_\mathsf{u}}$, $R \succ 0$, we denote the solution P to the discrete algebraic Riccati equation $Q + A^\top P A - A^\top P B(B^\top P B + R)^{-1} B^\top P A = 0$ by $\mathtt{DARE}(A, B, Q, R)$. We use the indexing shorthand $[K] := \{1, \dots, K\}$.

a) Problem Formulation: Let $\theta \in \mathbb{R}^{d_\Theta}$ be an unknown parameter. We study the fundamental limitations to learning to control the following parametric system model:

$$X_{t+1} = A(\theta)X_t + B(\theta)U_t + W_t,\ X_0 = 0, \quad t = 0, 1, \dots. \tag{1}$$

The noise process W_t is assumed to be iid mean zero Gaussian with fixed covariance matrices $\Sigma_W \succ 0$. The matrices $A(\theta) \in \mathbb{R}^{d_\mathsf{x}\times d_\mathsf{x}}$ and $B(\theta) \in \mathbb{R}^{d_\mathsf{x}\times d_\mathsf{u}}$ are known continuously differentiable functions of the unknown parameter. We also assume that these functions are L_{dyn}-Lipschitz in the spectral norm. The system $(A(\theta), B(\theta))$ is assumed to be stabilizable.

We assume that the learner is given access to $N \in \mathbb{N}$ experiments $(X_{0,n}, \dots, X_{T-1,n}), n \in [N]$ from (1) of length $T \in \mathbb{N}$. The input signal during these experiments is

$$U_{t,n} = F X_{t,n} + \tilde{U}_{t,n}, \tag{2}$$

where F renders the system stable[1], i.e. $\rho(A(\theta) + B(\theta)F) < 1$. Meanwhile, $\tilde{U}_{t,n}$ is an exploration component with energy budget $\sigma_{\tilde{u}}^2 NT$,[2] where $\sigma_{\tilde{u}} \in \mathbb{R}_+$. More precisely, $\tilde{U}_{t,n}$ may be selected as a function of past observations $(X_{0,n}, \ldots, X_{t,n})$, past trajectories $(X_{0,m}, \ldots, X_{T-1,m}), m < n$ and possible auxiliary randomization, while being constrained to an energy budget

$$\frac{1}{NT} \sum_{n=1}^{N} \sum_{t=0}^{T-1} \mathbf{E}_\theta \, \tilde{U}_{t,n}^\top \tilde{U}_{t,n} \leq \sigma_{\tilde{u}}^2. \tag{3}$$

This formulation allows both open- and closed-loop experiments, but normalizes the average exploratory input energy to $\sigma_{\tilde{u}}^2$. The subscript θ on the expectation denotes that the system is rolled out with parameter θ. For a fixed parameter θ, we denote the data collected from these experiments by the random variable $\mathcal{Z} := \{\{(X_{t,n}, U_{t,n}\}_{t=0}^{T-1}\}_{n=1}^{N}$.

The learner deploys a policy π which is a measurable function of the N offline experiments and the current state. In particular, the learner maps the offline data and the current state to the control input, $U_t = \pi(X_t; \mathcal{Z})$. This is the case if the learner outputs a non-adaptive state feedback controller designed with the offline data. The goal of the learner is to minimize the cost defined by:

$$V_T^\pi(\theta) := \frac{1}{T} \mathbf{E}_\theta^\pi \left[\sum_{t=0}^{T-1} \Big(\|X_t\|_Q^2 + \|U_t\|_R^2 \Big) + \|X_T\|_{Q_T(A(\theta),B(\theta))}^2 \right].$$

The expectation is over both the offline experiments, and a new evaluation rollout. Single subscripts on the states and actions, X_t and U_t, refer to the evaluation rollout at time t. The superscript on the expectation denotes that the inputs applied in the evaluation rollout follow the policy $U_t = \pi(X_t; \mathcal{Z})$. Note that due to the dependence of the terminal cost $Q_T(A(\theta), B(\theta))$ on the unknown parameter θ, the learner does not explicitly know the cost function it is minimizing. This is not an issue: it simply means that the learner must infer the objective function from the data.

The following assumption guarantees the existence of a static state feedback controller that minimizes $V_T^\pi(\theta)$.

Assumption 1.1: Assume $(A(\theta), B(\theta))$ is stabilizable, $(A(\theta), Q^{1/2})$ is detectable, $R \succ 0$ and $Q_T(A(\theta), B(\theta)) = P(\theta)$, where $P(\theta) = \texttt{DARE}(A(\theta), B(\theta), Q, R)$.

Under this assumption, the optimal policy for the known system is $U_t = K(\theta) X_t$, where $K(\theta)$ is the LQR controller:

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

In light of this, we focus on the case in which the search space of the learner is the class of linear time-invariant state feedback policies where the gain is a measurable function of the past N experiments. This set is denoted Π_{lin}.

The stochastic LQR cost $V_T^\pi(\theta)$ may be represented in terms of the gap between the control actions taken by the policy π and the optimal policy, as shown below.

[1]Access to a stabilizing controller is often assumed for unstable sysID [10]. Open-loop unstable identification leads to poor conditioning.

[2]The choice to place a budget on the exploratory input $\tilde{U}_{t,n}$ rather than the total input $U_{t,n}$ is for ease of exposition. The energy of the exploratory input is bounded by the total budget, which is sufficient for our bounds.

Lemma 1.1 (Lemma 11.2 of [11]): We have that

$$V_T^\pi(\theta) = \mathrm{tr}(P(\theta)\Sigma_W) + \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{E}_\theta^\pi \, \|U_t - K(\theta) X_t\|_{\Psi(\theta)}^2 ,$$

where $\Psi(\theta) := B^\top(\theta) P(\theta) B(\theta) + R$.

Using the above lemma, the objective of the learner may be restated from minimizing $V_T^\pi(\theta)$ to minimizing the excess cost:

$$\mathsf{EC}_T^\pi(\theta) := V_T^\pi(\theta) - \inf_{\hat{\pi}} V_T^{\hat{\pi}}(\theta) = \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{E}_\theta^\pi \|U_t - K(\theta) X_t\|_{\Psi(\theta)}^2. \tag{4}$$

The second equality follows from the representation of the stochastic LQR cost in Lemma 1.1 by cancelling the constant terms. Note that the infimum in the second term is given access to the true parameter value θ, and will therefore be attained by the optimal LQR controller. In particular, it does not rely upon the offline experimental data. We denote this optimal policy by $\pi_\theta(X_t; \mathcal{Z}) = K(\theta) X_t$.

Our objective is to lower bound the excess cost for any learning agent in the class Π_{lin}. To this end, we introduce the ε-local minimax excess cost:

$$\mathcal{EC}_T^{\mathsf{lin}}(\theta, \varepsilon) := \inf_{\pi \in \Pi_{\mathsf{lin}}} \sup_{\|\theta' - \theta\| \leq \varepsilon} \mathsf{EC}_T^\pi(\theta'). \tag{5}$$

To motivate this choice, first note that if we were instead interested in an excess cost bound for only a single value of θ that holds for all estimators, the optimal policy would trivially be the LQR, $\pi(X_t, \mathcal{Z}) = K(\theta) X_t$. This policy would result in a lower bound of zero. By instead requiring that the learner perform well on all parameter instances in a nearby neighborhood, we remove the possibility of the trivial solution, and can achieve meaningful lower bounds. The emphasis of the nearby neighborhood in (5) is essential. As the neighborhood defined by the ball of radius ε, $\mathcal{B}(\theta, \varepsilon) = \{\theta' | \, \|\theta' - \theta\| \leq \varepsilon\}$, becomes sufficiently small, we can provide instance-specific lower bounds. Therefore, the ε-local minimax excess cost is a much stronger notion than the standard *global* minimax excess cost, $\inf_{\pi \in \Pi_{\mathsf{lin}}} \sup_{\theta'} \mathsf{EC}_T^\pi(\theta')$, as it does not require our estimator to perform well on *all possible* parameter values but only those in a small (possibly infinitesimal) neighborhood. Indeed, the global minimax excess cost for learning the optimal controller of the class of unknown stable scalar systems is infinite, as shown in Corollary 2.2, and illustrated in Figure 1.

Our focus in obtaining the lower bound on $\mathcal{EC}_T^{\mathsf{lin}}(\theta, \varepsilon)$ is to gain an understanding of what system-theoretic quantities render the learning problem statistically challenging. To this end, our lower bound depends on familiar system-theoretic quantities, such as $P(\theta)$. The covariance of the state under the optimal LQR controller also appears in our analysis. Under the optimal LQR controller, the covariance of the state converges to the stationary covariance as $T \to \infty$:

$$\begin{aligned} \Sigma_X(\theta) &= \lim_{t \to \infty} \mathbf{E}_\theta^{\pi_\theta} \left[X_t X_t^\top \right] \\ &= \texttt{dlyap}((A(\theta) + B(\theta) K(\theta))^\top, \Sigma_W). \end{aligned}$$

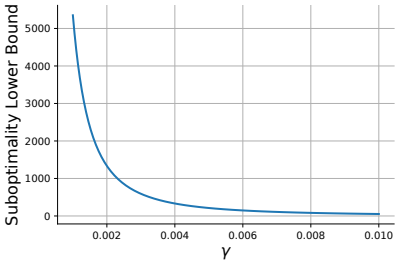

Fig. 1. Consider the scalar system $X_{t+1} = aX_t + bU_t + W_t$. We plot a lower bound arising from Corollary 2.1 letting $V = \begin{bmatrix} 0 & 1 \end{bmatrix}^\top$ for the system $a = 1 - \gamma$, $b = \gamma$ as γ ranges from 10^{-3} to 10^{-2} with $F = 0$, $\sigma_{\tilde{u}}^2 = 1$, $\Sigma_W = 1$. As $\gamma \to 0$ the optimally regulated system approaches marginal stability and controllability is lost. The problem of learning a controller therefore becomes challenging as $\gamma \to 0$, which is reflected by our excess cost lower bound; it approaches ∞. This illustrates the observation that systems which are difficult to control are also difficult to learn to control. This plot illustrates the result Corollary 2.2 in demonstrating that the global minimax excess cost is uninformative.

A. Contributions

Our first contribution is the following theorem. For the formal statements, see Theorem 2.2 and Corollary 2.1.

Theorem 1.1 (Main result, Informal): The ε-local minimax excess cost is lower bounded as

$$\text{excess cost} \geq \frac{\text{system-theoretic condition number}}{\text{\# data points} \times \text{signal-to-noise ratio}}.$$

The above theorem is similar to the lower bound in [12]. The primary difference is that while [12] focus on obtaining matching upper and lower bounds, we focus on expressing the system-theoretic condition number in terms of familiar quantities such as the covariance of the state under the optimal controller, and the solution to the Riccati equation. The signal-to-noise ratio depends on how easily the system is excited via both the exploratory input and the noise. This signal-to-noise ratio may be quantified in terms of the controllability gramian of the system, as well as the exploratory input budget.

We also study several consequences of the above result by restricting attention to the setting where all system parameters are unknown, i.e. $\text{vec}\begin{bmatrix} A(\theta) & B(\theta) \end{bmatrix} = \theta$. In this setting, Theorem 1.1 may be reduced to $\mathcal{EC}_T^{\text{lin}}(\theta, \varepsilon) \geq \frac{c(\theta,\varepsilon)}{NT}$, where $c(\theta, \varepsilon)$ is easily interpretable. In particular, we may reach the following conclusions:

- For classes of system where the operator norm of system-theoretic matrices such as the controllability gramian and the solution to the Riccati equation are constant with respect to dimension, we may take $c(\theta, \varepsilon) \propto d_{\mathsf{U}} d_{\mathsf{X}}$. We demonstrate that this is the optimal dependence on system dimension for this class of systems.
- There exist classes of systems for which we may take $c(\theta, \varepsilon) \propto \exp(d_{\mathsf{X}})$. This demonstrates that the excess cost of a learned LQR controller may grow exponentially in the dimension.
- We may take $c(\theta, \varepsilon)$ to grow with iterpretable system-theoretic quantities, such as the eigenvalues of both the solution to the Riccati equation, $P(\theta)$, and the state covariance under the optimal controller, $\Sigma_X(\theta)$. This suggests that the problem of learning to control a system with a small gap from the optimal controller is data intensive when controlling the underlying system is hard.

B. Related Work

a) System Identification: System identification is often a first step in designing a controller from experimental data, and has a longstanding history. The text [10] covers classic asymptotic results. Control oriented identification was studied in [13], [14]. Recently, there has been interest in finite sample analysis for fully-observed linear systems [6], [15]–[17], and partially-observed linear systems [15], [18]–[22]. Lower bounds for the sample complexity of system identification are presented in [23], [24]. For a more extensive discussion of prior work, we refer to the survey by [9].

b) Learning Controllers Offline: Learning a controller from offline data is a familiar paradigm for control theorists and practitioners. It typically consists of system identification, followed by robust [25] or certainty-equivalent [26] control design, see Figure 2. Recent work provides finite sample guarantees for such methods [6], [27]. Upper and lower bounds on the sample complexity of stabilization from offline data are presented in [28]. The RL community has a similar paradigm, known as offline RL [7]. Policy gradient approaches are a model-free algorithm suitable for offline RL, and are analyzed in [29]. Lower bounds on the variance of the gradient estimates in policy gradient approaches are supplied in [30]. Most similar to our work, [12] provide matching upper and lower bounds for offline linear control with the objective of designing optimal experiments. The bounds in [12] express the dependence of the excess cost on the Hessian of the control objective with respect to the unknown parameters, but do not illustrate how this quantity relates to familiar system-theoretic quantities. We instead focus on the LQR setting to understand the dependence of the excess cost on interpretable system-theoretic quantities.

c) Online LQR: The problem of learning the optimal LQR controller online has a rich history beginning with [31]. Regret minimization was introduced in [32], [33]. The study of regret in online LQR was re-initiated by [5], inspired by works in the RL community. Many works followed to propose algorithms which were computationally tractable [27], [34]–[39]. Lower bounds on the regret of online LQR are presented in [4], [8], [40]. The results in this paper follow a similar proof to [8]. The primary difference is that since our controller is designed via offline data, we may not make use of the exploration-exploitation tradeoff to upper bound the information available to the learner, as is done in [8].

II. Excess Cost Lower Bound

We now proceed to establish our lower bound. Missing proofs and further details may be found in [41]. As we are interested in the worst-case excess cost from any element of $\mathcal{B}(\theta, \varepsilon)$, we make the additional assumption that F stabilizes $(A(\theta'), B(\theta'))$ for all $\theta' \in \mathcal{B}(\theta, \varepsilon)$.[3] This also ensures that the optimal LQR controller exists for all $\theta' \in B(\theta, \varepsilon)$.

[3]We ultimately study the limit as ε becomes small. Therefore, this is not significantly stronger than assuming that F stabilizes $(A(\theta), B(\theta))$.

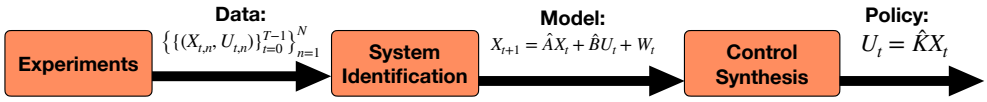

Fig. 2. A classic model-based pipeline for learning a controller from data.

To obtain a lower bound on the local minimax excess cost, we lower bound the maximization over $\theta' \in \mathcal{B}(\theta, \varepsilon)$ by an average over a distribution supported on $\mathcal{B}(\theta, \varepsilon)$. This reduces the problem to lower bounding a Bayesian complexity. Instead of fixing the parameter θ, we let Θ be a random vector taking values in $\mathbb{R}^{d_\Theta}$ and suppose that it has prior density λ. Doing so enables the use of information theoretic tools to lower bound the complexity of estimating the parameter from data. The lower bound on the maximization is shown in the following lemma.

Lemma 2.1: Fix $\varepsilon > 0$ and let λ be any prior on $\mathcal{B}(\theta, \varepsilon)$. Then for any $\pi \in \Pi^{\mathsf{lin}}$ with $\pi(X_t, \mathcal{Z}) = \hat{K}(\mathcal{Z})X_t$,

$$\sup_{\theta' \in \mathcal{B}(\theta)} \mathsf{EC}_T^\pi(\theta') \geq \mathbf{E}\,\mathrm{tr}\Big([\hat{K}(\mathcal{Z}) - K(\Theta)])^\top \Psi(\Theta)[\hat{K}(\mathcal{Z}) - K(\Theta)]\Sigma_\Theta^{\hat{K}(\mathcal{Z})}\Big),$$

where $\Sigma_\Theta^{\hat{K}(\mathcal{Z})} := \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}^\pi\big[X_t X_t^\top | \mathcal{Z}, \Theta\big]$. The expectation is over the prior $\Theta \sim \lambda$, and the randomness of both the offline rollouts and the evaluation rollout. We recall the shorthand $\Psi(\Theta) = B(\Theta)^\top P(\Theta) B(\Theta) + R$.

We may treat the data from offline experimentation, $\mathcal{Z}$, as an observation of the underlying parameter Θ. In particular, $\mathcal{Z}$ may be expressed as a random vector taking values in $\mathbb{R}^{NT(d_\mathsf{X}+d_\mathsf{U})}$ with conditional density $p(\cdot|\theta)$. The following Fisher information matrix and prior density concentration matrix measure estimation performance of Θ from the sample $\mathcal{Z}$ with respect to the square loss:

$$\mathtt{I}_p(\theta) := \int \left(\frac{\nabla_\theta p(z|\theta)}{p(z|\theta)}\right)\left(\frac{\nabla_\theta p(z|\theta)}{p(z|\theta)}\right)^\top p(z|\theta)dz, \quad (6)$$

$$\mathtt{J}(\lambda) := \int \left(\frac{\nabla_\theta \lambda(\theta)}{\lambda(\theta)}\right)\left(\frac{\nabla_\theta \lambda(\theta)}{\lambda(\theta)}\right)^\top \lambda(\theta)d\theta. \quad (7)$$

The first quantity (6) measures the information content of the sample $\mathcal{Z}$ with regards to Θ. The second quantity (7) measures the concentration of the prior density λ. As the gradient operator ∇_θ maps to a vector of dimension d_Θ, both $I_p(\theta)$ and $J(\lambda)$ are $d_\Theta \times d_\Theta$ dimensional. See [42] for further details about these integrals and their existence.

As we seek lower bounds for estimating $K(\Theta)$ instead of just Θ, we must account for the transformation from a quadratic loss over the error in esimating Θ to the error in estimating $K(\Theta)$, as appears in Lemma 2.1. To do so, we introduce the van Trees inequality [43], [44]. We first impose the following standard regularity conditions:

Assumption 2.1:

1) The prior λ is smooth with compact support.
2) The conditional density of $\mathcal{Z}$ given Θ, $p(z|\cdot)$, is continuously differentiable on the domain of λ for almost every z.
3) The score[4] has mean zero; $\int \left(\frac{\nabla_\theta p(z|\theta)}{p(z|\theta)}\right) p(z|\theta)dz = 0$.
4) $\mathtt{J}(\lambda)$ is finite and $\mathtt{I}_p(\theta)$ is a continuous function of θ on the domain of λ.
5) $\mathsf{vec}\, K$ is differentiable on the domain of λ.

The following theorem is a less general adaption from [44] which suffices for our needs.

Theorem 2.1 (Van Trees Inequality): Fix two random variables $(\mathcal{Z}, \Theta) \sim p(\cdot|\cdot)\lambda(\cdot)$ and suppose Assumption 2.1 holds. Let $\mathcal{G}$ be a $\sigma(\mathcal{Z})$-measurable event. Then for any $\sigma(\mathcal{Z})$-measurable $\hat{K}$:

$$\mathbf{E}\left[\mathsf{vec}(\hat{K}(\mathcal{Z}) - K(\Theta))\,\mathsf{vec}(\hat{K}(\mathcal{Z}) - K(\Theta))^\top \mathbf{1}_\mathcal{G}\right] \succeq \mathbf{E}[\mathsf{D}_\theta\,\mathsf{vec}\,K(\Theta)\mathbf{1}_\mathcal{G}]^\top [\mathbf{E}\,\mathtt{I}_p(\Theta) + \mathtt{J}(\lambda)]^{-1}\mathbf{E}[\mathsf{D}_\theta\,\mathsf{vec}\,K(\Theta)\mathbf{1}_\mathcal{G}]. \quad (8)$$

The notation $\mathsf{D}_\theta\,\mathsf{vec}\,K(\cdot)$ above follows the standard convention for a Jacobian: it stacks the transposed gradients of each element of $\mathsf{vec}\,K(\cdot)$ into a $d_\mathsf{X}d_\mathsf{U} \times d_\Theta$ dimensional matrix.

We see from Theorem 2.1 that the transformation to the error in estimating $K(\Theta)$ is accounted for by $\mathsf{D}_\theta\,\mathsf{vec}\,K(\cdot)$.

We now massage the lower bound in Lemma 2.1 to a form compatible with Theorem 2.1. Doing so requires us to express the lower bound as a quadratic form conditioned on some $\sigma(\mathcal{Z})$-measurable event $\mathcal{G}$. We therefore select an event $\mathcal{G}$ for which we may uniformly lower bound the quantities $\Psi(\Theta)$ and $\Sigma_\Theta^{\hat{K}(\mathcal{Z})}$. To this end, we define positive definite matrices $\Psi_{\theta,\varepsilon}$ and $\Sigma_{\theta,\varepsilon}$ that satisfy

$$\Psi(\theta') \succeq \Psi_{\theta,\varepsilon} \text{ and } \frac{1}{2}\Sigma_X(\theta') \succeq \Sigma_{\theta,\varepsilon} \quad \forall \theta' \in \mathcal{B}(\theta, \varepsilon). \quad (9)$$

The matrix $\Psi_{\theta,\varepsilon}$ will serve to uniformly lower bound $\Psi(\Theta)$. When the learned controller is close to the optimal controller, the covariance of the state under the learned controller will be close to the covariance of the state under the optimal controller, which is in turn lower bounded in terms of $\Sigma_{\theta,\varepsilon}$. In particular, if $\left\|\hat{K}(\mathcal{Z}) - K(\Theta)\right\|$ is sufficiently small, we can argue that $\Sigma_\Theta^{\hat{K}(\mathcal{Z})} \succeq \frac{1}{2}\Sigma_X(\Theta) \succeq \Sigma_{\theta,\varepsilon}$. The aforementioned condition on $\left\|\hat{K}(\mathcal{Z}) - K(\Theta)\right\|$ will hold only if there is a large amount of data available to fit $\hat{K}(\mathcal{Z})$. To achieve a bound that holds in the low data regime, we observe that the state covariance under the learned controller is always lower bounded by the noise covariance: $\Sigma_\Theta^\mathcal{Z} \succeq \Sigma_W$. For this reason, the subsequent results will be presented in two parts: one in which we condition on an event where $\left\|\hat{K}(\mathcal{Z}) - K(\theta)\right\|$ is small, and one that holds generally. To present these results concisely, the positive definite matrix $\Gamma_{\theta,\varepsilon}$ is used to denote either Σ_W or $\Sigma_{\theta,\varepsilon}$. The Kronecker product of these lower bounds arises frequently, motivating the shorthand

$$\Xi_{\theta,\varepsilon} := \Gamma_{\theta,\varepsilon} \otimes \Psi_{\theta,\varepsilon}. \quad (10)$$

[4]The score is the gradient of the log-likelihood. It evaluates to $\frac{\nabla_\theta p(z|\theta)}{p(z|\theta)}$.

Lemma 2.2 (Application of van Trees Inequality): For any smooth prior λ on $\mathcal{B}(\theta,\varepsilon)$ and any $\pi \in \Pi^{\mathsf{lin}}$ with $\pi(X_t,\mathcal{Z}) = \hat{K}(\mathcal{Z})X_t$,

$$\sup_{\theta'\in\mathcal{B}(\theta,\varepsilon)} \mathsf{EC}_T^\pi(\theta') \geq \frac{\operatorname{tr}\left(\Xi_{\theta,\varepsilon}\,\mathbf{E}[\mathsf{D}_\theta \operatorname{vec} K(\Theta)\mathbf{1}_{\mathcal{G}}]\,\mathbf{E}[\mathsf{D}_\theta \operatorname{vec} K(\Theta)\mathbf{1}_{\mathcal{G}}]^\top\right)}{\|\mathbf{E}\,\mathtt{I}_p(\Theta)+\mathtt{J}(\lambda)\|}, \tag{11}$$

where either:

1) $\Gamma_{\theta,\varepsilon} = \Sigma_W$ and $\mathcal{G} = \Omega$, or

2) $\Gamma_{\theta,\varepsilon} = \Sigma_{\theta,\varepsilon}$ and $\mathcal{G} = \mathcal{E}$, if $T \geq \sup_{\theta'\in\mathcal{B}(\theta,\varepsilon)} \frac{16\,\|\Sigma_X(\theta')\|}{\lambda_{\min}(\Sigma_X(\theta'))}$.

The event Ω is the entire sample space, i.e. $\mathbb{P}[\Omega]=1$, and

$$\mathcal{E} = \left\{ \sup_{\theta'\in\mathcal{B}(\theta,\varepsilon)} \left\|\hat{K}(\mathcal{Z}) - K(\theta')\right\| \leq \alpha \right\}$$

$$\alpha = \inf_{\theta'\in\mathcal{B}(\theta,\varepsilon)} \min\left\{ \frac{\|A_{cl}(\theta')\|}{\|B(\theta')\|}, \frac{\lambda_{\min}(\Sigma_X(\theta'))/24}{\|A_{cl}(\theta')\|\,\|B(\theta')\|\,\mathcal{J}(A_{cl}(\theta'))\,\|\Sigma_X(\theta')\|} \right\}.$$

Here, $A_{cl}(\theta) = A(\theta)+B(\theta)K(\theta)$ and $\mathcal{J}(A_{cl}(\theta)) = \sum_{t=0}^\infty \|A_{cl}(\theta)^t\|^2$.

The proof of the above result follows by applying Theorem 2.1 to the lower bound in Lemma 2.1 after replacing $\Psi(\Theta)$ by $\Psi_{\theta,\varepsilon}$ and $\Sigma_\Theta^{\hat{K}(\mathcal{Z})}$ by $\Gamma_{\theta,\varepsilon}\mathbf{1}_{\mathcal{G}}$.

Lemma 2.2 may be interpreted according to the following intuition. To design a controller that attains low cost, it is essential to distinguish between two nearby instances of the underlying parameter, θ and θ', from the experimental data, $\mathcal{Z}$. The Fisher Information term on the denominator of the bound in Lemma 2.2 captures the ease with which we can distinguish between θ and an infinitesimally perturbed θ' from the collected data $\mathcal{Z}$, and can be thought of as a signal-to-noise ratio. The derivative of the controller appearing on the numerator of the bound in Lemma 2.2 is a change of variables term that accounts for the extent to which infinitesimal perturbations of the underlying parameter impact the optimal controller gain. Sensitive perturbations are those which are difficult to detect from the collected data, yet lead to a large change in the controller gain. Such perturbations dictate the statistical hardness of learning a LQR controller. Motivated by this fact, we can select particularly sensitive perturbation directions of the underlying parameter which emphasize the hardness of the problem. To do so, we restrict the support of the prior λ to a lower dimensional subspace. Before presenting this result, it will be useful to see the expression for Fisher information matrix from this experimental setup. It can be shown via the chain rule of Fisher Information that

$$I_p(\theta) = \mathbf{E}_\theta \sum_{n=1}^N \sum_{t=0}^{T-1} \mathsf{D}_\theta \operatorname{vec}\begin{bmatrix}A(\theta) & B(\theta)\end{bmatrix}^\top \cdot [Z_{t,n}Z_{t,n}^\top \otimes \Sigma_W^{-1}]\,\mathsf{D}_\theta \operatorname{vec}\begin{bmatrix}A(\theta) & B(\theta)\end{bmatrix}, \tag{12}$$

where $Z_{t,n} = \begin{bmatrix}X_{t,n}\\ U_{t,n}\end{bmatrix}$. See, for instance, Lemma 3.1 of [8]. With this in hand, the following lemma provides a restriction to lower dimensional priors, which allows us to understand how poor conditioning of the information matrix along any particular parameter perturbation direction pushes through to a challenge in estimating the optimal controller.

Lemma 2.3: Let $V \in \mathbb{R}^{d_\Theta\times k}$ have orthonormal columns. For a smooth prior λ over $\left\{\theta + V\tilde{\theta} : \left\|\tilde{\theta}\right\| \leq \varepsilon\right\}$, and $\pi \in \Pi^{\mathsf{lin}}$ with $\pi(X_t,\mathcal{Z}) = \hat{K}(\mathcal{Z})X_t$,

$$\sup_{\theta'\in\mathcal{B}(\theta,\varepsilon)} \mathsf{EC}_T^\pi(\theta') \geq \frac{\operatorname{tr}\left(\Xi_{\theta,\varepsilon}\,\mathbf{E}[\mathsf{D}_\theta \operatorname{vec} K(\Theta)V\mathbf{1}_{\mathcal{G}}]\,\mathbf{E}[\mathsf{D}_\theta \operatorname{vec} K(\Theta)V\mathbf{1}_{\mathcal{G}}]^\top\right)}{\|V^\top(\mathbf{E}\,\mathtt{I}_p(\Theta)+\mathtt{J}(\lambda))V\|},$$

where $\Xi_{\theta,\varepsilon}$ is defined in (10) and $\mathcal{G}$ is defined in Lemma 2.2.

In the above lemma, the columns of V may be interpreted as perturbation directions of the system parameters.

We now upper bound the denominator arising in the above bound. In particular, we show how to bound the Fisher Information in any particular perturbation direction.

Lemma 2.4: For any $V \in \mathbb{R}^{d_\Theta\times k}$ with orthonormal columns,

$$\left\|V^\top \mathbf{E}[I_p(\Theta)]V\right\| \leq TN\bar{L},$$

where $\bar{L} = \sup_{\theta'\in\mathcal{B}(\theta,\varepsilon)} L(\theta')$ and

$$L(\theta') = \sup_{w\in\mathbf{span}(V),\|w\|\leq 1} \frac{4}{\lambda_{\min}(\Sigma_W)} \cdot \left(\nu_1(w)\left(\left\|\mathtt{dlyap}\big((A(\theta')+B(\theta')F)^\top,\Sigma_W\big)\right\| + \sigma_{\tilde{u}}^2\left(\sum_{t=0}^\infty \left\|(A(\theta')+B(\theta')F)^t B\right\|\right)^2\right) + 2\sigma_{\tilde{u}}^2\nu_2(w)\right). \tag{13}$$

Here, $\nu_1(w) = \|D_\theta \operatorname{vec} A(\theta')w\|^2 + 2\,\|D_\theta \operatorname{vec} B(\theta')w\|^2\,\|F\|^2$ and $\nu_2(w) = \|D_\theta \operatorname{vec} B(\theta')w\|^2$ are change of coordinate terms that quantify the impact of the perturbation direction on the information upper bound. We recall that $\sigma_{\tilde{u}}^2$ is the average exploratory input energy.

The quantity $\mathtt{dlyap}((A(\theta')+B(\theta')F)^\top,\Sigma_W)$ in the above bound is the controllability gramian from the noise to the state, while $\sigma_{\tilde{u}}^2(\sum_{t=0}^\infty \|(A(\theta')+B(\theta')F)^t B(\theta')\|)^2$ upper bounds the impact of exploratory input on the state during offline experimentation.

We now present our first main result: a non-asymptotic lower bound on the local minimax excess cost. As with Lemma 2.2, it is presented in two components: one that holds generally, and another that requires enough data such that any sufficiently good policy $\pi \in \Pi_{\mathsf{lin}}$ outputs a feedback controller $\hat{K}(\mathcal{Z})$ which is near optimal with high probability. Consequently, the burn-in times are larger for the second result, and the size of the prior, ε, is required to be small. We drop the dependence of A, B, P, Ψ, K, and Σ_X on θ when the argument is clear from context.

Theorem 2.2: Consider any matrix $V \in \mathbb{R}^{d_\Theta\times k}$ with $k \leq d_\Theta$ which has orthonormal columns. Let

$$G = \inf_{\theta',\tilde{\theta}\in\mathcal{B}(\theta,\varepsilon)} \operatorname{tr}\left(\Xi_{\theta,\varepsilon}\,\mathsf{D}_\theta \operatorname{vec} K(\theta')V\Big(\mathsf{D}_\theta \operatorname{vec} K(\tilde{\theta})V\Big)^\top\right),$$

and $\bar{L}$ be as in Lemma 2.4. Also let $\Xi_{\theta,\varepsilon}$ be as defined in (10). Then for any smooth prior λ over $\left\{\theta + V\tilde{\theta} : \left\|\tilde{\theta}\right\| \le \varepsilon\right\}$,

$$\mathcal{EC}_T^{\mathsf{lin}}(\theta,\varepsilon) \ge \frac{G}{8NT\bar{L}} \tag{14}$$

is satisfied for

1) $\Gamma_{\theta,\varepsilon} = \Sigma_W$ if $TN \ge \frac{\|J(\lambda)\|}{L}$.

2) $\Gamma_{\theta,\varepsilon} = \Sigma_{\theta,\varepsilon}$ if $T \ge \sup_{\theta' \in \mathcal{B}(\theta,\varepsilon)} \frac{16\|\Sigma_X(\theta')\|^2}{\lambda_{\min}(\Sigma_X(\theta'))}$, $TN \ge \frac{1}{L}\max\left\{\|J(\lambda)\|, \frac{G}{\lambda_{\min}(\Sigma_W)\lambda_{\min}(R)\alpha^2}\right\}$, and $\varepsilon \le \min\left\{\frac{\alpha}{2c_1}, c_2\right\}$, where

$$\begin{aligned}
c_1 &= 84\Phi^9\tau(A_{cl})L_{\mathrm{dyn}} \\
c_2 &= \frac{1}{10\tau(A_{cl})c_1}\min\left\{(1+\|A_{cl}\|)^{-2}, (1+\|P\|)^{-1}\right\} \\
\Phi &= (1+\max\left\{\|A\|, \|B\|, \|P\|, \|K\|, \left\|R^{-1}\right\|\right\}) \\
\tau(A_{cl}) &= \left(\sup_{k\ge 0}\left\{\left\|A_{cl}^k\right\|\rho(A_{cl})^{-k}\right\}\right)^2/(1-\rho(A_{cl})^2).
\end{aligned}$$

The theorem follows by bounding the probability of the event $\mathcal{G}$ in Lemma 2.3, then bounding the denominator using Lemma 2.4. The above result holds non-asymptotically. It will be helpful to present the result asymptotically, as the number of experiments tends to ∞ for an understanding of the dependence on control-theoretic quantities.

Corollary 2.1: For $\alpha \in (0, 1/2)$ and any $V \in \mathbb{R}^{d_\Theta \times k}$ with $k \le d_\Theta$ which has orthonormal columns, we have that

$$\liminf_{N\to\infty} \sup_{\theta' \in \mathcal{B}(\theta, N^{-\alpha})} N\mathsf{EC}_T^\pi(\theta') \ge \frac{G}{8TL(\theta)},$$

holds always for $\Gamma = \Sigma_W$ and for $\Gamma = \frac{1}{2}\Sigma_X$ if $T \ge \frac{16\|\Sigma_X\|^2}{\lambda_{\min}(\Sigma_X)}$, where L is as in Lemma 2.4 and

$$G = \mathrm{tr}\left((\Gamma \otimes \Psi)\,\mathsf{D}_\theta \,\mathsf{vec}\, K(\theta)V(\mathsf{D}_\theta \,\mathsf{vec}\, K(\theta)V)^\top\right).$$

Using a similar argument to the derivations above, it can be shown that the global minimax complexity is infinite.

Corollary 2.2: The global minimax excess cost is infinite for the class of scalar systems of the form: $X_{t+1} = aX_t + bU_t + W_t$, with $\theta = \begin{bmatrix} a & b \end{bmatrix}^\top$, and $Q = R = \Sigma_W = \sigma_{\tilde{u}}^2 = 1$. More precisely, for the class of stable scalar systems with the offline exploration policy $F = 0$, we have

$$\liminf_{N\to\infty} \sup_{a,b:|a|<1} N\mathsf{EC}_T^\pi(a,b) = \infty.$$

III. Consequences of the Lower Bound

In this section, we examine cases where the bound in Corollary 2.1 has interpretable dependence upon system properties. To do so, we restrict attention to the setting where all system parameters are unknown, i.e. $\mathsf{vec}\begin{bmatrix} A(\theta) & B(\theta)\end{bmatrix} = \theta$. In this setting, the quantity $\mathsf{D}_\theta \,\mathsf{vec}\begin{bmatrix} A(\theta) & B(\theta)\end{bmatrix}$ arising in the bounds from the previous section is the identity matrix.

The derivative of the controller multiplied by a matrix with orthonormal columns, $\mathsf{D}_\theta \,\mathsf{vec}\, K(\theta)V$, arises in the bounds from the previous section. In this section, this quantity is expressed in terms of the directional derivative of the controller in some direction v, denoted $d_vK(\theta)$. In particular, we represent the columns of V as $v = \mathsf{vec}\begin{bmatrix}\Delta_A & \Delta_B\end{bmatrix}$ for arbitrary perturbations Δ_A of A and Δ_B of B which satisfy $\left\|\begin{bmatrix}\Delta_A & \Delta_B\end{bmatrix}\right\|_F = 1$. The corresponding change in the closed-loop state matrix is denoted $\Delta_{A_{cl}} = \Delta_A + \Delta_B K$. Then $d_vK(\theta)$ is shown in Lemma B.1 of [4] to be

$$d_vK(\theta) = -\Psi^{-1}(\Delta_B^\top P A_{cl} + B^\top P \Delta_{A_{cl}} + B^\top P' A_{cl}), \tag{15}$$

where $P' = \mathtt{dlyap}(A_{cl}, A_{cl}^\top P \Delta_{A_{cl}} + \Delta_{A_{cl}}^\top P A_{cl})$. The subsequent sections study the bound from Corollary 2.1 under various perturbations $\begin{bmatrix}\Delta_A & \Delta_B\end{bmatrix}$.

A. Dimensional dependence

In the setting of online LQR for an unknown system, recent works [4], [8] obtaining lower bounds on the regret have used perturbation directions which cause tension between identification and control [45]. In particular, they considered the set of perturbation directions

$$\mathbf{\Delta} = \left\{\mathsf{vec}\begin{bmatrix}-\Delta K & \Delta\end{bmatrix} \middle| \Delta \in \mathbb{R}^{d_\mathsf{X} \times d_\mathsf{U}}, \left\|\begin{bmatrix}-\Delta K & \Delta\end{bmatrix}\right\|_F = 1\right\}. \tag{16}$$

For all such perturbations, $\Delta_{A_{cl}} = 0$, making it impossible to distinguish between the true parameters and the perturbed parameters online without sufficient exploratory input noise.

While the tension between identification and control is no longer present in the offline setting, $\mathbf{\Delta}$ retains the benefit that the directional derivative in (15) is easy to work with. In particular for any $v = \mathsf{vec}\begin{bmatrix}-\Delta K & \Delta\end{bmatrix} \in \mathbf{\Delta}$,

$$d_vK(\theta) = -\Psi^{-1}\Delta^\top P A_{cl}. \tag{17}$$

As the matrices Δ parametrizing $\mathbf{\Delta}$ are $d_\mathsf{X} \times d_\mathsf{U}$ dimensional, we may stack $d_\mathsf{X} d_\mathsf{U}$ orthogonal vectors v_i belonging $\mathbf{\Delta}$ into a matrix $V = \begin{bmatrix} v_1 & \dots & v_{d_\mathsf{X} d_\mathsf{U}}\end{bmatrix}$. This allows us to present a lower bound demonstrating the dependence of the LQR problem upon the system dimensions d_X and d_U.

Proposition 3.1: For $\alpha \in (0, 1/2)$,

$$\liminf_{N\to\infty} \sup_{\theta' \in \mathcal{B}(\theta, N^{-\alpha})} N\mathsf{EC}_T^\pi(\theta') \ge \frac{d_\mathsf{X} d_\mathsf{U} \lambda_{\min}(\Sigma_X - \Sigma_W)\lambda_{\min}(P)^2}{16T\,\|\Psi\|\left\|\begin{bmatrix}-K & I\end{bmatrix}\right\|^2 \tilde{L}},$$

as long as $T \ge \frac{16\|\Sigma_X\|^2}{\lambda_{\min}(\Sigma_X)}$. Here, $\tilde{L}$ is given by $L(\theta)$ as in (13) by replacing ν_1 with 1 and ν_2 with $1 + 2\|F\|^2$.

In addition to the system dimensions, we can interpret the remaining system-theoretic parameters. Note that $\tilde{L}$ bounds the information available from the offline experimentation. It depends on the norm of the controllability gramian from noise to the state, as well as $\sigma_{\tilde{u}}^2(\sum_{t=0}^\infty \|(A+BF)^t B\|)^2$, which bounds the impact of the exploratory input on the state. The Ψ in the denominator of the above bound may scale as $\lambda_{\max}(P)$, and therefore effectively cancels a $\lambda_{\min}(P)$ in the numerator for well-conditioned problems. This leaves a single $\lambda_{\min}(P)$ in the numerator. As $x^\top P x$ is the optimal objective value of the noiseless LQR problem starting from initial state x, the appearance of $\lambda_{\min}(P)$ in the bound captures the fact that as the system becomes harder to control, it also becomes harder to learn to control. Lastly, the variance term $\lambda_{\min}(\Sigma_X - \Sigma_W)$ implies that the excess

cost is large when the optimal closed-loop system has a large state covariance relative to the process noise covariance.

Remark 3.1: The dimensional dependence $d_{\mathsf{X}} d_{\mathsf{U}}$ is optimal up to constant factors for classes of under-actuated systems in which remaining system-theoretic quantities are constant with respect to system dimension. In particular, suppose that $\tilde{U}_{t,n} \sim \mathcal{N}(0, \sigma_u^2 I)$, and that $N \geq c d_{\mathsf{X}}$, for some universal constant c. In this case, Theorem 2 of [27] combined with Theorem 5.4 in [46] demonstrate the upper bound on the excess cost scales with $\frac{d_{\mathsf{X}} d_{\mathsf{U}}}{NT}$.

B. Exponential Lower Bounds

The previous section demonstrated a lower bound that scales with $d_{\mathsf{X}} d_{\mathsf{U}}$. Prior work [28] has shown that in the setting of online LQR, there exist classes of systems where the lower bounds on the regret scale exponentially with the state dimension. This is shown by demonstrating that particular system-theoretic terms, which are often treated as constant with respect to dimension, may grow exponentially with the state dimension. We demonstrate that in the setting of offline LQR, such systems still cause exponential dependence on dimension. Furthermore, because there are fewer restrictions upon the perturbation directions in the lower bound for the offline setting, we construct simpler systems which exhibit this behavior. In particular, consider the system

$$A = \begin{bmatrix} \rho & 2 & 0 & & 0 & 0 \\ & & & \ddots & & \\ 0 & 0 & 0 & & \rho & 2 \\ 0 & 0 & 0 & & 0 & \rho \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \tag{18}$$

with $0 < \rho < 1$, $F = 0$, $Q = I$, $R = 1$, and $\Sigma_W = I$. Using the perturbation direction $V = \mathsf{vec}\begin{bmatrix} 0 & B/\|B\|_F \end{bmatrix}$, Corollary 2.1 may be used to reach the following conclusion.

Proposition 3.2: For the system in (18) suppose $d_{\mathsf{X}} \geq 3$. Then for $\alpha \in (0, 1/2)$,

$$\liminf_{N \to \infty} \sup_{\theta' \in \mathcal{B}(\theta, N^{-\alpha})} N \mathsf{EC}_T^{\pi}(\theta') \geq \frac{\rho^2}{256 T \sigma_u^2} 4^{d_{\mathsf{X}} - 2}.$$

We have therefore demonstrated that accurately learning the LQR controller from offline data may require an amount of data that is exponential in the state dimension. The reason that this system is particularly challenging to learn to control is that a small misidentification of B causes the learner to apply slightly suboptimal control inputs, which are then amplified by the off-diagonal terms of A. The construction used, (18), avoids the two subsystem example that was used to derive exponential lower bounds for online LQR in [28]. A crucial reason that we are able to bypass such a construction in the offline setting is that the dominant statistical rate of $\frac{1}{NT}$ for offline LQR is present for any perturbation direction of the underlying parameters. In contrast, the regret in the online setting only has the dominant statistical rate in the directions defined by the perturbation set in (16).

C. Interesting System-Theoretic Quantities

A consequence of the result in Section III-B is that treating system-theoretic quantities as constant with respect to dimension, as is done in Section III-A, may fail to capture the difficulty of the problem. This leads to unfavorable aspects of the lower bound in Section III-A, such as the dependence of the denominator on $\|K\|$. Such an appearance indicates that for systems where the optimal LQR has a large gain, the lower bound becomes small. This is in contrast to our expectations, as a large optimal gain is often indicative of poor controllability (consider a scalar system, with $B \to 0$).

Motivated by the above discussion, we focus our attention on deriving bounds which have favorable dependence upon system-theoretic quantities. To do so, we examine a perturbation direction for which the lower bound from Corollary 2.1 reduces to easily interpretable quantities which align with our intuition. By taking $V = \mathsf{vec}\frac{\begin{bmatrix} A & B \end{bmatrix}}{\left\|\begin{bmatrix} A & B \end{bmatrix}\right\|_F}$, the directional derivative expression from (15) reduces to

$$d_V K(\theta) = \frac{2\Psi^{-1}(B^\top \mathtt{dlyap}(A_{cl}, P) A_{cl})}{\left\|\begin{bmatrix} A & B \end{bmatrix}\right\|_F}. \tag{19}$$

This leads to the following proposition.

Proposition 3.3: Suppose that R and $B^\top P B$ are simultaneously diagonalizable by U: $B^\top P B = U \Lambda_{B^\top P B} U^\top$ and $R = U \Lambda_R U^\top$, where $\Lambda_{B^\top P B}$ and Λ_R are diagonal. Also suppose that the diagonal entries of $\Lambda_{B^\top P B}$ are sorted in non-ascending order. Assume $T \geq \frac{16 \|\Sigma_X\|^2}{\lambda_{\min}(\Sigma_X)}$. Let $\tilde{L}$ be as in Proposition 3.1. Then for $\alpha \in (0, 1/2)$,

$$\liminf_{N \to \infty} \sup_{\theta' \in \mathcal{B}(\theta, N^{-\alpha})} N \mathsf{EC}_T^{\pi}(\theta') \geq \frac{\lambda_{\min}(\Sigma_X - \Sigma_W)}{2T \left\|\begin{bmatrix} A & B \end{bmatrix}\right\|_F^2 \tilde{L}} \cdot \inf_{i \in [d_{\mathsf{U}}]} \frac{\lambda_i(B^\top P B)}{\lambda_i(B^\top P B) + \Lambda_{R,ii}} \sum_{j=1}^{d_{\mathsf{U}}} \lambda_{n-j}(\mathtt{dlyap}(A_{cl}, P)).$$

The assumption that R and $B^\top P B$ are simultaneously diagonalizable is satisfied if R is chosen as a scalar multiple of the identity. As in Proposition 3.1, $\lambda_{\min}(\Sigma_X - \Sigma_W)$ highlights the dependence on the closed-loop state covariance, and $\tilde{L}$ describes the how easily system is excited offline. Rather than the appearance of $\left\|\begin{bmatrix} -K & I \end{bmatrix}\right\|$ in the denominator, as we saw in Proposition 3.1, we have $\left\|\begin{bmatrix} A & B \end{bmatrix}\right\|_F$. Therefore, the bound does not diminish as a result of a large optimal controller gain. Lastly, observe that $\sum_{j=1}^{du} \lambda_{n-j}(\mathtt{dlyap}(A_{cl}, P))$ replaces $\lambda_{\min}(P)$ from Proposition 3.1. This quantity captures the d_{U} smallest eigenvalues rather than just the smallest. If $d_{\mathsf{U}} = d_{\mathsf{X}}$, we get all eigenvalues of $\mathtt{dlyap}(A_{cl}, P)$. Further note that the eigenvalues of $\mathtt{dlyap}(A_{cl}, P)$ diverge as A_{cl} approaches marginal stability, leading to an infinite excess cost.

IV. Conclusion

We presented lower bounds for offline linear-quadratic control problems. The focus was to understand the fundamental limitations of learning controllers from offline data in terms of system-theoretic properties. Several interesting consequences arose, such as the fact that our lower bound achieves the optimal dimensional dependence $d_{\mathsf{X}} d_{\mathsf{U}}$ for underactuated systems. We also showed that there exist systems where the sample complexity is exponential with the system

dimension, d_{X}. We finally demonstrated that the lower bound scales in a naturally with familiar system-theoretic constants including the eigenvalues of the Riccati solution. Future work will consider the partially observed setting.

REFERENCES

[1] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334–1373, 2016.

[2] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[3] M. G. Azar, I. Osband, and R. Munos, "Minimax regret bounds for reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2017, pp. 263–272.

[4] M. Simchowitz and D. Foster, "Naive exploration is optimal for online lqr," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8937–8948.

[5] Y. Abbasi-Yadkori and C. Szepesvári, "Regret Bounds for the Adaptive Control of Linear Quadratic Systems," in *Proceedings of the 24th Annual Conference on Learning Theory*, 2011, pp. 1–26.

[6] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the Sample Complexity of the Linear Quadratic Regulator," *Foundations of Computational Mathematics*, pp. 1–47, 2019.

[7] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," *arXiv preprint arXiv:2005.01643*, 2020.

[8] I. Ziemann and H. Sandberg, "Regret lower bounds for learning linear quadratic gaussian systems," *arXiv preprint arXiv:2201.01680. Manuscript in preparation*, 2022.

[9] A. Tsiamis, I. Ziemann, N. Matni, and G. J. Pappas, "Statistical learning theory for control: A finite sample perspective," *arXiv preprint arXiv:2209.05423*, 2022.

[10] L. Ljung, *System identification*. Springer, 1998.

[11] T. Söderström, *Discrete-Time Stochastic systems: Estimation and Control*. Springer Science & Business Media, 2002.

[12] A. Wagenmaker, M. Simchowitz, and K. Jamieson, "Task-optimal exploration in linear dynamical systems," *arXiv preprint arXiv:2102.05214*, 2021.

[13] J. Chen and C. N. Nett, "The caratheodory-fejer problem and $\mathcal{H}_\infty$ a time domain approach," in *Proceedings of 32nd IEEE Conference on Decision and Control*. IEEE, 1993, pp. 68–73.

[14] A. J. Helmicki, C. A. Jacobson, and C. N. Nett, "Control oriented system identification: a worst-case/deterministic approach in $\mathcal{H}_\infty$," *IEEE Transactions on Automatic control*, vol. 36, no. 10, pp. 1163–1176, 1991.

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

[16] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite time identification in unstable linear systems," *Automatica*, vol. 96, pp. 342–353, 2018.

[17] T. Sarkar and A. Rakhlin, "Near Optimal Finite Time Identification of Arbitrary Linear Dynamical Systems," in *International Conference on Machine Learning*, 2019, pp. 5610–5618.

[18] S. Oymak and N. Ozay, "Non-asymptotic identification of lti systems from a single trajectory," in *2019 American control conference (ACC)*. IEEE, 2019, pp. 5655–5661.

[19] T. Sarkar, A. Rakhlin, and M. A. Dahleh, "Finite time lti system identification," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 1186–1246, 2021.

[20] A. Tsiamis and G. J. Pappas, "Finite sample analysis of stochastic system identification," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 3648–3654.

[21] B. Lee and A. Lamperski, "Non-asymptotic closed-loop system identification using autoregressive processes and hankel model reduction," in *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 2020, pp. 3419–3424.

[22] Y. Zheng and N. Li, "Non-asymptotic identification of linear dynamical systems using multiple trajectories," *IEEE Control Systems Letters*, vol. 5, no. 5, pp. 1693–1698, 2020.

[23] Y. Jedra and A. Proutiere, "Sample Complexity Lower Bounds for Linear System Identification," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 2676–2681.

[24] A. Tsiamis and G. J. Pappas, "Linear systems can be hard to learn," *arXiv preprint arXiv:2104.01120*, 2021.

[25] K. Zhou, J. Doyle, and K. Glover, *Robust and Optimal Control*, ser. Feher/Prentice Hall Digital and. Prentice Hall, 1996.

[26] H. A. Simon, "Dynamic Programming under Uncertainty with a Quadratic Criterion Function," *Econometrica, Journal of the Econometric Society*, pp. 74–81, 1956.

[27] H. Mania, S. Tu, and B. Recht, "Certainty Equivalence is Efficient for Linear Quadratic Control," in *Advances in Neural Information Processing Systems*, 2019, pp. 10 154–10 164.

[28] A. Tsiamis, I. Ziemann, M. Morari, N. Matni, and G. J. Pappas, "Learning to control linear systems can be hard," in *Conference on Learning Theory*. PMLR, 2022, pp. 3820–3857.

[29] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1467–1476.

[30] I. Ziemann, A. Tsiamis, H. Sandberg, and N. Matni, "How are policy gradient methods affected by the limits of control?" *arXiv preprint arXiv:2206.06863. To appear at CDC'22*, 2022.

[31] K. J. Åström and B. Wittenmark, "On self tuning regulators," *Automatica*, vol. 9, no. 2, pp. 185–199, 1973.

[32] T. L. Lai, "Asymptotically Efficient Adaptive Control in Stochastic Regression Models," *Advances in Applied Mathematics*, vol. 7, no. 1, pp. 23–45, 1986.

[33] T. L. Lai and C.-Z. Wei, "Extended Least squares and their Applications to Adaptive Control and Prediction in Linear Systems," *IEEE Transactions on Automatic Control*, vol. 31, no. 10, pp. 898–906, 1986.

[34] Y. Ouyang, M. Gagrani, and R. Jain, "Control of Unknown Linear Systems with Thompson Sampling," in *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2017, pp. 1198–1205.

[35] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "Regret Bounds for Robust Adaptive Control of the Linear Quadratic Regulator," in *Advances in Neural Information Processing Systems*, 2018, pp. 4188–4197.

[36] M. Abeille and A. Lazaric, "Improved Regret Bounds for Thompson Sampling in Linear Quadratic Control Problems," *Proceedings of Machine Learning Research*, vol. 80, 2018.

[37] A. Cohen, T. Koren, and Y. Mansour, "Learning Linear-Quadratic Regulators Efficiently with only $\sqrt{T}$ Regret," *arXiv preprint arXiv:1902.06223*, 2019.

[38] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Input Perturbations for Adaptive Control and Learning," *Automatica*, vol. 117, p. 108950, 2020.

[39] Y. Jedra and A. Proutiere, "Minimal expected regret in linear quadratic control," *arXiv preprint arXiv:2109.14429*, 2021.

[40] A. Cassel, A. Cohen, and T. Koren, "Logarithmic Regret for Learning Linear Quadratic Regulators Efficiently," *arXiv preprint arXiv:2002.08095*, 2020.

[41] B. D. Lee, I. Ziemann, A. Tsiamis, H. Sandberg, and N. Matni, "The fundamental limitations of learning linear-quadratic regulators," *arXiv preprint arXiv:2303.15637*, 2023.

[42] I. A. Ibragimov and R. Z. Has'minskii, *Statistical Estimation: Asymptotic Theory*. Springer Science & Business Media, 2013, vol. 16.

[43] H. L. van Trees, *Detection, Estimation, and Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory*. John Wiley & Sons, 2004.

[44] B.-Z. Bobrovsky, E. Mayer-Wolf, and M. Zakai, "Some Classes of Global Cramér-Rao Bounds," *The Annals of Statistics*, pp. 1421–1438, 1987.

[45] J. W. Polderman, "On the Necessity of Identifying the True Parameter in Adaptive LQ Control," *Systems & control letters*, vol. 8, no. 2, pp. 87–91, 1986.

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