Regret Analysis of Certainty Equivalence Policies in Continuous-Time Linear-Quadratic Systems

Mohamad Kazem Shirani Faradonbeh

***Abstract*—This work studies theoretical performance guarantees of a ubiquitous reinforcement learning policy for controlling the canonical model of stochastic linear-quadratic system. We show that randomized certainty equivalent policy addresses the exploration-exploitation dilemma for minimizing quadratic costs in linear dynamical systems that evolve according to stochastic differential equations. More precisely, we establish square-root of time regret bounds, indicating that randomized certainty equivalent policy learns optimal control actions fast from a single state trajectory. Further, linear scaling of the regret with the number of parameters is shown. The presented analysis introduces novel and useful technical approaches, and sheds light on fundamental challenges of continuous-time reinforcement learning.**

***Index Terms*—Adaptive control, Reinforcement learning, Optimal policies, Stochastic differential equations, Regret bounds, Learning-based control.**

I. Introduction

Linear state-space models are of the most popular settings for decision-making in continuous-time environments. A canonical problem is to minimize quadratic costs subject to state-dynamics that follow stochastic differential equations driven by control actions and Brownian noise. While applications are enormous [1], [2], [3], [4], little is known about data-driven methods for decision-making under uncertainty. A natural candidate is the randomized certainty equivalent policy that utilize randomizations together with the Certainty Equivalence principle, and will be the subject of this work.

While the existing literature is rich for reinforcement learning policies in *discrete-time* systems [5], [6], [7], [8], [9], [10], [11], [12], [13], study of efficient policies for *continuous-time* systems is immature. Early works focus on asymptotic consistency and propose some control policies with linearly growing regrets [14], [15], [16], [17]. Further, offline reinforcement learning algorithms that rely on multiple state trajectories are considered in some settings [18], [19], [20], [21]. However, performance analysis of *online* policies for learning optimal control actions from a *single* state trajectory are currently unavailable. A fundamental challenge (compared to offline methods) is that an online policy needs to *simultaneously* minimize the cost and estimate the unknown dynamics. These two objectives contradict each other and are subject to the trade-off between the exploration and exploitation; accurate estimation is necessary for good control and for efficiency, while sub-optimal control actions are required in order to have rich data for estimating accurately.

This work establishes that the popular randomized certainty equivalent reinforcement learning policy balances the trade-off between the exploration and exploitation. We present Algorithm 1, which is an episodic randomized certainty equivalent policy for stochastic continuous-time linear systems. We provide its regret analysis indicating efficiency; it learns the optimal control actions fast so that the regret at time T is $\widetilde{\mathcal{O}}\left(T^{1/2}\right)$. Therefore, the per-unit-time sub-optimality gap shrinks with the rate $\widetilde{\mathcal{O}}\left(T^{-1/2}\right)$ as time proceeds. The presented bound is tight and is obtained under minimal technical assumptions.

To obtain the results, we need to address important challenges. First, analysis of estimation error is needed for sample observations with ill-conditioned information matrices. Further, anti-concentration of singular values of random matrices, and full characterization of sub-optimalities in terms of model uncertainties are required. Thus, we develop novel techniques for establishing the rates of identifying the unknown system dynamics matrices based on the data of a single state-input trajectory. Leveraging that together with the effect of diminishing randomizations applied to the parameter estimates, we tightly bound the rates of narrowing down the sub-optimality gap. We also utilize useful results on self-normalized *stochastic integrals*, spectral properties of *random matrices*, and precisely capture the additional cost of sub-optimal control actions. En route, different tools from stochastic control, Ito calculus, and stochastic analysis are used, including Hamilton-Jacobi-Bellman equations, Ito Isometry, and martingale convergence theorems [22], [23], [24].

The outline of this paper is as follows. In Section II, we discuss the problem and the preliminary materials. Section III contains the statement and intuition of the randomized certainty equivalent Algorithm 1, followed by its theoretical and empirical analyses in Section IV. Proofs are provided in the appendices.

II. Problem Formulation: Continuous-Time Reinforcement Learning

In this work, we study reinforcement learning algorithms for an uncertain controlled multidimensional (Ito) stochastic differential equation. The state vector at time t is $x_t \in \mathbb{R}^p$, the control input is $u_t \in \mathbb{R}^q$, and we have

$$\mathrm{d}x_t = \left(A_\star x_t + B_\star u_t\right)\mathrm{d}t + C\mathrm{d}\mathbb{W}_t. \tag{1}$$

In the above dynamics equation, the stochastic disturbance $\{\mathbb{W}_t\}_{t\geq 0}$ is a standard Brownian motion. It starts from the origin in $\mathbb{R}^p$; $\mathbb{W}_0 = 0$. Moreover, $\mathbb{W}_t$ has independent normal increments: for all $0 \leq t_1 \leq t_2 \leq t_3 \leq t_4$, the vectors $\mathbb{W}_{t_2} - \mathbb{W}_{t_1}$ and $\mathbb{W}_{t_4} - \mathbb{W}_{t_3}$ are statistically independent, and

$$\mathbb{W}_{t_2} - \mathbb{W}_{t_1} \sim \boldsymbol{N}\left(0, (t_2 - t_1) I_p\right),$$

where $\boldsymbol{N}(\cdot, \cdot)$ is the multivariate normal distribution. The $p \times p$ matrix C indicates the effect of the Brownian stochastic disturbance on the continuous-time state evolution.

The goal is to design reinforcement learning algorithms to learn to control the dynamical system described in (1). More precisely, the transition matrix $A_\star$, the control input matrix $B_\star$, as well as C, *all are unknown*. We aim to minimize the quadratic cost function averaged over time;

$$\min_{\{u_t\}_{t\geq 0}} \limsup_{T\to\infty} \frac{1}{T}\int_0^T \left(x_t^\top Q x_t + u_t^\top R u_t\right) \mathrm{d}t, \tag{2}$$

where Q, R are positive definite cost matrices of proper dimensions. Above, the minimum is taken over non-anticipating closed-loop reinforcement learning policies, as elaborated below. At every time t, the policy determines u_t according to the information available at the time, which comprise the state observations $\{x_s\}_{0\leq s\leq t}$ and the previously taken actions $\{u_s\}_{0\leq s<t}$. Importantly, the policy faces the fundamental exploration-exploitation dilemma for minimizing the cost, because the dynamics matrices $A_\star, B_\star$ are unknown and need to be learned from the state and action data. The details of this challenge will be discussed in Section III. It is standard to assume that Q, R are known, the rationale being that the decision-maker is aware of the objective.

To proceed, we assume that the system is stabilizable:

Assumption 1. There exists $K_\star \in \mathbb{R}^{q\times p}$, such that real-parts of all eigenvalues of $A_\star + B_\star K_\star$ are negative.

Assumption 1 expresses that by applying $u_t = K_\star x_t$, the system can operate without unbounded growth in the state. Technically, if we apply the above feedback policy, and solve for the state vector in (1), it holds that

$$x_t = e^{(A_\star + B_\star K_\star)t} x_0 + \int_0^t e^{(A_\star + B_\star K_\star)(t-s)} C \mathrm{d}\mathbb{W}_s. \tag{3}$$

According to the above equation, if an eigenvalue of $A_\star + B_\star K_\star$ has non-negative real-part, the state x_t grows unbounded with t. Thus, Assumption 1 is necessary for the problem to be well-defined. Otherwise, state explosion renders the cost infinite for all policies [25], [22]. Note that $A_\star, B_\star$, and so $K_\star$, are unknown.

In the sequel, we examine effects of uncertainties about $A_\star, B_\star$ on the increase in cost compared to its optimal value. The benchmark for assessing reinforcement learning policies is the optimal policy that designs $u_t^\star$ according to the true dynamical model $A_\star, B_\star$. To see that, for generic dynamics matrices A, B, define the feedback matrix $\boldsymbol{K}_{A,B}$ based on $\boldsymbol{P}_{A,B}$, that solves

$$A^\top \boldsymbol{P}_{A,B} + \boldsymbol{P}_{A,B} A - \boldsymbol{P}_{A,B} B R^{-1} B^\top \boldsymbol{P}_{A,B} + Q = 0. \tag{4}$$

Definition 1. For generic dynamics matrices A, B, define $\boldsymbol{K}_{A,B} = -R^{-1} B^\top \boldsymbol{P}_{A,B}$, where $\boldsymbol{P}_{A,B}$ satisfies (4).

So, we show the unique existence of $\boldsymbol{P}_{A_\star,B_\star}$, and establish optimality of the control policy

$$u_t^\star = \boldsymbol{K}_{A_\star,B_\star} x_t^\star, \qquad \text{for all} \quad t \geq 0. \tag{5}$$

Theorem 1. *The matrix $\boldsymbol{P}_{A_\star,B_\star}$ in* (4) *uniquely exists, and the linear feedback policy in* (5) *is optimal.*

To see the intuition of (4) for obtaining the optimal policy in (5), note that the control action u_t directly influences the current cost value, and indirectly affects the future costs according to (1). So, the effects of control actions in the future need to be considered for minimizing the cost function in (2), and this consideration is performed by $\boldsymbol{P}_{A_\star,B_\star}$ [25], [22].

Next, we formulate sub-optimalities and increase in cost due to lack of knowledge about the optimal actions $u_t^\star$. For a reinforcement learning policy, its *regret* at time T is the cumulative increase in the cost. That is, the difference between the costs of the policy and that of the optimal feedback in (5) is integrated over $0 \leq t \leq T$:

$$\mathbf{Reg}(T) = \int_0^T \left(x_t^\top Q x_t + u_t^\top R u_t - x_t^{\star\top} Q x_t^\star - u_t^{\star\top} R u_t^\star\right) \mathrm{d}t.$$

Note that randomness of state and action leads to that of regret. We perform worst-case analysis and bound $\mathbf{Reg}(T)$ in terms of T, p, q. If the increasing observations of state and action over time will be effectively leveraged, the policy eventually takes near-optimal actions. So, $\mathbf{Reg}(T)$ is expected to scale sub-linearly with T. However, design of efficient policies with $\mathcal{O}\left(\sqrt{T}\right)$ regret and proving performance guarantees for them is challenging, as will be discussed shortly.

III. Randomized Certainty Equivalent Policy: Algorithm and Intuition

In this section, we discuss design and analysis of randomization of control policies based on the Certainty Equivalence principle for the continuous-time dynamical system in (1). We aim to have computationally fast algorithms with efficient performance guarantees for minimizing the cost in (2). That is, low-regret control policies subject to uncertainties about the true dynamics matrices $A_\star, B_\star$ in (1).

First, we explain the exploration-exploitation dilemma for the problem under consideration. Then, we investigate an estimation procedure for estimating the unknown dynamics matrices using the data of state-action trajectory. Based on that, the randomized certainty equivalent reinforcement learning policy that employs randomizations of the parameter estimates for balancing exploration versus exploitation is presented in

Algorithm 1. Finally, we provide theoretical and empirical performance analyses for the proposed algorithm.

In order to have a policy whose regret is not very large, we need $u_t \approx \boldsymbol{K}_{A_\star,B_\star} x_t$. Furthermore, since $A_\star, B_\star$ are unknown, the control policy needs to estimate them according to the available trajectory by the time, which is $\{x_s, u_s\}_{0 \leq s \leq t}$. However, if it holds that $u_s \approx \boldsymbol{K}_{A_\star,B_\star} x_s$, then the coordinates u_s of the data x_s, u_s cannot significantly contribute to the estimation procedure as they are approximately linear functions of the state coordinates x_s. This renders accurate estimation of $A_\star, B_\star$ infeasible and defeats the purpose. Note that accurate approximations of $A_\star, B_\star$ are needed for taking near-optimal control actions. The above-mentioned dilemma, known as the exploration-exploitation trade-off, is an important challenge and indicates the fact that a low-regret policy needs to carefully *randomize* the control inputs u_s, and so unavoidably *deviates* from the optimal feedback policy $u_s^\star = \boldsymbol{K}_{A_\star,B_\star} x_s^\star$.

Next, we derive an estimator for the unknown dynamics matrices $A_\star, B_\star$. Intuitively speaking, a framework similar to linear regression is used to obtain the parameter estimates A, B based on the trajectory of the state and the control input. To proceed, suppose that we aim to use samples of the trajectory at ϵ-apart discrete time points; $\{x_{k\epsilon}, u_{k\epsilon}\}_{k=0}^{n}$. So, for small ϵ, the dynamics in (1) gives

$$x_{(k+1)\epsilon} - x_{k\epsilon} \approx \left(A_\star x_{k\epsilon} + B_\star u_{k\epsilon}\right)\epsilon + C\left(\mathbb{W}_{(k+1)\epsilon} - \mathbb{W}_{k\epsilon}\right).$$

Fitting a linear regression, the least-squares estimate is

$$\arg\min_{A,B} \sum_{k=0}^{n-1} \left\| x_{(k+1)\epsilon} - x_{k\epsilon} - \left(A x_{k\epsilon} + B u_{k\epsilon}\right)\epsilon \right\|^2,$$

which, letting $y_s = \left[x_s^\top, u_s^\top\right]^\top$, leads to

$$[A, B] = \sum_{k=0}^{n-1} \left(x_{(k+1)\epsilon} - x_{k\epsilon}\right) y_{k\epsilon}^\top \left(\sum_{k=0}^{n-1} y_{k\epsilon} y_{k\epsilon}^\top \epsilon\right)^{-1}.$$

Therefore, letting $\epsilon \to 0$, we get the continuous-time estimator in (7). The latter is used to estimate the unknown matrices $A_\star, B_\star$ at the end of every episode of Algorithm 1, as explained below.

To introduce the episodes of the algorithm, we use the sequence $\{\tau_n\}_{n=0}^{\infty}$ that contains the time points at which the algorithm updates the parameter estimates. In fact, Algorithm 1 applies control actions $u_t = \boldsymbol{K}_{A_n,B_n} x_t$, during the episode $\tau_n \leq t < \tau_{n+1}$, where A_n, B_n are estimates of the true matrices $A_\star, B_\star$, based on the trajectory up to time τ_n. The episode lengths satisfy

$$\underline{\beta} \leq \inf_{n \geq 1} \frac{\tau_{n+1} - \tau_n}{\tau_n} \leq \sup_{n \geq 1} \frac{\tau_{n+1} - \tau_n}{\tau_n} \leq \overline{\beta}, \tag{6}$$

for some constants $\underline{\beta} > 0, \overline{\beta} < \infty$. The rationale for freezing the parameter estimates during the episodes is that the learning procedure can be deferred until collecting enough new observations. Clearly, smaller $\underline{\beta}, \overline{\beta}$ mean shorter episodes and more frequent updates in parameter estimates, which gives better exploration. Still, the episode lengths $\tau_{n+1} - \tau_n$ grow large to preclude unnecessary updates.

Further, to ensure that the policy is sufficiently committed to explore the environment, a random matrix Φ_n is added to the least-squares estimate, as shown in (7), where $\{\Phi_n\}_{n=0}^{\infty}$ are $p \times (p+q)$ random matrices, independent of everything else and of each others, and has independent standard Gaussian entries. The randomized certainty Equivalent reinforcement learning policy is provided in Algorithm 1.

Algorithm 1 Randomized Certainty Equivalent Policy

Let A_0, B_0 be the initial estimates and $\{\tau_n\}_{n=1}^{\infty}$ satisfy (6)
for $n = 1, 2, \cdots$ **do**
  while $\tau_{n-1} \leq t < \tau_n$ **do**
    Take action $u_t = \boldsymbol{K}_{A_{n-1},B_{n-1}} x_t$
  end while
  Let $y_s = \left[x_s^\top, u_s^\top\right]^\top$ and calculate

$$[A_n, B_n] = \left[\int_0^{\tau_n} y_s \mathrm{d}x_s^\top\right]^\top \left(\int_0^{\tau_n} y_s y_s^\top \mathrm{d}s\right)^{-1} + \frac{\Phi_n}{\tau_n^{1/4}}. \tag{7}$$

end for

The coefficients $\tau_n^{-1/4}$ of the sequence of random matrices are employed to serve a two-fold purpose. On one hand, the scaled random matrix $\tau_n^{-1/4}\Phi_n$ is sufficiently *large* for randomizing the parameter estimates to ensure that effective *exploration* occurs and the data is diverse to provide accurate estimates. At the same time, $\tau_n^{-1/4}\Phi_n$ is sufficiently *small* to prevent significant deviations from the least-squares estimates and from the optimal actions. Otherwise, large randomizations deteriorate the *exploitation*.

Note that implementation of Algorithm 1 is fast and requires minimal memory, as one needs to only update the integrals in (7) in an online fashion.

IV. Performance Analysis: Regret Bound

Next, we establish Theorem 2 indicating that Algorithm 1 efficiently minimizes the cost function so that the regret scales nearly as the square-root of time.

We suppose that when running Algorithm 1, the system evolves in a stable manner such that eigenvalues of the closed-loop matrices $A_\star + B_\star \boldsymbol{K}_{A_n,B_n}$ lie in the open left half-plane of the complex plane. This stability can be ensured in different ways. First, it suffices to find an initial stabilizing policy [26], [16], [17]. If such initial stabilizer is available, one can apply it and devote a (relatively short) time period to exploration, such that the collected data provide a coarse approximation of the true matrices $A_\star, B_\star$. Then, it is shown that such coarse-grained approximations are sufficient for stabilization [16], [27]. Otherwise, to find an initial stabilizer, we can employ Bayesian learning algorithms for a short time period to form a posterior

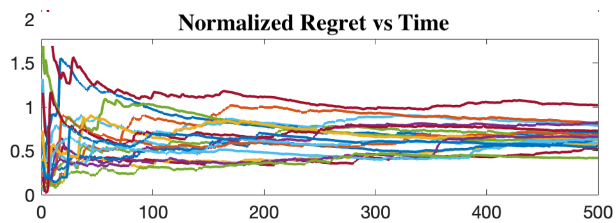

Fig. 1. The above graph presents curves of the normalized regret $T^{-1/2}\mathbf{Reg}(T)$ vs T, for Algorithm 1. Multiple replicates of the system are simulated, all of them corroborate Theorem 2 that the normalized regret remains bounded as time grows.

belief about $A_\star, B_\star$. Then, it is known that samples from the posterior belief guarantee high probability stabilization [27]. Since the sampling procedure can be repeated, we can assume that Algorithm 1 remains stable. Further details can be found in the references, as well as in the papers on discrete-time stabilization [5], [28], [29], [30], [31].

To establish regret bounds for Algorithm 1, we assume that the Brownian noise influences all state variables:

Assumption 2. The matrix C in (1) is full-rank.

This assumption is standard to ensure that the optimal actions can be learned over the course of interactions with the environment [32], [33], [17], [34], [21], [27]. Intuitively, it indicates that all state variables have significant roles and a smaller subset of them is *insufficient* for capturing the dynamics of the environment. Now, we present a theoretical performance guarantee for the randomized certainty equivalent policy.

Theorem 2. *For the regret of Algorithm 1, we have:*

$$\mathbf{Reg}(T) = \mathcal{O}\left(d^2 T^{1/2} \log T\right),$$

where $d = p + q$ is the total dimension of the system.

Above, note that d^2 scales linearly with the number of unknown parameters in $A_\star, B_\star$, which is $p(p+q)$. We experiment Algorithm 1 for flight control of X-29A airplane at 4000 ft altitude [35]. The lateral-directional state-space model is of dimensions $p = 4$, $q = 2$:

$$A_\star = \begin{bmatrix} -0.185 & 0.1475 & -0.9825 & 0.1120 \\ -0.347 & -1.710 & 0.9029 & -0.58\times 10^{-6} \\ 1.174 & -0.0825 & -0.1826 & -0.44\times 10^{-7} \\ 0.0 & 1.0 & 0.1429 & 0.0 \end{bmatrix},$$

$$B_\star = \begin{bmatrix} -0.4470\times 10^{-3} & 0.4020\times 10^{-3} \\ 0.3715 & 0.0549 \\ 0.0265 & -0.0135 \\ 0.0 & 0.0 \end{bmatrix}.$$

Moreover, we let $C = 0.2 \times I_4$, $Q = I_p$, $R = 0.1 \times I_q$, $\tau_n = 25 \times 1.2^n$, and run Algorithm 1 for the system in (1), where $A_\star, B_\star$ are unknown to the algorithm.

Figure 1 depicts the normalized regret versus time for Algorithm 1. The horizontal axis is T, while the vertical one corresponds $T^{-1/2}\mathbf{Reg}(T)$. It shows the result of Theorem 2 that the normalized regret is almost bounded.

V. Concluding Remarks and Future Work

This work studies the randomized certainty equivalent reinforcement learning policy for minimizing quadratic costs in continuous-time linear systems. We presented theoretical performance analysis of the randomized certainty equivalent algorithm showing that it is efficient. More precisely, we established regret bounds growing as square-root of time. Further, dependence on the problem dimension is quadratic indicating that the regret grows linearly with the number of parameters.

The presented results motivate interesting directions in the study of reinforcement learning algorithms for continuous-time environments. Finding regret bounds that hold uniformly over time, and deriving performance guarantees in high-dimensional systems with sparse or low-rank dynamics matrices, are interesting problems for future work. Moreover, extension of the presented analysis to reinforcement learning policies under imperfect state-observations, and policies for non-linear dynamical systems, are problems of interest for future investigations.

Outline of the Appendices

In the first appendix, we prove Theorem 2. Then, the auxiliary lemmas used in the proof are provided.

Appendix A
Proof of Theorem 2

Let $y_s = \left[x_s^\top, u_s^\top\right]^\top$ and $V_n = \int_0^{\tau_n} y_s y_s^\top \mathrm{d}s$. Replace for $\mathrm{d}x_t$ from (1) to obtain

$$\int_0^{\tau_n} y_s \mathrm{d}x_s^\top = \int_0^{\tau_n} y_s y_s^\top \left[A_\star, B_\star\right]^\top \mathrm{d}s + \int_0^{\tau_n} y_s \mathrm{d}\mathbb{W}_s^\top C^\top.$$

So, we have

$$\left[\int_0^{\tau_n} y_s \mathrm{d}x_s^\top\right]^\top V_n^{-1} = \left[A_\star, B_\star\right] + \left[V_n^{-1}\int_0^{\tau_n} y_s \mathrm{d}\mathbb{W}_s^\top C^\top\right]^\top.$$

The above, the triangle inequality, and (7), yield to

$$\left\|\left[A_n, B_n\right] - \left[A_\star, B_\star\right]\right\| \le \left\|V_n^{-1}\int_0^{\tau_n} y_s \mathrm{d}\mathbb{W}_s^\top C^\top\right\| + \left\|\tau_n^{-1/4}\Phi_n\right\|.$$

Since entries of $\tau_n^{-1/4}\Phi_n$ have $\boldsymbol{N}\left(0, {\tau_n}^{-1/2}\right)$ distribution, for $\beta > 0$ we have

$$\log \mathbb{P}\left(\left\|\tau_n^{-1/4}\Phi_n\right\| \ge p^{1/2}\left(p+q\right)^{1/2} {\tau_n}^{-1/4}\beta\right) = \mathcal{O}\left(-\beta^2\right).$$

This, by Borel-Cantelli Lemma and $\tau_n \to \infty$, gives

$$\left\|\tau_n^{-1/4}\Phi_n\right\| = \mathcal{O}\left(d{\tau_n}^{-1/4}\log^{1/2}\tau_n\right).$$

Thus, by Lemma 2, $\left\|\left[A_n, B_n\right] - \left[A_\star, B_\star\right]\right\|$ is at most

$$\mathcal{O}\left(d\left(\frac{\log \boldsymbol{\lambda}_{\max}\left(V_n\right)}{\boldsymbol{\lambda}_{\min}\left(V_n\right)}\right)^{1/2} + d\tau_n^{-1/4}\log^{1/2}\tau_n\right). \quad (8)$$

Now, Lemma 4 provides $\log \boldsymbol{\lambda}_{\max}(V_n) = \mathcal{O}(\log \tau_n)$. On the other hand, we will establish in the sequel that:

$$\liminf_{n\to\infty} \tau_n^{-1/2} \boldsymbol{\lambda}_{\min}(V_n) > 0. \tag{9}$$

Thus, (8) and (9) lead to

$$\|[A_n, B_n] - [A_\star, B_\star]\| = \mathcal{O}\left(d\tau_n^{-1/4} \log^{1/2} \tau_n\right). \tag{10}$$

Next, Lemma 3 implies that

$$\|\boldsymbol{K}_{A_n,B_n} - \boldsymbol{K}_{A_\star,B_\star}\|^2 = \mathcal{O}\left(d^2 \tau_n^{-1/2} \log \tau_n\right).$$

Since during the episode $\tau_{n-1} \leq t < \tau_n$ the parameter estimates are not updated; $K_t = \boldsymbol{K}_{A_{n-1},B_{n-1}}$, by Lemma 4, we have

$$\begin{aligned}
&\int_0^{\tau_n} \|(K_t - \boldsymbol{K}_{A_\star,B_\star})\, x_t\|^2 \mathrm{d}t \\
&= \mathcal{O}\left(\sum_{k=1}^{n} \frac{\tau_k - \tau_{k-1}}{\tau_{k-1}^{1/2}} d^2 \log \tau_{k-1}\right) \\
&= \mathcal{O}\left(\sum_{k=1}^{n} \left(\tau_k^{1/2} - \tau_{k-1}^{1/2}\right) d^2 \log \tau_n\right) \\
&= \mathcal{O}\left(\tau_n^{1/2} d^2 \log \tau_n\right),
\end{aligned}$$

where in the last two equalities above we used (6).

Moreover, since all eigenvalues of $A_\star + B_\star \boldsymbol{K}_{A_\star,B_\star}$ lie in the open left half-plane, the matrix E_t in Lemma 1 decays exponentially with t. So, we have

$$\int_0^{T} \left(x_t^\top E_{T-t} (K_t - \boldsymbol{K}_{A_\star,B_\star})\, x_t\right) \mathrm{d}t = \mathcal{O}\left(\log^2 T\right).$$

Thus, plugging the above two results in Lemma 1, we obtain the desired result in Theorem 2. In order to complete the proof, we show (9). To that end, Lemma 4 gives

$$\liminf_{k\to\infty} \tau_k^{-1} \boldsymbol{\lambda}_{\min}\left(\int_{\tau_{k-1}}^{\tau_k} x_t x_t^\top \mathrm{d}t\right) > 0. \tag{11}$$

Thus, for getting (9), it is enough to establish

$$\liminf_{n\to\infty} \boldsymbol{\lambda}_{\min}\left(\sum_{k=\ell}^{n-1} \frac{\tau_k}{\tau_n^{1/2}} \begin{bmatrix} I_p \\ \boldsymbol{K}_{A_k,B_k} \end{bmatrix} \begin{bmatrix} I_p \\ \boldsymbol{K}_{A_k,B_k} \end{bmatrix}^\top\right) > 0, \tag{12}$$

for some $0 \leq \ell < n-1$. For $\epsilon > 0$, consider the event that the above smallest eigenvalue is less than ϵ, and let $\mathcal{M}_n(\epsilon)$ be the set of matrices $[A_k, B_k]_{k=\ell}^{n-1}$ for which this event occurs. More precisely, define the $(p+q) \times p(n-\ell)$ matrix $P_{\ell,n}$ as

$$\left[\tau_\ell^{1/2} \tau_n^{-1/4} \begin{bmatrix} I_p \\ \boldsymbol{K}_{A_\ell,B_\ell} \end{bmatrix}, \cdots, \tau_{n-1}^{1/2} \tau_n^{-1/4} \begin{bmatrix} I_p \\ \boldsymbol{K}_{A_{n-1},B_{n-1}} \end{bmatrix}\right].$$

Then, let

$$\mathcal{M}_n(\epsilon) = \left\{[A_\ell, B_\ell, \cdots, A_{n-1}, B_{n-1}] : \boldsymbol{\lambda}_{\min}\left(P_{\ell,n} P_{\ell,n}^\top\right) \leq \epsilon\right\}.$$

Now, note that the set of all matrices

$$F_n = \begin{bmatrix} \tau_\ell^{1/2} \tau_n^{-1/4} I_p & \cdots & \tau_{n-1}^{1/2} \tau_n^{-1/4} I_p \\ \tau_\ell^{1/2} \tau_n^{-1/4} K_\ell & \cdots & \tau_{n-1}^{1/2} \tau_n^{-1/4} K_{n-1} \end{bmatrix},$$

for which there is $v \in \mathbb{R}^{p+q}$ satisfying $\|v\| = 1$, $F_n^\top v = 0$, is of dimension $p + q - 1 + (n-\ell)(q-1)$, as follows:

1) The set of unit $p+q$ dimensional vectors is (a sphere) of dimension $p+q-1$.
2) Write $v = \left[v_1^\top, v_2^\top\right]^\top$, for $v_1 \in \mathbb{R}^p$ and $v_2 \in \mathbb{R}^q$. So, $F_n^\top v = 0$, if and only if $K_k^\top v_2 = -v_1$, for all $k = \ell, \cdots, n-1$. This means every column of K_k is in a certain hyperplane in $\mathbb{R}^q$.

According to Lemma 5, the dimension of $\mathcal{M}_n(0)$ is at most

$$p + q - 1 + (q-1)(n-\ell) + (n-\ell)p^2,$$

and it lives in a $p(p+q)(n-\ell)$ dimensional space. So, the difference between the dimensions is

$$m = (pq - q + 1)(n - \ell) - p - q + 1.$$

Further, if ℓ is sufficiently large so that $\tau_\ell^{-1} \tau_n^{1/2} \epsilon < 1$, then for every $[A_k, B_k]_{k=\ell}^{n-1} \in \mathcal{M}_n(\epsilon)$, there exists some $\left[\widetilde{A}_k, \widetilde{B}_k\right]_{k=\ell}^{n-1} \in \mathcal{M}_n(0)$, such that

$$\max_{\ell \leq k \leq n-1} \left\|[A_k, B_k] - \left[\widetilde{A}_k, \widetilde{B}_k\right]\right\| = \mathcal{O}\left(\tau_k^{-1/2} \tau_n^{1/4} \epsilon^{1/2}\right).$$

Next, we use the above result to bound the probability of $\mathcal{M}_n(\epsilon)$. Note that the random matrices $\{\Phi_k\}_{k=0}^{n-1}$ are independent, and entries of $\tau_k^{-1/4} \Phi_k$ are independent identically distributed $\boldsymbol{N}\left(0, \tau_k^{-1/2}\right)$ random variables. Recall that the difference between the dimensions of $\mathcal{M}_n(0)$ and the space it is in, is m, as defined above. Hence, since $\tau_\ell \leq \tau_k$, we have

$$\mathbb{P}(\mathcal{M}_n(\epsilon)) = \left[\mathcal{O}\left(\tau_\ell^{1/4} \tau_\ell^{-1/2} \tau_n^{1/4} \epsilon^{1/2}\right) \wedge 1\right]^m.$$

Letting $\ell = n - 5$, we have $m \geq 5$. Further, if ϵ is small enough to satisfy $\mathcal{O}\left(\tau_n^{1/4} \tau_\ell^{-1/4} \epsilon^{1/2}\right) < 1$, we have

$$\sum_{n=5}^{\infty} \mathbb{P}(\mathcal{M}_n(\epsilon)) < \infty.$$

The above, by Borel-Cantelli Lemma, implies (12), which completes the proof.

Appendix B
Auxiliary Results

In this section, we present the lemmas used in the proof of Theorem 2. Lemma 1 expresses the regret in terms of the deviations of the control input from the optimal feedback. Then, in Lemma 2 we present the growth rates of stochastic integrals normalized by the empirical state-input covariance matrix. Lemma 3 establishes Lipschitz continuity of the optimal feedback matrix $\boldsymbol{K}_{A,B}$ with respect to A, B. Next, the state empirical covariance matrix is shown to converge to a positive definite limit in Lemma 4. Finally, the result of Lemma 5 provides the dimension of the optimality manifold. Proofs are omitted due to space constraints and can be found in [36].

Lemma 1. *Suppose that K_t is piecewise continuous and $u_t = K_t x_t$ is applied to the system in* (1)*. Define $\boldsymbol{K}_\star = \boldsymbol{K}_{A_\star,B_\star}$, $D_\star = A_\star + B_\star \boldsymbol{K}_\star$, $E_t = e^{D_\star^\top t} \boldsymbol{P}_{A_\star,B_\star} e^{D_\star t} B_\star$,*

$$\alpha_T = \int_0^T \Big(\left\| \left(K_t - \boldsymbol{K}_\star\right) x_t \right\|^2 - 2 x_t^\top E_{T-t} \left(K_t - \boldsymbol{K}_\star\right) x_t \Big) \mathrm{d}t.$$

Then, we have $\mathbf{Reg}\,(T) = \mathcal{O}\left(\alpha_T\right)$.

Lemma 2. *For $y_t = \left[x_t^\top, u_t^\top\right]^\top$, let $V_t = \int_0^t y_s y_s^\top \mathrm{d}s$. Then, it holds that*

$$\left\| \left(I + V_t\right)^{-1/2} \int_0^t y_s \mathrm{d}\mathbb{W}_s^\top \right\|^2 = \mathcal{O}\left(d^2 \log \boldsymbol{\lambda}_{\max}\left(V_t\right)\right).$$

Lemma 3. *There exists $\beta_\star < \infty$, such that*

$$\|\boldsymbol{K}_{A,B} - \boldsymbol{K}_{A_\star,B_\star}\| \le \beta_\star \|[A,B] - [A_\star, B_\star]\|.$$

Lemma 4. *In Algorithm 1, suppose that real-parts of all eigenvalues of $A_\star + B_\star \boldsymbol{K}_{A_n,B_n}$ are negative. Then, the following matrix is deterministic and positive definite:*

$$\lim_{n\to\infty} \frac{1}{\tau_{n+1} - \tau_n} \int_{\tau_n}^{\tau_{n+1}} x_t x_t^\top \mathrm{d}t.$$

Lemma 5. *For some fixed A_0, B_0, consider*

$$\mathcal{N}_0 = \left\{ [A,B] \in \mathbb{R}^{p \times (p+q)} : \boldsymbol{K}_{A,B} = \boldsymbol{K}_{A_0,B_0} \right\}.$$

The set $\mathcal{N}_0$ is a manifold of dimension p^2.

References

[1] D. T. Gillespie, "Stochastic simulation of chemical kinetics," *Annu. Rev. Phys. Chem.*, vol. 58, pp. 35–55, 2007.

[2] H. Schmidli, *Stochastic control in insurance*. Springer Science & Business Media, 2007.

[3] H. Pham, *Continuous-time stochastic control and optimization with financial applications*. Springer Science & Business Media, 2009, vol. 61.

[4] N. D. Lawrence, M. Girolami, M. Rattray, and G. Sanguinetti, *Learning and inference in computational systems biology*. MIT press, 2010.

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

[6] Y. Ouyang, M. Gagrani, and R. Jain, "Posterior sampling-based reinforcement learning for control of unknown linear systems," *IEEE Transactions on Automatic Control*, vol. 65, no. 8, pp. 3600–3607, 2019.

[7] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "On applications of bootstrap in continuous space reinforcement learning," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 1977–1984.

[8] ——, "On adaptive linear–quadratic regulators," *Automatica*, vol. 117, p. 108982, 2020.

[9] ——, "Input perturbations for adaptive control and learning," *Automatica*, vol. 117, p. 108950, 2020.

[10] A. Cassel, A. Cohen, and T. Koren, "Logarithmic regret for learning linear quadratic regulators efficiently," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1328–1337.

[11] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, "Logarithmic regret bound in partially observable linear dynamical systems," *arXiv preprint arXiv:2003.11227*, 2020.

[12] I. Ziemann and H. Sandberg, "Regret lower bounds for unbiased adaptive control of linear quadratic regulators," *IEEE Control Systems Letters*, vol. 4, no. 3, pp. 785–790, 2020.

[13] ——, "On uninformative optimal policies in adaptive lqr with unknown b-matrix," in *Learning for Dynamics and Control*. PMLR, 2021, pp. 213–226.

[14] P. Mandl, T. E. Duncan, and B. Pasik-Duncan, "On the consistency of a least squares identification procedure," *Kybernetika*, vol. 24, no. 5, pp. 340–346, 1988.

[15] T. E. Duncan and B. Pasik-Duncan, "Adaptive control of continuous-time linear stochastic systems," *Mathematics of Control, signals and systems*, vol. 3, no. 1, pp. 45–60, 1990.

[16] T. E. Duncan, L. Guo, and B. Pasik-Duncan, "Adaptive continuous-time linear quadratic gaussian control," *IEEE Transactions on automatic control*, vol. 44, no. 9, pp. 1653–1662, 1999.

[17] P. E. Caines and D. Levanony, "Stochastic ε-optimal linear quadratic adaptation: An alternating controls policy," *SIAM Journal on Control and Optimization*, vol. 57, no. 2, pp. 1094–1126, 2019.

[18] T. Bian, Y. Jiang, and Z.-P. Jiang, "Adaptive dynamic programming for stochastic systems with state and control dependent noise," *IEEE Transactions on Automatic control*, vol. 61, no. 12, pp. 4170–4175, 2016.

[19] K. Doya, "Reinforcement learning in continuous time and space," *Neural computation*, vol. 12, no. 1, pp. 219–245, 2000.

[20] H. Wang, T. Zariphopoulou, and X. Y. Zhou, "Reinforcement learning in continuous time and space: A stochastic control approach." *J. Mach. Learn. Res.*, vol. 21, pp. 198–1, 2020.

[21] M. Basei, X. Guo, A. Hu, and Y. Zhang, "Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon," *Available at SSRN 3848428*, 2021.

[22] J. Yong and X. Y. Zhou, *Stochastic controls: Hamiltonian systems and HJB equations*. Springer Science & Business Media, 1999, vol. 43.

[23] B. Oksendal, *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

[24] P. Baldi, *Stochastic Calculus: An Introduction Through Theory and Exercises*. Springer, 2017.

[25] G. Chen, G. Chen, and S.-H. Hsu, *Linear stochastic control systems*. CRC press, 1995, vol. 3.

[26] P. Caines, "Continuous time stochastic adaptive control: non-explosion, ε-consistency and stability," *Systems & control letters*, vol. 19, no. 3, pp. 169–176, 1992.

[27] M. K. S. Faradonbeh and M. S. S. Faradonbeh, "Bayesian algorithms learn to stabilize unknown continuous-time systems," *arXiv preprint arXiv:2112.15094*, 2021.

[28] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite-time adaptive stabilization of linear systems," *IEEE Transactions on Automatic Control*, vol. 64, no. 8, pp. 3498–3505, 2018.

[29] ——, "Randomized algorithms for data-driven stabilization of stochastic linear systems," in *2019 IEEE Data Science Workshop (DSW)*. IEEE, 2019, pp. 170–174.

[30] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, "Explore more and improve regret in linear quadratic regulators," *arXiv preprint arXiv:2007.12291*, 2020.

[31] X. Chen and E. Hazan, "Black-box control for linear dynamical systems," in *Conference on Learning Theory*. PMLR, 2021, pp. 1114–1143.

[32] D. Levanony and P. E. Caines, "On persistent excitation for linear systems with stochastic coefficients," *SIAM journal on control and optimization*, vol. 40, no. 3, pp. 882–897, 2001.

[33] A. Subrahmanyam and G. P. Rao, *Identification of Continuous-time Systems: Linear and Robust Parameter Estimation*. CRC Press, 2019.

[34] Z.-P. Jiang, T. Bian, and W. Gao, "Learning-based control: A tutorial and some recent results," *Foundations and Trends® in Systems and Control*, vol. 8, no. 3, 2020.

[35] J. T. Bosworth, *Linearized aerodynamic and control law models of the X-29A airplane and comparison with flight data*. National Aeronautics and Space Administration, Office of Management ..., 1992, vol. 4356.

[36] M. K. S. Faradonbeh and M. S. S. Faradonbeh, "Efficient estimation and control of unknown stochastic differential equations," *arXiv preprint arXiv:2109.07630*, 2021.