Journal of Machine Learning Research 23 (2022) 1-34 Submitted 6/20; Revised 4/22; Published 6/22

Logarithmic Regret for Episodic Continuous-Time Linear-Quadratic Reinforcement Learning over a Finite-Time Horizon

Matteo Basei MATTEO.BASEI@EDF.FR
EDF R&D Department, Paris, France.

Xin Guo XINGUO@BERKELEY.EDU
Department of Industrial Engineering and Operations Research
University of California, Berkeley, USA.

Anran Hu ANRAN_HU@BERKELEY.EDU
Department of Industrial Engineering and Operations Research
University of California, Berkeley, USA.

Yufei Zhang YUFEI.ZHANG@MATHS.OX.AC.UK
Mathematical Institute
University of Oxford, UK.

Editor: Csaba Szepesvari

Abstract

We study finite-time horizon continuous-time linear-quadratic reinforcement learning problems in an episodic setting, where both the state and control coefficients are unknown to the controller. We first propose a least-squares algorithm based on continuous-time observations and controls, and establish a logarithmic regret bound of magnitude $\mathcal{O}((\ln M)(\ln \ln M))$, with M being the number of learning episodes. The analysis consists of two components: perturbation analysis, which exploits the regularity and robustness of the associated Riccati differential equation; and parameter estimation error, which relies on sub-exponential properties of continuous-time least-squares estimators. We further propose a practically implementable least-squares algorithm based on discrete-time observations and piecewise constant controls, which achieves similar logarithmic regret with an additional term depending explicitly on the time stepsizes used in the algorithm.

Keywords: continuous-time, stochastic control, linear-quadratic, episodic reinforcement learning, regret analysis

1. Introduction

Reinforcement learning (RL) for linear quadratic (LQ) control problems has been one of the most active areas for both the control and the reinforcement learning communities. Over the last few decades, significant progresses have been made in the discrete-time setting.

1.1 Discrete-Time RL

In the area of adaptive control with unknown dynamics parameters, the goal is to find optimal stationary policy that stabilizes the unknown dynamics and minimizes the long term average cost (Ioannou and Fidan (2006); Landau et al. (2011)). For an infinite-time horizon LQ system, it has been shown that persistent excitation conditions Green and Moore (1986) are critical to the parameter identification. Meanwhile, algorithms with asymptotic convergence in both the parameter estimation and the optimal control have been developed in Goodwin et al. (1981), Kumar (1983) and Campi and Kumar (1998): the first one assumes that costs only depend on state variables and the other two consider both state and control costs and use a cost-biased least-squared estimation method. See Faradonbeh et al. (2018a, 2019) and references therein for recent developments of (randomised) adaptive control algorithms for LQ systems.

Following the seminal works of Auer and Ortner (2007); Auer et al. (2009) and Osband et al. (2013), *non-asymptotic* regret bound analysis for RL algorithms has been one of the main topics, and has been developed for tabular Markov decision problems.

The non-asymptotic analysis of adaptive LQ problem by Abbasi-Yadkori and Szepesvári (2011) utilizes the Optimism in the Face of Uncertainty principle to construct a sequence of improving confidence regions for the unknown model parameters, and solves a non-convex constrained optimization problem for each confidence region; their algorithm achieves an $\mathcal{O}(\sqrt{T})$ regret bound, with T being the number of time steps. To reduce the computational complexity and to avoid the non-convexity issue, Abeille and Lazaric (2018) and Ouyang et al. (2017) propose Thompson-sampling-based algorithms and derive $\mathcal{O}(\sqrt{T})$ regret bounds in the Bayesian setting; Dean et al. (2018) proposes a robust adaptive control algorithm to solve a convex sub-problem in each step and achieves an $\mathcal{O}(T^{2/3})$ regret bound. The gap between these regret bounds is removed by Mania et al. (2019) and Cohen et al. (2019) via two different approaches for the same $\mathcal{O}(\sqrt{T})$ frequentist regret bound. Later, Simchowitz and Foster (2020) establishes a lower bound on the regret of order $\mathcal{O}(\sqrt{d_u^2 d_x T})$, where d_u and d_x are the dimensions of the actions and the states, and shows that a simple variant of certainty equivalent control matches the lower bound in both T and the dimensions. Similar regret bounds have also been established under different settings and assumptions, such as Chen and Hazan (2021) in the adversarial setting and Lale et al. (2020a) without a stabilizing controller at the early stages of agent-environment interaction.

All the analyses are in discrete-time with an infinite time horizon. In all these problems, adaptive control algorithms are shown to achieve logarithmic regret bounds when additional information regarding the parameters of the system (often referred to as identifiability conditions) is available. Indeed, Faradonbeh et al. (2018b, 2020) prove that certainty equivalent adaptive regulator achieves logarithmic regret bounds if the system parameter satisfies certain sparsity or low-rankness conditions. Cassel et al. (2020) establishes logarithmic regret bounds when either the state transition matrix is unknown, or when the state-action transition matrix is unknown and the optimal policy is non-degenerate. In partially observable linear dynamical systems, which takes linear-quadratic Gaussian problem as a special case, Lale et al. (2020b) proposes an algorithm with a logarithmic regret bound, under the assumption that one has access to a set in which all controllers persistently excite the system to approximate the optimal control. Logarithmic regret bounds in the adversarial setting with known dynamics parameters have been established in Agarwal et al. (2019); Foster and Simchowitz (2020).

1.2 Continuous-Time RL

Most real-world control systems, such as those in aerospace, automotive industry and robotics, are naturally continuous-time dynamical systems. So are their related physical tasks, such as inverted pendulum problems, cart-pole balancing problems, and legged robot problems. Continuous-time finite-time horizon LQ control problems can be found in portfolio optimization Wang and Zhou (2020), algorithmic trading Cartea et al. (2018), production management of exhaustible resources Graber (2016), and biological movement systems Bailo et al. (2018).

Analysis for continuous-time LQ-RL and general RL problems, however, is fairly limited. The primary approach is to develop learning algorithms after discretizing both the time and the space spaces, and establish the convergence as discretization parameters tend to zero. For instance, Munos (2006) proposes a policy gradient algorithm and shows the convergence of the policy gradient estimate to the true gradient. Munos and Bourgine (1998); Munos (2000) design learning algorithms by discretizing Bellman equations of the underlying control problems and prove the asymptotic convergence of their algorithms. For the LQ system, attentions have been mostly on algorithms designs, including the integral reinforcement learning algorithm in Modares and Lewis (2014), and the policy iteration algorithm in Rizvi and Lin (2018). Yet, very little is known regarding the convergence rate or the regret bound of all these algorithms. Indeed, despite the natural analytical connection between LQ control and RL, the best known theoretical work for continuous-time LQ-RL is still due to Duncan et al. (1999), where an asymptotically sublinear regret for an ergodic model has been derived via a weighted least-squares-based estimation approach. Nevertheless, the exact order of the regret bound has not been studied.

Issues and challenges from non-asymptotic analysis. It is insufficient and improper to rely solely on the analysis and algorithms for the discrete-time RL to solve the continuous-time problems. There is a mismatch between the algorithms timescale for the former and the underlying systems timescale for the latter. When designing algorithms that make observations and take actions at discrete time points, it is important to take the model mismatch into consideration. For instance, the empirical studies in Tallec et al. (2019) suggest that vanilla Q-learning methods exhibit degraded performance as the time stepsize decreases, while a proper scaling of learning rates with stepsize leads to more robust performance.

The questions are therefore: A) How to quantify the precise impacts of the observation stepsize and action stepsize on algorithm performance? B) How to derive non-asymptotic regret analysis for learning algorithms in continuous-time LQ-RL (or general RL) system, analogous to the discrete-time LQ-RL counterpart?

There are technical reasons behind the limited theoretical progress in the continuous-time domain for RL, including LQ-RL. In addition to the known difficulty for analyzing stochastic control problems, the learning component compounds the problem complexity and poses new challenges.

For instance, the counterpart in the continuous-time problem to the algebraic equations in Mania et al. (2019) for the discrete-time version is the regularity and stability of the continuous-time Riccati equation and the regularity of feedback controls. While Riccati equation and its robustness and existence and uniqueness of optimal controls have been well studied in the control literature, regularity of feedback controls with respect to underlying models is completely new for control theory and crucial for algorithm design and its robustness analysis. Moreover, deriving the *exact* order of the regret bound requires developing new and different techniques than those used for the *asymptotic* regret analysis in Duncan et al. (1999).

Our work and contributions. This paper studies finite-time horizon continuous-time LQ-RL problems in an episodic setting.

- It first proposes a greedy least-squares algorithm based on continuous-time observations and controls. At each iteration, the algorithm estimates the unknown parameters by a regularized least-squares estimator based on observed trajectories, then designs linear feedback controls via the Riccati differential equation for the estimated model. It identifies conditions under which the unknown state transition matrix and state-action transition matrix are uniquely identifiable under the optimal policies. (Remark 2.1 and Proposition 2.1). By exploiting the identifiability of coefficients, this continuous-time least-squares algorithm is shown to have a logarithmic regret of the magnitude $\mathcal{O}((\ln M)(\ln\ln M))$, with M being the number of learning episodes (Theorem 2.2). To the best of our knowledge, this is the first non-asymptotic logarithmic regret bound for continuous-time LQ-RL problems with unknown state and control coefficients.

- It then proposes a practically implementable least-squares algorithm based on discrete-time observations and controls. At each iteration, the algorithm estimates the unknown parameters by observing continuous-time trajectories at discrete time points, then designs a piecewise constant linear feedback control via Riccati difference equations for an associated discrete-time LQ-RL problem. It shows that the regret of the discrete-time least-squares algorithm is of the magnitude $\mathcal{O}\big((\ln M)(\ln\ln M) + \sum_{\ell=0}^{\ln M} 2^\ell \tau_\ell^2\big)$, where τ_ℓ is the time stepsize used in the $(\ell+1)$-th update of model parameters (Theorem 2.3). Our analysis shows that scaling the regularization parameter of the discrete-time least-squares estimator with respect to time stepsize is critical for a robust performance of the algorithm in different timescales (Remark 2.3). To the best of our knowledge, this is the first discrete-time algorithm with rigorous regret bound for continuous-time LQ-RL problems.

Different from the least-squares algorithms for the ergodic LQ problems (see e.g., Duncan et al. (1999); Mania et al. (2019)), our continuous-time least-squares algorithm constructs feedback controls via Riccati differential equations instead of the algebraic equations in Mania et al. (2019). Here, the regularity and stability of the continuous-time Riccati equation is analyzed in order to establish the robustness of feedback controls.

Moreover, our analysis for the estimation error exploits extensively the sub-exponential tail behavior of the least-squares estimators. This probabilistic approach differs from the asymptotic sublinear regret analysis in Duncan et al. (1999); it establishes the exact order of the logarithmic regret bound by the concentration inequality for the error bound.

In addition, our analysis also exploits an important self-exploration property of finite-time horizon continuous-time LQ-RL problems, for which the time-dependent optimal feedback matrices ensure that the optimal state and control processes span the entire parameter space. This property allows us to design exploration-free learning algorithms with logarithmic regret bounds. Furthermore, we provide explicit conditions on models that guarantees the successful identification of the unknown parameters with optimal feedback policies. This is in contrast to the identification conditions for logarithmic regret bounds in discrete-time infinite-time-horizon LQ problems. Our conditions apply to arbitrary finite time-horizon problems, without imposing sparsity or low-rankness conditions on system parameters as in Faradonbeh et al. (2018b, 2020) or requiring these parameters to be partially known to the controller as in Cassel et al. (2020); Foster and Simchowitz (2020).

Finally, our analysis provides the precise parameter estimation error in terms of the sample size and time stepsize, and quantifies the performance gap between applying a piecewise-constant policy from an incorrect model and applying the optimal policy. The misspecification error scales linearly with respect to the stepsize, and the performance gap depends quadratically with respect to the time stepsize and the magnitude of parameter perturbations. Our analysis is based on the first-order convergence of Riccati difference equations and a uniform sub-exponential tail bound of discrete-time least-squares estimators.

Notation. For each $n \in \mathbb{N}$, we denote by $I = I_n$ the $n \times n$ identity matrix, and by $\mathbb{S}^n_0$ (resp. $\mathbb{S}^n_+$) the space of symmetric positive semidefinite (resp. definite) matrices. We denote by $|\cdot|$ the Euclidean norm of a given Euclidean space, by $\|\cdot\|_2$ the matrix norm induced by Euclidean norms, and by $A^\top$ and $\mathrm{tr}(A)$ the transpose and trace of a matrix A, respectively. For each $T > 0$, filtered probability space $(\Omega, \mathcal{F}, \mathbb{F} = \{\mathcal{F}_t\}_{t\in[0,T]}, \mathbb{P})$ satisfying the usual condition and Euclidean space $(E, |\cdot|)$, we introduce the following spaces:

- $C([0,T]; E)$ is the space of continuous functions $\phi : [0,T] \to E$ satisfying $\|\phi\|_{C([0,T];E)} = \sup_{t\in[0,T]} |\phi_t| < \infty$;
- $C^1([0,T]; E)$ is the space of continuously differentiable functions $\phi : [0,T] \to E$ satisfying $\|\phi\|_{C^1([0,T];E)} = \sup_{t\in[0,T]}(|\phi_t| + |\phi'_t|) < \infty$;
- $\mathcal{S}^2(E)$ is the space of E-valued $\mathbb{F}$-progressively measurable càdlàg processes $X : \Omega \times [0,T] \to E$ satisfying $\|X\|_{\mathcal{S}^2(E)} = \mathbb{E}[\sup_{t\in[0,T]} |X_t|^2]^{1/2} < \infty$;
- $\mathcal{H}^2(E)$ is the space of E-valued $\mathbb{F}$-progressively measurable processes $X : \Omega \times [0,T] \to E$ satisfying $\|X\|_{\mathcal{H}^2(E)} = \mathbb{E}[\int_0^T |X_t|^2 \,\mathrm{d}t]^{1/2} < \infty$.

For notation simplicity, we denote by $C \in [0, \infty)$ a generic constant, which depends only on the constants appearing in the assumptions and may take a different value at each occurrence.

2. Problem Formulation and Main Results

2.1 Linear-Quadratic Reinforcement Learning problem

In this section, we consider the linear-quadratic reinforcement learning (LQ-RL) problem, where the drift coefficient of the state dynamics is unknown to the controller.

More precisely, let $T \in (0, \infty)$ be a given terminal time, W be an n-dimensional standard Brownian motion defined on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and $\mathbb{F} = (\mathcal{F}_t)_{t\in[0,T]}$ be the filtration generated by W augmented by the $\mathbb{P}$-null sets. Let $x_0 \in \mathbb{R}^n$ be a given initial state and $(A^\star, B^\star) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times d}$ be fixed but unknown matrices, consider the following problem:

$$\inf_{U\in\mathcal{H}^2(\mathbb{R}^d)} J^{\theta^\star}(U), \quad \text{with} \quad J^{\theta^\star}(U)=\mathbb{E}\left[\int_0^T \left((X_t^{\theta^\star,U})^\top Q X_t^{\theta^\star,U}+(U_t)^\top R U_t\right)\mathrm{d}t\right], \tag{2.1}$$

where for each $U\in\mathcal{H}^2(\mathbb{R}^d)$, the process $X^{\theta^\star,U}\in\mathcal{S}^2(\mathbb{R}^n)$ satisfies the following controlled dynamics associated with the parameter $\theta^\star=(A^\star,B^\star)^\top$:

$$\mathrm{d}X_t=(A^\star X_t+B^\star U_t)\,\mathrm{d}t+\mathrm{d}W_t, \quad t\in[0,T]; \quad X_0=x_0, \tag{2.2}$$

with given matrices $Q\in\mathbb{S}^n_0$ and $R\in\mathbb{S}^d_+$. Note that we assume the loss functional (2.1) only involves a time homogeneous running cost to allow a direct comparison with infinite-time horizon RL problems (see e.g., Duncan et al. (1992)), but similar analysis can be performed if the cost functions are time inhomogeneous, a terminal cost is included, or the Brownian motion W in (2.2) is scaled by an known nonsingular diffusion matrix.

If the parameter $\theta^\star=(A^\star,B^\star)^\top$ are known to the controller, then (2.1)-(2.2) reduces to the classical LQ control problems. In this case, it is well known that (see e.g., Yong and Zhou (1999) and the references therein), the optimal control $U^{\theta^\star}$ of (2.1)-(2.2) is given in a feedback form by

$$U_t^{\theta^\star}=\psi^{\theta^\star}(t,X_t^{\theta^\star}), \quad \text{with } \psi^{\theta^\star}(t,x)=K_t^{\theta^\star}x,\ \forall(t,x)\in[0,T]\times\mathbb{R}^n, \tag{2.3}$$

where $K_t^{\theta^\star}=-R^{-1}(B^\star)^\top P_t^{\theta^\star}$ for all $t\in[0,T]$, $(P_t^{\theta^\star})_{t\in[0,T]}$ solves the Riccati equation

$$\tfrac{\mathrm{d}}{\mathrm{d}t}P_t+(A^\star)^\top P_t+P_tA^\star-P_t(B^\star R^{-1}(B^\star)^\top)P_t+Q=0, \quad t\in[0,T]; \quad P_T=0, \tag{2.4}$$

and $X^{\theta^\star}$ is the state process governed by the following dynamics:

$$\mathrm{d}X_t=(A^\star X_t+B^\star K_t^{\theta^\star}X_t)\,\mathrm{d}t+\mathrm{d}W_t, \quad t\in[0,T]; \quad X_0=x_0. \tag{2.5}$$

To solve the LQ-RL problem (2.1)-(2.2) with unknown $\theta^\star$, the controller searches for the optimal control while simultaneously learning the system, i.e., the matrices $A^\star, B^\star$. In an episodic (also known as reset or restart) learning framework, the controller improves her knowledge of the underlying dynamics X_t through successive learning episodes, in order to find a control that is close to the optimal one.

Mathematically, it goes as follows. Let $M\in\mathbb{N}$ be the total number of learning episodes. In the i-th learning episode, $i=1,\dots,M$, a feedback control ψ^i is exercised, and the state process X^{ψ^i} evolves according to the dynamics (2.2) controlled by the policy ψ^i:

$$\mathrm{d}X_t=(A^\star X_t+B^\star\psi^i(t,X_t))\mathrm{d}t+\mathrm{d}W_t^i, \quad t\in[0,T]; \quad X_0=x_0. \tag{2.6}$$

Here $W^i, i=1,2,\dots,M$ are independent n-dimensional Brownian motions defined on the same probability space $(\Omega,\mathcal{F},\mathbb{P})$. The (expected) cost of learning in the i-th episode is then given by

$$J^{\theta^\star}(U^{\psi^i})=\mathbb{E}\left[\int_0^T \left((X_t^{\psi^i})^\top Q X_t^{\psi^i}+(U_t^{\psi^i})^\top R U_t^{\psi^i}\right)\mathrm{d}t\right], \text{ with } U_t^{\psi^i}\coloneqq\psi^i(t,X_t^{\psi^i}),\ t\in[0,T], \tag{2.7}$$

and the (expected) regret of learning up to $M \in \mathbb{N}$ episodes (with the sequence of controls $(U^{\psi^i})_{i=1}^M$) is defined as follows:

$$R(M) = \sum_{i=1}^{M} \Big(J^{\theta^\star}(U^{\psi^i}) - J^{\theta^\star}(U^{\theta^\star}) \Big), \tag{2.8}$$

where $J^{\theta^\star}(U^{\theta^\star})$ is the optimal cost of (2.1)-(2.2) when $A^\star, B^\star$ are known. Intuitively, the regret characterizes the cumulative loss from taking sub-optimal policies in all episodes.

In the following, we shall propose several least-squares-based learning algorithms to solve (2.1)-(2.2), and prove that they achieve logarithmic regrets if $\theta^\star$ is identifiable (see Remark 2.1 for details).

2.2 Continuous-Time Least-Squares Algorithm and Its Regret Bound

In this section, we consider a continuous-time least-squares algorithm, which chooses the optimal feedback control based on the current estimation of the parameter, and updates the parameter estimation based on the whole trajectories of the state dynamics.

More precisely, let $\theta = (A, B)^\top \in \mathbb{R}^{(n+d)\times n}$ be the current estimate of the unknown parameter $\theta^\star$, then the controller would exercise the optimal feedback control ψ^θ for (2.1)-(2.2) with $\theta^\star$ replaced by θ, i.e.,

$$\psi^\theta(t,x) = K_t^\theta x, \quad K_t^\theta := -R^{-1}B^\top P_t^\theta, \quad \forall (t,x) \in [0,T]\times\mathbb{R}^n, \tag{2.9}$$

where P^θ satisfies the Riccati equation (2.4) with $\theta^\star$ replaced by θ:

$$\tfrac{\mathrm{d}}{\mathrm{d}t} P_t + A^\top P_t + P_t A - P_t(BR^{-1}B^\top)P_t + Q = 0, \quad t \in [0,T]; \quad P_T = 0. \tag{2.10}$$

This leads to the state process X^{ψ^θ} satisfying (cf. (2.6)):

$$\mathrm{d}X_t = (A^\star X_t + B^\star \psi^\theta(t, X_t))\,\mathrm{d}t + \mathrm{d}W_t, \quad t \in [0,T]; \quad X_0 = x_0. \tag{2.11}$$

We proceed to derive an ℓ_2-regularized least-squares estimation for $\theta^\star$ based on sampled trajectories of X^{ψ^θ}. Observing from (2.11) that

$$Z_t^{\psi^\theta}(\mathrm{d}X_t^{\psi^\theta})^\top = Z_t^{\psi^\theta}(Z_t^{\psi^\theta})^\top\theta^\star \mathrm{d}t + Z_t^{\psi^\theta}(\mathrm{d}W_t)^\top, \quad \text{with } Z_t^{\psi^\theta} = \begin{pmatrix} X^{\psi^\theta} \\ \psi^\theta(t, X_t^{\psi^\theta}) \end{pmatrix} \text{ for all } t \in [0,T].$$

Hence the martingale property of the Itô integral implies that

$$\theta^\star = \left(\mathbb{E}\left[\int_0^T Z_t^{\psi^\theta}(Z_t^{\psi^\theta})^\top\,\mathrm{d}t \right] \right)^{-1} \mathbb{E}\left[\int_0^T Z_t^{\psi^\theta}(\mathrm{d}X_t^{\psi^\theta})^\top \right], \tag{2.12}$$

provided that $\mathbb{E}\big[\int_0^T Z_t^{\psi^\theta}(Z_t^{\psi^\theta})^\top\,\mathrm{d}t\big]$ is invertible. This suggests a practical rule to improve one's estimate θ for the true parameter $\theta^\star$, by replacing the expectations in (2.12) with empirical averages over independent realizations. More precisely, let $m \in \mathbb{N}$ and $(X_t^{\psi^\theta,i}, \psi^\theta(t, X_t^{\psi^\theta,i}))_{t\in[0,T]}$, $i = 1, \ldots, m$, be trajectories of m independent realizations of the state and control processes, we shall update the estimate θ by the following rule, inspired by (2.12):

$$\theta \longleftarrow \left(\frac{1}{m}\sum_{i=1}^{m}\int_0^T Z_t^{\psi^\theta,i}(Z_t^{\psi^\theta,i})^\mathsf{T}\,\mathrm{d}t+\frac{1}{m}I\right)^{-1}\left(\frac{1}{m}\sum_{i=1}^{m}\int_0^T Z_t^{\psi^\theta,i}(\mathrm{d}X_t^{\psi^\theta,i})^\mathsf{T}\right), \tag{2.13}$$

where $Z_t^{\psi^\theta,i} := \begin{pmatrix} X_t^{\psi^\theta,i} \\ \psi^\theta(t,X_t^{\psi^\theta,i}) \end{pmatrix}$ for all $t\in[0,T]$ and $i=1,\ldots,m$, and I is the $(n+d)\times(n+d)$ identity matrix.

The regularization term $\frac{1}{m}I$ in (2.13) guarantees the required matrix inverse and vanishes as $m\to\infty$. The estimator (2.13) can be equivalently expressed as an ℓ_2-regularized least-squares estimator, as pointed out in Duncan et al. (1992) for the ergodic LQ-RL problem.

We summarize the continuous-time least-squares algorithm as follows.

Algorithm 1 Continuous-time least-squares algorithm

Input: Choose an initial estimation θ_0 of $\theta^\star$ and numbers of learning episodes $\{m_\ell\}_{\ell\in\mathbb{N}\cup\{0\}}$.
for $\ell=0,1,\cdots$ **do**
  Obtain the feedback control ψ^{θ_ℓ} as (2.9) with $\theta=\theta_\ell$.
  Execute the feedback control ψ^{θ_ℓ} for m_ℓ independent episodes, and collect the trajectory data $(X_t^{\psi^{\theta_\ell},i},\psi^{\theta_\ell}(t,X_t^{\psi^{\theta_\ell},i}))_{t\in[0,T]}$, $i=1,\ldots,m_\ell$.
  Obtain an updated estimation $\theta_{\ell+1}$ by using (2.13) and the m_ℓ trajectories collected above.
end for

Note that Algorithm 1 operates in cycles, with m_ℓ the number of episodes in the ℓ-th cycle. Hence, the regret of learning up to M episodes (cf. (2.8)) can be upper bounded by the accumulated regret at the end of the L-th cycle, where L is the smallest integer such that $\sum_{\ell=0}^L m_\ell \ge M$.

In this section, we analyze the regret of Algorithm 1 based on the following assumptions of the learning problem (2.1)-(2.2).

H.1 *(1)* $T\in(0,\infty)$*,* $n,d\in\mathbb{N}$*,* $x_0\in\mathbb{R}^n$*,* $A^\star\in\mathbb{R}^{n\times n}$*,* $B^\star\in\mathbb{R}^{n\times d}$*,* $Q\in\mathbb{S}_0^n$ *and* $R\in\mathbb{S}_+^d$*.*

(2) $\{v\in\mathbb{R}^d \mid (K_t^{\theta^\star})^\top v=0,\ \forall t\in[0,T]\}=\{0\}$*, with* $K^{\theta^\star}$ *defined in* (2.3)*.*

Before discussing the regret of Algorithm 1, we make the following remark of (H.1).

Remark 2.1 (Self-exploration of finite-time horizon RL problems) *(H.1(1)) is the standard assumption for finite-time horizon LQ-RL problems (see e.g., Hambly et al. (2020)), except that H.1(1) allows* Q *to be positive semidefinite, which is important for costs depending on partial states. (H.1(2)) corresponds to the identifiability of the true parameter* $\theta^\star$ *by executing the optimal policy* $K^{\theta^\star}$*. In fact, as shown in Proposition 3.10, under (H.1(1)), (H.1(2)) is equivalent to the following statement:*

(2') if $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^d$ satisfy $u^\top X^{\theta^\star} + v^\top U^{\theta^\star} = 0$ for $\mathrm{d}\mathbb{P} \otimes \mathrm{d}t$-almost everywhere in $\Omega \times [0, T]$, then $u = 0$ and $v = 0$, where $X^{\theta^\star}$ and $U^{\theta^\star}$ are the optimal state and control processes of (2.1)-(2.2) defined by (2.5) and (2.3), respectively,

Item (2') indicates an important self-exploration property of finite-time horizon continuous-time RL problems. In particular, the time-dependent optimal feedback matrix $K^{\theta^\star}$ and the non-degenerate noises guarantee the non-degeneracy of the space spanned by $X^{\theta^\star}$ and $U^{\theta^\star}$, enabling learning the parameters sufficiently well. This self-exploration property is critical for our design of exploration-free learning algorithms for (2.1)-(2.2) with a logarithmic regret (see Theorems 2.2 and 2.3).

One can easily show that (H.1(2)) holds if the optimal policy $(K^{\theta^\star})_{t\in[0,T]}$ is nondegenerate, i.e., $\sup_{t\in[0,T]} \lambda_{\min}\left((K_t^{\theta^\star})(K_t^{\theta^\star})^\top\right) > 0$. Similar nondegeneracy condition has been imposed in Cassel et al. (2020) for discrete-time ergodic LQ-RL problems. In particular, by assuming that the optimal stationary policy satisfies $\lambda_{\min}(K^\star(K^\star)^\top) > 0$ (along with other controllablity conditions), they propose learning algorithms with a logarithmic regret, under the assumption that only the control coefficient $B^\star$ is unknown. In contrast, we allow both the state coefficient $A^\star$ and the control coefficient $B^\star$ to be unknown.

Moreover, the following proposition provides sufficient conditions of (H.1(2)), which are special cases of Proposition 3.11.

Proposition 2.1 *Let $n, d \in \mathbb{N}$, $Q \in \mathbb{S}_0^n$ and $R \in \mathbb{S}_+^d$.*

(1) If $(B^\star)^\top Q B^\star \in \mathbb{S}_+^d$, then (H.1(2)) holds for all $T > 0$.

(2) Assume that the algebraic Riccati equation $(A^\star)^\top P + P A^\star - P(B^\star R^{-1}(B^\star)^\top)P + Q = 0$ admits a unique maximal solution $P_\infty^\star \in \mathbb{S}_+^n$. Let $K_\infty^\star = -R^{-1}(B^\star)^\top P_\infty^\star$, and for each $T > 0$, let $P^{\star,(T)} \in C([0,T]; \mathbb{S}_0^n)$ be defined in (2.4). Assume that $\lim_{T\to\infty} P_0^{\star,(T)} = P_\infty^\star$ and $K_\infty^\star (K_\infty^\star)^\top \in \mathbb{S}_+^d$. Then there exists $T_0 > 0$, such that (H.1(2)) holds for all $T \geq T_0$.

Proposition 2.1 provides two sets of conditions for (H.1(2)) under two different scenarios: Item (1) applies to an arbitrary finite $T > 0$, and Item (2) only applies to sufficiently large T. Item (2) assumes the asymptotic behavior of solutions to Riccati differential equations, which can be ensured by the stabilizability of the pair $(A^\star, B^\star)$ and detectability of the pair $(A^\star, Q^{1/2})$ (see (Bitmead and Gevers, 1991, Theorems 10.9 and 10.10)). Note that our subsequent analysis is based on (H.1), and does not require stabilizability assumptions.

Remark 2.2 (Stabilizability of $(A^\star, B^\star)$ and dependence on T) *Since the LQ-RL problem (2.1)-(2.2) is over the time horizon $[0, T]$ with a fixed $T < \infty$, in general one does not need additional conditions on $(A^\star, B^\star)$ for the well-definedness of (2.1)-(2.2). If $T = \infty$, then some stabilizability/controllability conditions of $(A^\star, B^\star)$ may be required for (2.1)-(2.2) to ensure a well-defined solution (see e.g., Dean et al. (2019)). Under these conditions, different algorithms have been shown to achieve sub-linear regret with respect to the number of decision steps (see e.g., Mania et al. (2019); Cohen et al. (2019)), and even logarithmic regrets provided that further identifiability assumptions are satisfied (see e.g., Faradonbeh et al. (2018b, 2020); Cassel et al. (2020); Lale et al. (2020b)); see Section 1.1 for more* *details. For $T<\infty$, the regrets of learning algorithms for* (2.1)-(2.2) *in general depend exponentially on the time horizon T (e.g., the constants C_0, C' in Theorem 2.2), as the moments of the optimal state process $X^{\theta^\star}$ and control process $U^{\theta^\star}$ may grow exponentially with respect to T. It would be interesting to quantify the precise dependence of the regret bounds on T. This would entail deriving precise a priori bounds of solutions to* (2.10) *and estimating the norm $\|(\mathbb{E}[\int_0^T Z_t^{\theta^\star}(Z_t^{\theta^\star})^\top \,\mathrm{d}t])^{-1}\|_2$ in terms of $(A^\star, B^\star, Q, T)$, and is left for future research.*

We are now ready to state the main result of this section, which shows that the regret of Algorithm 1 grows logarithmically with respect to the number of episodes.

Theorem 2.2 *Suppose (H.1) holds and let $\theta_0=(A_0,B_0)^\top \in \mathbb{R}^{(n+d)\times d}$ such that (H.1(2)) holds with θ_0. Then there exists a constant $C_0>0$ such that for all $C\geq C_0$, and $\delta\in(0,\frac{3}{\pi^2})$, if one sets $m_0=C(-\ln\delta)$ and $m_\ell=2^\ell m_0$ for all $\ell\in\mathbb{N}$, then with probability at least $1-\frac{\pi^2\delta}{3}$, the regret of Algorithm 1 given by* (2.8) *satisfies*

$$R(M)\leq C'\big((\ln M)(\ln\ln M)+(-\ln\delta)(\ln M)\big),\quad \forall M\in\mathbb{N},$$

where C' is a constant independent of M and δ.

To simplify the presentation, we analyze the performance of Algorithm 1 by assuming the number of learning episodes $\{m_\ell\}_\ell$ is doubled between two successive updates of the estimation of $\theta^\star$. Similar regret results can be established for Algorithm 1 with different choices of $\{m_\ell\}_\ell$. Under this specific choice of $\{m_\ell\}_\ell$, for any $M\in\mathbb{N}$, Algorithm 1 splits M episodes into $L=\lceil\log_2(\frac{M}{m_0}+1)\rceil-1$ cycles, where the ℓ-th cycle, $\ell=0,1,\ldots,L-1$, contains m_ℓ episodes, and the remaining $M-\sum_{\ell=0}^{L-1}m_\ell$ episodes are in the last cycle.

Sketched proof of Theorem 2.2. We outline the key steps of the proof of Theorem 2.2, and present the detailed arguments to Section 3.3. By exploiting the regularity and robustness of solutions to (2.10), we prove that the performance gap $J^{\theta^\star}(U^{\psi^\theta})-J^{\theta^\star}(U^{\theta^\star})$ is of the magnitude $\mathcal{O}(|\theta-\theta^\star|^2)$, for all *a-priori* bounded θ (Proposition 3.8). We then establish a uniform sub-exponential property for the (deterministic and stochastic) integrals in (2.12), which along with (H.1(2)) and Bernstein's inequality leads to the following estimate of the parameter estimation error: for all $\delta\in(0,1/2)$, all sufficiently large $m\in\mathbb{N}$, and all θ sufficiently close to $\theta^\star$,

$$|\hat\theta-\theta^\star|\leq\mathcal{O}\Big(\sqrt{\frac{-\ln\delta}{m}}+\frac{-\ln\delta}{m}+\frac{(-\ln\delta)^2}{m^2}\Big),\quad\text{with probability } 1-2\delta, \tag{2.14}$$

where $\hat\theta$ is generated by (2.13) with ψ^θ (Proposition 3.9). Then for each $\delta>0$, applying (2.14) with $\delta_\ell=\delta/(\ell+1)^2$ for all $\ell\in\mathbb{N}\cup\{0\}$ shows that with probability $1-2\sum_{\ell=0}^\infty\delta_\ell=1-\frac{\pi^2\delta}{3}$,

$$|\hat\theta_{\ell+1}-\theta^\star|^2\lesssim\frac{-\ln\delta_\ell}{m_\ell}+\frac{(-\ln\delta_\ell)^2}{m_\ell^2}+\frac{(-\ln\delta_\ell)^4}{m_\ell^4},\quad\forall\ell\in\mathbb{N}, \tag{2.15}$$

where $\lesssim$ means the inequality holds with a multiplicative constant independent of δ and ℓ. By the quadratic performance gap and the choice of $\{m_\ell\}_\ell$, the regret of Algorithm

1 up to the M-th episode can be bounded by the regret at the end of L-th cycle with $L = \lceil \log_2(\frac{M}{m_0}+1)\rceil - 1$:

$$R(M) \lesssim \sum_{\ell=0}^{L} m_\ell |\theta_\ell - \theta^\star|^2 \lesssim \sum_{\ell=0}^{L} (-\ln\delta_\ell)\left(1 + \frac{-\ln\delta_\ell}{m_\ell} + \frac{(-\ln\delta_\ell)^3}{m_\ell^3}\right). \tag{2.16}$$

Observe that the choices of $\{\delta_\ell\}_\ell$ and $\{m_\ell\}_\ell$ ensure that $\sup_{\delta\in(0,\frac{3}{\pi^2}),\ell\in\mathbb{N}} \frac{-\ln\delta_\ell}{m_\ell} < \infty$. Hence, the right-hand side of (2.16) is of the magnitude $\mathcal{O}\left(\sum_{\ell=0}^{L}(-\ln\delta_\ell)\right)$, which along with the choices of δ_ℓ and L leads to the desired regret bound; see the end of Section 3.3 for more details.

2.3 Discrete-Time Least-Squares Algorithm and Its Regret Bound

Note that Algorithm 1 in Section 2.2 requires executing feedback controls and observing corresponding state trajectories continuously. A common practice to solve continuous-time RL problems is by assuming that at each learning episode the dynamics only evolves in discrete time, and then estimate parameters according to discrete-time RL algorithms (see e.g., Munos and Bourgine (1998); Munos (2000, 2006); Tallec et al. (2019)). As the true dynamics evolves continuously, it is necessary to quantify the impact of reaction stepsize on the algorithm performance.

In this section, we analyze the performance of the above procedure for solving (2.1)-(2.2). We adapt regularized least-squares algorithms for discrete-time LQ problems to the present setting, and establish their regret bounds in terms of the discretization stepsize. Our analysis shows that a proper scaling of the regularization term in the least-squares estimation in terms of stepsize is critical for a robust performance with respect to different timescales.

More precisely, for a given cycle (i.e., the index ℓ in Algorithm 1), let $\theta = (A, B)^\top \in \mathbb{R}^{(n+d)\times n}$ be the current estimate of $\theta^\star$ in (2.2), and let $\{t_i\}_{i=0}^N$, $N \in \mathbb{N}$, be a uniform partition of $[0, T]$ with stepsize $\tau = T/N$. We then assume that (2.1)-(2.2) is piecewise constant between any two grid points $\{t_i\}_{i=0}^N$, choose actions and make observations every τ, and update the estimated parameter based on these observations. To this end, we consider the following discrete-time LQ control problem with parameter θ:

$$\inf_{U\in\mathcal{H}^2_N(\mathbb{R}^d)} J_N(U), \quad \text{with } J_N(U) = \mathbb{E}\left[\sum_{i=0}^{N-1}\left((X^{U,\tau}_{t_i})^\top Q X^{U,\tau}_{t_i} + U_{t_i}^\top R U_{t_i}\right)\tau\right], \tag{2.17}$$

where $\mathcal{H}^2_N(\mathbb{R}^d) = \{U \in \mathcal{H}^2(\mathbb{R}^d) \mid U_t = U_{t_i}, t \in [t_i, t_{i+1}), i = 0, \dots, N-1\}$, and $(X^{U,\tau}_{t_i})_{i=0}^{N-1}$ are defined by

$$X^{U,\tau}_{t_{i+1}} - X^{U,\tau}_{t_i} = (AX^{U,\tau}_{t_i} + BU_{t_i})\tau + W_{t_{i+1}} - W_{t_i}, \quad i = 0, \dots N-1; \quad X^{U,\tau}_0 = x_0. \tag{2.18}$$

Note that for simplicity, our strategy is constructed by assuming a discrete-time dynamics arising from an Euler discretization of (2.2) (with the estimated parameter θ); similar analysis can be performed with a high-order approximation of (2.1)-(2.2).

It is well-known that (see e.g., Bitmead and Gevers (1991)), the optimal control of (2.17)-(2.18) is given by the following feedback form:

$$U_t = \psi^{\theta,\tau}(t, X_t^{U,\tau}), \qquad \text{with } \psi^{\theta,\tau}(t,x) = K_t^{\theta,\tau} x, \quad \forall (t,x) \in [0,T) \times \mathbb{R}^n, \tag{2.19}$$

where $K^{\theta,\tau} : [0,T) \to \mathbb{R}^{d\times n}$ is the piecewise constant function (with stepsize $\tau = T/N$) defined by

$$\begin{aligned}
P_{t_{i-1}}^{\theta,\tau} &= \tau Q + (I+\tau A)^\top P_{t_i}^{\theta,\tau}(I+\tau A) - (I+\tau A)^\top P_{t_i}^{\theta,\tau} \tau B(R + \tau B^\top P_{t_i}^{\theta,\tau} B)^{-1} B^\top P_{t_i}^{\theta,\tau}(I+\tau A), \\
&\qquad \forall i = 0,\ldots,N-1; \quad P_T^{\theta,\tau} = 0, \\
K_t^{\theta,\tau} &= -(R + \tau B^\top P_{t_{i+1}}^{\theta,\tau} B)^{-1} B^\top P_{t_{i+1}}^{\theta,\tau}(I+\tau A), \quad t \in [t_i, t_{i+1}),\ i = 0,\ldots,N-1.
\end{aligned} \tag{2.20}$$

We then implement the piecewise constant strategy $\psi^{\theta,\tau}$ defined in (2.19) on the original system (2.2) for m episodes, and update the estimated parameter θ by observing (2.2) with stepsize $\tau = T/N$. More precisely, let $X^{\psi^{\theta,\tau}} \in \mathcal{S}^2(\mathbb{R}^n)$ be the state process associated with $\psi^{\theta,\tau}$:

$$\mathrm{d}X_t = (A^\star X_t + B^\star K_{t_i}^{\theta,\tau} X_t)\,\mathrm{d}t + \mathrm{d}W_t, \quad t \in [t_i, t_{i+1}],\ i = 0,\ldots,N-1; \quad X_0 = x_0, \tag{2.21}$$

and $(X_t^{\psi^{\theta,\tau},j})_{t\in[0,T]}$, $j = 1,\ldots,m$, $m \in \mathbb{N}$, be m independent trajectories of $X^{\psi^{\theta,\tau}} \in \mathcal{S}^2(\mathbb{R}^n)$, we update the parameter θ according to the following discrete-time least-squares estimator:

$$\theta \leftarrow \underset{\theta\in\mathbb{R}^{(n+d)\times n}}{\arg\min} \sum_{j=1}^m \sum_{i=0}^{N-1} \| X_{t_{i+1}}^{\psi^{\theta,\tau},j} - X_{t_i}^{\psi^{\theta,\tau},j} - \tau\theta^\top Z_{t_i}^{\psi^{\theta,\tau},j} \|_2^2 + \tau \mathrm{tr}(\theta^\top\theta), \tag{2.22}$$

with $Z_{t_i}^{\psi^{\theta,\tau},j} := \begin{pmatrix} X_{t_i}^{\psi^{\theta,\tau},j} \\ K_{t_i}^{\theta,\tau} X_{t_i}^{\psi^{\theta,\tau},j} \end{pmatrix}$ for all i, j. The update (2.22) is consistent with the agent's assumption that the state evolves according to (2.18) between two grid points. Setting the derivative (with respect to θ) of the right-hand side of (2.22) to zero leads to

$$-\sum_{j=1}^m \sum_{i=0}^{N-1} \tau Z_{t_i}^{\psi^{\theta,\tau},j} \left(\left(X_{t_{i+1}}^{\psi^{\theta,\tau},j} - X_{t_i}^{\psi^{\theta,\tau},j} \right)^\top - \tau (Z_{t_i}^{\psi^{\theta,\tau},j})^\top \theta \right) + \tau\theta = 0.$$

Dividing both sides by τ/m and rearranging the terms give the following equivalent expression of the discrete-time least squares estimator (2.22):

$$\theta \longleftarrow \left(\frac{1}{m} \sum_{j=1}^m \sum_{i=0}^{N-1} Z_{t_i}^{\psi^{\theta,\tau},j} (Z_{t_i}^{\psi^{\theta,\tau},j})^\top \tau + \frac{1}{m} I \right)^{-1} \left(\frac{1}{m} \sum_{j=1}^m \sum_{i=0}^{N-1} Z_{t_i}^{\psi^{\theta,\tau},j} \left(X_{t_{i+1}}^{\psi^{\theta,\tau},j} - X_{t_i}^{\psi^{\theta,\tau},j} \right)^\top \right). \tag{2.23}$$

Remark 2.3 (Scaling hyper-parameters with timescales) *In principle, when applying discrete-time RL algorithms in a continuous environment, it is critical to adopt a proper scaling of the hyper-parameters for a robust performance with respect to different timescales.*

Indeed, scaling the regularization term $\mathrm{tr}(\theta^\top\theta)$ *in* (2.22) *with respect to the stepsize* τ *is essential for the robustness of* (2.23) *for all small stepsizes* τ*. If one updates* θ *by minimizing the following* ℓ_2*-regularized loss function with a given hyper-parameter* $\alpha < 1$ *such that*

$$\underset{\theta\in\mathbb{R}^{(n+d)\times n}}{\arg\min} \sum_{j=1}^{m}\sum_{i=0}^{N-1} \|X_{t_{i+1}}^{\psi^{\theta,\tau},j} - X_{t_i}^{\psi^{\theta,\tau},j} - \tau\theta^\top Z_{t_i}^{\psi^{\theta,\tau},j}\|_2^2 + \tau^{\alpha}\mathrm{tr}(\theta^\top\theta), \tag{2.24}$$

then the corresponding discrete-time estimator is given by

$$\theta^\tau := \left(\frac{1}{m}\sum_{j=1}^{m}\sum_{i=0}^{N-1} Z_{t_i}^{\psi^{\theta,\tau},j}(Z_{t_i}^{\psi^{\theta,\tau},j})^\top\tau + \frac{1}{\tau^{1-\alpha}m}I\right)^{-1}\left(\frac{1}{m}\sum_{j=1}^{m}\sum_{i=0}^{N-1} Z_{t_i}^{\psi^{\theta,\tau},j}\left(X_{t_{i+1}}^{\psi^{\theta,\tau},j} - X_{t_i}^{\psi^{\theta,\tau},j}\right)^\top\right).$$

Observe that for any given $m\in\mathbb{N}$*, the estimator* θ^τ *degenerates to zero as the stepsize* τ *tends to zero. Hence, to ensure the viability of* θ^τ *across different timescales, the number of episodes* m *has to increase appropriately when* τ *tends to zero. In contrast, by choosing* $\alpha = 1$ *in* (2.24)*,* (2.23) *admits a continuous-time limit* (2.13) *as* $\tau \to 0$*, and leads to a learning algorithm in which the episode numbers and the time stepsize can be chosen independently (see Theorem 2.3).*

We now summarize the discrete-time least-squares algorithm as follows.

Algorithm 2 Discrete-time least-squares algorithm

Input: Choose an initial estimation θ_0 of $\theta^\star$, numbers of learning episodes $\{m_\ell\}_{\ell\in\mathbb{N}\cup\{0\}}$ and numbers of intervention points $\{N_\ell\}_{\ell\in\mathbb{N}\cup\{0\}}$.
for $\ell = 0, 1, \cdots$ **do**
  Obtain the piecewise constant control $\psi^{\theta_\ell,\tau_\ell}$ as (2.19) with $\tau = T/N_\ell$ and $\theta = \theta_\ell$.
  Execute the control $\psi^{\theta_\ell,\tau_\ell}$ for m_ℓ independent episodes, and collect the data $X_{t_i}^{\psi^{\theta_\ell,\tau_\ell},j}$, $i = 0,\ldots,N_\ell$, $j = 1,\ldots,m_\ell$.
  Obtain an updated estimation $\theta_{\ell+1}$ by using (2.23) and the data $(X_{t_i}^{\psi^{\theta_\ell,\tau_\ell},j})_{i=0,\ldots,N_\ell,j=1,\ldots,m_\ell}$.
end for

Again, as the ℓ-th cycle of Algorithm 2 contains m_ℓ episodes, for each $M\in\mathbb{N}$, the regret of learning up to M episodes (cf. (2.8)) can be upper bounded by the accumulated regret at the end of the L-th cycle, where L is the smallest integer such that $\sum_{\ell=0}^{L} m_\ell \geq M$. The following theorem is an analogue of Theorem 2.2 for Algorithm 2.

Theorem 2.3 *Suppose (H.1) holds and let* $\theta_0 = (A_0, B_0)^\top \in \mathbb{R}^{(n+d)\times d}$ *such that (H.1(2)) holds with* θ_0*. Then there exists* $C_0 > 0$ *and* $n_0\in\mathbb{N}$ *such that for all* $C \geq C_0$*, and* $\delta\in(0,\frac{3}{\pi^2})$*, if one sets* $m_0 = C(-\ln\delta)$*,* $m_\ell = 2^\ell m_0$ *and* $N_\ell \geq n_0$ *for all* $\ell\in\mathbb{N}\cup\{0\}$*, then with probability at least* $1-\frac{\pi^2\delta}{3}$*, the regret of Algorithm 2 given by* (2.8) *satisfies*

$$R(M) \leq C'\left((\ln M)(\ln\ln M) + (-\ln\delta)(\ln M) + (-\ln\delta)\sum_{\ell=0}^{\ln M} 2^\ell N_\ell^{-2}\right), \quad \forall M\in\mathbb{N}, \tag{2.25}$$

where C' *is a constant independent of* M*,* δ *and* $(N_\ell)_{\ell\in\mathbb{N}\cup\{0\}}$*.*

Remark 2.4 *Theorem 2.3 provides a general regret bound of Algorithm 2 with any time discretization steps* $\{N_\ell\}_{\ell\geq 0}$*, where* N_ℓ *is the number of intervention points in the* ℓ*-th cycle. Compared with Algorithm 1, the regret of Algorithm 2 has an additional term* $(-\ln\delta)\sum_{\ell=0}^{\ln M} 2^\ell N_\ell^{-2}$*: for each learning episode, one achieves a sub-optimal loss by adjusting her policy in the discrete time and also suffers from model misspecification error in parameter estimation from discrete-time observations. Specifically,*

- *if the time discretization step is fixed for all cycles, i.e.,* $N_\ell = T/\tau$ *for all* ℓ*, then the last term of* (2.25) *is of the magnitude:*

$$\mathcal{O}\left((-\ln\delta)\sum_{\ell=0}^{\ln M} 2^\ell N_\ell^{-2}\right) = \mathcal{O}\left((-\ln\delta)\tau^2\sum_{\ell=0}^{\ln M} 2^\ell\right) = \mathcal{O}((-\ln\delta)\tau^2 M),$$

  and consequently Algorithm 2 achieves a sub-optimal linear regret;

- *if the time discretization step of the* ℓ*-th cycle increases exponentially in terms of* ℓ*, e.g.,* $N_\ell = \sqrt{2}^\ell N_0$ *for* $\ell = 1,\dots,\ln M$*, then the last term of* (2.25) *is of the magnitude:*

$$\mathcal{O}\left((-\ln\delta)\sum_{\ell=0}^{\ln M} 2^\ell N_\ell^{-2}\right) = \mathcal{O}\left((-\ln\delta)\sum_{\ell=0}^{\ln M} N_0^{-2}\right) = \mathcal{O}((-\ln\delta)\ln M),$$

  which guarantees that the regret of Algorithm 2 is still logarithmic in M*.*

Sketched proof of Theorem 2.3. We point out the main differences between the proofs of Theorems 2.2-2.3, and give the detailed proof of Theorem 2.3 in Section 3.4. Compared with Theorem 2.2, the essential challenges in proving Theorem 2.3 are to quantify the precise dependence of the performance gap and the parameter estimation error on the stepsize. To this end, we first prove a first-order convergence of (2.20) to (2.9) as the stepsize tends to zero. Then by exploiting the affine structure of (2.21), we establish the following quadratic performance gap for a piecewise constant policy $\psi^{\theta,\tau}$ (Proposition 3.12):

$$J^{\theta^\star}(U^{\psi^{\theta,\tau}}) - J^{\theta^\star}(U^{\theta^\star}) \leq C(|\theta-\theta^\star|^2 + \tau^2). \tag{2.26}$$

The analysis of the parameter estimation error is somewhat involved, as the state trajectories are merely α-Hölder continuous in time with $\alpha < 1/2$. By leveraging the analytic expression of $X^{\psi^{\theta,\tau}}$, we first show the first-order convergence of $\hat\theta^\tau$ to $\theta^\star$ with

$$\hat\theta^\tau := \left(\mathbb{E}\left[\sum_{i=0}^{N-1} Z_{t_i}^{\psi^{\theta,\tau}}(Z_{t_i}^{\psi^{\theta,\tau}})^\top\tau\right]\right)^{-1}\left(\mathbb{E}\left[\sum_{i=0}^{N-1} Z_{t_i}^{\psi^{\theta,\tau}}\left(X_{t_{i+1}}^{\psi^{\theta,\tau}} - X_{t_i}^{\psi^{\theta,\tau}}\right)^\top\right]\right). \tag{2.27}$$

We then prove that (2.23) enjoys a uniform sub-exponential tail bound for all θ close to $\theta^\star$ and small τ. Comparing (2.23) with (2.27) and applying the above results allow for bounding the estimation error of (2.23) by (2.14) with an additional $\mathcal{O}(\tau)$ term (Proposition 3.14).

3. Proofs of Theorems 2.2 and 2.3

To simplify the notation, for any given $N, m \in \mathbb{N}$ and control $\psi : [0,T] \times \mathbb{R}^n \to \mathbb{R}^d$ that is affine in the spatial variable, we introduce the following random variables associated with continuous-time observations:

$$
\begin{aligned}
V^{\psi} &= \int_0^T Z_t^{\psi} (Z_t^{\psi})^\top \, \mathrm{d}t, & Y^{\psi} &= \int_0^T Z_t^{\psi} (\mathrm{d}X_t^{\psi})^\top, \\
V^{\psi,m} &= \frac{1}{m} \sum_{j=1}^m \int_0^T Z_t^{\psi,j} (Z_t^{\psi,j})^\top \, \mathrm{d}t, & Y^{\psi,m} &= \frac{1}{m} \sum_{j=1}^m \int_0^T Z_t^{\psi,j} (\mathrm{d}X_t^{\psi,j})^\top,
\end{aligned} \tag{3.1}
$$

and the random variables associated with discrete-time observations with stepsize $\tau = T/N$:

$$
\begin{aligned}
V^{\psi,\tau} &= \sum_{i=0}^{N-1} Z_{t_i}^{\psi} (Z_{t_i}^{\psi})^\top \tau, & Y^{\psi,\tau} &= \sum_{i=0}^{N-1} Z_{t_i}^{\psi} (X_{t_{i+1}}^{\psi} - X_{t_i}^{\psi})^\top, \\
V^{\psi,\tau,m} &= \frac{1}{m} \sum_{j=1}^m \sum_{i=0}^{N-1} Z_{t_i}^{\psi,j} (Z_{t_i}^{\psi,j})^\top \tau, & Y^{\psi,\tau,m} &= \frac{1}{m} \sum_{j=1}^m \sum_{i=0}^{N-1} Z_{t_i}^{\psi,j} (X_{t_{i+1}}^{\psi,j} - X_{t_i}^{\psi,j})^\top,
\end{aligned} \tag{3.2}
$$

where X^{ψ} is the state process associated with the parameter $\theta^\star$ and the control ψ (cf. (2.6)), $Z_t^{\psi} = \begin{pmatrix} X^{\psi} \\ \psi(t, X_t^{\psi}) \end{pmatrix}$ for all $t \in [0,T]$, and $(X^{\psi,j}, Z^{\psi,j})_{j=1}^m$ are independent copies of (X^{ψ}, Z^{ψ}).

3.1 Convergence and Stability of Riccati Equations and Feedback Controls

Lemma 3.1 *Suppose (H.1(1)) holds. Then for all $\theta = (A, B)^\top \in \mathbb{R}^{(n+d)\times n}$, the Riccati equation*

$$
\tfrac{\mathrm{d}}{\mathrm{d}t} P_t + A^\top P_t + P_t A - P_t B R^{-1} B^\top P_t + Q = 0, \quad t \in [0,T]; \quad P_T = 0. \tag{3.3}
$$

admits a unique solution $P^{\theta} \in C([0,T]; \mathbb{R}^{n\times n})$. Moreover, the map $\mathbb{R}^{(n+d)\times n} \ni \theta \mapsto P^{\theta} \in C^1([0,T]; \mathbb{R}^{n\times n})$ is continuously differentiable.

Proof It has been shown in (Yong and Zhou, 1999, Corollary 2.10 on p. 297) that under (H.1(1)), for all $\theta = (A, B)^\top \in \mathbb{R}^{(n+d)\times n}$, (3.3) admits a unique solution $P^{\theta} \in C([0,T]; \mathbb{R}^{n\times n})$ such that $P_t^{\theta} \in \mathbb{S}_0^n$ for all $t \in [0,T]$. It remains to to prove the continuous differentiability of $\theta \mapsto P^{\theta}$.

To this end, consider the Banach spaces $\mathbb{X} = \mathbb{R}^{(n+d)\times n} \times C^1([0,T]; \mathbb{R}^{n\times n})$ and $\mathbb{Y} = C([0,T]; \mathbb{R}^{n\times n}) \times \mathbb{R}^{n\times n}$, and the operator $\Phi : \mathbb{X} \to \mathbb{Y}$ defined by

$$
\mathbb{X} \ni (\theta, P) \mapsto \Phi(\theta, P) \coloneqq (F(\theta, P), P_T) \in \mathbb{Y},
$$

where $F(\theta, P)_t = \frac{\mathrm{d}}{\mathrm{d}t} P_t + A^\top P_t + P_t A - P_t B R^{-1} B^\top P_t + Q$ for all $t \in [0,T]$. Observe that for all $\theta \in \mathbb{R}^{(n+d)\times n}$, $\Phi(P^{\theta}, \theta) = 0$. Moreover, one can easily show that for any $(P, \theta) \in \mathbb{X}$, Φ is continuously Fréchet differentiable at (P, θ), and the partial derivative $\frac{\partial}{\partial P} \Phi(\theta, P) : C^1([0,T]; \mathbb{R}^{n\times n}) \to \mathbb{Y}$ is a bounded linear operator such that for all $\tilde{P} \in C^1([0,T]; \mathbb{R}^{n\times n})$,

$$
\frac{\partial}{\partial P} \Phi(\theta, P)(\tilde{P}) = \begin{pmatrix} \left(\frac{\mathrm{d}}{\mathrm{d}t} \tilde{P}_t + A^\top \tilde{P}_t + \tilde{P}_t A - \tilde{P}_t B R^{-1} B^\top P_t - P_t B R^{-1} B^\top \tilde{P}_t \right)_{t \in [0,T]} \\ \tilde{P}_T \end{pmatrix} \in \mathbb{Y}.
$$

Classical well-posedness results of linear differential equations and the boundedness of P imply that $\frac{\partial}{\partial P}\Phi(P,\theta) : C^1([0,T];\mathbb{R}^{n\times n}) \to \mathbb{Y}$ has a bounded inverse (and hence a bijection). Thus, applying the implicit function theorem (see (Ciarlet, 2013, Theorem 7.13-1)) to Φ proves that $\mathbb{R}^{(n+d)\times n} \ni \theta \mapsto P^\theta \in C^1([0,T];\mathbb{R}^{n\times n})$ is continuously differentiable. ∎

The following lemma establishes the stability of the Riccati difference operator, which is crucial for the subsequent convergence analysis.

Lemma 3.2 *Suppose (H.1(1)) holds. For each $\theta = (A,B)^\top \in \mathbb{R}^{(n+d)\times n}$ and $N \in \mathbb{N}$, let $\tau = T/N$ and the function $\Gamma^\theta_\tau : \mathbb{S}^n_0 \to \mathbb{S}^n_0$ such that for all $P \in \mathbb{S}^n_0$,*

$$\Gamma^\theta_\tau(P) \coloneqq \tau Q + (I+\tau A)^\top P(I+\tau A) - (I+\tau A)^\top P\tau B(R+\tau B^\top PB)^{-1}B^\top P(I+\tau A). \tag{3.4}$$

Then for all $P, P' \in \mathbb{S}^n_0$,

(1) $\|\Gamma^\theta_\tau(P)\|_2 \le \tau\|Q\|_2 + (1+\tau\|A\|_2)^2\|P\|_2$,

(2) $\|\Gamma^\theta_\tau(P) - \Gamma^\theta_\tau(P')\|_2 \le \big(1+\tau\|R^{-1}\|_2\|B\|_2^2\max\{\|P\|_2,\|P'\|_2\}\big)^2(1+\tau\|A\|_2)^2\|P-P'\|_2$.

Proof Item (1) follows directly from the definition of Γ^θ_τ and the identity that $\|\Gamma^\theta_\tau(P)\|_2 = \sup\{x^\top\Gamma^\theta_\tau(P)x \mid x \in \mathbb{R}^n, |x| = 1\}$. We now prove Item (2). Let $\delta P = P - P'$ and $\delta\Gamma(P) = \Gamma^\theta_\tau(P) - \Gamma^\theta_\tau(P')$, by (Bitmead and Gevers, 1991, Lemma 10.1),

$$\delta\Gamma(P) = F^\top\delta PF - F^\top\delta P\tau B(\tau B^\top P\tau B + \tau R)^{-1}\tau B^\top\delta PF,$$

with $F = (I - \tau B(\tau B^\top P'\tau B + \tau R)^{-1}\tau B^\top P')(I+\tau A)$. Thus for all $x \in \mathbb{R}^n$, $x^\top\delta\Gamma(P)x \le \|\delta P\|_2\|F\|_2^2|x|^2$, which along with $\|(\tau B^\top P'B + R)^{-1}\|_2 \le \|R^{-1}\|_2$ implies

$$x^\top\delta\Gamma(P)x \le \|\delta P\|_2(1+\tau\|R^{-1}\|_2\|B\|_2^2\|P'\|_2)^2(1+\tau\|A\|_2)^2|x|^2, \quad x \in \mathbb{R}^n.$$

Hence, interchanging the roles of P and P' in the above inequality and taking the supremum over $x \in \mathbb{R}^n$ lead to the desired estimate. ∎

The following proposition establishes the first-order convergence of the Riccati difference equation and the associated feedback controls, as the stepsize tends to zero.

Proposition 3.3 *Suppose (H.1(1)) holds and let Θ be a bounded subset of $\mathbb{R}^{(n+d)\times n}$. For each $\theta = (A,B)^\top \in \Theta$ and $N \in \mathbb{N}$, let $(P^{\theta,\tau}_i)_{i=0}^N$ such that $P^{\theta,\tau}_N = 0$ and $P^{\theta,\tau}_i = \Gamma^\theta_\tau(P^{\theta,\tau}_{i+1})$ for all $i = 0,\dots,N-1$, with Γ^θ_τ defined in (3.4) with $\tau = T/N$. Then there exists a constant $C \ge 0$ such that for all $\theta \in \Theta, N \in \mathbb{N}$,*

$$\sup_{i=0,\dots,N-1}\ \sup_{t\in[i\tau,(i+1)\tau)} \big(\|P^\theta_t - P^{\theta,\tau}_i\|_2 + \|K^\theta_t - K^{\theta,\tau}_i\|_2\big) \le C\tau,$$

where $P^\theta \in C^1([0,T];\mathbb{R}^{n\times n})$ satisfies (3.3), $K^\theta_t = -R^{-1}B^\top P^\theta_t$ for all $t \in [0,T]$ and $K^{\theta,\tau}_i = -(R+\tau B^\top P^{\theta,\tau}_{i+1}B)^{-1}B^\top P^{\theta,\tau}_{i+1}(I+\tau A)$ for all $i = 0,\dots,N-1$.

Proof Throughout this proof, we shall fix $\theta \in \Theta, N \in \mathbb{N}$, let $t_i = i\tau$ for all $i = 0, \dots, N$, and denote by C a generic constant independent of N and θ. By the continuity of the map $\theta \mapsto P^\theta$ (Lemma 3.1) and the boundedness of Θ, there exists a constant C such that $\|P^\theta\|_{C^1([0,T];\mathbb{R}^{n\times n})} \le C$ for all $\theta \in \Theta$, which implies $\|P^\theta_t - P^\theta_s\|_2 \le C|t-s|$ for all $t, s \in [0,T]$. Consequently, it suffices to prove $\|P^\theta_{t_i} - P^{\theta,\tau}_i\|_2 + \|K^\theta_{t_{i+1}} - K^{\theta,\tau}_i\|_2 \le C\tau$ for all $i = 0, \dots, N-1$.

We start by making two important observations. By Lemma 3.2 Item (1), $\|P^{\theta,\tau}_i\|_2 \le \tau C + (1 + C\tau)\|P^{\theta,\tau}_{i+1}\|_2$ for all $i = 0, \dots, N-1$, which along with Gronwall's inequality gives $\|P^{\theta,\tau}_i\|_2 \le C$ for all $i = 0, \dots, N$. Moreover, by (3.4), for all $P \in \mathbb{S}^n_0$,

$$
\begin{aligned}
\Gamma^\theta_\tau(P) = \tau Q + P + \tau(A^\top P + PA) + \tau^2 A^\top PA \\
- \tau\Big(PB(R + \tau B^\top PB)^{-1}B^\top P + \tau(A^\top H + HA^\top) + \tau^2 A^\top HA\Big),
\end{aligned}
$$

with $H \coloneqq PB(R + \tau B^\top PB)^{-1}B^\top P$. Hence for any given $i = 0, \dots, N-1$, we see from (3.3) that

$$
\begin{aligned}
&P^\theta_{t_i} - \Gamma^\theta_\tau(P^\theta_{t_{i+1}}) \\
&= \int_{t_i}^{t_{i+1}} \big(A^\top(P^\theta_t - P^\theta_{t_{i+1}}) + (P^\theta_t - P^\theta_{t_{i+1}})A\big)\,\mathrm{d}t - \int_{t_i}^{t_{i+1}} (P^\theta_t BR^{-1}B^\top P^\theta_t - P^\theta_{t_{i+1}} BR^{-1}B^\top P^\theta_{t_{i+1}})\,\mathrm{d}t \\
&\quad - \int_{t_i}^{t_{i+1}} (P^\theta_{t_{i+1}} BR^{-1}B^\top P^\theta_{t_{i+1}} - P^\theta_{t_{i+1}} B(R + \tau B^\top P^\theta_{t_{i+1}} B)^{-1}B^\top P^\theta_{t_{i+1}})\,\mathrm{d}t \\
&\quad + \tau^2(-A^\top P^\theta_{t_{i+1}} A + A^\top H^\theta_{i+1} + H^\theta_{i+1}A^\top + \tau A^\top H^\theta_{i+1} A),
\end{aligned}
$$

with $H^\theta_{i+1} = P^\theta_{t_{i+1}} B(R + \tau B^\top P^\theta_{t_{i+1}} B)^{-1}B^\top P^\theta_{t_{i+1}}$. Since $\|P^\theta\|_{C^1([0,T];\mathbb{R}^{n\times n})} \le C$ and $R \in \mathbb{S}^d_+$, we have $\|P^\theta_{t_i} - \Gamma^\theta_\tau(P^\theta_{t_{i+1}})\|_2 \le C\tau^2$ for all $i = 0, \dots, N-1$.

We are ready to show $\max_{i=0,\dots,N-1}(\|P^\theta_{t_i} - P^{\theta,\tau}_i\|_2 + \|K^\theta_{t_{i+1}} - K^{\theta,\tau}_i\|_2) \le C\tau$. For any given $i = 0, \dots, N-1$, by Lemma 3.2 Item (2) and the uniform boundedness of $(P^\theta_{t_i})_{i=0}^N$ and $(P^{\theta,\tau}_i)_{i=0}^N$,

$$
\begin{aligned}
\|P^\theta_{t_i} - P^{\theta,\tau}_i\|_2 &\le \|P^\theta_{t_i} - \Gamma^\theta_\tau(P^\theta_{t_{i+1}})\|_2 + \|\Gamma^\theta_\tau(P^\theta_{t_{i+1}}) - \Gamma^\theta_\tau(P^{\theta,\tau}_{i+1})\|_2 \\
&\le C\tau^2 + \big(1 + \tau C \max\{\|P^\theta_{t_{i+1}}\|_2, \|P^{\theta,\tau}_{i+1}\|_2\}\big)^2(1+\tau)\|P^\theta_{t_{i+1}} - P^{\theta,\tau}_{i+1}\|_2 \\
&\le C\tau^2 + \big(1 + \tau C\big)\|P^\theta_{t_{i+1}} - P^{\theta,\tau}_{i+1}\|_2,
\end{aligned}
$$

which along with Gronwall's inequality and $P^\theta_T = P^{\theta,\tau}_N = 0$ shows the desired convergence rate of $(P^{\theta,\tau}_i)_{i=1}^N$. Furthermore, for all $i = 0, \dots, N-1$,

$$
\begin{aligned}
\|K^\theta_{t_{i+1}} - K^{\theta,\tau}_i\|_2 &\le \|(R^{-1} - (R + \tau B^\top P^{\theta,\tau}_{i+1} B)^{-1})B^\top P^\theta_{t_{i+1}}\|_2 \\
&\quad + \|(R + \tau B^\top P^{\theta,\tau}_{i+1} B)^{-1}B^\top(P^\theta_{t_{i+1}} - P^{\theta,\tau}_{i+1}(I + \tau A))\|_2 \le C\tau,
\end{aligned}
$$

from the facts that $\|P^\theta_{t_i}\|_2 \le C$, $\|P^{\theta,\tau}_i\|_2 \le C$ and $\|P^\theta_{t_i} - P^{\theta,\tau}_i\|_2 \le C\tau$ for all i. ■

3.2 Concentration Inequalities for Least-Squares Estimators

In this section, we analyze the concentration behavior of the least-squares estimators (2.13) and (2.23). We first recall the definition of sub-exponential random variables (see e.g., Wainwright (2019)).

Definition 3.1 *A random variable X with mean $\mu = \mathbb{E}[X]$ is (ν, b)-sub-exponential for $\nu, b \in [0,\infty)$ if $\mathbb{E}[e^{\lambda(X-\mu)}] \le e^{\nu^2\lambda^2/2}$ for all $|\lambda| < 1/b$.*

Note that a $(\nu, 0)$-sub-exponential random variable is usually called a sub-Gaussian random variable. It is well-known that products of sub-Gaussian random variables are sub-exponential, and the class of sub-exponential random variables forms a vector space. Moreover, sub-exponential random variables enjoy the following concentration inequality (also known as Bernstein's inequality; see e.g., (Wainwright, 2019, Equation 2.18 p. 29)).

Lemma 3.4 *Let $m \in \mathbb{N}$, $\nu, b \in [0,\infty)$ and $(X_i)_{i=1}^m$ be independent (ν, b)-sub-exponential random variables with $\mu = \mathbb{E}[X_i]$ for all $i = 1, \ldots, m$. Then for all $\epsilon \ge 0$,*

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m X_i - \mu\right| \ge \epsilon\right) \le 2\exp\left(-\min\left\{\frac{m\epsilon^2}{2\nu^2}, \frac{m\epsilon}{2b}\right\}\right).$$

The following lemma shows double iterated Itô integrals are sub-exponential random variables.

Lemma 3.5 *Let $L \ge 0$ and $g, h : [0,T] \times [0,T] \to \mathbb{R}^{n\times n}$ be measurable functions such that $|g(t,s)| \le L$ and $|h(t,s)| \le L$ for all $t, s \in [0,T]$. Then there exist $\nu, b \in [0,\infty)$, depending polynomially on L, n, T, such that*

(1) $\int_0^T \left(\int_0^t g(t,s)\,\mathrm{d}W_s\right)^\top \mathrm{d}W_t$,

(2) $\int_0^T \left(\int_0^t g(t,s)\,\mathrm{d}W_s\right)^\top \left(\int_0^t h(t,s)\,\mathrm{d}W_s\right)\mathrm{d}t$

are (ν, b)-sub-exponential,

Proof We first prove Item (1) by assuming without loss of generality that $\|g(t,s)\|_2 \le L$ for all $t, s \in [0,T]$, and by defining $V^q := \int_0^T \left(\int_0^t q(t,s)\,\mathrm{d}W_s\right)^\top \mathrm{d}W_t$ for any bounded measurable function $q : [0,T]\times[0,T] \to \mathbb{R}^{n\times n}$. By similar arguments as (Cheridito et al., 2005, Lemma 3.2), we have for all $t \in [0,T]$ and $0 \le \lambda < \frac{1}{2T}$,

$$\mathbb{E}[\exp(2\lambda V^{\frac{g}{L}})] \le \mathbb{E}[\exp(2\lambda V^{I_n})] = \left(\frac{1}{\sqrt{1-2\lambda T}}\exp(-\lambda T)\right)^n.$$

As $\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \le e^{2\lambda^2}$ for all $|\lambda| \le 1/4$, we see $\mathbb{E}[\exp(2\lambda V^{g/L})] \le \exp(2n\lambda^2T^2)$ for all $0 \le \lambda < \frac{1}{4T}$. Consequently, for all $0 \le \lambda < \frac{1}{2LT}$,

$$\mathbb{E}[\exp(\lambda V^g)] = \mathbb{E}\left[\exp\left(2\frac{\lambda L}{2}V^{\frac{g}{L}}\right)\right] \le \exp\left(\frac{nL^2T^2\lambda^2}{2}\right).$$

Replacing g by $-g$ shows the above estimate holds for $|\lambda| < \frac{1}{2LT}$, which implies the desired sub-exponential property of V^g.

For Item (2), observe that for each $t \in [0,T]$, the Itô formula allows one to express the product $\big(\int_0^t g(t,s)\,\mathrm{d}W_s\big)^\top \big(\int_0^t h(t,s)\,\mathrm{d}W_s\big)$ as a linear combination of double iterated Itô integrals and deterministic integrals. Then the desired sub-exponential property follows from the stochastic Fubini theorem (see e.g., Veraar (2012)) and Item (1). ∎

The following theorem establishes the concentration properties of the random variables involved in the least-squares estimators.

Theorem 3.6 *Suppose (H.1(1)) holds and let Θ be a bounded subset of $\mathbb{R}^{(n+d)\times n}$. For each $\theta \in \Theta$ and $N \in \mathbb{N}$, let ψ^θ be defined in* (2.9)*, and $\psi^{\theta,\tau}$ be defined in* (2.20) *with stepsize $\tau = T/N$. Then there exist constants $C, \nu, b > 0$ such that for all $\theta \in \Theta$, $N, m \in \mathbb{N}$ and $\epsilon > 0$,*

$$\max\Big\{\mathbb{P}(|V^{\psi^\theta,m} - \mathbb{E}[V^{\psi^\theta}]| \geq \epsilon), \mathbb{P}(|Y^{\psi^\theta,m} - \mathbb{E}[Y^{\psi^\theta}]| \geq \epsilon),$$
$$\mathbb{P}(|V^{\psi^{\theta,\tau},\tau,m} - \mathbb{E}[V^{\psi^{\theta,\tau},\tau}]| \geq \epsilon), \mathbb{P}(|Y^{\psi^{\theta,\tau},\tau,m} - \mathbb{E}[Y^{\psi^{\theta,\tau},\tau}]| \geq \epsilon)\Big\} \leq C \exp\left(-\frac{1}{C}\min\left\{\frac{m\epsilon^2}{\nu^2}, \frac{m\epsilon}{b}\right\}\right),$$

where V^{ψ^θ}, Y^{ψ^θ}, $V^{\psi^\theta,m}$, $Y^{\psi^\theta,m}$ are defined in (3.1)*, and $V^{\psi^{\theta,\tau},\tau}$, $Y^{\psi^{\theta,\tau},\tau}$, $V^{\psi^{\theta,\tau},\tau,m}$, $Y^{\psi^{\theta,\tau},\tau,m}$ are defined in* (3.2)*.*

Proof We first show there exist $\nu, b > 0$ such that all entries of V^{ψ^θ}, Y^{ψ^θ} $Y^{\psi^{\theta,\tau},\tau}$, $V^{\psi^{\theta,\tau},\tau}$ are (ν,b)-sub-exponential for all $\theta \in \Theta$ and $N \in \mathbb{N}$. By (3.1), we have

$$V^{\psi^\theta} = \int_0^T \begin{pmatrix} X_t^{\psi^\theta} \\ K_t^\theta X_t^{\psi^\theta} \end{pmatrix} \begin{pmatrix} (X_t^{\psi^\theta})^\top & (K_t^\theta X_t^{\psi^\theta})^\top \end{pmatrix} \mathrm{d}t, \quad Y^{\psi^\theta} = V^{\psi^\theta}\theta^\star + \int_0^T \begin{pmatrix} X_t^{\psi^\theta} \\ K_t^\theta X_t^{\psi^\theta} \end{pmatrix} (\mathrm{d}W_t)^\top.$$

Moreover, applying the variation-of-constants formula (see e.g., (Mao, 2007, Theorem 3.1 p. 96)) to (2.11) shows that $X_t^{\psi^\theta} = \Phi_t^\theta\big(x_0 + \int_0^t (\Phi_s^\theta)^{-1}\,\mathrm{d}W_s\big)$ for all $t \in [0,T]$, where $\Phi^\theta \in C([0,T];\mathbb{R}^{n\times n})$ is the fundamental solution of $\mathrm{d}\Phi_t^\theta = (A^\star + B^\star K_t^\theta)\Phi_t^\theta\,\mathrm{d}t$. The continuity of $\mathbb{R}^{(n+d)\times n} \ni \theta \mapsto K^\theta \in C([0,T];\mathbb{R}^{d\times n})$ (cf. Proposition 3.1) and the boundedness of Θ implies that $K^\theta, \Phi^\theta, (\Phi^\theta)^{-1}$ are uniformly bounded for all $\theta \in \Theta$. Consequently, from Lemma 3.5, there exist $\nu, b > 0$ such that all entries of V^{ψ^θ} and Y^{ψ^θ} are (ν,b)-sub-exponential.

Similarly, by (2.21) and (3.2),

$$\begin{aligned} V^{\psi^{\theta,\tau},\tau} &= \int_0^T \sum_{i=0}^{N-1} \mathbf{1}_{[t_i,t_{i+1})}(t) \begin{pmatrix} X_{t_i}^{\psi^{\theta,\tau}} \\ K_{t_i}^{\theta,\tau} X_{t_i}^{\psi^{\theta,\tau}} \end{pmatrix} \begin{pmatrix} (X_{t_i}^{\psi^{\theta,\tau}})^\top & (K_{t_i}^{\theta,\tau} X_{t_i}^{\psi^{\theta,\tau}})^\top \end{pmatrix} \mathrm{d}t, \\ Y^{\psi^{\theta,\tau},\tau} &= \int_0^T \sum_{i=0}^{N-1} \mathbf{1}_{[t_i,t_{i+1})}(t) \begin{pmatrix} X_{t_i}^{\psi^{\theta,\tau}} \\ K_{t_i}^{\theta,\tau} X_{t_i}^{\psi^{\theta,\tau}} \end{pmatrix} \begin{pmatrix} (X_t^{\psi^{\theta,\tau}})^\top & (K_t^{\theta,\tau} X_t^{\psi^{\theta,\tau}})^\top \end{pmatrix} (\theta^\star)^\top \mathrm{d}t \\ &\quad + \int_0^T \sum_{i=0}^{N-1} \mathbf{1}_{[t_i,t_{i+1})}(t) \begin{pmatrix} X_{t_i}^{\psi^{\theta,\tau}} \\ K_{t_i}^{\theta,\tau} X_{t_i}^{\psi^{\theta,\tau}} \end{pmatrix} (\mathrm{d}W_t)^\top, \end{aligned}$$

where $X_t^{\psi^{\theta,\tau}} = \Phi_t^{\theta,\tau}\big(x_0 + \int_0^t (\Phi_s^{\theta,\tau})^{-1}\,\mathrm{d}W_s\big)$ for all $t \in [0,T]$, and $\Phi^{\theta,\tau} \in C([0,T];\mathbb{R}^{n\times n})$ is the fundamental solution of $\mathrm{d}\Phi_t^{\theta,\tau} = (A^\star + B^\star K_t^{\theta,\tau})\Phi_t^{\theta,\tau}\,\mathrm{d}t$. By Proposition 3.3, $K^{\theta,\tau}, \Phi^{\theta,\tau}, (\Phi^{\theta,\tau})^{-1}$ are uniformly bounded for all $\theta \in \Theta$ and $N \in \mathbb{N}$, which along with Lemma 3.5 leads to the desired sub-exponential properties of $Y^{\psi^{\theta,\tau},\tau}$ and $V^{\psi^{\theta,\tau},\tau}$.

Finally, since $\mathbb{P}(|\sum_{i=1}^\ell X_i| \geq \epsilon) \leq \sum_{i=1}^\ell \mathbb{P}(|X_i| \geq \epsilon/\ell)$ for all $\ell \in \mathbb{N}$ and random variables $(X_i)_{i=1}^\ell$, we can apply Lemma 3.4 to each component of V^{ψ^θ}, Y^{ψ^θ} $Y^{\psi^{\theta,\tau},\tau}$ and $V^{\psi^{\theta,\tau},\tau}$, and conclude the desired concentration inequality with a constant C depending polynomially on n, d. ■

3.3 Regret Analysis of Continuous-Time Least-Squares Algorithm

This section is devoted to the proof of Theorem 2.2, which consists of three steps: (1) We first quantify the performance gap between applying feedback controls for an incorrect model and that for the true model; our proof exploits the stability of Riccati equations established in Lemma 3.1; (2) We then estimate the parameter estimation error in terms of the number of learning episodes based on the sub-exponential tail behavior of the least-squares estimator (2.13); (3) Finally, we estimate the regret for the feedback controls $(\psi^{\theta_\ell})_{\ell\in\mathbb{N}}$ in Algorithm 1, thus establishing Theorem 2.2.

Step 1: Analysis of the performance gap. We start by establishing a quadratic expansion of the cost function at any open-loop control.

Proposition 3.7 *Suppose (H.1(1)) holds. Let $\psi^{\theta^\star}$ be defined in* (2.3)*, $X^{\theta^\star}$ be the state process associated with $\psi^{\theta^\star}$ (cf.* (2.5)*), and $U^{\theta^\star} \in \mathcal{H}^2(\mathbb{R}^d)$ be such that for all $t \in [0,T]$, $U_t^{\theta^\star} = \psi^{\theta^\star}(t, X_t^{\theta^\star})$. Then for all $U \in \mathcal{H}^2(\mathbb{R}^d)$,*

$$J^{\theta^\star}(U) - J^{\theta^\star}(U^{\theta^\star}) \leq \|Q\|_2 \|X^{\theta^\star,U} - X^{\theta^\star}\|^2_{\mathcal{H}^2(\mathbb{R}^n)} + \|R\|_2 \|U - U^{\theta^\star}\|^2_{\mathcal{H}^2(\mathbb{R}^d)}, \tag{3.5}$$

where $X^{\theta^\star,U}$ is the state process controlled by U (cf. (2.2)*), and $J^{\theta^\star} : \mathcal{H}^2(\mathbb{R}^d) \to \mathbb{R}$ is defined in* (2.1)*.*

Proof For notational simplicity, for all $U \in \mathcal{H}^2(\mathbb{R}^d)$ and $\epsilon > 0$, we write $U^\epsilon = U^{\theta^\star} + \epsilon(U - U^{\theta^\star})$, denote by $X^\epsilon = X^{\theta^\star,U^\epsilon}$ the associated state process defined by (2.2), and by $X^U = X^0 = X^{\theta^\star,U}$. The affineness of (2.2) implies that $X^\epsilon = (1-\epsilon)X^{\theta^\star} + \epsilon X^U$ for all

$\epsilon > 0$. Hence, for all $U \in \mathcal{H}^2(\mathbb{R}^d)$,

$$
\begin{aligned}
&\lim_{\epsilon \searrow 0} \frac{1}{\epsilon}\left(J^{\theta^\star}(U^\epsilon) - J^{\theta^\star}(U^{\theta^\star})\right) \\
&= \lim_{\epsilon \searrow 0} \frac{1}{\epsilon}\mathbb{E}\bigg[\int_0^T \bigg(\Big((1-\epsilon)X_t^{\theta^\star} + \epsilon X_t^U\Big)^\top Q\Big((1-\epsilon)X_t^{\theta^\star} + \epsilon X_t^U\Big) - (X_t^{\theta^\star})^\top Q X_t^{\theta^\star} \\
&\quad + \Big((1-\epsilon)U_t^{\theta^\star} + \epsilon U_t\Big)^\top R\left((1-\epsilon)U_t^{\theta^\star} + \epsilon U_t\right) - (U_t^{\theta^\star})^\top R U_t^{\theta^\star}\bigg)\,\mathrm{d}t\bigg] \\
&= \lim_{\epsilon \searrow 0} \epsilon\, \mathbb{E}\left[\int_0^T \left((X_t^U - X_t^{\theta^\star})^\top Q (X_t^U - X_t^{\theta^\star}) + (U_t - U_t^{\theta^\star})^\top R (U_t - U_t^{\theta^\star})\right)\mathrm{d}t\right] \\
&\qquad + 2\mathbb{E}\left[\int_0^T \left((X_t^U - X_t^{\theta^\star})^\top Q X_t^{\theta^\star} + (U_t - U_t^{\theta^\star})^\top R U_t^{\theta^\star}\right)\mathrm{d}t\right] \\
&= 2\mathbb{E}\left[\int_0^T \left((X_t^U - X_t^{\theta^\star})^\top Q X_t^{\theta^\star} + (U_t - U_t^{\theta^\star})^\top R U_t^{\theta^\star}\right)\mathrm{d}t\right],
\end{aligned}
$$

which is based on the fact that $X^U - X^{\theta^\star} \in \mathcal{H}^2(\mathbb{R}^n)$ and $U - U^\star \in \mathcal{H}^2(\mathbb{R}^d)$. As U^θ is the optimal control of $J^{\theta^\star}$, $J^{\theta^\star}(U) \geq J^{\theta^\star}(U^{\theta^\star})$ for all $U \in \mathcal{H}^2(\mathbb{R}^d)$. Hence for all $U \in \mathcal{H}^2(\mathbb{R}^d)$,

$$
\mathbb{E}\left[\int_0^T \left((X_t^U - X_t^{\theta^\star})^\top Q X_t^{\theta^\star} + (U_t - U_t^{\theta^\star})^\top R U_t^{\theta^\star}\right)\mathrm{d}t\right] = \lim_{\epsilon \searrow 0} \frac{1}{2\epsilon}(J^{\theta^\star}(U^\epsilon) - J^{\theta^\star}(U^{\theta^\star})) \geq 0. \tag{3.6}
$$

We now prove that the above quantity is in fact zero for all $U \in \mathcal{H}^2(\mathbb{R}^d)$. To this end, let $U \in \mathcal{H}^2(\mathbb{R}^d)$ be a given (open-loop) control, and consider $\tilde{U} = U^{\theta^\star} - (U - U^{\theta^\star})$. Then by the affineness of (2.2), $X^{\tilde{U}} - X^{\theta^\star}$ satisfies the following controlled dynamics:

$$
\mathrm{d}X_t = (A^\star X_t - B^\star (U - U^{\theta^\star})_t)\,\mathrm{d}t, \quad t \in [0,T]; \quad X_0 = 0. \tag{3.7}
$$

Moreover, one can verify by the affineness of (2.2) that $-(X^U - X^{\theta^\star})$ also satisfies the dynamics (3.7), which along with the uniqueness of solutions to (3.7) shows that $X^{\tilde{U}} - X^{\theta^\star} = -(X^U - X^{\theta^\star})$. Therefore, applying (3.6) with $U = \tilde{U}$ implies that

$$
\begin{aligned}
0 &\leq \mathbb{E}\left[\int_0^T \left((X_t^{\tilde{U}} - X_t^{\theta^\star})^\top Q X_t^{\theta^\star} + (\tilde{U}_t - U_t^{\theta^\star})^\top R U_t^{\theta^\star}\right)\mathrm{d}t\right] \\
&= -\mathbb{E}\left[\int_0^T \left((X_t^U - X_t^{\theta^\star})^\top Q X_t^{\theta^\star} + (U_t - U_t^{\theta^\star})^\top R U_t^{\theta^\star}\right)\mathrm{d}t\right] \leq 0.
\end{aligned}
$$

Hence for all $U \in \mathcal{H}^2(\mathbb{R}^d)$,

$$
\mathbb{E}\left[\int_0^T \left((X_t^U - X_t^{\theta^\star})^\top Q X_t^{\theta^\star} + (U_t - U_t^{\theta^\star})^\top R U_t^{\theta^\star}\right)\mathrm{d}t\right] = 0,
$$

which leads to the desired result (3.5) due to the following identify:

$$\begin{aligned}
J^{\theta^\star}(U) - J^{\theta^\star}(U^{\theta^\star}) &= \mathbb{E}\left[\int_0^T ((X_t^U)^\top Q X_t^U - (X_t^{\theta^\star})^\top Q X_t^{\theta^\star} + U_t^\top R U_t - (U_t^{\theta^\star})^\top R U_t^{\theta^\star})\,\mathrm{d}t\right] \\
&= \mathbb{E}\left[\int_0^T \Big((X_t^U)^\top Q X_t^U - (X_t^{\theta^\star})^\top Q X_t^{\theta^\star} + U_t^\top R U_t - (U_t^{\theta^\star})^\top R U_t^{\theta^\star}\Big)\,\mathrm{d}t\right] \\
&\quad - 2\mathbb{E}\left[\int_0^T \Big((X_t^U - X_t^{\theta^\star})^\top Q X_t^{\theta^\star} + (U_t - U_t^{\theta^\star})^\top R U_t^{\theta^\star}\Big)\,\mathrm{d}t\right] \\
&= \mathbb{E}\left[\int_0^T \Big((X_t^U - X_t^{\theta^\star})^\top Q (X_t^U - X_t^{\theta^\star}) + (U_t - U_t^{\theta^\star})^\top R (U_t - U_t^{\theta^\star})\,\mathrm{d}t\Big)\right].
\end{aligned}$$

∎

Armed with Proposition 3.7, the following proposition quantifies the quadratic performance gap of a greedy policy ψ^θ.

Proposition 3.8 *Suppose (H.1(1)) holds and let Θ be a bounded subset of $\mathbb{R}^{(n+d)\times n}$. For each $\theta \in \Theta$, let ψ^θ be defined in* (2.9)*, let X^{ψ^θ} be the state process associated with ψ^θ (cf.* (2.11)*), let $\psi^{\theta^\star}$ be defined in* (2.3)*, and let $X^{\theta^\star}$ be the state process associated with $\psi^{\theta^\star}$ (cf.* (2.5)*). Then there exists a constant C such that*

$$|J^{\theta^\star}(U^{\psi^\theta}) - J^{\theta^\star}(U^{\theta^\star})| \le C|\theta - \theta^\star|^2, \quad \forall \theta \in \Theta,$$

where $U_t^{\psi^\theta} = \psi^\theta(t, X_t^{\psi^\theta})$ and $U_t^{\theta^\star} = \psi^{\theta^\star}(t, X_t^{\theta^\star})$ for all $t \in [0,T]$, and $J^{\theta^\star}$ is defined in (2.1)*.*

Proof For all $\theta \in \Theta$, applying Proposition 3.7 with $U = U^{\psi^\theta}$ gives

$$\begin{aligned}
&J^{\theta^\star}(U^{\psi^\theta}) - J^{\theta^\star}(U^{\theta^\star}) \\
&\le \|Q\|_2 \|X^{\theta^\star, U^{\psi^\theta}} - X^{\theta^\star}\|^2_{\mathcal{H}^2(\mathbb{R}^n)} + \|R\|_2 \|U^{\psi^\theta} - U^{\theta^\star}\|^2_{\mathcal{H}^2(\mathbb{R}^d)}, \\
&\le \|Q\|_2 \|X^{\psi^\theta} - X^{\psi^{\theta^\star}}\|^2_{\mathcal{H}^2(\mathbb{R}^n)} + \|R\|_2 \|\psi^\theta(\cdot, X_\cdot^{\psi^\theta}) - \psi^{\theta^\star}(\cdot, X_\cdot^{\psi^{\theta^\star}})\|^2_{\mathcal{H}^2(\mathbb{R}^d)},
\end{aligned} \tag{3.8}$$

where the last inequality used the fact that $X^{\theta^\star, U^{\psi^\theta}} = X^{\psi^\theta}$ (see (2.11)), and the definitions of U^{ψ^θ} and $U^{\theta^\star}$. It remains to prove

$$\|X^{\psi^\theta} - X^{\psi^{\theta^\star}}\|_{\mathcal{H}^2(\mathbb{R}^n)} + \|\psi^\theta(\cdot, X_\cdot^{\psi^\theta}) - \psi^{\theta^\star}(\cdot, X_\cdot^{\psi^{\theta^\star}})\|_{\mathcal{H}^2(\mathbb{R}^d)} \le C|\theta - \theta^\star|,$$

for a constant C independent of θ.

Observe that by (2.9), for all $(t,x) \in [0,T]\times\mathbb{R}^n$, $\psi^\theta(t,x) = K_t^\theta x$ with $K_t^\theta = -R^{-1}B^\top P_t^\theta$. Now by Lemma 3.1 and the boundedness of Θ, there exists a constant $C \ge 0$ such that $\|P^\theta\|_{C([0,T;\mathbb{R}^{n\times n})} \le C$ and $\|P^\theta - P^{\theta^\star}\|_{C([0,T;\mathbb{R}^{n\times n})} \le C|\theta - \theta^\star|$ for all $\theta \in \Theta \cup \{\theta^\star\}$, which along with $K_t^\theta = -R^{-1}B^\top P_t^\theta$ implies that $\|K^\theta\|_{C([0,T;\mathbb{R}^{d\times n})} \le C$ and $\|K^\theta - K^{\theta^\star}\|_{C([0,T;\mathbb{R}^{d\times n})} \le C|\theta - \theta^\star|$. Moreover, observe from (2.5) and (2.11) that $X_0^{\theta^\star} = X_0^{\psi^\theta}$ and for all $t \in [0,T]$,

$$\mathrm{d}(X^{\psi^{\theta^\star}} - X^{\psi^\theta})_t = \Big((A^\star + B^\star K_t^{\theta^\star})(X^{\psi^{\theta^\star}} - X^{\psi^\theta})_t + B^\star(K_t^{\theta^\star} - K_t^\theta)X_t^{\psi^\theta}\Big)\,\mathrm{d}t,$$

which combined with the boundedness of $K^{\theta^\star}$ and Gronwall's inequality leads to

$$
\begin{aligned}
\|X^{\psi^{\theta^\star}} - X^{\psi^\theta}\|_{\mathcal{H}^2(\mathbb{R}^n)} &\le C\|X^{\psi^{\theta^\star}} - X^{\psi^\theta}\|_{\mathcal{S}^2(\mathbb{R}^n)} \\
&\le C\|(K^{\theta^\star} - K^\theta)X^{\psi^\theta}\|_{\mathcal{H}^2(\mathbb{R}^d)} \le C\|K^{\theta^\star} - K^\theta\|_{C([0,T;\mathbb{R}^{d\times n})}\|X^{\psi^\theta}\|_{\mathcal{H}^2(\mathbb{R}^n)} \\
&\le C|\theta - \theta^\star|, \quad \forall \theta \in \Theta,
\end{aligned}
$$

where the last inequality follows from $\|X^{\psi^\theta}\|_{\mathcal{H}^2(\mathbb{R}^n)} \le C$, as K^θ is uniformly bounded. The above inequality further implies

$$
\begin{aligned}
&\|\psi^\theta(\cdot, X^{\psi^\theta}_\cdot) - \psi^{\theta^\star}(\cdot, X^{\psi^{\theta^\star}}_\cdot)\|_{\mathcal{H}^2(\mathbb{R}^d)} = \|K^\theta_\cdot X^{\psi^\theta}_\cdot - K^{\theta^\star}_\cdot X^{\psi^{\theta^\star}}_\cdot\|_{\mathcal{H}^2(\mathbb{R}^d)} \\
&\le \|(K^\theta_\cdot - K^{\theta^\star}_\cdot)X^{\psi^\theta}_\cdot\|_{\mathcal{H}^2(\mathbb{R}^d)} + \|K^{\theta^\star}_\cdot(X^{\psi^\theta}_\cdot - X^{\psi^{\theta^\star}}_\cdot)\|_{\mathcal{H}^2(\mathbb{R}^d)} \\
&\le \|K^{\theta^\star} - K^\theta\|_{C([0,T;\mathbb{R}^{d\times n})}\|X^{\psi^\theta}\|_{\mathcal{H}^2(\mathbb{R}^n)} + \|K^{\theta^\star}\|_{C([0,T;\mathbb{R}^{d\times n})}\|X^{\psi^\theta} - X^{\psi^{\theta^\star}}\|_{\mathcal{H}^2(\mathbb{R}^n)} \\
&\le C|\theta - \theta^\star|, \quad \forall \theta \in \Theta,
\end{aligned}
$$

which along with (3.8) finishes the desired estimate. ■

Step 2: Error bound for parameter estimation.

Proposition 3.9 *Suppose (H.1(1)) holds and let $\Theta \subset \mathbb{R}^{(n+d)\times n}$ such that there exists $C_1 > 0$ satisfying $\|(\mathbb{E}[V^{\psi^\theta}])^{-1}\|_2 \le C_1$ and $|\theta| \le C_1$ for all $\theta \in \Theta$, with V^{ψ^θ} defined in (3.1). Then there exist constants $\bar{C}_1, \bar{C}_2 \ge 0$, such that for all $\theta \in \Theta$ and $\delta \in (0, 1/2)$, if $m \ge \bar{C}_1(-\ln\delta)$, then with probability at least $1 - 2\delta$,*

$$
|\hat{\theta} - \theta^\star| \le \bar{C}_2\left(\sqrt{\frac{-\ln\delta}{m}} + \frac{-\ln\delta}{m} + \frac{(-\ln\delta)^2}{m^2}\right), \tag{3.9}
$$

where $\hat{\theta}$ denotes the right-hand side of (2.13) with the control ψ^θ.

Proof Let us fix $\delta \in (0, 1/2)$ and $\theta \in \Theta$. By (2.12) and (2.13), we obtain

$$
\begin{aligned}
\|\hat{\theta} - \theta^\star\|_2 &= \|(V^{\psi^\theta,m} + \tfrac{1}{m}I)^{-1}Y^{\psi^\theta,m} - (\mathbb{E}[V^{\psi^\theta}])^{-1}\mathbb{E}[Y^{\psi^\theta}]\|_2 \\
&\le \|(V^{\psi^\theta,m} + \tfrac{1}{m}I)^{-1} - (\mathbb{E}[V^{\psi^\theta}])^{-1}\|_2\|Y^{\psi^\theta,m}\|_2 + \|(\mathbb{E}[V^{\psi^\theta}])^{-1}\|_2\|Y^{\psi^\theta,m} - \mathbb{E}[Y^{\psi^\theta}]\|_2.
\end{aligned}
$$

As $E^{-1} - F^{-1} = F^{-1}(F - E)E^{-1}$ for all nonsingular matrices E and F, we have

$$
\begin{aligned}
&\|\hat{\theta} - \theta^\star\|_2 \\
&\le \|(V^{\psi^\theta,m} + \tfrac{1}{m}I)^{-1}\|_2\|(\mathbb{E}[V^{\psi^\theta}])^{-1}\|_2\|Y^{\psi^\theta,m}\|_2\|V^{\psi^\theta,m} - \mathbb{E}[V^{\psi^\theta}] + \tfrac{1}{m}I\|_2 \\
&\quad + \|(\mathbb{E}[V^{\psi^\theta}])^{-1}\|_2\|Y^{\psi^\theta,m} - \mathbb{E}[Y^{\psi^\theta}]\|_2 \\
&\le C_1\big(\|(V^{\psi^\theta,m} + \tfrac{1}{m}I)^{-1}\|_2\|Y^{\psi^\theta,m}\|_2\|V^{\psi^\theta,m} - \mathbb{E}[V^{\psi^\theta}] + \tfrac{1}{m}I\|_2 + \|Y^{\psi^\theta,m} - \mathbb{E}[Y^{\psi^\theta}]\|_2\big),
\end{aligned} \tag{3.10}
$$

where the last inequality follows from the assumption $\|(\mathbb{E}[V^{\psi^\theta}])^{-1}\|_2 \le C_1$.

We now estimate each term in the right-hand side of (3.10), and denote by C a generic constant independent of $\theta \in \Theta, \delta \in (0,1/2), m \in \mathbb{N}$. By Theorem 3.6, with probability at least $1-2\delta$, $\|V^{\psi^\theta,m} - \mathbb{E}[V^{\psi^\theta}]\|_2 \le \delta_m$ and $\|Y^{\psi^\theta,m} - \mathbb{E}[Y^{\psi^\theta}]\|_2 \le \delta_m$, with the constant δ_m given by

$$\delta_m \coloneqq C \max\left\{ \left(\frac{-\ln\delta}{m}\right)^{\frac{1}{2}}, \frac{-\ln\delta}{m} \right\}. \tag{3.11}$$

Let m be a sufficiently large constant satisfying $\delta_m + 1/m \le 1/(2C_1)$, where C_1 is the constant such that $\|(\mathbb{E}[V^{\psi^\theta}])^{-1}\|_2 \le C_1$ for all $\theta \in \Theta$. Then with probability at least $1-2\delta$, $\|V^{\psi^\theta,m} - \mathbb{E}[V^{\psi^\theta}] + \frac{1}{m}I\|_2 \le \delta_m + \frac{1}{m} \le \frac{1}{2C_1}$, which in turn yields

$$\lambda_{\min}(V^{\psi^\theta,m} + \tfrac{1}{m}) \ge \lambda_{\min}(\mathbb{E}[V^{\psi^\theta}]) - \|V^{\psi^\theta,m} - \mathbb{E}[V^{\psi^\theta}] + \tfrac{1}{m}I\|_2 \ge \tfrac{1}{2C_1},$$

or equivalently $\|(V^{\psi^\theta,m} + \frac{1}{m})^{-1}\|_2 \le 2C_1$. Moreover, the continuity of $\mathbb{R}^{(n+d)\times n} \ni \theta \mapsto \mathbb{E}[Y^{\psi^\theta}] \in \mathbb{R}$ implies $\|Y^{\psi^\theta,m}\|_2 \le \|\mathbb{E}[Y^{\psi^\theta}]\|_2 + \|Y^{\psi^\theta,m} - \mathbb{E}[Y^{\psi^\theta}]\|_2 \le C + \|Y^{\psi^\theta,m} - \mathbb{E}[Y^{\psi^\theta}]\|_2$. Hence, by (3.10),

$$\begin{aligned}
&|\hat\theta - \theta^\star| \\
&\le C\big((1 + \|Y^{\psi^\theta,m} - \mathbb{E}[Y^{\psi^\theta}]\|_2)\|V^{\psi^\theta,m} - \mathbb{E}[V^{\psi^\theta}] + \tfrac{1}{m}I\|_2 + \|Y^{\psi^\theta,m} - \mathbb{E}[Y^{\psi^\theta}]\|_2\big) \\
&\le C\left((\delta_m + \tfrac{1}{m})(1+\delta_m) + \delta_m\right) \le C\left(\delta_m + \delta_m^2 + \tfrac{1}{m}\right).
\end{aligned}$$

Substituting (3.11) into the above estimate yields the desired estimate (3.22). As $\delta \in (0,1/2)$, it is clear that $\delta_m + 1/m \le 1/(2C_1)$ is satisfied for all $m \ge \bar{C}_1(-\ln\delta)$, with a sufficiently large C_1. ∎

Step 3: Proof of Theorem 2.2. The following proposition shows that for any given $\theta = (A,B)^\top \in \mathbb{R}^{(n+d)\times n}$, the full row rank of K^θ is equivalent to the well-definedness of (2.12) for all θ' sufficiently close to θ.

Proposition 3.10 *Suppose (H.1(1)) holds. For each $\theta \in \mathbb{R}^{(n+d)\times n}$, let V^{ψ^θ} be defined in (3.1). Then for any $\theta = (A,B)^\top \in \mathbb{R}^{(n+d)\times n}$, the following properties are equivalent:*

(1) $\{v \in \mathbb{R}^d \mid (K_t^\theta)^\top v = 0,\ \forall t \in [0,T]\} = \{0\}$, *with K^θ defined in (2.9);*

(2) $\mathbb{E}[V^{\psi^\theta}] \in \mathbb{S}_+^{n+d}$*;*

(3) there exist $\lambda_0, \varepsilon > 0$ such that $\lambda_{\min}(\mathbb{E}[V^{\psi^{\theta'}}]) \ge \lambda_0$ for all $\theta' \in \Phi_\varepsilon \coloneqq \{\theta' \in \mathbb{R}^{(n+d)\times n} \mid |\theta' - \theta| \le \varepsilon\}$, where $\lambda_{\min}(Z)$ is the minimum eigenvalue of $Z \in \mathbb{S}_0^{n+d}$.

Proof For (1) $\implies$ (2): By (3.1), $\mathbb{E}[V^{\psi^\theta}] \in \mathbb{S}_+^{n+d}$ if and only if there exists no nonzero $v \in \mathbb{R}^{n+d}$ such that

$$\mathbb{E}\left[\int_0^T v^\top Z_t^{\psi^\theta}(Z_t^{\psi^\theta})^\top v\,\mathrm{d}t\right] = \int_0^T v^\top \begin{pmatrix} I \\ K_t^\theta \end{pmatrix} \mathbb{E}\left[X_t^{\psi^\theta}(X_t^{\psi^\theta})^\top\right] \begin{pmatrix} I & (K_t^\theta)^\top \end{pmatrix} v\,\mathrm{d}t = 0, \tag{3.12}$$

where we applied Fubini's theorem for the first identity. By (2.5), $X_t^{\psi^\theta} = \Phi_t^\theta\big(x_0+\int_0^t(\Phi_s^\theta)^{-1}\,\mathrm{d}W_s\big)$ for all $t\in[0,T]$, where $\Phi^\theta\in C([0,T];\mathbb{R}^{n\times n})$ is the fundamental solution of $\mathrm{d}\Phi_t^\theta = (A^\star + B^\star K_t^\theta)\Phi_t^\theta\,\mathrm{d}t$, $K_t^\theta = -R^{-1}B^\top P_t^\theta$ for all $t\in[0,T]$, and P^θ satisfies (2.10). Hence,

$$\mathbb{E}\left[X_t^{\psi^\theta}(X_t^{\psi^\theta})^\top\right] = \Phi_t^\theta\left(x_0x_0^\top + \int_0^t(\Phi_s^\theta)^{-1}((\Phi_s^\theta)^{-1})^\top\,\mathrm{d}s\right)(\Phi_t^\theta)^\top\in\mathbb{S}_0^n,\quad \forall t\in[0,T].$$

Then by (3.12) and the continuity of $t\mapsto\mathbb{E}\left[X_t^{\psi^\theta}(X_t^{\psi^\theta})^\top\right]$ and $t\mapsto K_t^\theta$, $\mathbb{E}[V^{\psi^\theta}]\in\mathbb{S}_+^{n+d}$ if and only if there exists no nonzero $v\in\mathbb{R}^{n+d}$ such that

$$v^\top\begin{pmatrix} I\\ K_t^\theta\end{pmatrix}\Phi_t^\theta\left(x_0x_0^\top + \int_0^t(\Phi_s^\theta)^{-1}((\Phi_s^\theta)^{-1})^\top\,\mathrm{d}s\right)(\Phi_t^\theta)^\top\begin{pmatrix} I & (K_t^\theta)^\top\end{pmatrix}v=0,\quad\forall t\in[0,T],$$

where I is the $n\times n$ identity matrix. One can easily deduce by the invertibility of $(\Phi_t^\theta)^{-1}$ for all $t\in[0,T]$ that $\int_0^t(\Phi_s^\theta)^{-1}((\Phi_s^\theta)^{-1})^\top\,\mathrm{d}s\in\mathbb{S}_+^n$ for all $t>0$, which subsequently shows that $\mathbb{E}[V^{\psi^\theta}]\in\mathbb{S}_+^{n+d}$ if and only if there exists no nonzero $\tilde v\in\mathbb{R}^{n+d}$ such that $\begin{pmatrix} I & (K_t^\theta)^\top\end{pmatrix}\tilde v=0$ for all $t\in[0,T]$. Now let us denote without loss of generality that $\tilde v = \left(\begin{smallmatrix} u\\ v\end{smallmatrix}\right)$ for some $u\in\mathbb{R}^n$ and $v\in\mathbb{R}^d$. Then the above derivation shows that $\mathbb{E}[V^{\psi^\theta}]\in\mathbb{S}_+^{n+d}$ is equivalent to the following statement:

$$\text{if } u\in\mathbb{R}^n \text{ and } v\in\mathbb{R}^d \text{ satisfy } u+(K_t^\theta)^\top v=0 \text{ for all } t\in[0,T], \text{ then } u=0 \text{ and } v=0. \tag{3.13}$$

By (2.9), $(K_t^\theta)^\top = -P_t^\theta BR^{-1}$ for all $t\in[0,T]$ and $P_T^\theta=0$, implying that $K_T^\theta=0$. Then (3.13) can be rewritten as:

$$\text{if } v\in\mathbb{R}^d \text{ satisfies } (K_t^\theta)^\top v=0 \text{ for all } t\in[0,T], \text{ then } v=0.$$

For (2) $\iff$ (3): Item (3) clearly implies Item (2). On the other hand, for any given $\theta,\theta'\in\mathbb{R}^{(n+d)\times n}$,

$$\mathrm{d}(X^{\psi^\theta}-X^{\psi^{\theta'}})_t = \left((A^\star+B^\star K_t^\theta)(X^{\psi^\theta}-X^{\psi^{\theta'}})_t + B^\star(K_t^\theta-K_t^{\theta'})X_t^{\psi^{\theta'}}\right)\mathrm{d}t.$$

Then, we can easily deduce from the continuity of $t\mapsto K^\theta$ (see Lemma 3.1) that $\mathbb{R}^{(n+d)\times n}\ni\theta\mapsto Z^{\psi^\theta}\in\mathcal{H}^2(\mathbb{R}^{(n+d)\times n})$ is continuous, which implies the continuity of $\mathbb{R}^{(n+d)\times n}\ni\theta\mapsto V^{\psi^\theta}=\mathbb{E}\big[\int_0^T Z_t^{\psi^\theta}(Z_t^{\psi^\theta})^\top\,\mathrm{d}t\big]\in\mathbb{S}_0^{n+d}$. Hence, by the continuity of the minimum eigenvalue function, we can conclude Item (2) from Item (3). ■

The following proposition provides sufficient conditions for the nondegeneracy of K^θ.

Proposition 3.11 *Let $n,d\in\mathbb{N}$, $\theta=(A,B)^\top\in\mathbb{R}^{(n+d)\times n}$, $Q\in\mathbb{S}_0^n$ and $R\in\mathbb{S}_+^d$.*

(1) For all $T>0$, if $B^\top QB\in\mathbb{S}_+^d$, then $\{v\in\mathbb{R}^d\mid (K_t^\theta)^\top v=0,\ \forall t\in[0,T]\}=\{0\}$.

(2) Assume that the algebraic Riccati equation $A^\top P+PA-P(BR^{-1}B^\top)P+Q=0$ admits a unique maximal solution $P_\infty\in\mathbb{S}_+^n$. Let $K_\infty=-R^{-1}B^\top P_\infty$, and for each $T>0$, let $P^{(T)}\in C([0,T];\mathbb{S}_0^n)$ be defined in (2.10). Assume that $\lim_{T\to\infty}P_0^{(T)}=P_\infty$ and $K_\infty(K_\infty)^\top\in\mathbb{S}_+^d$. Then there exists $T_0>0$, such that for all $T\geq T_0$, $\{v\in\mathbb{R}^d\mid (K_t^\theta)^\top v=0,\ \forall t\in[0,T]\}=\{0\}$.

Proof To prove Item (1), suppose that $B^\top QB \in \mathbb{S}^n_+$ and $v \in \mathbb{R}^d$ such that $(K^\theta_t)^\top v = -P^\theta_t BR^{-1}v = 0$ for all $t \in [0,T]$, with P^θ defined in (2.10). Setting $u = R^{-1}v$, right multiplying (2.10) by Bu, and left multiplying (2.10) by $u^\top B^\top$ shows

$$u^\top B^\top(\tfrac{\mathrm{d}}{\mathrm{d}t}P^\theta_t)Bu + A^\top P^\theta_t Bu + u^\top B^\top P^\theta_t ABu - u^\top B^\top P^\theta_t BR^{-1}B^\top P^\theta_t Bu + u^\top B^\top QBu = 0.$$

As $P^\theta_t Bu = 0$ for all $t \in (0,T)$, $u^\top B^\top(\frac{\mathrm{d}}{\mathrm{d}t}P^\theta_t)Bu = u^\top B^\top P^\theta_t = 0$ for all $t \in (0,T)$, and hence $u^\top B^\top QBu = 0$. The assumption of $B^\top QB \in \mathbb{S}^n_+$ then gives $u = R^{-1}v = 0$, which along with the invertibility of R^{-1} shows that $v = 0$.

To prove Item (2), observe that $\lim_{T\to\infty}(-R^{-1}B^\top P^{(T)}_0) = K_\infty$. As $\lambda_{\min}(K_\infty(K_\infty)^\top) > 0$, there exists $T_0 > 0$ such that for all $T \geq T_0$, $\lambda_{\min}\left((-R^{-1}B^\top P^{(T)}_0)(-R^{-1}B^\top P^{(T)}_0)^\top\right) > 0$. Fix $T \geq T_0$ and consider $v \in \mathbb{R}^d$ such that $(K^\theta_t)^\top v = 0$ for all $t \in [0,T]$. Then the definitions of K^θ and $P^{(T)}$ imply the invertibility of $K^\theta_0(K^\theta_0)^\top$, which yields $v = (K^\theta_0(K^\theta_0)^\top)^{-1}K^\theta_0(K^\theta_0)^\top v = 0$. ■

Now we are ready for the proof of Theorem 2.2.

Proof [Proof of Theorem 2.2] As (H.1(2)) holds with $\theta^\star$ and θ_0, we can obtain from Proposition 3.10 that, there exist $C_1, \varepsilon > 0$ such that for all $\theta \in \Phi_\varepsilon := \{\theta \mid \mathbb{R}^{(n+d)\times n} \mid |\theta - \theta^\star| \leq \varepsilon\} \cup \{\theta_0\}$, we have $\|(\mathbb{E}[V^{\psi^\theta}])^{-1}\|_2 \leq C_1$. Then by Proposition 3.9, there exist constants $\bar{C}_1, \bar{C}_2 \geq 1$, such that for all $\theta \in \Theta_\varepsilon$ and $\delta' \in (0,1/2)$, if $m \geq \bar{C}_1(-\ln\delta')$, then with probability at least $1-2\delta'$,

$$|\hat{\theta} - \theta^\star| \leq \bar{C}_2\left(\sqrt{\frac{-\ln\delta'}{m}} + \frac{-\ln\delta'}{m} + \frac{(-\ln\delta')^2}{m^2}\right), \tag{3.14}$$

where $\hat{\theta}$ denotes the right-hand side of (2.13) with the control ψ^θ. In the following, we fix $\delta \in (0, 3/\pi^2)$ and $C \geq C_0$, with the constant $C_0 \in (0,\infty)$ satisfying

$$C_0 \geq \bar{C}_1\left(\sup_{\ell\in\mathbb{N}\cup\{0\},\delta\in(0,3/\pi^2)} \frac{-\ln(\delta/(\ell+1)^2)}{2^\ell(-\ln\delta)}\right)\Big/\min\left\{\left(\frac{\varepsilon}{3\bar{C}_2}\right)^2, 1\right\},$$

let $m_0 = C(-\ln\delta)$, and for each $\ell \in \mathbb{N}\cup\{0\}$, let $\delta_\ell = \delta/(\ell+1)^2$, $m_\ell = 2^\ell m_0$, and let $\theta_{\ell+1}$ be generated by (2.13) with $m = m_\ell$ and $\theta = \theta_\ell$. Note that the choices of C_0, m_ℓ, δ_ℓ ensure that $m_\ell \geq \bar{C}_1(-\ln\delta_\ell)$, and

$$\bar{C}_2\left(\sqrt{\frac{-\ln\delta_\ell}{m_\ell}} + \frac{(-\ln\delta_\ell)}{m_\ell} + \frac{(-\ln\delta_\ell)^2}{m_\ell^2}\right) \leq 3\bar{C}_2\sqrt{\frac{-\ln\delta_\ell}{m_\ell}} \leq \varepsilon, \quad \forall \ell \in \mathbb{N}\cup\{0\}. \tag{3.15}$$

We now prove with probability at least $1 - 2\sum_{\ell=0}^\infty \delta_\ell = 1 - \frac{\pi^2}{3}\delta$,

$$|\theta_{\ell+1} - \theta^\star| \leq \bar{C}_2\left(\sqrt{\frac{-\ln\delta_\ell}{m_\ell}} + \frac{(-\ln\delta_\ell)}{m_\ell} + \frac{(-\ln\delta_\ell)^2}{m_\ell^2}\right), \quad \forall \ell \in \mathbb{N}\cup\{0\}. \tag{3.16}$$

Let us consider the induction statement for each $k \in \mathbb{N}\cup\{0\}$: with probability at least $1 - 2\sum_{\ell=0}^k \delta_\ell$, (3.16) holds for all $\ell = 0, \ldots, k$. The fact that $\theta_0 \in \Theta_\varepsilon$ and (3.14) yields the induction statement for $k = 0$. Now suppose that the induction statement holds for some $k \in \mathbb{N} \cup \{0\}$. Then the induction hypothesis and (3.15) ensure that $|\theta_\ell - \theta^\star| \leq \varepsilon$ for all $\ell = 1, \ldots, k+1$ (and hence $\theta_{k+1} \in \Theta_\varepsilon$) with probability at least $1 - 2\sum_{\ell=0}^{k} \delta_\ell$. Conditioning on this event, we can apply (3.14) with $\theta = \theta_{k+1}$, $\delta' = \delta_{k+1} < 1/2$ and $m = m_{k+1} \geq \bar{C}_1(-\ln \delta_{k+1})$, and deduce with probability at least $1 - 2\delta_{k+1}$ that (3.16) holds for the index $\ell = k+1$. Combining this with the induction hypothesis yields (3.16) holds for the indices $\ell = 0, \ldots, k+1$, with probability at least $1 - 2\sum_{\ell=0}^{k+1} \delta_\ell$.

Observe that for all $i \in \mathbb{N}$, Algorithm 1 generates the i-th trajectory with control ψ^{θ_ℓ} if $i \in (\sum_{j=0}^{\ell-1} m_j, \sum_{j=0}^{\ell} m_j] = (m_0(2^\ell - 1), m_0(2^{\ell+1} - 1)]$ with some $\ell \in \mathbb{N} \cup \{0\}$. Then conditioning on the event (3.16), we can obtain from Proposition 3.8 that, for all $M \in \mathbb{N}$,

$$
\begin{aligned}
R(M) &\leq \sum_{\ell=0}^{\lceil \log_2(\frac{M}{m_0}+1) \rceil - 1} m_\ell \Big(J^{\theta^\star}(U^{\psi^{\theta_\ell}}) - J^{\theta^\star}(U^{\theta^\star}) \Big) \leq C' \sum_{\ell=0}^{\lceil \log_2(\frac{M}{m_0}+1) \rceil - 1} m_\ell |\theta_\ell - \theta^\star|^2 \\
&\leq C' m_0 + C' \sum_{\ell=0}^{\lceil \log_2(\frac{M}{m_0}+1) \rceil - 1} (-\ln \delta_\ell) \Big(1 + \frac{-\ln \delta_\ell}{m_\ell} + \frac{(-\ln \delta_\ell)^3}{m_\ell^3} \Big) \\
&\leq C'(-\ln \delta) + C' \sum_{\ell=1}^{\lceil \log_2 M \rceil} \Big(2 \ln \ell - \ln \delta \Big) \leq C' \left((\ln M)(\ln \ln M) + (\ln M)(-\ln \delta) \right),
\end{aligned}
\tag{3.17}
$$

with a constant C' independent of M and δ, where we have used $\sum_{\ell=1}^{n} \ln \ell = \ln(n!) \leq C' n \ln n$ due to Stirling's formula. ∎

3.4 Regret Analysis of Discrete-Time Least-Squares Algorithm

This section is devoted to the proof of Theorem 2.3. The main step is similar to the proof of Theorem 2.2 in Section 3.3. However, one needs to quantity the precise impact of the piecewise constant policies and the discrete-time observations on the performance gap and the parameter estimation error.

Step 1: Analysis of the performance gap. The following proposition shows the performance gap between applying a piecewise constant feedback control for an incorrect model and a continuous-time feedback control for the true model scales quadratically with respect to the stepsize and the parameter errors.

Proposition 3.12 *Suppose (H.1(1)) holds and let Θ be a bounded subset of $\mathbb{R}^{(n+d)\times n}$. For each $\theta \in \Theta$ and $N \in \mathbb{N}$, let $\psi^{\theta,\tau}$ be defined in* (2.19) *with stepsize $\tau = T/N$, let $X^{\psi^{\theta,\tau}}$ be the state process associated with $\psi^{\theta,\tau}$ (cf.* (2.21)*), let $\psi^{\theta^\star}$ be defined in* (2.3)*, and let $X^{\theta^\star}$ be the state process associated with $\psi^{\theta^\star}$ (cf.* (2.5)*). Then there exists $C > 0$ such that*

$$
|J^{\theta^\star}(U^{\psi^{\theta,\tau}}) - J^{\theta^\star}(U^{\theta^\star})| \leq C(N^{-2} + |\theta - \theta^\star|^2), \quad \forall \theta \in \Theta, N \in \mathbb{N}, \tag{3.18}
$$

where $U_t^{\psi^{\theta,\tau}} = \psi^{\theta,\tau}(t, X_t^{\psi^{\theta,\tau}})$ and $U_t^{\theta^\star} = \psi^{\theta^\star}(t, X_t^{\theta^\star})$ for all $t \in [0,T]$, and $J^{\theta^\star}$ is defined in (2.1).

Proof Let us fix $\theta \in \Theta$ and $N \in \mathbb{N}$. By applying Proposition 3.7 with $U = U^{\psi^{\theta,\tau}}$,

$$\begin{aligned} & J^{\theta^\star}(U^{\psi^{\theta,\tau}}) - J^{\theta^\star}(U^{\theta^\star}) \\ & \le \|Q\|_2 \|X^{\theta^\star, U^{\psi^{\theta,\tau}}} - X^{\theta^\star}\|^2_{\mathcal{H}^2(\mathbb{R}^n)} + \|R\|_2 \|U^{\psi^{\theta,\tau}} - U^{\theta^\star}\|^2_{\mathcal{H}^2(\mathbb{R}^d)}, \\ & \le \|Q\|_2 \|X^{\psi^{\theta,\tau}} - X^{\psi^{\theta^\star}}\|^2_{\mathcal{H}^2(\mathbb{R}^n)} + \|R\|_2 \|\psi^{\theta,\tau}(\cdot, X^{\psi^{\theta,\tau}}_\cdot) - \psi^{\theta^\star}(\cdot, X^{\psi^{\theta^\star}}_\cdot)\|^2_{\mathcal{H}^2(\mathbb{R}^d)}, \end{aligned} \tag{3.19}$$

where the last inequality used the fact that $X^{\theta^\star, U^{\psi^{\theta,\tau}}} = X^{\psi^{\theta,\tau}}$ (see (2.11)), and the definitions of $U^{\psi^{\theta,\tau}}$ and $U^{\theta^\star}$.

We then prove that there exists a constant C, independent of θ, N, such that

$$\|X^{\psi^{\theta,\tau}} - X^{\psi^{\theta^\star}}\|_{\mathcal{H}^2(\mathbb{R}^n)} + \|\psi^{\theta,\tau}(\cdot, X^{\psi^{\theta,\tau}}_\cdot) - \psi^{\theta^\star}(\cdot, X^{\psi^{\theta^\star}}_\cdot)\|_{\mathcal{H}^2(\mathbb{R}^d)} \le C(N^{-1} + |\theta - \theta^\star|).$$

By setting $\delta X = X^{\theta^\star} - X^{\psi^{\theta,\tau}}$, we obtain from (2.5) and (2.21) that

$$\mathrm{d}\delta X_t = (A^\star \delta X_t + B^\star K^{\theta^\star}_t \delta X_t + (K^{\theta^\star}_t - K^{\theta,\tau}_t) X^{\psi^{\theta,\tau}}_t)\, \mathrm{d}t, \quad t \in [0,T]; \quad \delta X_0 = 0. \tag{3.20}$$

Since $\|P^{\theta^\star}\|_{C([0,T];\mathbb{R}^{n\times n})} \le C$ and $K^{\theta^\star}_t = -R^{-1}B^\top P^{\theta^\star}_t$ for all $t \in [0,T]$, $\|K^{\theta^\star}\|_{C([0,T];\mathbb{R}^{d\times n})} \le C$. Moreover, by $\|P^{\theta,\tau}_{t_i}\|_2 \le C$ for all $i = 0, \ldots, N$ (see Proposition 3.3) and (2.20), we have $\|K^{\theta,\tau}_t\|_2 \le C$ for all $t \in [0,T]$, which along with a moment estimate of (2.21) yields $\|X^{\psi^{\theta,\tau}}\|_{\mathcal{S}^2(\mathbb{R}^n)} \le C$. Thus, by applying Gronwall's inequality to (3.20), Lemma 3.1 and Proposition 3.3, for all $\theta \in \Theta$ and $N \in \mathbb{N}$,

$$\begin{aligned} & \|X^{\theta^\star} - X^{\psi^{\theta,\tau}}\|_{\mathcal{H}^2(\mathbb{R}^n)} \le C\|X^{\theta^\star} - X^{\psi^{\theta,\tau}}\|_{\mathcal{S}^2(\mathbb{R}^n)} \\ & \le C\|(K^{\theta^\star}_t - K^{\theta,\tau}_t) X^{\psi^{\theta,\tau}}\|_{\mathcal{H}^2(\mathbb{R}^d)} \le C \max_{t\in[0,T]} \|K^{\theta^\star}_t - K^{\theta,\tau}_t\|_2 \\ & \le C \max_{t\in[0,T]} (\|K^{\theta}_t - K^{\theta,\tau}_t\|_2 + \|K^{\theta^\star}_t - K^{\theta}_t\|_2) \le C(N^{-1} + |\theta - \theta^\star|). \end{aligned} \tag{3.21}$$

The above inequality further implies

$$\begin{aligned} & \|\psi^{\theta,\tau}(\cdot, X^{\psi^{\theta,\tau}}_\cdot) - \psi^{\theta^\star}(\cdot, X^{\psi^{\theta^\star}}_\cdot)\|_{\mathcal{H}^2(\mathbb{R}^d)} = \|K^{\theta,\tau}_\cdot X^{\psi^{\theta,\tau}}_\cdot - K^{\theta^\star}_\cdot X^{\psi^{\theta^\star}}_\cdot\|_{\mathcal{H}^2(\mathbb{R}^d)} \\ & \le \|(K^{\theta,\tau}_\cdot - K^{\theta^\star}_\cdot) X^{\psi^{\theta,\tau}}_\cdot\|_{\mathcal{H}^2(\mathbb{R}^d)} + \|K^{\theta^\star}_\cdot (X^{\psi^{\theta,\tau}}_\cdot - X^{\psi^{\theta^\star}}_\cdot)\|_{\mathcal{H}^2(\mathbb{R}^d)} \\ & \le \|K^{\theta^\star} - K^{\theta,\tau}\|_{C([0,T;\mathbb{R}^{d\times n})} \|X^{\psi^{\theta,\tau}}\|_{\mathcal{H}^2(\mathbb{R}^n)} + \|K^{\theta^\star}\|_{C([0,T;\mathbb{R}^{d\times n})} \|X^{\psi^{\theta,\tau}} - X^{\psi^{\theta^\star}}\|_{\mathcal{H}^2(\mathbb{R}^n)} \\ & \le C(N^{-1} + |\theta - \theta^\star|), \quad \forall \theta \in \Theta, N \in \mathbb{N}, \end{aligned}$$

which along with (3.19) finishes the desired estimate. ■

Step 2: Error bound for parameter estimation. The following lemma shows that the difference between the expectations of $(V^{\psi^{\theta,\tau},\tau}, Y^{\psi^{\theta,\tau},\tau})$ and of $(V^{\psi^{\theta,\tau}}, Y^{\psi^{\theta,\tau}})$ scales linearly with respect to the stepsize.

Lemma 3.13 *Suppose (H.1(1)) holds and let Θ be a bounded subset of $\mathbb{R}^{(n+d)\times n}$. For each $\theta \in \Theta$ and $N \in \mathbb{N}$, let $\tau = T/N$, let $\psi^{\theta,\tau}$ be defined in* (2.19)*, let $V^{\psi^{\theta,\tau}}, Y^{\psi^{\theta,\tau}}$ be defined in* (3.1)*, and let $V^{\psi^{\theta,\tau},\tau}, Y^{\psi^{\theta,\tau},\tau}$ be defined in* (3.2)*. Then there exists a constant C such that*

$$|\mathbb{E}[V^{\psi^{\theta,\tau},\tau} - V^{\psi^{\theta,\tau}}]| + |\mathbb{E}[Y^{\psi^{\theta,\tau},\tau} - Y^{\psi^{\theta,\tau}}]| \le CN^{-1}, \quad \forall \theta \in \Theta, N \in \mathbb{N}.$$

Proof By (2.21), we have for all $i = 0, \dots, N-1$, $X_{t_{i+1}}^{\psi^{\theta,\tau}} - X_{t_i}^{\psi^{\theta,\tau}} = \int_{t_i}^{t_{i+1}} (\theta^\star)^\top Z_t^{\psi^{\theta,\tau}} \,\mathrm{d}t + W_{t_{i+1}} - W_{t_i}$, which implies

$$\begin{aligned}
\mathbb{E}[V^{\psi^{\theta,\tau}} - V^{\psi^{\theta,\tau},\tau}] &= \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \mathbb{E}[Z_t^{\psi^{\theta,\tau}} (Z_t^{\psi^{\theta,\tau}})^\top - Z_{t_i}^{\psi^{\theta,\tau}} (Z_{t_i}^{\psi^{\theta,\tau}})^\top] \,\mathrm{d}t, \\
\mathbb{E}[Y^{\psi^{\theta,\tau}} - Y^{\psi^{\theta,\tau},\tau}] &= \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \mathbb{E}[(Z_t^{\psi^{\theta,\tau}} - Z_{t_i}^{\psi^{\theta,\tau}})(Z_t^{\psi^{\theta,\tau}})^\top \theta^\star] \,\mathrm{d}t.
\end{aligned}$$

Hence it suffices to prove that $|\mathbb{E}[Z_t^{\psi^{\theta,\tau}} (Z_t^{\psi^{\theta,\tau}})^\top - Z_{t_i}^{\psi^{\theta,\tau}} (Z_{t_i}^{\psi^{\theta,\tau}})^\top]| \le CN^{-1}$ and $|\mathbb{E}[(Z_t^{\psi^{\theta,\tau}} - Z_{t_i}^{\psi^{\theta,\tau}})(Z_t^{\psi^{\theta,\tau}})^\top]| \le CN^{-1}$ for all $t \in [t_i, t_{i+1}]$ and $i = 0, \dots, N-1$.

Let us fix $i = 0, \dots, N-1$ and $t \in [t_i, t_{i+1}]$. In the following, we shall omit the superscripts of $X^{\psi^{\theta,\tau}}$ and $Z^{\psi^{\theta,\tau}}$ if no confusion occurs. As $t \in [t_i, t_{i+1}]$, by (2.21), we have $X_t = e^{Lt} X_{t_i} + \int_{t_i}^t e^{L(t-s)} \,\mathrm{d}W_s$ with $L := A^\star + B^\star K_{t_i}^{\theta,\tau}$. Thus,

$$\begin{aligned}
&X_t X_t^\top - X_{t_i} X_{t_i}^\top \\
&= (X_t - X_{t_i} + X_{t_i})(X_t - X_{t_i} + X_{t_i})^\top - X_{t_i} X_{t_i}^\top \\
&= (X_t - X_{t_i})(X_t - X_{t_i})^\top + X_{t_i}(X_t - X_{t_i})^\top + (X_t - X_{t_i}) X_{t_i}^\top \\
&= \left((e^{Lt} - I) X_{t_i} + \int_{t_i}^t e^{L(t-s)} \,\mathrm{d}W_s \right) \left((e^{Lt} - I) X_{t_i} + \int_{t_i}^t e^{L(t-s)} \,\mathrm{d}W_s \right)^\top \\
&\quad + X_{t_i} \left((e^{Lt} - I) X_{t_i} + \int_{t_i}^t e^{L(t-s)} \,\mathrm{d}W_s \right)^\top + \left((e^{Lt} - I) X_{t_i} + \int_{t_i}^t e^{L(t-s)} \,\mathrm{d}W_s \right) X_{t_i}^\top.
\end{aligned}$$

By taking expectations of both sides of the above identity, the martingale property of the Itô integral, and the Itô isometry,

$$\begin{aligned}
\mathbb{E}[X_t X_t^\top - X_{t_i} X_{t_i}^\top] &= (e^{Lt} - I)\mathbb{E}[X_{t_i} X_{t_i}^\top](e^{L^\top t} - I) + \int_{t_i}^t e^{L(t-s)} e^{L^\top (t-s)} \,\mathrm{d}s \\
&\quad + \mathbb{E}[X_{t_i} X_{t_i}^\top](e^{L^\top t} - I) + (e^{Lt} - I)\mathbb{E}[X_{t_i} X_{t_i}^\top] \le C(t - t_i),
\end{aligned}$$

where the last inequality follows from $\|X^{\psi^{\theta,\tau}}\|_{\mathcal{S}^2(\mathbb{R}^n)} \le C$. Since $\psi^{\theta,\tau}(t, X_t^{\psi^{\theta,\tau}}) = K_{t_i}^{\theta,\tau} X_t^{\psi^{\theta,\tau}}$ and $\|K_{t_i}^{\theta,\tau}\|_2 \le C$, one can easily show that $|\mathbb{E}[Z_t^{\psi^{\theta,\tau}} (Z_t^{\psi^{\theta,\tau}})^\top - Z_{t_i}^{\psi^{\theta,\tau}} (Z_{t_i}^{\psi^{\theta,\tau}})^\top]| \le CN^{-1}$. Furthermore, by $X_t^{\psi^{\theta,\tau}} = e^{Lt} X_{t_i}^{\psi^{\theta,\tau}} + \int_{t_i}^t e^{L(t-s)} \,\mathrm{d}W_s$ and the identity

$$Z_t^{\psi^{\theta,\tau}} (Z_t^{\psi^{\theta,\tau}})^\top - Z_{t_i}^{\psi^{\theta,\tau}} (Z_{t_i}^{\psi^{\theta,\tau}})^\top = (Z_t^{\psi^{\theta,\tau}} - Z_{t_i}^{\psi^{\theta,\tau}})(Z_t^{\psi^{\theta,\tau}})^\top + Z_{t_i}^{\psi^{\theta,\tau}} (Z_t^{\psi^{\theta,\tau}} - Z_{t_i}^{\psi^{\theta,\tau}})^\top,$$

we can show that

$$
\begin{aligned}
&|\mathbb{E}[(Z_t^{\psi^{\theta,\tau}} - Z_{t_i}^{\psi^{\theta,\tau}})(Z_t^{\psi^{\theta,\tau}})^\top]| \\
&\leq |\mathbb{E}[Z_t^{\psi^{\theta,\tau}}(Z_t^{\psi^{\theta,\tau}})^\top - Z_{t_i}^{\psi^{\theta,\tau}}(Z_{t_i}^{\psi^{\theta,\tau}})^\top]| + |\mathbb{E}[Z_{t_i}^{\psi^{\theta,\tau}}(Z_t^{\psi^{\theta,\tau}} - Z_{t_i}^{\psi^{\theta,\tau}})^\top]| \\
&\leq C\left(N^{-1} + \left|\mathbb{E}\left[Z_{t_i}^{\psi^{\theta,\tau}}(X_{t_i}^{\psi^{\theta,\tau}})^\top(e^{L^\top t} - I)\begin{pmatrix} I & (K_{t_i}^{\theta,\tau})^\top \end{pmatrix}\right]\right|\right) \leq CN^{-1},
\end{aligned}
$$

by the uniform boundedness of $\|X^{\psi^{\theta,\tau}}\|_{\mathcal{S}^2(\mathbb{R}^n)}$ and $K^{\theta,\tau}$. ■

Proposition 3.14 *Suppose (H.1(1)) holds, and let $\Theta \subset \mathbb{R}^{(n+d)\times n}$ such that there exists $C_1 > 0$ satisfying $\|(\mathbb{E}[V^{\psi^\theta}])^{-1}\|_2 \leq C_1$ and $|\theta| \leq C_1$ for all $\theta \in \Theta$, with V^{ψ^θ} defined in* (3.1)*. Then there exist constants $\bar{C}_1, \bar{C}_2 \geq 0$ and $n_0 \in \mathbb{N}$, such that for all $\theta \in \Theta$, $N \in \mathbb{N} \cap [n_0, \infty)$ and $\delta \in (0, 1/2)$, if $m \geq \bar{C}_1(-\ln\delta)$, then with probability at least $1 - 2\delta$,*

$$
|\hat{\theta} - \theta^\star| \leq \bar{C}_2\left(\sqrt{\frac{-\ln\delta}{m}} + \frac{-\ln\delta}{m} + \frac{(-\ln\delta)^2}{m^2} + \frac{1}{N}\right), \tag{3.22}
$$

where $\hat{\theta}$ denotes the right-hand side of (2.23) *with the control $\psi^{\theta,\tau}$ and stepsize $\tau = T/N$.*

Proof We first prove that there exists $n_0 \in \mathbb{N}$ such that for all $N \in \mathbb{N} \cap [n_0, \infty)$ and $\theta \in \Theta$, $\|(\mathbb{E}[V^{\psi^{\theta,\tau}}])^{-1}\|_2 \leq C$ for a constant $C > 0$ independent of θ and N. By (2.11) and (2.21), we have for all $\theta \in \Theta$ and $N \in \mathbb{N}$, $X_0^{\psi^\theta} = X_0^{\psi^{\theta,\tau}}$ and

$$
\mathrm{d}(X^{\psi^\theta} - X^{\psi^{\theta,\tau}})_t = \left((A^\star + B^\star K_t^\theta)(X^{\psi^\theta} - X^{\psi^{\theta,\tau}})_t + B^\star(K_t^\theta - K_t^{\theta,\tau})X_t^{\psi^{\theta,\tau}}\right)\mathrm{d}t, \quad t \in [0, T].
$$

Proposition 3.3 shows $\|K_t^\theta - K_t^{\theta,\tau}\|_2 \leq CN^{-1}$ for all $t \in [0, T]$, which along with Gronwall's inequality yields $\|X^{\psi^\theta} - X^{\psi^{\theta,\tau}}\|_{\mathcal{S}^2(\mathbb{R}^n)} \leq CN^{-1}$ for all $\theta \in \Theta$ and $N \in \mathbb{N}$. One can further prove that $\|U^{\psi^\theta} - U^{\psi^{\theta,\tau}}\|_{\mathcal{S}^2(\mathbb{R}^d)} \leq CN^{-1}$ with $U_t^{\psi^\theta} = \psi^\theta(t, X_t^{\psi^\theta})$ and $U_t^{\psi^{\theta,\tau}} = \psi^{\theta,\tau}(t, X_t^{\psi^{\theta,\tau}})$ for all $t \in [0, T]$. Thus, we have $|\mathbb{E}[V^{\psi^\theta}] - \mathbb{E}[V^{\psi^{\theta,\tau}}]| \leq CN^{-1}$, which along with $\|(\mathbb{E}[V^{\psi^\theta}])^{-1}\|_2 \leq C_1$ implies a uniform bound of $\|(\mathbb{E}[V^{\psi^{\theta,\tau}}])^{-1}\|_2$ for all sufficiently large N.

Let us fix $N \in \mathbb{N} \cap [n_0, \infty)$ and $\theta \in \Theta$ for the subsequent analysis. The invertibility of $\mathbb{E}[V^{\psi^{\theta,\tau}}]$ implies that $\theta^\star = (\mathbb{E}[V^{\psi^{\theta,\tau}}])^{-1}\mathbb{E}[Y^{\psi^{\theta,\tau}}]$)(cf. (2.12)). Then by (2.23), we can derive the following analogues of (3.10):

$$
\begin{aligned}
&\|\hat{\theta} - \theta^\star\|_2 \\
&= \|(V^{\psi^{\theta,\tau},\tau,m} + \tfrac{1}{m}I)^{-1}Y^{\psi^{\theta,\tau},\tau,m} - (\mathbb{E}[V^{\psi^{\theta,\tau}}])^{-1}\mathbb{E}[Y^{\psi^{\theta,\tau}}]\|_2 \\
&\leq \|(V^{\psi^{\theta,\tau},\tau,m} + \tfrac{1}{m}I)^{-1} - (\mathbb{E}[V^{\psi^{\theta,\tau}}])^{-1}\|_2\|Y^{\psi^{\theta,\tau},m,\tau}\|_2 + \|(\mathbb{E}[V^{\psi^{\theta,\tau}}])^{-1}\|_2\|Y^{\psi^{\theta,\tau},\tau,m} - \mathbb{E}[Y^{\psi^{\theta,\tau}}]\|_2 \\
&\leq \|(V^{\psi^{\theta,\tau},\tau,m} + \tfrac{1}{m}I)^{-1}\|\|(\mathbb{E}[V^{\psi^{\theta,\tau}}])^{-1}\|_2\|Y^{\psi^{\theta,\tau},m,\tau}\|_2\|V^{\psi^{\theta,\tau},\tau,m} - \mathbb{E}[V^{\psi^{\theta,\tau}}] + \tfrac{1}{m}I\|_2 \\
&\quad + \|(\mathbb{E}[V^{\psi^{\theta,\tau}}])^{-1}\|_2\|Y^{\psi^{\theta,\tau},\tau,m} - \mathbb{E}[Y^{\psi^{\theta,\tau}}]\|_2,
\end{aligned}
$$

where $V^{\psi^{\theta,\tau},\tau,m}$ and $Y^{\psi^{\theta,\tau},\tau,m}$ are defined in (3.2). Note that

$$\|V^{\psi^{\theta,\tau},\tau,m} - \mathbb{E}[V^{\psi^{\theta,\tau}}] + \tfrac{1}{m}I\|_2 \le \|V^{\psi^{\theta,\tau},\tau,m} - \mathbb{E}[V^{\psi^{\theta,\tau},\tau}]\|_2 + \|\mathbb{E}[V^{\psi^{\theta,\tau},\tau}] - \mathbb{E}[V^{\psi^{\theta,\tau}}]\|_2 + \tfrac{1}{m},$$
$$\|Y^{\psi^{\theta,\tau},\tau,m} - \mathbb{E}[Y^{\psi^{\theta,\tau}}]\|_2 \le \|Y^{\psi^{\theta,\tau},\tau,m} - \mathbb{E}[Y^{\psi^{\theta,\tau},\tau}]\|_2 + \|\mathbb{E}[Y^{\psi^{\theta,\tau},\tau}] - \mathbb{E}[Y^{\psi^{\theta,\tau}}]\|_2,$$

where for both inequalities, the first term on the right-hand side can be estimated by Theorem 3.6 (uniformly in N), and the second term is of the magnitude $\mathcal{O}(N^{-1})$ due to Lemma 3.13. Hence, proceeding along the lines of the proof of Proposition 3.9 leads to the desired result. ■

Step 3: Proof of Theorem 2.3. The proof follows from similar arguments as that of Theorem 2.2, and we only present the main steps here. As (H.1(2)) holds with θ_0 and $\theta^\star$, we can obtain from Propositions 3.10 and 3.14 that, there exists a bounded set $\Phi_\varepsilon \subset \mathbb{R}^{(n+d)\times n}$ and constants $\bar{C}_1, \bar{C}_2 \ge 1$, $n_0 \in \mathbb{N}$ that for all $\theta \in \Phi_\varepsilon$, $N \in \mathbb{N} \cap [n_0, \infty)$ and $\delta' \in (0, 1/2)$, if $m \ge \bar{C}_1(-\ln\delta)$, then with probability at least $1 - 2\delta'$,

$$|\hat{\theta} - \theta^\star| \le \bar{C}_2\left(\sqrt{\frac{-\ln\delta'}{m}} + \frac{-\ln\delta'}{m} + \frac{(-\ln\delta')^2}{m^2} + \frac{1}{N}\right), \tag{3.23}$$

where $\hat{\theta}$ denotes the right-hand side of (2.23) with the control $\psi^{\theta,\tau}$ and stepsize $\tau = T/N$. Then by proceeding along the lines of the proof of Theorem 2.2, there exists $C_0 > 0$ and $n_0 \in \mathbb{N}$, such that for any given $\delta \in (0, \frac{3}{\pi^2})$, if $m_0 = C(-\ln\delta)$ with $C \ge C_0$ and $N_\ell \ge n_0$ for all $\ell \in \mathbb{N} \cup \{0\}$, then with probability at least $1 - \frac{\pi^2}{3}\delta$,

$$|\theta_{\ell+1} - \theta^\star| \le \bar{C}_2\left(\sqrt{\frac{-\ln\delta_\ell}{m_\ell}} + \frac{(-\ln\delta_\ell)}{m_\ell} + \frac{(-\ln\delta_\ell)^2}{m_\ell^2} + \frac{1}{N_\ell}\right), \quad \forall \ell \in \mathbb{N} \cup \{0\}, \tag{3.24}$$

where $\delta_\ell = \delta/(\ell+1)^2$ and $m_\ell = 2^\ell m_0$ for all ℓ. Consequently, we can conclude the desired regret bound from Proposition 3.12 (cf. (3.17)), with an additional term $\sum_{\ell=0}^{\ln M} m_\ell N_\ell^{-2}$ due to the time discretization errors in (3.18) and (3.24).

References

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

Marc Abeille and Alessandro Lazaric. Improved regret bounds for thompson sampling in linear quadratic control problems. In *International Conference on Machine Learning*, pages 1–9. PMLR, 2018.

Naman Agarwal, Elad Hazan, and Karan Singh. Logarithmic regret for online control. *Advances in Neural Information Processing Systems*, 32, 2019.

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 49–56, 2007.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 89–96, 2009.

Rafael Bailo, Mattia Bongini, José A Carrillo, and Dante Kalise. Optimal consensus control of the Cucker-Smale model. *IFAC-PapersOnLine*, 51(13):1–6, 2018.

Robert R Bitmead and Michel Gevers. Riccati difference and differential equations: Convergence, monotonicity and stability. In *The Riccati Equation*, pages 263–291. Springer, 1991.

Marco C Campi and PR Kumar. Adaptive linear quadratic gaussian control: the cost-biased approach revisited. *SIAM Journal on Control and Optimization*, 36(6):1890–1907, 1998.

Alvaro Cartea, Sebastian Jaimungal, and Jason Ricci. Algorithmic trading, stochastic control, and mutually exciting processes. *SIAM Review*, 60(3):673–703, 2018.

Asaf Cassel, Alon Cohen, and Tomer Koren. Logarithmic regret for learning linear quadratic regulators efficiently. In *International Conference on Machine Learning*, pages 1328–1337. PMLR, 2020.

Xinyi Chen and Elad Hazan. Black-box control for linear dynamical systems. In *Conference on Learning Theory*, pages 1114–1143. PMLR, 2021.

Patrick Cheridito, H. Mete Soner, and Nizar Touzi. Small time path behavior of double stochastic integrals and applications to stochastic control. *The Annals of Applied Probability*, 15(4):2472–2495, 2005.

Philippe G Ciarlet. *Linear and nonlinear functional analysis with applications*, volume 130. Siam, 2013.

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear quadratic regulators efficiently with only $\sqrt{T}$ regret. *arXiv preprint arXiv:1902.06223*, 2019.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pages 4188–4197, 2018.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, pages 1–47, 2019.

Tyrone E Duncan, Petr Mandl, and Bożenna Pasik-Duncan. On least squares estimation in continuous time linear stochastic systems. *Kybernetika*, 28(3):169–180, 1992.

Tyrone E Duncan, Lei Guo, and Bozenna Pasik-Duncan. Adaptive continuous-time linear quadratic Gaussian control. *IEEE Transactions on Automatic Control*, 44(9):1653–1662, 1999.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite-time adaptive stabilization of linear systems. *IEEE Transactions on Automatic Control*, 64 (8):3498–3505, 2018a.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. On adaptive linear-quadratic regulators. *arXiv e-prints*, pages arXiv–1806, 2018b.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Randomized algorithms for data-driven stabilization of stochastic linear systems. In *2019 IEEE Data Science Workshop (DSW)*, pages 170–174. IEEE, 2019.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Input perturbations for adaptive control and learning. *Automatica*, 117:108950, 2020.

Dylan Foster and Max Simchowitz. Logarithmic regret for adversarial online control. In *International Conference on Machine Learning*, pages 3211–3221. PMLR, 2020.

Graham C Goodwin, Peter J Ramadge, and Peter E Caines. Discrete time stochastic adaptive control. *SIAM Journal on Control and Optimization*, 19(6):829–853, 1981.

P Jameson Graber. Linear quadratic mean field type control and mean field games with common noise, with application to production of an exhaustible resource. *Applied Mathematics & Optimization*, 74(3):459–486, 2016.

Michael Green and John B Moore. Persistence of excitation in linear systems. *Systems & control letters*, 7(5):351–360, 1986.

Ben M Hambly, Renyuan Xu, and Huining Yang. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *Available at SSRN*, 2020.

Petros Ioannou and Bariş Fidan. *Adaptive control tutorial*. SIAM, 2006.

PR Kumar. Optimal adaptive control of linear-quadratic-gaussian systems. *SIAM Journal on Control and Optimization*, 21(2):163–178, 1983.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Explore more and improve regret in linear quadratic regulators. *arXiv preprint arXiv:2007.12291*, 2020a.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Logarithmic regret bound in partially observable linear dynamical systems. *Advances in Neural Information Processing Systems*, 33:20876–20888, 2020b.

Ioan Doré Landau, Rogelio Lozano, Mohammed M'Saad, and Alireza Karimi. *Adaptive control: algorithms, analysis and applications*. Springer Science & Business Media, 2011.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalent control of LQR is efficient. *arXiv preprint arXiv:1902.07826*, 2019.

Xuerong Mao. *Stochastic differential equations and applications*. Elsevier, 2007.

Hamidreza Modares and Frank L. Lewis. Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning. *IEEE Transactions on Automatic Control*, 59(11):3051–3056, 2014.

Rémi Munos. A study of reinforcement learning in the continuous case by the means of viscosity solutions. *Machine Learning*, 40(3):265–299, 2000.

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

Rémi Munos and Paul Bourgine. Reinforcement learning for continuous stochastic control problems. In *Advances in Neural Information Processing Systems*, pages 1029–1035, 1998.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (More) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

Yi Ouyang, Mukul Gagrani, and Rahul Jain. Learning-based control of unknown linear systems with thompson sampling. *arXiv preprint arXiv:1709.04047*, 2017.

Syed Ali Asad Rizvi and Zongli Lin. Output feedback reinforcement learning control for the continuous-time linear quadratic regulator problem. In *2018 Annual American Control Conference (ACC)*, pages 3417–3422. IEEE, 2018.

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

Corentin Tallec, Léonard Blier, and Yann Ollivier. Making Deep Q-learning methods robust to time discretization. *arXiv preprint arXiv:1901.09732*, 2019.

Mark Veraar. The stochastic Fubini theorem revisited. *Stochastics An International Journal of Probability and Stochastic Processes*, 84(4):543–551, 2012.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Haoran Wang and Xun Yu Zhou. Continuous-time mean–variance portfolio selection: A reinforcement learning framework. *Mathematical Finance*, 30(4):1273–1308, 2020.

Jiongmin Yong and Xun Yu Zhou. *Stochastic Controls: Hamiltonian Systems and HJB Equations*, volume 43. Springer Science & Business Media, 1999.