Naive Exploration is Optimal for Online LQR

Max Simchowitz [1] **Dylan J. Foster** [2]

Abstract

We consider the problem of online adaptive control of the linear quadratic regulator, where the true system parameters are unknown. We prove new upper and lower bounds demonstrating that the optimal regret scales as $\widetilde{\Theta}(\sqrt{d_{\mathbf{u}}^2 d_{\mathbf{x}} T})$, where T is the number of time steps, $d_{\mathbf{u}}$ is the dimension of the input space, and $d_{\mathbf{x}}$ is the dimension of the system state. Notably, our lower bounds rule out the possibility of a $\mathrm{poly}(\log T)$-regret algorithm, which had been conjectured due to the apparent strong convexity of the problem. Our upper bound is attained by a simple variant of *certainty equivalent control*, where the learner selects control inputs according to the optimal controller for their estimate of the system while injecting exploratory random noise (Mania et al., 2019).

Central to our upper and lower bounds is a new approach for controlling perturbations of Riccati equations called the *self-bounding ODE method*, which we use to derive suboptimality bounds for the certainty equivalent controller synthesized from estimated system dynamics. This in turn enables regret upper bounds which hold for *any stabilizable instance* and scale with natural control-theoretic quantities.

1. Introduction

Reinforcement learning has recently achieved great success in application domains including Atari (Mnih et al., 2015), Go (Silver et al., 2016), and robotics (Lillicrap et al., 2015). All of these breakthroughs leverage data-driven methods for continuous control in large state spaces. Their success, along with challenges in deploying RL in the real world, has led to renewed interest on developing continuous control algorithms with improved reliability and sample efficiency. In particular, on the theoretical side, there has been a push to develop a non-asymptotic theory of data-driven continuous control, with an emphasis on understanding key algorithmic principles and fundamental limits.

In the non-asymptotic theory of reinforcement learning, much attention has been focused on the so-called "tabular" setting where states and actions are discrete, and the optimal rates for this setting are by now relatively well-understood (Jaksch et al., 2010; Dann & Brunskill, 2015; Azar et al., 2017). Theoretical results for continuous control setting have been more elusive, with progress spread across various models (Kakade et al., 2003; Munos & Szepesvári, 2008; Jiang et al., 2017; Jin et al., 2020), but the linear-quadratic regulator (LQR) problem has recently emerged as a candidate for a standard benchmark for continuous control and RL. For tabular reinforcement learning problems, it is widely understood that careful exploration is essential for sample efficiency. Recently, however, it was shown that for the online variant of the LQR problem, relatively simple exploration strategies suffice to obtain the best-known performance guarantees (Mania et al., 2019). In this paper, we address a curious question raised by these results: Is sophisticated exploration helpful for LQR, or is linear control in fact substantially easier than the general reinforcement learning setting? More broadly, we aim to shed light on the question:

To what extent to do sophisticated exploration strategies improve learning in online linear-quadratic control?

Is ε-Greedy Optimal for Online LQR? In the LQR problem, the system state $\mathbf{x}_t$ evolves according to

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + B\mathbf{u}_t + \mathbf{w}_t, \quad \text{where} \quad \mathbf{x}_1 = 0, \qquad (1.1)$$

and where $\mathbf{u}_t \in \mathbb{R}^{d_{\mathbf{u}}}$ is the learner's control input, $\mathbf{w}_t \in \mathbb{R}^{d_{\mathbf{x}}}$ is a noise process drawn as $\mathbf{w}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I)$, and $A \in \mathbb{R}^{d_{\mathbf{x}} \times d_{\mathbf{x}}}$, $B \in \mathbb{R}^{d_{\mathbf{x}} \times d_{\mathbf{u}}}$ are unknown system matrices.

Initially the learner has no knowledge of the system dynamics, and their goal is to repeatedly select control inputs and observing states over T rounds so as to minimize their total cost $\sum_{t=1}^{T} c(\mathbf{x}_t, \mathbf{u}_t)$, where $c(x, u) = x^\top R_{\mathbf{x}} x + u^\top R_{\mathbf{u}} u$ is a known quadratic function. In the online variant of the LQR problem, we measure performance via *regret* to the

[1]UC Berkeley [2]Massachusetts Institute of Technology. Correspondence to: Max Simchowitz <msimchow@berkeley.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

optimal linear controller:

$$\mathrm{Regret}_{A,B,T}[\pi] = \left[\sum_{t=1}^{T} c(\mathbf{x}_t, \mathbf{u}_t)\right] - T \min_K \mathcal{J}_{A,B}[K], \tag{1.2}$$

where K is a linear state feedback policy and—letting $\mathbb{E}_{A,B,K}[\cdot]$ denote expectation under this policy—where

$$\mathcal{J}_{A,B}[K] := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{A,B,K}\left[\sum_{t=1}^{T} c(\mathbf{x}_t, \mathbf{u}_t)\right],$$

is the average infinite-horizon cost of K, which is finite as long as K is *stabilizing* in the sense that $\rho(A+BK) < 1$, where $\rho(\cdot)$ denotes the spectral radius.[1] We further define $\mathcal{J}^\star_{A,B} := \min_K \mathcal{J}_{A,B}[K]$.

This setting has enjoyed substantial development beginning with the work of (Abbasi-Yadkori & Szepesvári, 2011), and following a line of successive improvements (Dean et al., 2018; Faradonbeh et al., 2018a; Cohen et al., 2019; Mania et al., 2019), the best known algorithms for online LQR have regret scaling as $\sqrt{T}$.

We investigate a question that has emerged from this research: The role of exploration in linear control. The first approach in this line of work, (Abbasi-Yadkori & Szepesvári, 2011), proposed a sophisticated though computationally inefficient strategy based on *optimism in the face of uncertainty*, upon which (Cohen et al., 2019) improved to ensure optimal $\sqrt{T}$-regret and polynomial runtime. Another approach which enjoys $\sqrt{T}$-regret, due to (Mania et al., 2019), employs a variant of the classical ε-greedy exploration strategy (Sutton & Barto, 2018) known in control literature as *certainty equivalence*: At each timestep, the learner computes the greedy policy for the current estimate of the system dynamics, then follows this policy, adding exploration noise proportional to ε. While appealing in its simplicity, ε-greedy has severe drawbacks for general reinforcement learning problems: For tabular RL, it leads to exponential blowup in the time horizon (Kearns et al., 2000), and for multi-armed bandits, bandit linear optimization, and contextual bandits, it leads to suboptimal dependence on the time horizon T (Langford & Zhang, 2007).

This begs the question: Can we improve beyond $\sqrt{T}$ regret for online LQR using more sophisticated exploration strategies? Or is exploration in LQR simply much easier than in general reinforcement learning settings? One natural hope would be to achieve logarithmic (i.e. $\mathrm{poly}(\log T)$) regret. After all, online LQR has strongly convex loss functions, and this is a sufficient condition for logarithmic regret in many simpler online learning and optimization problems (Vovk, 2001; Hazan et al., 2007; Rakhlin & Sridharan, 2014), as well as LQR with known dynamics but potentially changing costs (Agarwal et al., 2019b). More subtly, the $\sqrt{T}$ online LQR regret bound of (Mania et al., 2019) requires that the pair $(A_\star, B_\star)$ be *controllable*;[2] it was not known if naive exploration attains this rate for arbitrary *stabilizable* problem instances, or if it necessarily leverages controllability to ensure its efficiency.

[1] For potentially asymmetric matrix $A \in \mathbb{R}^{d\times d}$, $\rho(A) := \max\{|\lambda| \mid \lambda \text{ is an eigenvalue for } A\}$.

1.1. Contributions

We prove new upper and lower bounds which characterize the minimax optimal regret for online LQR as $\widetilde{\Theta}(\sqrt{d_{\mathbf{u}}^2 d_{\mathbf{x}} T})$. Beyond dependence on the horizon T, dimensions $d_{\mathbf{x}}, d_{\mathbf{u}}$, and logarithmic factors, our bounds depend only on *operator* norms of transparent, control theoretic quantities, which do not hide additional dimension dependence. Our main lower bound is Theorem 1, which implies that no algorithm can improve upon $\sqrt{T}$ regret for online LQR, and so simple ε-greedy exploration is indeed *rate-optimal*.

Theorem 1 (informal). For every sufficiently non-degenerate problem instance and every (potentially randomized) algorithm, there exists a nearby problem instance on which the algorithm must suffer regret at least $\widetilde{\Omega}(\sqrt{d_{\mathbf{u}}^2 d_{\mathbf{x}} T})$.

Perhaps more surprisingly, our main upper bound shows that a simple variant of certainty equivalence is also *dimension-optimal*, in that it asymptotically matches the $\sqrt{d_{\mathbf{u}}^2 d_{\mathbf{x}} T}$ lower bound of Theorem 1.

Theorem 2 (informal). Certainty equivalent control with continual ε-greedy exploration (Algorithm 1) has regret at most $\widetilde{O}\left(\sqrt{d_{\mathbf{u}}^2 d_{\mathbf{x}} T} + d_{\mathbf{x}}^2\right)$ for every stabilizable online LQR instance.

Our upper bound *does not* require controllability, and is the first bound for *any* algorithm to attain the optimal dimension dependence. In comparison, result of (Mania et al., 2019) guarantees $\sqrt{(d_{\mathbf{x}} + d_{\mathbf{u}})^3 T}$ regret *and* imposes strong additional assumptions. In the many control settings where $d_{\mathbf{u}} \ll d_{\mathbf{x}}$, our bound constitutes a significant improvement. Other approaches *not* based on certainty equivalence suffer considerably larger dimension dependence (Cohen et al., 2019). Together, Theorem 1 and Theorem 2 characterize the asymptotic minimax regret for online LQR, showing that there is little room for improvement over naive exploration.

Our results leverage a new perturbation bound for controllers synthesized via certainty equivalence. Unlike prior bounds due to (Mania et al., 2019), our guarantee depends

[2] $(A_\star, B_\star)$ are said to be controllable if and only the *controllability Gramian* $\mathcal{C}_n\mathcal{C}_n^\top := \sum_{i=0}^n A_\star^i B_\star B_\star^\top (A_\star^i)^\top$ is strictly positive definite for some $n \geq 0$. For any n for which $\mathcal{C}_n \succ 0$, the upper bounds of (Mania et al., 2019) scale polynomially in $n, 1/\lambda_{\min}(\mathcal{C}_n\mathcal{C}_n^\top)$. Controllability implies stabilizability, but the converse is not true.

only on natural control-theoretic quantities, and crucially does not require controllability of the system.

Theorem 3 (informal). Fix an instance (A, B). Let $(\widehat{A}, \widehat{B})$, and let $\widehat{K}$ denote the optimal infinite horizon controller from instance $(\widehat{A}, \widehat{B})$. Then if $(\widehat{A}, \widehat{B})$ are sufficiently close to (A, B), we have

$$\mathcal{J}_{A,B}[\widehat{K}] - \mathcal{J}^{\star}_{A,B} \leq 142\|P\|_{\mathrm{op}}^{8} \cdot (\|\widehat{A} - A\|_{\mathrm{F}}^{2} + \|\widehat{B} - B\|_{\mathrm{F}}^{2}),$$

where P is the solution to the DARE for the system (A, B).

For simplicity, the bound above assumes the various normalization conditions on the noise and cost matrices, described in Section 1.4. With these conditions, our perturbation bound only requires that the operator norm distance between $(\widehat{A}, \widehat{B})$ and (A, B) be at most $1/\mathrm{poly}(\|P\|_{\mathrm{op}})$. Hence, we establish perturbation bounds for which both the scaling of the deviation and the region in which the bound applies can be quantified in terms of a single quantity: the norm of DARE solution P. We prove this bound through a new technique we term the *Self-Bounding ODE method*, described below. Beyond removing the requirement of controllability, we believe this method is simpler and more transparent than past approaches.

1.2. Our Approach

Both our lower and upper bounds are facilitated by the *self-bounding ODE method*, a new technique for establishing perturbation bounds for the Riccati equations that characterize the optimal value function and controller for LQR. The method sharpens existening perturbation bounds, weakens controllability and stability assumptions required by previous work (Dean et al., 2018; Faradonbeh et al., 2018a; Cohen et al., 2019; Mania et al., 2019), and yields an upper bound whose leading terms depend only on the horizon T, dimension parameters $d_{\mathbf{x}}, d_{\mathbf{u}}$, and the control-theoretic parameters sketched in the prequel.

In more detail, if (A, B) is stabilizable and $R_{\mathbf{x}}, R_{\mathbf{u}} \succ 0$, there exists a unique PSD solution $P_{\infty}(A, B)$ for the *discrete algebraic Riccati equation* (DARE),

$$P = A^{\top}PA + R_{\mathbf{x}} - A^{\top}PB(R_{\mathbf{u}} + B^{\top}PB)^{-1}B^{\top}PA \quad (1.3)$$

The unique optimal infinite-horizon controller is given by

$$K_{\infty}(A, B) = -(R_{\mathbf{u}} + B^{\top}P_{\infty}(A, B)B)^{-1}B^{\top}P_{\infty}(A, B)A,$$

and the matrix $P_{\infty}(A, B)$ induces a positive definite quadratic form which can be interpreted as a value function for the LQR problem.

Both our upper and lower bounds make use of novel perturbation bounds to control the change in P_{∞} and K_{∞} when we move from a nominal instance (A, B) to a nearby instance $(\widehat{A}, \widehat{B})$. For our upper bound, these are used to show that a good estimator for the nominal instance leads to a good controller, while for our lower bounds, they show that the converse is true. The self-bounding ODE method allows us to prove perturbation guarantees that depend only on the norm of the value function $\|P_{\infty}(A, B)\|_{\mathrm{op}}$ for the nominal instance, which is a weaker assumption that subsumes previous conditions. The key observation underpinning the method is that the norm of the directional derivative of $\frac{d}{dt}P_{\infty}(A(t), B(t))\big|_{t=u}$ at a point $t = u$ along a line $(A(t), B(t))$ is bounded in terms of the magnitude of $\|P_{\infty}(A(u), B(u))\|$; we call this the *self-bounding* property. From this relation, we show that bounding the norm of the derivatives reduces to solving a scalar ordinary differential equation, whose derivative saturates the scalar analogue of this self-bounding property. Notably, this technique does not require that the system be controllable, and in particular does not yield guarantees which depend on the smallest singular value of the controllability matrix as in (Mania et al., 2019). Moreover, given estimates $(\widehat{A}, \widehat{B})$ and an upper-bound on their deviation from the true system $(A_{\star}, B_{\star})$, our bound allows the learner to check whether the certainty-equivalent controller synthesized from $\widehat{A}, \widehat{B}$ stabilizes the true system and satisfies the preconditions for our perturbation bounds.

On the lower bound side, we begin with a nominal instance (A_0, B_0) and consider a packing of alternative instances within a small neighborhood. Specifically, if K_0 is the optimal controller for (A_0, B_0), we consider perturbations of the form $(A_{\Delta}, B_{\Delta}) = (A_0 - \Delta K_0, B_0 + \Delta)$ for $\Delta \in \mathbb{R}^{d_{\mathbf{u}} d_{\mathbf{x}}}$. The self-bounding ODE method facilitates a perturbation analysis which implies that the optimal controller K_{Δ} on each alternative (A_{Δ}, B_{Δ}) deviates from K_0 by $\|K_0 - K_{\Delta}\|_{\mathrm{F}} \geq \Omega(\|\Delta\|_{\mathrm{F}})$ for non-degenerate instances. Using this reasoning, we show that any low-regret algorithm can approximately recover the perturbation Δ.

On the other hand, if the learner selects inputs $\mathbf{u}_t = K_0\mathbf{x}_t$ according to the optimal control policy for the nominal instance, all alternatives are *indistinguishable* from the nominal instance. Indeed, the structure of our perturbations ensures that $A_{\Delta} + B_{\Delta}K_0 = A_0 + B_0K_0$ for all choices of Δ. Thus, since low regret implies identification of the perturbation, any low regret learner must substantially deviate from the nominal controller K_0. Equivalently, this can be understood as a consequence of the fact that playing $\mathbf{u}_t = K_0\mathbf{x}_t$ yields a degenerate covariance matrix for the random variable $(\mathbf{x}_t, \mathbf{u}_t)$, and thus some deviation from K_0 is required to ensure this covariance is full rank. The regret scales proportionally to the deviation from K_0, which scales proportionally to the minimum eigenvalues of the aforementioned covariance matrix, but the estimation error rate scales as $1/T$ (the typical "fast rate") times the *inverse* of these eigenvalues. Balancing the tradeoffs leads to the

"slow" $\sqrt{T}$ lower bound. Crucially, our argument exploits a fundamental tension between control and indentification in linear systems, first described by Polderman (1986), and summarized in Polderman (1989).

Our upper bound refines the certainty equivalent control strategy proposed in (Mania et al., 2019) by re-estimating the system parameters on a doubling epoch schedule to advantage of the endogenous excitation supplied by the $\mathbf{w}_t$-sequence. A careful analysis of the least squares estimator shows that the error in a $d_{\mathbf{x}}d_{\mathbf{u}}$-dimensional subspace decays as $\mathcal{O}\left(1/\sqrt{t}\right)$, and in the remaining $d_{\mathbf{x}}^2$ dimensions decays at a *fast rate* of $\mathcal{O}\left(1/t\right)$.

Related Work Non-asymptotic guarantees for learning linear dynamical systems have been the subject of intense recent interest (Dean et al.; Hazan et al., 2017; Tu & Recht, 2018; Hazan et al., 2018; Simchowitz et al., 2018; Sarkar & Rakhlin, 2019; Simchowitz et al., 2019; Mania et al., 2019; Sarkar et al., 2019). The online LQR setting we study was introduced by (Abbasi-Yadkori & Szepesvári, 2011), which considers the problem of controlling an unknown linear system under stationary stochastic noise.[3] They showed that an algorithm based on the optimism in the face of uncertainty (OFU) principle enjoys $\sqrt{T}$, but their algorithm is computationally inefficient and their regret bound depends exponentially on dimension. The problem was revisited by (Dean et al., 2018), who showed that an explicit explore-exploit scheme based on ε-greedy exploration and certainty equivalence achieves $T^{2/3}$ regret efficiently, and left the question of obtaining $\sqrt{T}$ regret efficiently as an open problem. This issue was subsequently addressed by (Faradonbeh et al., 2018a) and (Mania et al., 2019), who showed that certainty equivalence obtains $\sqrt{T}$ regret, and (Cohen et al., 2019), who achieve $\sqrt{T}$ regret using a semidefinite programming relaxation for the OFU scheme. The regret bounds in (Faradonbeh et al., 2018a) do not specify dimension dependence, and (for $d_{\mathbf{x}} \geq d_{\mathbf{u}}$), the dimension scaling of (Cohen et al., 2019) can be as large as $\sqrt{d_{\mathbf{x}}^{16}T}$;[4] (Mania et al., 2019) incurs an almost-optimal dimension dependence of $\sqrt{d_{\mathbf{x}}^3 T}$ (suboptimal when $d_{\mathbf{u}} \ll d_{\mathbf{x}}$), but at the expense of imposing a strong controllability assumption.

The question of whether regret for online LQR could be improved further (for example, to $\log T$) remained open, and was left as a conjecture by (Faradonbeh et al., 2018b). Our lower bounds resolve this conjecture by showing that $\sqrt{T}$-regret is optimal. Moreover, by refining the upper bounds of (Mania et al., 2019), our results show that the asymptotically optimal regret is $\widetilde{\Theta}(\sqrt{d_{\mathbf{u}}^2 d_{\mathbf{x}} T})$, and that this achieved by certainty equivalence. Beyond attaining the optimal dimension dependence, our upper bounds also enjoy refined dependence on problem parameters, and do not require a-priori knowledge of these parameters.

Logarithmic regret bounds are ubiquitous in online learning and optimization problems with strongly convex loss functions (Vovk, 2001; Hazan et al., 2007; Rakhlin & Sridharan, 2014). (Agarwal et al., 2019b) demonstrate that for the problem of controlling an *known* linear dynamic system with adversarially chosen, strongly convex costs, logarithmic regret is also attainable. Our $\sqrt{T}$ lower bound shows that the situation for the online LQR with an *unknown* system parallels that of bandit convex optimization, where (Shamir, 2013) showed that $\sqrt{T}$ is optimal even for strongly convex quadratics. That is, in spite of strong convexity of the losses, issues of partial observability prevent fast rates in both settings.

Our lower bound carefully exploits the online LQR problem structure to show that $\sqrt{T}$ is optimal. To obtain optimal dimension dependence for the lower bound, we build on well-known lower bound technique for adaptive sensing based on Assouad's lemma (Arias-Castro et al., 2012) (see also (Assouad, 1983; Yu, 1997)).

Finally, a parallel line of research provides Bayesian and frequentist regret bounds for online LQR based on Thompson sampling (Ouyang et al., 2017; Abeille & Lazaric, 2017), with (Abeille & Lazaric, 2018) demonstrating $\sqrt{T}$-regret for the scalar setting. Unfortunately, Thompson sampling is not computationally efficient for the LQR.

1.3. Organization

Section 1.4 introduces basic notation and definitions. Section 2 introduces our main results: In Section 2.1 and Section 2.2 we state our main lower and upper bounds respectively and give an overview of the proof techniques, and in Section 2.3 we instantiate and compare these bounds for the simple special case of strongly stable systems. In Section 3 we introduce the self-bounding ODE method and show how it is used to prove key perturbation bounds used in our main results. All additional proofs and proof details are given in the appendix, whose organization is described at length in Appendix A. Future directions and open problems are discussed in Section 4.

1.4. Preliminaries

Assumptions We restrict our attention *stabilizable* systems (A, B) for which there exists a stabilizing controller K such that $\rho(A + BK) < 1$. Note that this does not require that the system be controllable. We further assume that $R_{\mathbf{u}} = I$ and $R_{\mathbf{x}} \succeq I$. The first can be enforced by a change of basis in input space, and the second can be enforced by

[3] A more recent line of work studies a more general *non-stochastic* noise regime (see (Agarwal et al., 2019a) et seq.), which we do not consider in this work.

[4] The regret bound of (Cohen et al., 2019) scales as $d_{\mathbf{x}}^3\sqrt{T} \cdot (\mathcal{J}^\star_{A_\star,B_\star})^5$; typically, $\mathcal{J}^\star_{A_\star,B_\star}$ scales linearly in $d_{\mathbf{x}}$

rescaling the state space, increasing the regret by at most a multiplicative factor of $\min\{1, 1/\sigma_{\min}(R_{\mathbf{x}})\}$. We also assume that the process noise $\mathbf{w}_t$ has identity covariance. We note that non-identity noise can be adressed via a change of variables, and in Appendix I.8 we sketch extensions of our results to (a) independent, sub-Gaussian noise with bounded below covariance, and (b) more general martingale noise, where we remark on how to achieve optimal rates in the regime $d_{\mathbf{x}} \lesssim d_{\mathbf{u}}^2$.

Algorithm Protocol and Regret Formally, the learner's (potentially randomized) decision policy is modeled as a sequence of mappings $\pi = (\pi_t)_{t=1}^T$, where each function π_t maps the history $(\mathbf{x}_1, \dots, \mathbf{x}_t, \mathbf{u}_1, \dots, \mathbf{u}_{t-1})$ and an internal random seed ξ to an output control signal $\mathbf{u}_t$. For a linear system evolving according to Eq. (1.1) and policy π, we let $\mathbb{P}_{A,B,\pi}$ and $\mathbb{E}_{A,B,\pi}[\cdot]$ denote the probability and expectation with respect to the dynamics (1.1) and randomization of π. For such a policy, we use the notation $\mathrm{Regret}_{A,B,T}[\pi]$ as in Eq. (1.2) for regret, which is a random variable with law $\mathbb{P}_{A,B,\pi}[\cdot]$. We prove high-probability upper bounds on $\mathrm{Regret}_{A,B,T}[\pi]$, and prove lower bounds on the expected regret $\mathbb{E}\mathrm{Regret}_{A,B,T}[\pi] := \mathbb{E}_{A,B,\pi}[\mathrm{Regret}_{A,B,T}[\pi]]$.[5]

Additional Notation For vectors $x \in \mathbb{R}^d$, $\|x\|$ denotes the ℓ_2 norm. For matrices $X \in \mathbb{R}^{d_1 \times d_2}$, $\|X\|_{\mathrm{op}}$ denotes the spectral norm, and $\|X\|_{\mathrm{F}}$ the Frobenius norm. When $d_1 \le d_2$, $\sigma_1(X), \dots, \sigma_{d_1}(X)$ denote the singular values of X, arranged in decreasing order. We say $f \lesssim g$ to denote that $f(x) \le Cg(x)$ for a universal constant C, and $f \lessapprox g$ to denote informal inequality. We write $f \eqsim g$ if $g \lesssim f \lesssim g$.

For "starred" systems $(A_\star, B_\star)$, we adopt the shorthand $P_\star := P_\infty(A_\star, B_\star)$, $K_\star := K_\infty(A_\star, B_\star)$ for the optimal controller, $\mathcal{J}_\star := \mathcal{J}^\star_{A_\star,B_\star} := \mathcal{J}_{A_\star,B_\star}[K_\star]$ for optimal cost, and $A_{\mathrm{cl},\star} := A_\star + B_\star K_\star$ for the optimal closed loop system. We define $\Psi_\star := \max\{1, \|A_\star\|_{\mathrm{op}}, \|B_\star\|_{\mathrm{op}}\}$ and $\Psi_{B_\star} := \max\{1, \|B_\star\|_{\mathrm{op}}\}$. For systems (A_0, B_0), we let $\mathcal{B}_{\mathrm{op}}(\epsilon; A_0, B_0) = \{(A, B) \mid \|A - A_0\|_{\mathrm{op}} \vee \|B - B_0\|_{\mathrm{op}} \le \epsilon\}$ denote the set of nearby systems in operator norm.

2. Main Results

We now state our main upper and lower bounds for online LQR and give a high-level overview of the proof techniques behind both results. At the end of the section, we instantiate and compare the two bounds for the simple special case of strongly stable systems.

[5] One might consider as a stronger benchmark described the expected loss of the optimal policy for *fixed horizon* T. A fortiori, our lower bounds apply for this benchmark as well: In view of the proof of Lemma F.3 in Appendix G.2, this benchmark differs from $T\,\mathcal{J}^\star_{A,B}$ by a constant factor which depends on (A, B) but *does not grow with* T.

Both our upper and lower bounds start with the following question: Suppose that the learner is selecting near optimal control inputs $\mathbf{u}_t \approx K_\star \mathbf{x}_t$, where $K_\star = K_\infty(A_\star, B_\star)$ is the optimal controller for the system $(A_\star, B_\star)$. What information can she glean about the system?

2.1. Lower Bound

We provide a *local minimax* lower bound, which captures the difficulty of ensuring low regret on both a *nominal instance* $(A_\star, B_\star)$ and on the hardest nearby alternative. For a distance parameter $\epsilon > 0$, we define the local minimax complexity at scale ϵ as

$$\mathcal{R}_{A_\star,B_\star,T}(\epsilon) := \min_{\pi} \max_{A,B} \Big\{ \mathbb{E}\mathrm{Regret}_{A,B,T}[\pi] : \|A - A_\star\|_{\mathrm{F}}^2 \vee \|B - B_\star\|_{\mathrm{F}}^2 \le \epsilon \Big\}.$$

Local minimax complexity captures the idea certain instances $(A_\star, B_\star)$ are more difficult than others, and allows us to provide lower bounds that scale only with control-theoretic parameters of the nominal instance. Of course, the local minimax lower bound immediately implies a lower bound on the global minimax complexity as well.[6]

Intuition Behind the Lower Bound. We show that if the learner plays near-optimally on every instance in the neighborhood of $(A_\star, B_\star)$, then there is a $d_{\mathbf{x}} d_{\mathbf{u}}$-dimensional subspace of system parameters that the learner must explore by deviating from $K_\star$ when the underlying instance is $(A_\star, B_\star)$. Even though the system parameters can be estimated at a fast rate, such deviations preclude logarithmic regret.

In more detail, if the learner plays near-optimally, she is not be able to distinguish between whether the instance she is interacting with is $(A_\star, B_\star)$, or another system of the form

$$(A, B) = (A_\star - K_\star \Delta, B_\star + \Delta), \tag{2.1}$$

for some perturbation $\Delta \in \mathbb{R}^{d_{\mathbf{x}} \times d_{\mathbf{u}}}$. This is because all the obsevations $(\mathbf{x}_t, \mathbf{u}_t)$ generated by the optimal controller lie in the subspace $\{(x, u) : u - K_\star x = 0\}$, and likewise all observations generated by any near-optimal controller approximately lie in this subspace. Since the learner cannot distinguish between $(A_\star, B_\star)$ and (A, B), she will also play $\mathbf{u}_t \approx K_\star \mathbf{x}_t$ on (A, B). This leads to poor regret when the instance is (A, B), since the optimal controller in this case has $\mathbf{u}_t = K_\infty(A, B)\mathbf{x}_t$. This is made concrete by the next lemma, which shows to a first-order approximation that if Δ is large, the distance between $K_\star$ and $K_\infty(A, B)$ must also be large.

[6] Some care must be taken in defining the global complexity, or it may well be infinite. One sufficient definition, which captures prior work, is to consider minimax regret over all instances subject to a global bound on $\|P_\star\|$, $\|B_\star\|$, and so on.

Lemma 2.1 (Derivative Computation (Abeille & Lazaric (2018), Proposition 2)). *Let $(A_\star, B_\star)$ be stabilizable, and recall $A_{\mathrm{cl},\star} := A_\star + B_\star K_\star$. Then,*

$$\frac{d}{dt} K_\infty(A_\star - t\Delta K_\star, B_\star + t\Delta)\big|_{t=0} = -(R_{\mathbf{u}} + B_\star^\top P_\star B_\star)^{-1} \cdot \Delta^\top P_\star A_{\mathrm{cl},\star}.$$

In particular, when the closed loop system $A_{\mathrm{cl},\star}$ is (approximately) well-conditioned, the optimal controllers for $(A_\star, B_\star)$ and for (A, B) are $\Omega(\|\Delta\|_{\mathrm{F}})$-apart, and so the learner cannot satisfy both $\mathbf{u}_t \approx K_\star \mathbf{x}_t$ and $\mathbf{u}_t \approx K_\infty(A,B)\mathbf{x}_t$ simultaneously. More precisely, for the learner to ensure $\sum_t \|\mathbf{x}_t - K_\infty(A,B)\mathbf{u}_t\|_{\mathrm{F}}^2 \lessapprox d_{\mathbf{x}} d_{\mathbf{u}} \epsilon^2$ on every instance, she must deviate from optimal by at least $\sum_{t=1}^T \|\mathbf{x}_t - K_\star \mathbf{u}_t\|_{\mathrm{F}}^2 \gtrapprox d_{\mathbf{u}} T/\epsilon^2$ on the optimal instance; the $d_{\mathbf{u}}$ factor here comes from the necessity of exploring all control-input directions. Balancing these terms leads to the final $\Omega(\sqrt{T d_{\mathbf{u}} d_{\mathbf{x}}^2})$ lower bound (proven in Appendix F).

Theorem 1. *Let $c_1, p > 0$ denote universal constants. For $m \in [d_{\mathbf{x}}]$, define $\nu_m := \sigma_m(A_{\mathrm{cl},\star})/\|R_{\mathbf{u}} + B_\star^\top P_\star B_\star\|_{\mathrm{op}}$. Then if $\nu_m > 0$, we have*

$$\mathcal{R}_{A_\star, B_\star, T}(\epsilon_T) \gtrsim \sqrt{d_{\mathbf{u}}^2 m T} \cdot \frac{1 \wedge \nu_m^2}{\|P_\star\|_{\mathrm{op}}^2},$$

where $\epsilon_T = \sqrt{d_{\mathbf{u}}^2 m / T}$, provided that T is at least $c_1(\|P_\star\|_{\mathrm{op}}^p (d_{\mathbf{u}} m \vee \frac{d_{\mathbf{x}}^2 \Psi_{B_\star}^4 (1 \vee \nu_m^{-4})}{m d_{\mathbf{u}}^2} \vee d_{\mathbf{x}} \log(1 + d_{\mathbf{x}} \|P_\star\|_{\mathrm{op}}))$.

Let us briefly discuss some key features of Theorem 1.

- The only system-dependent parameters appearing in the lower bound are the operator norm bounds $\Psi_{B_\star}$ and $\|P_\star\|_{\mathrm{op}}$, which only depend on the nominal instance. The latter parameter is finite whenever the system is stabilizable, and does not explicitly depend on the spectral radius or strong stability parameters.
- The lower bound takes $\epsilon_T \propto T^{-1/2}$, so the alternative instances under consideration converge to the nominal instance $(A_\star, B_\star)$ as $T \to \infty$.
- The theorem can be optimized for each instance by tuning the dimension parameter $m \in [d_{\mathbf{x}}]$: The leading $\sqrt{d_{\mathbf{u}}^2 m T}$ term is increasing in m, while the parameter ν_m scales with $\sigma_m(A_{\mathrm{cl},\star})$ and thus is decreasing in m. The simplest case is when $\sigma_m(A_{\mathrm{cl},\star})$ is bounded away from 0 for $m \gtrsim d_{\mathbf{x}}$; here we obtain the optimal $\sqrt{d_{\mathbf{u}}^2 d_{\mathbf{x}} T}$ lower bound. In particular, if $d_{\mathbf{u}} \le d_{\mathbf{x}}/2$, we can choose $m = \frac{1}{2} d_{\mathbf{x}}$ to get $\sigma_m(A_{\mathrm{cl},\star}) \ge \sigma_{\min}(A_\star)$.

2.2. Upper Bound

While playing near-optimally prevents the learner from ruling out perturbations of the form Eq. (2.1), she can rule perturbations in orthogonal directions. Indeed, if $\mathbf{u}_t \approx K_\star \mathbf{x}_t$, then $\mathbf{x}_{t+1} \approx (A_\star + B_\star K_\star)\mathbf{x}_t + \mathbf{w}_t$. As a result, the persistent noise process $\mathbf{w}_t$ allows the learner recover the closed loop dynamics matrix $A_{\mathrm{cl},\star} = A_\star + B_\star K_\star$ to Frobenius error $d_{\mathbf{x}}\epsilon$ after just $T \gtrapprox 1/\epsilon^2$ steps, regardless of whether she incorporates additional exploration (Simchowitz et al., 2018). Hence, for perturbations perpendicular to those in Eq. (2.1), the problem closely resembles a setting where $\log T$ is achievable.

Our main algorithm, Algorithm 1, is detailed in Appendix H. It is an ε-greedy scheme that takes advantage of this principle. The full pseudocode and analysis are deferred to Appendix H, but we sketch the intuition here. The algorithm takes as input a stabilizing controller K_0 and proceeds in epochs k of length $\tau_k = 2^k$. After an initial burn-in period ending with epoch k_{safe}, the algorithm can ensure the reliability of its synthesized controllers, and uses a (projected) least-squares estimate $(\widehat{A}_k, \widehat{B}_k)$ of $(A_\star, B_\star)$ to synthesize a controller $\widehat{K}_k = K_\infty(\widehat{A}_k, \widehat{B}_k)$ known as the *certainty equivalent* controller. The learner then selects inputs by adding white Gaussian noise with variance σ_k^2: $\mathbf{u}_t = \widehat{K}_t \mathbf{x}_t + \mathcal{N}(0, \sigma_k^2 I)$. We show that this scheme exploits the rapid estimation along directions orthogonal to those in Eq. (2.1), leading to optimal dimension dependence.

To begin, we show (Theorem 3) that the cost of the certainty-equivalent controller is bounded by the estimation error for $\widehat{A}_k$ and $\widehat{B}_k$, i.e.

$$\mathcal{J}_{A_\star, B_\star}[\widehat{K}_k] - \mathcal{J}_\star \lesssim \mathrm{poly}(\|P_\star\|_{\mathrm{op}}) \cdot (\|\widehat{A}_k - A_\star\|_{\mathrm{F}}^2 + \|\widehat{B}_k - B_\star\|_{\mathrm{F}}^2),$$

once $(\widehat{A}_k, \widehat{B}_k)$ are sufficiently accurate, as guaranteed by the burn-in period. Through a regret decomposition based on the Hanson-Wright inequality (Lemma I.1), we next show that the bulk of the algorithm's regret scales as the sum of the suboptimality in the controller for a given epoch, plus the cost of the exploratory noise: $\sum_{k=k_{\mathrm{safe}}}^{\log_2 T} \tau_k \left(\mathcal{J}_{A_\star, B_\star}[\widehat{K}_k] - \mathcal{J}_\star\right) + d_{\mathbf{u}} \tau_k \sigma_k^2 \lessapprox \sum_{k=k_{\mathrm{safe}}}^{\log_2 T} \tau_k \left(\|\widehat{A}_k - A_\star\|_{\mathrm{F}}^2 + \|\widehat{B}_k - B_\star\|_{\mathrm{F}}^2\right) + d_{\mathbf{u}} \tau_k \sigma_k^2$. In the above, we also incur a term of approximately $\sum_{k=k_{\mathrm{safe}}}^{\log_2 T} \sqrt{(d_{\mathbf{x}} + d_{\mathbf{u}})\tau_k} \lesssim \sqrt{T(d_{\mathbf{x}} + d_{\mathbf{u}})}$, which is lower order than the overall regret of $\sqrt{T d_{\mathbf{x}} d_{\mathbf{u}}^2}$. This term arises from the random fluctuations of the costs around their expectation, and crucially, the Hanson-Wright inequality allows us to pay of the *square root* of the dimension.[7]

[7] The use of the Hanson-Wright crucially leverages independence of the noise process; for general sub-Gaussian martingale noise, an argument based on martingale concentration would mean that the fluctuations contribute $(d_{\mathbf{x}} + d_{\mathbf{u}})\sqrt{T}$ to the regret up to logarithmic factors, yielding an overall regret of $\sqrt{\max\{d_{\mathbf{x}}, d_{\mathbf{u}}^2\} d_{\mathbf{x}} T}$. This is suboptimal regret for $d_{\mathbf{x}} \gg d_{\mathbf{u}}^2$, but still an improvement over the $\sqrt{(d_{\mathbf{x}} + d_{\mathbf{u}})^3 T}$-bound of (Mania et al., 2019). It is un-

Paralleling the lower bound, the analysis crucially relies on the exploratory noise to bound the error in the $d_{\mathbf{x}}d_{\mathbf{u}}$-dimensional subspace corresponding to Eq. (2.1), as the error in this subspace grows as $\frac{d_{\mathbf{x}}d_{\mathbf{u}}}{\sigma_k^2\tau_k}$. However, for the directions parallel to those in Eq. (2.1), the estimation error is at most $d_{\mathbf{x}}^2/\tau_k$, and so the total regret is bounded as $\mathrm{Regret}_{A_\star,B_\star,T}[\mathsf{Alg}] \lessapprox \sum_{k=k_{\mathrm{safe}}}^{\log_2 T} \tau_k \left(\frac{d_{\mathbf{x}}^2}{\tau_k} + \frac{d_{\mathbf{x}}d_{\mathbf{u}}}{\tau_k\sigma_k^2}\right) + d_{\mathbf{u}}\tau_k\sigma_k^2 \approx d_{\mathbf{x}}^2\log T + \sum_{k=1}^{\log_2 T}\frac{d_{\mathbf{x}}d_{\mathbf{u}}}{\sigma_k^2} + d_{\mathbf{u}}\tau_k\sigma_k^2$. Trading off $\sigma_k^2 = \sqrt{d_{\mathbf{x}}/\tau_k}$ gives regret $d_{\mathbf{x}}^2\log T + \sum_{k=1}^{\log_2 T}\sqrt{d_{\mathbf{x}}d_{\mathbf{u}}^2\tau_k} \approx d_{\mathbf{x}}^2\log T + \sqrt{d_{\mathbf{x}}d_{\mathbf{u}}^2 T}$. We emphasize that to ensure that the $d_{\mathbf{x}}^2$ term in this bound scales only with $\log T$ due to rapid exploration perpendicular to Eq. (2.1), and it is crucial that the algorithm uses doubling epochs to take advantage of this. We now state the full guarantee.

Theorem 2. *When Algorithm 1 is invoked with stabilizing controller K_0 and confidence parameter $\delta \in (0, 1/T)$, it guarantees that with probability at least $1-\delta$, $\mathrm{Regret}_T[\mathsf{Alg}; A_\star, B_\star]$ is bounded as*

$$\lesssim \sqrt{d_{\mathbf{u}}^2 d_{\mathbf{x}} T \cdot \Psi_{B_\star}^2 \|P_\star\|_{\mathrm{op}}^{11} \log\tfrac{\|P_\star\|_{\mathrm{op}}}{\delta}} + d^2\cdot\mathcal{P}_0\Psi_{B_\star}^6\|P_\star\|_{\mathrm{op}}^{11}(1+\|K_0\|_{\mathrm{op}}^2)\log\frac{d\Psi_{B_\star}\mathcal{P}_0}{\delta}\log^2\frac{1}{\delta},$$

where $\mathcal{P}_0 := \mathcal{J}_{A_\star,B_\star}[K_0]/d_{\mathbf{x}}$ is the normalized cost of K_0, and $d = d_{\mathbf{x}} + d_{\mathbf{u}}$.

Ignoring dependence on problem parameters, the upper bound of Theorem 2 scales asymptotically as $\sqrt{d_{\mathbf{u}}^2 d_{\mathbf{x}} T}$, matching our lower bound. Like the lower bound, the theorem depends on the instance $(A_\star, B_\star)$ only through the operator norm bounds $\Psi_{B_\star}$ and $\|B_\star\|_{\mathrm{op}}$. Similar to previous work (Dean et al., 2018; Mania et al., 2019), the regret bound has additional dependence on the stabilizing controller K_0 through $\|K_0\|_{\mathrm{op}}$ and $\mathcal{P}_0$, but these parameters only affect the lower-order terms.

2.3. Consequences for Strongly Stable Systems

To emphasize the dependence on dimension and time horizon in our results, we now present simplified findings for a special class of *strongly stable* systems.

Definition 2.1 (Strongly Stable System (Cohen et al., 2018))**.** We say that $A_\star$ is (γ,κ)-strongly stable if there exists a transform T such that $\|T\|_{\mathrm{op}}\cdot\|T^{-1}\|_{\mathrm{op}} \le \kappa$ and $\|TA_\star T^{-1}\|_{\mathrm{op}} \le 1-\gamma$. When $A_\star$ is (γ,κ)-strongly stable, we define $\gamma_{\mathrm{sta}} := \gamma/\kappa^2$.

For the simplified results in this section we make the following assumption.

clear if one can do better in this setting without improved concentration bounds for quadratic forms of martingale vectors, because it is unclear how an algorithm can ameliorate these random fluctuations.

Assumption 1. *The nominal instance $(A_\star, B_\star)$ is such that $A_\star$ is (γ,κ)-strongly stable and $\|B_\star\|_{\mathrm{op}} \le 1$. Furthermore, $R_{\mathbf{x}} = R_{\mathbf{u}} = I$.*

For strongly stable systems under Assumption 1, our main lower bound (Theorem 1) takes the following particularly simple form.

Corollary 1 (Lower Bound for Strongly Stable Systems)**.** *Suppose that Assumption 1 holds, and that $d_{\mathbf{u}} \le \frac{1}{2}d_{\mathbf{x}}$ and $\sigma_{\min}(A_\star) > 0$.[8] Then for any $T \ge (d_{\mathbf{x}}d_{\mathbf{u}} + d_{\mathbf{x}}\log d_{\mathbf{x}})\mathrm{poly}(1/\gamma_{\mathrm{sta}}, 1/\sigma_{\min}(A_\star))$, we have*

$$\mathcal{R}_{A_\star,B_\star,T}(\varepsilon_T) \gtrsim \sqrt{d_{\mathbf{u}}^2 d_{\mathbf{x}} T}\cdot\sigma_{\min}(A_\star)^2\gamma_{\mathrm{sta}}^4,$$

where $\varepsilon_T := \sqrt{d_{\mathbf{u}}^2 d_{\mathbf{x}}/T}$.

The upper bound from Theorem 2 takes on a similarly simple form, and is seen to be nearly matching.

Corollary 2 (Upper Bound for Strongly Stable Systems)**.** *Suppose that Assumption 1 holds. Then Algorithm 1 with stabilizing controller $K_0 = 0$ and confidence parameter $\delta \in (0, 1/T)$, ensures that probability at least $1-\delta$, $\mathrm{Regret}_T[\mathsf{Alg}; A_\star, B_\star]$ is bounded as*

$$\lesssim \sqrt{d_{\mathbf{u}}^2 d_{\mathbf{x}} T\cdot\gamma_{\mathrm{sta}}^{-11}\log\frac{1}{\delta\gamma_{\mathrm{sta}}}} + (d_{\mathbf{x}}+d_{\mathbf{u}})^2\gamma_{\mathrm{sta}}^{-12}\log\frac{d}{\delta\gamma_{\mathrm{sta}}}\log^2\frac{1}{\delta}.$$

We observe that the leading $\sqrt{d_{\mathbf{u}}^2 d_{\mathbf{x}} T}$ terms in the upper and lower bounds differ only by factors polynomial in γ_{sta}, as well as a $\sigma_{\min}(A_\star)$ factor incurred by the lower bound. The lower order term $(d_{\mathbf{x}}+d_{\mathbf{u}})^2$ in the upper bound appears unavoidable, but we leave a complementary lower bound for future work. Both corollaries hold because strong stability immediately implies a bound on $\|P_\star\|_{\mathrm{op}}$.

Proof of Corollary 1 and Corollary 2. First, observe that under Assumption 1, $\Psi_{B_\star} \le 1$. Next, note that if $d_{\mathbf{u}} < d_{\mathbf{x}}/2$, then for $m = \lceil d_{\mathbf{x}}/2\rceil$, $\sigma_m(A_{\mathrm{cl},\star}) = \sigma_m(A_\star + B_\star K_\star) \ge \sigma_{m+d_{\mathbf{u}}}(A_\star + B_\star K_\star) \ge \sigma_{\min}(A_\star)$. This gives $\nu_m \ge \sigma_{\min}(A_\star)/(1+\|P_\star\|_{\mathrm{op}})$. Finally, Lemma B.7 (stated and proven in Appendix B.3.1) gives $\|P_\star\|_{\mathrm{op}} \le \gamma_{\mathrm{sta}}^{-1}$. Plugging these three observations into Theorem 1 and Theorem 2 concludes the proof. □

3. Perturbation Bounds via the Self-Bounding ODE Method

Both Theorem 1 and Theorem 2 scale only with the natural system parameter $\|P_\star\|_{\mathrm{op}}$, and avoid explicit dependence

[8] The assumption $d_{\mathbf{u}} \le \frac{1}{2}d_{\mathbf{x}}$ can be replaced with $d_{\mathbf{u}} \le \alpha d_{\mathbf{x}}$ for any $\alpha < 1$, and can be removed entirely for special instances. See Corollary 7 in Appendix G.7 for more details.

on the spectral radius or strong stability parameters found in prior work. This is achieved using the *self-bounding ODE* method, a new technique for deriving bounds on perturbations to the DARE solution $P_\infty(A, B)$ and corresponding controller $K_\infty(A, B)$ as the matrices A and B are varied. This method gives a general recipe for establishing perturbation bounds for solutions to implicit equations. It depends only on the norms of the system matrices and DARE solution $P_\infty(A, B)$, and it applies to all stabilizable systems, even those that are not controllable.

In this section we give an overview of the self-bounding ODE method and use it to prove a simplified version of the main perturbation bound used in our main upper and lower bounds. To state the perturbation bound, we first define the following problem-dependent constants.

$$C_{\mathrm{safe}}(A, B) = 54\|P_\infty(A, B)\|_{\mathrm{op}}^5, \quad \text{and}$$
$$C_{\mathrm{est}}(A, B) = 142\|P_\infty(A, B)\|_{\mathrm{op}}^8. \tag{3.1}$$

The parameter $C_{\mathrm{safe}}(A, B)$ determines the radius of admissible perturbations, while the parameter $C_{\mathrm{est}}(A, B)$ determines the quality of controllers synthesized from the resulting perturbation. The main perturbation bound is as follows.

Theorem 3. *Let $(A_\star, B_\star)$ be a stabilizable system. Given an alternate pair of matrices $(\widehat{A}, \widehat{B})$, for each $\circ \in \{\mathrm{op}, F\}$ define $\epsilon_\circ := \max\{\|\widehat{A} - A_\star\|_\circ, \|\widehat{B} - B_\star\|_\circ\}$. Then if $\epsilon_{\mathrm{op}} \le 1/C_{\mathrm{safe}}(A_\star, B_\star)$,*

1. $\|P_\infty(\widehat{A}, \widehat{B})\|_{\mathrm{op}} \lesssim \|P_\star\|_{\mathrm{op}}$ *and* $\|K_\star - K_\infty(\widehat{A}, \widehat{B})\|_{\mathrm{op}} \lesssim \frac{1}{\|P_\star\|_{\mathrm{op}}^{3/2}}$.
2. $\mathcal{J}_{A_\star, B_\star}[K_\infty(\widehat{A}, \widehat{B})] - \mathcal{J}^\star_{A_\star, B_\star} \le C_{\mathrm{est}}(A_\star, B_\star)\epsilon_{\mathrm{F}}^2$.

This theorem is a simplification of a stronger version, Theorem 5, stated and proven in Appendix B.1. Additional perturbation bounds are detailed in Appendix B.1; notably, Theorem 11 shows that the condition $\epsilon_{\mathrm{op}} \le 1/C_{\mathrm{safe}}(A_\star, B_\star)$ can be replaced by a condition that can be certificated from an approximate estimate of the system. In the remainder of this section, we sketch how to use the self-bounding ODE method to prove the following slightly more general version of the first part of Theorem 3.

Proposition 4. *Let $(A_\star, B_\star)$ be a stabilizable system and let $(\widehat{A}, \widehat{B})$ be an alternate pair of matrices. Then, if $u := 8\|P_\star\|_{\mathrm{op}}^2\epsilon_{\mathrm{op}} < 1$, the pair $(\widehat{A}, \widehat{B})$ is stabilizable and the following bounds hold:*

1. $\|P_\infty(\widehat{A}, \widehat{B})\|_{\mathrm{op}} \le (1-u)^{-1/2}\|P_\star\|_{\mathrm{op}}$.
2. *For each $\circ \in \{\mathrm{op}, F\}$, $\|K_\infty(\widehat{A}, \widehat{B}) - K_\star\|_\circ \le 7(1-u)^{-7/4}\|P_\star\|_{\mathrm{op}}^{7/2}\epsilon_\circ$.*

To begin proving the proposition, set $\Delta_A := \widehat{A} - A_\star$ and $\Delta_B := \widehat{B} - B_\star$. We consider a linear curve between the two instances, parameterized by $t \in [0, 1]$:

$$(A(t), B(t)) = (A_\star + t\Delta_A, B_\star + t\Delta_B)\,. \tag{3.2}$$

At each point t for which $(A(t), B(t))$ is stabilizable, the DARE has a unique solution, which allows us to define associated optimal cost matrices, controllers, and closed-loop dynamics matrices:

$$P(t) := P_\infty(A(t), B(t)), \quad K(t) := K_\infty(A(t), B(t))$$
$$\text{and} \quad A_{\mathrm{cl}}(t) := A(t) + B(t)K(t). \tag{3.3}$$

Our strategy will be to show that $P(t)$ and $K(t)$ are in fact smooth curves, and then obtain uniform bounds on $\|P'(t)\|_\circ$ and $\|K'(t)\|_\circ$ over the interval $[0, 1]$, yielding perturbation bounds via the mean value theorem. To start, we express the derivatives of the DARE in terms of Lyapunov equations.

Definition 3.1 (Discrete Lyapunov Equation)**.** Let $X, Y \in \mathbb{R}^{d_{\mathtt{x}} \times d_{\mathtt{x}}}$ with $Y = Y^\top$ and $\rho(X) < 1$. We let $\mathcal{T}_X[P] := X^\top PX - X$, and let $\mathsf{dlyap}(X, Y)$ denote the unique PSD solution $\mathcal{T}_X[P] = Y$. We let $\mathsf{dlyap}[X] := \mathsf{dlyap}(X, I)$.

The following lemma (proven in Appendix C.2) serves as the basis for our computations, and also establishes the requisite smoothness required to take derivatives.

Lemma 3.1 (Derivative and Smoothness of the DARE)**.** *Let $(A(t), B(t))$ be an analytic curve, and define $\Delta_{A_{\mathrm{cl}}}(t) := A'(t) + B'(t)K_\infty(A(t), B(t))$. Then for any t such that $(A(t), B(t))$ is stabilizable, the functions $P(u)$ and $K(u)$ are analytic in a neighborhood around t, and we have $P'(u) = \mathsf{dlyap}(A_{\mathrm{cl}}(u), Q_1(u))$, where $Q_1(u) := A_{\mathrm{cl}}(u)^\top P(u)\Delta_{A_{\mathrm{cl}}}(u) + \Delta_{A_{\mathrm{cl}}}(u)^\top P(u)A_{\mathrm{cl}}(u)$.*

Lemma 3.1 expresses $P'(t)$ as the solution to an ordinary differential equation. While the lemma guarantees local existence of the derivatives, it is not clear that the entire curve $(A(t), B(t)), t \in [0, 1]$ is stabilizable. However, since ODEs are locally guaranteed to have solutions, we should only expect trouble when the corresponding ODE becomes ill-defined, i.e. if $P'(t)$ escapes to infinity. We circumvent this issue by observing that $P'(t)$ satisfies the following self-bounding property.

Lemma 3.2 (Bound on First Derivatives)**.** *Let $(A(t), B(t))$ be an analytic curve. Then, for all t at which $(A(t), B(t))$ is stabilizable, we have $\|P'(t)\|_\circ \le 4\|P(t)\|_{\mathrm{op}}^3\epsilon_\circ$, and $\|K'(t)\|_\circ \le 7\|P(t)\|_{\mathrm{op}}^{7/2}\epsilon_\circ$.*

The bound on $P'(t)$ above follows readily from the expression for $P'(t)$ derived in Lemma 3.1, and the bound on $K'(t)$ uses that K is an explicit, analytic function of P; see Appendix C.2 for a full proof. Intuitively, the self-bounding property states that if P does not escape to infinity, then

$P'(t)$ cannot escape either. Since the rate of growth for $P(t)$ is in turn bounded by $P'(t)$, this suggests that there is an interval for t on which P and P' self-regulate one another, ensuring a well-behaved solution.

3.1. Norm Bounds for Self-Bounding ODEs

Informally, the self-bounding ODE method argues that if a vector-valued ODE $y(t)$ satisfies a self-bounding property of the form $\|y'(t)\| \le g(\|y(t)\|)$ wherever it is defined, then the ODE can be compared to a scalar ODE $z'(t) \approx g(z(t))$ with initial condition $z(0) \approx \|y(0)\|$. Specifically, it admits a solution $y(t)$ which is well-defined on an interval roughly as large as that of $z(t)$. We develop the method in a general setting where $y(t)$ (when defined) is the zero of a sufficiently regular function.

Definition 3.2 (Valid Implicit Function)**.** A function $F(\cdot,\cdot) : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}^d$ is a called a *valid implicit function* with domain $\mathcal{U} \subseteq \mathbb{R}^d$ if F is continuously differentiable, and if for any continuously differentiable curve $x(t)$ and any $t \in [0,1]$, either (a) $F(x(t), y) = 0$ has no solution $y \in \mathcal{U}$, or (b) it has a unique solution $y(t) \in \mathcal{U}$, and there exists an open interval around t and a $\mathcal{C}^1$ curve $y(u)$ defined on this interval for which $F(x(u), y(u)) = 0$.

This setting captures as a special case the characterization of $P(t)$ from Lemma 3.1. As a consequence of the lemma, we may take $F = \mathcal{F}_{\mathsf{DARE}}$, where, identifying $\mathbb{S}^{d_{\mathbf{x}}}$ as a $\binom{d_{\mathbf{x}}+1}{2}$-dimensional euclidean space, $\mathcal{F}_{\mathsf{DARE}} : (\mathbb{R}^{d_{\mathbf{x}}^2} \times \mathbb{R}^{d_{\mathbf{x}} d_{\mathbf{u}}}) \times \mathbb{S}^{d_{\mathbf{x}}} \to \mathbb{S}^{d_{\mathbf{x}}}$ is the function whose zero-solution defines the DARE: $\mathcal{F}_{\mathsf{DARE}}((A,B),P) := A^\top PA - P - A^\top PB(R_{\mathbf{u}} + B^\top PB)^{-1}B^\top PA + R_{\mathbf{x}}$. Then $\mathcal{F}_{\mathsf{DARE}}$ is a valid implicit function with unique solutions in the set of positive-definite matrices $\mathcal{U} := \mathbb{S}^{d_{\mathbf{x}}}_{++}$. To proceed, we introduce our self-bounding condition.

Definition 3.3 (Self-bounding)**.** Let $g : \mathbb{R} \to \mathbb{R}_{\ge 0}$ be non-negative and non-decreasing, let F be a valid implicit function with domain $\mathcal{U}$, and let $\|\cdot\|$ be a norm. For a continuously differentiable curve $x(t)$ defined on $[0,1]$, we say that F is $(g, \|\cdot\|)$-*self bounded* on $x(t)$ if $F(x(0), y) = 0$ has a solution $y \in \mathcal{U}$ and $\|y'(t)\| \le g(\|y\|)$ for all $t \in [0,1]$ for which $F(x(t), y)\ y \in \mathcal{U}$. We call the tuple $(F, \mathcal{U}, g, \|\cdot\|, x(\cdot))$ a *self-bounding* tuple.

Lemma 3.2 shows that $\mathcal{F}_{\mathsf{DARE}}$ is $(g, \|\cdot\|_{\mathrm{op}})$-self bounding on the curve the $(A(t), B(t))$ with $g(z) = cz^3$ for $c \propto \epsilon_{\mathrm{op}}$. For functions $g(z)$ with this form we have the following general bound on $\|y(t)\|$.

Corollary 3. *Let $(F, \mathcal{U}, g, \|\cdot\|, x(\cdot))$ be a self-bounding tuple, where $g(z) = cz^p$ for $c > 0$ and $p > 1$. Then, if $\alpha := c(p-1)\|y(0)\|^{p-1} < 1$, there exists a unique continuously differentiable function $y(t) \in \mathcal{U}$ defined on $[0,1]$ which satisfies $F(x(t), y(t)) = 0$, and this solution satisfies $\forall t \in [0,1]$, $\|y(t)\| \le (1-\alpha)^{-1/(p-1)}\|y(0)\|$, and $\|y'(t)\| \le c(1-\alpha)^{-p/(p-1)}\|y(0)\|^p$.*

Corollary 3 is a consequence of a similar result for general functions g (Theorem 13, in Appendix D). The condition on the parameter α directly arises from the requirement that the scalar ODE $w'(u) = cw(u)^3$ has a solution on $[0,1]$.

Finishing the Proof of Proposition 4 Finally, we use Corollary 3 to conclude the proof of Proposition 4.

Proof of Proposition 4. Lemma 3.2 states that for any $t \in [0,1]$ for which $(A(t), B(t))$ is stabilizable (i.e., $\mathcal{F}_{\mathsf{DARE}}([A(t), B(t)], \cdot)$ has a solution), we have the bound

$$\|P'(t)\|_{\mathrm{op}} \le 4\|P(t)\|_{\mathrm{op}}^3 \epsilon_{\mathrm{op}}.$$

Applying Corollary 3 with $p = 2$ and $c = 4\epsilon_{\mathrm{op}}$, we see that if $\alpha := 8\epsilon_{\mathrm{op}}\|P_\star\|_{\mathrm{op}}^2 < 1$, then $P(t)$ is continuously differentiable on the interval $[0,1]$ and $\forall t \in [0,1]$, $\|P(t)\|_{\mathrm{op}} \le \|P_\star\|_{\mathrm{op}}/\sqrt{1-\alpha}$. By Lemma 3.2, $K(t)$ is well defined as well, and satisfies $\max_{t\in[0,1]} \|K'(t)\|_\circ \le 7\epsilon_\circ \max_{t\in[0,1]} \|P(t)\|_{\mathrm{op}}^{7/2} \le (1-\alpha)^{-7/4}\|P_\star\|_{\mathrm{op}}$. The desired bound on $\|K_\infty(A_\star, B_\star) - K_\infty(\widehat{A}, \widehat{B})\|_\circ$ follows from the mean value theorem. □

4. Concluding Remarks

We have established that the asymptotically optimal regret for the online LQR problem is $\widetilde{\Theta}(\sqrt{d_{\mathbf{u}}^2 d_{\mathbf{x}} T})$, and that this rate is attained by ε-greedy exploration. We are hopeful that the our new analysis techniques, especially our perturbation bounds, will find broader use within the non-asymptotic theory of control and beyond. Going forward our work raises a number of interesting conceptual questions. Are there broader classes of "easy" reinforcement learning problems beyond LQR for which naive exploration attains optimal sample complexity, or is LQR a fluke? Conversely, is there a more demanding (eg, robust) version of the LQR problem for which more sophisticated exploration techniques such as robust synthesis (Dean et al., 2018) or optimism in the face of uncertainty (Abbasi-Yadkori & Szepesvári, 2011; Cohen et al., 2019) are required to attain optimal regret? On the purely technical side, recall that while our upper and lower bound match in terms of dependence on $d_{\mathbf{u}}$, $d_{\mathbf{x}}$, and T, they differ in their polynomial dependence on $\|P_\star\|_{\mathrm{op}}$. Does closing this gap require new algorithmic techniques, or will a better analysis suffice?

Acknowledgements Max Simchowitz is generously supported by an Open Philanthropy graduate student fellowship. Dylan Foster acknowledges the support of NSF TRIPODS award #1740751.

References

Abbasi-Yadkori, Y. and Szepesvári, C. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 1–26, 2011.

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Abeille, M. and Lazaric, A. Thompson sampling for linear-quadratic control problems. In *Artificial Intelligence and Statistics*, pp. 1246–1254, 2017.

Abeille, M. and Lazaric, A. Improved regret bounds for thompson sampling in linear quadratic control problems. In *International Conference on Machine Learning*, pp. 1–9, 2018.

Adamczak, R. A note on the Hanson-Wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20, 2015.

Agarwal, N., Bullins, B., Hazan, E., Kakade, S., and Singh, K. Online control with adversarial disturbances. In *International Conference on Machine Learning*, pp. 111–119, 2019a.

Agarwal, N., Hazan, E., and Singh, K. Logarithmic regret for online control. In *Advances in Neural Information Processing Systems 32*, pp. 10175–10184. 2019b.

Arias-Castro, E., Candes, E. J., and Davenport, M. A. On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory*, 59(1):472–481, 2012.

Assouad, P. Deux remarques sur l'estimation. *Comptes rendus des séances de l'Académie des sciences. Série 1, Mathématique*, 296(23):1021–1024, 1983.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 263–272. JMLR. org, 2017.

Bertsekas, D. P. *Dynamic Programming and Optimal Control, Vol. I*. Athena Scientific, 2005.

Bof, N., Carli, R., and Schenato, L. Lyapunov theory for discrete time systems. *arXiv preprint arXiv:1809.05289*, 2018.

Boyd, S. Lecture 13: Linear quadratic Lyapunov theory. *EE363 Course Notes, Stanford University*, 2008.

Cohen, A., Hasidim, A., Koren, T., Lazic, N., Mansour, Y., and Talwar, K. Online linear quadratic control. In *International Conference on Machine Learning*, pp. 1028–1037, 2018.

Cohen, A., Koren, T., and Mansour, Y. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pp. 1300–1309, 2019.

Dann, C. and Brunskill, E. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 2818–2826, 2015.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, pp. 1–47.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pp. 4188–4197, 2018.

Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. Input perturbations for adaptive regulation and learning. *arXiv preprint arXiv:1811.04258*, 2018a.

Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. On optimality of adaptive linear-quadratic regulators. *arXiv preprint arXiv:1806.10749*, 2018b.

Fazel, M., Ge, R., Kakade, S., and Mesbahi, M. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pp. 1466–1475, 2018.

Hazan, E., Agarwal, A., and Kale, S. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

Hazan, E., Singh, K., and Zhang, C. Learning linear dynamical systems via spectral filtering. In *Advances in Neural Information Processing Systems*, pp. 6702–6712, 2017.

Hazan, E., Lee, H., Singh, K., Zhang, C., and Zhang, Y. Spectral filtering for general linear dynamical systems. In *Advances in Neural Information Processing Systems*, pp. 4634–4643, 2018.

Hsu, D., Kakade, S., Zhang, T., et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1704–1713. JMLR. org, 2017.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. *Conference on Learning Theory (COLT)*, 2020.

Kakade, S., Kearns, M. J., and Langford, J. Exploration in metric state spaces. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 306–312, 2003.

Kearns, M. J., Mansour, Y., and Ng, A. Y. Approximate planning in large POMDPs via reusable trajectories. In *Advances in Neural Information Processing Systems*, pp. 1001–1007, 2000.

Langford, J. and Zhang, T. The epoch-greedy algorithm for contextual multi-armed bandits. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pp. 817–824. Citeseer, 2007.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Lincoln, B. and Rantzer, A. Relaxing dynamic programming. *IEEE Transactions on Automatic Control*, 51(8): 1249–1260, 2006.

Mania, H., Tu, S., and Recht, B. Certainty equivalence is efficient for linear quadratic control. In *Advances in Neural Information Processing Systems*, pp. 10154–10164, 2019.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529, 2015.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (May):815–857, 2008.

Ouyang, Y., Gagrani, M., and Jain, R. Control of unknown linear systems with Thompson sampling. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1198–1205. IEEE, 2017.

Polderman, J. W. On the necessity of identifying the true parameter in adaptive LQ control. *Systems & control letters*, 8(2):87–91, 1986.

Polderman, J. W. Adaptive LQ control: Conflict between identification and control. *Linear algebra and its applications*, 122:219–244, 1989.

Rakhlin, A. and Sridharan, K. Online nonparametric regression. In *Conference on Learning Theory*, 2014.

Ran, A. and Vreugdenhil, R. Existence and comparison theorems for algebraic riccati equations for continuous-and discrete-time systems. *Linear Algebra and its applications*, 99:63–83, 1988.

Rudelson, M. and Vershynin, R. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

Sarkar, T. and Rakhlin, A. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pp. 5610–5618, 2019.

Sarkar, T., Rakhlin, A., and Dahleh, M. A. Finite-time system identification for partially observed LTI systems of unknown order. *arXiv preprint arXiv:1902.01848*, 2019.

Shamir, O. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*, pp. 3–24, 2013.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

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

Simchowitz, M., Boczar, R., and Recht, B. Learning linear dynamical systems with semi-parametric least squares. In *Conference on Learning Theory*, pp. 2714–2802, 2019.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. 2018.

Tilli, P. Singular values and eigenvalues of non-Hermitian block Toeplitz matrices. *Linear Algebra and its Applications*, 272(1-3):59–89, 1998.

Tu, S. and Recht, B. Least-squares temporal difference learning for the linear quadratic regulator. In *International Conference on Machine Learning*, pp. 5005–5014, 2018.

Vovk, V. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

Yu, B. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pp. 423–435. Springer, 1997.