Data-Driven Exploration for a Class of Continuous-Time Indefinite Linear–Quadratic Reinforcement Learning Problems

Yilie Huang, *Member, IEEE*, Xun Yu Zhou, *Fellow, IEEE*

***Abstract*— We study reinforcement learning (RL) for a class of continuous-time stochastic linear–quadratic (LQ) control problems, where volatilities depend on both states and controls while states are scalar-valued and there are no running control rewards, the last feature also leading to the so-called *indefinite* LQ controls. We propose a model-free, data-driven exploration mechanism that adaptively adjusts entropy regularization by the critic and policy variance by the actor. Unlike the constant or deterministic exploration schedules employed in the existing literature, which require extensive tuning for implementations and ignore learning progress during iterations, our adaptive exploratory approach boosts learning efficiency with minimal tuning. Despite its flexibility, our method achieves a sublinear regret bound that matches the best-known model-free results for this class of LQ problems, which were previously derived only with fixed exploration schedules. Numerical experiments demonstrate that adaptive explorations accelerate convergence and improve regret performance compared to the non-adaptive model-free and model-based counterparts.**

***Index Terms*— Stochastic Indefinite Linear–Quadratic Control, Continuous-Time Reinforcement Learning, Actor, Critic, Data-Driven Exploration, Model-Free Policy Gradient, Regret Bounds.**

I. Introduction

Linear–quadratic (LQ) control is a cornerstone of optimal control theory, widely applied in engineering, economics, and robotics among many others due to its analytical tractability and practical relevance. Its theory has been extensively developed in the classical model-based setting where all the LQ coefficients are known and given [1], [2]. Assuming the objective is to *minimize* the cost functional, a key assumption in the LQ theory is that the control cost weighting matrix must be positive definite. Intuitively, this is because if increasing the level of control is costless or even beneficial, then the optimal control would just be infinitely large leading to an ill-posed problem. [3] finds, however, that certain stochastic LQ control problems with indefinite cost matrices—including those associated with control—can still be well-posed if the diffusion coefficients depend on the control variable. This seemingly surprising result actually has a simple explanation: while benefiting from an indefinite control cost, a larger control is disadvantageous in terms of increasing the level of uncertainty because the volatility term depends on control multiplicatively. Therefore, one needs to carefully choose the optimal level of control, rendering a well-posed optimization problem. [3] coins the term "indefinite stochastic LQ control" for such a problem, and establishes conditions for well-posedness and solvability via newly introduced generalized Riccati equations. Subsequent works extend the theory in several directions, including infinite-time horizons [4] and connections with linear matrix inequalities [5], asymptotic properties of the Riccati solutions [6], and semidefinite programming for numerical computations [7]. Research on indefinite LQ controls and the related generalized Riccati equations has been active to this day; see, e.g., [8]–[11].

All the research on indefinite LQ controls, and indeed most studies on stochastic controls, are undertaken in a model-based paradigm where the system coefficients and objective/cost functionals are fully known. In real-world applications, however, complete knowledge of model parameters is rarely available. Many environments exhibit LQ-like structure yet key parameters in dynamics and objectives may be partially known or entirely unknown. While one can use historical data to estimate those model parameters as commonly done in the so-called "plug-in" approach, estimating some of the parameters to a required accuracy may not be possible. For instance, it is statistically impossible to estimate a stock return rate for a very small period – a phenomenon termed "mean-blur" [12], [13]. On the other hand, objective function values can be highly sensitive to estimated model parameters, rendering inferior performances or even complete irrelevance of model-based solutions [14].

Given the limitations of model-based controls, particularly the challenges associated with unknown model coefficients, reinforcement learning (RL) rises to provide a promising remedy. Simply put, RL is stochastic control with *unknown* model parameters. Yet it never attempts to estimate any model coefficients, and solves the problem in a fundamentally different way compared to the classical, model-based stochastic control.

Initially submitted on 29 June 2025. The authors were supported by the Nie Center for Intelligent Asset Management at Columbia University. Their work is also part of a Columbia–CityU/HK collaborative project that is supported by the InnoHK Initiative, the Government of the HKSAR, and the AIFT Lab. Yilie Huang was also supported by a start-up fund from The Hong Kong Polytechnic University.

Yilie Huang: Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (yilie.huang@polyu.edu.hk; tel. +852 2766 4010; fax +852 2362 9045).

Xun Yu Zhou (corresponding author): Department of Industrial Engineering and Operations Research & Data Science Institute, Columbia University, 500 West 120th Street, New York, NY 10027, USA (xz2574@columbia.edu; tel. +1 212 854 5237; fax +1 212 854 8103).

Within RL, LQ control has also served as a key benchmark for studying learning in controlled systems. For deterministic LQ problems, convergence guarantees for learning-based controllers have been obtained; see, for example, [15], [16]. For systems with identically and independently distributed noises, regret and finite-sample guarantees for learning-based controllers have been developed, including policy-gradient type methods; see, for example, [17]–[21]. These results, however, are all established in discrete time, typically under the assumption that the control does not directly affect the volatility level of the state dynamics, in contrast to the continuous-time, state- and control-dependent diffusion setting considered in this paper. Moreover, their exploration mechanisms are tailored to discrete-time models and do not directly extend to continuous-time diffusion dynamics.

A centerpiece of RL, including the works mentioned above, is exploration, with which the agent interacts with the unknown environment and improves her policies. The simplest exploration strategy is perhaps the so-called ϵ-greedy for bandits problem [22]. More sophisticated and efficient exploration strategies include intrinsic motivation methods that reward exploring novel or unpredictable states [23]–[29], count-based exploration that encourages the agent to explore less frequently visited states [30], [31], go-explore that explicitly stores and revisits promising past states to enhance sample efficiency [32], [33], imitation-based exploration which leverages expert demonstrations [34], and safe exploration methods that rely on predefined safety constraints or risk models [35], [36]. While these methods are shown to be successful in their respective contexts, they are all for discrete-time problems and seem difficult to be extended to the continuous-time counterparts. However, most real-life applications are inherently continuous-time with continuous state spaces and possibly continuous control spaces (e.g., autonomous driving, robot navigation, and video game playing). On the other hand, transforming a continuous-time problem into a discrete-time one upfront by discretization has shown to be problematic when the time step size becomes very small [37]–[39].

Thus, there is a pressing need for developing exploration strategies tailored to continuous-time RL, particularly and foremost in the LQ setting, for achieving both stability and learning efficiency. Recent works, such as [40], [41], have employed constant or deterministic exploration schedules to regulate exploration parameters and proved to achieve sublinear regret bounds. However, these approaches come with notable drawbacks. They require extensive manual tuning of the exploration hyperparameters, which considerably increases computational costs, yet these tuning efforts are typically not reflected in theoretical analyses and results including those concerning regret bounds. Moreover, pre-determined exploration strategies lack adaptability by nature, as they do not react to the incoming data nor to the current learning states, typically resulting in slower convergence and increased variance from excessive exploration or premature convergence to suboptimal policies from insufficient exploration.

This paper presents a companion research of [40]. We continue to study the same class of stochastic LQ problems in [40], which is also of an indefinite LQ type because control is absent in the objective functional. We develop a data-driven adaptive exploration framework for both critic (the value functions) and actor (the control policies). Specifically, we propose an adaptive exploration strategy that dynamically adjusts the entropy regularization parameter for the critic and the stochastic policy variance for the actor based on the agent's learning progress. Moreover, we provide theoretical guarantees by proving the almost sure convergence of both the adaptive exploration and policy parameters, along with a sublinear regret bound of $O(N^{\frac{3}{4}})$. This bound matches the best-known model-free result from [40], which assumes a nonzero initial state and uses a fixed/deterministic exploration schedule, whereas we remove this assumption and learn entirely from data. Finally, we validate our theoretical results through numerical experiments, demonstrating a faster convergence and improved regret bounds compared with a model-based benchmark with fixed exploration. For a comparison with the previous paper [40], we design experiments with both fixed and randomly generated model parameters. While theoretically the regret bounds in the two papers are of the same order, the numerical experiments show that our adaptive exploration strategy consistently outperforms and achieves lower regrets.

A distinctive advantage of the RL approach over the traditional model-based one for indefinite LQ control is that the latter relies heavily on the generalized (or indefinite) Riccati equation, which is a highly complex and unconventional Riccati equation. In particular, it involves a positive definiteness constraint and may be singular due to the indefiniteness of the control weighting cost [3]. A central task for solving indefinite LQ control problems has been to solve this equation analytically or numerically (e.g. [5]–[7], [42], [43]), which remains open for the most general case in terms of its well-posedness. By contrast, RL does not need to involve this equation at all, in the same way that RL gets around the HJB PDE for a general stochastic control problem.

The remainder of this paper is organized as follows. Section II formulates the problem. Section III discusses the limitations of deterministic exploration strategies and presents our data-driven adaptive exploration mechanism. Section IV describes the continuous-time LQ-RL algorithm with adaptive exploration. Section V establishes theoretical guarantees, proving convergence properties and regret bounds. Section VI presents numerical experiments that validate the effectiveness of our approach and compare it against model-free and model-based benchmarks with deterministic exploration schedules. Finally, Section VII concludes.

II. Problem Formulation and Preliminaries

A. Stochastic LQ Control: Classical Setting

We consider the same stochastic LQ control problem in [40]. The one-dimensional state process $x^u = \{x^u(t) \in \mathbb{R} : 0 \leq t \leq T\}$ evolves under the multi-dimensional control process $u = \{u(t) \in \mathbb{R}^l : 0 \leq t \leq T\}$ according to the stochastic differential equation (SDE):

$$\begin{aligned} \mathrm{d}x^u(t) =& (Ax^u(t) + B^\top u(t))\mathrm{d}t \\ &+ \sum_{j=1}^{m} (C_j x^u(t) + D_j^\top u(t))\mathrm{d}W^{(j)}(t), \end{aligned} \tag{1}$$

where A and C_j are scalars, B and D_j are $l \times 1$ vectors, and the initial condition is $x^u(0) = x_0$. The noise process $W = \{(W^{(1)}(t), \dots, W^{(m)}(t))^\top \in \mathbb{R}^m : 0 \le t \le T\}$ is an m-dimensional standard Brownian motion. We assume $\sum_{j=1}^m D_j D_j^\top > 0$ to ensure well-posedness of the problem soon to be formulated.

The objective is to find a control process u that maximizes the expected quadratic reward:

$$\max_u \mathbb{E}\left[\int_0^T -\frac{1}{2}Qx^u(t)^2\mathrm{d}t - \frac{1}{2}Hx^u(T)^2\right], \tag{2}$$

where $Q \ge 0$ and $H \ge 0$ are scalar weighting parameters. Note that we *maximize* the reward functional (with negative signs), instead of minimizing a cost functional as in the usual LQ literature, to be consistent with the setting of [40]. Because there is no penalty on control, the problem is an instance of the indefinite LQ control [3]. In the classical model-based setting, such an indefinite problem requires a careful analysis of the associated *indefinite* Riccati equation (e.g., [3]). In our model-free continuous-time RL framework, we do not solve the Riccati equation, and develop the same theory and algorithm whether the problem is definite or indefinite.

We are able to treat only the case of one-dimensional state processes because the method in this paper crucially depends on the feature that optimal control turns out to be time-*invariant*. Specifically, the optimal feedback control in general takes the form $u_{CL}(t,x) = -\big(\sum_{j=1}^m D_j^\top P(t) D_j\big)^{-1}\big(B^\top P(t) + \sum_{j=1}^m D_j^\top P(t) C_j\big)x$, where P solves the associated matrix Riccati ODE; see [2, Chapter 6]. When x and therefore P are multi-dimensional, this control is time variant and our method fails. However, when x is one-dimensional, the $P(t)$'s in the above expression cancel out and the control becomes time independent. The case of multi-dimensional state processes remains an outstanding open question.

B. Stochastic LQ Control: RL Setting

We now consider the *model-free*, RL version of the problem just formulated, following [44], without assuming a complete knowledge of the model parameters or relying on parameter estimation.

In this case, there is no Riccati equation to solve (because its coefficients are unknown) and no explicit solutions to compute as in [2, Chapter 6]. Instead, the RL agent randomizes controls and considers distribution-valued control processes of the form $\pi = \{\pi(\cdot,t) \in \mathcal{P}(\mathbb{R}^l) : 0 \le t \le T\}$, where $\mathcal{P}(\mathbb{R}^l)$ is the set of all probability density functions over $\mathbb{R}^l$. The agent employs the control $u = \{u(t) \in \mathbb{R}^l : 0 \le t \le T\}$, where $u(t) \in \mathbb{R}^l$ is sampled from $\pi(\cdot,t)$ at each t, to control the original state dynamic.

According to [44], the system dynamic under a randomized control π follows the SDE:

$$\mathrm{d}x^\pi(t) = \widetilde{b}(x^\pi(t), \pi(\cdot,t))\mathrm{d}t + \sum_{j=1}^m \widetilde{\sigma}_j(x^\pi(t), \pi(\cdot,t))\mathrm{d}W^{(j)}(t), \tag{3}$$

where

$$\widetilde{b}(x,\pi) := Ax + B^\top \int_{\mathbb{R}^l} u\pi(u)\mathrm{d}u, \tag{4}$$

$$\widetilde{\sigma}_j(x,\pi) := \sqrt{\int_{\mathbb{R}^l} (C_j x + D_j^\top u)^2 \pi(u)\mathrm{d}u}. \tag{5}$$

To encourage exploration, an entropy term is added to the objective function. The entropy-regularized value function associated with a given policy π is defined as

$$\begin{aligned} J(t,x;\pi) = \mathbb{E}\bigg[\int_t^T \Big(-\frac{1}{2}Q(x^\pi(s))^2 + \gamma p^\pi(s)\Big)\mathrm{d}s \\ - \frac{1}{2}H(x^\pi(T))^2 \Big| x^\pi(t) = x\bigg], \end{aligned} \tag{6}$$

where $p^\pi(t) = -\int_{\mathbb{R}^l} \pi(t,u)\log\pi(t,u)\mathrm{d}u$ denotes the differential entropy of π, and $\gamma \ge 0$ is the so-called temperature parameter, which represents the trade-off between exploration and exploitation. The optimal value function is then given by $V(t,x) = \max_\pi J(t,x;\pi)$.

Applying standard stochastic LQ control techniques as in [2, Chapter 6], [40, Section 2.2] derives explicitly the optimal value function and control for the same problem studied in this paper, leading directly to

$$V(t,x) = -\frac{1}{2}k_1(t)x^2 + k_3(t), \tag{7}$$

$$\pi(u \mid t,x) = \mathcal{N}\left(u \Big| \bar{\mu}x, \Sigma\right), \tag{8}$$

where $\mathcal{N}(\cdot|\bar{\mu}x,\Sigma)$ represents a multivariate Gaussian distribution with mean $\bar{\mu}x$ and covariance matrix Σ with $\bar{\mu} = -\Big(\sum_{j=1}^m D_j D_j^\top\Big)^{-1}\Big(B + \sum_{j=1}^m C_j D_j\Big)$ and $\Sigma = \frac{\gamma}{k_1(t)}(\sum_{j=1}^m D_j D_j^\top)^{-1}$. The functions k_1 and k_3 are determined by ordinary differential equations:

$$\begin{aligned} k_1'(t) =& -\bigg[2A + 2B^\top\bar{\mu} + \sum_{j=1}^m\Big(D_j^\top\bar{\mu}\bar{\mu}^\top D_j + 2C_j D_j^\top\bar{\mu} \\ &+ C_j^2\Big)\bigg]k_1(t) - Q, \qquad k_1(T) = H, \\ k_3'(t) =& \frac{k_1(t)}{2}\sum_{j=1}^m D_j^\top\Sigma D_j - \frac{\gamma}{2}\log\Big((2\pi e)^l\det(\Sigma)\Big), \\ & k_3(T) = 0. \end{aligned} \tag{9}$$

It is clear from (9) that $k_1(t) > 0$ is independent of γ, and k_1, k_3 along with their derivatives are uniformly bounded over $[0,T]$. [1]

We stress that the above are theoretical results that cannot be used to compute the optimal solutions because none of the parameters in the expressions is known; yet they offer valuable *structural* insights for developing the LQ-RL algorithms. Specifically, the optimal value function (critic) is quadratic in the state variable x, while the optimal (randomized) feedback control policy (actor) follows a Gaussian distribution, with its

[1] Write $c := 2A + 2B^\top\bar{\mu} + \sum_{j=1}^m\big(D_j^\top\bar{\mu}\bar{\mu}^\top D_j + 2C_j D_j^\top\bar{\mu} + C_j^2\big)$. Then the explicit solution of k_1 is $k_1(t) = -Q/c + (H + Q/c)\exp\big(-c(t-T)\big)$ if $c \ne 0$ and $k_1(t) = H - Q(t-T)$ if $c = 0$.

mean exhibiting a linear relationship with x. This intrinsic structure will play a pivotal role in reducing the complexity of function parameterization/approximation in our subsequent learning procedure.

C. Exploration in LQ-RL: Actor and Critic Perspectives

In our RL formulation above, both the actor and the critic engage in controls of exploration. The actor does this through the variance in the Gaussian policy (8), which represents the level of additional randomness injected into the interactions with the environment. The critic, on the other hand, regulates exploration via the entropy-regularized objective (6), where the temperature parameter γ determines, if indirectly, the extent of exploration. Managing exploration properly is both important and challenging: excessive exploration can cause instability and hinder convergence, and insufficient exploration may prevent effective policy improvement. The next section introduces a data-driven adaptive exploration strategy designed to address these challenges and optimize policy learning.

III. Data-driven Exploration Strategy for LQ-RL

This section presents the core contribution of this work: data-driven explorations by both the actor and the critic.

A. Limitations of Deterministic Exploration Strategies

A critical challenge in RL is to design exploration strategies that are both effective and efficient. In the setting of continuous-time LQ, existing methods [40], [41] employ non-adaptive strategies by setting exploration levels of both actors and critics either as fixed constants or some deterministic schedules of time. While these strategies are proved to have theoretical guarantees, they exhibit several essential limitations in learning efficiency and practical implementations.

One major limitation of deterministic explorations lies in their arbitrary nature. Whether fixed as a constant or a predefined sequence, such a schedule often requires extensive manual tuning in actual implementations. The tuning process in turn may introduce significant computational and time costs for learning; yet these costs are rarely incorporated into theoretical analyses, creating a disconnect between theoretical guarantees and practical performance.

Another drawback of deterministic explorations is their lack of adaptability, as they follow fixed exploration trajectories regardless of the current states of the iterates, often leading to both excessive and insufficient explorations. While maintaining a constant exploration level is clearly ineffective, a deterministic schedule also suffers from a fundamental limitation: it adjusts the exploration parameter in a *predetermined* direction. Consequently, when the current exploration level is either too high or too low, the fixed schedule not only fails to correct this but may indeed exacerbate the situation by moving in the wrong direction. We will illustrate this issue with the experiments presented in Section VI-C.

To address these limitations, we propose an endogenous, data-driven approach that adaptively updates both the actor and critic exploration parameters, which will be shown in the subsequent sections.

B. Adaptive Critic Exploration

a) Critic Parameterization: Motivated by the form of the optimal value function in (7), we parameterize the value function with learnable parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ and temperature parameter $\gamma \in \mathbb{R}$:

$$J(t,x;\boldsymbol{\theta},\gamma) = -\frac{1}{2}k_1(t;\boldsymbol{\theta})x^2 + k_3(t;\boldsymbol{\theta},\gamma), \tag{10}$$

where both k_1 and k_3 and their derivatives are continuous. Both functions can be neural nets; e.g. $k_1(\cdot;\boldsymbol{\theta})$ can be, for example, a positive neural network with a sigmoid activation. Furthermore, these functions are constructed so that there are appropriate positive constants c_1, c_2, c_3 such that

$$\frac{1}{c_2} \le k_1(t;\boldsymbol{\theta}) \le c_2, \left|k_1'(t;\boldsymbol{\theta})\right| \le c_1, \left|k_3'(t;\boldsymbol{\theta},\gamma)\right| \le c_3, \tag{11}$$

for all $0 \le t \le T$, $|\boldsymbol{\theta}| \le c^{(\boldsymbol{\theta})}$, and $|\gamma| \le c^{(\gamma)}$, where $c^{(\boldsymbol{\theta})} > 0$ and $c^{(\gamma)} > 0$ are (fixed) hyperparameters. The values of c_1, c_2, c_3 are determined by the specific parametric forms of $k_1(\cdot;\boldsymbol{\theta})$ and $k_3(\cdot;\boldsymbol{\theta},\gamma)$, as well as $c^{(\boldsymbol{\theta})}$ and $c^{(\gamma)}$. [2] The boundedness assumptions (11) follow from the fact that when the model parameters are known, the corresponding functions satisfy the same conditions, as shown in Section II-B.

In what follows, we introduce the subscript n to denote values at the n-th iteration. For instance, γ_n represents the updated value of γ at iteration n, which evolves dynamically as learning progresses.

b) Data-Driven Temperature Parameter: The temperature parameter γ plays a crucial role in RL in controlling the weight on the entropy-regularized term in the objective function (6). It governs the level of exploration by the critic for the purpose of balancing between exploration and exploitation. A larger γ_n increases the weight on entropy and encourages more exploration, whereas our choice of γ_n (see (12)) makes it gradually decrease as learning progresses, shifting the emphasis from exploration towards exploitation.

In the related literature, γ has been either taken as a given, exogenous constant hyperparameter [40], [45] or an ad hoc deterministic sequence of n [41], which in turn requires extensive tuning in implementations. By contrast, we derive a data-driven, adaptive temperature schedule endogenously, guided by the regret analysis. Specifically, the update of γ_n is chosen to ensure convergence (which will be shown in the proof of Theorem 2) and determined by the learned critic parameter $\boldsymbol{\theta}_n$ and a predefined deterministic sequence b_n as follows

$$\gamma_n = \frac{c_\gamma \int_0^T k_1(t;\boldsymbol{\theta}_n)\mathrm{d}t}{b_n T}, \quad \text{for } n = 0,1,2,\ldots. \tag{12}$$

Here $c_\gamma > 0$ is selected to be sufficiently large so that $\frac{1}{c_\gamma}$ is smaller than the minimum eigenvalue of $(\sum_{j=1}^m D_j D_j^\top)^{-1}$, and the sequence $\{b_n\}$ is positive and monotonically increasing with $b_n \uparrow \infty$, so that γ_n gradually decreases as

[2] As will be shown in the subsequent sections, the values of $c_1, c_2, c_3, c^{(\boldsymbol{\theta})}, c^{(\gamma)}$ do not affect the order of regret bound. The key requirement here is the boundedness of k_1, k_1', and k_3'. Our analysis uses only the bounds in (11), *not* the explicit functional forms of k_1 and k_3.

more exploration has been done. The normalization term $\frac{\int_0^T k_1(t;\boldsymbol{\theta}_n)\mathrm{d}t}{T}$ couples the temperature update with the current critic scale, which makes γ_n depend on the critic θ_n in the general parameterization because the latter is learned from trajectory data through policy evaluation. This normalization is tailored to our analysis; see the proof of Theorem 2.

Incidentally, it follows from (11) that

$$\gamma_n = \frac{c_\gamma \int_0^T k_1(t;\boldsymbol{\theta}_n)\,\mathrm{d}t}{b_n T} \leq \frac{c_\gamma c_2}{b_n} \to 0 \tag{13}$$

as $n \to +\infty$.

C. Adaptive Actor Exploration

a) Actor Parameterization: Motivated by the structure of the optimal policy in (8), we parameterize the actor using learnable parameters $\boldsymbol{\phi} \in \mathbb{R}^l$ and $\Gamma \in \mathbb{S}^l_{++}$:

$$\pi(u \mid x; \boldsymbol{\phi}, \Gamma) = \mathcal{N}(u \mid \boldsymbol{\phi} x, \Gamma). \tag{14}$$

Write the corresponding entropy as $p^\pi(t) = p(t; \boldsymbol{\phi}, \Gamma)$.

b) Adaptive Actor Exploration: The variance parameter Γ represents the level of exploration controlled by the actor. A higher variance encourages broader exploration, allowing the agent to sample diverse actions, while a lower variance focuses more on exploitation. As will be shown in Sections V and VI, the adaptive iteration derived gradually reduces Γ_n as the learning signal stabilizes, shifting the actor from exploration towards exploitation. In the existing literature [40], [41], Γ_n is chosen to be a deterministic sequence of n. Such a predetermined schedule does not account for the agent's experience and observed data. Introducing an adaptive Γ_n enables a more flexible and efficient exploration strategy.

To achieve this, we apply the continuous-time policy gradient (PG) martingale orthogonality condition of [46, Section 3.2, Theorem 5], which specializes in our case to the following moment condition:

$$\mathbb{E}\bigg[\int_0^T \bigg\{ \frac{\partial \log \pi}{\partial \Gamma}(u(t) \mid x(t); \boldsymbol{\phi}, \Gamma)\,(\mathrm{d}J(t, x(t); \boldsymbol{\theta}, \gamma) - \frac{1}{2}Qx(t)^2\mathrm{d}t + \gamma p(t; \boldsymbol{\phi}, \Gamma)\mathrm{d}t) + \gamma \frac{\partial p}{\partial \Gamma}(t; \boldsymbol{\phi}, \Gamma)\mathrm{d}t \bigg\}\bigg] = 0. \tag{15}$$

To enhance numerical efficiency in the resulting stochastic approximation algorithm, we reparametrize Γ as Γ^{-1}. The chain rule implies that the derivative of Γ^{-1} with respect to Γ is a deterministic and time-invariant term, which can thus be omitted while preserving the validity of (15). Consequently,

$$\mathbb{E}\bigg[\int_0^T \bigg\{ \frac{\partial \log \pi}{\partial \Gamma^{-1}}(u(t) \mid x(t); \boldsymbol{\phi}, \Gamma)\,(\mathrm{d}J(t, x(t); \boldsymbol{\theta}, \gamma) - \frac{1}{2}Qx(t)^2\mathrm{d}t + \gamma p(t; \boldsymbol{\phi}, \Gamma)\mathrm{d}t) + \gamma \frac{\partial p}{\partial \Gamma^{-1}}(t; \boldsymbol{\phi}, \Gamma)\mathrm{d}t \bigg\}\bigg] = 0.$$

The corresponding stochastic approximation algorithm then gives rise to the updating rule for the actor exploration parameter Γ:

$$\Gamma_{n+1} \leftarrow \Gamma_n - a_n^{(\Gamma)} Z_n(T), \tag{16}$$

where $a_n^{(\Gamma)}$ is the learning rate, and

$$Z_n(s) = \int_0^s \bigg\{ \frac{\partial \log \pi}{\partial \Gamma^{-1}}(u_n(t) \mid t, x_n(t); \boldsymbol{\phi}_n, \Gamma_n) \Big[\mathrm{d}J(t, x_n(t); \boldsymbol{\theta}_n, \gamma_n) - \frac{1}{2}Qx_n(t)^2\mathrm{d}t + \gamma_n p(t; \boldsymbol{\phi}_n, \Gamma_n)\,\mathrm{d}t\Big] + \gamma_n \frac{\partial p}{\partial \Gamma^{-1}}(t; \boldsymbol{\phi}_n, \Gamma_n)\,\mathrm{d}t \bigg\}, \tag{17}$$

where $0 \leq s \leq T$.

In sharp contrast with [40], the fact that Γ_n is no longer a deterministic sequence that monotonically decreases to zero introduces significant analytical challenges. In particular, its learning rates must be carefully chosen to establish the convergence and convergence rate of Γ_n; these are done in Section V-A. These properties are essential for guaranteeing that the algorithm ultimately attains a sublinear regret bound, and their implications for overall performance will be discussed in the subsequent sections.

IV. A Continuous-Time RL Algorithm with Data-driven Exploration

This section presents the development of continuous-time RL algorithms with data-driven exploration. We discuss the policy evaluation and policy improvement steps, followed by a projection technique, time discretization, and the final pseudocode.

A. Policy Evaluation

To update the critic parameter $\boldsymbol{\theta}$, we employ policy evaluation (PE), a fundamental component of RL that focuses on learning the value function of a given policy.

Based on the parameterizations of the value function and policy in (10) and (14) respectively, we update $\boldsymbol{\theta}$ by applying the general continuous-time PE algorithm of [47, Section 4.2, Proposition 4], i.e., specializing the martingale-orthogonality conditions therein to our LQ model, which yields

$$\boldsymbol{\theta}_{n+1} \leftarrow \boldsymbol{\theta}_n + a_n^{(\boldsymbol{\theta})} \int_0^T \frac{\partial J}{\partial \boldsymbol{\theta}}(t, x_n(t); \boldsymbol{\theta}_n, \gamma_n) \Big[-\frac{1}{2}Qx_n(t)^2\mathrm{d}t + \mathrm{d}J(t, x_n(t); \boldsymbol{\theta}_n, \gamma_n) + \gamma_n p(t; \boldsymbol{\phi}_n, \Gamma_n)\mathrm{d}t\Big], \tag{18}$$

where $a_n^{(\boldsymbol{\theta})}$ is the corresponding learning rate. We also remark that the above PE method applies equally to definite and indefinite LQ problems and requires no additional treatment for the latter.

B. Policy Improvement

The remaining learnable parameter is the mean of the stochastic Gaussian policy, $\boldsymbol{\phi}$. We again apply the continuous-time PG result of [46, Section 3.2, Theorem 5] to update $\boldsymbol{\phi}$ via stochastic approximation:

$$\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + a_n^{(\boldsymbol{\phi})} Y_n(T), \tag{19}$$

where $a_n^{(\boldsymbol{\phi})}$ is the learning rate and

$$Y_n(s) = \int_0^s \Bigg\{ \frac{\partial \log \pi}{\partial \boldsymbol{\phi}} \left(u_n(t) \mid t, x_n(t); \boldsymbol{\phi}_n, \Gamma_n\right) \Bigg[-\frac{1}{2} Q x_n(t)^2 \mathrm{d}t + \mathrm{d}J\left(t, x_n(t); \boldsymbol{\theta}_n, \gamma_n\right) + \gamma_n p\left(t; \boldsymbol{\phi}_n, \Gamma_n\right) \mathrm{d}t \Bigg] + \gamma_n \frac{\partial p}{\partial \boldsymbol{\phi}} \left(t; \boldsymbol{\phi}_n, \Gamma_n\right) \mathrm{d}t \Bigg\}, \tag{20}$$

where $0 \leq s \leq T$.

C. Projections

The parameter updates for $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, and Γ follow stochastic approximation (SA), a foundational method introduced by [48] and extensively studied in subsequent works [49]–[51]. However, directly applying SA in our setting presents challenges such as extreme state values and unbounded estimation errors, which can destabilize the learning process. To address these issues, we incorporate the projection technique proposed by [52], ensuring that parameter updates remain in a bounded region at every iteration.

Define $\Pi_K(x) := \arg\min_{y \in K} |y - x|^2$, the projection of a point x onto a set K. We now introduce the projection sets for all the critic and actor parameters.

First, we define the projection sets for the critic parameter $\boldsymbol{\theta}$ and the temperature parameter γ:

$$K^{(\boldsymbol{\theta})} = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d \mid |\boldsymbol{\theta}| \leq c^{(\boldsymbol{\theta})} \right\}, K^{(\gamma)} = \left\{ \gamma \in \mathbb{R} \mid |\gamma| \leq c^{(\gamma)} \right\}, \tag{21}$$

which are employed to enforce the boundedness conditions in (11). The update rules incorporating projection for $\boldsymbol{\theta}$ and γ are modified to

$$\boldsymbol{\theta}_{n+1} \leftarrow \Pi_{K^{(\boldsymbol{\theta})}} \Bigg(\boldsymbol{\theta}_n + a_n^{(\boldsymbol{\theta})} \int_0^T \frac{\partial J}{\partial \boldsymbol{\theta}} (t, x_n(t); \boldsymbol{\theta}_n, \gamma_n) \Bigg[\mathrm{d}J\left(t, x_n(t); \boldsymbol{\theta}_n, \gamma_n\right) - \frac{1}{2} Q x_n(t)^2 \mathrm{d}t + \gamma_n p\left(t; \boldsymbol{\phi}_n, \Gamma_n\right) \mathrm{d}t \Bigg] \Bigg), \tag{22}$$

$$\gamma_{n+1} \leftarrow \Pi_{K^{(\gamma)}} \left(\frac{c_\gamma \int_0^T k_1(t; \boldsymbol{\theta}_n) \mathrm{d}t}{b_n T} \right). \tag{23}$$

Next, we define the projection sets for the actor parameters $\boldsymbol{\phi}$ and Γ:

$$K_n^{(\boldsymbol{\phi})} = \left\{ \boldsymbol{\phi}_n \in \mathbb{R}^l \middle| |\boldsymbol{\phi}_n| \leq c_n^{(\boldsymbol{\phi})} \right\}, \quad K_n^{(\Gamma)} = \left\{ \Gamma_n \in \mathbb{S}_{++}^l \middle| |\Gamma_n| \leq c_n^{(\Gamma)}, \Gamma_n - \frac{1}{b_n} I \in \mathbb{S}_{++}^l \right\}, \tag{24}$$

where $\{c_n^{(\boldsymbol{\phi})} \uparrow \infty\}$, $\{c_n^{(\Gamma)} \uparrow \infty\}$, and $\{b_n \uparrow \infty\}$ are positive, monotonically increasing sequences, with b_n the same deterministic sequence in (12). The specifics of these sequences are given in Theorem 1. Clearly, the sequences of sets $K_n^{(\boldsymbol{\phi})}$ and $K_n^{(\Gamma)}$ expand over time, eventually covering the entire spaces $\mathbb{R}^l$ and $\mathbb{S}_{++}^l$, respectively. This ensures that our algorithm remains model-free without needing to have prior knowledge of the model parameters. The update rules with projection for $\boldsymbol{\phi}$ and Γ are now

$$\boldsymbol{\phi}_{n+1} \leftarrow \Pi_{K_{n+1}^{(\boldsymbol{\phi})}} \left(\boldsymbol{\phi}_n + a_n^{(\boldsymbol{\phi})} Y_n(T) \right), \tag{25}$$

$$\Gamma_{n+1} \leftarrow \Pi_{K_{n+1}^{(\Gamma)}} \left(\Gamma_n - a_n^{(\Gamma)} Z_n(T) \right), \tag{26}$$

where $Y_n(T)$ and $Z_n(T)$ are respectively determined by (20) and (17).

D. Discretization

While our RL framework is continuous-time in both formulation and analysis, numerical implementation requires discretization at the final stage. To facilitate this, we divide the time horizon $[0, T]$ into uniform intervals of size Δt_n at the n-th iteration, resulting in discretized rules:[3]

$$\boldsymbol{\theta}_{n+1} \leftarrow \Pi_{K^{(\boldsymbol{\theta})}} \Bigg(\boldsymbol{\theta}_n + a_n^{(\boldsymbol{\theta})} \sum_{k=0}^{\lfloor \frac{T}{\Delta t_n} - 1 \rfloor} \frac{\partial J}{\partial \boldsymbol{\theta}} (t_k, x_n(t_k); \boldsymbol{\theta}_n, \gamma_n) \Bigg[J\left(t_{k+1}, x_n(t_{k+1}); \boldsymbol{\theta}_n, \gamma_n\right) - J\left(t_k, x_n(t_k); \boldsymbol{\theta}_n, \gamma_n\right) - \frac{1}{2} Q x_n(t_k)^2 \Delta t_n + \gamma_n p\left(t_k; \boldsymbol{\phi}_n, \Gamma_n\right) \Delta t_n \Bigg] \Bigg), \tag{27}$$

$$\gamma_{n+1} \leftarrow \Pi_{K^{(\gamma)}} \left(\frac{c_\gamma}{b_n T} \sum_{k=0}^{\lfloor \frac{T}{\Delta t_n} - 1 \rfloor} k_1(t_k; \boldsymbol{\theta}_n) \Delta t_n \right). \tag{28}$$

Moreover, $Y_n(T)$ and $Z_n(T)$ in the update rules (25) and (26) are approximated by

$$\hat{Y}_n(T) = \sum_{k=0}^{\lfloor \frac{T}{\Delta t_n} - 1 \rfloor} \Delta t_n \Bigg\{ \frac{\partial \log \pi}{\partial \boldsymbol{\phi}} \left(u_n(t_k) \mid t_k, x_n(t_k); \boldsymbol{\phi}_n, \Gamma_n\right) \Bigg[\left(J\left(t_{k+1}, x_n(t_{k+1}); \boldsymbol{\theta}_n, \gamma_n\right) - J\left(t_k, x_n(t_k); \boldsymbol{\theta}_n, \gamma_n\right)\right) \frac{1}{\Delta t_n} - \frac{1}{2} Q x_n(t_k)^2 + \gamma_n p\left(t_k; \boldsymbol{\phi}_n, \Gamma_n\right) \Bigg] + \gamma_n \frac{\partial p}{\partial \boldsymbol{\phi}} \left(t_k; \boldsymbol{\phi}_n, \Gamma_n\right) \Bigg\}, \tag{29}$$

$$\hat{Z}_n(T) = \sum_{k=0}^{\lfloor \frac{T}{\Delta t_n} - 1 \rfloor} \Delta t_n \Bigg\{ \frac{\partial \log \pi}{\partial \Gamma^{-1}} \left(u_n(t_k) \mid t_k, x_n(t_k); \boldsymbol{\phi}_n, \Gamma_n\right) \Bigg[\left(J\left(t_{k+1}, x_n(t_{k+1}); \boldsymbol{\theta}_n, \gamma_n\right) - J\left(t_k, x_n(t_k); \boldsymbol{\theta}_n, \gamma_n\right)\right) \frac{1}{\Delta t_n} - \frac{1}{2} Q x_n(t_k)^2 + \gamma_n p\left(t_k; \boldsymbol{\phi}_n, \Gamma_n\right) \Bigg] + \gamma_n \frac{\partial p}{\partial \Gamma^{-1}} \left(t_k; \boldsymbol{\phi}_n, \Gamma_n\right) \Bigg\}. \tag{30}$$

[3] Analysis on discretization errors is covered and discussed in Theorems 2 - 4, and Remark 2.

E. LQ-RL Algorithm Featuring Data-Driven Exploration

The preceding analysis gives rise to the following RL algorithm that integrates adaptive exploration:

Algorithm 1 Data-driven Exploration LQ-RL Algorithm

Input: Initial values of learnable parameters $\boldsymbol{\theta}_0, \boldsymbol{\phi}_0, \gamma_0, \Gamma_0$
for $n = 0$ to N **do**
  Initialize $k = 0$, time $t = t_k = 0$, and state $x_n(t_k) = x_0$.
  while $t < T$ **do**
    Sample action $u_n(t_k)$ from policy by (Eq. (14)).
    Update state $x_n(t_{k+1})$ by LQ dynamics (Eq. (1)).
    Update time: $t_{k+1} \leftarrow t_k + \Delta t$, and set $t \leftarrow t_{k+1}$.
  end while
  Collect trajectory data: $\{(t_k, x_n(t_k), u_n(t_k))\}_{k\geq 0}$.
  Update value function parameters $\boldsymbol{\theta}_{n+1}$ by (Eq. (27)).
  Update policy mean parameters $\boldsymbol{\phi}_{n+1}$ by (Eq. (29)).
  Update critic exploration parameter γ_{n+1} by (Eq. (28)).
  Update actor exploration parameter Γ_{n+1} by (Eq. (30)).
end for
Output: Final parameters $\boldsymbol{\theta}_N, \boldsymbol{\phi}_N, \gamma_N, \Gamma_N$.

Algorithm 1 is on-policy: at iteration n, the target policy π_n is updated based on data generated by π_n itself. Off-policy variants within the same continuous-time actor–critic framework are also possible. For example, in the mean–variance portfolio selection application [53] one learns a deterministic policy using data generated by Gaussian exploratory policies. More general off-policy learning theory has been established in [54] for continuous-time q-learning.

V. Regret Analysis

This section contains the main contribution of this work: establishing a sublinear regret bound for the LQ-RL algorithm with data-driven exploration, presented in Algorithm 1. We prove that the algorithm achieves a regret bound of $O(N^{\frac{3}{4}})$ (up to a logarithmic factor), matching the best-known model-free bound for this class of continuous-time LQ problems as reported in [40].

To quickly recap the major differences from [40]: 1) the actor exploration parameter Γ_n therein is a deterministic monotonically decreasing sequence of n while we apply policy gradient for updating Γ_n; 2) the static temperature parameter γ is used in [40], whereas in our algorithm γ is adaptively updated; 3) [40] only considers the case when the initial state is nonzero ($x_0 \neq 0$), and our analysis removes this assumption. These differences beg for a substantially different theoretical analysis compared with [40]. In particular, the stochastic update of Γ_n calls for proving the corresponding almost-sure convergence and explicit convergence rates (which feed into the regret bound), and the case of $x_0 = 0$ imposes additional constraints on the hyperparameters rendering more technical difficulties in both the convergence and regret analysis.

In the remainder of the paper, we use c (and its variants) to denote generic positive constants. These constants depend solely on the model parameters $A, B, C_j, D_j, Q, H, x_0, T, \gamma, m, l$, and their values may vary from one instance to another.

A. Convergence Analysis on γ_n and Γ_n

To start, we explore the almost sure convergence of the critic and actor exploration parameters γ_n and Γ_n, together with the convergence rate of Γ_n in terms of the mean-squared error (MSE). Throughout the rest of the paper, for real numbers a, b we write $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$.

Theorem 1: Consider Algorithm 1 with fixed positive hyperparameters c_1, c_2, c_3 and a sufficiently large fixed constant c_γ. Define the sequences:

$$a_n^{(\Gamma)} = a_n^{(\boldsymbol{\phi})} = \frac{\alpha^{\frac{3}{4}}}{(n+\beta)^{\frac{3}{4}}}, \quad b_n = 1 \vee \frac{(n+\beta)^{\frac{1}{4}}}{\alpha^{\frac{1}{4}}},$$
$$c_n^{(\boldsymbol{\phi})} = 1 \vee (\log\log n)^{\frac{1}{6}}, c_n^{(\Gamma)} = 1 \vee \log n, \Delta t_n = T(n+1)^{-\frac{5}{8}},$$

where $\alpha > 0$ and $\beta > 0$ are constants. Then, the following results hold:

(a) As $n \to \infty$, γ_n converges almost surely to 0, and Γ_n converges almost surely to the zero matrix $\mathbf{0}$.

(b) For all n, $\mathbb{E}[|\Gamma_n|^2] \leq c\frac{(\log n)^{p_2}(\log\log n)^{\frac{4}{3}}}{n^{\frac{1}{2}}}$, where c and p_2 are positive constants.

These results guarantee that the adaptive exploration parameters γ_n and Γ_n diminish almost surely, and the convergence rate of Γ_n is essential for the proof of the regret bound. The remainder of this subsection is devoted to the proof of Theorem 1.

Define the conditional expectation of $Z_n(T)$ given the current parameter iterates as

$$h^{(\Gamma)}(\boldsymbol{\phi}_n, \Gamma_n; \boldsymbol{\theta}_n, \gamma_n) = \mathbb{E}[Z_n(T) \mid \boldsymbol{\theta}_n, \boldsymbol{\phi}_n, \gamma_n, \Gamma_n],$$

as well as the deviation

$$\xi_n^{(\Gamma)} = Z_n(T) - h^{(\Gamma)}(\boldsymbol{\phi}_n, \Gamma_n; \boldsymbol{\theta}_n, \gamma_n). \tag{31}$$

The update rule (26) can be rewritten as

$$\Gamma_{n+1} = \Pi_{K_{n+1}^{(\Gamma)}}\left(\Gamma_n - a_n^{(\Gamma)}[h^{(\Gamma)}(\boldsymbol{\phi}_n, \Gamma_n; \boldsymbol{\theta}_n, \gamma_n) + \xi_n^{(\Gamma)}]\right). \tag{32}$$

The proof of Theorem 1 is complex and will be subsequently broken into several steps for the reader's convenience. Here let us highlight the key components of the proof. First, we analyze the variance of the increment $Z_n(T)$ to ensure that the stochastic fluctuations along iterations remain controlled (Section V-A.1). Second, we study the conditional mean, $h^{(\Gamma)}(\boldsymbol{\phi}_n, \Gamma_n; \boldsymbol{\theta}_n, \gamma_n)$, and obtain its bound (Section V-A.2). Third, combining these results with a projected stochastic-approximation argument, we establish the almost sure convergence of Γ_n (Section V-A.3). Finally, we derive an iterative inequality for $\mathbb{E}[|\Gamma_n|^2]$ and prove the desired convergence rate by mathematical induction (Section V-A.4).

Now we present the proof of Theorem 1.

1) A Variance Upper Bound: The following result establishes an upper bound on the variance of the increment $Z_n(T)$.

Lemma 1: There exists a constant $c > 0$ depending solely on the model parameters such that

$$\begin{aligned}&\mathrm{Var}\left(Z_n(T)\middle|\boldsymbol{\theta}_n, \boldsymbol{\phi}_n, \gamma_n, \Gamma_n\right)\\ \leq& c\left(1 + |\boldsymbol{\phi}_n|^8 + |\Gamma_n|^8 + (\log b_n)^8 + \gamma_n^8\right)\exp\{c|\boldsymbol{\phi}_n|^4\}.\end{aligned} \tag{33}$$

Proof: The proof is similar to that of [40, Lemma B.2] except that we need to account for the estimates on Γ_n; so we will be brief here. It follows from (17) together with Ito's lemma (applied to $J(t, x_n(t); \boldsymbol{\theta}_n, \gamma_n)$) that

$$\mathrm{d}Z_n(t) \triangleq Z_n^{(1)}(t)\mathrm{d}t + \sum_{j=1}^{m} Z_n^{(2,j)}(t)\mathrm{d}W_n^{(j)}(t), \tag{34}$$

for which we have the following estimates

$$\mathbb{E}[|Z_n^{(1)}(t)|^2|\boldsymbol{\theta}_n, \boldsymbol{\phi}_n, \gamma_n, \Gamma_n, x_n(t)] \le c(1+|\Gamma_n|^4)(1+x_n(t)^4 + |\boldsymbol{\phi}_n|^4 x_n(t)^4 + \gamma_n^2 + \gamma_n^2(\log(\det(\Gamma_n)))^2),$$

$$\mathbb{E}[|Z_n^{(2,j)}(t)|^2|\boldsymbol{\theta}_n, \boldsymbol{\phi}_n, \gamma_n, \Gamma_n, x_n(t)] \le c(1+|\Gamma_n|^3)(1+ x_n(t)^4 + |\boldsymbol{\phi}_n|^2 x_n(t)^4 + |\Gamma_n|^2 x_n(t)^4 + |\boldsymbol{\phi}_n|^2|\Gamma_n|^2 x_n(t)^2).$$

Taking expectations in the above conditional on $x_n(t)$ and applying [40, Lemma B.1], we get

$$\begin{aligned}&\mathbb{E}[|Z_n^{(1)}(t)|^2 + \sum_{j=1}^{m}|Z_n^{(2,j)}(t)|^2|\boldsymbol{\theta}_n, \boldsymbol{\phi}_n, \gamma_n, \Gamma_n] \\ \le& c\left(1+|\boldsymbol{\phi}_n|^8 + |\Gamma_n|^8 + (\log b_n)^8 + \gamma_n^8\right)\exp\{c|\boldsymbol{\phi}_n|^4\}.\end{aligned} \tag{35}$$

This establishes the desired inequality. ∎

2) A Mean Increment: We now analyze the mean increment $h^{(\Gamma)}(\boldsymbol{\phi}_n, \Gamma_n; \boldsymbol{\theta}_n, \gamma_n)$ in the updating rule (32). Integrating and taking conditional expectation in (34), the stochastic integral term vanishes; so only $Z_n^{(1)}$ contributes to the mean increment. Moreover, since the control is sampled from a Gaussian distribution, all the required moments of the control can be computed explicitly, yielding

$$\begin{aligned}&h^{(\Gamma)}(\Gamma_n; \boldsymbol{\theta}_n, \gamma_n)\\ =&\frac{1}{2}\int_0^T k_1(t;\boldsymbol{\theta}_n)\mathrm{d}t\Gamma_n\left(\sum_j D_j D_j^\top\right)\Gamma_n - \frac{\gamma_n T}{2}\Gamma_n.\end{aligned} \tag{36}$$

Given that $h^{(\Gamma)}(\Gamma_n; \boldsymbol{\theta}_n, \gamma_n)$ is quadratic in Γ_n, we apply (11) to establish the bound:

$$\left|h^{(\Gamma)}(\Gamma_n; \boldsymbol{\theta}_n, \gamma_n)\right| \le c(1+|\Gamma_n|^2+\gamma_n^2). \tag{37}$$

3) Almost Sure Convergence of γ_n and Γ_n: We now prove Part (a) of Theorem 1. Indeed we will present and prove a more general result that involves a bias term $\beta_n^{(\Gamma)} = \mathbb{E}[\xi_n^{(\Gamma)} \mid \mathcal{G}_n]$ (where $\mathcal{G}_n$ is defined in Theorem 2), which captures implementation errors such as the discretization error (see Remark 2). When $\beta_n^{(\Gamma)} = \mathbf{0}$, the result simplifies to Theorem 1-(a). Moreover, by the definition of $\xi_n^{(\Gamma)}$ in (31) and Lemma 1, the conditional expectation and moment assumptions on $\xi_n^{(\Gamma)}$ stated in Theorem 2 are automatically satisfied.

Theorem 2: Assume $\xi_n^{(\Gamma)}$ satisfies $\mathbb{E}\left[\xi_n^{(\Gamma)}\middle|\mathcal{G}_n\right] = \beta_n^{(\Gamma)}$ and

$$\begin{aligned}&\mathbb{E}\left[\left|\xi_n^{(\Gamma)} - \beta_n^{(\Gamma)}\right|^2\middle|\mathcal{G}_n\right]\\ \le& c(1+|\boldsymbol{\phi}_n|^8 + |\Gamma_n|^8 + (\log b_n)^8 + \gamma_n^8)\exp\{c|\boldsymbol{\phi}_n|^4\},\end{aligned} \tag{38}$$

where $c>0$ is a constant independent of n, and $\{\mathcal{G}_n\}$ is the filtration generated by $\{\boldsymbol{\theta}_m, \boldsymbol{\phi}_m, \gamma_m, \Gamma_m; m = 0, 1, 2, ..., n\}$. Moreover, assume

$$\begin{aligned}&(i) 0 < a_n^{(\Gamma)} \le 1\ \forall n, \sum_n a_n^{(\Gamma)} = \infty, \sum_n a_n^{(\Gamma)}|\beta_n^{(\Gamma)}| < \infty;\\ &(ii) c_n^{(\boldsymbol{\phi})} \uparrow \infty, c_n^{(\Gamma)} \uparrow \infty, \text{ and for } 0 \le q_1, q_2, q_3, 0 \le q_4 \le 4,\\ &\sum (a_n^{(\Gamma)})^2 (c_n^{(\Gamma)})^{q_1}(c_n^{(\boldsymbol{\phi})})^{q_2}(\log b_n)^{q_3} e^{c(c_n^{(\boldsymbol{\phi})})^{q_4}} < \infty;\\ &(iii) b_n \ge 1, b_n \uparrow \infty, \sum \frac{a_n^{(\Gamma)}}{b_n} = \infty, \sum |\frac{1}{b_n} - \frac{1}{b_{n+1}}| < \infty.\end{aligned} \tag{39}$$

Then, $\gamma_n \xrightarrow{a.s.} 0$ and $\Gamma_n \xrightarrow{a.s.} \mathbf{0}$.

The convergence of γ_n is obvious so one needs to focus on that of Γ_n. By viewing its iteration as a projected stochastic-approximation scheme and adapting the almost sure convergence analysis of [52] under relaxed conditions, we obtain the convergence. A full proof is relegated to Appendix I.

Remark 1: A case satisfying the assumptions in (39) is when $a_n^{(\Gamma)} = 1 \wedge \frac{\alpha^{\frac{3}{4}}}{(n+\beta)^{\frac{3}{4}}}$ and $b_n = 1 \vee \frac{(n+\beta)^{\frac{1}{4}}}{\alpha^{\frac{1}{4}}}$, where constants $\alpha > 0, \beta > 0$. $c_n^{(\boldsymbol{\phi})} = 1 \vee (\log\log n)^{\frac{1}{6}}, c_n^{(\Gamma)} = 1 \vee \log n, \beta_n^{(\boldsymbol{\phi})} = 0$. This is because $\sum \frac{1}{n} = \infty$; $\sum \frac{(\log n)^p (\log\log n)^q}{n^r} < \infty$, for any $p, q > 0$ and $r > 1$; and $\sum |n^{-\frac{1}{4}} - (n+1)^{-\frac{1}{4}}| < \frac{1}{4}\sum n^{-\frac{5}{4}} < \infty$ (by the mean–value theorem). Additionally, it is interesting to note that the assumptions $\sum \frac{a_n^{(\Gamma)}}{b_n} = \infty$ and $\sum |\frac{1}{b_n} - \frac{1}{b_{n+1}}| < \infty$ are new compared to [40]. This is due to the implementation of data-driven exploration.

4) Convergence Rate of Γ_n: We need the following lemmas for the proof of Theorem 3, now that Γ_n is updated from data rather than following a deterministic schedule as in [40]. Specifically, Lemma 2 is used to deal with the term b_n in (47), while Lemma 3 provides the order of $|G_n|^2 = |\Gamma_n^* - \Gamma_{n+1}^*|^2$.

Lemma 2: For any $W > 0$ and any $0 < q < 1$, there exist positive numbers $\alpha > \frac{1}{W}$ and $\beta \ge \max(\frac{1}{W\alpha - 1}, W^{\frac{2q-1}{1-q}}\alpha^{\frac{q}{1-q}})$ such that the sequences $\hat{a}_n = \frac{\alpha}{n+\beta}$ and $a_n = \frac{\alpha^q}{(n+\beta)^q}$ satisfy $\hat{a}_n \le \hat{a}_{n+1}(1+W\hat{a}_{n+1})$ and $a_n \le a_{n+1}(1+Wa_{n+1})$ for all $n \ge 0$.

Proof: First,

$$\hat{a}_n \le \hat{a}_{n+1}(1+W\hat{a}_{n+1}) \Leftrightarrow n+1+\beta \le W\alpha n + W\alpha\beta,$$

the latter being true for any n when $\alpha > \frac{1}{W} > 0$, $\beta \ge \frac{1}{W\alpha - 1} > 0$.

Next, we have

$$\begin{aligned}&a_n \le a_{n+1}(1+Wa_{n+1})\\ \Leftrightarrow&(n+\beta+1)^q - (n+\beta)^q \le W\alpha^q\left(\frac{n+\beta}{n+\beta+1}\right)^q.\end{aligned} \tag{40}$$

For the latter inequality in (40), notice that the left-hand side decreases in n while the right-hand side increases in n. So to show that this inequality is true for all n, it is sufficient to show that it is true when $n = 0$, which is $(\beta+1)^q - \beta^q \le W\alpha^q \frac{\beta^q}{(\beta+1)^q}$.

Now, for $\beta \ge \frac{1}{W\alpha - 1}$, we have $\beta + 1 \le W\alpha\beta$. On the other hand, the mean–value theorem yields $(\beta+1)^q - \beta^q \le q\beta^{q-1}$. Hence,

$$W^{\frac{2q-1}{1-q}}\alpha^{\frac{q}{1-q}} \le \beta \Rightarrow (\beta+1)^q - \beta^q \le W\alpha^q\frac{\beta^q}{(\beta+1)^q}.$$

∎

Lemma 3: Let the sequence $\hat{G}_n = n^{-q} - (n+1)^{-q} > 0$ for some $q > 0$. Then $\hat{G}_n^2 = O(n^{-2q-2})$.

Proof: We have

$$\hat{G}_n^2 = \left[\frac{(n+1)^q - n^q}{(n+1)^q n^q}\right]^2 < \left[\frac{(n+1)^q - n^q}{n^{2q}}\right]^2.$$

Applying the mean–value theorem to the function $f(x) = x^q$ and noting that $f'(x) = qx^{q-1}$ is a decreasing function for $x > 0$, we conclude

$$\left[\frac{(n+1)^q - n^q}{n^{2q}}\right]^2 < \frac{(qn^{q-1})^2}{n^{4q}} = q^2 n^{-2q-2}.$$

■

The following theorem specializes to Theorem 1-(b) when the bias term $\beta_n^{(\Gamma)} = 0$.

Theorem 3: Under the same setting of Theorem 2, if the sequence $\{\hat{a}_n\}$, defined as $\{\frac{a_n^{(\Gamma)}}{b_n}\}$, satisfies

$$\hat{a}_n \le \hat{a}_{n+1}(1 + w\hat{a}_{n+1})$$

for some sufficiently small constant $w > 0$, and the sequences $\{\frac{b_n^3}{a_n^{(\Gamma)}}|\beta_n^{(\Gamma)}|^2\}$ and $\{\frac{b_n^3|G_n|^2}{(a_n^{(\Gamma)})^3}\}$ are non-decreasing in n, then there exists an increasing sequence $\{\hat{\eta}_n^{(\Gamma)}\}$ and a constant $c' > 0$ such that

$$\mathbb{E}[|\Gamma_{n+1} - \Gamma_{n+1}^*|^2] \le c'\hat{a}_n\hat{\eta}_n^{(\Gamma)}.$$

In particular, if the parameters $b_n, a_n^{(\Gamma)}, c_n^{(\phi)}, c_n^{(\Gamma)}, \beta_n^{(\Gamma)}$ are set as in Remark 1, then

$$\mathbb{E}[|\Gamma_n|^2] \le c\frac{(e \vee \log n)^{p_2}(1 \vee \log\log n)^{\frac{4}{3}}}{n^{\frac{1}{2}}}$$

where p_2 is the same constant appearing in Theorem 2.

The proof of this theorem is lengthy and here we highlight the key ideas while providing the full proof in Appendix II. Consider the squared error $\rho_n := \mathbb{E}[|\Gamma_n - \Gamma_n^*|^2]$ and use the recursion for Γ_n to derive a recursive inequality of the form $\rho_{n+1} \le (1 - c\hat{a}_n)\rho_n + 4\hat{a}_n^2\hat{\eta}_n^{(\Gamma)}$ for a suitably defined step-size sequence $\{\hat{a}_n\}$ and auxiliary sequence $\{\hat{\eta}_n^{(\Gamma)}\}$. Applying the monotonicity of these sequences and adapting the stochastic-approximation rate analysis in [55] yield the asserted bound.

B. Convergence Analysis of ϕ_n

In this subsection, we investigate the convergence properties of the actor parameter $\boldsymbol{\phi}_n$ for both cases $x_0 \neq 0$ and $x_0 = 0$.

Theorem 4: Under the same setting of Theorem 1, we have

(a) As $n \to \infty$, $\boldsymbol{\phi}_n$ converges almost surely to $\boldsymbol{\phi}^* = -\left(\sum_{j=1}^m D_j D_j^\top\right)^{-1}\left(B + \sum_{j=1}^m C_j D_j\right)$.

(b) The expected squared error satisfies

$$\mathbb{E}[|\boldsymbol{\phi}_n - \boldsymbol{\phi}^*|^2] \le c\frac{(\log n)^{p_1}(\log\log n)^{\frac{4}{3}}}{n^{\frac{1}{2}}}, \quad \text{if } x_0 \neq 0,$$

$$\mathbb{E}[|\boldsymbol{\phi}_n - \boldsymbol{\phi}^*|^2] \le c\frac{(\log n)^{p_1'}(\log\log n)^{\frac{4}{3}}}{n^{\frac{1}{4}}}, \quad \text{if } x_0 = 0,$$

where c, p_1, and p_1' are positive constants only dependent on the model parameters.

A proof proceeds by separating the cases of $x_0 \neq 0$ and $x_0 = 0$. When $x_0 \neq 0$, one adapts the projected stochastic-approximation framework used in [40, Theorem 4.1] to our setting. When $x_0 = 0$, one applies the core ideas from the proof of Theorem 1. The full proof is provided in Appendix III.

The established convergence of the learned policy underpins the regret analysis of Algorithm 1. Theorem 4 posits that the MSE of $\boldsymbol{\phi}_n$ depends on whether the initial state x_0 is zero or nonzero; yet both cases ultimately lead to the same sublinear regret bound, which will be shown in the next subsection.

C. Regret Bound

The regret quantifies the difference between the value function of the learned policy and that of the oracle optimal policy. We say that a regret bound is sublinear if the regret grows strictly slower than linearly in the number of iterations, N. The average regret over N iterations of an algorithm with sublinear regret converges to 0 as $N \to \infty$, yielding an almost oracle performance in the long run.

For a stochastic Gaussian policy $\pi = \mathcal{N}(\cdot|\boldsymbol{\phi}x, \Gamma)$, we denote

$$\bar{J}(\boldsymbol{\phi}, \Gamma) = \mathbb{E}\Big[\int_0^T \left(-\frac{1}{2}Qx^\pi(s)^2\right)\mathrm{d}s - \frac{1}{2}Hx^\pi(T)^2\Big|x^\pi(0) = x_0\Big]. \tag{41}$$

This function evaluates policy $\mathcal{N}(\cdot|\boldsymbol{\phi}x, \Gamma)$ under the original objective *without* entropy regularization. The oracle value of the original problem is $\bar{J}(\boldsymbol{\phi}^*, \mathbf{0})$.

Theorem 5: Under the same setting of Theorem 1, Algorithm 1 leads to the following regret bounds over N iterations:

$$\begin{aligned}
&\sum_{n=1}^N \mathbb{E}\big[\bar{J}(\boldsymbol{\phi}^*, \mathbf{0}) - \bar{J}(\boldsymbol{\phi}_n, \Gamma_n)\big] \\
&\le c + cN^{\frac{3}{4}}(\log N)^{\frac{p_1+p_2}{2}+1}(\log\log N), \quad \text{if } x_0 \neq 0, \\
&\sum_{n=1}^N \mathbb{E}\big[\bar{J}(\boldsymbol{\phi}^*, \mathbf{0}) - \bar{J}(\boldsymbol{\phi}_n, \Gamma_n)\big] \\
&\le c + cN^{\frac{3}{4}}(\log N)^{\frac{p_1'+p_2}{2}\vee(p_1'+\frac{3}{2})}(\log\log N), \quad \text{if } x_0 = 0,
\end{aligned}$$

where $c > 0$ is a constant independent of N, and p_1, p_1', p_2 are the same constants given in Theorems 1 and 4.

We defer the full proof to Appendix IV, while outlining the key ideas here. Using a Feynman–Kac representation, we express the value function under the Gaussian policy as a smooth function of (ϕ, Γ) and decompose the one-step regret into a small number of terms associated with the critic and actor errors, treating the cases $x_0 \neq 0$ and $x_0 = 0$ separately. Each term is then bounded using Hölder- and Cauchy–Schwarz-type inequalities together with the mean-squared convergence rates from Theorems 1 and 4. Summing these bounds yields the stated $O(N^{3/4})$ regret rate.

Remark 2: Discretization errors arising from the time step size Δt_n are captured by the terms $\beta_n^{(\phi)}$ and $\beta_n^{(\Gamma)}$. Theorems 2 - 4 suggest that $\beta_n^{(\phi)}$ and $\beta_n^{(\Gamma)}$ should be set to decrease at the order of $n^{-\frac{3}{8}}$, which can be achieved by setting $\Delta t_n = T(n+1)^{-\frac{5}{8}}$. We omit the details here, but refer to [40], [41], [56] for discussions on time-discretization analyses.

VI. Numerical Experiments

This section reports four numerical experiments to evaluate the performance of our proposed exploratory adaptive LQ-RL algorithm. The first one validates our theoretical findings by examining the convergence behaviors of ϕ and Γ and confirming the sublinear regret bound. The second one compares our model-free approach with a model-based benchmark [41] adapted to our setting that incorporates state- and control-dependent volatility. The third one evaluates the effect of exploration strategies by comparing our approach to a deterministic exploration schedule [40]. We examine scenarios where exploration parameters are either overly conservative or excessively aggressive, demonstrating the advantages of dynamically adjusting exploration based on observed data rather than relying on predefined schedules. The final experiment extends the third one by introducing random model parameters and random initial exploration parameter values, capturing real-world situations where no prior knowledge of the environment is available. This last experiment further demonstrates the effectiveness of the data-driven exploration mechanism in adapting to the environment.

For all the experiments, we set the control dimension and Brownian motion dimension to be $l = 1$ and $m = 1$. Under this setting, the LQ dynamics are simplified to

$$\mathrm{d}x^u(t) = (Ax^u(t) + Bu(t))\mathrm{d}t + (Cx^u(t) + Du(t))\mathrm{d}W(t). \tag{42}$$

Moreover, we set all the model parameters A, B, C, D, Q, H, x_0, and T to be 1 and use a time step of $\Delta t = 0.01$ for the first three experiments. See also Appendix V for the replicability of our experiments.

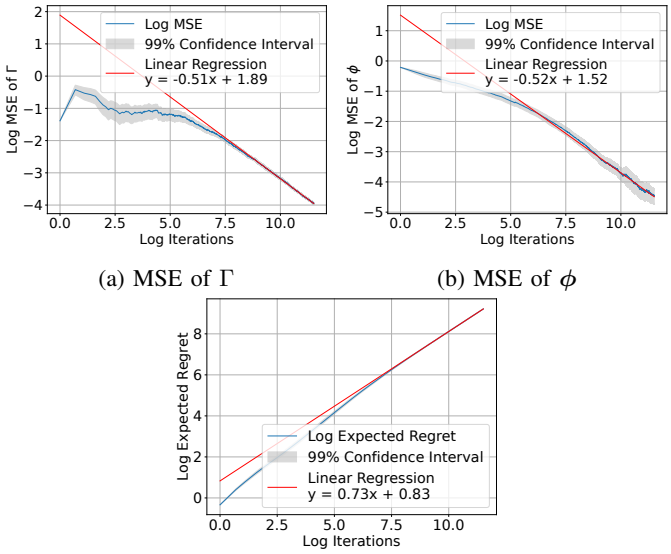
A. Numerical Validation of Theoretical Results

To validate the theoretical results on the convergence rates and regret bounds, we focus on the actor parameters Γ and ϕ. We conduct 100 independent runs, each consisting of 100,000 iterations to capture long-term behavior. Convergence rates and regret are plotted in Figure 1 using a log-log scale, where the logarithm of the MSE (for convergence) or regret is plotted against the logarithm of the number of iterations.

The theoretical convergence rates in Theorems 1 and 4 have both orders of $n^{-1/2}$, while the regret bound in Theorem 5 is of order $N^{3/4}$, all up to logarithmic factors. On a log–log plot, a rate of order n^{α} appears as an approximately straight line with slope α. With linear regression, which is commonly used to estimate empirical convergence rates or regrets (see, e.g., [57], [58]), Figures 1a and 1b show slopes of about -0.51 and -0.52 for Γ_n and ϕ_n, respectively, and Figure 1c indicates a regret slope of about 0.73, all of which are close to the theoretical values $-\frac{1}{2}$ and $\frac{3}{4}$. Therefore, these numerical results are consistent with the theoretical analysis and confirm the long-term effectiveness of our learning algorithm.

B. Model-Free vs. Model-Based

Under the same setting, we compare our model-free LQ-RL algorithm with data-driven exploration to the recently developed model-based method with a deterministic exploration schedule [41], which estimates parameters A and B in the drift term under the assumption of a constant volatility. To cover the settings with state- and control-dependent volatilities, [40] modified the algorithm in [41] by incorporating estimations of also C and D, which we adopt in our experiment.



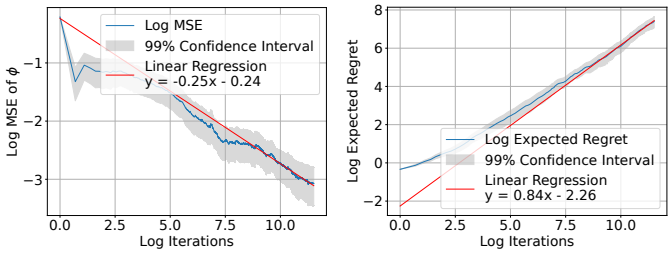
(a) MSE of Γ (b) MSE of ϕ

(c) Regret

Fig. 1: **Log-log plot of Algorithm 1.**

Method	MSE slope of ϕ_n	Regret slope
Model-free (this paper)	-0.52	0.73
Model-based benchmark	-0.25	0.84

TABLE I: Empirical slopes (log–log) of the MSE of ϕ_n and the regret for the model-free and model-based algorithms.



(a) MSE of ϕ in model-based algorithm (b) Regret of model-based algorithm

Fig. 2: **Log-log plot of model-based LQ-RL algorithm with fixed exploration schedule.**

Figure 2a shows that the model-based counterpart exhibits a significantly slower convergence rate for ϕ with a slope of -0.25. Moreover, Figure 2b indicates a worse regret with a slope of 0.84, considerably larger than that of ours.[4]

These differences are also summarized in Table I for a direct comparison.

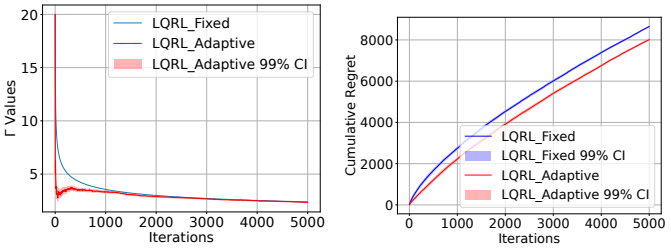
[4] Γ follows a fixed schedule in the model-based benchmark; hence its plot is uninformative and omitted.

C. Adaptive vs. Fixed Explorations

Now, we compare two model-free continuous-time LQ-RL algorithms: ours with data-driven exploration (LQRL_Adaptive) and the one by [40] with fixed exploration schedules (LQRL_Fixed). Both algorithms follow the same fundamental framework, with the key distinction being how exploration is carried out. In Section III-A, we outlined several drawbacks associated with deterministic exploration schedules. To illustrate them more concretely, we analyze two scenarios where a pre-determined exploration is either excessive or insufficient.

For comparison, we examine the trajectories of the actor exploration parameter Γ and the regrets over iterations.[5] Both algorithms share exactly the same experimental setup, and we conduct 1,000 independent runs to observe the overall performances.

a) Excessive Exploration with a Near-Optimal Policy: We first consider a scenario where the policy parameter is initialized at $\phi_0 = -1.8$, which is close to its optimal value $\phi^* = -2$. As the initial policy is already nearly optimal, exploration is not highly necessary. However, the initial exploration levels by both the critic and actor are set excessively high at $\gamma_0 = \Gamma_0 = 20$.



(a) Trajectory of Γ (b) Regret

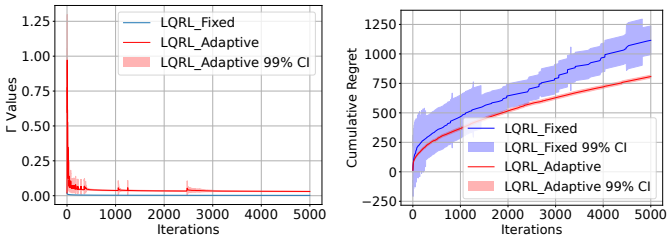
Fig. 3: Comparison of exploration level and regret under excessive initial exploration.

Figure 3a shows that while both algorithms start with an unnecessarily high level of Γ, LQRL_Adaptive rapidly adjusts Γ downward, having effectively learned that exploration is less needed in this setting. By contrast, LQRL_Fixed follows a predetermined decay in Γ and remains at a consistently higher level. As a result, Figure 3b demonstrates that LQRL_Adaptive achieves lower regret, indicating a more efficient learning process by dynamically adapting exploration to the circumstances.

b) Insufficient Exploration with a Poor Policy: We now examine an opposite scenario where ϕ_0 starts at 0, relatively far away from its optimal value $\phi^* = -2$. The initial exploration parameters are set too low at $\gamma_0 = \Gamma_0 = 0.02$, creating a situation where increased exploration is crucial for effective learning.

Figure 4a shows that LQRL_Adaptive rapidly increases Γ from its initial low value of 0.02 to nearly 1, maintaining a higher exploration level than LQRL_Fixed throughout. By contrast, LQRL_Fixed follows a predetermined *decaying* schedule, reducing Γ even further over iterations, which is

[5]The temperature parameter γ is not compared, as it remains fixed in [40].



(a) Trajectory of Γ (b) Regret

Fig. 4: Comparison of exploration level and regret under insufficient initial exploration.

counterproductive in this case where greater exploration is called for. As a result, Figure 4b demonstrates an even larger performance gap in regret compared to the previous scenario, further underscoring the benefits of adaptively adjusting exploration for learning.

D. Adaptive vs. Fixed Explorations: Randomized Model Parameters

In the final experiment, we evaluate the robustness of the comparison between LQRL_Adaptive and LQRL_Fixed in a setting where model parameters and initial exploration levels are randomly generated, mimicking real-world scenarios with no prior knowledge of the environment and no pre-tuning of exploration parameters. Specifically, for each simulation run, the model parameters A, B, C, and D are sampled from a uniform distribution $\mathcal{U}(-5, 5)$, while the initial exploration levels of γ_0 and Γ_0 follow $\mathcal{U}(0, 5)$. The initial policy parameter is set to be $\phi_0 = 0$ as a neutral starting point. To ensure the reliability of the experiment, we conduct $10,000$ independent runs.

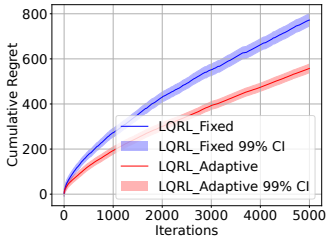


Fig. 5: Comparison of regrets under randomized environments.

Figure 5 demonstrates that LQRL_Adaptive significantly outperforms LQRL_Fixed, with the performance gap widening as the number of iterations increases. While the former is expected to be more computationally costly due to multiple adaptive update steps, its wall-clock runtime was less than 5% higher than that of the latter in this experiment (approximately 4h25 vs. 4h13 on our machine). So the extra updates of γ_n and Γ_n incur negligible computational overhead vis-à-vis the substantial performance gain.

VII. Conclusions

This paper is a continuation of [40], aiming to develop a model-free continuous-time RL framework for a class of

indefinite stochastic LQ problems with state- and control-dependent volatilities. A key contribution, compared to [40], is the introduction of a data-driven exploration mechanism that adaptively adjusts both the temperature parameter controlled by the critic and the stochastic policy variance managed by the actor. This approach overcomes the limitations of fixed exploration schedules, which often require extensive tuning and incur unnecessary exploration costs. While the theoretically proved sublinear regret bound of $O(N^{\frac{3}{4}})$ is the same as that in [40], numerical experiments confirm that our adaptive exploration strategy significantly improves learning efficiency compared to [40].

Research on *model-free* continuous-time RL is still in the very early innings. To our best knowledge [40], [59] are the only works on regret analysis in the realm, largely because the problem is extremely challenging. We focus on the same LQ problem in [40] because this is the problem that we are able to solve at the moment, one nevertheless that may serve as a starting point for tackling more complex problems. Meanwhile, it remains an interesting open question whether adaptive exploration strategies also theoretically improve the regret bound, which is observed from our numerical experiments. Besides serving as a theoretical benchmark, the class of LQ problems studied here arises naturally in several applied domains. A prominent example is continuous-time mean–variance portfolio selection [53], where the state variable, namely the wealth process, is inherently one-dimensional and the problem is automatically of an LQ structure. Other examples include surplus or cash-flow management in finance and production or inventory control in manufacturing. In those cases, the proposed model-free actor–critic method can be applied directly using sampled and observed trajectories, without requiring model identification. Finally, our framework relies on entropy regularization and policy variance for exploration, which may be enhanced by goal-based constraints/rewards to improve sampling efficiency. All these provide promising avenues for future research, driving further advancements in data-driven exploration and continuous-time RL.

APPENDIX I
PROOF OF THEOREM 2

Proof: It is straightforward to see that γ_n almost surely converges to 0. To prove the almost sure convergence of Γ_n, the key idea is to establish inductive upper bounds on $|\Gamma_n|^2$, namely to bound $|\Gamma_{n+1}|^2$ by a function of $|\Gamma_n|^2$.

Define the deviation term $U_n^{(\Gamma)} = \Gamma_n - \Gamma_n^*$, where Γ_n^* is given by:

$$\Gamma_n^* = \frac{\gamma_n T}{\int_0^T k_1(t;\boldsymbol{\theta}_n)\mathrm{d}t}(\sum_{j=1}^m D_j D_j^\top)^{-1}. \tag{43}$$

By the adaptive temperature parameter γ_n in (12), Γ_n^* can be rewritten as:

$$\Gamma_n^* = \frac{c_\gamma}{b_n}(\sum_{j=1}^m D_j D_j^\top)^{-1}, \tag{44}$$

where we recall $c_\gamma > 0$ is such that $\frac{1}{c_\gamma}$ is smaller than the minimum eigenvalue of $(\sum_{j=1}^m D_j D_j^\top)^{-1}$, denoted as $\lambda_{\min}((\sum_{j=1}^m D_j D_j^\top)^{-1})$. Thus $\Gamma_n^* > \frac{1}{b_n} I$. Since $b_n \uparrow \infty$, Γ_n^* decreases in n, leading to the positive matrix difference: $G_n = \Gamma_n^* - \Gamma_{n+1}^* > 0$.

We now show the almost sure convergence of $U_n^{(\Gamma)}$ to 0. Consider sufficiently large n such that $\Gamma_n^* \in K_{n+1}^{(\Gamma)}$. By applying the general projection inequality $|\Pi_K(y) - x|^2 \le |y - x|^2$, we have

$$\begin{aligned}
&\mathbb{E}\left[|U_{n+1}^{(\Gamma)}|^2 \middle| \mathcal{G}_n\right] \\
\le&\mathbb{E}\left[|U_n^{(\Gamma)} - a_n^{(\Gamma)}[h^{(\Gamma)}(\Gamma_n;\boldsymbol{\theta}_n,\gamma_n) + \xi_n^{(\Gamma)}] + G_n|^2 \middle| \mathcal{G}_n\right] \\
\le&|U_n^{(\Gamma)}|^2 - 2a_n^{(\Gamma)}\langle U_n^{(\Gamma)}, h^{(\Gamma)}(\Gamma_n;\boldsymbol{\theta}_n,\gamma_n)\rangle + (1 + |U_n^{(\Gamma)}|^2) \\
&(a_n^{(\Gamma)}|\beta_n^{(\Gamma)}| + |G_n|) + 4(a_n^{(\Gamma)})^2\Bigg(|h^{(\Gamma)}(\Gamma_n;\boldsymbol{\theta}_n,\gamma_n)|^2 \\
&+ |\beta_n^{(\Gamma)}|^2 + |\frac{G_n}{a_n^{(\Gamma)}}|^2 + \mathbb{E}\left[\left|\xi_n^{(\Gamma)} - \beta_n^{(\Gamma)}\right|^2 \middle| \mathcal{G}_n\right]\Bigg).
\end{aligned}$$

Recall that $\frac{1}{b_n}| \le |\Gamma_n| \le c_n^{(\Gamma)}$ almost surely. By (21), (24), (37), and (38), we can further derive that

$$\mathbb{E}\left[|U_{n+1}^{(\Gamma)}|^2 \middle| \mathcal{G}_n\right] \le (1 + \kappa_n^{(\Gamma)})|U_n^{(\Gamma)}|^2 - \zeta_n^{(\Gamma)} + \eta_n^{(\Gamma)},$$

where $\kappa_n^{(\Gamma)} = a_n^{(\Gamma)}|\beta_n^{(\Gamma)}| + |G_n|$,
$\zeta_n^{(\Gamma)} = 2a_n^{(\Gamma)}\langle U_n^{(\Gamma)}, h^{(\Gamma)}(\Gamma_n;\boldsymbol{\theta}_n,\gamma_n)\rangle$, and

$$\begin{aligned}
\eta_n^{(\Gamma)} =&a_n^{(\Gamma)}|\beta_n^{(\Gamma)}| + |G_n| + 4(a_n^{(\Gamma)})^2\Bigg\{c(1 + (c_n^{(\Gamma)})^4) \\
&+ c(1 + (c_n^{(\boldsymbol{\phi})})^8 + (c_n^{(\Gamma)})^8 + (\log b_n)^8)\exp\{c(c_n^{(\boldsymbol{\phi})})^4\} \\
&+ |\beta_n^{(\Gamma)}|^2 + |\frac{G_n}{a_n^{(\Gamma)}}|^2\Bigg\}.
\end{aligned}$$

Next, we prove that $\sum|G_n| < \infty$ and $\sum|G_n|^2 < \infty$. By the assumption (iii) in (39) along with (44), we obtain

$$\sum_{n=0}^{\infty}|G_n| = \sum_{n=1}^{\infty}|\Gamma_n^* - \Gamma_{n+1}^*| < c\sum_{j=1}^{m}|\frac{1}{b_n} - \frac{1}{b_{n+1}}| < \infty. \tag{45}$$

Furthermore, since $b_n \uparrow \infty$, $n_0 = \inf\{n' \in \mathbb{N} : |G_n| < 1 \text{ for all } n \ge n'\} < \infty$. Therefore, (45) yields

$$\sum_{n=1}^{\infty}|G_n|^2 < \sum_{n=1}^{n_0-1}|G_n|^2 + \sum_{n=n_0}^{\infty}|G_n| < \infty. \tag{46}$$

By the assumptions (i)-(ii) of (39), as well as (45) and (46), we know $\sum\kappa_n^{(\Gamma)} < \infty$ and $\sum\eta_n^{(\Gamma)} < \infty$. It then follows from [60, Theorem 1] that $\left|U_n^{(\Gamma)}\right|^2$ converges to a finite limit and $\sum\zeta_n^{(\Gamma)} < \infty$ almost surely.

It suffices to show $|U_n^{(\Gamma)}| \to 0$ almost surely. The equations (24) and (36) imply

$$\zeta_n^{(\Gamma)} \ge \frac{a_n^{(\Gamma)}T}{b_n c_2}\lambda_{\min}(\sum_{j=1}^m D_j D_j^\top)|U_n^{(\Gamma)}|^2.$$

Now, suppose $|U_n^{(\Gamma)}|^2 \to c$ almost surely, where $0 < c < \infty$ is a constant. Then there exists a set $Z \in \mathcal{F}$ with $\mathbb{P}(Z) > 0$ so that for every $\omega \in Z$, there exists a measurable set $Z \in \mathcal{F}$

with $\mathbb{P}(Z) = 1$ such that, for every $\omega \in Z$, there exists a constant $0 < \delta(\omega) < c$ satisfying

$$|U_n^{(\Gamma)}|^2 = |\Gamma_n - \Gamma_n^*|^2 \geq c - \delta(\omega) > 0 \quad \text{for sufficiently large } n.$$

Thus, by the assumption (iii) of (39),

$$\sum \zeta_n^{(\Gamma)} \geq \frac{T}{c_2}(c - \delta(\omega))\lambda_{\min}\Big(\sum_{j=1}^m D_j D_j^\top\Big) \sum \frac{a_n^{(\Gamma)}}{b_n} = \infty.$$

This is a contradiction, proving that $\Gamma_n \xrightarrow{a.s.} \Gamma_n^*$. However, Γ_n^* converges to $\mathbf{0}$ almost surely. Hence, $\Gamma_n \xrightarrow{a.s.} \mathbf{0}$ as well. ∎

Appendix II
Proof of Theorem 3

Proof: First, it follows from (24) and (36) that

$$\begin{aligned}&\langle U_n^{(\Gamma)}, h^{(\Gamma)}(\Gamma_n; \boldsymbol{\theta}_n, \gamma_n)\rangle \\ \geq &\frac{T}{2b_n c_2}\lambda_{\min}\Big(\sum_{j=1}^m D_j D_j^\top\Big)|U_n^{(\Gamma)}|^2 := \frac{\tilde{c}}{b_n}|U_n^{(\Gamma)}|^2,\end{aligned} \tag{47}$$

where $\tilde{c} = \frac{T}{2c_2}\lambda_{\min}(\sum_{j=1}^m D_j D_j^\top) > 0$.

Define $n_1 = \inf\{n \in \mathbb{N} : \Gamma_n^* \in K_{n+1}^{(\Gamma)}\}$. For $n \geq n_1$, we follow a similar reasoning as in the proof of Theorem 2 to get

$$\begin{aligned}\mathbb{E}\left[|U_{n+1}^{(\Gamma)}|^2 \middle| \mathcal{G}_n\right] \leq &(1 - \tilde{c}\frac{a_n^{(\Gamma)}}{b_n})|U_n^{(\Gamma)}|^2 + 4\frac{(a_n^{(\Gamma)})^2}{b_n^2}b_n^2 \\ &\Bigg(|h^{(\Gamma)}(\Gamma_n; \boldsymbol{\theta}_n, \gamma_n)|^2 + (1 + |\frac{b_n}{2\tilde{c}a_n^{(\Gamma)}}|)|\beta_n^{(\Gamma)}|^2 \\ &+ |\frac{b_n G_n^2}{2\tilde{c}(a_n^{(\Gamma)})^3}| + |\frac{G_n}{a_n^{(\Gamma)}}|^2 + \mathbb{E}\left[\left|\xi_{n+1}^{(\Gamma)} - \beta_n^{(\Gamma)}\right|^2 \middle| \mathcal{G}_n\right]\Bigg).\end{aligned} \tag{48}$$

Moreover, the assumptions in (39) imply that $(1 + |\frac{b_n}{2\tilde{c}a_n^{(\Gamma)}}|)|\beta_n^{(\Gamma)}|^2 \leq c(\frac{b_n}{a_n^{(\Gamma)}}|\beta_n^{(\Gamma)}|^2)$ and $\frac{|G_n|^2}{(a_n^{(\Gamma)})^2} \leq c(\frac{b_n|G_n|^2}{(a_n^{(\Gamma)})^3})$ for some constant $c > 0$. When $n \geq n_1$, in view of (21), (24), (37), and (38), it follows from (48) that

$$\mathbb{E}\left[|U_{n+1}^{(\Gamma)}|^2 \middle| \mathcal{G}_n\right] \leq (1 - \tilde{c}\hat{a}_n)|U_n^{(\Gamma)}|^2 + 4\hat{a}_n^2\hat{\eta}_n^{(\Gamma)},$$

where

$$\begin{aligned}\hat{\eta}_n^{(\Gamma)} = &cb_n^2\Bigg(1 + (c_n^{(\boldsymbol{\phi})})^8 + (c_n^{(\Gamma)})^8 + (\log b_n)^8 + \frac{b_n|\beta_n^{(\Gamma)}|^2}{a_n^{(\Gamma)}} \\ &+ \frac{b_n|G_n|^2}{(a_n^{(\Gamma)})^3}\Bigg)\exp\{c(c_n^{(\boldsymbol{\phi})})^4\},\end{aligned} \tag{49}$$

which is monotonically increasing because $c_n^{(\boldsymbol{\phi})}$, $c_n^{(\Gamma)}$, b_n are monotonically increasing and $\frac{b_n^3}{a_n^{(\Gamma)}}|\beta_n^{(\Gamma)}|^2$, $\frac{b_n^3|G_n|^2}{(a_n^{(\Gamma)})^3}$ are non-decreasing by the assumptions. Taking expectations on both sides of the above and denoting $\rho_n = \mathbb{E}[|U_n^{(\Gamma)}|^2]$, we get

$$\rho_{n+1} \leq (1 - \tilde{c}\hat{a}_n)\rho_n + 4\hat{a}_n^2\hat{\eta}_n^{(\Gamma)} \tag{50}$$

when $n \geq n_1$.

Next, we show $\rho_{n+1} \leq c'\hat{a}_n\hat{\eta}_n^{(\Gamma)}$ for all $n \geq 0$, where $c' = \max\{\frac{\rho_1}{\hat{a}_0\hat{\eta}_0^{(\Gamma)}}, \frac{\rho_2}{\hat{a}_1\hat{\eta}_1^{(\Gamma)}}, \cdots, \frac{\rho_{n_1+1}}{\hat{a}_{n_1}\hat{\eta}_{n_1}^{(\Gamma)}}, \frac{4}{\tilde{c}}\} + 1$. Indeed, it is true when $n \leq n_1$. Assume that $\rho_{k+1} \leq c'a_k\hat{\eta}_n^{(\Gamma)}$ is true for $n_1 \leq k \leq n-1$. Then (50) yields

$$\begin{aligned}\rho_{n+1} &\leq (1 - \tilde{c}\hat{a}_n)\rho_n + 4\hat{a}_n^2\hat{\eta}_n^{(\Gamma)} \\ &\leq (1 - \tilde{c}\hat{a}_n)c'\hat{a}_{n-1}\hat{\eta}_{1,n-1} + 4\hat{a}_n^2\hat{\eta}_n^{(\Gamma)} \\ &\leq (1 - \tilde{c}\hat{a}_n)c'\hat{a}_n(1 + w\hat{a}_n)\hat{\eta}_n^{(\Gamma)} + 4\hat{a}_n^2\hat{\eta}_n^{(\Gamma)} \\ &= c'\hat{a}_n\hat{\eta}_n^{(\Gamma)} + c'\hat{\eta}_n^{(\Gamma)}\hat{a}_n^2\left(w - \tilde{c} - \tilde{c}w\hat{a}_n + \frac{4}{c'}\right).\end{aligned}$$

Consider the function

$$f(x) = c'\hat{\eta}_n^{(\Gamma)}x^2\left(w - \tilde{c} - \tilde{c}wx + \frac{4}{c'}\right),$$

which has two roots at $x_{1,2} = 0$ and one root at $x_3 = \frac{w - (\tilde{c} - \frac{4}{c'})}{cw}$. Because $\tilde{c} - \frac{4}{c'} > 0$, we can choose $0 < w < \tilde{c} - \frac{4}{c'}$ so that $x_3 < 0$. So $f(x) < 0$ when $x > 0$, leading to

$$c'\hat{\eta}_n^{(\Gamma)}\hat{a}_n^2\left(w - \tilde{c} - \tilde{c}w\hat{a}_n + \frac{4}{c'}\right) < 0, \quad \forall n$$

because $\hat{a}_n > 0$. We have now proved $\mathbb{E}[|U_{n+1}^{(\boldsymbol{\phi})}|^2] \leq c'\hat{a}_n\hat{\eta}_n^{(\Gamma)}$.

In particular, under the setting of Remark 1 and by Lemma 3, we can verify that $\frac{b_n^3}{a_n^{(\Gamma)}}|\beta_n^{(\Gamma)}|^2$ and $\frac{b_n^3|G_n|^2}{(a_n^{(\Gamma)})^3}$ are non-decreasing sequences of n, and $\hat{a}_n = \Theta(n^{-1})$. Then

$$\hat{\eta}_n^{(\Gamma)} \leq cn^{\frac{1}{2}}(e \vee \log n)^{p_2}(1 \vee \log\log n)^{\frac{4}{3}}, \tag{51}$$

where c and p_2 are positive constants. Since $n \geq 1$, it follows that

$$\begin{aligned}\mathbb{E}[|\Gamma_{n+1} - \Gamma_{n+1}^*|^2] \leq &c\frac{(e \vee \log n)^{p_2}(1 \vee \log\log n)^{\frac{4}{3}}}{n^{\frac{1}{2}}} \\ \leq &c\frac{(e \vee \log(n+1))^{p_2}(1 \vee \log\log(n+1))^{\frac{4}{3}}}{(n+1)^{\frac{1}{2}}}\frac{(n+1)^{\frac{1}{2}}}{n^{\frac{1}{2}}} \\ \leq &c\sqrt{2}\frac{(e \vee \log(n+1))^{p_2}(1 \vee \log\log(n+1))^{\frac{4}{3}}}{(n+1)^{\frac{1}{2}}}.\end{aligned}$$

Finally, noting that Γ_n^* converges to $\mathbf{0}$ with the rate of $\frac{1}{b_n}$ by (44), we have

$$\begin{aligned}\mathbb{E}[|\Gamma_n|^2] = &\mathbb{E}[|\Gamma_n - \Gamma_n^* + \Gamma_n^*|^2] \leq \mathbb{E}[|\Gamma_n - \Gamma_n^*|^2] + \mathbb{E}[|\Gamma_n^*|^2] \\ \leq &c'\frac{(e \vee \log n)^{p_2}(1 \vee \log\log n)^{\frac{4}{3}}}{n^{\frac{1}{2}}} + c'\frac{\alpha^{\frac{1}{2}}}{(n+\beta)^{\frac{1}{2}}} \\ \leq &c\frac{(e \vee \log n)^{p_2}(1 \vee \log\log n)^{\frac{4}{3}}}{n^{\frac{1}{2}}}.\end{aligned}$$

The proof is complete. ∎

Appendix III
Proof of Theorem 4

Proof: When $x_0 \neq 0$, the proof is similar to that of [40, Theorem 4.1] with only minor modifications needed. When $x_0 = 0$, the proof becomes more delicate as the convergence behavior of $\boldsymbol{\phi}$ now depends explicitly on the actor exploration parameter Γ. Here, we highlight the differences needed.

Denote

$$h^{(\boldsymbol{\phi})}(\boldsymbol{\phi}_n, \Gamma_n; \boldsymbol{\theta}_n, \gamma_n) = \mathbb{E}[Y_n(T) \mid \boldsymbol{\theta}_n, \boldsymbol{\phi}_n, \gamma_n, \Gamma_n],$$

and the noise

$$\xi_n^{(\phi)} = Y_n(T) - h^{(\phi)}(\boldsymbol{\phi}_n, \Gamma_n; \boldsymbol{\theta}_n, \gamma_n).$$

The updating rule (25) for $\boldsymbol{\phi}$ is now

$$\boldsymbol{\phi}_{n+1} = \Pi_{K_{n+1}^{(\phi)}}(\boldsymbol{\phi}_n + a_n^{(\phi)}[h^{(\phi)}(\boldsymbol{\phi}_n, \Gamma_n; \boldsymbol{\theta}_n.\gamma_n) + \xi_n^{(\phi)}]). \tag{52}$$

Similar to [40, Section B], we obtain

$$h^{(\phi)}(\boldsymbol{\phi}_n, \Gamma_n; \boldsymbol{\theta}_n, \gamma_n) = -l(\boldsymbol{\phi}_n, \Gamma_n; \boldsymbol{\theta}_n, \gamma_n)(\boldsymbol{\phi}_n - \boldsymbol{\phi}^*), \tag{53}$$

where

$$l(\boldsymbol{\phi}_n, \Gamma_n; \boldsymbol{\theta}_n, \gamma_n) = (\sum_{j=1}^m D_j D_j^\top) \int_0^T k_1(t; \boldsymbol{\theta}_n) \mathbb{E}[x_n(t)^2] \mathrm{d}t. \tag{54}$$

When $x_0 \neq 0$, we have $l(\boldsymbol{\phi}_n, \Gamma_n; \boldsymbol{\theta}_n, \gamma_n) \geq \bar{c} I$, where $0 < \bar{c}' < 1$ is independent of Γ_n. The same proof of [40, Theorem 4.1] then applies.

However, when $x_0 = 0$,

$$\begin{aligned} l(\boldsymbol{\phi}_n, \Gamma_n; \boldsymbol{\theta}_n, \gamma_n) &\geq (\sum_{j=1}^m D_j D_j^\top) \int_0^T \frac{c'|\Gamma_n| t}{c_2} \mathrm{d}t \\ &= \frac{(\sum_{j=1}^m D_j D_j^\top) T^2 c'}{2c_2} |\Gamma_n| \geq \bar{c}' |\Gamma_n| I. \end{aligned} \tag{55}$$

Thus l no longer admits a uniform lower bound strictly away from $\mathbf{0}$ because $\Gamma_n \to \mathbf{0}$. To get around this, for establishing part (a), we make use of the condition $\sum \frac{a_n^{(\phi)}}{b_n} = \infty$, which ensures that the argument in proving [40, Theorem B.3] remains valid. (Notice that the hyperparameters $a_n^{(\phi)}$ and b_n specified in Theorem 1 also satisfy this condition.) On the other hand, for proving part (b), we reinterpret $\hat{a}_n = \frac{a_n^{(\phi)}}{b_n}$ as the effective learning rate and follow the strategy outlined in the proof of Theorem 3. This adjustment allows the analysis in [40, Appendix B.6] to be carried over to our setting. ■

Appendix IV
Proof of Theorem 5

Proof: Following [40, Lemma B.7, B.8], we have

$$\bar{J}(\boldsymbol{\phi}, \Gamma) = f(a(\boldsymbol{\phi})) + (\sum_{j=1}^m D_j^\top \Gamma D_j) g(a(\boldsymbol{\phi})), \tag{56}$$

where

$$a(\boldsymbol{\phi}) = 2A + 2B^\top \boldsymbol{\phi} + \sum_{j=1}^m (C_j^2 + 2C_j D_j^\top \boldsymbol{\phi} + D_j^\top \boldsymbol{\phi}\boldsymbol{\phi}^\top D_j), \tag{57}$$

and

$$f(a) = \begin{cases} \frac{x_0^2(-H-QT)}{2} & \text{if } a = 0, \\ \frac{1}{2a}(Q - e^{aT}Q - He^{aT}a)x_0^2 & \text{if } a \neq 0, \end{cases} \tag{58}$$

$$g(a) = \begin{cases} \frac{T(-2H-QT)}{4} & \text{if } a = 0, \\ \frac{1}{2a^2}(QTa + Q + Ha - e^{aT}Q - He^{aT}a) & \text{if } a \neq 0. \end{cases} \tag{59}$$

Clearly, both f and g are continuously differentiable, non-increasing, and strictly negative everywhere.

Next, we present the proofs under $x_0 \neq 0$ and $x_0 = 0$.

Case I: $x_0 \neq 0$. By (56), we have

$$\begin{aligned} &\bar{J}(\boldsymbol{\phi}^*, 0) - \bar{J}(\boldsymbol{\phi}_n, \Gamma_n) \\ =&[f(a(\boldsymbol{\phi}^*)) - f(a(\boldsymbol{\phi}_n))] - (\sum_{j=1}^m D_j^\top \Gamma_n D_j) g(a(\boldsymbol{\phi}^*)) \\ &+ (\sum_{j=1}^m D_j^\top \Gamma_n D_j)[g(a(\boldsymbol{\phi}^*)) - g(a(\boldsymbol{\phi}_n))]. \end{aligned} \tag{60}$$

The goal is to establish bounds for the three terms on the right-hand side of (60), adapting the arguments of the proof of [40, Theorem 4.2] to the current setting. The first term, $f(a(\boldsymbol{\phi}^*)) - f(a(\boldsymbol{\phi}_n))$, can be bounded exactly as in the proof of [40, Theorem 4.2], since it does not involve Γ_n. For the second term, $\left(\sum_j D_j^\top \Gamma_n D_j\right) g(a(\boldsymbol{\phi}^*))$, and the third term, $\left(\sum_j D_j^\top \Gamma_n D_j\right)[g(a(\boldsymbol{\phi}^*)) - g(a(\boldsymbol{\phi}_n))]$, although Γ_n is now stochastic rather than deterministic, the bounds can still be derived by utilizing part (b) of Theorem 1 and applying the Hölder and Cauchy–Schwarz inequalities. These modifications allow the original argument in the proof of [40, Theorem 4.2] to follow through.

Case II: $x_0 = 0$**.** It follows from (58) that $f(a(\boldsymbol{\phi})) = 0$ for any $\boldsymbol{\phi}$. Hence

$$\begin{aligned} &\bar{J}(\boldsymbol{\phi}^*, 0) - \bar{J}(\boldsymbol{\phi}_n, \Gamma_n) \\ =& -(\sum_{j=1}^m D_j^\top \Gamma_n D_j) g(\boldsymbol{\phi}^*) + (\sum_{j=1}^m D_j^\top \Gamma_n D_j)(g(\boldsymbol{\phi}^*) - g(\boldsymbol{\phi}_n)). \end{aligned} \tag{61}$$

The bound for the term $\left(\sum_j D_j^\top \Gamma_n D_j\right) g(a(\boldsymbol{\phi}^*))$ can be derived exactly as with the second term in Case I. Moreover, utilizing the MSE of $\boldsymbol{\phi}_n$ for $x_0 = 0$, we can also derive the bound for the other term, $(\sum_j D_j^\top \Gamma_n D_j)(g(\boldsymbol{\phi}^*) - g(\boldsymbol{\phi}_n))$, analogously to Case I.

■

Appendix V
More Details for Numerical Experiments

To ensure full replicability, we set the random seed from 1 to 100 for each independent run in both LQRL_Adaptive and the model-based benchmark in the first and second experiments. For the third experiment comparing LQRL_Adaptive and LQRL_Fixed, we set the random seed from 1 to 1,000 for each independent run in both scenarios. For the fourth experiment under the randomized environment, we set the random seed from 1 to 10,000.

The setup for our first experiment using the model-free algorithm with data-driven exploration (LQRL_Adaptive), Algorithm 1, is as follows. The initial policy mean parameter is set to $\boldsymbol{\phi}_0 = -1.1$, with the actor and critic exploration levels initialized at $\Gamma_0 = 0.5$ and $\gamma_0 = 2$, respectively. The learning rates are specified by $a_n^{(\boldsymbol{\phi})} = \frac{0.05}{(n+1)^{3/4}}$ and $a_n^{(\Gamma)} = \frac{1}{(n+1)^{3/4}}$. For computational efficiency, the projection sets are defined as $[-2.25, -1.1]$ for $\boldsymbol{\phi}_n$ and $[0, 1]$ for Γ_n. Although the projections were originally specified for theoretical convergence guarantees, they are slightly modified in our experiments to accelerate convergence. The projection sets for

$\boldsymbol{\theta}$ and γ remain unbounded. The deterministic sequence b_n is chosen as $b_n = 20\max(1,(n+1)^{1/4})$. Additionally, we use the simplified critic specification $k_1(t;\boldsymbol{\theta}) \equiv 1$ and $k_3(t;\boldsymbol{\theta},\gamma) \equiv 0$, which trivially satisfies the boundedness conditions in (11). Under this specialization, the temperature update reduces to $\gamma_n = c_\gamma / b_n$. Thus, the numerical implementation corresponds to a special case of the general adaptive critic update (12). This simplification is used to avoid introducing a critic neural-network so as to save computational cost. In this sense, the experiments provide a conservative implementation of the proposed framework, showing that strong empirical performance can already be obtained under a simple bounded critic specification. Note that this does not contradict the general theoretical framework, since the convergence and regret analysis only requires the bounds in (11) and does not rely on the explicit functional forms of k_1 and k_3. More general choices, including neural-network parameterizations satisfying (11), can be incorporated directly into Algorithm 1 through the policy-evaluation update (27).

The model-based benchmark is adapted from [41] and extended by [40, Algorithm A.1] to accommodate state- and control-dependent volatility in our setting. The initial parameter estimations are set to be $A = B = C = D = 10$, resulting in the same initial $\phi_0 = -1.1$ as in LQRL_Adaptive. The initial exploration level $\Gamma_0 = 0.5$ is also matched. The deterministic exploration schedule follows $\Gamma_n = 0.5(n+1)^{-1}$, consistent with [40]. Finally, the projection set for ϕ_n remains $[-2.25, -1.1]$, aligning with the setup used in LQRL_Adaptive.

For all the plots in the first and second experiments, for both our adaptive model-free algorithm and model-based benchmarks, regression lines are fitted using data from iterations 5,000 to 100,000. This avoids the influence of early-stage learning noise. In all the regret plots, we use the median and exclude extreme values; similar conclusions hold when using the mean instead.

In our third experiment, comparing LQRL_Adaptive and LQRL_Fixed, most settings remain identical to the first experiment, except that the bounds $c_n^{(\phi)}$ and $c_n^{(\Gamma)}$ are set to be 20 to accommodate the scenarios under consideration. For the fourth experiment, to account for potentially large values of ϕ and Γ due to random model parameters, we further increase the bounds $c_n^{(\phi)}$ and $c_n^{(\Gamma)}$ to be 100.

[1] B. D. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*. Courier Corporation, 2007.

[2] J. Yong and X. Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*. New York, NY: Spinger, 1999.

[3] S. Chen, X. Li, and X. Y. Zhou, "Stochastic linear quadratic regulators with indefinite control weight costs," *SIAM Journal on Control and Optimization*, vol. 36, no. 5, pp. 1685–1702, 1998.

[4] M. Ait Rami, X. Y. Zhou, and J. Moore, "Well-posedness and attainability of indefinite stochastic linear quadratic control in infinite time horizon," *Systems & Control Letters*, vol. 41, no. 2, pp. 123–133, 2000.

[5] M. A. Rami and X. Y. Zhou, "Linear matrix inequalities, riccati equations, and indefinite stochastic linear quadratic controls," *IEEE Transactions on Automatic Control*, vol. 45, no. 6, pp. 1131–1143, 2000.

[6] M. A. Rami, X. Chen, J. B. Moore, and X. Y. Zhou, "Solvability and asymptotic behavior of generalized riccati equations arising in indefinite stochastic lq controls," *IEEE Transactions on Automatic Control*, vol. 46, no. 3, pp. 428–440, 2001.

[7] D. D. Yao, S. Zhang, and X. Y. Zhou, "A primal-dual semi-definite programming approach to linear quadratic control," *IEEE Transactions on Automatic Control*, vol. 46, no. 9, pp. 1442–1447, 2001.

[8] H. Li, Q. Qi, and H. Zhang, "Stabilization control for itô stochastic system with indefinite state and control weight costs," *International Journal of Control*, vol. 95, no. 2, pp. 295–302, 2022.

[9] B. Gashi and H. Hua, "Optimal regulators for a class of nonlinear stochastic systems," *International Journal of Control*, vol. 96, no. 1, pp. 136–146, 2023.

[10] G. Wang and W. Wang, "Indefinite linear-quadratic optimal control of mean-field stochastic differential equation with jump diffusion: an equivalent cost functional method," *IEEE Transactions on Automatic Control*, 2024.

[11] F. Wu, X. Li, and X. Zhang, "Stochastic linear quadratic optimal control problems with regime-switching jumps in infinite horizon," *SIAM Journal on Control and Optimization*, vol. 63, no. 2, pp. 852–891, 2025.

[12] R. C. Merton, "On estimating the expected return on the market: An exploratory investigation," *Journal of Financial Economics*, vol. 8, no. 4, pp. 323–361, 1980.

[13] D. G. Luenberger, *Investment Science*. Oxford University Press, 1998.

[14] B. Rustem and M. Howe, *Algorithms for worst-case design and applications to risk management*. Princeton University Press, 2009.

[15] S. Bradtke, "Reinforcement learning applied to linear quadratic regulation," *Advances in Neural Information Processing Systems*, vol. 5, 1992.

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

[17] 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*, pp. 1–26, JMLR Workshop and Conference Proceedings, 2011.

[18] M. Abeille and A. Lazaric, "Improved regret bounds for Thompson sampling in linear quadratic control problems," in *International Conference on Machine Learning*, pp. 1–9, PMLR, 2018.

[19] B. Hambly, R. Xu, and H. Yang, "Policy gradient methods for the noisy linear quadratic regulator over a finite horizon," *SIAM Journal on Control and Optimization*, vol. 59, no. 5, pp. 3359–3391, 2021.

[20] W. Wang, J. Han, Z. Yang, and Z. Wang, "Global convergence of policy gradient for linear-quadratic mean-field control/game in continuous time," in *International Conference on Machine Learning*, pp. 10772–10782, PMLR, 2021.

[21] M. Zhou and J. Lu, "Single timescale actor-critic method to solve the linear quadratic regulator with convergence guarantees," *Journal of Machine Learning Research*, vol. 24, no. 222, pp. 1–34, 2023.

[22] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 2018.

[23] A. Aubret, L. Matignon, and S. Hassas, "A survey on intrinsic motivation in reinforcement learning," *arXiv preprint arXiv:1908.06976*, 2019.

[24] J. Schmidhuber, "Formal theory of creativity, fun, and intrinsic motivation (1990–2010)," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 230–247, 2010.

[25] J. Achiam and S. Sastry, "Surprise-based intrinsic motivation for deep reinforcement learning," *arXiv preprint arXiv:1703.01732*, 2017.

[26] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," in *International Conference on Machine Learning*, pp. 2778–2787, PMLR, 2017.

[27] Y. Burda, H. Edwards, D. Pathak, A. Storkey, T. Darrell, and A. A. Efros, "Large-scale study of curiosity-driven learning," *arXiv preprint arXiv:1808.04355*, 2018.

[28] Y. Burda, H. Edwards, A. Storkey, and O. Klimov, "Exploration by random network distillation," *arXiv preprint arXiv:1810.12894*, 2018.

[29] I. Osband, J. Aslanides, and A. Cassirer, "Randomized prior functions for deep reinforcement learning," *Advances in neural information processing systems*, vol. 31, 2018.

[30] P. Ménard, O. D. Domingues, A. Jonsson, E. Kaufmann, E. Leurent, and M. Valko, "Fast active learning for pure exploration in reinforcement learning," in *International Conference on Machine Learning*, pp. 7599–7608, PMLR, 2021.

[31] H. Tang, R. Houthooft, D. Foote, A. Stooke, O. Xi Chen, Y. Duan, J. Schulman, F. DeTurck, and P. Abbeel, "# exploration: A study of count-based exploration for deep reinforcement learning," *Advances in neural information processing systems*, vol. 30, 2017.

[32] A. Ecoffet, J. Huizinga, J. Lehman, K. O. Stanley, and J. Clune, "Go-explore: A new approach for hard-exploration problems," *arXiv preprint arXiv:1901.10995*, 2019.

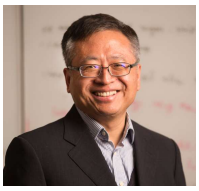
[33] A. Ecoffet, J. Huizinga, J. Lehman, K. O. Stanley, and J. Clune, "First return, then explore," *Nature*, vol. 590, no. 7847, pp. 580–586, 2021.

[34] M. Vecerik, T. Hester, J. Scholz, F. Wang, O. Pietquin, B. Piot, N. Heess, T. Rothörl, T. Lampe, and M. Riedmiller, "Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards," *arXiv preprint arXiv:1707.08817*, 2017.
[35] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.
[36] W. Saunders, G. Sastry, A. Stuhlmueller, and O. Evans, "Trial without error: Towards safe reinforcement learning via human intervention," *arXiv preprint arXiv:1707.05173*, 2017.
[37] R. Munos, "Policy gradient in continuous time," *Journal of Machine Learning Research*, vol. 7, pp. 771–791, 2006.
[38] C. Tallec, L. Blier, and Y. Ollivier, "Making deep Q-learning methods robust to time discretization," in *International Conference on Machine Learning*, pp. 6096–6104, PMLR, 2019.
[39] S. Park, J. Kim, and G. Kim, "Time discretization-invariant safe action repetition for policy gradient methods," *Advances in Neural Information Processing Systems*, vol. 34, pp. 267–279, 2021.
[40] Y. Huang, Y. Jia, and X. Y. Zhou, "Sublinear regret for a class of continuous-time linear-quadratic reinforcement learning problems," *SIAM Journal on Control and Optimization*, vol. 63, no. 5, pp. 3452–3474, 2025.
[41] L. Szpruch, T. Treetanthiploet, and Y. Zhang, "Optimal scheduling of entropy regularizer for continuous-time linear-quadratic reinforcement learning," *SIAM Journal on Control and Optimization*, vol. 62, no. 1, pp. 135–166, 2024.
[42] Y. Hu and X. Y. Zhou, "Indefinite stochastic riccati equations," *SIAM Journal on Control and Optimization*, vol. 42, no. 1, pp. 123–137, 2003.
[43] Z. Qian and X. Y. Zhou, "Existence of solutions to a class of indefinite stochastic riccati equations," *SIAM journal on Control and Optimization*, vol. 51, no. 1, pp. 221–229, 2013.
[44] H. Wang, T. Zariphopoulou, and X. Y. Zhou, "Reinforcement learning in continuous time and space: A stochastic control approach," *Journal of Machine Learning Research*, vol. 21, no. 198, pp. 1–34, 2020.
[45] L. Szpruch, T. Treetanthiploet, and Y. Zhang, "Exploration-exploitation trade-off for continuous-time episodic reinforcement learning with linear-convex models," *arXiv preprint arXiv:2112.10264*, 2021.
[46] Y. Jia and X. Y. Zhou, "Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms," *Journal of Machine Learning Research*, vol. 23, no. 275, pp. 1–50, 2022.
[47] Y. Jia and X. Y. Zhou, "Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach," *Journal of Machine Learning Research*, vol. 23, no. 154, pp. 1–55, 2022.
[48] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, pp. 400–407, 1951.
[49] T. L. Lai, "Stochastic approximation," *The Annals of Statistics*, vol. 31, no. 2, pp. 391–406, 2003.
[50] V. S. Borkar, *Stochastic approximation: A dynamical systems viewpoint*, vol. 48. Springer, 2009.
[51] M. Chau and M. C. Fu, "An overview of stochastic approximation," *Handbook of Simulation Optimization*, pp. 149–178, 2014.
[52] S. Andradóttir, "A stochastic approximation algorithm with varying bounds," *Operations Research*, vol. 43, no. 6, pp. 1037–1048, 1995.
[53] Y. Huang, Y. Jia, and X. Y. Zhou, "Mean–variance portfolio selection by continuous-time reinforcement learning: Algorithms, regret analysis, and empirical study," *arXiv preprint arXiv:2412.16175*, 2024.
[54] Y. Jia and X. Y. Zhou, "q-learning in continuous time," *Journal of Machine Learning Research*, vol. 24, no. 161, pp. 1–61, 2023.
[55] M. Broadie, D. Cicek, and A. Zeevi, "General bounds and finite-time improvement for the Kiefer-Wolfowitz stochastic approximation algorithm," *Operations Research*, vol. 59, no. 5, pp. 1211–1224, 2011.
[56] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, vol. 23. Springer, 1992.
[57] O. Shamir and T. Zhang, "Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes," in *International conference on machine learning*, pp. 71–79, PMLR, 2013.
[58] G. Nie, M. Agarwal, A. K. Umrawal, V. Aggarwal, and C. J. Quinn, "An explore-then-commit algorithm for submodular maximization under full-bandit feedback," in *Uncertainty in Artificial Intelligence*, pp. 1541–1551, PMLR, 2022.
[59] W. Tang and X. Y. Zhou, "Regret of exploratory policy improvement and q-learning," *arXiv preprint arXiv:2411.01302*, 2024.
[60] H. Robbins and D. Siegmund, "A convergence theorem for non negative almost supermartingales and some applications," in *Optimizing Methods in Statistics*, pp. 233–257, Elsevier, 1971.

Yilie Huang (Member, IEEE) was born in Guangzhou, China. He received the B.S. degree in mathematics and applied mathematics from Zhejiang University, Hangzhou, China, in 2017, and the M.S. and Ph.D. degrees in operations research from Columbia University, New York, USA, in 2018 and 2024, respectively.

He is currently an Assistant Professor with the Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong. Prior to this appointment, he was a Postdoctoral Research Scientist with the Department of Industrial Engineering and Operations Research, Columbia University, New York, USA. His research lies at the intersection of stochastic control, reinforcement learning, diffusion models for generative AI, and financial engineering, with a focus on the development, analysis, and empirical evaluation of continuous-time reinforcement learning algorithms for decision making, control, and financial systems under uncertainty.

Dr. Huang is a Member of the Institute for Operations Research and the Management Sciences (INFORMS), the Society for Industrial and Applied Mathematics (SIAM), the Association for Computing Machinery (ACM), and is a Chartered Financial Analyst (CFA) charterholder.



Xun Yu Zhou (Fellow, IEEE) was born in Suzhou, China. He received his B.Sc. in mathematics in 1984, and his Ph.D. in operations research and control theory in 1989, both from Fudan University, Shanghai, China.

He is currently the Liu Family Professor of Industrial Engineering and Operations Research and Director of the Nie Center for Intelligent Asset Management at Columbia University in New York. He was the Nomura Professor of Mathematical Finance, the Director of Nomura Center for Mathematical Finance and the Director of Oxford-Nie Financial Big Data Lab at the University of Oxford during 2007-2016 before joining Columbia.

He is known for his work in indefinite stochastic LQ control theory and application to dynamic mean–variance portfolio selection, in asset allocation and pricing under prospect theory, and in general time-inconsistent problems. His current research focuses on reinforcement learning for controlled diffusion processes and applications to GenAI and intelligent wealth management solutions. He has addressed the 2010 International Congress of Mathematicians, and has been awarded the Wolfson Research Award from The Royal Society (UK), the Outstanding Paper Prize from SIAM, the Humboldt Distinguished Lecturer, the Alexander von Humboldt Research Fellowship, the Archimedes Lecturer at Columbia, and Distinguished Faculty Teaching Award at Columbia University.

Professor Zhou is a Fellow of SIAM, a member of INFORMS, and a Life Member of the Bachelier Finance Society. He is or was on the editorial boards of numerous journals including *IEEE Transactions on Automatic Control* and *SIAM Journal on Control and Optimization*.