# Sublinear Regret for a Class of Continuous-Time Linear–Quadratic Reinforcement Learning Problems

Yilie Huang * Yanwei Jia† Xun Yu Zhou‡

**Abstract**

We study reinforcement learning (RL) for a class of continuous-time linear–quadratic (LQ) control problems for diffusions, where states are scalar-valued and running control rewards are absent but volatilities of the state processes depend on both state and control variables. We apply a model-free approach that relies neither on knowledge of model parameters nor on their estimations, and devise an actor–critic algorithm to learn the optimal policy parameter directly. Our main contributions include the introduction of an exploration schedule and a regret analysis of the proposed algorithm. We provide the convergence rate of the policy parameter to the optimal one, and prove that the algorithm achieves a regret bound of $O(N^{\frac{3}{4}})$ up to a logarithmic factor, where $N$ is the number of learning episodes. We conduct a simulation study to validate the theoretical results and demonstrate the effectiveness and reliability of the proposed algorithm. We also perform numerical comparisons between our method and those of the recent model-based stochastic LQ RL studies adapted to the state- and control-dependent volatility setting, demonstrating a better performance of the former in terms of regret bounds.

***Keywords:*** C ontinuous-Time RL, Stochastic Linear–Quadratic Control, Model-Free Policy Gradient, Regret Bounds, Exploration Scheduling, Actor–Critic Algorithm

## 1 Introduction

Linear–quadratic (LQ) control, where the system dynamics are linear in the state and control variables while the rewards are quadratic in them, takes up a center stage in classical model-based control theory when the model parameters are assumed to be given and known. The reason is

*Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, USA. Email: yh2971@columbia.edu.

†Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong. Email: yanweijia@cuhk.edu.hk.

‡Department of Industrial Engineering and Operations Research & Data Science Institute, Columbia University, New York, NY 10027, USA. Email: xz2574@columbia.edu.

twofold: LQ control can be solved explicitly and elegantly, and it can be used to approximate more complicated nonlinear control problems. Detailed theoretical accounts in the continuous-time setting can be found in Anderson and Moore (2007) for deterministic control (i.e., dynamics are described by ordinary differential equations) and in Yong and Zhou (1999) for stochastic control (i.e., dynamics are governed by stochastic differential equations).

Many real-world applications often present themselves with partially known or entirely unknown environments. Specifically in the LQ context, one may know that a problem is *structurally* LQ, namely the system responds linearly to state and control whereas the reward is quadratic (e.g. a variance is involved) in these variables, yet *without knowing some or any of the model parameters*. The so-called plug-in method has been traditionally used to solve such a problem, namely, one first estimates the model parameters based on observed data and then plugs in the estimated parameters and applies the classical optimal control theory to derive the solutions. Such an approach is *model-based/-driven* because it takes learning the model as its core mission. It is well known, however, that the plug-in method has significant drawbacks, especially in that optimal controls are typically very sensitive to the model parameters, yet estimating some of the parameters accurately when data are limited is a daunting, sometimes impossible, task (e.g., the return rate of a stock (Merton, 1980; Luenberger, 1998)).

Reinforcement learning (RL) has been developed to tackle complex control problems in largely unknown environments. Its successful applications range from strategic board games such as chess and Go (Silver et al., 2016, 2017) to robotic systems (Gu et al., 2017; Khan et al., 2020). However, RL has been predominantly studied for discrete-time, Markov decision processes (MDPs) with discrete state and control spaces, even though most real-life applications are inherently continuous-time with continuous state space and possibly continuous control space (e.g., autonomous driving, stock trading, and video game playing). More importantly, while one can turn a continuous-time problem into a discrete-time MDP upfront by time discretization, such an approach is very sensitive to time step size and performs poorly with small time steps Munos (2006); Tallec et al. (2019); Park et al. (2021).

While there were studies directly on continuous-time RL, these had been rare and far between (Baird, 1994; Doya, 2000; Vamvoudakis and Lewis, 2010; Lee and Sutton, 2021; Kim et al., 2021) up to just recent years, overall lacking a systematic and unified theory. Starting with Wang et al. (2020) that introduces an entropy-regularized relaxed control framework for continuous-time RL, a series of subsequent papers (Jia and Zhou, 2022a,b, 2023) develop theories on policy evaluation,

policy gradient, and $q$-learning respectively within this framework. This strand of research is characterized by focusing on learning the optimal control policies directly without attempting to estimate or learn the model parameters, underlining a model-free (up to the unknown dynamics being governed by diffusion processes) and data-driven approach. The mathematical foundation of the entire theory is the martingale property of certain stochastic processes, the enforcement of which naturally leads to various temporal difference and actor–critic algorithms to train and learn $q$-functions, value functions, and optimal (stochastic) policies. Subsequently, there has been active follow-up research with various extensions and applications; see, e.g. Huang et al. (2022); Dai et al. (2023); Wang et al. (2023); Frikha et al. (2023); Wei and Yu (2023).

A crucial question in RL is the convergence and regret bounds of RL algorithms that provide theoretical guidance and guarantee their effectiveness and reliability. For LQ problems, such theoretical results exist, for example, for deterministic systems (Bradtke, 1992; Fazel et al., 2018; Malik et al., 2019) as well as systems with identically and independently distributed noises (Abbasi-Yadkori and Szepesvári, 2011; Abeille and Lazaric, 2018; Cohen et al., 2018, 2019; Hambly et al., 2021; Wang et al., 2021; Cassel and Koren, 2021; Yang et al., 2019; Zhou and Lu, 2023; Chen et al., 2023; Simchowitz and Foster, 2020), covering finite-horizon, infinite-horizon, and ergodic cases. These studies are nevertheless all for discrete-time models, with control not affecting the level of randomness in the state dynamics. Some of them, e.g. Abbasi-Yadkori and Szepesvári (2011); Abeille and Lazaric (2018); Hambly et al. (2021); Wang et al. (2021); Zhou and Lu (2023), design their algorithms based on policy gradient. However, the gradient representations therein rely on estimations of the drift parameters; hence, the methods are essentially model-based. In addition, the semidefinite programming formulation in (Cohen et al., 2018, 2019) does not seem applicable to continuous-time systems.

The algorithm proposed and analyzed in the present paper belongs to the class of actor–critic algorithms originally put forward by Konda and Tsitsiklis (1999). Such algorithms for discrete-time systems have been studied in Wu et al. (2020); Xu et al. (2021); Cen et al. (2022), and in particular, for ergodic LQ problems in Yang et al. (2019); Chen et al. (2023) and for episodic linear MDPs in Cai et al. (2020); Zhong and Zhang (2023). The "optimal" regret of these algorithms is mostly of the order $O(\sqrt{N})$, where $N$ is the number of episodes or timesteps. However, it is unclear whether they still work for the diffusion case where the volatility also depends on state and control. For general continuous-time diffusion environments, however, the aforementioned series of papers (Wang et al. (2020); Jia and Zhou (2022a,b, 2023)) and their follow-up study have not addressed the problems

of convergence and regret bounds. These remain highly significant yet challenging open questions due to the model-free nature of the underlying approach and the stochastic approximation type of algorithms involved.

Recently, there has been some progress on regret analysis for continuous-time stochastic LQ RL in (Basei et al., 2022; Szpruch et al., 2024) that achieve sublinear regrets of their respective algorithms. Both papers assume that the diffusion coefficients are *constant* independent of state and control, which is vital for their approaches to work. Again, in essence, these works are model-based because they apply either least-square or Bayesian methods to estimate model parameters and use the corresponding estimation errors to deduce the regret bounds. In particular, there is an intrinsic drawback in these model-based methods specific to LQ problems: optimal control policies are *linear* feedbacks of the state; hence, these methods may suffer from the unidentifiability problem. In addition, they require both the batch size and the number of timesteps to increase exponentially over episodes, adding substantial computational and memory costs.

This paper endeavors to design an RL algorithm with a provable sublinear regret for a class of stochastic LQ RL problems in the model-free framework of Wang et al. (2020); Jia and Zhou (2022a,b, 2023). We allow the diffusion coefficients to depend on both state and control, the latter being of particular practical significance (e.g. the wealth equation in continuous-time finance (Zhou and Li, 2000)). Indeed, this type of stochastic LQ problems have led to a very active research area called "*indefinite* stochastic LQ control" in the classical, model-based literature, starting from (Chen et al., 1998; Rami and Zhou, 2000).

**Main Contributions** In this paper, we propose a policy gradient based actor–critic algorithm to solve a special class of continuous-time, finite-horizon stochastic LQ problems under the model-free, episodic RL setting, where the state processes are one dimensional and there is no running control award. Our main contributions are

(1) We provide a convergence and regret analysis when the volatility of the state process is affected by both state and control. The regret is upper bounded by the order of $O(N^{\frac{3}{4}})$ (up to a logarithmic factor), where $N$ is the number of episodes. While it may not yet be the best regret bound, to our best knowledge, it is the first sublinear regret result obtained in the entropy-regularized exploratory framework of Wang et al. (2020), with state- and action-dependent volatility.

(2) We take a model-free approach to develop our algorithm, i.e., a policy gradient based (soft)

actor–critic algorithm, and base our analysis on the stochastic approximation scheme. In particular, the policy gradient in this paper is a "model-free gradient" instead of a "model-based gradient" commonly taken in discrete-time RL. As a result, we do not need to estimate model primitives in the entire analysis, circumventing the issues discussed earlier arising from estimating/learning those model parameters.

(3) We propose a novel exploration schedule. Note that stochastic policies are considered in this paper for both conceptual and technical reasons. Conceptually, stochastic policies reach more action areas otherwise not necessarily explored by deterministic policies. Technically, we apply the policy gradient method developed in Jia and Zhou (2022b) that works only for stochastic policies. Gaussian exploration policies are shown to be optimal in achieving the ideal balance between exploration and exploitation, whose variance represents the level of exploration. We propose a decreasing schedule of variances for the Gaussian exploration over iterations, guided by the desired regret bound.

The remainder of the paper is structured as follows. Section 2 formulates the problem and provides some preliminary results necessary for the subsequent development. Section 3 describes and explains the steps leading to our RL algorithm. Section 4 presents the main theoretical results on convergence and a regret bound of the proposed algorithm. Section 5 reports the results of numerical experiments. Finally, Section 6 concludes. The appendices contain proofs of the main results and a detailed description of the numerical experiments.

# 2 Problem Formulation and Preliminaries

## 2.1 Classical Stochastic LQ Control

We begin by recalling the classical stochastic LQ control formulation and main results. Denote by $x^u = \{x^u(t) \in \mathbb{R} : 0 \leqslant t \leqslant T\}$ the state process under a control process $u = \{u(t) \in \mathbb{R}^l : 0 \leqslant t \leqslant T\}$, whose dynamics are described by the following stochastic differential equation (SDE):

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

where $A$ and $C_j$ are scalars, while $B$ and $D_j$ are $l \times 1$ vectors. The initial state is $x^u(0) = x_0 \neq 0$, and $W = \{(W^{(1)}(t), \cdots, W^{(m)}(t))^\top \in \mathbb{R}^m : 0 \leqslant t \leqslant T\}$ is an $m$-dimensional standard Brownian

motion.

The goal of the control problem is to choose a control process $u$ to maximize the expected value of a quadratic objective functional:

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

where $Q \geqslant 0$ and $H \geqslant 0$ are given scalar weighting parameters. One can define the optimal value function

$$V_{CL}(t,x) = \max_u \mathbb{E}\left[\int_t^T -\frac{1}{2}Qx^u(s)^2\mathrm{d}s - \frac{1}{2}Hx^u(T)^2\Big|x^u(t) = x\right]. \tag{3}$$

While the existing works on stochastic LQ RL assume the diffusion coefficient to be a constant, control- and state-dependent diffusion terms appear in many applications. On the other hand, the state is one-dimensional and running control reward is absent in our problem, which are crucial assumptions for our approach to work. This class of problems cover important applications such as the mean–variance portfolio selection (Zhou and Li, 2000; Dai et al., 2023; Huang et al., 2022).

If the model parameters $A$, $B$, $C_j$, $D_j$, $Q$, and $H$ are all known with the assumption that $\sum_{j=1}^m D_j D_j^\top$ is positive definite, this problem can be solved explicitly as detailed in (Yong and Zhou, 1999, Chapter 6). Specifically, the optimal value function and optimal feedback control policy are respectively

$$V_{CL}(t,x) = -\frac{1}{2}\left[\frac{Q}{\Lambda} + (H - \frac{Q}{\Lambda})e^{\Lambda(t-T)}\right]x^2, \tag{4}$$

$$u_{CL}(t,x) = -(\sum_{j=1}^m D_j D_j^\top)^{-1}(B + \sum_{j=1}^m C_j D_j)x, \tag{5}$$

where

$$\begin{aligned}\Lambda = &-2A + 2B^\top(\sum_{j=1}^m D_j D_j^\top)^{-1}B + 4B^\top(\sum_{j=1}^m D_j D_j^\top)^{-1}(\sum_{j=1}^m C_j D_j)\\ &- \sum_{j=1}^m C_j^2 + 2(\sum_{j=1}^m C_j D_j)^\top(\sum_{j=1}^m D_j D_j^\top)^{-1}(\sum_{j=1}^m C_j D_j)\\ &- \sum_{j=1}^m \left(D_j^\top(\sum_{j=1}^m D_j D_j^\top)^{-1}B + D_j^\top(\sum_{j=1}^m D_j D_j^\top)^{-1}(\sum_{j=1}^m C_j D_j)\right)^2.\end{aligned}$$

We remark that if $x$ is of a higher dimension and/or there is a control running reward, then the optimal policy will depend on the solution of the differential Riccati equation and hence become time-dependent, for which our current method fails.

## 2.2 RL Theory for LQ

In most real-life problems, it is often unrealistic to assume precise knowledge of the parameters such as $A$, $B$, $C_j$, and $D_j$. These problems call for RL which differs fundamentally from the traditional estimate-then-optimize methods. The essence of RL is to strike an exploration–exploitation balance by strategically exploring the unknown environment (Sutton and Barto, 2018). To achieve this, RL employs randomized controls to capture exploration where control processes $u$ are sampled from a process $\pi = \{\pi(\cdot, t) \in \mathcal{P}(\mathbb{R}^l) : 0 \leqslant t \leqslant T\}$ of probability distributions with $\mathcal{P}(\mathbb{R}^l)$ being the space of all probability density functions over $\mathbb{R}^l$, and adds an entropy term in the objective function to encourage exploration. Such an entropy regularization is linked to soft-max approximation and Boltzmann exploration (Haarnoja et al., 2018; Ziebart et al., 2008). Wang et al. (2020) is the first to present a rigorous mathematical formulation of entropy regularized RL for (continuous-time) controlled diffusion processes.

Following Wang et al. (2020), under a given randomized control $\pi$ the dynamic of the LQ RL satisfies SDE:

$$\mathrm{d}x^\pi(t) = \widetilde{b}(x^\pi(t), \pi(\cdot, t))\mathrm{d}t + \sum_{j=1}^{m} \widetilde{\sigma}_j(x^\pi(t), \pi(\cdot, t))\mathrm{d}W^{(j)}(t), \tag{6}$$

where

$$\widetilde{b}(x, \psi) := Ax + B^\top \int_{\mathbb{R}^l} u\psi(u)\mathrm{d}u, \tag{7}$$

$$\widetilde{\sigma}_j(x, \psi) := \sqrt{\int_{\mathbb{R}^l} (C_j x + D_j^\top u)^2 \psi(u)\mathrm{d}u}, \quad (x, \psi) \in \mathbb{R} \times \mathcal{P}(\mathbb{R}^l). \tag{8}$$

The entropy-regularized value function of $\pi$ is

$$J(t, x; \pi) = \mathbb{E}\left[\int_t^T \left(-\frac{1}{2}Qx^\pi(s)^2 + \gamma p^\pi(s)\right)\mathrm{d}s - \frac{1}{2}Hx^\pi(T)^2 \Big| x^\pi(t) = x\right], \tag{9}$$

where $p^\pi(t) = -\int_{\mathbb{R}^l} \pi(t, u)\log\pi(t, u)\mathrm{d}u$ is the differential entropy of $\pi$ and $\gamma \geqslant 0$, known as the temperature parameter, is the weight on exploration. The optimal value function is then

$$V(t, x) = \max_\pi J(t, x; \pi). \tag{10}$$

By Wang et al. (2020), the optimal value function and optimal randomized/stochastic (feedback) policy are determined by

$$V(t, x) = -\frac{1}{2}k_1(t)x^2 + k_3(t), \tag{11}$$

$$\pi(u \mid t, x) = \mathcal{N}\left(u \middle| -(\sum_{j=1}^{m} D_j D_j^\top)^{-1}(B + \sum_{j=1}^{m} C_j D_j)x, \frac{\gamma}{k_1(t)}(\sum_{j=1}^{m} D_j D_j^\top)^{-1}\right), \tag{12}$$

where $k_1 > 0$ and $k_3$ are certain functions of $t$ that can be determined completely by the model primitives, and $\mathcal{N}(\cdot|\mu, \Sigma)$ is the multivariate Gaussian density with mean $\mu$ and covariance $\Sigma$. These *theoretical* results cannot be used to compute the solution of the exploratory problem because the model parameters are unknown, yet they reveal the *structure* of the solution inherent to LQ RL (i.e. the optimal value function is quadratic in $x$ and optimal stochastic policy is Gaussian) that can be utilized to significantly reduce the complexity of function parameterization/approximation in learning.

Throughout this paper (including the appendices) we use $c$ or its variants for generic constants (depending only on the model parameters $A$, $B$, $C_j$, $D_j$, $Q$, $H$, $x_0$, $T$, $\gamma$, $m$ and $l$) whose values may change from line to line.

# 3 A Continuous-Time RL Algorithm

This section presents the steps of designing a continuous-time RL algorithm for solving our LQ problem, including function parameterization, policy evaluation and policy gradient.[1] We will introduce various techniques such as exploration scheduling and projection for deriving the convergence rate of the policy parameter and the regret bound. We will also describe time discretization for final implementation.

## 3.1 Function Parameterization

Inspired by (11) and (12), we parameterize the value function with parameters $\boldsymbol{\theta} \in \mathbb{R}^d$:

$$J(t, x; \boldsymbol{\theta}) = -\frac{1}{2}k_1(t; \boldsymbol{\theta})x^2 + k_3(t; \boldsymbol{\theta}), \tag{13}$$

and parameterize the policy with parameters $\boldsymbol{\phi} = (\phi_1, \phi_2)^\top$:

$$\pi(u \mid x; \boldsymbol{\phi}) = \mathcal{N}(u \mid \phi_1 x, \phi_2), \tag{14}$$

where $(\phi_1, \phi_2) \in \mathbb{R}^l \times \mathbb{S}^l_{++}$.

[1] As will be explained below, policy evaluation is actually not necessary for the LQ problem considered in this paper. However, we still include it in the discussion and the algorithm for future extension to general problems where policy evaluation is generally needed and indeed a crucial step.

Note that (12) suggests that the optimal feedback policy is time-dependent, whose variance depends explicitly on $t$. In our parameterization, the time-dependent variance of the Gaussian policies is replaced by a decaying schedule, called an *exploration schedule*, of $\phi_2$ as a function of the number of iterations, to be presented shortly.

Henceforth we assume that there are positive constants $c_1, c_2, c_3$ such that $1/c_2 \leqslant k_1(t;\boldsymbol{\theta}) \leqslant c_2$, $|k_1'(t;\boldsymbol{\theta})| \leqslant c_1$ and $|k_3'(t;\boldsymbol{\theta})| \leqslant c_3$, for any $0 \leqslant t \leqslant T$. These assumptions are consistent with the fact that the corresponding functions satisfy the same conditions when the model parameters are known.

### 3.2 Policy Evaluation

Policy evaluation (PE) is generally a key step in RL to learn the value function of a *given* control policy.

The general continuous-time PE method developed in (Jia and Zhou, 2022a) dictates that one first parameterizes the value function $J(\cdot,\cdot;\pi)$ and the policy $\pi$ by (13) and (14) respectively (with a slight abuse of notation), with the corresponding $p^\pi(t) = p(t;\boldsymbol{\phi})$, and then updates $\boldsymbol{\theta}$ in an offline learning setting:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \int_0^T \frac{\partial J}{\partial \boldsymbol{\theta}}(t, x(t); \boldsymbol{\theta}) \left[ \mathrm{d}J(t, x(t); \boldsymbol{\theta}) - \left( \frac{1}{2} Q x(t)^2 \mathrm{d}t - \gamma p(t;\boldsymbol{\phi}) \mathrm{d}t \right) \right], \tag{15}$$

where $\alpha$ is the learning rate.

Intriguingly, however, our subsequent theoretical proofs indicate that the convergence and regret results depend only on the bounds (i.e. the constants $c_1$, $c_2$ and $c_3$) of the functions $k_1$ and $k_2$, *not* on the specific forms of these functions. This feature is due to the special class of LQ control problems we are tackling. As a result, in our numerical experiments we actually fix a value function (or equivalently $\boldsymbol{\theta}$) throughout without updating it.

### 3.3 Policy Iteration

Having learned the value function associated with a Gaussian policy, the next step is to improve the policy by updating $\boldsymbol{\phi} = (\phi_1, \phi_2)^\top$. For $\phi_1$, we employ the continuous-time policy gradient (PG) method established in (Jia and Zhou, 2022b) to get the following updating rule:

$$\phi_1 \leftarrow \phi_1 + \alpha Z_1(T), \tag{16}$$

where $\alpha$ is the learning rate, and

$$Z_1(s) = \int_0^s \left\{ \frac{\partial \log \pi}{\partial \phi_1} \left( u(t) \mid x(t); \boldsymbol{\phi} \right) \Big[ \mathrm{d}J\left(t, x(t); \boldsymbol{\theta}\right) - \frac{1}{2} Q x(t)^2 \mathrm{d}t \right.$$
$$\left. + \gamma p\left(t, \boldsymbol{\phi}\right) \mathrm{d}t \Big] + \gamma \frac{\partial p}{\partial \phi_1}\left(t, \boldsymbol{\phi}\right) \mathrm{d}t \right\}, \quad 0 \leqslant s \leqslant T. \tag{17}$$

As discussed earlier, the other parameter, $\phi_2$, controls the level of exploration. In our algorithm, we set a deterministic schedule of this parameter which decreases to 0 as the number of iterations grows. Specifically, we set $\phi_{2,n} = \frac{I_l}{b_n}$ where $I_l$ is the identity matrix of dimension $l$ and $b_n \uparrow \infty$ is specified in Theorem 1 below. The order of $b_n$ in iteration $n$ is carefully chosen along with those of the other hyperparameters, such as the learning rates, in order to achieve the desired sublinear regret bound of the RL algorithm.

### 3.4 Projections

Our updating rules for the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are types of stochastic approximation (SA), a technique pioneered by (Robbins and Monro, 1951). To tailor the general SA algorithms to our specific requirements—primarily to circumvent issues like extreme state values and unbounded estimation errors—we include projection, a technique originally proposed by (Andradóttir, 1995). The projection maps do not depend on prior environmental knowledge, allowing our method to remain model-free while ensuring that the learning regions expand to cover the entire parameter space over time.

Define $\Pi_K(x) := \arg\min_{y \in K} |y - x|^2$ to be a general projection mapping a point $x$ onto a given set $K$. Let

$$\begin{aligned} K_{\boldsymbol{\theta},n} &= \left\{ \boldsymbol{\theta}_n \in \mathbb{R}^d \middle| 1/c_2 \leqslant k_1(t; \boldsymbol{\theta}_n) \leqslant c_2, |k_1'(t; \boldsymbol{\theta}_n)| \leqslant c_1, |k_3'(t; \boldsymbol{\theta}_n)| \leqslant c_3 \right\}, \\ K_{1,n} &= \left\{ \phi_{1,n} \in \mathbb{R}^l \middle| |\phi_{1,n}| \leqslant c_{1,n} \right\}, \end{aligned} \tag{18}$$

where $c_1, c_2, c_3$ are hyperparameters, and $\{c_{1,n}\}$ is an increasing sequence to be specified in Theorem 1 below. Note here the choice of $K_{\boldsymbol{\theta},n}$ is specific to the special class of LQ problems under consideration – it is independent of $n$ as the regret analysis does not rely on the convergence of $\boldsymbol{\theta}_n$. For a general problem, $K_{\boldsymbol{\theta},n}$ needs to be an expanding sequence of sets.

With projection the updating rules for $\boldsymbol{\theta}$ and $\phi_1$ in (15) and (16) are modified to

$$\begin{aligned}\boldsymbol{\theta}_{n+1} \leftarrow \Pi_{K_{\boldsymbol{\theta},n+1}}\Bigg(\boldsymbol{\theta}_n + a_n \int_0^T \frac{\partial J}{\partial \boldsymbol{\theta}}(t, x_n(t); \boldsymbol{\theta}_n) \\ \left[\mathrm{d}J\left(t, x_n(t); \boldsymbol{\theta}_n\right) - \frac{1}{2} Q x_n(t)^2 \mathrm{d}t + \gamma p\left(t, \boldsymbol{\phi}_n\right) \mathrm{d}t\right]\Bigg),\end{aligned} \tag{19}$$

$$\phi_{1,n+1} \leftarrow \Pi_{K_{1,n+1}}\Big(\phi_{1,n} + a_n Z_{1,n}(T)\Big), \tag{20}$$

where

$$\begin{aligned}Z_{1,n}(s) = \int_0^s \Bigg\{ \frac{\partial \log \pi}{\partial \phi_1}\left(u_n(t) \mid x_n(t); \boldsymbol{\phi}_n\right) \Big[\mathrm{d}J\left(t, x_n(t); \boldsymbol{\theta}_n\right) - \frac{1}{2} Q x_n(t)^2 \mathrm{d}t \\ + \gamma p\left(t, \boldsymbol{\phi}_n\right) \mathrm{d}t\Big] + \gamma \frac{\partial p}{\partial \phi_1}\left(t, \boldsymbol{\phi}_n\right) \mathrm{d}t \Bigg\}, \quad 0 \leqslant s \leqslant T.\end{aligned} \tag{21}$$

### 3.5 Discretization

Our approach for continuous-time RL is characterized by carrying out the entire analysis in the continuous-time setting and discretizing time only at the final implementation stage. The iterations in (19) and (20) involve integrals that can be computed only by approximated discretized summations as well as the $\mathrm{d}J$ term that can be approximated by the temporal difference between two consecutive time steps. [2] We therefore discretize the interval $[0, T]$ into uniform time intervals of length $\Delta t$, leading to the following schemes:

$$\begin{aligned}\boldsymbol{\theta}_{n+1} \leftarrow \Pi_{K_{\boldsymbol{\theta},n+1}}\Bigg(\boldsymbol{\theta}_n + a_n \sum_{k=0}^{\lfloor \frac{T}{\Delta t} - 1 \rfloor} \frac{\partial J}{\partial \boldsymbol{\theta}}(t_k, x_n(t_k); \boldsymbol{\theta}_n) \Big[J\left(t_{k+1}, x_n(t_{k+1}); \boldsymbol{\theta}_n\right) \\ - J\left(t_k, x_n(t_k); \boldsymbol{\theta}_n\right) - \frac{1}{2} Q x_n(t_k)^2 \Delta t + \gamma p\left(t_k, \boldsymbol{\phi}_n\right) \Delta t\Big]\Bigg),\end{aligned} \tag{22}$$

$$\begin{aligned}\phi_{1,n+1} \leftarrow \Pi_{K_{1,n+1}}\Bigg(\phi_{1,n} + a_n \sum_{k=0}^{\lfloor \frac{T}{\Delta t} - 1 \rfloor} \Bigg\{ \frac{\partial \log \pi}{\partial \phi_1}\left(u_n(t_k) \mid t_k, x_n(t_k); \boldsymbol{\phi}_n\right) \\ \Big[J\left(t_{k+1}, x_n(t_{k+1}); \boldsymbol{\theta}_n\right) - J\left(t_k, x_n(t_k); \boldsymbol{\theta}_n\right) - \frac{1}{2} Q x_n(t_k)^2 \Delta t \\ + \gamma p\left(t_k, \boldsymbol{\phi}_n\right) \Delta t\Big] + \gamma \frac{\partial p}{\partial \phi_1}\left(t_k, \boldsymbol{\phi}_n\right) \Delta t \Bigg\}\Bigg).\end{aligned} \tag{23}$$

[2] The discretization errors will be taken into consideration in the appendices; see Theorem 3, Theorem 4, and Remark 2.

### 3.6 RL-LQ Algorithm

The analysis above leads to the following RL algorithm for the LQ problem:

**Algorithm 1** RL-LQ Algorithm

**Input**

| | |
|---|---|
| $\boldsymbol{\theta}_0$, $\phi_{1,0}$ | Initial values of trainable parameters for value function and policy. |
| $\phi_{2,n}$ | Deterministic sequence of $\phi_{2,n} = \frac{I_l}{b_n}$ specified in Theorem 1. |

**for** $n = 1$ `to` $N$ **do**

Initialize $k = 0$, time $t = t_k = 0$, state $x_n(t_k) = x_0$.

**while** $t < T$ **do**

Generate action $u_n(t_k) \sim \boldsymbol{\pi}\left(\cdot \mid t_k, x_n(t_k); \boldsymbol{\phi}_n\right)$ following policy (14) .

Apply action $u_n(t_k)$ and get new state $x_n(t_{k+1})$ by dynamic (1).

Update time $t_{k+1} \leftarrow t_k + \Delta t$ and $t \leftarrow t_{k+1}$.

**end while**

Collect whole trajectory $\{(t_k, x_n(t_k), u_n(t_k))\}_{k \geqslant 0}$.

Update value function parameters $\boldsymbol{\theta}$ using (22).

Update policy parameter $\phi_1$ using (23).

**end for**

**Output**

| | |
|---|---|
| $\boldsymbol{\theta}_N, \phi_{1,N}, \phi_{2,N}$ | Parameters for value function and policy. |

## 4 Regret Analysis

This section presents the main result of the paper – a sublinear regret bound of the RL-LQ algorithm, Algorithm 1. For that, we need to first examine the convergence property and convergence rate of the parameter $\phi_{1,n}$, whose analysis forms the theoretical underpinning of the algorithm. Proofs of the results in this section are provided in Appendices A and B.

### 4.1 Convergence of $\phi_{1,n}$

The following theorem shows the convergence and convergence rate of the parameter $\phi_{1,n}$.

**Theorem 1.** *In Algorithm 1, let the hyperparameters $c_1$, $c_2$, $c_3$ and $\gamma$ be fixed positive constants. Set*

$$a_n = \frac{\alpha^{\frac{3}{4}}}{(n+\beta)^{\frac{3}{4}}}, \quad b_n = 1 \vee \frac{(n+\beta)^{\frac{1}{4}}}{\alpha^{\frac{1}{4}}}, \quad c_{1,n} = 1 \vee (\log\log n)^{\frac{1}{6}},$$

*where $\alpha > 0$ and $\beta > 0$ are constants. Then,*

*(a) as $n \to \infty$, $\phi_{1,n}$ converges almost surely to*

$$\phi_1^* = -(\sum_{j=1}^{m} D_j D_j^\top)^{-1}(B + \sum_{j=1}^{m} C_j D_j).$$

*(b) for any $n$, $\mathbb{E}[|\phi_{1,n} - \phi_1^*|^2] \leqslant c\frac{(\log n)^p (\log\log n)^{\frac{4}{3}}}{n^{\frac{1}{2}}}$, for some positive constants $c$ and $p$.*

These results ensure the convergence of the learned policy. Moreover, it is a prerequisite for deriving the regret bound of Algorithm 1.

### 4.2 Regret Bound

A regret bound measures the cumulative derivation (over episodes) of the value functions of the learned policies from the oracle optimal value function. A sublinear regret bound guarantees an almost optimal performance of the RL policy in the long run.

Denote

$$\bar{J}(\phi_1, \phi_2) = \mathbb{E}\left[\int_0^T \left(-\frac{1}{2}Qx^\pi(s)^2\right) \mathrm{d}s - \frac{1}{2}Hx^\pi(T)^2 \Big| x^\pi(0) = x_0\right], \tag{24}$$

where $\pi = \mathcal{N}(\cdot|\phi_1 x, \phi_2)$.

So $\bar{J}(\phi_1, \phi_2)$ is the value of the Gaussian policy $\mathcal{N}(\cdot|\phi_1 x, \phi_2)$ assessed using the *original* objective function (i.e. one *without* the entropy regularization term). Clearly, $\bar{J}(\phi_1^*, 0)$ is the oracle value of the original problem.

**Theorem 2.** *Under the assumptions of Theorem 1, applying Algorithm 1 results in a cumulative regret bound over $N$ episodes given by:*

$$\sum_{n=1}^{N} \mathbb{E}[\bar{J}(\phi_1^*, 0) - \bar{J}(\phi_{1,n}, \phi_{2,n})] \leqslant c + cN^{\frac{3}{4}}(\log N)^{\frac{p+1}{2}}(\log\log N)^{\frac{2}{3}},$$

*where $c > 0$ is a constant independent of $N$, and $p$ is the same constant appearing in Theorem 1.*

# 5 Numerical Experiments

This section reports the results of numerically evaluating the convergence rate of $\phi_{1,n}$ and the sublinear regret bound of our RL-LQ algorithm, compared with a benchmark algorithm. The benchmark adapts the model-based methods in (Basei et al., 2022; Szpruch et al., 2024) to our setting of state- and control-dependent volatility.

## 5.1 Simulation Setup

In our simulation study, we consider the case where $l = 1$ for the control dimension and $m = 1$ for the Brownian motion dimension. The controlled system (1) simplifies to the following form:

$$\mathrm{d}x^u(t) = (Ax^u(t) + Bu(t))\mathrm{d}t + (Cx^u(t) + Du(t))\mathrm{d}W(t). \tag{25}$$

In addition, we set the model parameters $A, B, C, D, Q, H, x_0, T$ to be all 1, and set the exploration schedule $b_n = 0.2(n+1)^{1/4}$. Other sequences such as that of the learning rate $\{a_n\}$ are configured according to the assumptions stated in Theorem 1 and Subsection 3.1; for details see Appendix C.2. In each experiment we execute both our proposed Algorithm 1 and the benchmark Algorithm 2 over $N = 400,000$ iterations, while we replicate the experiment independently for 120 times to draw statistical conclusions.

## 5.2 A Modified Model-Based Algorithm

The algorithms proposed in (Basei et al., 2022; Szpruch et al., 2024) are designed to estimate parameters $A$ and $B$ in the drift term under the constant volatility setting. To compare with our algorithm tailored for state- and control-dependent volatilities, we extend their methods to include estimating the parameters $C$ and $D$. The details of this modified algorithm are described in Appendix C.1.

## 5.3 Analysis of Numerical Results

Figures 1 and 2 compare the mean-squared convergence rates of $\phi_{1,n}$ for our model-free Algorithm 1 and the model-based Algorithm 2, using a log-log plot of Mean Squared Error (MSE) versus iterations. The fitted linear regression shows our model's slope of $-0.5$, confirming Theorem 1 and outperforming the model-based benchmark slope of $-0.08$.

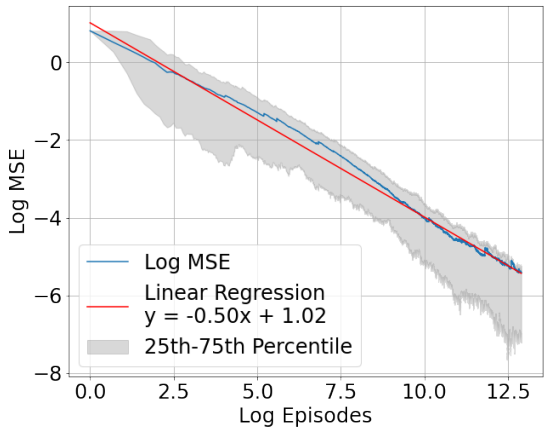


Figure 1: Log-log plot of MSE of learned $\phi_{1,n}$ in Algorithm 1

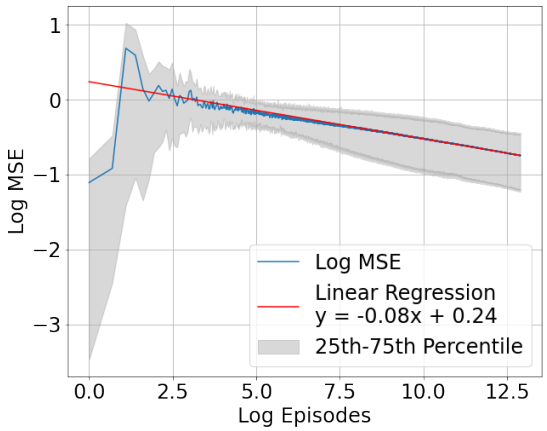


Figure 2: Log-log plot of MSE of learned $\phi_{1,n}$ in the benchmark algorithm

A comparison of regrets between Algorithms 1 and 2 is presented in Figures 3 and 4. The former yields a regret slope of around 0.73, which is close to the theoretical bound stipulated in Theorem 2 and superior to the slope of 0.88 achieved by the latter.

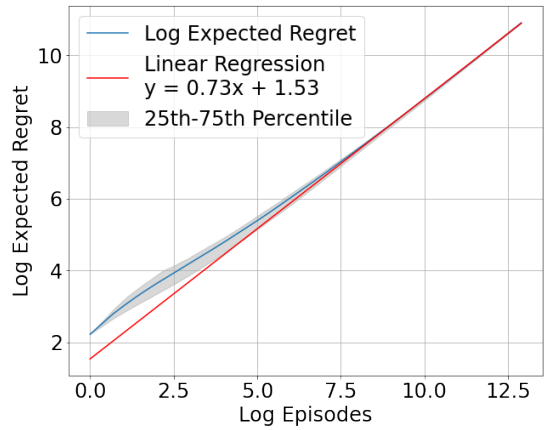


Figure 3: Log-log plot of the regret of Algorithm 1

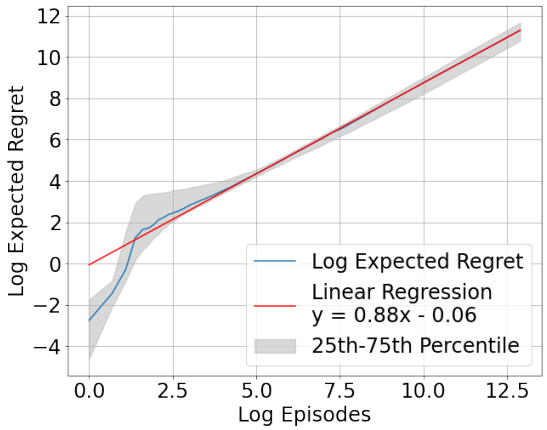


Figure 4: Log-log plot of the regret of the benchmark algorithm

These experimental results support the theoretical claims and demonstrate the outperformance of our RL-LQ algorithm compared with its model-based counterpart in terms of both the convergence rates of the policy parameters and the regret bounds.

## 6 Conclusions

This paper is the first to derive a convergence rate and a regret bound within the model-free framework of continuous-time entropy-regularized RL for controlled diffusion processes initiated by Wang et al. (2020). Here, by model-free, we mean that neither theory nor algorithm involves estimating model parameters. While it deals with the LQ case, it treats the case in which the diffusion term depends both on state and control, one that has not been studied in the RL literature to our best knowledge.

There are several limitations in the setting or results of the paper. First, the state is one-dimensional and the quadratic objective functional (2) has no running reward from controls, which are key assumptions needed to simplify our analysis so that it suffices to consider only (time-invariant) stationary policies. While these assumptions are satisfied in some applications (e.g. in portfolio choice), imposing them is far from satisfactory. We hope that the present paper represents the *first step* towards solving the general LQ problem and some of the ideas here can inspire a more general convergence/regret analysis for continuous-time RL. Second, we are unable to achieve a better sublinear regret, e.g., a square-root one, which is typical in episodic RL algorithms for tabular or linear MDPs. We are not certain whether that is due to our approach or it is more fundamental due to the diffusion nature of the system dynamics. Finally, extending the analysis to non-LQ problems with general function approximations is an enormous open question. All these point to exciting research opportunities in the (hopefully near) future.

## References


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

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

Anderson, B. D. and Moore, J. B. (2007). *Optimal control: Linear quadratic methods*. Courier Corporation.

Andradóttir, S. (1995). A stochastic approximation algorithm with varying bounds. *Operations Research*, 43(6):1037–1048.

Baird, L. C. (1994). Reinforcement learning in continuous time: Advantage updating. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 4, pages 2448–2453. IEEE.

Basei, M., Guo, X., Hu, A., and Zhang, Y. (2022). Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon. *Journal of Machine Learning Research*, 23(178):1–34.

Bradtke, S. (1992). Reinforcement learning applied to linear quadratic regulation. *Advances in Neural Information Processing Systems*, 5.

Broadie, M., Cicek, D., and Zeevi, A. (2011). General bounds and finite-time improvement for the Kiefer-Wolfowitz stochastic approximation algorithm. *Operations Research*, 59(5):1211–1224.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. (2020). Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR.

Cassel, A. B. and Koren, T. (2021). Online policy gradient for model free learning of linear quadratic regulators with $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pages 1304–1313. PMLR.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. (2022). Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578.

Chen, S., Li, X., and Zhou, X. Y. (1998). Stochastic linear quadratic regulators with indefinite control weight costs. *SIAM Journal on Control and Optimization*, 36(5):1685–1702.

Chen, X., Duan, J., Liang, Y., and Zhao, L. (2023). Global convergence of two-timescale actor-critic for solving linear quadratic regulator. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7087–7095.

Cohen, A., Hasidim, A., Koren, T., Lazic, N., Mansour, Y., and Talwar, K. (2018). Online linear quadratic control. In *International Conference on Machine Learning*, pages 1029–1038. PMLR.

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

Dai, M., Dong, Y., and Jia, Y. (2023). Learning equilibrium mean-variance strategy. *Mathematical Finance*, 33(4):1166–1212.

Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural Computation*, 12(1):219–245.

Fazel, M., Ge, R., Kakade, S., and Mesbahi, M. (2018). Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR.

Frikha, N., Germain, M., Laurière, M., Pham, H., and Song, X. (2023). Actor-critic learning for mean-field control in continuous time. *arXiv preprint arXiv:2303.06993*.

Gu, S., Holly, E., Lillicrap, T., and Levine, S. (2017). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3389–3396. IEEE.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870. PMLR.

Hambly, B., Xu, R., and Yang, H. (2021). Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *SIAM Journal on Control and Optimization*, 59(5):3359–3391.

Huang, Y., Jia, Y., and Zhou, X. (2022). Achieving mean–variance efficiency by continuous-time reinforcement learning. In *Proceedings of the Third ACM International Conference on AI in Finance*, pages 377–385.

Jia, Y. and Zhou, X. Y. (2022a). Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. *Journal of Machine Learning Research*, 23(154):1–55.

Jia, Y. and Zhou, X. Y. (2022b). Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *Journal of Machine Learning Research*, 23(154):1–55.

Jia, Y. and Zhou, X. Y. (2023). q-Learning in continuous time. *Journal of Machine Learning Research*, 24(161):1–61.

Khan, M. A.-M., Khan, M. R. J., Tooshil, A., Sikder, N., Mahmud, M. P., Kouzani, A. Z., and Nahid, A.-A. (2020). A systematic review on reinforcement learning-based robotics within the last decade. *IEEE Access*, 8:176598–176623.

Kim, J., Shin, J., and Yang, I. (2021). Hamilton-Jacobi deep Q-learning for deterministic continuous-time systems with Lipschitz continuous controls. *Journal of Machine Learning Research*, 22(206):1–34.

Kloeden, P. E. and Platen, E. (1992). Numerical solution of stochastic differential equations.

Konda, V. and Tsitsiklis, J. (1999). Actor-critic algorithms. *Advances in Neural Information Processing Systems*, 12.

Lee, J. and Sutton, R. S. (2021). Policy iterations for reinforcement learning problems in continuous time and space—Fundamental theory and methods. *Automatica*, 126:109421.

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

Luenberger, D. G. (1998). *Investment Science*. Oxford University Press.

Malik, D., Pananjady, A., Bhatia, K., Khamaru, K., Bartlett, P., and Wainwright, M. (2019). Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2916–2925. PMLR.

Merton, R. C. (1980). On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics*, 8(4):323–361.

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

Munos, R. (2006). Policy gradient in continuous time. *Journal of Machine Learning Research*, 7:771–791.

Park, S., Kim, J., and Kim, G. (2021). Time discretization-invariant safe action repetition for policy gradient methods. *Advances in Neural Information Processing Systems*, 34:267–279.

Rami, M. and Zhou, X. Y. (2000). Linear matrix inequalities, Riccati equations, and indefinite stochastic linear quadratic controls. *IEEE Transactions on Automatic Control*, 45(6):1131–1143.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407.

Robbins, H. and Siegmund, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. Elsevier.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., and Lanctot, M. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., and Bolton, A. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359.

Simchowitz, M. and Foster, D. (2020). Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pages 8937–8948. PMLR.

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

Szpruch, L., Treetanthiploet, T., and Zhang, Y. (2024). Optimal scheduling of entropy regularizer for continuous-time linear-quadratic reinforcement learning. *SIAM Journal on Control and Optimization*, 62(1):135–166.

Tallec, C., Blier, L., and Ollivier, Y. (2019). Making deep Q-learning methods robust to time discretization. In *International Conference on Machine Learning*, pages 6096–6104. PMLR.

Vamvoudakis, K. G. and Lewis, F. L. (2010). Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica*, 46(5):878–888.

Wang, B., Gao, X., and Li, L. (2023). Reinforcement learning for continuous-time optimal execution: actor-critic algorithm and error analysis. *Available at SSRN 4378950*.

Wang, H., Zariphopoulou, T., and Zhou, X. Y. (2020). Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, 21(198):1–34.

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

Wei, X. and Yu, X. (2023). Continuous-time q-learning for McKean-Vlasov control problems. *arXiv preprint arXiv:2306.16208*.

Wu, Y. F., Zhang, W., Xu, P., and Gu, Q. (2020). A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628.

Xu, T., Yang, Z., Wang, Z., and Liang, Y. (2021). Doubly robust off-policy actor-critic: Convergence and optimality. In *International Conference on Machine Learning*, pages 11581–11591. PMLR.

Yang, Z., Chen, Y., Hong, M., and Wang, Z. (2019). Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. *Advances in Neural Information Processing Systems*, 32.

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

Zhong, H. and Zhang, T. (2023). A theoretical analysis of optimistic proximal policy optimization in linear Markov decision processes. *Advances in Neural Information Processing Systems*, 36.

Zhou, M. and Lu, J. (2023). Single timescale actor-critic method to solve the linear quadratic regulator with convergence guarantees. *Journal of Machine Learning Research*, 24(222):1–34.

Zhou, X. Y. and Li, D. (2000). Continuous-time mean-variance portfolio selection: A stochastic LQ framework. *Applied Mathematics and Optimization*, 42(1):19–33.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., and Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA.

# A A Proof of Theorem 1

Let $x_n = \{x_n(t) : 0 \leqslant t \leqslant T\}$ be the sample state trajectory in the $n$-th iteration that follows the dynamics:

$$\mathrm{d}x_n(t) = (Ax_n(t) + B^\top u_n(t))\mathrm{d}t + \sum_{j=1}^{m}(C_j x_n(t) + D_j^\top u_n(t))\mathrm{d}W_n^{(j)}(t), \quad 0 \leqslant t \leqslant T, \tag{26}$$

where $W_n = \{(W_n^{(1)}(t), \cdots, W_n^{(m)}(t))^\top \in \mathbb{R}^m : 0 \leqslant t \leqslant T\}$ is a standard Brownian motions in the $n$-th iteration, and the policy $u_n(t) \mid x_n(t) \sim \mathcal{N}(\cdot|\phi_{1,n}x_n(t), \phi_{2,n})$ independent of $W_n$.

Recall $Z_1(\cdot)$ defined by (17), and $Z_{1,n}(T)$ defined by (21) as the value of $Z_1(T)$ at the $n$-th iteration. The expectation of $Z_{1,n}(T)$ conditioned on the parameters is denoted by

$$h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n) = \mathbb{E}[Z_{1,n}(T) \mid \boldsymbol{\theta}_n, \boldsymbol{\phi}_n],$$

and the noise contained in $Z_{1,n}(T)$ is defined as

$$\xi_{1,n} = Z_{1,n}(T) - h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n).$$

Hence, the updating rule for $\phi_1$ is given by:

$$\phi_{1,n+1} = \Pi_{K_{1,n+1}}(\phi_{1,n} + a_n[h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n) + \xi_{1,n}]). \tag{27}$$

To prove Theorem 1, we need a series of lemmas to adapt the general stochastic approximation techniques and results (e.g. (Andradóttir, 1995) and (Broadie et al., 2011)) to our specific setting.

## A.1 Moment Estimates

Let $\{x^{\boldsymbol{\phi}}(t) : 0 \leqslant t \leqslant T\}$ be the state process under the policy (14) following the dynamic (6). The following lemma gives some moment estimates of $x^{\boldsymbol{\phi}}(t)$ in terms of $\boldsymbol{\phi} = (\phi_1, \phi_2)$.

**Lemma 1.** *There exists a constant $c > 0$ (that only depends on $A, B, C, D$) such that*

$$\mathbb{E}[x^{\boldsymbol{\phi}}(t)] = x_0 e^{(A+B^\top\phi_1)t},$$

$$\mathbb{E}[x^{\boldsymbol{\phi}}(t)^2] = (\sum_{j=1}^{m} D_j^\top \phi_2 D_j)\int_0^t e^{a(\phi_1)(t-s)}\mathrm{d}s + x_0^2 e^{a(\phi_1)t},$$

*where*

$$a(\phi_1) = 2A + 2B^\top \phi_1 + \sum_{j=1}^{m} (C_j^2 + 2C_j D_j^\top \phi_1 + D_j^\top \phi_1 \phi_1^\top D_j). \tag{28}$$

*Moreover, we have*

$$\mathbb{E}[x^{\boldsymbol{\phi}}(t)^2] \leqslant c(1 + |\phi_2|t) \exp\{c|\phi_1|^2 t\},$$

$$\mathbb{E}[x^{\boldsymbol{\phi}}(t)^4] \leqslant c(1 + |\phi_2|^2 t) \exp\{c|\phi_1|^4 t\},$$

$$\mathbb{E}[x^{\boldsymbol{\phi}}(t)^6] \leqslant c(1 + |\phi_2|^3 t) \exp\{c|\phi_1|^6 t\}.$$

*Proof.* We have

$$\begin{aligned} x^{\boldsymbol{\phi}}(t) = x(0) &+ \int_0^t \left( A x^{\boldsymbol{\phi}}(s) + B^\top \phi_1 x^{\boldsymbol{\phi}}(s) \right) \mathrm{d}s \\ &+ \int_0^t \sum_{j=1}^{m} \sqrt{C_j^2 x^{\boldsymbol{\phi}}(s)^2 + 2C_j D_j^\top \phi_1 x^{\boldsymbol{\phi}}(s)^2 + D_j^\top (\phi_1 \phi_1^\top x^{\boldsymbol{\phi}}(s)^2 + \phi_2) D_j} \mathrm{d}W^{(j)}(s). \end{aligned}$$

Taking expectation on both sides, we have

$$\mathbb{E}[x^{\boldsymbol{\phi}}(t)] = x_0 + \int_0^t (A\mathbb{E}[x^{\boldsymbol{\phi}}(s)] + B^\top \phi_1 \mathbb{E}[x^{\boldsymbol{\phi}}(s)]) \mathrm{d}s,$$

leading to

$$\mathbb{E}[x^{\boldsymbol{\phi}}(t)] = x_0 e^{(A + B^\top \phi_1)t}. \tag{29}$$

Next, applying Ito's formula to $x^{\boldsymbol{\phi}}(t)^2$ and taking expectation on both sides, we obtain

$$\mathbb{E}[x^{\boldsymbol{\phi}}(t)^2] = x_0^2 + \int_0^t \left( a(\phi_1)\mathbb{E}[x^{\boldsymbol{\phi}}(s)^2] + \sum_{j=1}^{m} D_j^\top \phi_2 D_j \right) \mathrm{d}s,$$

yielding

$$\mathbb{E}[x^{\boldsymbol{\phi}}(t)^2] = (\sum_{j=1}^{m} D_j^\top \phi_2 D_j) \int_0^t e^{a(\phi_1)(t-s)} \mathrm{d}s + x_0^2 e^{a(\phi_1)t}. \tag{30}$$

Now we prove the inequality related to $\mathbb{E}[x^{\boldsymbol{\phi}}(t)^6]$. Hölder's inequality yields

$$
\begin{aligned}
&\mathbb{E}[x^{\boldsymbol{\phi}}(t)^6] \\
=&\mathbb{E}\bigg[\bigg(x(0)+\int_0^t Ax^{\boldsymbol{\phi}}(s)+B^\top\phi_1 x^{\boldsymbol{\phi}}(s)\mathrm{d}s \\
&+\int_0^t\sum_{j=1}^m\sqrt{C_j^2x^{\boldsymbol{\phi}}(s)^2+2C_jD_j^\top\phi_1x^{\boldsymbol{\phi}}(s)^2+D_j^\top(\phi_1\phi_1^\top x^{\boldsymbol{\phi}}(s)^2+\phi_2)D_j}\mathrm{d}W^{(j)}(s)\bigg)^6\bigg] \\
\leqslant& cx_0^6+c\mathbb{E}\bigg[\bigg(\int_0^t Ax^{\boldsymbol{\phi}}(s)+B^\top\phi_1x^{\boldsymbol{\phi}}(s)\mathrm{d}s\bigg)^6\bigg]+c\mathbb{E}\bigg[\bigg(\int_0^t\sum_{j=1}^m \\
&\sqrt{C_j^2x^{\boldsymbol{\phi}}(s)^2+2C_jD_j^\top\phi_1x^{\boldsymbol{\phi}}(s)^2+D_j^\top(\phi_1\phi_1^\top x^{\boldsymbol{\phi}}(s)^2+\phi_2)D_j}\mathrm{d}W^{(j)}(s)\bigg)^6\bigg] \\
\leqslant& cx_0^6+c(A+B^\top\phi_1)^6\mathbb{E}\bigg[\bigg(\int_0^t x^{\boldsymbol{\phi}}(s)\mathrm{d}s\bigg)^6\bigg] \\
&+c\mathbb{E}\bigg[\sum_{j=1}^m\bigg(\int_0^t C_j^2x^{\boldsymbol{\phi}}(s)^2+2C_jD_j^\top\phi_1x^{\boldsymbol{\phi}}(s)^2+D_j^\top(\phi_1\phi_1^\top x^{\boldsymbol{\phi}}(s)^2+\phi_2)D_j\mathrm{d}s\bigg)^3\bigg] \\
\leqslant& cx_0^6+c(1+|\phi_1|^6)\mathbb{E}\bigg[\int_0^t x^{\boldsymbol{\phi}}(s)^6\mathrm{d}s\bigg]+c\mathbb{E}\bigg[\int_0^t(1+|\phi_1|^6)x^{\boldsymbol{\phi}}(s)^6+|\phi_2|^3\mathrm{d}s\bigg].
\end{aligned}
$$

It follows from Grönwall's inequality that

$$
\begin{aligned}
\mathbb{E}[x^{\boldsymbol{\phi}}(t)^6]\leqslant& c(1+|\phi_2|^3t)\exp\{\int_0^t c(1+|\phi_1|^6)\mathrm{d}s\} \\
=&c(1+|\phi_2|^3t)\exp\{c(1+|\phi_1|^6)t\} \\
\leqslant& ce^{cT}(1+|\phi_2|^3t)\exp\{c|\phi_1|^6t\}.
\end{aligned}
$$

The proofs for the inequalities of $\mathbb{E}[x^{\boldsymbol{\phi}}(t)^2]$ and $\mathbb{E}[x^{\boldsymbol{\phi}}(t)^4]$ are similar. □

The next lemma concerns the variance of the increment $Z_{1,n}(T)$.

**Lemma 2.** *There exists a constant $c>0$ that depends only on the model primitives such that*

$$
\mathrm{Var}\left(Z_{1,n}(T)\middle|\boldsymbol{\theta}_n,\phi_{1,n},\phi_{2,n}\right)\leqslant cb_n\left(1+|\phi_{1,n}|^8+(\log b_n)^8\right)\exp\{c|\phi_{1,n}|^6\}. \tag{31}
$$

*Proof.* Applying Ito's lemma to the process $J\left(t, x_n(t); \boldsymbol{\theta}_n\right)$, where $x_n$ follows (26), we have

$$
\begin{aligned}
&\mathrm{d}J\left(t, x_n(t); \boldsymbol{\theta}_n\right) \\
=&\bigg(-\frac{1}{2}k_1'(t;\boldsymbol{\theta}_n)x_n(t)^2 + k_3'(t;\boldsymbol{\theta}_n) - (Ax_n(t) + B^\top u_n(t))k_1(t;\boldsymbol{\theta}_n)x_n(t) \\
&- \frac{\sum_{j=1}^m (C_j x_n(t) + D_j^\top u_n(t))^2}{2} k_1(t;\boldsymbol{\theta}_n)\bigg)\mathrm{d}t \\
&- \sum_{j=1}^m \bigg((C_j x_n(t) + D_j^\top u_n(t))k_1(t;\boldsymbol{\theta}_n)x_n(t)\bigg)\mathrm{d}W_n^{(j)}(t).
\end{aligned}
$$

In addition,

$$
p(t, \boldsymbol{\phi}_n) = \frac{1}{2}\log(\det(\phi_{2,n})) + \frac{l}{2}\log(2\pi e).
$$

Hence

$$
\begin{aligned}
\mathrm{d}Z_{1,n}(t) =&\phi_{2,n}^{-1}(u_n(t) - \phi_{1,n}x_n(t))x_n(t)\bigg[\bigg(-\frac{1}{2}k_1'(t;\boldsymbol{\theta}_n)x_n(t)^2 + k_3'(t;\boldsymbol{\theta}_n) \\
&- (Ax_n(t) + B^\top u_n(t))k_1(t;\boldsymbol{\theta}_n)x_n(t) \\
&- \frac{\sum_{j=1}^m (C_j x_n(t) + D_j^\top u_n(t))^2}{2} k_1(t;\boldsymbol{\theta}_n)\bigg)\mathrm{d}t \\
&- \sum_{j=1}^m \bigg((C_j x_n(t) + D_j^\top u_n(t))k_1(t;\boldsymbol{\theta}_n)x_n(t)\bigg)\mathrm{d}W_n^{(j)}(t) - \frac{1}{2}Qx_n(t)^2\mathrm{d}t \\
&+ \gamma\bigg(\frac{1}{2}\log(\det(\phi_{2,n})) + \frac{l}{2}\log(2\pi e)\bigg)\mathrm{d}t\bigg] \\
=&\phi_{2,n}^{-1}(u_n(t) - \phi_{1,n}x_n(t))x_n(t)\bigg[\bigg(-\frac{1}{2}k_1'(t;\boldsymbol{\theta}_n)x_n(t)^2 + k_3'(t;\boldsymbol{\theta}_n) \\
&- (Ax_n(t) + B^\top u_n(t))k_1(t;\boldsymbol{\theta}_n)x_n(t) \\
&- \frac{\sum_{j=1}^m (C_j x_n(t) + D_j^\top u_n(t))^2}{2} k_1(t;\boldsymbol{\theta}_n)\bigg) \\
&- \frac{1}{2}Qx_n(t)^2 + \gamma\bigg(\frac{1}{2}\log(\det(\phi_{2,n})) + \frac{l}{2}\log(2\pi e)\bigg)\bigg]\mathrm{d}t \\
&- \phi_{2,n}^{-1}(u_n(t) - \phi_{1,n}x_n(t))x_n(t) \\
&\quad \sum_{j=1}^m \bigg((C_j x_n(t) + D_j^\top u_n(t))k_1(t;\boldsymbol{\theta}_n)x_n(t)\bigg)\mathrm{d}W_n^{(j)}(t) \\
\triangleq& Z_{1,n}^{(1)}(t)\mathrm{d}t + \sum_{j=1}^m Z_{1,n}^{(2,j)}(t)\mathrm{d}W_n^{(j)}(t).
\end{aligned}
\tag{32}
$$

We now estimate

$$\mathbb{E}[|Z_{1,n}^{(1)}(t)|^2|\boldsymbol{\theta}_n,\phi_{1,n},\phi_{2,n},x_n(t)] \leqslant c\Big[(1+|\phi_{1,n}|^4)|\phi_{2,n}^{-1}|x_n(t)^6+(1+|\phi_{1,n}|^2)x_n(t)^4 + (1+|\phi_{2,n}|^2+(\log(\det(\phi_{2,n})))^4)|\phi_{2,n}^{-1}|x_n(t)^2\Big],$$

and

$$\mathbb{E}[|Z_{1,n}^{(2,j)}(t)|^2|\boldsymbol{\theta}_n,\phi_{1,n},\phi_{2,n},x_n(t)] \leqslant c\Big[1+|\phi_{2,n}^{-1}|(1+|\phi_{1,n}|^2)x_n(t)^6+x_n(t)^4\Big].$$

By Lemma 1, taking expectations in the above with respect to $x_n(t)$, we deduce

$$\begin{aligned}
&\mathbb{E}[|Z_{1,n}^{(1)}(t)|^2+\sum_{j=1}^m|Z_{1,n}^{(2,j)}(t)|^2|\boldsymbol{\theta}_n,\phi_{1,n},\phi_{2,n}]\\
\leqslant& c\Big[(1+|\phi_{1,n}|^4)|\phi_{2,n}^{-1}|(1+|\phi_{2,n}t|^3)\exp\{c|\phi_{1,n}|^6t\}\\
&+(1+|\phi_{1,n}|^2)(1+|\phi_{2,n}t|^2)\exp\{c|\phi_{1,n}|^4t\}\\
&+(1+|\phi_{2,n}|^2+(\log(\det(\phi_{2,n})))^4)|\phi_{2,n}^{-1}|(1+|\phi_{2,n}t|)\exp\{c|\phi_{1,n}|^2t\}\Big].
\end{aligned} \tag{33}$$

Recalling that $\phi_{2,n}=\frac{I_l}{b_n}$ set in Algorithm 1, we arrive at (31). □

## A.2 Mean Increment

We now analyze the mean increment
$h_1(\phi_{1,n},\phi_{2,n};\boldsymbol{\theta}_n)$ in the updating rule (27). First, note that $\int_0^{\cdot}Z_{1,n}^{(2,j)}(t)\mathrm{d}W_n^{(j)}(t)$ is a martingale by virtue of Lemma 1 and (33). Taking integration and expectation in (32), we get

$$\mathbb{E}[Z_{1,n}(s)]=-\int_0^s k_1(t;\boldsymbol{\theta}_n)(B+(\sum_{j=1}^m C_jD_j)+(\sum_{j=1}^m D_jD_j^\top)\phi_{1,n})\mathbb{E}[x_n(t)^2]\mathrm{d}t, \tag{34}$$

where $0\leqslant s\leqslant T$. Hence

$$\begin{aligned}
h_1(\phi_{1,n},\phi_{2,n};\boldsymbol{\theta}_n)&=-(B+(\sum_{j=1}^m C_jD_j)+(\sum_{j=1}^m D_jD_j^\top)\phi_{1,n})\int_0^T k_1(t;\boldsymbol{\theta}_n)\mathbb{E}[x_n(t)^2]\mathrm{d}t\\
&=-l(\phi_{1,n},\phi_{2,n};\boldsymbol{\theta}_n)(\phi_{1,n}-\phi_1^*),
\end{aligned} \tag{35}$$

where

$$l(\phi_{1,n},\phi_{2,n};\boldsymbol{\theta}_n)=(\sum_{j=1}^m D_jD_j^\top)\int_0^T k_1(t;\boldsymbol{\theta}_n)\mathbb{E}[x_n(t)^2]\mathrm{d}t. \tag{36}$$

Next we study $h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n)$. It follows from (28) that $a(\phi_1)$ is a quadratic function of $\phi_1$ and

$$a(\phi_1) \geqslant 2A + \sum_{j=1}^{m} C_j^2 - (B + (\sum_{j=1}^{m} C_j D_j))^\top (\sum_{j=1}^{m} D_j D_j^\top)^{-1} (B + (\sum_{j=1}^{m} C_j D_j)).$$

Hence, by (30), we have

$$\begin{aligned} \mathbb{E}[x_n(t)^2] &= (\sum_{j=1}^{m} D_j^\top \phi_{2,n} D_j) \int_0^t e^{a(\phi_{1,n})(t-s)} \mathrm{d}s + x_0^2 e^{a(\phi_{1,n})t} \\ &\geqslant x_0^2 e^{a(\phi_{1,n})t} \geqslant c, \end{aligned} \tag{37}$$

where $c > 0$ is a constant independent of $n$. Thus,

$$\begin{aligned} l(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n) &= (\sum_{j=1}^{m} D_j D_j^\top) \int_0^T k_1(t; \boldsymbol{\theta}_n) \mathbb{E}[x_n(t)^2] \mathrm{d}t \\ &\geqslant (\sum_{j=1}^{m} D_j D_j^\top) \int_0^T (1/c_2) c \mathrm{d}t \\ &= (\sum_{j=1}^{m} D_j D_j^\top)(1/c_2) cT \geqslant \bar{c} I_l, \end{aligned} \tag{38}$$

where $0 < \bar{c} < 1$ is a constant independent of $n$.

On the other hand, since $a(\phi_1)$ is a quadratic function of $\phi_1$, we have

$$|a(\phi_1)| \leqslant c(1 + |\phi_1|^2), \quad \forall \phi_1 \in \mathbb{R}^l,$$

for some constant $c > 0$. So

$$\begin{aligned} \mathbb{E}[x_n(t)^2] &= (\sum_{j=1}^{m} D_j^\top \phi_{2,n} D_j) \int_0^t e^{a(\phi_{1,n})(t-s)} \mathrm{d}s + x_0^2 e^{a(\phi_{1,n})t} \\ &\leqslant c(1 + |\phi_{2,n}|) e^{c|\phi_{1,n}|^2 T}, \end{aligned}$$

leading to

$$\begin{aligned} l(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n) &= (\sum_{j=1}^{m} D_j D_j^\top) \int_0^T k_1(t; \boldsymbol{\theta}_n) \mathbb{E}[x_n(t)^2] \mathrm{d}t \\ &\leqslant (\sum_{j=1}^{m} D_j D_j^\top) c_2 T c (1 + |\phi_{2,n}|) e^{c|\phi_{1,n}|^2 T}. \end{aligned}$$

Recalling that $\phi_{2,n} = \frac{I_L}{b_n}$, we arrive at

$$\begin{aligned} |h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n)| &\leqslant |\phi_{1,n} - \phi_1^*||l(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n)| \\ &\leqslant c(1 + |\phi_{1,n}|)e^{c|\phi_{1,n}|^2} \end{aligned} \tag{39}$$

for some constant $c > 0$.

### A.3 Almost Sure Convergence of $\phi_{1,n}$

We are now ready to prove Part (a) of Theorem 1 regarding the almost sure convergence of $\phi_{1,n}$. Here and henceforth we will prove more general results by allowing a bias term $\beta_{1,n} = \mathbb{E}[\xi_{1,n} \mid \mathcal{G}_n]$ that may account for errors arising from practical implementations (e.g. the discretization errors; see Remark 2 for details). The following theorem specializes to Part (a) of Theorem 1 when $\beta_{1,n} = \mathbf{0}$.

**Theorem 3.** *Assume the noise term $\xi_{1,n}$ satisfies $\mathbb{E}\left[\xi_{1,n}\middle|\mathcal{G}_n\right] = \beta_{1,n}$ and*

$$\mathbb{E}\left[|\xi_{1,n} - \beta_{1,n}|^2 \middle| \mathcal{G}_n\right] \leqslant c b_n(1 + |\phi_{1,n}|^8 + (\log b_n)^8) \exp\{c|\phi_{1,n}|^6\}, \tag{40}$$

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

$$\begin{aligned} &(i) \quad 0 < a_n \leqslant 1 \text{ for all } n, \quad \sum_n a_n = \infty, \quad \sum_n a_n |\beta_{1,n}| < \infty; \\ &(ii) \quad c_{1,n} \uparrow \infty, \quad \sum_n a_n^2 b_n^q (\log b_n)^{q_2} e^{c c_{1,n}^{q_3}} < \infty \\ &\qquad \text{for any } c > 0,\ 0 \leqslant q \leqslant 1\ ,\ 0 \leqslant q_2 \leqslant 8 \text{ and } 0 \leqslant q_3 \leqslant 6; \\ &(iii) \quad b_n \geqslant 1 \text{ for all } n, \quad b_n \uparrow \infty. \end{aligned} \tag{41}$$

*Then $\phi_{1,n}$ almost surely converges to the unique equilibrium point*

$$\phi_1^* = -(\sum_{j=1}^m D_j D_j^\top)^{-1}(B + \sum_{j=1}^m C_j D_j).$$

*Proof.* The main idea is to derive inductive upper bounds of $|\phi_{1,n} - \phi_1^*|^2$, namely, we will bound $|\phi_{1,n+1} - \phi_1^*|^2$ in terms of $|\phi_{1,n} - \phi_1^*|^2$.

First, for any closed, convex set $K \subset \mathbb{R}$ and $x \in K, y \in \mathbb{R}$, it follows from a property of projection that the function $f(t) = |t\Pi_K(y) + (1-t)x - y|^2$, $t \in \mathbb{R}$, achieves minimum at $t = 1$. The first-order

condition at $t = 1$ then yields

$$2|\Pi_K(y) - y|^2 - 2\langle \Pi_K(y) - y, x - y\rangle = 0.$$

Therefore,

$$\begin{aligned}|\Pi_K(y) - x|^2 =& |\Pi_K(y) - y + y - x|^2 = |y - x|^2 + |\Pi_K(y) - y|^2 \\ &+ 2\langle \Pi_K(y) - y, y - x\rangle \\ =& |y - x|^2 - |\Pi_K(y) - y|^2 \leqslant |y - x|^2.\end{aligned}$$

Taking $n$ sufficiently large such that $\phi_1^* \in K_{1,n+1}$, we have

$$|\phi_{1,n+1} - \phi_1^*|^2 \leqslant |\phi_{1,n} + a_n[h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n) + \xi_{1,n}] - \phi_1^*|^2.$$

Denoting $U_{1,n} = \phi_{1,n} - \phi_1^*$, we deduce

$$\begin{aligned}&\mathbb{E}\left[|U_{1,n+1}|^2 \middle| \mathcal{G}_n\right] \\ \leqslant& \mathbb{E}\left[|U_{1,n} + a_n[h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n) + \xi_{1,n}]|^2 \middle| \mathcal{G}_n\right] \\ =& |U_{1,n}|^2 + 2a_n\langle U_{1,n}, h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n) + \beta_{1,n}\rangle + a_n^2\mathbb{E}\left[|h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n) + \xi_{1,n}|^2 \middle| \mathcal{G}_n\right] \\ =& |U_{1,n}|^2 + 2a_n\langle U_{1,n}, h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n) + \beta_{1,n}\rangle \\ &+ a_n^2\mathbb{E}\left[|h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n) + (\xi_{1,n} - \beta_{1,n}) + \beta_{1,n}|^2 \middle| \mathcal{G}_n\right] \\ \leqslant& |U_{1,n}|^2 + 2a_n\langle U_{1,n}, h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n)\rangle + 2a_n|\beta_{1,n}||U_{1,n}| \\ &+ 3a_n^2\left(|h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n)|^2 + |\beta_{1,n}|^2 + \mathbb{E}\left[|\xi_{1,n} - \beta_{1,n}|^2 \middle| \mathcal{G}_n\right]\right) \\ \leqslant& |U_{1,n}|^2 + 2a_n\langle U_{1,n}, h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n)\rangle + a_n|\beta_{1,n}|(1 + |U_{1,n}|^2) \\ &+ 3a_n^2\left(|h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n)|^2 + |\beta_{1,n}|^2 + \mathbb{E}\left[|\xi_{1,n} - \beta_{1,n}|^2 \middle| \mathcal{G}_n\right]\right).\end{aligned}$$

Recall that $|\phi_{1,n}| \leqslant c_{1,n}$ almost surely. By (39), we have

$$|h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n)|^2 \leqslant c(1 + |\phi_{1,n}|)^2 e^{2c|\phi_{1,n}|^2} \leqslant c(1 + c_{1,n}^2)e^{cc_{1,n}^2}.$$

In addition, the assumption (40) yields

$$\begin{aligned}\mathbb{E}\left[|\xi_{1,n} - \beta_{1,n}|^2 \middle| \mathcal{G}_n\right] \leqslant& cb_n(1 + |\phi_{1,n}|^8 + (\log b_n)^8)\exp\{c|\phi_{1,n}|^6\} \\ \leqslant& cb_n(1 + c_{1,n}^8 + (\log b_n)^8)\exp\{cc_{1,n}^6\}.\end{aligned}$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}\left[|U_{1,n+1}|^2\Big|\mathcal{G}_n\right] \\
\leqslant&|U_{1,n}|^2 + 2a_n\langle U_{1,n}, h_1(\phi_{1,n},\phi_{2,n};\boldsymbol{\theta}_n)\rangle + a_n|\beta_{1,n}|(1+|U_{1,n}|^2) \\
&+ 3a_n^2\left(|h_1(\phi_{1,n},\phi_{2,n};\boldsymbol{\theta}_n)|^2 + |\beta_{1,n}|^2 + \mathbb{E}\left[|\xi_{1,n}-\beta_{1,n}|^2\Big|\mathcal{G}_n\right]\right) \\
\leqslant&(1+a_n|\beta_{1,n}|)|U_{1,n}|^2 + 2a_n\langle U_{1,n}, h_1(\phi_{1,n},\phi_{2,n};\boldsymbol{\theta}_n)\rangle + a_n|\beta_{1,n}| \\
&+ 3a_n^2\left(c(1+c_{1,n}^2)e^{cc_{1,n}^2} + |\beta_{1,n}|^2 + cb_n(1+c_{1,n}^8+(\log b_n)^8)\exp\{cc_{1,n}^6\}\right) \\
=&: (1+\kappa_{1,n})|U_{1,n}|^2 - \zeta_{1,n} + \eta_{1,n},
\end{aligned}
$$

where $\kappa_{1,n} = a_n|\beta_{1,n}|$, $\zeta_{1,n} = -2a_n\langle U_{1,n}, h_1(\phi_{1,n},\phi_{2,n};\boldsymbol{\theta}_n)\rangle$, and

$$
\begin{aligned}
\eta_{1,n} = a_n|\beta_{1,n}| + 3a_n^2\Big(&c(1+c_{1,n}^2)e^{cc_{1,n}^2} + |\beta_{1,n}|^2 \\
&+ cb_n(1+c_{1,n}^8+(\log b_n)^8)\exp\{cc_{1,n}^6\}\Big).
\end{aligned} \tag{42}
$$

By assumptions (i)-(ii) of (41), we know $\sum\kappa_{1,n} < \infty$ and $\sum\eta_{1,n} < \infty$. It then follows from (Robbins and Siegmund, 1971, Theorem 1) that $|U_{1,n}|^2$ converges to a finite limit and $\sum\zeta_{1,n} < \infty$ almost surely.

It remains to show $|U_{1,n}| \to 0$ almost surely. It follows from (35) and (38) that

$$
\begin{aligned}
\zeta_{1,n} &= -2a_n\langle U_{1,n}, h_1(\phi_{1,n},\phi_{2,n};\boldsymbol{\theta}_n)\rangle \\
&= 2a_n\langle(\phi_{1,n}-\phi_1^*), l(\phi_{1,n},\phi_{2,n};\boldsymbol{\theta}_n)(\phi_{1,n}-\phi_1^*)\rangle \\
&= 2a_n(\phi_{1,n}-\phi_1^*)^\top l(\phi_{1,n},\phi_{2,n};\boldsymbol{\theta}_n)(\phi_{1,n}-\phi_1^*) \\
&\geqslant 2\bar{c}a_n|\phi_{1,n}-\phi_1^*|^2.
\end{aligned}
$$

Now, suppose $|U_{1,n}| \nrightarrow 0$ almost surely. Then there exists a set $Z \in \mathcal{F}$ with $\mathbb{P}(Z) > 0$ so that for every $\omega \in Z$, there is $\delta(\omega) > 0$ such that $|U_{1,n}| = |\phi_{1,n}-\phi_1^*| \geqslant \delta(\omega) > 0$ for sufficiently large $n$. Thus,

$$
\sum\zeta_{1,n} \geqslant \sum 2\bar{c}a_n|\phi_{1,n}-\phi_1^*|^2 \geqslant 2\bar{c}\delta(\omega)^2\sum a_n = \infty.
$$

This is a contradiction. Therefore, we have proved that $|U_{1,n}|$ converges to 0 almost surely, or $\phi_{1,n}$ converges to $\phi_1^*$ almost surely. □

**Remark 1.** *An instance satisfying the assumptions in* (41) *is* $a_n = 1 \wedge \frac{\alpha^{\frac{3}{4}}}{(n+\beta)^{\frac{3}{4}}}$, *where constants* $\alpha > 0$, $\beta > 0$. $b_n = 1 \vee \frac{(n+\beta)^{\frac{1}{4}}}{\alpha^{\frac{1}{4}}}, c_{1,n} = 1 \vee (\log\log n)^{\frac{1}{6}}$, *and* $\beta_{1,n} = 0$. *This is because* $\sum \frac{1}{n} = \infty$, *and* $\sum \frac{(\log n)^p (\log\log n)^q}{n^r} < \infty$, *for any* $p, q > 0$ *and* $r > 1$.

## A.4 Mean-Squared Error of $\phi_{1,n} - \phi_1^*$

In this section, we establish the convergence rate of $\phi_{1,n}$ to $\phi_1^*$ stipulated in part (b) of Theorem 1.

The following lemma shows a general recursive relation satisfied by some sequences of learning rates.

**Lemma 3.** *For any* $w > 0$, *there exists positive numbers* $\alpha > \frac{1}{w}$ *and* $\beta \geqslant \max(\frac{1}{w\alpha - 1}, w^2\alpha^3)$ *such that the sequence* $a_n = \frac{\alpha^{\frac{3}{4}}}{(n+\beta)^{\frac{3}{4}}}$ *satisfies* $a_n \leqslant a_{n+1}(1 + wa_{n+1})$ *for all* $n \geqslant 0$.

*Proof.* We have the following equivalences:

$$\begin{aligned} & a_n \leqslant a_{n+1}(1 + wa_{n+1}) \\ \Leftrightarrow & \frac{\alpha^{\frac{3}{4}}}{(n+\beta)^{\frac{3}{4}}} \leqslant \frac{\alpha^{\frac{3}{4}}}{(n+1+\beta)^{\frac{3}{4}}} + w\left(\frac{\alpha}{n+1+\beta}\right)^{\frac{2}{3}} \\ \Leftrightarrow & (n+\beta+1)^{\frac{3}{4}} - (n+\beta)^{\frac{3}{4}} \leqslant w\alpha^{\frac{3}{4}}\left(\frac{n+\beta}{n+\beta+1}\right)^{\frac{3}{4}}. \end{aligned} \tag{43}$$

Consider the last inequality in (43) and notice that the left hand side is a decreasing function of $n$ and the right hand side is an increasing function of $n$. 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

$$(\beta+1)^{\frac{3}{4}} - \beta^{\frac{3}{4}} \leqslant w\alpha^{\frac{3}{4}} \frac{\beta^{\frac{3}{4}}}{(\beta+1)^{\frac{3}{4}}}. \tag{44}$$

To this end, it follows from $\beta \geqslant w^2\alpha^3$ and $w\alpha\beta \geqslant \beta + 1$ that

$$w\beta^4 \geqslant w^3\alpha^3\beta^3 \geqslant (\beta+1)^3.$$

Hence

$$w^{\frac{1}{4}} \frac{\beta^{\frac{3}{4}}}{(\beta+1)^{\frac{3}{4}}} \geqslant \frac{3}{4}\beta^{-\frac{1}{4}} \geqslant (\beta+1)^{\frac{3}{4}} - \beta^{\frac{3}{4}},$$

where the last inequality is due to the mean value theorem. Now the desired inequality (44) follows

from the fact that $\alpha > \frac{1}{w}$. $\square$

The following result covers part (b) of Theorem 1 as a special case.

**Theorem 4.** *Under the assumption of Theorem 3, if the sequence $\{a_n\}$ further satisfies*

$$a_n \leqslant a_{n+1}(1 + wa_{n+1})$$

*for some sufficiently small constant $w > 0$ and $\{\frac{b_n}{a_n}|\beta_{1,n}|^2\}$ is non-decreasing in $n$, then there exists an increasing sequence $\{\hat{\eta}_{1,n}\}$ and a constant $c' > 0$ such that*

$$\mathbb{E}[|\phi_{1,n+1} - \phi_1^*|^2] \leqslant c' a_n \hat{\eta}_{1,n}.$$

*In particular, if we set the parameters $a_n, b_n, c_{1,n}, \beta_{1,n}$ as in Remark 1, then*

$$\mathbb{E}[|\phi_{1,n+1} - \phi_1^*|^2] \leqslant c\frac{(1 \vee \log n)^p (1 \vee \log\log n)^{\frac{4}{3}}}{n^{\frac{1}{2}}}$$

*for any $n$, where $c$ and $p$ are positive constants that only depend on model primitives.*

*Proof.* Denote $n_0 = \inf\{n \geqslant 0 : \phi_1^* \in K_{1,n+1}\}$ and $U_{1,n} = \phi_{1,n} - \phi_1^*$. It follows from (35) and (38) that

$$\langle U_{1,n}, h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n)\rangle = -U_{1,n}^\top l(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n) U_{1,n} \leqslant -\bar{c}|U_{1,n}|^2. \tag{45}$$

When $n \geqslant n_0$, this together with a similar argument to the proof of Theorem 3 yields

$$\begin{aligned}
&\mathbb{E}\left[|U_{1,n+1}|^2 \middle| \mathcal{G}_n\right] \\
\leqslant &|U_{1,n}|^2 + 2a_n\langle U_{1,n}, h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n)\rangle + 2a_n|\beta_{1,n}||U_{1,n}| \\
&+ 3a_n^2\left(|h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n)|^2 + |\beta_{1,n}|^2 + \mathbb{E}\left[|\xi_{1,n} - \beta_{1,n}|^2 \middle| \mathcal{G}_n\right]\right) \\
\leqslant &|U_{1,n}|^2 - 2a_n\bar{c}|U_{1,n}|^2 + 2a_n|\beta_{1,n}||U_{1,n}| \\
&+ 3a_n^2\left(|h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n)|^2 + |\beta_{1,n}|^2 + \mathbb{E}\left[|\xi_{1,n} - \beta_{1,n}|^2 \middle| \mathcal{G}_n\right]\right) \\
\leqslant &|U_{1,n}|^2 - 2a_n\bar{c}|U_{1,n}|^2 + a_n(\frac{1}{\bar{c}}|\beta_{1,n}|^2 + \bar{c}|U_{1,n}|^2) \\
&+ 3a_n^2\left(|h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n)|^2 + |\beta_{1,n}|^2 + \mathbb{E}\left[|\xi_{1,n} - \beta_{1,n}|^2 \middle| \mathcal{G}_n\right]\right) \\
= &(1 - \bar{c}a_n)|U_{1,n}|^2 + 3a_n^2\Bigg(|h_1(\phi_{1,n}, \phi_{2,n}; \boldsymbol{\theta}_n)|^2 + (1 + \frac{1}{3\bar{c}a_n})|\beta_{1,n}|^2 \\
&+ \mathbb{E}\left[|\xi_{1,n} - \beta_{1,n}|^2 \middle| \mathcal{G}_n\right]\Bigg).
\end{aligned} \tag{46}$$

Now, by the proof of Theorem 3,

$$\begin{aligned}
&|h_1(\phi_{1,n},\phi_{2,n};\boldsymbol{\theta}_n)|^2+\mathbb{E}\left[|\xi_{1,n}-\beta_{1,n}|^2\Big|\mathcal{G}_n\right]\\
\leqslant& c\bigg((1+c_{1,n}^2)e^{cc_{1,n}^2}+b_n(1+c_{1,n}^8+(\log b_n)^8)\exp\{cc_{1,n}^6\}\bigg)\\
\leqslant& cb_n(1+c_{1,n}^8+(\log b_n)^8)\exp\{cc_{1,n}^6\}.
\end{aligned}$$

Moreover, the assumptions in (41) imply that $(1+\frac{1}{3\bar{c}a_n})|\beta_{1,n}|^2\leqslant c\frac{b_n}{a_n}|\beta_{1,n}|^2$ for some constant $c>0$. When $n\geqslant n_0$, it follows from (46) that

$$\mathbb{E}\left[|U_{1,n+1}|^2\Big|\mathcal{G}_n\right]\leqslant(1-\bar{c}a_n)|U_{1,n}|^2+3a_n^2\hat{\eta}_{1,n},$$

where

$$\hat{\eta}_{1,n}=cb_n(1+c_{1,n}^8+(\log b_n)^8+\frac{|\beta_{1,n}|^2}{a_n})\exp\{cc_{1,n}^6\},\tag{47}$$

which is monotonically increasing because $c_{1,n}$, $b_n$ are monotonically increasing and $\frac{b_n}{a_n}|\beta_{1,n}|^2$ is non-decreasing by the assumptions. Taking expectation on both sides of the above and denoting $\rho_n=\mathbb{E}[|U_{1,n}|^2]$, we get

$$\rho_{n+1}\leqslant(1-\bar{c}a_n)\rho_n+3a_n^2\hat{\eta}_{1,n}\tag{48}$$

when $n\geqslant n_0$.

Next, we show $\rho_{n+1}\leqslant c'a_n\hat{\eta}_{1,n}$ for all $n\geqslant 0$ by induction, where $c'=max\{\frac{\rho_1}{a_0\hat{\eta}_{1,0}},\frac{\rho_2}{a_1\hat{\eta}_{1,1}},\cdots,\frac{\rho_{n_0+1}}{a_{n_0}\hat{\eta}_{1,n_0}},\frac{3}{\bar{c}}\}+1$. Indeed, it is true when $n\leqslant n_0$. Assume that $\rho_{k+1}\leqslant c'a_k\hat{\eta}_{1,k}$ is true for $n_0\leqslant k\leqslant n-1$. Then (48) yields

$$\begin{aligned}
\rho_{n+1}&\leqslant(1-\bar{c}a_n)\rho_n+3a_n^2\hat{\eta}_{1,n}\\
&\leqslant(1-\bar{c}a_n)c'a_{n-1}\hat{\eta}_{1,n-1}+3a_n^2\hat{\eta}_{1,n}\\
&\leqslant(1-\bar{c}a_n)c'a_n(1+wa_n)\hat{\eta}_{1,n}+3a_n^2\hat{\eta}_{1,n}\\
&=c'a_n\hat{\eta}_{1,n}+c'\hat{\eta}_{1,n}a_n^2\bigg(w-\bar{c}-\bar{c}wa_n+\frac{3}{c'}\bigg).
\end{aligned}$$

Consider the function

$$f(x)=c'\hat{\eta}_{1,n}x^2\bigg(w-\bar{c}-\bar{c}wx+\frac{3}{c'}\bigg),$$

which has two roots at $x_{1,2}=0$ and one root at $x_3=\frac{w-(\bar{c}-\frac{3}{c'})}{cw}$. Because $\bar{c}-\frac{3}{c'}>0$, we can choose

$0 < w < \bar{c} - \frac{3}{c'}$ so that $x_3 < 0$. So $f(x) < 0$ when $x > 0$, leading to

$$c'\hat{\eta}_{1,n}a_n^2\left(w - \bar{c} - \bar{c}wa_n + \frac{3}{c'}\right) < 0, \quad \forall n$$

because $a_n > 0$. We have now proved $\mathbb{E}[|U_{1,n+1}|^2] \leqslant c'a_n\hat{\eta}_{1,n}$.

In particular, under the setting of Remark 1, we can verify that $(\frac{b_n}{a_n})|\beta_{1,n}|^2$ is a non-decreasing sequence of $n$, and $a_n = \Theta(n^{-\frac{3}{4}})$. Then

$$\begin{aligned}
\hat{\eta}_{1,n} =& cb_n(1 + c_{1,n}^8 + (\log b_n)^8 + \frac{|\beta_{1,n}|^2}{a_n})\exp\{cc_{1,n}^6\} \\
\leqslant& cn^{\frac{1}{4}}(1 + (1 \vee \log\log n)^{\frac{4}{3}} + (0 \vee \log n)^8)(e \vee \log n)^c \\
\leqslant& cn^{\frac{1}{4}}(1 \vee \log n)^p(1 \vee \log\log n)^{\frac{4}{3}},
\end{aligned} \tag{49}$$

where $c$ and $p$ are positive constants. The proof is now complete.

□

# B A Proof of Theorem 2

This section is dedicated to proving Theorem 2, which pivots around analyzing the value function $\bar{J}$ in terms of $\phi_1$ and $\phi_2$.

## B.1 Analyzing $\bar{J}(\phi_1, \phi_2)$

Recall that $\bar{J}(\phi_1, \phi_2) = J(0, x_0; \pi)$ with $\gamma = 0$, where $\pi = \mathcal{N}(\cdot|\phi_1 x, \phi_2)$.

**Lemma 4.** *The value function can be expressed as*

$$\bar{J}(\phi_1, \phi_2) = f(a(\phi_1)) + (\sum_{j=1}^{m} D_j^\top \phi_2 D_j)g(a(\phi_1)),$$

*where $a(\phi_1)$ is defined by* (28) *and the functions $f$ and $g$ are defined as follows:*

$$f(a) = \begin{cases} \frac{x_0^2(-H-QT)}{2} & \text{if } a = 0, \\ \frac{1}{2a}(Q - e^{aT}Q - He^{aT}a)x_0^2 & \text{if } a \neq 0, \end{cases} \tag{50}$$

$$g(a)=\begin{cases}\frac{T(-2H-QT)}{4} & \text{if } a=0,\\ \frac{1}{2a^2}(QTa+Q+Ha-e^{aT}Q-He^{aT}a) & \text{if } a\neq 0.\end{cases} \tag{51}$$

*Proof.* The value function of the policy $\mathcal{N}(\cdot|\phi_1 x,\phi_2)$ with $\gamma=0$ is (with a slight abuse of notation)

$$J(t,x;\phi_1,\phi_2)=\mathbb{E}\left[-\frac{1}{2}\int_t^T Qx^{\boldsymbol{\phi}}(s)^2\mathrm{d}s-\frac{1}{2}Hx^{\boldsymbol{\phi}}(T)^2\Big|x^{\boldsymbol{\phi}}(t)=x\right],$$

where $\boldsymbol{\phi}=(\phi_1,\phi_2)$ and $\{x^{\boldsymbol{\phi}}(s):t\leqslant s\leqslant T\}$ is the corresponding exploratory state process. By the Feynman–Kac formula, $J(\cdot,\cdot;\phi_1,\phi_2)$ satisfies

$$\begin{cases} v_t+\frac{1}{2}\int_{\mathcal{R}^l}\sum_{j=1}^m(C_jx+D_j^\top u)^2\mathcal{N}(u|\phi_1x,\phi_2)\mathrm{d}u v_{xx}\\ \qquad+\left(Ax+B^\top\int_{\mathcal{R}^l}u\mathcal{N}(u|\phi_1x,\phi_2)\mathrm{d}u\right)v_x-\frac{1}{2}Qx^2=0,\\ v(T,x)=-\frac{1}{2}Hx^2.\end{cases} \tag{52}$$

The solution to the above PDE is

$$\begin{aligned}J(t,x;\phi_1,\phi_2)=&\frac{1}{2}\left[\frac{Q}{a(\phi_1)}-e^{a(\phi_1)(T-t)}(\frac{Q}{a(\phi_1)}+H)\right]x^2-\frac{1}{2}(\sum_{j=1}^m D_j^\top\phi_2D_j)\\ &\left[\frac{Q}{a(\phi_1)}t+\frac{e^{a(\phi_1)(T-t)}}{a(\phi_1)}(\frac{Q}{a(\phi_1)}+H)-(\frac{QT}{a(\phi_1)}+\frac{Q}{a(\phi_1)^2}+\frac{H}{a(\phi_1)})\right]\end{aligned} \tag{53}$$

if $a(\phi_1)\neq 0$, and

$$\begin{aligned}J(t,x;\phi_1,\phi_2)=&\frac{1}{2}(Qt-QT-H)x^2+\frac{1}{4}(\sum_{j=1}^m D_j^\top\phi_2D_j)(Qt^2-2QTt-2Ht)\\ &-\frac{1}{4}(\sum_{j=1}^m D_j^\top\phi_2D_j)(QT^2+2HT)\end{aligned} \tag{54}$$

if $a(\phi_1)=0$.

The desired result follows immediately noting that $\bar{J}(\phi_1,\phi_2)=J(0,x_0;\phi_1,\phi_2)$. □

**Lemma 5.** *Both the functions $f$ and $g$ defined respectively by* (50) *and* (51) *are continuously differentiable, monotonically non-increasing, and strictly negative everywhere.*

*Proof.* First of all, it is straightforward to check by L'Hôpital's rule that both $f$ and $g$ are continuous

at 0; hence they are continuous functions. Next, when $a \neq 0$,

$$f'(a) = -\frac{Qx_0^2(1 + aTe^{aT} - e^{aT})}{2a^2} - \frac{x_0^2}{2}HTe^{aT} \leqslant 0,$$

where the inequality is due to the facts that $1 + xe^x - e^x \leqslant 0 \ \forall x$ and that $Q, H \geqslant 0$. Moreover, again by L'Hôpital's rule we obtain

$$f'(0) = -\frac{Tx_0^2(2H + QT)}{4} = \lim_{a \to 0} f'(a),$$

implying that $f$ is continuously differentiable at 0, and hence continuously differentiable everywhere and monotonically non-increasing. Similarly we can prove that $g$ is also continuously differentiable everywhere and monotonically non-increasing.

Finally, it is evident that

$$\lim_{a \to -\infty} f(a) = \lim_{a \to -\infty} g(a) = 0, \quad \lim_{a \to \infty} f(a) = \lim_{a \to \infty} g(a) = -\infty.$$

Thus, in view of the proved monotonicity, we have $f(a) < 0$ and $g(a) < 0$ for any $a$. □

## B.2 Regret Analysis

We now proceed to prove Theorem 2.

*Proof.* By Lemma 4, we can write

$$\begin{aligned}
&\bar{J}(\phi_1^*, 0) - \bar{J}(\phi_{1,n}, \phi_{2,n}) \\
=&f(a(\phi_1^*)) - [f(a(\phi_{1,n})) + (\sum_{j=1}^{m} D_j^\top \phi_{2,n} D_j) g(a(\phi_{1,n}))] \\
=&-(\sum_{j=1}^{m} D_j^\top \phi_{2,n} D_j) g(a(\phi_1^*)) + [f(a(\phi_1^*)) - f(a(\phi_{1,n}))] \\
&+ (\sum_{j=1}^{m} D_j^\top \phi_{2,n} D_j)[g(a(\phi_1^*)) - g(a(\phi_{1,n}))].
\end{aligned} \tag{55}$$

Because $\phi_{2,n} = \frac{I_l}{b_n}$ and $g < 0$ (by Lemma 5), we have

$$\mathbb{E}[-(\sum_{j=1}^{m} D_j^\top \phi_{2,n} D_j) g(a(\phi_1^*))] = -g(a(\phi_1^*)) \frac{D}{b_n} \leqslant \frac{c}{n^{\frac{1}{4}}}, \tag{56}$$

where $D = (\sum_{j=1}^{m} D_j^\top D_j)$ and $c > 0$ is independent of $n$.

Next, by the definition of function $a(\phi_1)$ in (28), we have $|a(\phi_1) - a(\phi_1^*)| \leqslant \bar{c}_1 |\phi_1 - \phi_1^*|^2$, where $\bar{c}_1 > 0$ is a constant that depends on the model primitives. Furthermore, it follows from (28) and (18) that $|a(\phi_{1,n})| \leqslant \bar{c}_2(1 + c_{1,n}^2)$ for some constant $\bar{c}_2 > 0$. In addition, by the monotonicity of the functions $f$ and $g$ and assumptions (41), we have

$$|f(a(\phi_{1,n}))| \leqslant |f(\bar{c}_2(1 + c_{1,n}^2))| \leqslant c(1 + e^{\bar{c}_2(1+c_{1,n}^2)T})(1 + \frac{1}{c_{1,n}^2}) \leqslant \bar{c}_3 \log n,$$

$$|g(a(\phi_{1,n}))| \leqslant |g(\bar{c}_2(1 + c_{1,n}^2))| \leqslant c(1 + e^{\bar{c}_2(1+c_{1,n}^2)T})(1 + \frac{1}{c_{1,n}^4}) \leqslant \bar{c}_4 \log n,$$

where $\bar{c}_3 > 0$ and $\bar{c}_4 > 0$ are constants independent of $n$.

Furthermore, it follows from Lemma 5 that for a given fixed $\delta > 0$, the inequalities $|f'(a(\phi_1))| \leqslant \bar{c}_5(\delta)$ and $|g'(a(\phi_1))| \leqslant \bar{c}_6(\delta)$ hold for any $\phi_1$ satisfying $|\phi_1 - \phi_1^*| \leqslant \delta$, where $\bar{c}_5(\delta) > 0$ and $\bar{c}_6(\delta) > 0$ are constants that depend on the value of $\delta$. On the other hand, Theorem 4 yields

$$\mathbb{E}[|\phi_{1,n+1} - \phi_1^*|^2] \leqslant \bar{c}_7 \frac{(1 \vee \log n)^p (1 \vee \log\log n)^{\frac{4}{3}}}{n^{\frac{1}{2}}},$$

where $\bar{c}_7 > 0$ is a constant.

Now, we consider a positive sequence

$$\delta_{1,n} = \left( \frac{|f(a(\phi_1^*))| + \bar{c}_3 \log n}{\bar{c}_5(\delta)\bar{c}_1} \frac{\bar{c}_7 (1 \vee \log n)^p (1 \vee \log\log n)^{\frac{4}{3}}}{n^{\frac{1}{2}}} \right)^{\frac{1}{4}}, \quad n = 1, 2, \cdots.$$

It is straightforward to see that $\delta_{1,n} \to 0$ as $n \to \infty$.

Thus there exists the finite $n_1 = \inf\{n' \in \mathbb{N} : \delta_{1,n} < \delta \text{ for all } n \geqslant n'\}$. Denote $\delta_n = \delta$ for $n < n_1$ and $\delta_n = \delta_{1,n}$ for $n \geqslant n_1$. When $n \geqslant n_1$, we deduce

$$
\begin{aligned}
&\mathbb{E}[f(a(\phi_1^*)) - f(a(\phi_{1,n}))] \\
=&\mathbb{E}[f(a(\phi_1^*)) - f(a(\phi_{1,n}))]\mathbf{1}_{\{|\phi_{1,n}-\phi_1^*|\leqslant\delta_n\}} + \mathbb{E}[f(a(\phi_1^*)) - f(a(\phi_{1,n}))]\mathbf{1}_{\{|\phi_{1,n}-\phi_1^*|>\delta_n\}} \\
=&\int_{|\phi_{1,n}-\phi_1^*|\leqslant\delta_n} f(a(\phi_1^*)) - f(a(\phi_{1,n}))\mathrm{d}\mathbb{P} + \int_{|\phi_{1,n}-\phi_1^*|>\delta_n} f(a(\phi_1^*)) - f(a(\phi_{1,n}))\mathrm{d}\mathbb{P} \\
=&\int_{|\phi_{1,n}-\phi_1^*|\leqslant\delta_n} f'(a(\tilde{\phi}_{1,n}))(a(\phi_1^*) - a(\phi_{1,n}))\mathrm{d}\mathbb{P} \\
&+ \int_{|\phi_{1,n}-\phi_1^*|>\delta_n} f(a(\phi_1^*)) - f(a(\phi_{1,n}))\mathrm{d}\mathbb{P} \\
\leqslant&\bar{c}_5(\delta)\bar{c}_1\delta_n^2 + (|f(a(\phi_1^*))| + |f(a(\phi_{1,n}))|)\mathbb{P}(|\phi_{1,n} - \phi_1^*| > \delta_n) \\
\leqslant&\bar{c}_5(\delta)\bar{c}_1\delta_n^2 + \frac{|f(a(\phi_1^*))| + \bar{c}_3\log n}{\delta_n^2}\mathbb{E}[|\phi_{1,n} - \phi_1^*|^2] \\
\leqslant&\bar{c}_5(\delta)\bar{c}_1\delta_n^2 + \frac{|f(a(\phi_1^*))| + \bar{c}_3\log n}{\delta_n^2}\frac{\bar{c}_7(1\vee\log n)^p(1\vee\log\log n)^{\frac{4}{3}}}{n^{\frac{1}{2}}} \\
=&2\sqrt{\bar{c}_7\bar{c}_5(\delta)\bar{c}_1(|f(a(\phi_1^*))| + \bar{c}_3\log n)}\frac{(1\vee\log n)^{\frac{p}{2}}(1\vee\log\log n)^{\frac{2}{3}}}{n^{\frac{1}{4}}},
\end{aligned}
\tag{57}
$$

where the third equality follows from the mean–value theorem and the fact that $\tilde{\phi}_{1,n} \in \{\phi_1 \in \mathbb{R}^l : |\phi_1 - \phi_1^*| < |\phi_{1,n} - \phi_1^*|\}$ satisfies $f(a(\phi_1^*)) - f(a(\phi_{1,n})) = f'(a(\tilde{\phi}_{1,n}))(a(\phi_1^*) - a(\phi_{1,n}))$.

When $n < n_1$, by the same argument as in (57) with $\delta_n$ replaced by $\delta$, we have,

$$
\begin{aligned}
&\mathbb{E}[f(a(\phi_1^*)) - f(a(\phi_{1,n}))] \\
\leqslant&\bar{c}_5(\delta)\bar{c}_1\delta^2 + \frac{|f(a(\phi_1^*))| + \bar{c}_3\log n}{\delta^2}\frac{\bar{c}_7(1\vee\log n)^p(1\vee\log\log n)^{\frac{4}{3}}}{n^{\frac{1}{2}}}.
\end{aligned}
\tag{58}
$$

Similarly, we consider another positive sequence

$$
\delta_{2,n} = \left(\frac{|g(a(\phi_1^*))| + \bar{c}_4\log n}{\bar{c}_6(\delta)\bar{c}_1}\frac{\bar{c}_7(1\vee\log n)^p(1\vee\log\log n)^{\frac{4}{3}}}{n^{\frac{1}{2}}}\right)^{\frac{1}{4}}, \quad n = 1, 2, \cdots.
$$

Because $\delta_{2,n} \to 0$ as $n \to \infty$, there exists the finite

$$
n_2 = \inf\{n' \in \mathbb{N} : \delta_{2,n} < \delta \text{ for all } n \geqslant n'\}.
$$

Set $\delta_n' = \delta$ for $n < n_2$ and $\delta_n' = \delta_{2,n}$ for $n \geqslant n_2$. When $n \geqslant n_2$, we have

$$
\begin{aligned}
&\mathbb{E}[(\sum_{j=1}^{m} D_j^\top \phi_{2,n} D_j)(g(a(\phi_1^*)) - g(a(\phi_{1,n})))] \\
=&\mathbb{E}[(\sum_{j=1}^{m} D_j^\top \phi_{2,n} D_j)(g(a(\phi_1^*)) - g(a(\phi_{1,n})))]\mathbf{1}_{\{|\phi_{1,n}-\phi_1^*|\leqslant\delta_n'\}} \\
&+ \mathbb{E}[(\sum_{j=1}^{m} D_j^\top \phi_{2,n} D_j)(g(a(\phi_1^*)) - g(a(\phi_{1,n})))]\mathbf{1}_{\{|\phi_{1,n}-\phi_1^*|>\delta_n'\}} \\
=&\frac{D}{b_n}\int_{|\phi_{1,n}-\phi_1^*|\leqslant\delta_n'} (g'(a(\check{\phi}_{1,n}))(a(\phi_1^*) - a(\phi_{1,n})))\mathrm{d}\mathbb{P} \\
&+ \frac{D}{b_n}\int_{|\phi_{1,n}-\phi_1^*|>\delta_n'} (g(a(\phi_1^*)) - g(a(\phi_{1,n})))\mathrm{d}\mathbb{P} \\
\leqslant&\frac{D}{b_n}\left[\bar{c}_6(\delta)\bar{c}_1\delta_n'^2 + (|g(a(\phi_1^*))| + |g(a(\phi_{1,n}))|)\mathbb{P}(|\phi_{1,n}-\phi_1^*|>\delta_n')\right] \\
\leqslant&\frac{D}{b_n}\left[\bar{c}_6(\delta)\bar{c}_1\delta_n'^2 + \frac{|g(a(\phi_1^*))| + \bar{c}_4\log n}{\delta_n'^2}\mathbb{E}[|\phi_{1,n}-\phi_1^*|^2]\right] \\
\leqslant&\frac{D}{b_n}\left[\bar{c}_6(\delta)\bar{c}_1\delta_n'^2 + \frac{|g(a(\phi_1^*))| + \bar{c}_4\log n}{\delta_n'^2}\frac{\bar{c}_7(1\vee\log n)^p(1\vee\log\log n)^{\frac{4}{3}}}{n^{\frac{1}{2}}}\right] \\
=&\frac{2D}{b_n}\sqrt{\bar{c}_7\bar{c}_6(\delta)\bar{c}_1(|g(a(\phi_1^*))| + \bar{c}_4\log n)}\frac{(1\vee\log n)^{\frac{p}{2}}(1\vee\log\log n)^{\frac{2}{3}}}{n^{\frac{1}{4}}}.
\end{aligned}
\tag{59}
$$

For $n < n_2$, by the same argument as in (59) with $\delta_n'$ replaced by $\delta$, we have

$$
\begin{aligned}
&\mathbb{E}[\phi_{2,n}(g(a(\phi_1^*)) - g(a(\phi_{1,n})))] \\
\leqslant&\frac{D}{b_n}\left[\bar{c}_6(\delta)\bar{c}_1\delta^2 + \frac{|g(a(\phi_1^*))| + \bar{c}_4\log n}{\delta^2}\frac{\bar{c}_7(1\vee\log n)^p(1\vee\log\log n)^{\frac{4}{3}}}{n^{\frac{1}{2}}}\right].
\end{aligned}
\tag{60}
$$

Finally, combining (55) – (60) yields

$$
\begin{aligned}
&\sum_{n=1}^{N} \mathbb{E}[\bar{J}(\phi_1^*, 0) - \bar{J}(\phi_{1,n}, \phi_{2,n})] \\
=&\sum_{n=1}^{N} \mathbb{E}[-\phi_{2,n} g(a(\phi_1^*))] + \sum_{n=1}^{n_1-1} \mathbb{E}[f(a(\phi_1^*)) - f(a(\phi_{1,n}))] \\
&+ \sum_{n=n_1}^{N} \mathbb{E}[f(a(\phi_1^*)) - f(a(\phi_{1,n}))] \\
&+ \sum_{n=1}^{n_2-1} \mathbb{E}[\phi_{2,n}(g(a(\phi_1^*)) - g(a(\phi_{1,n}))] + \sum_{n=n_2}^{N} \mathbb{E}[\phi_{2,n}(g(a(\phi_1^*)) - g(a(\phi_{1,n}))] \\
\leqslant&\sum_{n=1}^{N} \frac{c}{n^{\frac{1}{4}}} + \sum_{n=1}^{n_1-1} \bar{c}_5(\delta)\bar{c}_1\delta^2 + \frac{|f(a(\phi_1^*))| + \bar{c}_3 \log n}{\delta^2} \frac{\bar{c}_7(1 \vee \log n)^p (1 \vee \log\log n)^{\frac{4}{3}}}{n^{\frac{1}{2}}} \\
&+ 2\sum_{n=n_1}^{N} \sqrt{\bar{c}_7\bar{c}_5(\delta)\bar{c}_1(|f(a(\phi_1^*))| + \bar{c}_3 \log n)} \frac{(1 \vee \log n)^{\frac{p}{2}} (1 \vee \log\log n)^{\frac{2}{3}}}{n^{\frac{1}{4}}} \\
&+ \sum_{n=1}^{n_2-1} \frac{D}{b_n} \left[ \bar{c}_6(\delta)\bar{c}_1\delta^2 + \frac{|g(a(\phi_1^*))| + \bar{c}_4 \log n}{\delta^2} \frac{\bar{c}_7(1 \vee \log n)^p (1 \vee \log\log n)^{\frac{4}{3}}}{n^{\frac{1}{2}}} \right] \\
&+ 2\sum_{n=n_2}^{N} \frac{D}{b_n} \sqrt{\bar{c}_7\bar{c}_6(\delta)\bar{c}_1(|g(a(\phi_1^*))| + \bar{c}_4 \log n)} \frac{(1 \vee \log n)^{\frac{p}{2}} (1 \vee \log\log n)^{\frac{2}{3}}}{n^{\frac{1}{4}}} \\
\leqslant& c + cN^{\frac{3}{4}} (\log N)^{\frac{p+1}{2}} (\log\log N)^{\frac{2}{3}}.
\end{aligned}
$$

The proof is complete. □

**Remark 2.** *The discretization errors in our model, impacted by the step size $\Delta t_n$ and parameter values $\boldsymbol{\theta}_n$ and $\boldsymbol{\phi}_n$, can be captured by the term $\beta_{1,n}$. Theorems 3 and 4 show that we only need to set $\beta_{1,n}$ to be of the order $n^{-\frac{3}{8}}$ by a suitable decreasing schedule of $\Delta t_n$. We omit the details here, but refer to Kloeden and Platen (1992); Szpruch et al. (2024) for discussions and analyses on time-discretization techniques.*

# C Specifics of Numerical Experiments

This section presents the implementation details of the numerical experiments outlined in Section 5. For clarity and simplicity, we set $l = m = 1$ both our model-free continuous-time RL algorithm and the adapted model-based counterpart. To facilitate reproducibility, we fix the random seeds for all 120 independent experiments, ranging sequentially from 1 to 120.

The section is organized into three subsections: the first one presents and explains the modified

algorithm that adapts the model-based methods (Basei et al., 2022; Szpruch et al., 2024) to our setting involving state and control dependent volatility. The second one details the experimental conditions, parameter settings, and the overall framework used to validate our claims and assess the performance of our algorithmic enhancements. The third one describes the computational resources used for these experiments.

### C.1 A Modified Model-Based Algorithm

The key component in the algorithms developed by (Basei et al., 2022; Szpruch et al., 2024) is to estimate the parameters $A$ and $B$ in the drift term whereas the diffusion term is assumed to be constant. Clearly, these algorithms cannot be used directly to our setting where the diffusion term is state- and control-dependent. Here we extend them to also including estimates of the parameters $C$ and $D$.

Specifically, in the $n$-th iteration, the policy is defined as

$$u_n(t,x) = \bar{\phi}_{1,n} x, \tag{61}$$

where $\bar{\phi}_{1,n}$ is distributed according to $\mathcal{N}(\cdot| - \frac{B_n + C_n D_n}{D_n^2}, v_n)$, with $v_n$ being a deterministic sequence defined by $v_n = \frac{1}{n+1}$, and $B_n, C_n, D_n$ being the current estimation of the parameter $B, C, D$.

Applying this feedback policy to the classical dynamic (1) yields

$$\begin{aligned} \mathrm{d}x_n(t) =& (A + B\bar{\phi}_{1,n})x_n(t)\mathrm{d}t + (C + D\bar{\phi}_{1,n})x_n(t)\mathrm{d}W_n^{(j)}(t) \\ :=& P_n x_n(t)\mathrm{d}t + R_n x_n(t)\mathrm{d}W_n^{(j)}(t), \end{aligned} \tag{62}$$

where $W_n$ is the Browian motion for the $n$-th iteration.

Given an observed state trajectory $\{x_n(t) : 0 \leqslant t \leqslant T\}$ following (62), we discretize it uniformly into $m$ intervals resulting in the "snapshots" of the state, $\{x_n(t_0), x_n(t_1), \ldots, x_n(t_m)\}$, and then employ a statistical approach to estimate $P_n$ and $R_n$:

$$\begin{aligned} \hat{R}_n^2 =& \frac{\sum_{k=1}^m (\log x_n(t_k) - \log x_n(t_{k-1}))^2}{T}, \\ \hat{P}_n =& \frac{\log x_n(t_m) - \log x_n(t_0)}{T} + \frac{1}{2}\hat{R}_n^2. \end{aligned} \tag{63}$$

Parameters $A_n$ and $B_n$ are subsequently estimated via linear regression, using $\bar{\phi}_{1,n}$ as the independent variable and $\hat{P}_n$ as the dependent variable. Similarly, parameters $C_n$ and $D_n$ are determined

using quadratic regression, with $\bar{\phi}_{1,n}^2$ and $\bar{\phi}_{1,n}$ as the independent variables and $\hat{R}_n^2$ as the dependent variable. We enhance parameter estimation accuracy by incorporating an experience replay mechanism (Lillicrap et al., 2015; Mnih et al., 2015), which utilizes all historical data for ongoing updates.

The pseudocode for implementing this modified model-based algorithm is presented below:

**Algorithm 2** Modified Model-Based Algorithm

**Input**

$A_0, B_0$ Initial drift parameters.

$C_0, D_0$ Initial diffusion parameters.

---

Initialize experience replay buffer for $\hat{P}_n$, $\hat{R}_n^2$, $\bar{\phi}_{1,n}$.

Collect two trajectories with distinct values of $\bar{\phi}_{1,-1}$ and $\bar{\phi}_{1,0}$.

**for** $n = 1$ `to` $N$ **do**

Draw $\bar{\phi}_{1,n}$ from $\mathcal{N}(-\frac{B_n + C_n D_n}{D_n^2}, v_n)$.

Initialize $k = 0$, $t = t_k = 0$, $x_n(t_k) = x_0$.

**while** $t < T$ **do**

Apply action $u_n(t_k) = \bar{\phi}_{1,n} x_n(t_k)$ using policy (61).

Simulate new state $x_n(t_{k+1})$ via dynamics (62).

Advance time to $t_{k+1} \leftarrow t_k + \Delta t$ and update $t$.

**end while**

Record trajectory $\{(t_k, x_n(t_k))\}_{k \geqslant 0}$.

Estimate $\hat{P}_n$, $\hat{R}_n^2$ using (63).

Update $A_n, B_n$ using linear regression with $\hat{P}_n$.

Update $C_n, D_n$ using quadratic regression with $\hat{R}_n^2$.

**end for**

---

**Output**

$A_N, B_N$ Final estimated drift parameters.

$C_N, D_N$ Final estimated diffusion parameters.

## C.2 Experiment Setup

The specific setup for the experiment applying the model-free Algorithm 1 is as follows:

- The initial value for $\phi_1$ is $\phi_{1,0} = -0.5$.
- The leaning rate of $\phi_1$ is $a_n = \frac{0.05}{(n+1)^{\frac{3}{4}}}$.
- The projection for $\phi_{1,n}$ is set to be a constant set of $[-2.2, -0.5]$ for computational efficiency.[3]
- The exploration schedule is $\phi_{2,n} = \frac{1}{b_n}$ where $b_n = 0.2(n+1)^{\frac{1}{4}}$.
- The functions $\hat{k}_1(t;\boldsymbol{\theta}) = 1$ and $\hat{k}_3(t;\boldsymbol{\theta}) = 1$ for simplicity, which satisfy the assumptions in Subsection 3.1.[4]
- The parameters $\boldsymbol{\theta}$ need not to be learned, as the value function is not updated.
- The temperature parameter $\gamma = 1$.
- The initial state $x_0 = 1$.
- The time horizon $T = 1$.
- The time step $\Delta t = 0.01$.
- The total number of iteration for each experiment $N = 400,000$.

The specifics of implementing the adapted model-based Algorithm 2 are:

- The initial state $x_0 = 1$.
- The time horizon $T = 1$.
- The time step $\Delta t = 0.01$.
- The total number of iterations for each experiment is $N = 400,000$.

[3]The projection was originally set to prove the theoretical convergence rate and regret bound. For implementation the theoretical projection grows too slow; instead it could be tuned.

[4]Recall that our results do not dependent on the form of the value function.

### C.3 Compute Resources

All experiments were performed on a MacBook Pro (16-inch, 2019) equipped with a 2.4 GHz 8-Core Intel Core i9 processor, 32 GB of 2667 MHz DDR4 memory, and dual graphics processors, comprising an AMD Radeon Pro 5500M with 8 GB and an Intel UHD Graphics 630 with 1536 MB. Not having a high-powered server, this consumer-grade laptop was sufficient to handle the computational task of conducting 120 independent experiments sequentially, each running for 400,000 iterations. The model-free actor–critic algorithm required approximately 26 hours for a complete sequential run, whereas the model-based plugin algorithm took about 83 hours. This significant difference in running times also demonstrates the efficiency of our model-free approach compared with the model-based one.