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

Rate-matching the regret lower-bound in the linear quadratic regulator with unknown dynamics

Feicheng Wang and Lucas Janson

***Abstract*— The theory of adaptive learning-based control currently suffers from a mismatch between its empirical performance and the theoretical characterization of its performance, with consequences for, e.g., the understanding of sample efficiency, safety, and robustness. The linear quadratic regulator with unknown dynamics is a fundamental adaptive control setting with significant structure in its dynamics and cost function, yet even in this setting the ratio between the best-known regret or estimation error upper bounds and their corresponding best-known lower bounds is unbounded due to polylogarithmic factors in T. This gap has not been closed in any of the many papers theoretically studying the linear quadratic regulator with unknown dynamics, and indeed similar gaps have plagued other areas of theoretical online learning such as reinforcement learning. The contribution of this paper is to close that gap by establishing a novel regret upper-bound of $O_p(\sqrt{T})$, and simultaneously establishes an estimation error bound on the dynamics of $O_p(T^{-1/4})$. The two keys to our improved proof technique are (1) a more precise upper- and lower-bound on the system Gram matrix by establishing exact rates of eigenvalues from different sub-spaces and (2) a self-bounding argument for the expected estimation error of the optimal controller. Our technique may shed light on removing polylogarithmic factors in other adaptive learning problems.**

I. INTRODUCTION

We have witnessed an increasing drive to apply adaptive learning-based control in real-world data-driven systems such as self-driving cars [1] and automatic robots [2]. Yet real-world deployment comes with increased risks and costs, and as such has been hindered by the field's limited understanding of the gap between theoretical bounds and the empirical performance of adaptive control. One line of attack for this problem is to deepen our understanding of relatively simple yet fundamental systems such as the linear quadratic regulator (LQR) with unknown dynamics.

A. Problem statement

In the LQR problem, the system dynamics are represented by a linear state-space model starting from $t = 0$:

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

where $x_t \in \mathbb{R}^n$ represents the state of the system at time t and starts at some initial state x_0, $u_t \in \mathbb{R}^d$ represents the action or control applied at time t, $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2 I_n)$ is the system noise, and $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times d}$ are matrices determining the system's linear dynamics. The goal is to find an algorithm U that, at each time t, outputs a control $u_t = U(H_t)$ that is computed using the entire thus-far-observed history of the system $H_t = \{x_t, u_{t-1}, x_{t-1}, \ldots, u_1, x_1, u_0, x_0\}$ to maximize the system's function while minimizing control effort. The cost of the LQR problem up to a given finite time T is quadratic:

$$\mathcal{J}(U, T) = \sum_{t=1}^{T} \left(x_t^\top Q x_t + u_t^\top R u_t\right) \tag{2}$$

for some known positive definite matrices $Q \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{d\times d}$. When the system dynamics A and B are also known and $T \to \infty$, the cost-minimizing algorithm is known: $u_t^* = U^*(H_t) = Kx_t$, where $K \in \mathbb{R}^{d\times n}$ is the efficiently-computable solution to a system of equations that only depend on A, B, Q, and R. Like the Gaussian linear model in supervised learning, the aforementioned LQR problem is foundational to control theory because it is conceptually simple yet it describes some real-world systems remarkably well. In fact, many systems are close to linear over their normal range of operation, and linearity is an important factor in system design [3].

In this paper we consider the case when the system dynamics A and B are unknown. Intuitively, one might hope that after enough time observing a system controlled by almost any algorithm, one should be able to estimate A and B (and hence K) fairly well and thus be able to apply an algorithm quite close to U^*. Indeed the key challenge in LQR with unknown dynamics, as in any adaptive control or reinforcement learning problem, is to trade off exploration (actions that help estimate A and B) with exploitation (actions that minimize cost). We will quantify the cost of an algorithm by its regret, which is the difference in cost achieved by the algorithm and that achieved by the oracle optimal controller U^*:

$$\mathcal{R}(U, T) = \mathcal{J}(U, T) - \mathcal{J}(U^*, T).$$

A key goal of the theoretical study of adaptive control is to accurately characterize the performance of controlled systems. Unfortunately, the best-known upper-bound for the regret of LQR with unknown dynamics is $O_p(\sqrt{T}\,\mathrm{polylog}(T))$, which contains a polylogarithmic factor of T that is not present in the best-known lower-bound of $\Omega_p(\sqrt{T})$. This means that the ratio of the best-known upper-bound and the best-known lower-bound, which we would like to be as small as possible in order to claim a tight characterization of realistic linear adaptive control performance, is currently unbounded in T.

The authors are grateful for partial support from NSF CBET-2112085

F. Wang is with the Department of Statistics, Harvard University, MA 02138, USA f_wang@g.harvard.edu

L. Janson is with the Department of Statistics, Harvard University, MA 02138, USA ljanson@fas.harvard.edu

B. Technical contribution

This paper for the first time proves that this ratio is bounded by establishing a regret upper bound of $O_p(\sqrt{T})$, where the improvement comes from a new bound on the system Gram matrix combined with a novel self-bounding argument for the expected estimation error. As part of our proof, we show that the algorithm that achieves our optimal rate of regret also produces data that can be used for system identification (estimation of A and B) at a rate of $\left\|\hat{A}-A\right\|_2 = \left\|\hat{B}-B\right\|_2 = O_p(T^{-1/4})$, which is also tighter than the best-known bounds of $O_p(T^{-1/4}\operatorname{polylog}(T))$ for data from an algorithm achieving $O_p(\sqrt{T}\operatorname{polylog}(T))$ regret.

The key to removing the extra $\operatorname{polylog}(T)$ terms is to prove a high-probability upper- and lower-bound ($\bar{\Gamma}$ and $\underline{\Gamma}$) on the system Gram matrix, so that the norm of their ratio $\left\|\bar{\Gamma}\underline{\Gamma}^{-1}\right\|$ is bounded by a constant. This idea is potentially applicable to other adaptive learning problems with extra logarithmic terms in their high-probability estimation bounds.

The following two key insights comprise the main technical innovation of this paper.

1) The first step in tightly bounding the Gram matrix is to separately bound the eigenvalues for the sub-spaces spanned by $\begin{bmatrix} I_n \\ K \end{bmatrix}$ and $\begin{bmatrix} -K^\top \\ I_d \end{bmatrix}$. In the sub-space spanned by $\begin{bmatrix} I_n \\ K \end{bmatrix}$, the eigenvalues of both $\bar{\Gamma}$ and $\underline{\Gamma}$ are $\Theta(T)$. In the sub-space spanned by $\begin{bmatrix} -K^\top \\ I_d \end{bmatrix}$, the eigenvalues of $\underline{\Gamma}$ are $\Theta(\sqrt{T})$, but the eigenvalues of $\bar{\Gamma}$ have a complicated expression of unclear order. See Lemma 3 for the detailed expression and formal result.
2) The second step is to establish a self-bounding argument in Theorem 6 which, combined with Lemma 4, proves the exact order of the complicated expression is $\Theta(\sqrt{T})$. Our self-bounding argument, which establishes a connection between high-probability tail bounds and expectations, is potentially applicable to proofs in other contexts.

These two steps establish that the eigenvalues in all directions of $\bar{\Gamma}$ and $\underline{\Gamma}$ match, thereby bounding $\left\|\bar{\Gamma}\underline{\Gamma}^{-1}\right\|$.

C. Related works

Many works have studied optimal rates of regret in online and reinforcement learning problems. In bandits, matching upper- and lower-bounds have been found as $\Theta_p(\log(T))$ for the distribution-dependent regret [4]–[6] and $\Theta_p(\sqrt{T})$ for the distribution-free regret [6]–[8] .

For Markov decision processes (MDPs), most work has considered finite state and action spaces. In this setting, a matching upper- and lower-bound of $\Theta_p(\log(T))$ is known for the distribution-dependent regret [9]–[11], while the best-known upper-bound of $O_p(\sqrt{T}\operatorname{polylog}(T))$ for the distribution-free regret [12]–[16] has a polylogarithmic gap with the best-known lower-bound of $\Omega_p(\sqrt{T})$ [12], [13].

The LQR system is an MDP with continuous state and action spaces, and has received increasing interest recently. For the LQR system with unknown dynamics, [17] proved a $\Omega_p(\sqrt{T})$ lower-bound for the regret along with an upper-bound of $O_p(\sqrt{T\log(\frac{1}{\delta})})$ with probability $1-\delta$ under the condition $\delta < 1/T$, so that the upper-bound contains an implicit additional $\log^{1/2}(T)$ term. Other $O_p(\sqrt{T}\operatorname{polylog}(T))$ regret upper-bounds for LQR with unknown dynamics have been established elsewhere [18]–[24] and some work has tightened these bounds when the dynamics are partially known [25]–[27], but to the best of our knowledge, no existing work has matched the $\Omega_p(\sqrt{T})$ lower-bound in the case of unknown dynamics until the present paper. Our proof borrows many insightful results and ideas from a number of these prior works, especially [17], [24], [28], [30].

D. Algorithm and assumptions

Throughout the paper, we make only one assumption on the true system parameters:

Assumption 1 (Stability)**.** *Assume the system is stabilizable, i.e., there exists K_0 such that the spectral radius (maximum absolute eigenvalue) of $A+BK_0$ is strictly less than 1.*

Under Assumption 1, it is well known that there is a unique optimal controller $u_t = Kx_t$ [31] which can be computed from A and B, where

$$K = -(R+B^\top PB)^{-1}B^\top PA \tag{3}$$

and P is the unique positive definite solution to the discrete algebraic Riccati equation (DARE):

$$P = A^\top PA - A^\top PB(R+B^\top PB)^{-1}B^\top PA + Q. \tag{4}$$

In this paper we will consider the same algorithm as in [24], reproduced here as Algorithm 1, which is a noisy certainty equivalent control algorithm [17], [18].

In particular, at every round t, we generate an estimate $\hat{K}_t$ for K, and then apply control $u_t = \hat{K}_t x_t + \eta_t$ as a substitute of the optimal unknown control $u_t = Kx_t$, where $\eta_t \sim \mathcal{N}(0, t^{-1/2}I_d)$ is a noise term whose variance shrinks at a carefully chosen rate in t so as to rate-optimally trade off exploration and exploitation. Note that Algorithm 1 is step-wise and online, i.e., it does not rely on independent restarts or episodes of any kind and does not depend on the time horizon T. The two things it does rely on, which are standard in the literature (see, e.g., [32]), are the knowledge of a stabilizing controller K_0 and an upper-bound C_K on the spectral norm of the optimal controller K; C_x and σ_η are also inputs but can take any positive numbers.

E. Notation

Throughout our proofs, we use $X \lesssim Y$ (resp. $X \gtrsim Y$) as shorthand for the inequality $X \le CY$ (resp. $X \ge CY$) for some constant C. $X \eqsim Y$ means both $X \lesssim Y$ and $X \gtrsim Y$. We will almost always establish such relations between quantities that (at least may) depend on T and show that they hold with at least some stated probability $1-\delta$; in such cases, we will always make all dependence on both T and δ explicit, i.e., the hidden constant(s) C will never depend on T or δ, though they may depend on any other parameters of

Algorithm 1 Stepwise Noisy Certainty Equivalent Control

INPUT: Initial state x_0, stabilizing control matrix K_0, scalars $C_x > 0$, $C_K > \|K\|$, $\sigma_\eta > 0$.

1: Let $u_0 = K_0 x_0 + \eta_0$ and $u_1 = K_0 x_1 + \eta_1$, with $\eta_0, \eta_1 \overset{iid}{\sim} \mathcal{N}(0, \sigma_\eta^2 I_d)$.

2: **for** $t = 2, 3, \ldots$ **do**

3: Compute

$$(\hat{A}_{t-1}, \hat{B}_{t-1}) \in \operatorname*{argmin}_{(A', B')} \sum_{k=0}^{t-2} \|x_{k+1} - A' x_k - B' u_k\|^2 \tag{5}$$

and if stabilizable, plug them into the DARE (Eqs. (3) and (4)) to compute $\hat{K}_t$, otherwise set $\hat{K}_t = K_0$.

4: If $\|x_t\| \gtrsim C_x \log(t)$ or $\left\|\hat{K}_t\right\| \gtrsim C_K$, reset $\hat{K}_t = K_0$.

5: Let

$$u_t = \hat{K}_t x_t + \eta_t, \qquad \eta_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\eta^2 t^{-1/2} I_n) \tag{6}$$

6: **end**

the system or algorithm, including A, B, Q, R, σ_ϵ^2, σ_η^2, K_0, C_x, C_K.

F. Outline

For the remainder of this paper, we will primarily focus on establishing a new $O_p\left(T^{-1/4}\right)$ bound on the estimation error of $\hat{A}_t$, $\hat{B}_t$, and $\hat{K}_t$ from Algorithm 1, with an emphasis on our two main technical contributions Lemma 3 and Theorem 6. After that, we will leverage this tighter estimation error bound to establish our rate-matching $O_p\left(\sqrt{T}\right)$ bound on the regret of Algorithm 1.

II. Main results

Our $O_p\left(T^{-1/4}\right)$ bound on the estimation error starts with a key result from [28], which relates the estimation error to the system Gram matrix via a lower- and upper-bound for it. The rest of the proof is primarily comprised of two parts. In the first part, we prove a more precise upper- and lower-bound on the system Gram matrix so that the two bounds are almost of the same order, which is crucial in removing the polylog(T) in the estimation error bound. In the second part, we take the estimation error bound from plugging in the Gram matrix bounds from the first part and transform it into a self-bounding argument for the expected estimation error of the estimated dynamics that yields the $O_p(T^{-1/4})$ final rate for the estimation error.

To streamline notation, define $z_t = \begin{bmatrix} x_t \\ u_t \end{bmatrix}$ and $\Theta = [A, B]$, and correspondingly define $\hat{\Theta}_t = [\hat{A}_t, \hat{B}_t]$. Then by Theorem 2.4 of [28], given a fixed $\delta \in (0, 1)$, $T \in \mathbb{N}$ and $0 \preceq \underline{\Gamma} \preceq \bar{\Gamma} \in \mathbb{R}^{(n+d)\times(n+d)}$ such that $\mathbb{P}\left[\sum_{t=0}^{T-1} z_t z_t^\top \succeq T\underline{\Gamma}\right] \geq 1 - \delta$ and $\mathbb{P}[\sum_{t=0}^{T-1} z_t z_t^\top \preceq T\bar{\Gamma}] \geq 1 - \delta$, when $T \gtrsim \log\left(\frac{1}{\delta}\right) + 1 + \log\det(\bar{\Gamma}\underline{\Gamma}^{-1})$, $\hat{\Theta}_T$ satisfies:

$$\mathbb{P}\left[\left\|\hat{\Theta}_T - \Theta\right\| \gtrsim \sqrt{\frac{1 + \log\det\bar{\Gamma}\underline{\Gamma}^{-1} + \log\left(\frac{1}{\delta}\right)}{T\lambda_{\min}(\underline{\Gamma})}}\right] \leq \delta. \tag{7}$$

Here, $\lambda_{\min}(\underline{\Gamma})$ is the minimum eigenvalue of $\underline{\Gamma}$. Similar upper-bounds to those that already exist in the literature (which contain extra polylog(T) terms compared to the best know lower-bound) can be achieved by taking $\underline{\Gamma} \eqsim T^{-1/2} I_{n+d}$ and $\bar{\Gamma} \eqsim \log^2(T) I_{n+d}$, and we restate this result here (due to the space limit, we defer the proof of Lemma 2 to our online report [29]).

Lemma 2 (Estimation error bound with polylog(T) term)**.** *Algorithm 1 applied to a system described by Eq.* (1) *under Assumption 1 satisfies, when* $0 < \delta < 1/2$*, for any* $T \gtrsim \log(1/\delta)$*,*

$$\mathbb{P}\left[\left\|\hat{\Theta}_T - \Theta\right\| \gtrsim T^{-1/4}\sqrt{\left(\log T + \log\left(\frac{1}{\delta}\right)\right)}\right] \leq \delta. \tag{8}$$

Improving this $O_p\left(T^{-1/4}\log^{1/2}(T)\right)$ bound to the desired $O_p\left(T^{-1/4}\right)$ requires tighter lower- and upper-bounds $\underline{\Gamma}$ and $\bar{\Gamma}$ for $\sum_{t=0}^{T-1} z_t z_t^\top$, which will be facilitated by the following key lemma.

Lemma 3. *Algorithm 1 applied to a system described by Eq.* (1) *under Assumption 1 satisfies, for any* $0 < \delta < 1/2$ *and* $T \gtrsim \log^3(1/\delta)$*, with probability at least* $1 - \delta$*:*

$$\begin{aligned} T\underline{\Gamma} &:= \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} T^{1/2} \begin{bmatrix} -K^\top \\ I_d \end{bmatrix}^\top \\ &\precsim \sum_{t=0}^{T-1} z_t z_t^\top \\ &\precsim \left(\frac{1}{\delta}\begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + \right. \\ &\quad \left. \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} \lambda_{\max}\left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top\right) \begin{bmatrix} -K^\top \\ I_d \end{bmatrix}^\top\right) := T\bar{\Gamma}, \end{aligned} \tag{9}$$

where $\Delta_t := (\hat{K}_t - K)x_t + \eta_t$.

Lemma 3 contains our first main technical contribution. Aside from $\lambda_{\max}\left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top\right)$, we established an almost exact eigendecomposition of the Gram matrix $\sum_{t=0}^{T-1} z_t z_t^\top$ into different sub-spaces $\begin{bmatrix} I_n \\ K \end{bmatrix}$ and $\begin{bmatrix} -K^\top \\ I_d \end{bmatrix}$ with explicit eigenvalue bounds of $\Theta(T)$ and $\Theta(\sqrt{T})$, respectively.

Proof. (sketch) Due to space limitations, we defer the detailed proof of Lemma 3 to our online report [29] and only introduce the proof outline here. $G_T := \sum_{t=0}^{T-1} z_t z_t^\top$

can be represented as a summation of two parts:

$$G_T = \sum_{t=0}^{T-1} z_t z_t^\top = \begin{bmatrix} I \\ K \end{bmatrix} \sum_{t=0}^{T-1} x_t x_t^\top \begin{bmatrix} I \\ K \end{bmatrix}^\top + \sum_{t=0}^{T-1} \begin{bmatrix} 0_n & x_t \Delta_t^\top \\ \Delta_t x_t^\top & \Delta_t \Delta_t^\top + K x_t \Delta_t^\top + \Delta_t x_t^\top K^\top \end{bmatrix}. \tag{10}$$

Consider the dominating part $\begin{bmatrix} I \\ K \end{bmatrix} \sum_{t=0}^{T-1} x_t x_t^\top \begin{bmatrix} I \\ K \end{bmatrix}^\top$ (whose smallest eigenvalue scales with T) and the remainder part $\sum_{t=0}^{T-1} \begin{bmatrix} 0_n & x_t \Delta_t^\top \\ \Delta_t x_t^\top & \Delta_t \Delta_t^\top + K x_t \Delta_t^\top + \Delta_t x_t^\top K^\top \end{bmatrix}$ separately. In our online report [29], we show that with probability at least $1-\delta$:

$$\begin{bmatrix} I \\ K \end{bmatrix} T \begin{bmatrix} I \\ K \end{bmatrix}^\top \preceq \begin{bmatrix} I \\ K \end{bmatrix} \sum_{t=0}^{T-1} x_t x_t^\top \begin{bmatrix} I \\ K \end{bmatrix}^\top \preceq 1/\delta \begin{bmatrix} I \\ K \end{bmatrix} T \begin{bmatrix} I \\ K \end{bmatrix}^\top \tag{11}$$

These bounds reflect the intuition that x_t should converge to a stationary distribution, making each of the summands $x_t x_t^\top$ of constant order.

a) Lower bound: Eq. (11) provides a partial lower bound for G_T: with probability at least $1-\delta$,

$$G_T \succeq \begin{bmatrix} I \\ K \end{bmatrix} \sum_{t=0}^{T-1} x_t x_t^\top \begin{bmatrix} I \\ K \end{bmatrix}^\top \succsim \begin{bmatrix} I \\ K \end{bmatrix} T \begin{bmatrix} I \\ K \end{bmatrix}^\top. \tag{12}$$

This part only covers the subspace spanned by $\begin{bmatrix} I \\ K \end{bmatrix}$; we still need to consider a general bound for the whole matrix $G_T = \sum_{t=0}^{T-1} z_t z_t^\top$. Noting that the magnitude of $z_t = (x_t, u_t)^\top$ is lower-bounded by that of the innovation term $(\varepsilon_{t-1}, \eta_t)^\top$ with standard error at least $\Omega(t^{-1/2})$ (in all directions), Lemma 34 of [24] gives a high probability lower-bound $G_T \succsim T^{1/2} I_{n+d}$. Combining this and Eq. (12), with high probability we have:

$$\begin{aligned} G_T + G_T &\succsim \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + T^{1/2} I_{n+d} \\ &\succsim \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} T^{1/2} \begin{bmatrix} -K^\top \\ I_d \end{bmatrix}^\top. \end{aligned}$$

b) Upper bound: The argument for our upper-bound divides $\mathbb{R}^{n+d}$ into two orthogonal subspaces spanned by the columns of $\begin{bmatrix} I_n \\ K \end{bmatrix}$ and $\begin{bmatrix} -K^\top \\ I_d \end{bmatrix}$, and essentially bounds $\xi^\top G_T \xi$ separately by order T and $\lambda_{\max}\left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top\right)$ for the two subspaces, respectively. In particular, for any ξ_1 in the span of $\begin{bmatrix} I_n \\ K \end{bmatrix}$ and ξ_2 in the span of $\begin{bmatrix} -K^\top \\ I_d \end{bmatrix}$,

$$\begin{aligned} &(\xi_1+\xi_2)^\top G_T (\xi_1+\xi_2) \\ &\quad \le 2\xi_1^\top G_T \xi_1 + 2\xi_2^\top G_T \xi_2 \\ &\quad \lesssim 2\xi_1^\top G_T \xi_1 + 2\left\|\xi_2\right\|^2 \lambda_{\max}\left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top\right) \\ &\quad \lesssim \frac{1}{\delta} T \left\|\xi_1\right\|^2 + \left\|\xi_2\right\|^2 \lambda_{\max}\left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top\right), \end{aligned}$$

where the second inequality follows from Eq. (10) because ξ_2 is orthogonal to $\begin{bmatrix} I \\ K \end{bmatrix} \sum_{t=0}^{T-1} x_t x_t^\top \begin{bmatrix} I \\ K \end{bmatrix}^\top$, and the third inequality follows from $\mathbb{P}\left[G_T \precsim \frac{1}{\delta} T I_{n+d}\right] \ge 1-\delta$. This last expression can in turn be bounded by

$$(\xi_1+\xi_2)^\top \left(\frac{1}{\delta} \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} \lambda_{\max}\left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top\right) \begin{bmatrix} -K^\top \\ I_d \end{bmatrix}^\top \right) (\xi_1+\xi_2),$$

establishing the upper-bound from Eq. (9). □

In Lemma 3, the upper bound $\bar{\Gamma}$ and lower bound $\underline{\Gamma}$ have similar forms. Plugging them into Eq. (7) gives that when $T \gtrsim \log^3(1/\delta)$,

$$\mathbb{P}\left[\left\|\hat{\Theta}_T - \Theta\right\| \gtrsim \sqrt{\frac{1 + \log\left(\lambda_{\max}\left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top\right) T^{-1/2}\right) + \log\left(\frac{1}{\delta}\right)}{T^{1/2}}} \right] \le \delta. \tag{13}$$

The following Lemmas 4 and 5 will connect the key term $\lambda_{\max}\left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top\right)$ in the estimation error bound of Eq. (13) with the estimation error itself, setting up the self-bounding argument that is key to our main estimation error bound in Theorem 6.

Lemma 4. *Algorithm 1 applied to a system described by Eq.* (1) *under Assumption 1 satisfies, for any* $0 < \delta < 1/2$, $T \gtrsim \log^2(1/\delta)$,

$$\mathbb{P}\left\{ \lambda_{\max}\left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top\right) \gtrsim 1/\delta \left(\sum_{t=1}^{T-1} \mathbb{E}\left(t^{1/2} \left\|\hat{K}_t - K\right\|^4 \right) + \log^2(1/\delta) + T^{1/2} \right) \right\} \le 2\delta.$$

Here, $\mathbb{E}(\cdot)$ denotes the expected value function. Due to the space limit, we defer the detailed proof of Lemma 4 to our online report [29]. Lemma 5 below can be obtained by substituting Lemma 4 into Eq. (13).

Lemma 5. *Algorithm 1 applied to a system described by Eq.* (1) *under Assumption 1 satisfies, for any* $0 < \delta < 1/2$ *and* $T \gtrsim \log^3(1/\delta)$,

$$\mathbb{P}\left[T^{1/2} \left\|\hat{\Theta}_T - \Theta\right\|^2 \gtrsim \log\left(T^{-1/2} \left(\sum_{t=1}^{T-1} \mathbb{E}\left(t^{1/2} \left\|\hat{K}_t - K\right\|^4 \right) \right) + 1 \right) + \log\left(\frac{1}{\delta}\right) \right] \le 3\delta. \tag{14}$$

Due to the space limit, we defer the detailed proof of Lemma 5 to our online report [29]. We are now ready to state the main result of this paper:

Theorem 6. *Algorithm 1 applied to a system described by Eq.* (1) *under Assumption 1 satisfies*

$$\left\|\hat{\Theta}_T - \Theta\right\| = O_p(T^{-1/4}) \text{ and } \left\|\hat{K}_T - K\right\| = O_p(T^{-1/4}). \tag{15}$$

Theorem 6 contains our second main technical contribution: we establish a self-bounding argument that connects between high-probability tail bounds and expectations. In particular, we bound $\mathbb{E}\left(T\left\|\hat{\Theta}_T - \Theta\right\|^4\right)$ with an integral of the high probability bound in Eq. (14).

Proof. (sketch) Due to space limitations, we defer the detailed proof of Theorem 6 to our online report [29] and only introduce the proof outline here. Intuitively speaking, we aim to prove

$$\left\|\hat{K}_T - K\right\| \lesssim \left\|\hat{\Theta}_T - \Theta\right\|, \tag{16}$$

and

$$\mathbb{E}\left(T\left\|\hat{\Theta}_T - \Theta\right\|^4\right) \lesssim \log\left(\mathbb{E}\left(T\left\|\hat{K}_T - K\right\|^4\right)\right).$$

The previous two equations imply that $\mathbb{E}\left(T\left\|\hat{K}_T - K\right\|^4\right)$ is at most of constant order, otherwise, it cannot be bounded by the log of itself. By Proposition 4 of [17], Eq. (16) holds as long as $\left\|\hat{\Theta}_T - \Theta\right\| \le \epsilon_0$, where ϵ_0 is some fixed constant determined by the system parameters. We want to focus on the event $1_{\|\hat{\Theta}_T - \Theta\| \le \epsilon_0}$ to transfer $T^{1/2}\left\|\hat{\Theta}_T - \Theta\right\|^2$ to $T^{1/2}\left\|\hat{K}_T - K\right\|^2$.

We can estimate $\mathbb{E}\left(T\left\|\hat{\Theta}_T - \Theta\right\|^4 1_{\|\hat{\Theta}_T - \Theta\| \le \epsilon_0}\right)$ by calculating the integral using the tail bound from Lemma 5, which gives us, when $T \ge T_0$ (T_0 is a large enough constant)

$$\mathbb{E}\left(T\left\|\hat{\Theta}_T - \Theta\right\|^4 1_{\|\hat{\Theta}_T - \Theta\| \le \epsilon_0}\right) \lesssim \left(\log\left(T^{-1/2}\left(\sum_{t=1}^{T-1} t^{-1/2}\mathbb{E}\left(t\left\|\hat{K}_t - K\right\|^4\right)\right)\right) + 1\right)^2 + 1.$$

On the right-hand side, consider the maximum of $\mathbb{E}\left(t\left\|\hat{K}_t - K\right\|^4\right)$ from T_0 to $T_{\max} \ge T$, and bound other terms from 1 to $T_0 - 1$ by constant (remind that Algorithm 1 ensures $\left\|\hat{K}_t\right\| \le C_K$). We have

$$\mathbb{E}\left(T\left\|\hat{\Theta}_T - \Theta\right\|^4 1_{\|\hat{\Theta}_T - \Theta\| \le \epsilon_0}\right) \lesssim \left(\log\left(1 + \max_{T_0 \le s \le T_{\max}} \mathbb{E}\left(s\left\|\hat{K}_s - K\right\|^4\right)\right) + 1\right)^2 + 1$$

By Eq. (16), we can transfer $\left\|\hat{\Theta}_T - \Theta\right\|$ on the left hand side to $\left\|\hat{K}_T - K\right\|$

$$\begin{aligned}&\mathbb{E}\left(T\left\|\hat{\Theta}_T - \Theta\right\|^4 1_{\|\hat{\Theta}_T - \Theta\| \le \epsilon_0}\right)\\ &\gtrsim \mathbb{E}\left(T\left\|\hat{K}_T - K\right\|^4 1_{\|\hat{\Theta}_T - \Theta\| \le \epsilon_0}\right)\\ &\ge \mathbb{E}\left(T\left\|\hat{K}_T - K\right\|^4\right) - 1.\end{aligned}$$

The final inequality holds because by Lemma 2, the probabilty that $\left\|\hat{\Theta}_T - \Theta\right\|^2 > \epsilon_0$ decays exponentially with T. Thus,

$$\mathbb{E}\left(T\left\|\hat{K}_T - K\right\|^4\right) \lesssim \left(\log\left(1 + \max_{T_0 \le s \le T_{\max}} \mathbb{E}\left(s\left\|\hat{K}_s - K\right\|^4\right)\right) + 1.\right)^2 + 1$$

The right hand side is constant. Taking the maximum over T from T_0 to $T_{\max}$ on the left hand side:

$$\max_{T_0 \le s \le T_{\max}} \mathbb{E}\left(s\left\|\hat{K}_s - K\right\|^4\right) \lesssim \left(\log\left(1 + \max_{T_0 \le s \le T_{\max}} \mathbb{E}\left(s\left\|\hat{K}_s - K\right\|^4\right)\right) + 1\right)^2 + 1$$

Thus

$$\max_{T_0 \le s \le T_{\max}} \mathbb{E}\left(s\left\|\hat{K}_s - K\right\|^4\right) \lesssim 1.$$

The hidden constant only depends on T_0, and hence the same inequality holds for <u>any</u> $T_{\max}$:

$$\max_{s \ge T_0} \mathbb{E}\left(s\left\|\hat{K}_s - K\right\|^4\right) \lesssim 1.$$

Plugging this back to Eq. (14) gives that when $T \gtrsim \log^3(1/\delta)$,

$$\mathbb{P}\left[T^{1/2}\left\|\hat{\Theta}_T - \Theta\right\|^2 \gtrsim \log\left(\frac{1}{\delta}\right)\right] \le 3\delta.$$

Thus,

$$\left\|\hat{\Theta}_T - \Theta\right\| = O_p(T^{-1/4}),$$

and $\left\|\hat{K}_T - K\right\| = O_p(T^{-1/4})$ is a direct corollary from Eq. (16). □

Section 2.2 of [18] conjectures that the average regret of the LQR problem $\mathcal{R}(U,T)/T$ is determined by the summation of quadratic terms of estimation error $\left\|\hat{K}_T - K\right\|^2$ and exploration noise $\|\eta_T\|^2$. By the design of Algorithm 1, $\|\eta_T\|^2 = O_p\left(T^{-1/2}\right)$. By Theorem 6, $\left\|\hat{K}_T - K\right\|^2 = O_p\left(T^{-1/2}\right)$. This leads to the second main result of this paper: a regret upper-bound that exactly rate-matches the regret lower-bound of $\Omega\left(\sqrt{T}\right)$ established in [17].

Theorem 7. *Algorithm 1 applied to a system described by Eq.* (1) *under Assumption 1 satisfies*

$$\mathcal{R}(U,T) = O_p\left(\sqrt{T}\right). \tag{17}$$

Due to space limitations, we defer the detailed proof of Theorem 7 to our online report [29]. The key of our proof is to demonstrate $\mathcal{R}(U,T)$ is determined by the larger one of $\sum_{t=1}^{T}\eta_t^\top R\eta_t$ and $\sum_{t=1}^{T}x_t^\top(\hat{K}_t-K)^\top(R+B^\top PB)(\hat{K}_t-K)x_t$. Intuitively speaking, when the exploration (first term) gets larger, the second term gets smaller. Indeed, the order of these two terms strikes a delicate balance at $O_p\left(\sqrt{T}\right)$ with $\eta_t \overset{iid}{\sim} \mathcal{N}(0,\sigma_\eta^2 t^{-1/2}I_n)$.

Since our algorithm (see Algorithm 1) is identical to the one analyzed in [24], readers are encouraged to consult Section 4 of [24] for comprehensive experimental outcomes. In particular, Figures 1b and I.1b in [24] empirically demonstrate that the regret follows an order of $\sqrt{T}$.

III. Discussion

This paper provides progress in understanding the practical performance of adaptive learning-based control by, for the LQR problem with unknown dynamics, proving a regret upper-bound of $O_p(\sqrt{T})$, which is the first to have a bounded ratio with the best-known lower-bound of $\Omega_p(\sqrt{T})$ established in [17]. There are related settings such as non-linear LQR [33] and non-stationary LQR [34] whose best-known regret upper-bounds are $O_p(\sqrt{T}\,\mathrm{polylog}(T))$, and we hope our work can shed light on removing the $\mathrm{polylog}(T)$ terms in these settings as well.

References

[1] Kiran, B. Ravi, et al. "Deep reinforcement learning for autonomous driving: A survey." IEEE Transactions on Intelligent Transportation Systems 23.6 (2021): 4909-4926.

[2] Levine, Sergey, et al. "End-to-end training of deep visuomotor policies." The Journal of Machine Learning Research 17.1 (2016): 1334-1373.

[3] Recht, Benjamin. "A tour of reinforcement learning: The view from continuous control." Annual Review of Control, Robotics, and Autonomous Systems 2 (2019): 253-279.

[4] Lai, Tze Leung, and Herbert Robbins. "Asymptotically efficient adaptive allocation rules." Advances in applied mathematics 6.1 (1985): 4-22.

[5] Auer P, Cesa-Bianchi N, Fischer P. Finite-time analysis of the multiarmed bandit problem[J]. Machine learning, 2002, 47: 235-256.

[6] Garivier A, Hadiji H, Menard P, et al. KL-UCB-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints[J]. 2018.

[7] Li Y, Wang Y, Zhou Y. Nearly minimax-optimal regret for linearly parameterized bandits[C]//Conference on Learning Theory. PMLR, 2019: 2173-2174.

[8] Hajiesmaili M, Talebi M S, Lui J, et al. Adversarial bandits with corruptions: Regret lower bound and no-regret algorithm[J]. Advances in Neural Information Processing Systems, 2020, 33: 19943-19952.

[9] Ok J, Proutiere A, Tranos D. Exploration in structured reinforcement learning[J]. Advances in Neural Information Processing Systems, 2018, 31.

[10] Tirinzoni A, Pirotta M, Lazaric A. A fully problem-dependent regret lower bound for finite-horizon mdps[J]. arXiv preprint arXiv:2106.13013, 2021.

[11] Xu H, Ma T, Du S. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap[C]//Conference on Learning Theory. PMLR, 2021: 4438-4472.

[12] Auer P, Jaksch T, Ortner R. Near-optimal regret bounds for reinforcement learning[J]. Advances in neural information processing systems, 2008, 21.

[13] Azar M G, Osband I, Munos R. Minimax regret bounds for reinforcement learning[C]//International Conference on Machine Learning. PMLR, 2017: 263-272.

[14] Agrawal S, Jia R. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds[J]. Advances in Neural Information Processing Systems, 2017, 30.

[15] Simchowitz M, Jamieson K G. Non-asymptotic gap-dependent regret bounds for tabular mdps[J]. Advances in Neural Information Processing Systems, 2019, 32.

[16] Xiong Z, Shen R, Du S S. Randomized exploration is near-optimal for tabular mdp[J]. arXiv preprint arXiv:2102.09703, 2021.

[17] Simchowitz M, Foster D. Naive exploration is optimal for online lqr[C]//International Conference on Machine Learning. PMLR, 2020: 8937-8948.

[18] Mania H, Tu S, Recht B. Certainty equivalence is efficient for linear quadratic control[J]. Advances in Neural Information Processing Systems, 2019, 32.

[19] Abbasi-Yadkori Y, Szepesvári C. Regret bounds for the adaptive control of linear quadratic systems[C]//Proceedings of the 24th Annual Conference on Learning Theory. JMLR Workshop and Conference Proceedings, 2011: 1-26.

[20] Faradonbeh M K S, Tewari A, Michailidis G. Finite time analysis of optimal adaptive policies for linear-quadratic systems[J]. arXiv preprint arXiv:1711.07230, 2017.

[21] Cohen A, Koren T, Mansour Y. Learning Linear-Quadratic Regulators Efficiently with only $\sqrt{T}$ Regret[C]//International Conference on Machine Learning. PMLR, 2019: 1300-1309.

[22] Ouyang Y, Gagrani M, Jain R. Learning-based control of unknown linear systems with thompson sampling[J]. arXiv preprint arXiv:1709.04047, 2017.

[23] Abeille M, Lazaric A. Improved regret bounds for thompson sampling in linear quadratic control problems[C]//International Conference on Machine Learning. PMLR, 2018: 1-9.

[24] Wang F, Janson L. Exact asymptotics for linear quadratic adaptive control[J]. The Journal of Machine Learning Research, 2021, 22(1): 12136-12247.

[25] Jedra Y, Proutiere A. Minimal expected regret in linear quadratic control[C]//International Conference on Artificial Intelligence and Statistics. PMLR, 2022: 10234-10321.

[26] Cassel A, Cohen A, Koren T. Logarithmic regret for learning linear quadratic regulators efficiently[C]//International Conference on Machine Learning. PMLR, 2020: 1328-1337.

[27] Foster D, Simchowitz M. Logarithmic regret for adversarial online control[C]//International Conference on Machine Learning. PMLR, 2020: 3211-3221.

[28] Simchowitz M, Mania H, Tu S, et al. Learning without mixing: Towards a sharp analysis of linear system identification[C]//Conference On Learning Theory. PMLR, 2018: 439-473.

[29] Wang, Feicheng, and Lucas Janson. Rate-matching the regret lower-bound in the linear quadratic regulator with unknown dynamics (extended version). [Online]. Available: https://feicheng-wang.github.io/2023cdcLQRregret.pdf

[30] Fazel M, Ge R, Kakade S, et al. Global convergence of policy gradient methods for the linear quadratic regulator[C]//International conference on machine learning. PMLR, 2018: 1467-1476.

[31] Arnold W F, Laub A J. Generalized eigenproblem algorithms and software for algebraic Riccati equations[J]. Proceedings of the IEEE, 1984, 72(12): 1746-1754.

[32] Dean S, Mania H, Matni N, et al. Regret bounds for robust adaptive control of the linear quadratic regulator[J]. Advances in Neural Information Processing Systems, 2018, 31.

[33] Kakade S, Krishnamurthy A, Lowrey K, et al. Information theoretic regret bounds for online nonlinear control[J]. Advances in Neural Information Processing Systems, 2020, 33: 15312-15325.

[34] Luo Y, Gupta V, Kolar M. Dynamic regret minimization for control of non-stationary linear dynamical systems[J]. Proceedings of the ACM on Measurement and Analysis of Computing Systems, 2022, 6(1): 1-72.

[35] Li Y, Das S, Shamma J, et al. Safe adaptive learning-based control for constrained linear quadratic regulators with regret guarantees[J]. arXiv preprint arXiv:2111.00411, 2021.