# Online Reinforcement Learning in Stochastic Continuous-Time Systems

**Mohamad Kazem Shirani Faradonbeh** MOHAMADKSF@SMU.EDU
*Department of Mathematics, Southern Methodist University, Dallas, TX.*

**Mohamad Sadegh Shirani Faradonbeh** SSHIRANI@STANFORD.EDU
*Graduate School of Business, Stanford University, Stanford, CA.*

**Editors:** Gergely Neu and Lorenzo Rosasco

## Abstract

Linear dynamical systems that obey stochastic differential equations are canonical models. While optimal control of known systems has a rich literature, the problem is technically hard under model uncertainty and there are hardly any such result. We initiate study of this problem and aim to learn (and *simultaneously* deploy) optimal actions for minimizing a quadratic cost function. Indeed, this work is the first that comprehensively addresses the crucial challenge of balancing exploration versus exploitation in continuous-time systems. We present online policies that learn optimal actions fast by carefully randomizing the parameter estimates, and establish their performance guarantees: a regret bound that grows with *square-root of time* multiplied by the *number of parameters*. Implementation of the policy for a flight-control task demonstrates its efficacy. Further, we prove sharp stability results for inexact system dynamics and tightly specify the infinitesimal regret caused by sub-optimal actions. To obtain the results, we conduct a novel eigenvalue-sensitivity analysis for matrix perturbation, establish upper-bounds for comparative ratios of stochastic integrals, and introduce the new method of policy differentiation. Our analysis sheds light on fundamental challenges in continuous-time reinforcement learning and suggests a useful cornerstone for similar problems.

**Keywords:** Ito Processes; Online Policies; Regret Bounds; Stability Analysis; Exploration vs Exploitation; Randomized Estimates.

## 1. Introduction

State-space models are widely-used for decision-making in dynamic environments. A popular such model is the one that represents the continuous-time dynamics of the environment by linear stochastic differential equations. In this setting, the multidimensional state of the system is driven by the control action and the Brownian noise, according to an Ito stochastic differential equation. The range of application areas is extensive, including chemistry, biology, finance, insurance, and engineering (Gillespie, 2007; Schmidli, 2007; Pham, 2009; Lawrence et al., 2010).

In many applications, uncertainties about the true dynamics necessitate reinforcement learning policies that adaptively learn optimal actions. Unlike the continuous-time setting, reinforcement learning policies are extensively studied in discrete-time systems. The literature is rich and includes efficient algorithms that use optimism in the face of uncertainty, posterior sampling, or bootstrap (Abbasi-Yadkori and Szepesvári, 2011; Abeille and Lazaric, 2018; Ouyang et al., 2019; Faradonbeh et al., 2019a, 2020c; Dean et al., 2020; Faradonbeh et al., 2020a,b), and regret bounds are shown in the presence of domain knowledge and partial observations (Cassel et al., 2020; Ziemann and Sandberg, 2020a,b; Asghari et al., 2020; Lale et al., 2020b).

On the other hand, the existing literature for continuous-time systems is still immature, mainly due to technical difficulties that will be discussed shortly. Early papers focus on estimation after an infinitely long interaction with the environment (Mandl et al., 1988; Mandl, 1989; Duncan and Pasik-Duncan, 1990; Duncan et al., 1992, 1999). Recently, sub-optimal policies with linear regret bounds are proposed and consistency is shown for systems with full-rank input matrices (Caines and Levanony, 2019). Ensuing papers study offline algorithms for computing the control actions according to a batch of data, using methods such as dynamic programming and entropy regularization (Doya, 2000; Rizvi and Lin, 2018; Wang et al., 2020; Basei et al., 2021).

However, *online* reinforcement learning policies that can learn optimal actions from a *single* state trajectory without imposing undue costs, are currently unavailable. The existing results are merely asymptotic, require restrictive assumptions, and provide linear regret bounds (Duncan et al., 1999; Caines and Levanony, 2019). The only exception is a recent paper that appeared after the first version of this work (Shirani Faradonbeh et al., 2022). A fundamental challenge of online policies is that they need to *simultaneously* minimize the cost and estimate the unknown parameters. These two goals contradict and constitute the exploration-exploitation dilemma; accurate estimation is necessary for optimal decision-making, while sub-optimal actions are required for obtaining accurate estimates. This crux remains unsolved in continuous-time systems as conventional frameworks are incapable of relating exploring actions to estimation accuracy and optimal policies. In fact, the discrete-time analysis fails in Ito processes that the evolution is infinitesimal, and the signal is highly dominated by noise.

The main contributions of this paper can be summarized as follows. In Algorithm 1, we propose an efficient online reinforcement learning policy based on randomized estimates of the unknown system matrices. Then, we establish the rates for the error in learning the dynamics matrices, and for the regret that the algorithm incurs. Algorithm 1 is easy to implement, yet it learns the optimal actions *fast* so that its regret at time $T$ is $\mathcal{O}\left(T^{1/2}\log T\right)$ (Theorem 4). So, the per-unit-time sub-optimality caused by uncertain system parameters decays with time as $\widetilde{\mathcal{O}}\left(t^{-1/2}\right)$ under Algorithm 1, which is the first efficiency result for online policies. Furthermore, we study stability of linear systems for inaccurate system matrices and establish the stabilizability margin (Theorem 2). Finally, a sharp regret expression is provided that fully captures sub-optimalities due to inaccuracies in approximating the optimal actions (Theorem 3). The results provide both the average-case and worst-case analyses, the presented bounds are tight, and the technical assumptions are minimal.

To study online reinforcement learning policies, one needs to address the following technical challenges. First, sensitivity analysis of (complex) eigenvalues of matrices with perturbed entries is needed. Further, anti-concentration results on singular values of partially-random matrices are required. Finally, we need to accurately characterize the time-varying sub-optimalities in cost function in terms of model uncertainties. Thus, we develop multiple novel techniques for (i) *matrix-perturbation* analysis, (ii) spectral properties of *random matrices*, (iii) comparative ratios of *stochastic integrals*, and also (iv) introduce *policy differentiation* to precisely capture the infinitesimal cost of sub-optimal actions. Note that (iii) and (iv) above do *not* appear in discrete-time settings. The former two are different in differential and difference equations, such that the existing literature fails to properly address them in continuous-time systems. We also use various tools from Ito calculus and stochastic analysis, including Hamilton-Jacobi-Bellman partial differential equations, Ito Isometry, as well as (dominated and martingale) convergence theorems for continuous-time stochastic processes (Yong and Zhou, 1999; Oksendal, 2013; Baldi, 2017).

This paper is organized as follows. In Section 2, we discuss the problem and preliminary material. Then, in Section 3, we study system stability when the control action is designed based on dynamics matrices *other than* the true ones and establish stabilizability guarantees. Next, we examine effects of sub-optimal actions and the regret they cause, in Section 4. Section 5 contains the randomized-estimates policy of Algorithm 1, as well as its theoretical analysis showing efficiency. Experimental results are presented in Section 6, and the paper is concluded in Section 7. Because of space limitations, all proofs are delegated to the appendices, as outlined on page 17.

The following notation will be used in this work. The smallest (largest) eigenvalue of $A$, in magnitude, is $\boldsymbol{\lambda}_{\min}(A)$ ($\boldsymbol{\lambda}_{\max}(A)$). For $v \in \mathbb{C}^d$, its norm is defined as $\|v\|^2 = \sum_{i=1}^{d} |v_i|^2$. Moreover, we write $|||A|||$ for the operator norm of matrices; $|||A||| = \sup_{\|v\|=1} \|Av\|$, and $A^\dagger$ for Moore-Penrose generalized inverse. The sigma-field generated by the stochastic process $\{Y_s\}_{0\leq s\leq t}$ is denoted by $\sigma(Y_{0:t})$. A multivariate normal distribution with mean $\mu$ and covariance matrix $\Sigma$ is shown by $\mathcal{N}(\mu, \Sigma)$. For $\lambda \in \mathbb{C}$, we use $\Re(\lambda), \Im(\lambda)$ to denote the real and imaginary parts of $\lambda$, respectively. The symbol $\vee$ (resp., $\wedge$) is used to show the maximum (resp., minimum). Finally, $\mathcal{O}(\cdot)$ refers to the order of magnitude.

## 2. Problem Statement

We study reinforcement learning policies for a multidimensional Ito stochastic differential equation with unknown drift matrices. That is, the state vector at time $t$ is $X_t \in \mathbb{R}^{d_X}$, which follows

$$\mathrm{d}X_t = (A_\star X_t + B_\star U_t)\,\mathrm{d}t + C\mathrm{d}W_t, \tag{1}$$

the vector $U_t \in \mathbb{R}^{d_U}$ is the control action at time $t$, and the disturbance $\{W_t\}_{t\geq 0}$ is a standard Brownian motion in a $d_W$ dimensional space. Technically, by fixing the probability space $(\Omega, \mathbb{F}, \mathbb{P})$ which is completed by adding the null-sets of $\mathbb{P}$, let all stochastic objects belong to this probability space, and let $\mathbb{E}[\cdot]$ be the expectation with respect to $\mathbb{P}$ (unless otherwise explicitly stated). The Brownian motion $\{W_t\}_{t\geq 0}$ starts from the origin, and has continuous sample paths as well as independent normally distributed increments. That is, $W_0 = 0$, for all $0 \leq t_1 \leq t_2 \leq t_3 \leq t_4$, the vectors $W_{t_2} - W_{t_1}$ and $W_{t_4} - W_{t_3}$ are statistically independent, and for all non-negative reals $s < t$, it holds that $W_t - W_s \sim \mathcal{N}(0, (t-s) I_{d_W})$. Furthermore, $C \in \mathbb{R}^{d_X \times d_W}$ reflects the effect of $W_t$ on the state evolution.

We aim to design computationally tractable and provably efficient reinforcement learning policies for the system in (1). The transition matrix $A_\star$, the input matrix $B_\star$, and the noise-coefficient matrix $C$, *all are unknown*. The goal is to minimize the expected average cost

$$\mathcal{J}_{\boldsymbol{\pi}} = \limsup_{T\to\infty} \frac{1}{T}\,\mathbb{E}\left[\int\limits_0^T c_t(\boldsymbol{\pi})\mathrm{d}t\right],$$

where $c_t(\boldsymbol{\pi}) = X_t^\top Q X_t + U_t^\top R U_t$ is the cost of policy $\boldsymbol{\pi}$ at $t$, its value being determined by the positive definite matrices $Q, R$ of proper dimensions, as explained below.

The policy $\boldsymbol{\pi}$ is non-anticipative closed-loop: At every time $t$, $\boldsymbol{\pi}$ determines $U_t$ according to the information available at the time. More precisely, $\boldsymbol{\pi}$ maps the state observations (i.e., $X_s$ for $s \in [0, t]$) and the previously taken actions (i.e., $U_s$ for $s$ in the semi-open interval $[0, t[$) to the current control action $U_t$. This mapping can be stochastic or deterministic. Importantly, $\boldsymbol{\pi}$ faces the fundamental exploration-exploitation dilemma for minimizing the expected average cost, because the dynamics matrices $A_\star, B_\star$ are unknown and need to be learned based on the state and action observations. The details of this dilemma is discussed in Section 5. We assume that $Q, R$ are known to the policy, the rationale being that the decision-makers are aware of the objective they aim to achieve, while their uncertainty about the environment impedes them from deciding optimally.

The benchmark for assessing reinforcement learning policies is the optimal policy $\boldsymbol{\pi}^\star$ that designs $U_t$ *having access* to $A_\star, B_\star$. To define $\boldsymbol{\pi}^\star$, let $\Phi_{A_\star,B_\star}(\cdot) : \mathbb{R}^{d_X \times d_X} \to \mathbb{R}^{d_X \times d_X}$ be

$$\Phi_{A_\star,B_\star}(M) = A_\star^\top M + M A_\star - M B_\star R^{-1} B_\star^\top M + Q.$$

This function is vital for finding $\boldsymbol{\pi}^\star$. To see the intuition, first note that an action $U_t$ directly influences the current cost $c_t(\boldsymbol{\pi})$, and indirectly affects the future cost values according to (1). So, to capture future consequences of decisions, $\Phi_{A_\star,B_\star}(\cdot)$ is employed (Yong and Zhou, 1999). To proceed toward identifying $\boldsymbol{\pi}^\star$, let the positive semidefinite matrix $M = \mathcal{K}(A_\star, B_\star)$ solve $\Phi_{A_\star,B_\star}(M) = 0$. To investigate existence and uniqueness of $\mathcal{K}(A_\star, B_\star)$, we need the followings.

**Definition 1 (Notations $\overline{\boldsymbol{\lambda}}(\cdot), \mathcal{E}(\cdot)$)** *Let $\overline{\boldsymbol{\lambda}}(D)$ be the largest real-part of the eigenvalues of an arbitrary square matrix $D$: $\overline{\boldsymbol{\lambda}}(D) = \max\{\Re(\lambda) : \det(D - \lambda I) = 0\}$. Further, for arbitrary matrices $A \in \mathbb{R}^{d_X \times d_X}$, $B \in \mathbb{R}^{d_X \times d_U}$, define $\mathcal{E}(A, B) = \|A - A_\star\| + \|B - B_\star\|$. So, $\mathcal{E}(A, B)$ measures the deviation of $A, B$ from the true dynamics matrices $A_\star, B_\star$.*

Note that unlike $\boldsymbol{\lambda}_{\max}(\cdot)$ that considers only *magnitudes* of the eigenvalues, $\overline{\boldsymbol{\lambda}}(\cdot)$ reflects the signs of the eigenvalues as well, and so can be either positive, zero, or negative. However, they are related according to $\boldsymbol{\lambda}_{\max}\left(e^D\right) = e^{\overline{\boldsymbol{\lambda}}(D)}$.

We assume that the true dynamics matrices $A_\star, B_\star$ are stabilizable, in the following sense:

**Assumption 1 (Stabilizability)** *There exists some $L \in \mathbb{R}^{d_U \times d_X}$, such that $\overline{\boldsymbol{\lambda}}(A_\star + B_\star L) < 0$.*

Assumption 1 expresses that by applying $U_t = L X_t$, the system can operate without any explosion. To see that, solve the differential equation (1) under the feedback policy $U_t = L X_t$ to obtain

$$X_t = e^{(A_\star + B_\star L)t} X_0 + \int_0^t e^{(A_\star + B_\star L)(t-s)} C \mathrm{d}W_s. \tag{2}$$

So, because of $\overline{\boldsymbol{\lambda}}(A_\star + B_\star L) < 0$, $X_t$ does not grow unbounded with $t$. Importantly, existence of a stabilizing matrix $L$ is *necessary* for the problem to be well-defined. Otherwise, state explosion renders the average cost infinite for *all* decision-making policies (Chen et al., 1995; Yong and Zhou, 1999). Now, recall that $\Phi_{A_\star,B_\star}(\mathcal{K}(A_\star, B_\star)) = 0$, let $\mathcal{L}(A_\star, B_\star) = -R^{-1} B_\star^\top \mathcal{K}(A_\star, B_\star)$, and define the linear feedback policy

$$\boldsymbol{\pi}^\star : \qquad U_t = \mathcal{L}(A_\star, B_\star) X_t, \qquad \forall t \geq 0. \tag{3}$$

We show that Assumption 1 suffices for unique existence of $\mathcal{K}(A_\star, B_\star)$ and for optimality of $\boldsymbol{\pi}^\star$.

**Theorem 1 (Optimal policy)** *Under Assumption 1, the matrix $\mathcal{K}(A_\star, B_\star)$ uniquely exists, and the policy $\boldsymbol{\pi}^\star$ in (3) gives*

$$\mathcal{J}_{\boldsymbol{\pi}^\star} = \inf_{\boldsymbol{\pi}} \mathcal{J}_{\boldsymbol{\pi}} = \mathbf{tr}\left(\mathcal{K}(A_\star, B_\star) C C^\top\right), \qquad \textit{and} \qquad \overline{\boldsymbol{\lambda}}(A_\star + B_\star \mathcal{L}(A_\star, B_\star)) < 0.$$

To compute $\mathcal{K}\left(A_\star, B_\star\right)$, it suffices to solve the differential equation $\dot{M}_t = \Phi_{A_\star,B_\star}\left(M_t\right)$ starting from a positive semidefinite $M_0$, or equivalently calculate the integral $M_t = M_0 + \int\limits_0^t \Phi_{A_\star,B_\star}\left(M_s\right)\mathrm{d}s$. In the proof of Theorem 1, we show that $\lim\limits_{t\to\infty} M_t = \mathcal{K}\left(A_\star, B_\star\right)$.

Next, we focus on learning $A_\star, B_\star$ and the additional cost compared to the cost of $\boldsymbol{\pi}^\star$ that we are charged for, because of uncertainties about $A_\star, B_\star$. To that end, we formulate sub-optimalities in the performance of decision-making policies and the penalty due to lack of knowledge about the optimal actions $U_t = \mathcal{L}\left(A_\star, B_\star\right) X_t$. For a general policy $\boldsymbol{\pi}$, the *regret* of $\boldsymbol{\pi}$ at time $T$ is denoted by $\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right)$ and is defined as the cumulative increase in cost by time $T$. That is, the difference between the instantaneous costs of $\boldsymbol{\pi}$ and $\boldsymbol{\pi}^\star$ in (3) is integrated over the time interval $[0, T]$:

$$\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right) = \int\limits_0^T \left[c_t\left(\boldsymbol{\pi}\right) - c_t\left(\boldsymbol{\pi}^\star\right)\right]\mathrm{d}t. \tag{4}$$

Note that under $\boldsymbol{\pi}$, the state trajectory is generated by (1) for $U_t = \boldsymbol{\pi}\left(\left\{U_s\right\}_{0\leq s<t}, \left\{X_s\right\}_{0\leq s\leq t}\right)$, at all times $t \geq 0$. So, $c_t\left(\boldsymbol{\pi}\right) - c_t\left(\boldsymbol{\pi}^\star\right)$ includes the differences between the control actions as well as the state trajectories of $\boldsymbol{\pi}, \boldsymbol{\pi}^\star$. Clearly, the random state evolution in (1) renders $\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right)$ random. So, regret analyses for reinforcement learning policies include worst-case analyses that establish upper-bounds for $\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right)$, as well as average-case analyses that provide bounds for $\mathbb{E}\left[\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right)\right]$. Further, for unknown $A_\star, B_\star$, we hope that the increasing observations of state and action over time will be effectively leveraged so that eventually, the policy makes near-optimal decisions. So, as $t$ grows, we desire $c_t\left(\boldsymbol{\pi}\right) - c_t\left(\boldsymbol{\pi}^\star\right)$ to shrink and so $\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right)$ to scale sub-linearly with $T$. In the sequel, we study $\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right)$, $\mathbb{E}\left[\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right)\right]$ and their dependence on $T$ and the problem parameters.

Another quantity of interest is the *accuracy* of estimating the unknown dynamics. So, letting $A_t, B_t$ be estimates of $A_\star, B_\star$ based on the state-action observations by time $t$; i.e., $\left\{X_s, U_s\right\}_{0\leq s\leq t}$, we study the decay rate of the estimation error $\mathcal{E}\left(A_t, B_t\right)$, as defined in Definition 1. Note that similar to regret, $\mathcal{E}\left(A_t, B_t\right)$ is stochastic.

## 3. Stability Analysis for Perturbed Dynamics Matrices

In this section, we study the effects of uncertainties about the dynamical model on system stability. We specify the minimal information one needs to possess in order to ensure stabilization, and show that a coarse-grained approximation of the truth is sufficient for this purpose. Results of this section will be used later in the design of randomized-estimates policy in Section 5. Importantly, the following stability analysis is general, captures effects of all involved quantities, and provides tight results in the sense that the conditions of Theorem 2 are required for guaranteeing stabilization. In addition, the results presented here are of independent interests, because stability is required for letting the system operate for a reasonable time period, regardless of optimality of the control actions.

To proceed, note that if hypothetically the optimal linear feedback in (3) is applied to the system in (1), then stability is guaranteed. More precisely, applying $\mathcal{L}\left(A_\star, B_\star\right)$, the resulting closed-loop transition matrix $D_\star = A_\star + B_\star\mathcal{L}\left(A_\star, B_\star\right)$ has all its eigenvalues on the open left half-plane of the complex plane, as stated in Theorem 1. The issue is that the true dynamics matrices $A_\star, B_\star$ are *unknown* and need to be learned. However, if some matrices $A, B$ meet the conditions we shortly discuss, one can stabilize the system by applying the linear feedback $U_t = \mathcal{L}\left(A, B\right) X_t$.

To proceed, let $D = A + B\mathcal{L}(A,B)$ be the closed-loop transition matrix of a system with dynamics matrices $A, B$, under the feedback $U_t = \mathcal{L}(A,B) X_t$. Then, let $\rho > 0$ and $\zeta < \infty$ satisfy

$$\overline{\boldsymbol{\lambda}}(D) \leq -\rho, \qquad\qquad \left\|\left\|\mathcal{K}(A,B)\right\|\right\| \leq \zeta. \tag{5}$$

The quantities in (5) are required for studying stability of the matrix $A_\star + B_\star\mathcal{L}(A,B)$, as follows. Remember that we have a closed-loop stability result in Theorem 1: $\overline{\boldsymbol{\lambda}}(D) < 0$. So, the first inequality in (5) quantifies the extent to which $\mathcal{L}(A,B)$ is able to stabilize the system of parameters $A, B$. Intuitively, $\overline{\boldsymbol{\lambda}}(D)$ is the best (i.e., most negative) upper-bound one can hope for the eigenvalues of $A_\star + B_\star\mathcal{L}(A,B)$, since the optimal feedback $\mathcal{L}(A,B)$ is *purposefully* designed for the certainly known matrices $A, B$. Later on, we show that $\rho$ enjoys a positive uniform lower-bound as long as $A, B$ live in some neighborhoods of $A_\star, B_\star$. The second inequality in (5) is somewhat guaranteed by the first one, as we will show in the proof of Theorem 1 (in (22)) that

$$\mathcal{K}(A,B) = \int\limits_0^\infty e^{D^\top t}\left(Q + \mathcal{L}(A,B)^\top R\mathcal{L}(A,B)\right)e^{Dt}\mathrm{d}t. \tag{6}$$

Therefore, $\overline{\boldsymbol{\lambda}}(D) \leq -\rho < 0$ implies that for some $\zeta < \infty$, we have $\left\|\left\|\mathcal{K}(A,B)\right\|\right\| \leq \zeta$. So, $\zeta$ is merely used for simplifying the expressions.

Towards stability analysis, we need further information of $D = A + B\mathcal{L}(A,B)$ that the Jordan form of this matrix provides. Suppose that eigenvalues of $D$ are $\lambda_1, \cdots, \lambda_k$, and let the Jordan decomposition be $D = P^{-1}\Lambda P$; i.e., $\Lambda = \mathrm{diag}(\Lambda_1, \cdots, \Lambda_k)$ is a block-diagonal matrix and all diagonal entries of $\Lambda_i$ are $\lambda_i$, the immediate off-diagonal entries above the diagonal of $\Lambda_i$ are 1, and all other entries of $\Lambda_i$ are 0 (as shown in (24)). Now, we introduce a very important quantity for determining the stability margin. For the above-mentioned blocks $\Lambda_i$, $i = 1, \cdots, k$, let $\boldsymbol{\mu}_i$ denote the dimension of the square matrix $\Lambda_i$, and refer to the largest value among $\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k$ by $\boldsymbol{\mu}_D$.

**Definition 2 (Largest block-size $\boldsymbol{\mu}_D$)** *Letting $P$ and $\Lambda_i \in \mathbb{C}^{\boldsymbol{\mu}_i \times \boldsymbol{\mu}_i}$ be as in the Jordan decomposition $D = P^{-1}\Lambda P$ explained above, define $\boldsymbol{\mu}_D = \max\limits_{1 \leq i \leq k} \boldsymbol{\mu}_i$.*

The quantity $\boldsymbol{\mu}_D$ is the largest size of the blocks $\Lambda_1, \cdots, \Lambda_k$ in the Jordan form and crucially determines the *order* of stability margin, as established in the following theorem.

**Theorem 2 (Stability margin)** *Let $P, \boldsymbol{\mu}_D$ and $\rho, \zeta$ be as in Definition 2 and (5), respectively. Then, following Definition 1, for $\delta > 0$, we have $\overline{\boldsymbol{\lambda}}(A_\star + B_\star\mathcal{L}(A,B)) < -\delta$, as long as*

$$\mathcal{E}(A,B) < \left(1 \wedge \frac{1}{\left\|\left\|\mathcal{L}(A,B)\right\|\right\|}\right)\frac{(\rho-\delta) \wedge (\rho-\delta)^{\boldsymbol{\mu}_D}}{\boldsymbol{\mu}_D^{1/2}\left\|\left\|P\right\|\right\|\left\|\left\|P^{-1}\right\|\right\|}. \tag{7}$$

Note that the definition of $\mathcal{L}(A,B)$ before (3) shows that $\left\|\left\|\mathcal{L}(A,B)\right\|\right\|^{-1}$ in (7) can be replaced with $\boldsymbol{\lambda}_{\min}(R)\,\zeta^{-1}\left\|\left\|B\right\|\right\|^{-1}$. Theorem 2 states that if $\mathcal{E}(A,B)$ is sufficiently small to satisfy (7), then $A_\star + B_\star\mathcal{L}(A,B)$ is stable and all of its eigenvalues in the complex plane lie on the left-hand-side of the vertical line $\Re = -\delta$. In addition, (7) reflects effects of different factors, as follows. First, the stability margin on the right-hand-side of (7) decreases as $\left\|\left\|\mathcal{L}(A,B)\right\|\right\|$ increases. To see the intuition, note that the difference between the closed-loop matrices is $A_\star - A + (B_\star - B)\mathcal{L}(A,B)$, which shows the multiplicative effect of $\mathcal{L}(A,B)$. Further, Definition 2 indicates that $\boldsymbol{\mu}_D^{1/2}\left\|\left\|P\right\|\right\|\left\|\left\|P^{-1}\right\|\right\|$

quantifies non-diagonality of $D$, is at least 1, and becomes 1 for diagonal $D$ (where $P$ is the identity matrix and $\boldsymbol{\mu}_D = 1$). Therefore, more non-diagonal closed-loop matrices $D$ lead to smaller stability margins and make stabilization harder to be learned.

By (5), the dependence on $\rho$ corroborates the intuition that systems whose optimal closed-loop matrices has eigenvalues of larger real-parts (i.e., smaller $\rho$), are harder to stabilize. Moreover, the expression $(\rho - \delta) \wedge (\rho - \delta)^{\boldsymbol{\mu}_D}$ indicates that $\boldsymbol{\mu}_D$ is very important and determines the *rates* of bounding the eigenvalues of $A_\star + B_\star \mathcal{L}(A, B)$. The rates for $\rho - \delta < 1$ and $\rho - \delta > 1$ are different, because of a similar phenomena in the sensitivity of eigenvalues of matrices to perturbations in their entries. This result is of independent interest as it is a generalization of Bauer-Fike Theorem (Bauer and Fike, 1960) to *asymmetric* matrices. Indeed, we establish in the proof of Theorem 2 that *larger blocks in Jordan forms can lead to drastically higher eigenvalue-sensitivities against entries.*

To close this section, we provide uniform lower and upper bounds for $\rho > 0$ and $\zeta < \infty$, respectively. For that purpose, similar to Definition 2, define the largest block size $\boldsymbol{\mu}_\star = \boldsymbol{\mu}_{D_\star}$ based on the Jordan decomposition $D_\star = A_\star + B_\star \mathcal{L}(A_\star, B_\star) = P_\star^{-1} \Lambda_\star P_\star$. Then, we show in the proof of Theorem 2 that $\mathcal{E}(A, B) \leq \epsilon_0$ is sufficient for stabilization, and it holds that $\rho \geq \boldsymbol{\lambda}_{\min}(Q)\, 4^{-1} \vert\!\vert\!\vert \mathcal{K}(A_\star, B_\star) \vert\!\vert\!\vert^{-1}$, and $\zeta \leq 2 \vert\!\vert\!\vert \mathcal{K}(A_\star, B_\star) \vert\!\vert\!\vert$, as long as

$$\left(1 \vee \vert\!\vert\!\vert \mathcal{L}(A_\star, B_\star) \vert\!\vert\!\vert\right) \epsilon_0 = \frac{\left(-\overline{\boldsymbol{\lambda}}(D_\star)\right) \wedge \left(-\overline{\boldsymbol{\lambda}}(D_\star)\right)^{\boldsymbol{\mu}_\star}}{\boldsymbol{\mu}_\star^{1/2} \vert\!\vert\!\vert P_\star^{-1} \vert\!\vert\!\vert \, \vert\!\vert\!\vert P_\star \vert\!\vert\!\vert} \wedge \left[4 \int_0^\infty \vert\!\vert\!\vert e^{D_\star t} \vert\!\vert\!\vert^2 \mathrm{d}t\right]^{-1}. \tag{8}$$

## 4. Tight Regret Expressions and Policy Differentiation

In this section, we investigate sub-optimalities and provide a sharp expression for the regret that non-optimal control actions cause. Such an investigation is vital since reinforcement learning policies need to learn the unknown dynamics $A_\star, B_\star$ and so they require to take non-optimal actions.

To proceed, let $U_t$ be the control action of the policy $\boldsymbol{\pi}$ at time $t$. In the following theorem, we quantify $\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}(T)$ in terms of deviations $U_t - \mathcal{L}(A_\star, B_\star) X_t$, and introduce $\boldsymbol{\alpha}_T$ that fully assesses the regret of $\boldsymbol{\pi}$. In fact, $\boldsymbol{\alpha}_T$ unifies average-case and worst-case analyses by capturing both $\mathbb{E}[\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}(T)]$ and $\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}(T)$. Further, Theorem 3 provides scalings with different problem parameters and shows that $\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}(T) - \mathbb{E}[\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}(T)]$ scales linearly with the dimension of the Brownian motion.

**Theorem 3 (Regret analysis)** *Suppose that $L_t$ is a bounded piecewise-continuous function of $t$, and $\boldsymbol{\pi}$ is the policy $U_t = L_t X_t$. Then, we have $\mathbb{E}[\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}(T)] = \mathbb{E}[\boldsymbol{\alpha}_T]$, and*

$$\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}(T) \quad = \quad \boldsymbol{\alpha}_T + \mathcal{O}\left(\boldsymbol{\omega}_{\mathcal{R}} \boldsymbol{\alpha}_T^{1/2} \log \boldsymbol{\alpha}_T\right),$$

*where* $\boldsymbol{\omega}_{\mathcal{R}} = \frac{\vert\!\vert\!\vert C \vert\!\vert\!\vert \, \vert\!\vert\!\vert \mathcal{K}(A_\star, B_\star) \vert\!\vert\!\vert^{3/2} d_W}{\boldsymbol{\lambda}_{\min}(Q)^{1/2} \boldsymbol{\lambda}_{\min}(R)^{1/2}}$, $D_\star = A_\star + B_\star \mathcal{L}(A_\star, B_\star)$, $E_t = e^{D_\star^\top t} \mathcal{K}(A_\star, B_\star) e^{D_\star t}$, *and*

$$\boldsymbol{\alpha}_T = \int_0^T \left\| R^{1/2} \left(L_t - \mathcal{L}(A_\star, B_\star)\right) X_t \right\|^2 \mathrm{d}t - 2 \int_0^T \left(X_t^\top E_{T-t} B_\star \left(L_t - \mathcal{L}(A_\star, B_\star)\right) X_t\right) \mathrm{d}t.$$

To establish Theorem 3, we utilize the theory of continuous-time martingales and (in Lemma 2) develop novel results on comparative ratios of stochastic integrals. More importantly, we construct the new framework of *policy differentiation* for finding sharp regret bounds. Broadly speaking, policy differentiation precisely evaluates the regret in terms of infinitesimal sub-optimalities and integrates these infinitesimal deviations to obtain $\boldsymbol{\alpha}_T$, in which the integrand $\left\| R^{1/2} \left( L_t - \mathcal{L}\left(A_\star, B_\star\right) \right) X_t \right\|^2$ plays a role similar to the *derivative* of the regret. This framework can be used for analogous regret analyses in other continuous-time reinforcement learning problems.

The boundedness and piecewise continuity conditions in Theorem 3 are somewhat natural because the optimal policy in (3) is a time-invariant feedback, and so one gains nothing by violating these conditions. Furthermore, since by Theorem 1 we have $\overline{\boldsymbol{\lambda}}\left(D_\star\right) < 0$, the matrix $E_t$ exponentially decays as $t$ grows. Thus, the second integral in the definition of $\boldsymbol{\alpha}_T$ is dominated by the first one. Moreover, note that the above-mentioned matrix appears in $\boldsymbol{\alpha}_T$ in the form of $E_{T-t}$, i.e., with a time inversion. This reflects the fact that sub-optimal control feedbacks $U_t = L_t X_t$ have descending effects on the regret $\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right)$ as we move from $T$ backward in time that $t$ descends.

Putting the discussions in the above two paragraphs all together, we conclude as follows. Theorem 3 shows that the sub-optimality $\boldsymbol{\pi}$ incurs at time $t$, scales as the **square** of the deviation $L_t - \mathcal{L}\left(A_\star, B_\star\right)$. On top of that, the constant $\boldsymbol{\omega}_{\mathcal{R}}$ reflects effects of different parameters and indicates, for example, that $\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right) - \boldsymbol{\alpha}_T$ scales linearly with $d_W$.

The results of Theorem 3 are insightful along different directions. First, the exact equality $\mathbb{E}\left[\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right)\right] = \mathbb{E}\left[\boldsymbol{\alpha}_T\right]$ can be used for establishing minimax *lower-bounds* for regret by finding the fastest rates $L_t - \mathcal{L}\left(A_\star, B_\star\right)$ can shrink. Further, since $\boldsymbol{\alpha}_T^{1/2} \log \boldsymbol{\alpha}_T = \mathcal{O}\left(\boldsymbol{\alpha}_T\right)$, both the average-case criteria $\mathbb{E}\left[\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right)\right]$ as well as the worst-case regret $\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right)$ are captured by $\boldsymbol{\alpha}_T$. In other words, Theorem 3 indicates that the fluctuations of $\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right)$ around its expectation $\mathbb{E}\left[\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right)\right]$ are in magnitude smaller than the expected value. Thus, studying $\boldsymbol{\alpha}_T$ is *sufficient and necessary* for regret analysis and there is a tight and reciprocal relationship between $\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right), \boldsymbol{\alpha}_T$, for **all** policies.

Moreover, as time goes by, a reinforcement learning policy becomes *progressively* more capable of narrowing down the sub-optimality gap by estimating the unknown dynamics $A_\star, B_\star$. Indeed, as soon as having sufficiently long trajectories to learn $A_\star, B_\star$ accurate enough to satisfy $\left\| R^{1/2} \left( L_t - \mathcal{L}\left(A_\star, B_\star\right) \right) X_t \right\| < 1$, the regret grows much slower since $\boldsymbol{\alpha}_T$ integrates the *squares* of these deviations. For example, if the estimation accuracy satisfies the ideal square-root rate $\mathcal{E}\left(A_t, B_t\right) = \mathcal{O}\left(t^{-1/2}\right)$, then the regret is a logarithmic function of time; $\boldsymbol{\alpha}_T = \mathcal{O}\left(\log T\right)$. However, due to the trade-off between the exploration and exploitation that will be elaborated shortly, this is not the case and to obtain the above error rate, $U_t$ needs to persistently deviate from $\mathcal{L}\left(A_\star, B_\star\right) X_t$, which causes a *linearly* growing regret (see Proposition 1 in the appendices).

Theorem 3 provides both a general result for analyzing reinforcement learning policies, as well as a useful insight on how to design them to minimize the regret. We utilize this insight to design Algorithm 1 and to establish Theorem 4 in the next section. Indeed, we randomize the parameter estimates so that $U_t$ appropriately deviates from $\boldsymbol{\pi}^\star$, leading to $\mathcal{E}\left(A_t, B_t\right) = \widetilde{\mathcal{O}}\left(t^{-1/4}\right)$. So, we obtain the efficient regret bound $\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right) = \mathcal{O}\left(\boldsymbol{\alpha}_T\right) = \widetilde{\mathcal{O}}\left(T^{1/2}\right)$.

## 5. Randomized-Estimates Policy

In this section, we discuss a fast and tractable algorithm with an efficient performance for cost minimization subject to uncertainties about the dynamics matrices $A_\star, B_\star$. First, we explain the fundamental exploration-exploitation dilemma. Then, we investigate a procedure for estimating the unknown dynamics using the data of state-action trajectory. Based on that, an online reinforcement learning policy that employs randomizations of the parameter estimates for balancing exploration versus exploitation is presented in Algorithm 1. Next, a regret bound is established in Theorem 4 indicating that Algorithm 1 efficiently minimizes the cost function so that the regret scales as the square-root of the time. We also specify the rates of identifying the dynamics matrices.

According to Theorem 3, the policy needs to ensure that $U_t \approx \mathcal{L}\left(A_\star, B_\star\right) X_t$ in order to incur a small regret. Furthermore, since $A_\star, B_\star$ are unknown, the policy needs to estimate them based on the data $\{X_s, U_s\}_{0\leq s\leq t}$. However, if $U_s \approx \mathcal{L}\left(A_\star, B_\star\right) X_s$, then the $U_s$ coordinates of the data point $X_s, U_s$ become (almost) uninformative as they are (approximately) linear transformations of the $X_s$ coordinates. This defeats the purpose and renders accurate estimation of $A_\star, B_\star$ infeasible. Note that accurate approximations of $A_\star, B_\star$ are needed for taking near-optimal control actions. This, known as the exploration-exploitation dilemma, is the main obstacle in online reinforcement learning and shows that a low-regret policy needs to carefully *diversify* the actions $\{U_s\}_{0\leq s\leq t}$ by *deviating* from $\{\mathcal{L}\left(A_\star, B_\star\right) X_s\}_{0\leq s\leq t}$.

### 5.1. Design of the algorithm and intuitions

Now, we discuss the learning procedure in Algorithm 1, based on extensions of the least-squares estimates. Suppose that instead of the full data $X_s, U_s$ for real values of $s \geq 0$, one has access to samples at a discrete set of time points; $X_{k\epsilon}, U_{k\epsilon}$ for $k = 0, 1, \cdots$. Then, (1) for a small $\epsilon$ gives the approximate data generation mechanism $X_{(k+1)\epsilon} - X_{k\epsilon} = \left(A_\star X_{k\epsilon} + B_\star U_{k\epsilon}\right)\epsilon + C\left(W_{(k+1)\epsilon} - W_{k\epsilon}\right)$. So, an approach is to estimate $A_\star, B_\star$ by minimizing $\sum_k \left\|X_{(k+1)\epsilon} - X_{k\epsilon} - \left(AX_{k\epsilon} + BU_{k\epsilon}\right)\epsilon\right\|^2$ over $A, B$. Letting $Y_s = \left[X_s^\top, U_s^\top\right]^\top$ and $\epsilon \to 0$, we get the continuous-time estimate based on the full trajectory $\{X_s, U_s\}_{0\leq s\leq t}$. The result is shown in (9) below and will be used by Algorithm 1.

To ensure that the system evolves stably, Algorithm 1 projects the estimates on the following stabilization oracle $\mathcal{S}_0$ in lights of Theorem 2. In the sequel, we explain how one can learn $\mathcal{S}_0$ fast.

**Definition 3** *For a fixed $\delta_0 > 0$, let $\mathcal{S}_0$ be a set containing matrices $A, B$ that satisfy (7) for $\delta = \delta_0$.*

Intuitively, the system is stabilized by having access to $\mathcal{S}_0$, despite uncertainties about the true dynamics matrices $A_\star, B_\star$. Note that the condition in (7) is verifiable since $\rho, \zeta$ depend on the *known* parameter estimates $A, B$. Availability of a stabilization set is a common assumption in the literature of online reinforcement learning policies for linear systems (Mandl et al., 1988; Mandl, 1989; Caines, 1992; Abeille and Lazaric, 2018; Ouyang et al., 2019; Faradonbeh et al., 2020a; Ziemann and Sandberg, 2020a,b). For example, if an initial stabilizing feedback $L_0$ is available, we can devote a period to only explore by applying sub-optimal control actions, and use the resulting observations to learn $\mathcal{S}_0$. Such procedures, that ignore the main objective of regret minimization for a short time period, are shown to be fast and effective in the sense that the probability of failing to learn to stabilize, decays exponentially with time (Shirani Faradonbeh et al., 2022). In systems that are in operation prior to running Algorithm 1 or in open-loop-stable systems, this condition automatically holds (in the latter case, $L_0 = 0$ is an initial stabilizer). Similarly, in systems with a *reset* option that can immediately steer the system-state to small values, $\mathcal{S}_0$ can be learned fast (Duncan et al., 1999; Caines and Levanony, 2019; Basei et al., 2021).

In absence of initial stabilizer and state-reset options, learning $\mathcal{S}_0$ will be more challenging. In Section 3 we saw that an $\epsilon_0$ neighborhood of $A_\star, B_\star$ is sufficient for bounding $\rho, \zeta$, for $\epsilon_0$ in (8). That is, coarse-grained approximations of $A_\star, B_\star$ suffice for stabilization. So, $\mathcal{S}_0$ can be learned from short state trajectories derived by applying randomized control actions (Caines and Levanony, 2019; Faradonbeh et al., 2018, 2019b; Lale et al., 2020a; Chen and Hazan, 2021; Gramlich and

Ebenbauer, 2022). Importantly, fast and reliable learning-based stabilization can be ensured via Bayesian methods (Faradonbeh and Faradonbeh, 2022). These methods retain a Gaussian posterior about the unknown true dynamics $A_\star, B_\star$, and design stabilizing feedbacks as if samples from the posterior coincide with the truth. Importantly, even if failed at their first attempts, these Bayesian methods can be utilized again with no need to repeat the state observation and learning procedure. Technically, if a failure is detected (e.g., if the state magnitude $\|X_t\|$ keeps growing), one can can resample from the posterior until successful stabilization (Faradonbeh and Faradonbeh, 2022). Note that effectiveness of the above methods does not depend on availability of any prior distribution, although an informative prior can help to improve learning from shorter state trajectories.

Finally, we will show (in Theorem 4) that the algorithm learns $A_\star, B_\star$ with the rate $t^{-1/4}$. So, the projection on $\mathcal{S}_0$ will be automatically performed after the time $t_0 = \widetilde{\mathcal{O}}\left(\epsilon_0^{-4}\right)$.

The algorithm proceeds as follows. For some fixed $\gamma > 1$, the exponentially growing sequence $\{\gamma^n\}_{n=0}^{\infty}$ contains the time instants at which the algorithm updates the parameter estimates. In fact, Algorithm 1 applies control actions $U_t = \mathcal{L}\left(A_n, B_n\right) X_t$ during the time period $\gamma^n \leq t < \gamma^{n+1}$, where $A_n, B_n$ are estimates of $A_\star, B_\star$ , based on the trajectory up to time $\gamma^n$.

Further, to ensure that the policy commits sufficiently to explore the environment, a random matrix $\Theta_n$ is added to the parameter estimates at time $t = \gamma^n$. Then, Algorithm 1 projects the resulting $d_X \times (d_X + d_U)$ matrix onto $\mathcal{S}_0$. Formally, let $\Pi_{\mathcal{S}_0}(\cdot)$ denote projection on $\mathcal{S}_0$; i.e., it gives the closest matrix in $\mathcal{S}_0$ according to the distance induced by the Frobenius norm. Then, define

$$[A_n, B_n] = \Pi_{\mathcal{S}_0}\left( \left[ \int_0^{\gamma^n} Y_s \mathrm{d}X_s^\top \right]^\top \left[ \int_0^{\gamma^n} Y_s Y_s^\top \mathrm{d}s \right]^\dagger + \Theta_n \right), \tag{9}$$

where $Y_s = \left[X_s^\top, U_s^\top\right]^\top$ and the $d_X \times (d_X + d_U)$ matrices $\{\Theta_n\}_{n=0}^{\infty}$ are independent of everything else and of each others. Further, the random matrix $\Theta_n$ that is used at time $t = \gamma^n$ has independent Gaussian entries of mean zero and standard deviation $\sigma_n = \sigma_0 \left(\gamma^{-n} n\right)^{1/4} = \sigma_0 \left(t^{-1} \log_\gamma t\right)^{1/4}$, for some fixed $\sigma_0$. This decay rate of $\sigma_n$ is delicately adjusted for two purposes. On one hand, $\Theta_n$ is sufficiently large for randomizing the parameter estimates to ensure that effective exploration occurs and the current data is diverse enough so that we obtain accurate estimates in the future. On the other hand, $\Theta_n$ is sufficiently small to let the current estimates remain accurate and prevent significant deviations. Otherwise, large values of $\Theta_n$ deteriorate the current efficient exploitation. More precisely, this value of $\sigma_n$ is selected by minimizing its role in the regret that consists of summations of some terms of the form $\gamma^n \left(\sigma_n^2 + \sigma_n^{-2} n\right)$.

**Algorithm 1 : Randomized-Estimates Policy**

```
Select $\gamma > 1$ and $A_0, B_0 \in \mathcal{S}_0$ arbitrarily
For $0 \leq t < 1 = \gamma^0$, apply $U_t = \mathcal{L}(A_0, B_0) X_t$
    for $n = 0, 1, 2, \cdots$ do
        Obtain parameter estimates $A_n, B_n$ by (9)
        while $\gamma^n \leq t < \gamma^{n+1}$ do
            Take control action $U_t = \mathcal{L}(A_n, B_n) X_t$
        end
    end
```

### 5.2. Analysis of the algorithm and performance guarantees

The memory that Algorithm 1 occupies is remarkably small since it can update the parameter estimates by only storing the values of the two integrals in (9) in an online fashion. Furthermore, calculations can be done quite fast, making update of the parameter estimates at time $\gamma^n$ immediately effective. Note that the computational complexity of numerically obtaining a sufficiently accurate $\mathcal{L}\left(A_n, B_n\right)$ is at worst $\mathcal{O}\left(\left(d_X + d_U\right)^3 n\right)$, for the matrix operations in the definition of $\boldsymbol{\pi}^\star$ in (3), and for the fact that the (integration or differentiation) procedures described after Theorem 1 converge exponentially fast, while the next theorem indicates that an accuracy of $\mathcal{O}\left(\gamma^{-n/4}\right)$ is sufficient (and necessary).

The rationale for freezing the parameter estimates for exponentially growing time intervals $\gamma^n \leq t < \gamma^{n+1}$ is that Algorithm 1 can defer the learning step until collecting enough observations $Y_s$ so that a new update of parameter estimates is effectively more accurate than the previous one.

The following result provides performance guarantees for the randomized-estimates policy.

**Theorem 4 (Analysis of Algorithm 1)** *Let the policy $\boldsymbol{\pi}$ and the estimates $A_n, B_n$ be those in Algorithm 1. Assume that $n$ satisfies $\gamma^n \leq T < \gamma^{n+1}$. Then, using Definition 1 and (4), we have*

$$\mathcal{E}\left(A_n, B_n\right)^2 = \mathcal{O}\left(\boldsymbol{\omega}_{\mathcal{E}} T^{-1/2} \log T\right), \qquad \boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right) = \mathcal{O}\left(\boldsymbol{\omega}_{\boldsymbol{\pi}} T^{1/2} \log T\right),$$

*where*

$$\boldsymbol{\omega}_{\mathcal{E}} = \left(d_X + d_U\right)\left(\frac{d_X}{\log \gamma} + \frac{d_W \left\|C\right\|^2}{\boldsymbol{\lambda}_{\min}\left(CC^\top\right)}\right), \qquad \boldsymbol{\omega}_{\boldsymbol{\pi}} = \frac{(\gamma - 1)\left\|C\right\|^2 \left\|\mathcal{K}\left(A_\star, B_\star\right)\right\|^6 \left\|R\right\|}{\boldsymbol{\lambda}_{\min}\left(Q\right)^2 \boldsymbol{\lambda}_{\min}\left(R\right)^4}\ \boldsymbol{\omega}_{\mathcal{E}}.$$

Theorem 4 indicates efficiency of Algorithm 1: At time $T$, the sub-optimality gap is as small as $\mathcal{O}\left(\boldsymbol{\omega}_{\boldsymbol{\pi}} T^{-1/2} \log T\right)$. It also provides $\boldsymbol{\omega}_{\boldsymbol{\pi}}, \boldsymbol{\omega}_{\mathcal{E}}$ that reflect the dependence of estimation error and regret on different parameters in the problem. So, the regret scales linearly with the number of unknown parameters in $A_\star, B_\star$, while the estimation error dwindles linearly with the dimension.

To establish Theorem 4, we study the learning step in (9) and determine the rates Algorithm 1 estimates $A_\star, B_\star$. For that purpose, we prove concentration bounds for the empirical covariance matrices of the state vectors and also show anti-concentration of the Gram matrix $\int\limits_0^t Y_s Y_s^\top \mathrm{d}s$ of the signal $Y_s = \left[X_s^\top, U_s^\top\right]^\top$, as $t$ grows. Then, we establish bounds on the comparative ratios of stochastic integrals and use that for controlling the estimation error. Furthermore, we show that the optimal feedback matrices have a Lipschitz property with respect to the dynamics matrices and leverage that for finding the deviation rates from the optimal feedback. Finally, we utilize policy differentiation and Theorem 3 for getting the regret bounds in Theorem 4.

Note that for obtaining descending estimation errors and sub-linear regret bounds we need $\boldsymbol{\lambda}_{\min}\left(CC^\top\right) > 0$. This is a standard requirement in estimation and control of stochastic linear systems and expresses that all coordinates of the state vectors are randomized by the Brownian motion $\left\{W_t\right\}_{t \geq 0}$ in a relatively short time (Levanony and Caines, 2001; Subrahmanyam and Rao, 2019; Caines and Levanony, 2019). So, all state variables have significant roles in the dynamics. From a modeling point of view, $\boldsymbol{\lambda}_{\min}\left(CC^\top\right) > 0$ indicates that the stochastic differential equation in (1) is irreducible in the sense that a smaller subset of state variables is *insufficient* for capturing the stochastic dynamical behavior of the environment.

To close this section, observe that the estimation error of Algorithm 1 shrinks as $T^{-1/4}$. So, it does not decay with the ideal square-root rate because the main priority of Algorithm 1 is to minimize its regret by exploring minimally. However, if the randomization matrices $\Theta_n$ are persistent and their standard deviations do not diminish as $n$ grows, then we obtain the square-root consistency. This is formalized in Proposition 1 in Appendix E. Of course, the compromise is that $\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right)$ grows *linearly* with $T$ if $\Theta_n$ does not dwindle as $n$ grows.

## 6. Numerical Illustrations of Estimation Error and Regret

Now, we provide numerical analyses for showcasing the performance of Algorithm 1 for estimating the unknown dynamics matrices and learning the optimal policy. For this purpose, we assume that the true continuous-time system matrices are lateral-directional state-space matrices of X-29A airplane at 4000 ft altitude (Bosworth, 1992). The system is of dimension $d_X = 4$, is controlled by two dimensional commands; $d_U = 2$, and the transition and input matrices in (1) are

$$A_\star = \begin{bmatrix} -0.1850 & 0.1475 & -0.9825 & 0.1120 \\ -0.3467 & -1.710 & 0.9029 & -0.5843 \times 10^{-6} \\ 1.174 & -0.0825 & -0.1826 & -0.4428 \times 10^{-7} \\ 0.0 & 1.0 & 0.1429 & 0.0 \end{bmatrix}, B_\star = \begin{bmatrix} -0.4470 \times 10^{-3} & 0.4020 \times 10^{-3} \\ 0.3715 & 0.0549 \\ 0.0265 & -0.0135 \\ 0.0 & 0.0 \end{bmatrix}.$$

Note that the dimensions of the control action and the state vector, as well as open-loop *instability* of $A_\star$ and the small entries of $B_\star$, render control of the above-mentioned airplane challenging. Further, we let the coefficient matrix of the Brownian disturbance $W_t$ be $C = 0.2 \times I_4$, and employ Algorithm 1 to learn to control a quadratic cost with the weight matrices $Q = 10 \times I_{d_X}, R = I_{d_U}$. The online reinforcement learning policy of Algorithm 1 is run for 500 seconds, while the parameter estimates are updated at times $t = \gamma^n$, for integer values of $n$ and $\gamma = 1.2$. To find an initial stabilizing feedback, we run a Bayesian learning algorithm for 25 seconds (Faradonbeh and Faradonbeh, 2022).

The normalized rates of the estimation error are plotted in Figure 1, versus the continuous time $T$. To illustrate the rates in Theorem 4, the figure contains multiple trajectories of $T^{1/2}\mathcal{E}\left(A_n, B_n\right)^2$, while $n, T$ satisfy $\gamma^n \leq T < \gamma^{n+1}$. The normalized estimation errors in the left panel are (almost) bounded as time grows, corroborating Theorem 4. The right panel depicts normalized regret versus time: the horizontal axis is $T$ and the vertical one represents $T^{-1/2}\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right)$, where $\boldsymbol{\pi}$ is the online reinforcement learning policy in Algorithm 1. Again, it is clear that the statement of Theorem 4 holds, as the normalized regret remains bounded as time grows.

## 7. Concluding Remarks

We studied online reinforcement learning policies for unknown continuous-time stochastic linear systems and presented algorithms that learn to minimize quadratic costs. Three important problems are fully investigated, followed by the intuitions and implications of the presented analyses.

First, we studied stabilization of stochastic linear systems based on inexact dynamics matrices and proved Theorem 2 that specifies the coarse-grained accuracy for guaranteeing stability. Then, proposing the novel approach of policy differentiation, we established a reciprocal result in Theorem 3 for the regret that policies cause by taking sub-optimal actions. More importantly, we presented the online reinforcement learning Algorithm 1 and established its performance guarantees. Indeed, Theorem 4 expresses that the estimation error rate of Algorithm 1 is $dT^{-1/4}\log^{1/2} T$ and it enjoys the efficient regret bound $\mathcal{O}\left(d^2 T^{1/2} \log T\right)$, where $T$ is the time and $d$ is the dimension.

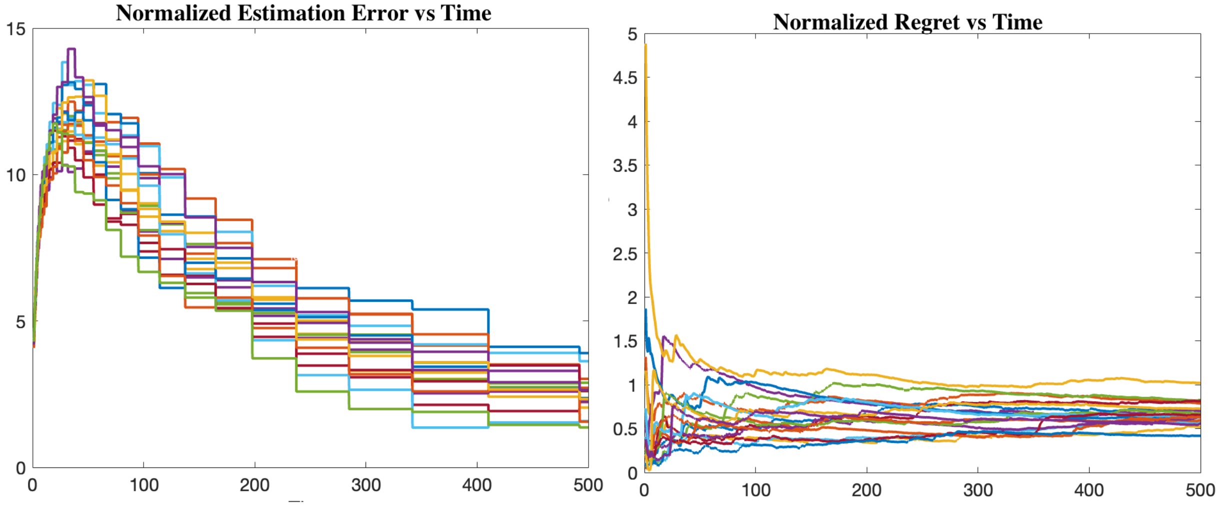

Figure 1: **Left:** Normalized estimation error $T^{1/2}\mathcal{E}\left(A_n, B_n\right)^2$ is plotted vs $T$, such that $\gamma^n \leq T < \gamma^{n+1}$, for some integer $n$. Multiple replicates of the normalized estimation error are reported in the graph, which clearly remain bounded as time grows. Therefore, the graph depicts Theorem 4 about the rates Algorithm 1 learns the dynamics matrices.
**Right:** The graph presents curves of the normalized regret $T^{-1/2}\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right)$ versus time $T$, while $\boldsymbol{\pi}$ is the policy in Algorithm 1. Multiple replicates of the system are simulated, all corroborating Theorem 4 that the normalized regret remains (almost) bounded.

As an initiating paper on design and analysis of online reinforcement learning policies for continuous-time stochastic systems, this study introduces interesting directions for future work. That includes establishing regret lower-bounds, investigating high-dimensional systems with structured dynamics such as low-rank or sparse matrices, and designing efficient policies under imperfect state-observations. Another interesting avenue for future studies that the authors expect the presented techniques apply to, is that of learning to control systems with non-linear dynamics or arbitrary cost functions.

## 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. JMLR Workshop and Conference Proceedings, 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.

Seyed Mohammad Asghari, Yi Ouyang, and Ashutosh Nayyar. Regret bounds for decentralized learning in cooperative multi-agent dynamical systems. In *Conference on Uncertainty in Artificial Intelligence*, pages 121–130. PMLR, 2020.

Paolo Baldi. *Stochastic Calculus: An Introduction Through Theory and Exercises*. Springer, 2017.

Matteo Basei, Xin Guo, Anran Hu, and Yufei Zhang. Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon. *Available at SSRN 3848428*, 2021.

Friedrich L Bauer and Charles T Fike. Norms and exclusion theorems. *Numerische Mathematik*, 2 (1):137–141, 1960.

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

PE Caines. Continuous time stochastic adaptive control: non-explosion, $\varepsilon$-consistency and stability. *Systems & control letters*, 19(3):169–176, 1992.

Peter E Caines and David Levanony. Stochastic $\varepsilon$-optimal linear quadratic adaptation: An alternating controls policy. *SIAM Journal on Control and Optimization*, 57(2):1094–1126, 2019.

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.

Siew Chan, GC Goodwin, and Kwai Sin. Convergence properties of the riccati difference equation in optimal filtering of nonstabilizable systems. *IEEE Transactions on Automatic Control*, 29(2): 110–118, 1984.

Goong Chen, Guanrong Chen, and Shih-Hsun Hsu. *Linear stochastic control systems*, volume 3. CRC press, 1995.

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

C De Souza, M Gevers, and G Goodwin. Riccati equations in optimal filtering of nonstabilizable systems having singular state transition matrices. *IEEE Transactions on Automatic control*, 31 (9):831–838, 1986.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4): 633–679, 2020.

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

Tyrone E Duncan and Bozenna Pasik-Duncan. Adaptive control of continuous-time linear stochastic systems. *Mathematics of Control, signals and systems*, 3(1):45–60, 1990.

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 and Mohamad Sadegh Shirani Faradonbeh. Bayesian algorithms learn to stabilize unknown continuous-time systems. *IFAC-PapersOnLine*, 55(12):377–382, 2022.

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, 2018.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. On applications of bootstrap in continuous space reinforcement learning. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 1977–1984. IEEE, 2019a.

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, 2019b.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. On adaptive linear–quadratic regulators. *Automatica*, 117:108982, 2020a.

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

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Optimism-based adaptive regulation of linear-quadratic systems. *IEEE Transactions on Automatic Control*, 66(4): 1802–1808, 2020c.

Daniel T Gillespie. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.*, 58:35–55, 2007.

Dennis Gramlich and Christian Ebenbauer. Fast identification and stabilization of unknown linear systems. *arXiv preprint arXiv:2208.10392*, 2022.

Panqanamala Ramana Kumar and Pravin Varaiya. *Stochastic systems: Estimation, identification, and adaptive control*. SIAM, 2015.

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. *arXiv preprint arXiv:2003.11227*, 2020b.

Neil D Lawrence, Mark Girolami, Magnus Rattray, and Guido Sanguinetti. *Learning and inference in computational systems biology*. MIT press, 2010.

David Levanony and Peter E Caines. On persistent excitation for linear systems with stochastic coefficients. *SIAM journal on control and optimization*, 40(3):882–897, 2001.

Petr Mandl. Consistency of estimators in controlled systems. In *Stochastic Differential Systems*, pages 227–234. Springer, 1989.

Petr Mandl, Tyrone E Duncan, and Bożenna Pasik-Duncan. On the consistency of a least squares identification procedure. *Kybernetika*, 24(5):340–346, 1988.

Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

Yi Ouyang, Mukul Gagrani, and Rahul Jain. Posterior sampling-based reinforcement learning for control of unknown linear systems. *IEEE Transactions on Automatic Control*, 65(8):3600–3607, 2019.

Huyên Pham. *Continuous-time stochastic control and optimization with financial applications*, volume 61. Springer Science & Business Media, 2009.

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.

Hanspeter Schmidli. *Stochastic control in insurance*. Springer Science & Business Media, 2007.

Mohamad Kazem Shirani Faradonbeh, Mohamad Sadegh Shirani Faradonbeh, and Mohsen Bayati. Thompson sampling efficiently learns to control diffusion processes. *Advances in Neural Information Processing Systems*, 35:3871–3884, 2022.

Allamaraju Subrahmanyam and Ganti Prasada Rao. *Identification of Continuous-time Systems: Linear and Robust Parameter Estimation*. CRC Press, 2019.

Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. Reinforcement learning in continuous time and space: A stochastic control approach. *J. Mach. Learn. Res.*, 21:198–1, 2020.

Jiongmin Yong and Xun Yu Zhou. *Stochastic controls: Hamiltonian systems and HJB equations*, volume 43. Springer Science & Business Media, 1999.

Ingvar Ziemann and Henrik Sandberg. On a phase transition of regret in linear quadratic control: The memoryless case. *IEEE Control Systems Letters*, 5(2):695–700, 2020a.

Ingvar Ziemann and Henrik Sandberg. Regret lower bounds for unbiased adaptive control of linear quadratic regulators. *IEEE Control Systems Letters*, 4(3):785–790, 2020b.

## Contents

## Appendix A. Proof of Theorem 1 (Optimal policy)

Fixing $\epsilon > 0$, suppose that the control inputs $U_t$ are frozen in intervals of length $\epsilon$ and can change only at times $k\epsilon$, for $k = 0, 1, \cdots$. That is, for all times $t$ satisfying $k\epsilon \leq t < (k+1)\epsilon$, the action vector is fixed; $U_t = U_{k\epsilon}$. Next, we proceed towards finding a decision-making policy for minimizing the expected average cost. Note that due to the above-mentioned freezing during $\epsilon$-length intervals, the resulting decision-making policies can be sub-optimal, and indeed provide an upper bound for the optimal cost value. However, we will address this possible sub-optimality at the end of the proof, and with a slight abuse of notation, we still use $\boldsymbol{\pi}^\star$ to denote the above-mentioned policy.

Next, fix an arbitrary time horizon $T$, and denote the minimum cost-to-go at time $t$ by

$$\mathcal{V}_t\left(X_t\right) = \inf \mathbb{E}\left[\int\limits_t^T c_s\left(\boldsymbol{\pi}^\star\right)\mathrm{d}s \middle| \mathcal{F}_t\right],$$

where the infimum is taken over non-anticipating policies that freeze the control action in $\epsilon$-length intervals, as elaborated above, and the information at time $t$ is $\mathcal{F}_t = \sigma\left(X_{0:t}, U_{0:t}\right)$; the sigma-field generated by the state and action vectors up to the time. Now, finding an optimal policy is equivalent to applying dynamic programming principle and writing Bellman optimality equations (Kumar and Varaiya, 2015; Chen et al., 1995). So, we have

$$\mathcal{V}_{k\epsilon}\left(X_{k\epsilon}\right) = \min_{U_{k\epsilon}} \mathbb{E}\left[\int\limits_{k\epsilon}^{(k+1)\epsilon} c_{X_t, U_{k\epsilon}}\left(\boldsymbol{\pi}^\star\right)\mathrm{d}t + \mathcal{V}_{(k+1)\epsilon}\left(X_{(k+1)\epsilon}\right) \middle| \mathcal{F}_{k\epsilon}\right], \tag{10}$$

subject to the dynamics equation in (1).

For the sake of simplicity, suppose that $T/\epsilon$ is an integer. Solving (10) for $k = T/\epsilon - 1$, we get the optimal control action $U^\star_{k\epsilon} = 0$. Accordingly, this gives

$$\mathcal{V}_{(k+1)\epsilon}\left(X_{(k+1)\epsilon}\right) = X^\top_{(k+1)\epsilon} Q X_{(k+1)\epsilon}\epsilon,$$

for $k = T/\epsilon - 2$, which, after substituting in (10), becomes

$$\mathcal{V}_{k\epsilon}\left(X_{k\epsilon}\right) = \min_{U_{k\epsilon}} \int\limits_{k\epsilon}^{(k+1)\epsilon} \mathbb{E}\left[X_t^\top Q X_t \middle| \mathcal{F}_{k\epsilon}\right]\mathrm{d}t + U^\top_{k\epsilon} R U_{k\epsilon}\epsilon + \mathbb{E}\left[X^\top_{(k+1)\epsilon} Q X_{(k+1)\epsilon} \middle| \mathcal{F}_{k\epsilon}\right]\epsilon, \tag{11}$$

where we applied Fubini's Theorem to derive

$$\mathbb{E}\left[\int\limits_{k\epsilon}^{(k+1)\epsilon} c_{X_t, U_{k\epsilon}}\left(\boldsymbol{\pi}^\star\right)\mathrm{d}t \middle| \mathcal{F}_{k\epsilon}\right] = \int\limits_{k\epsilon}^{(k+1)\epsilon} \mathbb{E}\left[X_t^\top Q X_t \middle| \mathcal{F}_{k\epsilon}\right]\mathrm{d}t + U^\top_{k\epsilon} R U_{k\epsilon}\epsilon. \tag{12}$$

However, solving the dynamics (1) for $k\epsilon \leq t \leq (k+1)\epsilon$, we obtain

$$X_t = e^{A_\star(t-k\epsilon)} X_{k\epsilon} + \int\limits_{k\epsilon}^{t} e^{A_\star(t-s)} C \mathrm{d}W_s + \int\limits_{k\epsilon}^{t} e^{A_\star(t-s)} \mathrm{d}s B_\star U_{k\epsilon},$$

which together with Ito's Lemma, $\mathrm{d}W_s \mathrm{d}W_s^\top = I_{d_W} \mathrm{d}s$ (Oksendal, 2013), yields to

$$\begin{aligned}\mathbb{E}\left[X_t^\top Q X_t \middle| \mathcal{F}_{k\epsilon}\right] &= \int_{k\epsilon}^{t} \mathbf{tr}\left(e^{A_\star^\top (t-s)} Q e^{A_\star (t-s)} C C^\top\right) \mathrm{d}s \\ + & \left(e^{A_\star (t-k\epsilon)} X_{k\epsilon} + \int_{k\epsilon}^{t} e^{A_\star (t-s)} \mathrm{d}s B_\star U_{k\epsilon}\right)^\top Q \left(e^{A_\star (t-k\epsilon)} X_{k\epsilon} + \int_{k\epsilon}^{t} e^{A_\star (t-s)} \mathrm{d}s B_\star U_{k\epsilon}\right).\end{aligned}$$

Plugging these results in the dynamic programming equation in (11), the expression in front of the minimum becomes the following quadratic function of $U_{k\epsilon}$:

$$\begin{aligned}& X_{k\epsilon}^\top \widetilde{Q} X_{k\epsilon} + 2 X_{k\epsilon}^\top \widetilde{G} U_{k\epsilon} + U_{k\epsilon}^\top \widetilde{R} U_{k\epsilon} \\ + & \left(\widetilde{A} X_{k\epsilon} + \widetilde{B} U_{k\epsilon}\right)^\top P_{k+1} \left(\widetilde{A} X_{k\epsilon} + \widetilde{B} U_{k\epsilon}\right) + \mathbf{tr}\left(\widetilde{P}_{k+1} C C^\top\right),\end{aligned}$$

where $P_{k+1} = Q\epsilon$, and

$$\begin{aligned}\widetilde{A} &= e^{A_\star \epsilon}, \\ \widetilde{B} &= \int_0^\epsilon e^{A_\star s} \mathrm{d}s B_\star, \\ \widetilde{Q} &= \int_0^\epsilon e^{A_\star^\top t} Q e^{A_\star t} \mathrm{d}t, \\ \widetilde{G} &= \int_0^\epsilon e^{A_\star^\top t} Q \left(\int_0^t e^{A_\star (t-s)} \mathrm{d}s\right) B_\star \mathrm{d}t, \\ \widetilde{R} &= R\epsilon + \int_0^\epsilon B_\star^\top \left(\int_0^t e^{A_\star^\top (t-s)} Q e^{A_\star (t-s)} \mathrm{d}s\right) B_\star \mathrm{d}t, \\ \widetilde{P}_{k+1} &= P_{k+1} \int_0^\epsilon e^{A_\star^\top s} e^{A_\star s} \mathrm{d}s + \int_0^\epsilon \left(\int_0^t e^{A_\star^\top s} Q e^{A_\star s} \mathrm{d}s\right) \mathrm{d}t.\end{aligned}$$

Note that in the last equation above, we used Ito Isometry (Baldi, 2017) to find $\widetilde{P}_{k+1}$. Now, performing the minimization the optimal control action is

$$U_{k\epsilon}^\star = -\left(\widetilde{B}^\top P_{k+1} \widetilde{B} + \widetilde{R}\right)^{-1} \left(\widetilde{B}^\top P_{k+1} \widetilde{A} + \widetilde{G}^\top\right) X_{k\epsilon},$$

and (11) leads to

$$\mathcal{V}_{k\epsilon}\left(X_{k\epsilon}\right) = X_{k\epsilon}^\top P_k X_{k\epsilon} + \mathbf{tr}\left(C C^\top \left[P_{k+1} \int_0^\epsilon e^{A_\star^\top s} e^{A_\star s} \mathrm{d}s + \int_0^\epsilon \left(\int_0^t e^{A_\star^\top s} Q e^{A_\star s} \mathrm{d}s\right) \mathrm{d}t\right]\right), \quad (13)$$

where $P_k$ is calculated according to the discrete time Riccati equation

$$P_k = \widetilde{Q} + \widetilde{A}^\top P_{k+1}\widetilde{A} - \left(\widetilde{G} + \widetilde{A}^\top P_{k+1}\widetilde{B}\right)\left(\widetilde{B}^\top P_{k+1}\widetilde{B} + \widetilde{R}\right)^{-1}\left(\widetilde{B}^\top P_{k+1}\widetilde{A} + \widetilde{G}^\top\right). \quad (14)$$

It is shown that if there is some matrix $L$ such that $\boldsymbol{\lambda}_{\max}\left(\widetilde{A} + \widetilde{B}L\right) < 1$, then as $k \to -\infty$, the matrix $P_k$ in the above discrete time Riccati equation converges to a uniquely existing matrix $P$ that solves the algebraic Riccati equation

$$P = \widetilde{Q} + \widetilde{A}^\top P\widetilde{A} - \left(\widetilde{G} + \widetilde{A}^\top P\widetilde{B}\right)\left(\widetilde{B}^\top P\widetilde{B} + \widetilde{R}\right)^{-1}\left(\widetilde{B}^\top P\widetilde{A} + \widetilde{G}^\top\right), \quad (15)$$

regardless of the terminal matrix for the largest value $k+1$, which here corresponds to $P_{T/\epsilon}$ (Chan et al., 1984; De Souza et al., 1986; Faradonbeh et al., 2018).

Next, we show that if $\epsilon$ is sufficiently small, then the matrix $L$ mentioned above exists. To that end, write

$$\begin{aligned}\widetilde{A} &= e^{A_\star\epsilon} = \sum_{n=0}^{\infty}\frac{A_\star^n\epsilon^n}{n!} = I_{d_X} + \epsilon M(\epsilon)A_\star,\\ \widetilde{B} &= \sum_{n=0}^{\infty}\int_0^\epsilon \frac{A_\star^n s^n}{n!}\mathrm{d}sB_\star = \sum_{n=0}^{\infty}\frac{A_\star^n\epsilon^{n+1}}{(n+1)!}B_\star = \epsilon M(\epsilon)B_\star,\end{aligned}$$

where

$$M(\epsilon) = \sum_{n=1}^{\infty}\frac{A_\star^{n-1}\epsilon^{n-1}}{n!} = I_{d_X} + \epsilon\sum_{n=2}^{\infty}\frac{A_\star^{n-1}\epsilon^{n-2}}{n!}.$$

Then, letting $L$ be as in Assumption 1, if $\epsilon$ is small enough, it holds that

$$\overline{\boldsymbol{\lambda}}\left(M(\epsilon)\left(A_\star + B_\star L\right)\right) < 0. \quad (16)$$

That is because the eigenvalues of the matrix $M(\epsilon)\left(A_\star + B_\star L\right)$ are continuous functions of $\epsilon$, and for $\epsilon = 0$ we have $\overline{\boldsymbol{\lambda}}\left(M(0)\left(A_\star + B_\star L\right)\right) = \overline{\boldsymbol{\lambda}}\left(\left(A_\star + B_\star L\right)\right) < 0$, according to Assumption 1. Hence, $\widetilde{A} + \widetilde{B}L = I_{d_X} + M(\epsilon)\left(A_\star + B_\star L\right)\epsilon$ implies that eigenvalues of $\widetilde{A} + \widetilde{B}L$ are exactly one plus the eigenvalues of $M(\epsilon)\left(A_\star + B_\star L\right)\epsilon$. So, it holds that

$$\boldsymbol{\lambda}_{\max}\left(\widetilde{A} + \widetilde{B}L\right)^2 \leq 1 + 2\epsilon\overline{\boldsymbol{\lambda}}\left(M(\epsilon)\left(A_\star + B_\star L\right)\right) + \boldsymbol{\lambda}_{\max}\left(M(\epsilon)\left(A_\star + B_\star L\right)\right)^2\epsilon^2. \quad (17)$$

Now, putting (16) and (17) together, if $\epsilon$ is small enough, then $\boldsymbol{\lambda}_{\max}\left(\widetilde{A} + \widetilde{B}L\right) < 1$. Henceforth, suppose that $\epsilon$ is sufficiently small so that the latter inequality holds true.

As long as $\epsilon > 0$ is small enough as described above, letting the time horizon $T$ tend to infinity, the $\epsilon$-length frozen optimal policy for minimizing the expected average cost is

$$U_{k\epsilon}^\star = -\left(\widetilde{B}^\top P\widetilde{B} + \widetilde{R}\right)^{-1}\left(\widetilde{B}^\top P\widetilde{A} + \widetilde{G}^\top\right)X_{k\epsilon}, \quad (18)$$

where $P$ is the unique solution of (15). On the other hand, for a fixed time horizon $T$, as $\epsilon$ shrinks the discrete-time Riccati equation in (14) becomes a continuous-time Riccati equation as follows. First, we have

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{\widetilde{A}-I_{d_X}}{\epsilon} &= A_\star,\\ \lim_{\epsilon\to 0}\frac{\widetilde{B}}{\epsilon} &= B_\star,\\ \lim_{\epsilon\to 0}\frac{\widetilde{Q}}{\epsilon} &= Q,\\ \lim_{\epsilon\to 0}\frac{\widetilde{G}}{\epsilon} &= 0,\\ \lim_{\epsilon\to 0}\frac{\widetilde{R}}{\epsilon} &= R.\end{aligned}$$

Using these limits, letting $\epsilon \to 0$ in (14) leads to

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{P_k-P_{k+1}}{\epsilon} &= \lim_{\epsilon\to 0}\frac{\widetilde{Q}}{\epsilon}+\lim_{\epsilon\to 0}\frac{\widetilde{A}^\top P_{k+1}\widetilde{A}-P_{k+1}}{\epsilon}\\ &- \lim_{\epsilon\to 0}\left(\frac{\widetilde{G}+\widetilde{A}^\top P_{k+1}\widetilde{B}}{\epsilon}\right)\left(\frac{\widetilde{B}^\top P_{k+1}\widetilde{B}+\widetilde{R}}{\epsilon}\right)^{-1}\left(\frac{\widetilde{B}^\top P_{k+1}\widetilde{A}+\widetilde{G}^\top}{\epsilon}\right)\\ &= Q+A_\star^\top P_{k+1}+P_{k+1}A_\star-P_{k+1}B_\star R^{-1}B_\star^\top P_{k+1}.\end{aligned}$$

That is, the backward differential equation

$$-\frac{\mathrm{d}P_t}{\mathrm{d}t}=\Phi_{A_\star,B_\star}\left(P\right), \tag{19}$$

with the terminal condition $P_T=0$. Thus, as $\epsilon\to 0$, the optimal policy becomes

$$U_t^\star=-R^{-1}B_\star^\top P_t X_t,$$

where $P_t$ is the solution of (19). Similarly, letting $\epsilon\to 0$ in (15), we get the optimal policy $U_t^\star=\mathcal{L}\left(A_\star,B_\star\right)X_t$ for minimizing the infinite horizon expected average cost, where

$$\mathcal{L}\left(A_\star,B_\star\right)=-R^{-1}B_\star^\top\mathcal{K}\left(A_\star,B_\star\right),$$

and $\mathcal{K}\left(A_\star,B_\star\right)$ solves $\Phi_{A_\star,B_\star}\left(P\right)=0$. Equivalently, letting $P_{t,T}$ be the solution of (19) when the time horizon is $T$, it holds that $\lim\limits_{T\to\infty} P_{0,T}=\mathcal{K}\left(A_\star,B_\star\right)$, where $\mathcal{K}\left(A_\star,B_\star\right)$ solves $\Phi_{A_\star,B_\star}\left(P\right)=0$. Note that all these relationships rely on the convergence of discrete time Riccati equation (14) to the algebraic Riccati equation (15), as $T\to\infty$.

Next, subtracting $\mathcal{V}_{(k+1)\epsilon}\left(X_{k\epsilon}\right)$ from both sides of (10), dividing by $\epsilon$, and letting $\epsilon\to 0$, Ito Isomery implies that

$$-\frac{\partial\mathcal{V}_t\left(X_t\right)}{\partial t}\mathrm{d}t=\min_{U_t}c_{X_t,U_t}\left(\boldsymbol{\pi}^\star\right)\mathrm{d}t+\mathbb{E}\left[\mathrm{d}X_t^\top\frac{\partial\mathcal{V}_t\left(X_t\right)}{\partial X_t}+\frac{1}{2}\mathrm{d}X_t^\top\frac{\partial^2\mathcal{V}_t\left(X_t\right)}{\partial X_t\partial X_t^\top}\mathrm{d}X_t\middle|\mathcal{F}_t\right],$$

where we used the limits of the matrices $\widetilde{Q}, \widetilde{G}, \widetilde{R}$ as $\epsilon \to 0$ to find the expression on the right-hand-side of the above equality. Note that the above partial derivatives exist according to (13) together with Dominated Convergence Theorem. Hence, substituting for $\mathrm{d}X_t$ from the dynamics (1), and leveraging Ito's Lemma, we obtain the Hamilton-Jacobi-Bellman (Yong and Zhou, 1999) equation

$$-\frac{\partial \mathcal{V}_t\left(X_t\right)}{\partial t} = \min_{U_t} c_{X_t,U_t}\left(\boldsymbol{\pi}^\star\right) + \frac{\partial \mathcal{V}_t\left(X_t\right)}{\partial X_t}^\top \left(A_\star X_t + B_\star U_t\right) + \frac{1}{2}\mathbf{tr}\left(\frac{\partial^2 \mathcal{V}_t\left(X_t\right)}{\partial X_t \partial X_t^\top} CC^\top\right). \tag{20}$$

Further, letting $\epsilon \to 0$, the expression in (13) gives

$$\mathcal{V}_t\left(X_t\right) = X_t^\top P_t X_t + \int_t^T \mathbf{tr}\left(CC^\top P_s\right)\mathrm{d}s, \tag{21}$$

where $P_t$ solve (19). This can be equivalently obtained using the fact that a quadratic function of the form $\mathcal{V}_t\left(X_t\right) = X_t^\top F_t X_t + \varphi_t$ solves the partial differential equation (20), as long as

$$\begin{aligned}-\frac{\mathrm{d}\varphi_t}{\mathrm{d}t} - X_t^\top \frac{\mathrm{d}F_t}{\mathrm{d}t} X_t &= \min_{U_t} X_t^\top Q X_t + U_t^\top R U_t \\ &+ 2X_t^\top F_t\left(A_\star X_t + B_\star U_t\right) + \mathbf{tr}\left(F_t CC^\top\right),\end{aligned}$$

which after solving for $U_t$ gives the optimal policy $U_t^\star = -R^{-1}B_\star^\top F_t X_t$, as well as

$$\begin{aligned}-\frac{\mathrm{d}\varphi_t}{\mathrm{d}t} - X_t^\top \frac{\mathrm{d}F_t}{\mathrm{d}t} X_t &= X_t^\top Q X_t + 2X_t^\top F_t\left(A_\star X_t\right) \\ &- X_t^\top F_t B_\star R^{-1} B_\star^\top F_t X_t + \mathbf{tr}\left(F_t CC^\top\right).\end{aligned}$$

Because the equation above needs to hold for an arbitrary $X_t$, it splits to

$$-\frac{\mathrm{d}F_t}{\mathrm{d}t} = \Phi_{A_\star,B_\star}\left(F_t\right), \frac{\mathrm{d}\varphi_t}{\mathrm{d}t} = -\mathbf{tr}\left(F_t CC^\top\right),$$

that is, $F_t$ solves (19). Further, note that cost-to-go at time $T$ is zero because time-to-go is zero, which provides the terminal condition $\mathcal{V}_T\left(X_T\right) = 0$, implying that $\varphi_t = \int_t^T \mathbf{tr}\left(CC^\top F_s\right)\mathrm{d}s$. Therefore, the solutions $F_t, \varphi_t$ of (20) lead to the same expression as in (21).

Finally, the expected average cost of the policy $U_t = \mathcal{L}\left(A_\star, B_\star\right) X_t$ is the limit of the expected average cost of the policy $U_t = -R^{-1}B_\star^\top P_{t,T} X_t$, as $T \to \infty$;

$$\begin{aligned}&\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\int_0^T c_s\left(\boldsymbol{\pi}^\star\right)\mathrm{d}s\right] \\ = \; &\limsup_{T\to\infty} \frac{1}{T}\int_0^T \mathbf{tr}\left(CC^\top P_{s,T}\right)\mathrm{d}s \\ = \; &\mathbf{tr}\left(CC^\top \lim_{T\to\infty} P_{s,T}\right) \\ = \; &\mathbf{tr}\left(CC^\top \mathcal{K}\left(A_\star, B_\star\right)\right).\end{aligned}$$

Moreover, suppose that $C = 0$, and apply the policy $U_t = \mathcal{L}(A_\star, B_\star) X_t$. Then, the state trajectory becomes $X_t = e^{D_\star t} X_0$, where $D_\star = A_\star + B_\star \mathcal{L}(A_\star, B_\star)$. So, by (21), we have

$$\begin{aligned} & X_0^\top \mathcal{K}(A_\star, B_\star) X_0 \\ = & \int_0^\infty X_t^\top \left(Q + \mathcal{L}(A_\star, B_\star)^\top R \mathcal{L}(A_\star, B_\star)\right) X_t \mathrm{d}t \\ = & X_0^\top \int_0^\infty e^{D_\star^\top t} \left(Q + \mathcal{L}(A_\star, B_\star)^\top R \mathcal{L}(A_\star, B_\star)\right) e^{D_\star t} \mathrm{d}t X_0, \end{aligned}$$

for an arbitrary initial state $X_0$. Thus, (6) holds:

$$\mathcal{K}(A_\star, B_\star) = \int_0^\infty e^{D_\star^\top t} \left(Q + \mathcal{L}(A_\star, B_\star)^\top R \mathcal{L}(A_\star, B_\star)\right) e^{D_\star t} \mathrm{d}t. \tag{22}$$

Since $Q$ is positive definite, the above equality implies that $\overline{\boldsymbol{\lambda}}(D_\star) < 0$, as well as

$$\begin{aligned} & D_\star^\top \mathcal{K}(A_\star, B_\star) + \mathcal{K}(A_\star, B_\star) D_\star \\ + & Q + \mathcal{L}(A_\star, B_\star)^\top R \mathcal{L}(A_\star, B_\star) = 0. \end{aligned} \tag{23}$$

So far, we have shown that by restricting our search for an optimal decision-making policy to the class of policies that the control action is frozen during intervals of length $\epsilon$, and then letting $\epsilon$ decay to vanish, we obtain optimal policies given by (19). Next, we show that these policies are optimal in the larger class of all control policies satisfying the information criteria at every time. That is, for all $t$, the control action $U_t$ can be determined using $\mathcal{F}_t = \sigma(X_{0:t}, U_{0:t})$. For this purpose, first note that the decision-making policy $U_t = R^{-1} B_\star^\top \mathcal{K}(A_\star, B_\star) X_t$ provides an upper-bound for the optimal expected average cost. That is,

$$\inf_{\boldsymbol{\pi}} \mathcal{J}_{\boldsymbol{\pi}} \leq \mathbf{tr}\left(CC^\top \mathcal{K}(A_\star, B_\star)\right).$$

Now, suppose that there is another policy, denoted by $\widetilde{\pi}$, that satisfies $\mathcal{J}_{\widetilde{\pi}} \leq \mathbf{tr}\left(CC^\top \mathcal{K}(A_\star, B_\star)\right)$. Define cost-to-go of the policy $\widetilde{\pi}$ by

$$\widetilde{\mathcal{V}}_t(X_t) = \mathbb{E}\left[\int_t^T c_s(\widetilde{\pi}) \mathrm{d}s \middle| \mathcal{F}_t\right],$$

where $T$ is large enough to satisfy $\widetilde{\mathcal{V}}_t(X_t) \leq 2X_t^\top \mathcal{K}(A_\star, B_\star) X_t + 2T\mathbf{tr}\left(CC^\top \mathcal{K}(A_\star, B_\star)\right)$, for all $0 \leq t \leq 1$. Note that such $T$ exists since $\widetilde{\pi}$ provides a smaller expected average cost than the policy $U_t = R^{-1} B_\star^\top \mathcal{K}(A_\star, B_\star) X_t$, and the desired upper-bound for $\widetilde{\mathcal{V}}_t(X_t)$ is $2\mathcal{V}_t(X_t)$; two times the cost-to-go of the policy $U_t = R^{-1} B_\star^\top \mathcal{K}(A_\star, B_\star) X_t$. Next, writing

$$\widetilde{\mathcal{V}}_t(X_t) = \mathbb{E}\left[\int_t^{t+\epsilon} c_{X_s, U_s}(\widetilde{\pi}) \mathrm{d}s + \widetilde{\mathcal{V}}_{t+\epsilon}(X_{t+\epsilon}) \middle| \mathcal{F}_t\right],$$

subtract $\widetilde{\mathcal{V}}_{t+\epsilon}\left(X_t\right)$ from both sides, and divide by $\epsilon$. Letting $\epsilon$ decay to zero, the upper-bound for $\widetilde{\mathcal{V}}_t\left(X_t\right)$ in terms of $\mathcal{V}_t\left(X_t\right)$ implies that according to Dominated Convergence Theorem, the following derivatives exist and it holds that

$$\begin{aligned} -\frac{\partial \widetilde{\mathcal{V}}_t\left(X_t\right)}{\partial t} &= c_{X_t,\widetilde{\pi}\left(\mathcal{F}_t\right)}\left(\boldsymbol{\pi}^\star\right) \\ + \quad & \frac{\partial \widetilde{\mathcal{V}}_t\left(X_t\right)^\top}{\partial X_t}\left(A_\star X_t + B_\star \widetilde{\pi}\left(\mathcal{F}_t\right)\right) + \frac{1}{2}\mathbf{tr}\left(\frac{\partial^2 \widetilde{\mathcal{V}}_t\left(X_t\right)}{\partial X_t \partial X_t^\top} C C^\top\right). \end{aligned}$$

Now, note that since $c_t\left(\boldsymbol{\pi}^\star\right)$ as well as $B_\star U_t$ are continuous functions of $U_t$, the above partial differential equation for $\widetilde{\mathcal{V}}_t\left(X_t\right)$ indicates that $\widetilde{\pi}\left(\mathcal{F}_t\right)$ is a continuous function of $X_t$. This, together with the fact that $W_t$ is an almost surely continuous function of time $t$, in lights of the dynamics equation in (1), leads to continuity of state trajectory $X_t$; i.e., $U_t = \widetilde{\pi}\left(\mathcal{F}_t\right)$ is continuous as $t$ varies. Thus, decision-making policies that freeze for $\epsilon$-length intervals provide accurate approximations of $U_t = \widetilde{\pi}\left(\mathcal{F}_t\right)$ in a sense that there exists a sequence $\left\{U_t^{(n)}\right\}_{n=1}^{\infty}$ such that $U_t^{(n)}$ freezes during intervals of the length $1/n$, and it holds that

$$\limsup_{n\to\infty}\limsup_{T\to\infty}\frac{1}{T}\int\limits_0^T \mathbb{E}\left[\left\|U_t^{(n)} - \widetilde{\pi}\left(\mathcal{F}_t\right)\right\|\right]\mathrm{d}t = 0.$$

Therefore, we have

$$\mathcal{J}_{\widetilde{\pi}} \geq \inf_{\epsilon>0}\inf \mathcal{J}_{\boldsymbol{\pi}} = \mathbf{tr}\left(\mathcal{K}\left(A_\star, B_\star\right) C C^\top\right),$$

where the inner infimum is taken over all policies that freeze during $\epsilon$-length intervals. This shows that the policy $U_t = -R^{-1}B_\star^\top \mathcal{K}\left(A_\star, B_\star\right) X_t$ is an optimal one, which completes the proof.

## Appendix B. Proof of Theorem 2 (Stability margin)

First, we study eigenvalues of the sum of two matrices. Suppose that $M, \Delta$ are arbitrary square matrices of the same size, and let $M = P^{-1} \Lambda P$ be the Jordan decomposition. That is, $\lambda_1, \cdots, \lambda_k$ are eigenvalues of $M$, $\Lambda \in \mathbb{C}^{d_X \times d_X}$ is a block diagonal matrix with blocks $\Lambda_1, \cdots, \Lambda_k$, and

$$\Lambda_i = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_i & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \lambda_i & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda_i \end{bmatrix} \in \mathbb{C}^{\boldsymbol{\mu}_i \times \boldsymbol{\mu}_i}. \tag{24}$$

Further, similar to Definition 2, let $\boldsymbol{\mu}_M = \max\limits_{1 \le i \le k} \boldsymbol{\mu}_i$. We prove that $\overline{\boldsymbol{\lambda}}\left(M - \Delta\right)$ is at most

$$\overline{\boldsymbol{\lambda}}\left(M\right) + \boldsymbol{\mu}_M^{1/2} \left|\!\left|\!\left| P \Delta P^{-1} \right|\!\right|\!\right| \vee \left( \boldsymbol{\mu}_M^{1/2} \left|\!\left|\!\left| P \Delta P^{-1} \right|\!\right|\!\right| \right)^{1/\boldsymbol{\mu}_M}. \tag{25}$$

To show the above inequality, first let $\lambda$ be an eigenvalue of $M - \Delta$ that satisfies $\Re\left(\lambda\right) > \overline{\boldsymbol{\lambda}}\left(M\right)$. So, $M - \lambda I$ is an invertible matirx, and there exists at least one vector $v$, such that $v \neq 0$ and $\left(M - \Delta - \lambda I\right) P^{-1} v = 0$. Then, $\left(M - \lambda I\right) P^{-1} v = \Delta P^{-1} v$ implies that

$$v = \left(\Lambda - \lambda I\right)^{-1} P \Delta P^{-1} v. \tag{26}$$

Because $\Lambda = \operatorname{diag}\left(\Lambda_1, \cdots, \Lambda_k\right)$, the matrix $\Lambda - \lambda I$ is block diagonal as well, and we have $\left(\Lambda - \lambda I\right)^{-1} = \operatorname{diag}\left( \left(\Lambda_1 - \lambda I_{\boldsymbol{\mu}_1}\right)^{-1}, \cdots, \left(\Lambda_k - \lambda I_{\boldsymbol{\mu}_k}\right)^{-1} \right)$. Further, it is straightforward to see that $\left(\Lambda_i - \lambda I_{\boldsymbol{\mu}_i}\right)^{-1}$ is

$$- \begin{bmatrix} \left(\lambda - \lambda_i\right)^{-1} & \left(\lambda - \lambda_i\right)^{-2} & \cdots & \left(\lambda - \lambda_i\right)^{-\boldsymbol{\mu}_i} \\ 0 & \left(\lambda - \lambda_i\right)^{-1} & \cdots & \left(\lambda - \lambda_i\right)^{-\boldsymbol{\mu}_i + 1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \left(\lambda - \lambda_i\right)^{-1} \end{bmatrix}.$$

Therefore, we have

$$\left|\!\left|\!\left| \left(\Lambda_i - \lambda I_{\boldsymbol{\mu}_i}\right)^{-1} \right|\!\right|\!\right| \le \boldsymbol{\mu}_i^{1/2} \left( \left|\lambda - \lambda_i\right| \wedge \left|\lambda - \lambda_i\right|^{\boldsymbol{\mu}_i} \right)^{-1}.$$

Using this bound for the operator norms of blocks of the block-diagonal matrix $\left(\Lambda - \lambda I\right)^{-1}$, since $\boldsymbol{\mu}_i \le \boldsymbol{\mu}_M$ and $\Re\left(\lambda\right) > \overline{\boldsymbol{\lambda}}\left(M\right)$, the equation in (26) implies

$$\begin{aligned} 1 \quad &\le \quad \left|\!\left|\!\left| \left(\Lambda - \lambda I\right)^{-1} P \Delta P^{-1} \right|\!\right|\!\right| \le \left|\!\left|\!\left| \left(\Lambda - \lambda I\right)^{-1} \right|\!\right|\!\right| \left|\!\left|\!\left| P \Delta P^{-1} \right|\!\right|\!\right| \\ &\le \quad \boldsymbol{\mu}_M^{1/2} \left|\!\left|\!\left| P \Delta P^{-1} \right|\!\right|\!\right| \left( \left(\Re(\lambda) - \overline{\boldsymbol{\lambda}}\left(M\right)\right) \wedge \left(\Re(\lambda) - \overline{\boldsymbol{\lambda}}\left(M\right)\right)^{\boldsymbol{\mu}_M} \right)^{-1}. \end{aligned}$$

So, letting $\lambda$ be an eigenvalue of $M - \Delta$ that satisfies $\Re(\lambda) = \overline{\boldsymbol{\lambda}}\left(M - \Delta\right)$, we obtain (25).

Now, using (25), we compare $A_\star + B_\star \mathcal{L}\left(A, B\right)$ and $D = A + B \mathcal{L}\left(A, B\right)$. Since $A_\star + B_\star \mathcal{L}\left(A, B\right) - D$ is

$$\Delta_\star = A_\star - A - \left(B_\star - B\right) R^{-1} B \mathcal{K}\left(A, B\right), \tag{27}$$

using (5), and letting $M = D$ in (25), we have

$$\overline{\boldsymbol{\lambda}}\left(A_\star + B_\star \mathcal{L}\left(A, B\right)\right) \le -\rho + \boldsymbol{\mu}_D^{1/2} |||P^{-1}||| \, |||P||| \, |||\Delta_\star||| \vee \left(\boldsymbol{\mu}_D^{1/2} |||P^{-1}||| \, |||P||| \, |||\Delta_\star|||\right)^{1/\boldsymbol{\mu}_D} .$$

So, in order to have $\overline{\boldsymbol{\lambda}}\left(A_\star + B_\star \mathcal{L}\left(A, B\right)\right) < -\delta$, it suffices to show that

$$\boldsymbol{\mu}_D^{1/2} |||P^{-1}||| \, |||P||| \, |||\Delta_\star||| < \rho - \delta \wedge \left(\rho - \delta\right)^{\boldsymbol{\mu}_D} . \tag{28}$$

However, since $|||\Delta_\star||| \le \mathcal{E}\left(A, B\right)\left(1 \vee \frac{|||B|||\zeta}{\boldsymbol{\lambda}_{\min}(R)}\right)$, (7) provides (28), which leads to the desired result.

### B.1. Proof of sufficiency of (8) for stabilization bounds

Next, we show that $\mathcal{E}\left(A, B\right) \le \epsilon_0$ is sufficient for stabilization and express uniform bounds for $\rho, \zeta$ in (5). Let $D_\star = A_\star + B_\star \mathcal{L}\left(A_\star, B_\star\right) = P_\star^{-1} \Lambda_\star P_\star$ be the Jordan decomposition as defined in the beginning of the proof, and define the largest block size $\boldsymbol{\mu}_\star = \boldsymbol{\mu}_{D_\star}$, similar to Definition 2. Further, suppose that the following is satisfied:

$$\epsilon_0 \le \frac{1}{1 \vee |||\mathcal{L}\left(A_\star, B_\star\right)|||} \left( \frac{\left(-\overline{\boldsymbol{\lambda}}\left(D_\star\right)\right) \wedge \left(-\overline{\boldsymbol{\lambda}}\left(D_\star\right)\right)^{\boldsymbol{\mu}_\star}}{\boldsymbol{\mu}_\star^{1/2} |||P_\star^{-1}||| \, |||P_\star|||} \wedge \left[ 4 \int\limits_0^\infty |||e^{D_\star t}|||^2 \mathrm{d}t \right]^{-1} \right) . \tag{29}$$

The inequality in (29) implies that if we write $D_1 = A + B\mathcal{L}\left(A_\star, B_\star\right) = A_\star + B_\star \mathcal{L}\left(A_\star, B_\star\right) + \Delta_1 = D_\star + \Delta_1$, then, the matrix $\Delta_1 = A - A_\star + \left(B - B_\star\right) \mathcal{L}\left(A_\star, B_\star\right)$ satisfies

$$|||\Delta_1||| < \frac{\left(-\overline{\boldsymbol{\lambda}}\left(D_\star\right)\right) \wedge \left(-\overline{\boldsymbol{\lambda}}\left(D_\star\right)\right)^{\boldsymbol{\mu}_\star}}{\boldsymbol{\mu}_\star^{1/2} |||P_\star||| \, |||P_\star^{-1}|||}.$$

So, taking $M = D_\star$, the bound in (25) implies that $\overline{\boldsymbol{\lambda}}\left(D_1\right) < 0$. Hence, we can employ Lemma 4 to study consequences of applying the linear feedback $\mathcal{L}\left(A_\star, B_\star\right)$ to a system of dynamics matrices $A, B$, and get

$$\mathcal{K}\left(A, B\right) \le M = \mathcal{K}\left(A, B\right) + \int\limits_0^\infty e^{D_1^\top t} F e^{D_1 t} \mathrm{d}t,$$

where

$$F = \left[\mathcal{L}\left(A_\star, B_\star\right) - \mathcal{L}\left(A, B\right)\right]^\top R \left[\mathcal{L}\left(A_\star, B_\star\right) - \mathcal{L}\left(A, B\right)\right].$$

Above, we used the fact that the initial state $X_0 = x$ in Lemma 4 is arbitrary, and so, the involved matrices are themselves equal. Further, similar to Lemma 4, it is straightforward to see that

$$M = \int\limits_0^\infty e^{D_1^\top t} \left[Q + \mathcal{L}\left(A_\star, B_\star\right)^\top R \mathcal{L}\left(A_\star, B_\star\right)\right] e^{D_1 t} \mathrm{d}t.$$

This leads to

$$\begin{aligned} & Q + \mathcal{L}\left(A_\star, B_\star\right)^\top R \mathcal{L}\left(A_\star, B_\star\right) = -D_1^\top M - M D_1 \\ = \; & -D_\star^\top M - M D_\star - \Delta_1^\top M - M \Delta_1. \end{aligned}$$

Because $\overline{\boldsymbol{\lambda}}\left(D_\star\right) < 0$, the latter equation and (6) provide

$$\begin{aligned} M &= \int_0^\infty e^{D_\star^\top t} \left[Q + \mathcal{L}\left(A_\star, B_\star\right)^\top R \mathcal{L}\left(A_\star, B_\star\right) + \Delta_1^\top M + M\Delta_1\right] e^{D_\star t} \mathrm{d}t \\ &= \int_0^\infty e^{D_\star^\top t} \left[Q + \mathcal{L}\left(A_\star, B_\star\right)^\top R \mathcal{L}\left(A_\star, B_\star\right)\right] e^{D_\star t} \mathrm{d}t + \int_0^\infty e^{D_\star^\top t} \left[\Delta_1^\top M + M\Delta_1\right] e^{D_\star t} \mathrm{d}t \\ &= \mathcal{K}\left(A_\star, B_\star\right) + \int_0^\infty e^{D_\star^\top t} \left[\Delta_1^\top M + M\Delta_1\right] e^{D_\star t} \mathrm{d}t. \end{aligned}$$

Therefore, it holds that $|||M||| \leq |||\mathcal{K}\left(A_\star, B_\star\right)||| + 2|||\Delta_1||| |||M||| \int_0^\infty |||e^{D_\star t}|||^2 \mathrm{d}t$, which, according to (29) and $\mathcal{K}\left(A, B\right) \leq M$, yields to

$$|||\mathcal{K}\left(A, B\right)||| \leq |||M||| \leq 2|||\mathcal{K}\left(A_\star, B_\star\right)|||. \tag{30}$$

To proceed, suppose that $v \in \mathbb{C}^{d_X}$ satisfies $\|v\| = 1$ and $Dv = \lambda v$. Now, (6) implies that

$$\begin{aligned} v^* \mathcal{K}\left(A, B\right) v &= \int_0^\infty v^* e^{D^\top t} \left[Q + \mathcal{L}\left(A, B\right)^\top R \mathcal{L}\left(A, B\right)\right] e^{Dt} v \mathrm{d}t \\ &= \int_0^\infty \left\| \left[Q + \mathcal{L}\left(A, B\right)^\top R \mathcal{L}\left(A, B\right)\right]^{\frac{1}{2}} e^{\lambda t} v \right\|^2 \mathrm{d}t, \end{aligned}$$

where $v^*$ is the transposed complex conjugate of $v$. Thus, maximizing the left-hand-side above while taking minimum on the right-hand-side, it holds that

$$|||\mathcal{K}\left(A, B\right)||| \geq \boldsymbol{\lambda}_{\min}\left(Q\right) \int_0^\infty e^{2\Re(\lambda) t} \mathrm{d}t \geq \frac{\boldsymbol{\lambda}_{\min}\left(Q\right)}{2\Re\left(-\lambda\right)}. \tag{31}$$

Putting (30) and (31) together, we obtain $\overline{\boldsymbol{\lambda}}\left(D\right) \leq -\boldsymbol{\lambda}_{\min}\left(Q\right)\left(4|||\mathcal{K}\left(A_\star, B_\star\right)|||\right)^{-1}$. This and (30) imply that $\mathcal{E}\left(A, B\right) \leq \epsilon_0$ is sufficient for (5), with $\rho = \boldsymbol{\lambda}_{\min}\left(Q\right) 4^{-1} |||\mathcal{K}\left(A_\star, B_\star\right)|||^{-1}$, $\zeta = 2|||\mathcal{K}\left(A_\star, B_\star\right)|||$. ■

## Appendix C. Proof of Theorem 3 (Regret analysis)

Let $M = Q + \mathcal{L}\left(A_\star, B_\star\right)^\top R \mathcal{L}\left(A_\star, B_\star\right)$. Recall that $\boldsymbol{\pi}$ applies $U_t = L_t X_t$ at time $t$. Now, for a given $T$, suppose that $\epsilon > 0$ is a fixed small real, and let $N = \lceil T/\epsilon \rceil$. Then, define the sequence of policies $\{\boldsymbol{\pi}_i\}_{i=0}^N$:

$$\boldsymbol{\pi}_i = \begin{cases} U_t = L_t X_t & t < i\epsilon \\ U_t = \mathcal{L}\left(A_\star, B_\star\right) X_t & t \geq i\epsilon \end{cases}.$$

Note that as long as one concerns about times $t \leq T$, it holds that $\boldsymbol{\pi}^\star = \boldsymbol{\pi}_0, \boldsymbol{\pi}_N = \boldsymbol{\pi}$. Clearly, since $\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}_0}(T) = 0$, we have $\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}(T) = \sum\limits_{i=0}^{N-1} \left(\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}_{i+1}}(T) - \boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}_i}(T)\right)$. Thus, Lemma 1 gives $\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}(T) = \sum\limits_{i=0}^{N-1} \left(X_{i\epsilon}^\top F_{i\epsilon} X_{i\epsilon} + 2X_{i\epsilon}^\top g_{i\epsilon} + \beta_{i\epsilon}\right)$, where the matrix $F_{i\epsilon}$, the vector $g_{i\epsilon}$, and the scalar $\beta_{i\epsilon}$ are defined in (41), (42), and (43), respectively. Now, letting $\epsilon \to 0$, since $L_t$ is piecewise continuous, we have

$$\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}(T) = \int\limits_0^T \left(X_t^\top \widetilde{F}_t X_t + 2X_t^\top \widetilde{g}_t + \widetilde{\beta}_t\right) \mathrm{d}t, \tag{32}$$

where $\widetilde{F}_t = \lim\limits_{\epsilon\to 0, i\epsilon \to t} \epsilon^{-1} F_{i\epsilon}$, $\widetilde{g}_t = \lim\limits_{\epsilon\to 0, i\epsilon \to t} \epsilon^{-1} g_{i\epsilon}$, and $\widetilde{\beta}_t = \lim\limits_{\epsilon\to 0, i\epsilon \to t} \epsilon^{-1} \beta_{i\epsilon}$. Note that the above limits exist, since $F_{i\epsilon}, g_{i\epsilon}, \beta_{i\epsilon}$ are continuous. To calculate $\widetilde{F}_t, \widetilde{g}_t, \widetilde{\beta}_t$, using Lemma 1 and the piecewise continuity of $L_t$, we obtain $\widetilde{\beta}_t = 0$,

$$\begin{aligned} \widetilde{F}_t &= S_t + 2H_t^\top \int\limits_t^T e^{D_\star^\top (s-t)} M e^{D_\star (s-t)} \mathrm{d}s, \\ \widetilde{g}_t &= \int\limits_t^T \left( H_t^\top e^{D_\star^\top (s-t)} M \int\limits_t^s e^{D_\star (s-u)} C \mathrm{d}W_u \right) \mathrm{d}s, \end{aligned}$$

where $S_t = L_t^\top R L_t - \mathcal{L}\left(A_\star, B_\star\right)^\top R \mathcal{L}\left(A_\star, B_\star\right)$, and

$$H_t = \lim_{\epsilon \to 0} \frac{e^{(A_\star + B_\star L_t)\epsilon} - e^{D_\star \epsilon}}{\epsilon} = B_\star \left(L_t - \mathcal{L}\left(A_\star, B_\star\right)\right).$$

Now, by (6) and $\int\limits_T^\infty e^{D_\star^\top (s-t)} M e^{D_\star (s-t)} \mathrm{d}s = E_{T-t}$, the expression for $\widetilde{F}_t$ becomes

$$S_t + H_t^\top \mathcal{K}\left(A_\star, B_\star\right) + \mathcal{K}\left(A_\star, B_\star\right) H_t - H_t^\top E_{T-t} - E_{T-t} H_t. \tag{33}$$

So, after doing some algebra (see (51)), we get

$$\begin{aligned} S_t &+ H_t^\top \mathcal{K}\left(A_\star, B_\star\right) + \mathcal{K}\left(A_\star, B_\star\right) H_t \\ &= \left(L_t - \mathcal{L}\left(A_\star, B_\star\right)\right)^\top R \left(L_t - \mathcal{L}\left(A_\star, B_\star\right)\right). \end{aligned} \tag{34}$$

Since $W_u$ has independent increments and in $\widetilde{g}_t$ we have $u \geq t$, Fubini's Theorem gives

$$\mathbb{E}\left[X_t^\top \widetilde{g}_t\right] = \mathbb{E}\left[\mathbb{E}\left[X_t^\top \widetilde{g}_t \middle| \sigma\left(W_{0:t}\right)\right]\right] = \mathbb{E}\left[X_t^\top \mathbb{E}\left[\widetilde{g}_t \middle| \sigma\left(W_{0:t}\right)\right]\right] = 0.$$

Hence, (32), (33), (34), and Fubini's Theorem imply that $\mathbb{E}\left[\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}\left(T\right)\right] = \mathbb{E}\left[\boldsymbol{\alpha}_T\right]$.

To proceed towards establishing the second result, apply Stochastic Fubini Theorem (Oksendal, 2013; Baldi, 2017) to get

$$\begin{aligned} \int_0^T X_t^\top \widetilde{g}_t \mathrm{d}t &= \int_0^T \int_t^T \int_t^s \left( X_t^\top H_t^\top e^{D_\star^\top (s-t)} M e^{D_\star (s-u)} C \right) \mathrm{d}W_u \mathrm{d}s \mathrm{d}t \\ &= \int_0^T \int_0^u \int_u^T \left( X_t^\top H_t^\top e^{D_\star^\top (s-t)} M e^{D_\star (s-u)} C \right) \mathrm{d}s \mathrm{d}t \mathrm{d}W_u = \int_0^T Y_u^\top \mathrm{d}W_u, \end{aligned}$$

where, using the expression for $H_t$, the vector $Y_u$ can be written as

$$Y_u^\top = \int_0^u \int_u^T \left( X_t^\top H_t^\top e^{D_\star^\top (s-t)} M e^{D_\star (s-u)} C \right) \mathrm{d}s \mathrm{d}t = \int_0^u \left( X_t^\top \left( L_t - \mathcal{L}\left(A_\star, B_\star\right)\right)^\top P_{t,u}^\top \right) \mathrm{d}t,$$

for $P_{t,u}^\top = \int_u^T B_\star^\top e^{D_\star^\top (s-t)} M e^{D_\star (s-u)} C \mathrm{d}s$. Now, letting $V_T = \int_0^T \|Y_u\|^2 \mathrm{d}u$, for $V_T < 1$, Ito Isometry (Baldi, 2017), and for $V_T \geq 1$, Lemma 2, imply that

$$\int_0^T Y_u^\top \mathrm{d}W_u = \mathcal{O}\left( d_W V_T^{1/2} \log^{1/2} V_T \right). \tag{35}$$

However, by using the triangle inequality and Fubini's Theorem, we obtain

$$\begin{aligned} V_T &\leq \int_0^T \int_0^u \left\| P_{t,u} \left( L_t - \mathcal{L}\left(A_\star, B_\star\right)\right) X_t \right\|^2 \mathrm{d}t \mathrm{d}u \\ &= \int_0^T \left( X_t^\top \left( L_t - \mathcal{L}\left(A_\star, B_\star\right)\right)^\top \left[ \int_t^T P_{t,u}^\top P_{t,u} \mathrm{d}u \right] \left( L_t - \mathcal{L}\left(A_\star, B_\star\right)\right) X_t \right) \mathrm{d}t \\ &\leq \int_0^T \boldsymbol{\lambda}_{\max} \left( \int_t^T R^{-1/2} P_{t,u}^\top P_{t,u} R^{-1/2} \mathrm{d}u \right) \left\| R^{1/2} \left( L_t - \mathcal{L}\left(A_\star, B_\star\right)\right) X_t \right\|^2 \mathrm{d}t. \end{aligned}$$

The second part of the integrand above appears in $\boldsymbol{\alpha}_T$. So, we proceed by finding an upper-bound for the first part. For this purpose, we use the triangle inequality and (6) to get the equation

$$\begin{aligned} \boldsymbol{\lambda}_{\max} \left( \int_t^T P_{t,u}^\top P_{t,u} \mathrm{d}u \right) &\leq \int_t^T \left|\!\left|\!\left| B_\star^\top e^{D_\star^\top (u-t)} \right|\!\right|\!\right|^2 \left|\!\left|\!\left| \int_u^T e^{D_\star^\top (s-u)} M e^{D_\star (s-u)} \mathrm{d}s \right|\!\right|\!\right|^2 \left|\!\left|\!\left| C \right|\!\right|\!\right|^2 \mathrm{d}u \\ &\leq \left|\!\left|\!\left| B_\star \right|\!\right|\!\right|^2 \left|\!\left|\!\left| \mathcal{K}\left(A_\star, B_\star\right) \right|\!\right|\!\right|^2 \left|\!\left|\!\left| C \right|\!\right|\!\right|^2 \int_0^\infty \left|\!\left|\!\left| e^{D_\star^\top u} \right|\!\right|\!\right|^2 \mathrm{d}u. \end{aligned}$$

Therefore, by using (31), we get

$$V_T \leq \frac{\|B_\star\|^2 \|\mathcal{K}\left(A_\star, B_\star\right)\|^3 \|C\|^2}{\boldsymbol{\lambda}_{\min}\left(Q\right)\boldsymbol{\lambda}_{\min}\left(R\right)}\boldsymbol{\alpha}_T,$$

since $E_t$ decays exponentially with $t$. So, (35) gives the desired result. ■

## Appendix D. Proof of Theorem 4 (Analysis of Algorithm 1)

In order to establish Theorem 4, we study the estimation procedure in (9) and specify the accuracy at which the algorithm is able to estimate $A_\star, B_\star$. To that end, Lemma 5 and Lemma 6 are utilized to study the Gram matrix $V_n = \int\limits_0^{\gamma^n} Y_s Y_s^\top \mathrm{d}s$ in (9), while Lemma 2 is used for bounding the estimation error. Then, by leveraging Lemma 3, we find the rates of deviating from the optimal policy in (3). Finally, the resulting regret of Algorithm 1 is investigated in lights of Theorem 3.

By using (1) to substitute for $\mathrm{d}X_t$, $\left[\int\limits_0^{\gamma^n} Y_s \mathrm{d}X_s^\top\right]^\top V_n^\dagger$ is

$$\left[\int\limits_0^{\gamma^n} Y_s Y_s^\top \left[A_\star, B_\star\right]^\top \mathrm{d}s + \int\limits_0^{\gamma^n} Y_s \mathrm{d}W_s^\top C^\top\right]^\top V_n^\dagger.$$

In (38), we show that $V_n$ is non-singular. So, we have

$$\left[\int\limits_0^{\gamma^n} Y_s \mathrm{d}X_s^\top\right]^\top V_n^{-1} = \left[A_\star, B_\star\right] + \left[V_n^{-1}\int\limits_0^{\gamma^n} Y_s \mathrm{d}W_s^\top C^\top\right]^\top. \tag{36}$$

Because $\left[A_\star, B_\star\right] \in \mathcal{S}_0$, (9) and (36) lead to

$$\mathcal{E}\left(A_n, B_n\right) \leq \left\|V_n^{-1}\int\limits_0^{\gamma^n} Y_s \mathrm{d}W_s^\top C^\top\right\| + \|\Theta_n\|.$$

Since entries of $\Theta_n$ are $\mathcal{N}\left(0, \gamma^{-n/2} n^{1/2}\right)$, we have

$$\mathbb{P}\left(\|\Theta_n\| \geq d_X^{1/2}\left(d_X + d_U\right)^{1/2}\gamma^{-n/4} n^{1/2}\right) = \mathcal{O}\left(e^{-n^{1/2}}\right).$$

This, by Borel-Cantelli Lemma, leads to

$$\|\Theta_n\| = \mathcal{O}\left(d_X^{1/2}\left(d_X + d_U\right)^{1/2}\gamma^{-n/4} n^{1/2}\right).$$

Thus, letting $d = d_X + d_U$, according to Lemma 2, $\mathcal{E}\left(A_n, B_n\right)$ is at most

$$d^{1/2}\mathcal{O}\left(d_W^{1/2}\|C\|\left(\frac{\log \boldsymbol{\lambda}_{\max}\left(V_n\right)}{\boldsymbol{\lambda}_{\min}\left(V_n\right)}\right)^{1/2} + d_X^{1/2}\gamma^{-n/4} n^{1/2}\right). \tag{37}$$

Now, Lemma 5 provides $\mathcal{O}\left(\log \boldsymbol{\lambda}_{\max}\left(V_n\right)\right) = n\log\gamma$. Further, we will shortly show that

$$\liminf_{n\to\infty} \gamma^{-n/2}\boldsymbol{\lambda}_{\min}\left(V_n\right) \geq \boldsymbol{\lambda}_{\min}\left(CC^\top\right). \tag{38}$$

Thus, (37) and (38) yield to the upper-bound

$$\mathcal{E}\left(A_n, B_n\right) = \mathcal{O}\left(d^{1/2}\left(d_X^{1/2} + \frac{d_W^{1/2}\|C\|\log^{1/2}\gamma}{\boldsymbol{\lambda}_{\min}\left(CC^\top\right)^{1/2}}\right)\gamma^{-n/4} n^{1/2}\right).$$

This gives the first result in Theorem 4. To prove the other statement, let $\beta_\star$ be as defined in Lemma 3. So, Lemma 3 implies that

$$\left|\!\left|\!\left|\mathcal{L}\left(A_n, B_n\right) - \mathcal{L}\left(A_\star, B_\star\right)\right|\!\right|\!\right|^2 = \mathcal{O}\left(\left(d_X + d_U\right)\beta_\star^2\left(d_X + \frac{d_W \left|\!\left|\!\left|C\right|\!\right|\!\right|^2 \log\gamma}{\boldsymbol{\lambda}_{\min}\left(CC^\top\right)}\right)\gamma^{-n/2} n\right).$$

Now, since during the time period $\gamma^{n-1} \leq t < \gamma^n$ the feedback matrix is frozen to $\mathcal{L}\left(A_{n-1}, B_{n-1}\right)$, according to Lemma 5, we have

$$\int\limits_0^{\gamma^n} \left\|R^{1/2}\left(L_t - \mathcal{L}\left(A_\star, B_\star\right)\right)X_t\right\|^2 \mathrm{d}t = \mathcal{O}\left(\sum_{k=1}^{n} \beta_L \gamma^{k-1}\gamma^{-(k-1)/2} k\right),$$

where

$$\beta_L = d\beta_\star^2\left(d_X + \frac{d_W \left|\!\left|\!\left|C\right|\!\right|\!\right|^2 \log\gamma}{\boldsymbol{\lambda}_{\min}\left(CC^\top\right)}\right)(\gamma - 1)\left|\!\left|\!\left|R\right|\!\right|\!\right|\left|\!\left|\!\left|C\right|\!\right|\!\right|^2.$$

Moreover, since by Theorem 1 we have $\overline{\boldsymbol{\lambda}}\left(D_\star\right) < 0$, the matrix $E_t$ in Theorem 3 decays exponentially with $t$. So, it holds that

$$\int\limits_0^{T} \left(X_t^\top E_{T-t} B_\star \left(L_t - \mathcal{L}\left(A_\star, B_\star\right)\right)X_t\right)\mathrm{d}t = \mathcal{O}\left(\log^2 T\right).$$

Therefore, according to Theorem 3, we have the following:

$$\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}}(T) = \mathcal{O}\left(\sum_{k=1}^{\lceil(\log T)/(\log\gamma)\rceil} \gamma^{(k-1)/2} k\right) = \mathcal{O}\left(\frac{\beta_L}{\log\gamma} T^{1/2}\log T\right).$$

This, according to $\beta_\star$ in Lemma 3, completes the proof.

To prove (38), let $D_{k-1} = A_\star + B_\star \mathcal{L}\left(A_{k-1}, B_{k-1}\right)$. Then, by Lemma 5, we have

$$\liminf_{k\to\infty} \gamma^{-k}\boldsymbol{\lambda}_{\min}\left(\int\limits_{\gamma^{k-1}}^{\gamma^k} X_t X_t^\top \mathrm{d}t\right) \geq \eta_k \boldsymbol{\lambda}_{\min}\left(CC^\top\right), \tag{39}$$

where $\eta_k = \left(1 - \gamma^{-1}\right)\left(\int\limits_0^1 \left|\!\left|\!\left|e^{-D_{k-1}s}\right|\!\right|\!\right|^2 \mathrm{d}s\right)^{-1}$. Hence, (39) implies that to establish (38), it suffices to show that the following inequality holds for some $0 \leq \ell < n-1$:

$$\liminf_{n\to\infty} \boldsymbol{\lambda}_{\min}\left(\sum_{k=\ell}^{n-1} \gamma^{k-n/2}\begin{bmatrix} I_{d_X} \\ \mathcal{L}\left(A_k, B_k\right)\end{bmatrix}\begin{bmatrix} I_{d_X} \\ \mathcal{L}\left(A_k, B_k\right)\end{bmatrix}^\top\right) \geq \max_{\ell \leq k \leq n-1} \frac{1}{\eta_k}. \tag{40}$$

For an arbitrary fixed $\epsilon > 0$, consider the event that the above-mentioned smallest eigenvalue is less than $\epsilon$, and let $\mathcal{M}_n(\epsilon)$ be the set of matrices $\left[A_k, B_k\right]_{k=\ell}^{n-1}$ for which this event occurs:

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

where the $(d_X + d_U) \times d_X(n - \ell)$ matrix $P_{\ell,n}$ is

$$\left[\gamma^{\frac{\ell}{2}-\frac{n}{4}} \begin{bmatrix} I_{d_X} \\ \mathcal{L}\left(A_\ell, B_\ell\right) \end{bmatrix}, \cdots, \gamma^{\frac{n-1}{2}-\frac{n}{4}} \begin{bmatrix} I_{d_X} \\ \mathcal{L}\left(A_{n-1}, B_{n-1}\right) \end{bmatrix}\right].$$

Now, note that the set of all matrices

$$F_n = \begin{bmatrix} \gamma^{\ell/2-n/4} I_{d_X} & \cdots & \gamma^{(n-1)/2-n/4} I_{d_X} \\ \gamma^{\ell/2-n/4} L_\ell & \cdots & \gamma^{(n-1)/2-n/4} L_{n-1} \end{bmatrix},$$

that there exists $v \in \mathbb{R}^{d_X+d_U}$ satisfying $\|v\| = 1$ and $F_n^\top v = 0$, is of dimension $d_X + d_U - 1 + (n - \ell)(d_U - 1)$. To show that, on one hand, the set of unit $d_X + d_U$ dimensional vectors is (a sphere) of dimension $d_X + d_U - 1$. On the other hand, by writing $v = \left[v_1^\top, v_2^\top\right]^\top$, for $v_1 \in \mathbb{R}^{d_X}$ and $v_2 \in \mathbb{R}^{d_U}$, clearly, $F_n^\top v = 0$ is equivalent to $L_k^\top v_2 = -v_1$, for all $k = \ell, \cdots, n-1$. The latter enforces every column of $L_k$ to be in a certain hyperplane in $\mathbb{R}^{d_U}$.

Thus, according to Lemma 6, the dimension of $\mathcal{M}_n(0)$ is at most $d_X + (d_U - 1)(n - \ell + 1) + (n - \ell)d_X^2$. Further, if $\ell$ is sufficiently large so that $\gamma^{-\ell+n/2}\epsilon < 1$, then for every $[A_k, B_k]_{k=\ell}^{n-1} \in \mathcal{M}_n(\epsilon)$, there exists some $\left[\widetilde{A}_k, \widetilde{B}_k\right]_{k=\ell}^{n-1} \in \mathcal{M}_n(0)$, such that for all $k = \ell, \cdots, n-1$, it holds that

$$\left|\!\left|\!\left| [A_k, B_k] - \left[\widetilde{A}_k, \widetilde{B}_k\right] \right|\!\right|\!\right| = \mathcal{O}\left(\gamma^{-k/2+n/4}\epsilon^{1/2}\right).$$

The random matrices $\{\Theta_k\}_{k=0}^{n-1}$ are independent, and entries of $\Theta_k$ are independent identically distributed $\mathcal{N}\left(0, \gamma^{-k/2}k^{1/2}\right)$ random variables. Hence, we have

$$\mathbb{P}(\mathcal{M}_n(\epsilon)) = \left[\mathcal{O}\left(\gamma^{\ell/4}\ell^{-1/4}\gamma^{-\ell/2+n/4}\epsilon^{1/2}\right) \wedge 1\right]^m,$$

where $m = (d_X d_U - d_U + 1)(n - \ell) - d_X - d_U + 1$. To see that, note that $\mathcal{M}_n(0)$ is a $d_X + (d_U - 1)(n - \ell + 1) + (n - \ell)d_X^2$ dimensional object in a $d_X(d_X + d_U)(n - \ell)$ dimensional space. So, the exponent is at least $m$. Letting $\ell = n - 5$, we have $m \geq 5$. Further, as $n$ grows, $\mathcal{O}\left(\ell^{-1/4}\gamma^{(n-\ell)/4}\epsilon^{1/2}\right) < 1$ holds for $\epsilon = \max\limits_{\ell \leq k \leq n-1} \eta_k{}^{-1}$. So, we have $\sum\limits_{n=5}^{\infty} \mathbb{P}(\mathcal{M}_n(\epsilon)) = \sum\limits_{n=5}^{\infty} \mathcal{O}\left(n^{-1/4}\right)^5 < \infty$, which by Borel-Cantelli Lemma implies (40). ■

## Appendix E. Estimation Rates under Persistent Randomization

**Proposition 1** *Assume that in Algorithm 1 the variance of entries of $\Theta_n$ is $\sigma_n^2$, where*

$$\liminf_{n\to\infty} \sigma_n > 0.$$

*Then, letting $\boldsymbol{\omega}_{\mathcal{E}}$ be as in Theorem 4, $Y_s = \left[X_s^\top, U_t^\top\right]^\top$, and $V_n = \int\limits_0^{\gamma^n} Y_s Y_s^\top \mathrm{d}s$, we have*

$$\left|\!\left|\!\left| \left( \int\limits_0^{\gamma^n} Y_s \mathrm{d}X_s^\top \right)^\top V_n^\dagger - [A_\star, B_\star] \right|\!\right|\!\right|^2 = \mathcal{O}\left(\boldsymbol{\omega}_{\mathcal{E}} \gamma^{-n} n^2\right).$$

**Proof** By (36), it suffices to study $\Delta = V_n^\dagger \int\limits_0^{\gamma^n} Y_s \mathrm{d}W_s^\top C^\top$. In the sequel, we show that

$$\liminf_{n\to\infty} n\gamma^{-n} \boldsymbol{\lambda}_{\min}\left(V_n\right) \geq \boldsymbol{\lambda}_{\min}\left(CC^\top\right).$$

So, putting Lemma 2 and Lemma 5 together, we obtain the desired result, since they give

$$|\!|\!|\Delta|\!|\!|^2 = \mathcal{O}\left( \left(d_X + d_U\right) d_W |\!|\!|C|\!|\!|^2 \frac{\gamma^{-n} n^2 \log \gamma}{\boldsymbol{\lambda}_{\min}\left(CC^\top\right)} \right).$$

Thus, by (39), it is enough to show that for some $0 \leq \ell < n-1$,

$$\liminf_{n\to\infty} \boldsymbol{\lambda}_{\min} \left( \sum_{k=\ell}^{n-1} \gamma^{k-n} n \begin{bmatrix} I_{d_X} \\ \mathcal{L}\left(A_k, B_k\right) \end{bmatrix} \begin{bmatrix} I_{d_X} \\ \mathcal{L}\left(A_k, B_k\right) \end{bmatrix}^\top \right)$$

is at least $\epsilon = \max\limits_{\ell \leq k \leq n-1} {\eta_k}^{-1}$. Let $\mathcal{M}_n(\epsilon)$ be the set of $[A_k, B_k]_{k=\ell}^{n-1}$ that the above does not hold: $\mathcal{M}_n(\epsilon) = \left\{ [A_\ell, B_\ell, \cdots, A_{n-1}, B_{n-1}] : \boldsymbol{\lambda}_{\min}\left(P_{\ell,n} P_{\ell,n}^\top\right) \leq \epsilon \right\}$, where $P_{\ell,n}$ is

$$\left[ \gamma^{\frac{\ell-n}{2}} n^{\frac{1}{2}} \begin{bmatrix} I_{d_X} \\ \mathcal{L}\left(A_\ell, B_\ell\right) \end{bmatrix}, \cdots, \gamma^{-\frac{1}{2}} n^{\frac{1}{2}} \begin{bmatrix} I_{d_X} \\ \mathcal{L}\left(A_{n-1}, B_{n-1}\right) \end{bmatrix} \right].$$

Similar to the proof of Theorem 4, for $[A_k, B_k]_{k=\ell}^{n-1} \in \mathcal{M}_n(\epsilon)$, there is $\left[\widetilde{A}_k, \widetilde{B}_k\right]_{k=\ell}^{n-1} \in \mathcal{M}_n(0)$, such that $\left|\!\left|\!\left| [A_k, B_k] - \left[\widetilde{A}_k, \widetilde{B}_k\right] \right|\!\right|\!\right|^2 = \mathcal{O}\left(\gamma^{n-k} n^{-1} \epsilon\right)$. Thus, $\liminf\limits_{n\to\infty} \sigma_n > 0$, together with the dimension of $\mathcal{M}_n(0)$ that we calculated in the proof of Theorem 4, leads to

$$\mathbb{P}(\mathcal{M}_n(\epsilon)) = \left[ \mathcal{O}\left( \gamma^{(n-\ell)/2} n^{-1/2} \epsilon^{1/2} \right) \wedge 1 \right]^m,$$

for $m = \left(d_X d_U - d_U + 1\right)(n - \ell) - d_X - d_U + 1$. Finally, $\ell = n - 4$ gives $\sum\limits_{n=4}^{\infty} \mathbb{P}(\mathcal{M}_n(\epsilon)) < \infty$. Therefore, Borel-Cantelli Lemma implies the desired result. ■

## Appendix F. Auxiliary Lemmas

In this section, we state the auxiliary lemmas used for establishing the main results and provide their proofs, each subsection corresponding to one lemma.

First, in Lemma 1 in Subsection F.1, we provide expressions for the difference between the regrets of two policies. Study of self-normalized stochastic integrals is the content of Lemma 2, while Lemma 3 on Lipschitz continuity of the optimal feedback with respect to the dynamics matrices is established in Subsection F.3.

Next, in Lemma 4, we consider the total cumulative cost for the case of applying a sub-optimal time-invariant linear feedback policy to a deterministic system. Then, Lemma 5 focuses on explicit calculation of the empirical covariance matrix of the state vectors. Finally, in Lemma 6 in Subsection F.6 we specify the set of dynamics matrices that possess the same optimal linear feedback matrix.

### F.1. Difference in regrets of two policies

**Lemma 1** *For fixed* $0 \leq t_1 \leq t_2 \leq T$*, define the policies* $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2$ *according to*

$$\boldsymbol{\pi}_i = \begin{cases} U_t = L X_t & t < t_i \\ U_t = \mathcal{L}\left(A_\star, B_\star\right) X_t & t \geq t_i \end{cases}.$$

*Further, let* $D_\star = A_\star + B_\star \mathcal{L}\left(A_\star, B_\star\right)$*,* $D = A_\star + B_\star L$*,* $M_\star = Q + \mathcal{L}\left(A_\star, B_\star\right)^\top R \mathcal{L}\left(A_\star, B_\star\right)$*,* $M = Q + LRL$*,* $\Delta_t = e^{D(t-t_1)} - e^{D_\star(t-t_1)}$*,* $Z_t = \int\limits_{t_1}^{t} \left[e^{D(t-s)} - e^{D_\star(t-s)}\right] C \mathrm{d}W_s$*, and* $S = M - M_\star = L^\top R L - \mathcal{L}\left(A_\star, B_\star\right)^\top R \mathcal{L}\left(A_\star, B_\star\right)$*.*

*Then, we have* $\boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}_2}\left(T\right) - \boldsymbol{\mathcal{R}}_{\boldsymbol{\pi}_1}\left(T\right) = X_{t_1}^\top F_{t_1} X_{t_1} + 2 X_{t_1}^\top g_{t_1} + \beta_{t_1}$*, where* $F_{t_1}, g_{t_1}$*, and* $\beta_{t_1}$ *are*

$$\begin{aligned} F_{t_1} &= \int\limits_{t_1}^{t_2} \left(e^{D_\star^\top(t-t_1)} S e^{D_\star(t-t_1)} + 2\Delta_t^\top M e^{D_\star(t-t_1)} + \Delta_t^\top M \Delta_t\right) \mathrm{d}t \\ &+ \int\limits_{t_2}^{T} \left(2\Delta_{t_2}^\top e^{D_\star^\top(t-t_2)} M_\star e^{D_\star(t-t_1)} + \Delta_{t_2}^\top e^{D_\star^\top(t-t_2)} M_\star e^{D_\star(t-t_2)} \Delta_{t_2}\right) \mathrm{d}t, \end{aligned} \tag{41}$$

$$\begin{aligned} g_{t_1} &= \int\limits_{t_1}^{t_2} \left(S \int\limits_{t_1}^{t} e^{D_\star(t-s)} C \mathrm{d}W_s + \Delta_t^\top M \int\limits_{t_1}^{t} e^{D_\star(t-s)} C \mathrm{d}W_s + e^{D_\star^\top(t-t_1)} M Z_t\right) \mathrm{d}t \\ &+ \int\limits_{t_1}^{t_2} \Delta_t^\top M Z_t \mathrm{d}t + \int\limits_{t_2}^{T} \Delta_{t_2}^\top e^{D_\star^\top(t-t_2)} M_\star \left(e^{D_\star(t-t_2)} Z_{t_2} + \int\limits_{t_1}^{t} e^{D_\star(t-s)} C \mathrm{d}W_s\right) \mathrm{d}t \\ &+ \int\limits_{t_2}^{T} \left(e^{D_\star^\top(t-t_1)} M_\star e^{D_\star(t-t_2)} Z_{t_2}\right) \mathrm{d}t, \end{aligned} \tag{42}$$

$$\begin{aligned}\beta_{t_1} \;&=\; \int\limits_{t_1}^{t_2} \left( \left\| S^{1/2} \int\limits_{t_1}^{t} e^{D_\star(t-s)} C \mathrm{d}W_s \right\|^2 + 2 Z_t^\top M \int\limits_{t_1}^{t} e^{D_\star(t-s)} C \mathrm{d}W_s + Z_t^\top M Z_t \right) \mathrm{d}t \\ &+\; \int\limits_{t_2}^{T} \left( 2 Z_{t_2}^\top e^{D_\star^\top (t-t_2)} M_\star \int\limits_{t_1}^{t} e^{D_\star(t-s)} C \mathrm{d}W_s + Z_{t_2}^\top e^{D_\star^\top (t-t_2)} M_\star e^{D_\star(t-t_2)} Z_{t_2} \right) \mathrm{d}t. \quad (43)\end{aligned}$$

**Proof** Letting $X_t^{\boldsymbol{\pi}_i}$ be the state of the system under the policy $\boldsymbol{\pi}_i$, clearly, for $t \leq t_1$, it holds that $X_t^{\boldsymbol{\pi}_1} = X_t^{\boldsymbol{\pi}_2}$. So, we use $X_{t_1}$ for both states at time $t_1$. Moreover, for $t_1 \leq t \leq t_2$, we have

$$\begin{aligned} X_t^{\boldsymbol{\pi}_1} \;&=\; e^{D_\star(t-t_1)} X_{t_1} + \int\limits_{t_1}^{t} e^{D_\star(t-s)} C \mathrm{d}W_s, \\ X_t^{\boldsymbol{\pi}_2} \;&=\; e^{D(t-t_1)} X_{t_1} + \int\limits_{t_1}^{t} e^{D(t-s)} C \mathrm{d}W_s, \end{aligned}$$

where $Y_t = X_t^{\boldsymbol{\pi}_2} - X_t^{\boldsymbol{\pi}_1}$. So, by denoting the instantaneous cost of policy $\boldsymbol{\pi}_i$ at time $t$ by $c_t\left(\boldsymbol{\pi}_i\right)$, we get $Y_t = \Delta_t X_{t_1} + Z_t$, as well as

$$\begin{aligned} \int\limits_{t_1}^{t_2} \left(c_t\left(\boldsymbol{\pi}_2\right) - c_t\left(\boldsymbol{\pi}_1\right)\right) \mathrm{d}t \;&=\; \int\limits_{t_1}^{t_2} \left[ \left(X_t^{\boldsymbol{\pi}_1} + Y_t\right)^\top M \left(X_t^{\boldsymbol{\pi}_1} + Y_t\right) - X_t^{\boldsymbol{\pi}_1 \top} M_\star X_t^{\boldsymbol{\pi}_1} \right] \mathrm{d}t \\ &=\; \int\limits_{t_1}^{t_2} \left[ X_t^{\boldsymbol{\pi}_1 \top} S X_t^{\boldsymbol{\pi}_1} + 2 Y_t^\top M X_t^{\boldsymbol{\pi}_1} + Y_t^\top M Y_t \right] \mathrm{d}t. \qquad (44) \end{aligned}$$

On the other hand, for $t \geq t_2$, we have

$$\begin{aligned} \int\limits_{t_2}^{T} \left(c_t\left(\boldsymbol{\pi}_2\right) - c_t\left(\boldsymbol{\pi}_1\right)\right) \mathrm{d}t \;&=\; \int\limits_{t_2}^{T} \left[ \left(X_t^{\boldsymbol{\pi}_1} + Y_t\right)^\top M_\star \left(X_t^{\boldsymbol{\pi}_1} + Y_t\right) - X_t^{\boldsymbol{\pi}_1 \top} M_\star X_t^{\boldsymbol{\pi}_1} \right] \mathrm{d}t \\ &=\; \int\limits_{t_2}^{T} \left[ 2 Y_t^\top M_\star X_t^{\boldsymbol{\pi}_1} + Y_t^\top M_\star Y_t \right] \mathrm{d}t. \qquad (45) \end{aligned}$$

and

$$\begin{aligned} X_t^{\boldsymbol{\pi}_i} \;&=\; e^{D_\star(t-t_2)} X_{t_2}^{\boldsymbol{\pi}_i} + \int\limits_{t_2}^{t} e^{D_\star(t-s)} C \mathrm{d}W_s, \\ Y_t \;&=\; e^{D_\star(t-t_2)} \left[ X_{t_2}^{\boldsymbol{\pi}_2} - X_{t_2}^{\boldsymbol{\pi}_1} \right] = e^{D_\star(t-t_2)} \left[ \Delta_{t_2} X_{t_1} + Z_{t_2} \right]. \end{aligned}$$

Thus, putting (44) and (45) together, we obtain the desired results. ■

### F.2. Upper-bounding comparative ratios of stochastic integrals

**Lemma 2** *Suppose that $Y_t \in \mathbb{R}^m$ is a vector-valued stochastic process such that $Y_t$ is $\mathcal{F}_t$-measurable for the natural filtration $\mathcal{F}_t = \sigma\left(\{W_s\}_{0 \leq s \leq t}\right)$. Then, letting $V_t = \int\limits_0^t Y_s Y_s^\top \mathrm{d}s$, we have*

$$\left|\!\left|\!\left| \left(I + V_t\right)^{-1/2} \int\limits_0^t Y_s \mathrm{d}W_s^\top \right|\!\right|\!\right|^2 = \mathcal{O}\left(m d_W \log \boldsymbol{\lambda}_{\max}\left(V_t\right)\right).$$

**Proof** First, fix $t > 0$, and for an arbitrary $\epsilon > 0$, let $n = \lfloor t/\epsilon \rfloor$. So, for $k = 0, 1, \cdots, n$, consider the sequence of matrices $M_k = \epsilon^{-1} I + \sum\limits_{i=0}^{k} Y_{i\epsilon} Y_{i\epsilon}^\top$. Then, for $k = 1, \cdots, n$, consider the sequence of scalars $\beta_k$ defined according to $\beta_k = Y_{k\epsilon}^\top M_{k-1}^{-1} Y_{k\epsilon}$. Using the formula for determinants of the products of matrices, we have

$$\det M_k = \det\left[M_{k-1}\left(I + M_{k-1}^{-1} Y_{k\epsilon} Y_{k\epsilon}^\top\right)\right] = \det\left(M_{k-1}\right) \det\left(I + M_{k-1}^{-1} Y_{k\epsilon} Y_{k\epsilon}^\top\right).$$

Since all eigenvalues of $I + M_{k-1}^{-1} Y_{k\epsilon} Y_{k\epsilon}^\top$ are unit, except one of them which is $1 + \beta_k$, we have $(1 + \beta_k) \det M_{k-1} = \det M_k$. On the other hand, matrix inversion formula gives

$$M_k^{-1} = M_{k-1}^{-1} - \frac{1}{1 + Y_{k\epsilon}^\top M_{k-1}^{-1} Y_{k\epsilon}} M_{k-1}^{-1} Y_{k\epsilon} Y_{k\epsilon}^\top M_{k-1}^{-1},$$

which leads to

$$Y_{k\epsilon}^\top M_k^{-1} Y_{k\epsilon} = Y_{k\epsilon}^\top \left(M_{k-1} + Y_{k\epsilon} Y_{k\epsilon}^\top\right)^{-1} Y_{k\epsilon} = \beta_k - \frac{\beta_k^2}{1 + \beta_k} = 1 - \frac{1}{1 + \beta_k} = 1 - \frac{\det M_{k-1}}{\det M_k}.$$

Further, by using the inequality $1 - \beta \leq -\log \beta$ for $\beta > 0$, the latter equality gives

$$Y_{k\epsilon}^\top M_k^{-1} Y_{k\epsilon} \leq \log\det M_k - \log\det M_{k-1}. \tag{46}$$

Now, let $F_k = \sum\limits_{i=0}^{k} Y_{i\epsilon}\left(W_{(i+1)\epsilon} - W_{i\epsilon}\right)^\top$. Using the facts that $Y_{k\epsilon}, F_{k-1}$, and $M_k$ all are $\mathcal{F}_{k\epsilon}$-measurable, the Brownian motion $W_t$ has independent increments, and its covariance matrix is a multiple of identity, properties of conditional expectations give

$$\begin{aligned}
&\mathbb{E}\left[F_k^\top M_k^{-1} F_k\right] = \mathbb{E}\left[\mathbb{E}\left[F_k^\top M_k^{-1} F_k \middle| \mathcal{F}_{k\epsilon}\right]\right] \\
= \;& \mathbb{E}\left[\mathbb{E}\left[\left(F_{k-1} + Y_{k\epsilon}\left(W_{(k+1)\epsilon} - W_{k\epsilon}\right)^\top\right)^\top M_k^{-1}\left(F_{k-1} + Y_{k\epsilon}\left(W_{(k+1)\epsilon} - W_{k\epsilon}\right)^\top\right) \middle| \mathcal{F}_{k\epsilon}\right]\right] \\
= \;& \mathbb{E}\left[F_{k-1}^\top M_k^{-1} F_{k-1} + \mathbb{E}\left[\left(W_{(k+1)\epsilon} - W_{k\epsilon}\right) Y_{k\epsilon}^\top M_k^{-1} Y_{k\epsilon}\left(W_{(k+1)\epsilon} - W_{k\epsilon}\right)^\top \middle| \mathcal{F}_{k\epsilon}\right]\right] \\
= \;& \mathbb{E}\left[F_{k-1}^\top M_k^{-1} F_{k-1} + \left(Y_{k\epsilon}^\top M_k^{-1} Y_{k\epsilon}\right)\epsilon I\right].
\end{aligned}$$

So, using (46) together with the fact that (as positive semidefinite matrices) the order $M_{k-1} \leq M_k$ holds, we get the telescopic relationships

$$\boldsymbol{\lambda}_{\max}\left(\mathbb{E}\left[F_k^\top M_k^{-1} F_k\right]\right) - \boldsymbol{\lambda}_{\max}\left(\mathbb{E}\left[F_{k-1}^\top M_{k-1}^{-1} F_{k-1}\right]\right) \leq \epsilon\left(\log\frac{\det\left(\epsilon M_k\right)}{\det\left(\epsilon M_{k-1}\right)}\right).$$

Since $F_k^\top (M_k)^{-1} F_k$ is positive semidefinite, its trace is larger than its largest eigenvalue. Hence, adding up for $k = 0, 1, \cdots, n$, by interchanging trace and expectation, we obtain

$$\mathbb{E}\left[\boldsymbol{\lambda}_{\max}\left(F_n^\top (M_n)^{-1} F_n\right)\right] \leq \mathbb{E}\left[\mathbf{tr}\left(F_n^\top (M_n)^{-1} F_n\right)\right] \leq d_W \boldsymbol{\lambda}_{\max}\left(\mathbb{E}\left[F_n^\top (M_n)^{-1} F_n\right]\right),$$

which, by $\epsilon M_0 \geq I$, leads to

$$\mathbb{E}\left[\boldsymbol{\lambda}_{\max}\left(F_n^\top (\epsilon M_n)^{-1} F_n\right)\right] \leq m d_W \log \boldsymbol{\lambda}_{\max}(\epsilon M_n).$$

Thus, according to Doob's Martingale Convergence Theorem (Oksendal, 2013; Baldi, 2017), we have

$$\left|\!\left|\!\left| (\epsilon M_n)^{-1/2} F_n \right|\!\right|\!\right|^2 = \mathcal{O}\left(m d_W \log \boldsymbol{\lambda}_{\max}(\epsilon M_n)\right).$$

Finally, letting $\epsilon \to 0$, we obtain the desired result, because $\epsilon M_n, F_n$ are $\epsilon$-approximations of the corresponding integrals. ■

### F.3. Lipschitz continuity of optimal feedback

**Lemma 3** *Using the Jordan decomposition $D_\star = A_\star + B_\star \mathcal{L}(A_\star, B_\star) = P_\star^{-1} \Lambda_\star P_\star$, define $\boldsymbol{\mu}_\star = \boldsymbol{\mu}_{D_\star}$, similar to Definition 2, and suppose that $\mathcal{E}(A, B) \leq \kappa_\star$, for*

$$\kappa_\star = \frac{1}{1 \vee |\!|\!|\mathcal{L}(A_\star, B_\star)|\!|\!|}\left(\frac{\left(-\overline{\boldsymbol{\lambda}}(D_\star)\right) \wedge \left(-\overline{\boldsymbol{\lambda}}(D_\star)\right)^{\boldsymbol{\mu}_\star}}{\boldsymbol{\mu}_\star^{1/2} |\!|\!|P_\star^{-1}|\!|\!| |\!|\!|P_\star|\!|\!|} \wedge \left[4 \int_0^\infty |\!|\!|e^{D_\star t}|\!|\!|^2 \mathrm{d}t\right]^{-1}\right).$$

*Then, letting*

$$\beta_\star = \frac{2 |\!|\!|\mathcal{K}(A_\star, B_\star)|\!|\!|}{\boldsymbol{\lambda}_{\min}(R)}\left[1 + \frac{4 |\!|\!|B_\star|\!|\!|}{\boldsymbol{\lambda}_{\min}(Q)} |\!|\!|\mathcal{K}(A_\star, B_\star)|\!|\!| \left(1 \vee \frac{2\left(|\!|\!|B_\star|\!|\!| + \kappa_\star\right) |\!|\!|\mathcal{K}(A_\star, B_\star)|\!|\!|}{\boldsymbol{\lambda}_{\min}(R)}\right)\right],$$

*we have*

$$|\!|\!|\mathcal{L}(A, B) - \mathcal{L}(A_\star, B_\star)|\!|\!| \leq \beta_\star \mathcal{E}(A, B).$$

*In general, without the condition $\mathcal{E}(A, B) \leq \kappa_\star$, the constant $\beta_\star$ is replaced with*

$$\beta \quad = \quad \frac{|\!|\!|\mathcal{K}(A, B)|\!|\!|}{\boldsymbol{\lambda}_{\min}(R)} + \frac{2 |\!|\!|B_\star|\!|\!| |\!|\!|\mathcal{K}(A_0, B_0)|\!|\!|^2}{\boldsymbol{\lambda}_{\min}(Q) \boldsymbol{\lambda}_{\min}(R)}\left(1 \vee \frac{\left(|\!|\!|B_\star|\!|\!| + \mathcal{E}(A, B)\right) |\!|\!|\mathcal{K}(A_0, B_0)|\!|\!|}{\boldsymbol{\lambda}_{\min}(R)}\right),$$

*for some convex combination $[A_0, B_0] = \eta [A, B] + (1 - \eta) [A_\star, B_\star]$, and $0 \leq \eta \leq 1$.*

**Proof** Fix the matrices $A, B$, and consider the matrix-valued curve

$$\varphi = \left\{(1 - \eta) [A_\star, B_\star] + \eta [A, B]\right\}_{0 \leq \eta \leq 1}.$$

For an arbitrary $A_0, B_0 \in \varphi$, we find the derivative of the matrix $\mathcal{K}(A_0, B_0)$ at $A_0, B_0$, assuming that the matrices $A_0, B_0$ vary along $\varphi$. For this purpose, letting $\Delta_A = A - A_\star$, $\Delta_B = B - B_\star$, we first calculate $\mathcal{K}\left(A_1, B_1\right)$ for $A_1 = A_0 + \eta\Delta_A, B_1 = B_0 + \eta\Delta_B$, and then let $\eta \to 0$. First, letting $P = \mathcal{K}\left(A_1, B_1\right) - \mathcal{K}\left(A_0, B_0\right)$, we get

$$\begin{aligned}
& \mathcal{K}\left(A_0, B_0\right) B_1 R^{-1} B_1^\top \mathcal{K}\left(A_0, B_0\right) \\
= \; & \eta \mathcal{K}\left(A_0, B_0\right) \Delta_B R^{-1} B_1^\top \mathcal{K}\left(A_0, B_0\right) + \mathcal{K}\left(A_0, B_0\right) B_0 R^{-1} B_1^\top \mathcal{K}\left(A_0, B_0\right) \\
= \; & \eta^2 \mathcal{K}\left(A_0, B_0\right) \Delta_B R^{-1} \Delta_B^\top \mathcal{K}\left(A_0, B_0\right) + \eta \mathcal{K}\left(A_0, B_0\right) \Delta_B R^{-1} B_0^\top \mathcal{K}\left(A_0, B_0\right) \\
+ \; & \eta \mathcal{K}\left(A_0, B_0\right) B_0 R^{-1} \Delta_B^\top \mathcal{K}\left(A_0, B_0\right) \\
+ \; & \mathcal{K}\left(A_0, B_0\right) B_0 R^{-1} B_0^\top \mathcal{K}\left(A_0, B_0\right).
\end{aligned}$$

The above expression, because of

$$\begin{aligned}
& \mathcal{K}\left(A_1, B_1\right) B_1 R^{-1} B_1^\top \mathcal{K}\left(A_1, B_1\right) \\
= \; & \mathcal{K}\left(A_1, B_1\right) B_1 R^{-1} B_1^\top P + \mathcal{K}\left(A_1, B_1\right) B_1 R^{-1} B_1^\top \mathcal{K}\left(A_0, B_0\right) \\
= \; & P B_1 R^{-1} B_1^\top P + \mathcal{K}\left(A_0, B_0\right) B_1 R^{-1} B_1^\top P \\
+ \; & P B_1 R^{-1} B_1^\top \mathcal{K}\left(A_0, B_0\right) + \mathcal{K}\left(A_0, B_0\right) B_1 R^{-1} B_1^\top \mathcal{K}\left(A_0, B_0\right),
\end{aligned}$$

implies that the followings hold true:

$$\begin{aligned}
& \mathcal{K}\left(A_1, B_1\right) B_1 R^{-1} B_1^\top \mathcal{K}\left(A_1, B_1\right) - \mathcal{K}\left(A_0, B_0\right) B_0 R^{-1} B_0^\top \mathcal{K}\left(A_0, B_0\right) \\
= \; & P B_1 R^{-1} B_1^\top P + \mathcal{K}\left(A_0, B_0\right) B_1 R^{-1} B_1^\top P + P B_1 R^{-1} B_1^\top \mathcal{K}\left(A_0, B_0\right) \\
+ \; & \eta^2 \mathcal{K}\left(A_0, B_0\right) \Delta_B R^{-1} \Delta_B^\top \mathcal{K}\left(A_0, B_0\right) + \eta \mathcal{K}\left(A_0, B_0\right) \Delta_B R^{-1} B_0^\top \mathcal{K}\left(A_0, B_0\right) \\
+ \; & \eta \mathcal{K}\left(A_0, B_0\right) B_0 R^{-1} \Delta_B^\top \mathcal{K}\left(A_0, B_0\right).
\end{aligned} \tag{47}$$

By plugging (47) and

$$\begin{aligned}
& A_1^\top \mathcal{K}\left(A_1, B_1\right) + \mathcal{K}\left(A_1, B_1\right) A_1 = A_1^\top \mathcal{K}\left(A_0, B_0\right) + A_1^\top P + \mathcal{K}\left(A_0, B_0\right) A_1 + P A_1 \\
= \; & A_0^\top \mathcal{K}\left(A_0, B_0\right) + \eta \Delta_A^\top \mathcal{K}\left(A_0, B_0\right) + A_1^\top P + \mathcal{K}\left(A_0, B_0\right) A_0 + \eta \mathcal{K}\left(A_0, B_0\right) \Delta_A + P A_1,
\end{aligned}$$

in $\Phi_{A_i, B_i}\left(\mathcal{K}\left(A_i, B_i\right)\right) = 0$ for $i = 0, 1$, we obtain

$$\begin{aligned}
0 \; = \; & \left[A_1^\top - \mathcal{K}\left(A_0, B_0\right) B_1 R^{-1} B_1^\top\right] P + P \left[A_1 - B_1 R^{-1} B_1^\top \mathcal{K}\left(A_0, B_0\right)\right] - P B_1 R^{-1} B_1^\top P \\
+ \; & \eta \Delta_A^\top \mathcal{K}\left(A_0, B_0\right) + \eta \mathcal{K}\left(A_0, B_0\right) \Delta_A - \eta^2 \mathcal{K}\left(A_0, B_0\right) \Delta_B R^{-1} \Delta_B^\top \mathcal{K}\left(A_0, B_0\right) \\
- \; & \eta \mathcal{K}\left(A_0, B_0\right) \Delta_B R^{-1} B_0^\top \mathcal{K}\left(A_0, B_0\right) - \eta \mathcal{K}\left(A_0, B_0\right) B_0 R^{-1} \Delta_B^\top \mathcal{K}\left(A_0, B_0\right),
\end{aligned}$$

or equivalently,

$$0 = \widetilde{A}^\top P + P \widetilde{A} - P B_1 R^{-1} B_1^\top P + \widetilde{Q}, \tag{48}$$

for $\widetilde{A} = A_1 - B_1 R^{-1} B_1^\top \mathcal{K}\left(A_0, B_0\right)$, and

$$\begin{aligned}
\widetilde{Q} \; = \; & \eta \Delta_A^\top \mathcal{K}\left(A_0, B_0\right) + \eta \mathcal{K}\left(A_0, B_0\right) \Delta_A - \eta^2 \mathcal{K}\left(A_0, B_0\right) \Delta_B R^{-1} \Delta_B^\top \mathcal{K}\left(A_0, B_0\right) \\
= \; & \eta \mathcal{K}\left(A_0, B_0\right) \left[\Delta_A + \Delta_B \mathcal{L}\left(A_0, B_0\right)\right] + \eta \left[\mathcal{L}\left(A_0, B_0\right)^\top \Delta_B^\top + \Delta_A^\top\right] \mathcal{K}\left(A_0, B_0\right) \\
- \; & \eta^2 \mathcal{K}\left(A_0, B_0\right) \Delta_B R^{-1} \Delta_B^\top \mathcal{K}\left(A_0, B_0\right).
\end{aligned}$$

Suppose that $\eta$ is sufficiently small so that $\overline{\boldsymbol{\lambda}}\left(\widetilde{A}\right) < 0$. Note that it is possible thanks to stabilizability of $A_0, B_0$, Theorem 1, and $\lim\limits_{\eta \to 0} \widetilde{A} = A_0 + B_0 \mathcal{L}\left(A_0, B_0\right) = D_0$. So, since $P B_1 R^{-1} B_1^\top P$ is a positive semidefinite matrix, (48) implies that

$$\begin{aligned} P &= \int\limits_0^\infty e^{\widetilde{A}^\top t} \left(-P B_1 R^{-1} B_1^\top P + \widetilde{Q}\right) e^{\widetilde{A} t} \mathrm{d}t \\ &\leq \int\limits_0^\infty e^{\widetilde{A}^\top t} \widetilde{Q} e^{\widetilde{A} t} \mathrm{d}t \leq \left( \left\lVert\!\left| \widetilde{Q} \right|\!\right\rVert \int\limits_0^\infty \left\lVert\!\left| e^{\widetilde{A} t} \right|\!\right\rVert^2 \mathrm{d}t \right) I_{d_X}. \end{aligned}$$

This, because of $\lim\limits_{\eta \to 0} \widetilde{Q} = 0$, leads to $\lim\limits_{\eta \to 0} P = 0$. Thus, letting $M = \Delta_A + \Delta_B \mathcal{L}\left(A_0, B_0\right)$ and $\Delta_{\mathcal{K}(A_0,B_0)} = \lim\limits_{\eta \to 0} \eta^{-1} P$, (48) gives the following for $\Delta_{\mathcal{K}(A_0,B_0)}$:

$$\int\limits_0^\infty e^{D_0^\top t} \left( \mathcal{K}\left(A_0, B_0\right) M + M^\top \mathcal{K}\left(A_0, B_0\right) \right) e^{D_0 t} \mathrm{d}t. \tag{49}$$

By

$$\mathcal{K}\left(A, B\right) - \mathcal{K}\left(A_\star, B_\star\right) = \int\limits_0^1 \Delta_{(1-\eta)[A_\star, B_\star] + \eta[A, B]} \mathrm{d}\eta,$$

(31), (49), and the Cauchy-Schwarz inequality provide

$$\begin{aligned} & \left\lVert\!\left| \mathcal{K}\left(A, B\right) - \mathcal{K}\left(A_\star, B_\star\right) \right|\!\right\rVert \\ \leq\; & \mathcal{E}\left(A, B\right) \sup_{[A_0, B_0] \in \varphi} 2 \left\lVert\!\left| \mathcal{K}\left(A_0, B_0\right) \right|\!\right\rVert \left(1 \vee \left\lVert\!\left| \mathcal{L}\left(A_0, B_0\right) \right|\!\right\rVert\right) \int\limits_0^\infty \left\lVert\!\left| e^{D_0 t} \right|\!\right\rVert^2 \mathrm{d}t \\ \leq\; & \mathcal{E}\left(A, B\right) \frac{2}{\boldsymbol{\lambda}_{\min}\left(Q\right)} \sup_{[A_0, B_0] \in \varphi} \left\lVert\!\left| \mathcal{K}\left(A_0, B_0\right) \right|\!\right\rVert^2 \left(1 \vee \left\lVert\!\left| \mathcal{L}\left(A_0, B_0\right) \right|\!\right\rVert\right). \end{aligned}$$

Next, note that $\mathcal{E}\left(A, B\right) \leq \kappa_\star$, together with (29) and (30), implies that

$$\left\lVert\!\left| \mathcal{K}\left(A, B\right) - \mathcal{K}\left(A_\star, B_\star\right) \right|\!\right\rVert \leq \mathcal{E}\left(A, B\right) \frac{8 \left\lVert\!\left| \mathcal{K}\left(A_\star, B_\star\right) \right|\!\right\rVert^2}{\boldsymbol{\lambda}_{\min}\left(Q\right)} \left( 1 \vee \frac{2 \left( \left\lVert\!\left| B_\star \right|\!\right\rVert + \kappa_\star \right) \left\lVert\!\left| \mathcal{K}\left(A_\star, B_\star\right) \right|\!\right\rVert}{\boldsymbol{\lambda}_{\min}\left(R\right)} \right).$$

Therefore, using (30), and putting the above inequality together with

$$\begin{aligned} & \left\lVert\!\left| \mathcal{L}\left(A, B\right) - \mathcal{L}\left(A_\star, B_\star\right) \right|\!\right\rVert \\ =\; & \left\lVert\!\left| R^{-1} \left[ \left(B_\star - B_1\right) \mathcal{K}\left(A, B\right) + B_\star \left( \mathcal{K}\left(A_\star, B_\star\right) - \mathcal{K}\left(A, B\right) \right) \right] \right|\!\right\rVert \\ \leq\; & \left\lVert\!\left| R^{-1} \right|\!\right\rVert \left[ \left\lVert\!\left| B_\star - B_1 \right|\!\right\rVert \left\lVert\!\left| \mathcal{K}\left(A, B\right) \right|\!\right\rVert + \left\lVert\!\left| B_\star \right|\!\right\rVert \left\lVert\!\left| \mathcal{K}\left(A_\star, B_\star\right) - \mathcal{K}\left(A, B\right) \right|\!\right\rVert \right], \end{aligned}$$

we get the first desired result. To establish the second result, it suffices to let $A_0, B_0$ be the one for which the above supremum over $\varphi$ is achieved. ■

## F.4. Effects of sub-optimal linear feedback policies

**Lemma 4** *Consider a noiseless linear dynamical system with the stabilizable dynamics matrices* $A, B$*. That is,* $\mathrm{d}X_t = (AX_t + BU_t)\,\mathrm{d}t$*, starting from* $X_0 = x$*. Then, if we apply the linear feedback*

$$\boldsymbol{\pi}: \quad U_t = LX_t,$$

*as long as* $\overline{\boldsymbol{\lambda}}\left(A + BL\right) < 0$*, it holds that*

$$\int_0^\infty c_t\left(\boldsymbol{\pi}\right)\mathrm{d}t = x^\top \mathcal{K}\left(A, B\right) x + \int_0^\infty \left\| R^{1/2}\left(L - \mathcal{L}\left(A, B\right)\right) e^{(A+BL)t} x \right\|^2 \mathrm{d}t.$$

**Proof** Denote $D_1 = A + B\mathcal{L}\left(A, B\right)$ and $D_2 = A + BL$. So, the dynamics equation $\mathrm{d}X_t = (AX_t + BLX_t)\,\mathrm{d}t$ implies that $X_t = e^{D_2 t}x$, which leads to

$$\begin{aligned}\int_0^\infty c_t\left(\boldsymbol{\pi}\right)\mathrm{d}t &= \int_0^\infty X_t^\top \left(Q + L^\top R L\right) X_t \mathrm{d}t \\ &= \int_0^\infty x^\top e^{D_2^\top t}\left(Q + L^\top R L\right) e^{D_2 t} x \mathrm{d}t = x^\top P x,\end{aligned}$$

where

$$\begin{aligned} P &= \int_0^\infty e^{D_2^\top t}\left(Q + L^\top R L\right) e^{D_2 t}\mathrm{d}t \\ &= \int_0^\epsilon e^{D_2^\top t}\left(Q + L^\top R L\right) e^{D_2 t}\mathrm{d}t \\ &+ e^{D_2^\top \epsilon}\left(\int_0^\infty e^{D_2^\top t}\left(Q + L^\top R L\right) e^{D_2 t}\mathrm{d}t\right) e^{D_2 \epsilon} \\ &= \int_0^\epsilon e^{D_2^\top t}\left(Q + L^\top R L\right) e^{D_2 t}\mathrm{d}t + e^{D_2^\top \epsilon} P e^{D_2 \epsilon},\end{aligned}$$

which yields to

$$\begin{aligned} Q + L^\top R L &= \lim_{\epsilon \to 0} \frac{1}{\epsilon}\int_0^\epsilon e^{D_2^\top t}\left(Q + L^\top R L\right) e^{D_2 t}\mathrm{d}t \\ &= \lim_{\epsilon \to 0} \frac{1}{\epsilon}\left[P - e^{D_2^\top \epsilon}P + e^{D_2^\top \epsilon}P - e^{D_2^\top \epsilon} P e^{D_2 \epsilon}\right] \\ &= -D_2^\top P - P D_2.\end{aligned}$$

Similar to (23), it holds that $D_1^\top \mathcal{K}(A,B) + \mathcal{K}(A,B) D_1 + Q + \mathcal{L}(A,B)^\top R \mathcal{L}(A,B)$. So, subtracting the latter two equalities, we get

$$\begin{aligned} & (D_2 - D_1)^\top \mathcal{K}(A,B) + \mathcal{K}(A,B)(D_2 - D_1) \\ + \quad & D_2^\top (P - \mathcal{K}(A,B)) + (P - \mathcal{K}(A,B)) D_2 + S = 0, \end{aligned} \tag{50}$$

where

$$S = L^\top R L - \mathcal{L}(A,B)^\top R \mathcal{L}(A,B).$$

Because $\overline{\boldsymbol{\lambda}}(D_2) < 0$, solving (50) for $P - \mathcal{K}(A,B)$, and using the fact $D_2 - D_1 = B[L - \mathcal{L}(A,B)]$, we have

$$P - \mathcal{K}(A,B) = \int\limits_0^\infty e^{D_2^\top t} F e^{D_2 t} \mathrm{d}t,$$

where

$$F = S + [L - \mathcal{L}(A,B)]^\top B^\top \mathcal{K}(A,B) + \mathcal{K}(A,B) B [L - \mathcal{L}(A,B)].$$

Then, using $B^\top \mathcal{K}(A,B) = -R\mathcal{L}(A,B)$, after doing some algebra we obtain

$$\begin{aligned} S \quad + \quad & [L - \mathcal{L}(A,B)]^\top B^\top \mathcal{K}(A,B) + \mathcal{K}(A,B) B [L - \mathcal{L}(A,B)] \\ = \quad & [L - \mathcal{L}(A,B)]^\top R [L - \mathcal{L}(A,B)]. \end{aligned} \tag{51}$$

Thus, $P - \mathcal{K}(A,B)$ is

$$\int\limits_0^\infty e^{D_2^\top t} [L - \mathcal{L}(A,B)]^\top R [L - \mathcal{L}(A,B)] e^{D_2 t} \mathrm{d}t,$$

which implies the desired result. ■

### F.5. Convergence of empirical covariance matrix of the state vectors

**Lemma 5** *Suppose that for $t \geq \gamma$, the linear feedback $L$ is applied to the system (1) such that $\overline{\boldsymbol{\lambda}}(D) < 0$, where $D = A_\star + B_\star L$. Then, we have*

$$\lim_{T \to \infty} \frac{1}{T} \int\limits_\gamma^{\gamma+T} X_t X_t^\top \mathrm{d}t = \int\limits_0^\infty e^{Ds} C C^\top e^{D^\top s} \mathrm{d}s.$$

**Proof** First, denote

$$V_T = \frac{1}{T} \int\limits_\gamma^{\gamma+T} X_t X_t^\top \mathrm{d}t.$$

Then, define the matrix $Y_t = X_t X_t^\top$, and apply Ito's Formula (Baldi, 2017) to find $\mathrm{d}Y_t$:

$$\mathrm{d}Y_t = \mathrm{d}X_t X_t^\top + X_t \mathrm{d}X_t^\top + \mathrm{d}X_t \mathrm{d}X_t^\top.$$

Plugging in for $\mathrm{d}X_t$ from (1), we obtain

$$\begin{aligned}\mathrm{d}Y_t &= \left(DX_t\mathrm{d}t + C\mathrm{d}W_t\right)X_t^\top \\ &+ X_t\left(DX_t\mathrm{d}t + C\mathrm{d}W_t\right)^\top + CC^\top\mathrm{d}t,\end{aligned}$$

where we used the facts $\mathrm{d}t\mathrm{d}t = 0$, $\mathrm{d}W_t\mathrm{d}t = 0$, and Ito Isometry $\mathrm{d}W_t\mathrm{d}W_t^\top = \mathrm{d}tI_{d_W}$ (Baldi, 2017). Thus, we have

$$Y_{\gamma+T} - Y_\gamma = \int\limits_\gamma^{\gamma+T} \mathrm{d}Y_t\mathrm{d}t \;=\; \int\limits_\gamma^{\gamma+T}\left(DX_tX_t^\top + X_tX_t^\top D^\top + CC^\top\right)\mathrm{d}t + TM_{\gamma,T},$$

where

$$M_{\gamma,T} = \frac{1}{T}\int\limits_\gamma^{\gamma+T} X_t\mathrm{d}W_t^\top C^\top + \frac{1}{T}\left(\int\limits_\gamma^{\gamma+T} X_t\mathrm{d}W_t^\top C^\top\right)^\top.$$

This can equivalently be written as

$$\frac{1}{T}\left(X_{\gamma+T}X_{\gamma+T}^\top - X_\gamma X_\gamma^\top\right) \;=\; DV_T + V_TD^\top + CC^\top + M_{\gamma,T}.$$

Since $\overline{\boldsymbol{\lambda}}(D) < 0$, the latter equality implies that $V_T$ is

$$\int\limits_0^\infty e^{Ds}\left(CC^\top + M_{\gamma,T} + \frac{1}{T}X_\gamma X_\gamma^\top - \frac{1}{T}X_{\gamma+T}X_{\gamma+T}^\top\right)e^{D^\top s}\mathrm{d}s.$$

Now, according to the following statements, the above leads to the desired result, because the terms corresponding to $M_{\gamma,T}, X_\gamma, X_{\gamma+T}$ vanish as $T$ grows.

1. Clearly, it holds that $\lim\limits_{T\to\infty} T^{-1/2}\|X_\gamma\| = 0$.

2. Since $\overline{\boldsymbol{\lambda}}(D) < 0$, the expression

$$X_{\gamma+T} = e^{DT}X_\gamma + \int\limits_\gamma^{\gamma+T} e^{D(\gamma+T-s)}C\mathrm{d}W_s$$

   implies that $\lim\limits_{T\to\infty} T^{-1/2}\|X_{\gamma+T}\| = 0$.

3. Putting $\overline{\boldsymbol{\lambda}}(D) < 0$ together with Doob's Martingale Convergence Theorem (Oksendal, 2013; Baldi, 2017), we get $\lim\limits_{T\to\infty} M_{\gamma,T} = 0$.

■

### F.6. Manifolds of dynamical systems with equal optimal feedback matrices

**Lemma 6** *Consider the set of dynamics matrices $A, B$ that share optimal feedback with $A_0, B_0$:*

$$\mathcal{M}_0 = \left\{ [A, B] \in \mathbb{R}^{d_X \times (d_X + d_U)} : \mathcal{L}(A, B) = \mathcal{L}(A_0, B_0) \right\}.$$

*Then, $\mathcal{M}_0$ is a manifold of dimension $d_X^2$.*

**Proof** Suppose that for the matrix $[A, B] = [A_0, B_0] + \epsilon [M, N]$, it holds that $\mathcal{L}(A, B) = \mathcal{L}(A_0, B_0)$. We find the derivative of $\mathcal{L}(A_0, B_0)$ along the direction $[M, N]$. First, using the expressions in (23) for $A, B$ and for $A_0, B_0$, we get

$$\begin{aligned}
& (D_0 + \epsilon M + \epsilon N \mathcal{L}(A_0, B_0))^\top \mathcal{K}(A, B) \\
+ \; & \mathcal{K}(A, B)(D_0 + \epsilon M + \epsilon N \mathcal{L}(A_0, B_0)) \\
= \; & -Q - \mathcal{L}(A_0, B_0)^\top R \mathcal{L}(A_0, B_0) \\
= \; & D_0^\top \mathcal{K}(A_0, B_0) + \mathcal{K}(A_0, B_0) D_0,
\end{aligned}$$

where $D_0 = A_0 + B_0 \mathcal{L}(A_0, B_0)$. Simplifying the above expressions and letting $\epsilon \to 0$, for the matrix

$$\Delta = \lim_{\epsilon \to 0} \epsilon^{-1} (\mathcal{K}(A, B) - \mathcal{K}(A_0, B_0)),$$

we have

$$\begin{aligned}
& D_0^\top \Delta + \Delta D_0 + (M + N \mathcal{L}(A_0, B_0))^\top \mathcal{K}(A_0, B_0) \\
+ \; & \mathcal{K}(A_0, B_0)(M + N \mathcal{L}(A_0, B_0)) = 0.
\end{aligned}$$

Thus, since according to Theorem 1, $\overline{\boldsymbol{\lambda}}(D_0) < 0$, it yields to

$$\Delta = \int_0^\infty e^{D_0^\top t} F e^{D_0 t} \mathrm{d}t,$$

where

$$\begin{aligned}
F \; = \; & (M + N \mathcal{L}(A_0, B_0))^\top \mathcal{K}(A_0, B_0) \\
+ \; & \mathcal{K}(A_0, B_0)(M + N \mathcal{L}(A_0, B_0)).
\end{aligned}$$

On the other hand, $\mathcal{L}(A, B) = -R^{-1} B^\top \mathcal{K}(A, B)$ gives

$$\begin{aligned}
0 \; = \; & \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left( B^\top \mathcal{K}(A, B) - B_0^\top \mathcal{K}(A_0, B_0) \right) \\
= \; & \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[ \left( B^\top - B_0^\top \right) \mathcal{K}(A, B) \right. \\
- \; & \left. B_0^\top (\mathcal{K}(A_0, B_0) - \mathcal{K}(A, B)) \right] \\
= \; & N^\top \mathcal{K}(A_0, B_0) + B_0^\top \Delta.
\end{aligned}$$

So, $\mathcal{M}_0$ is a manifold, and its tangent space consists of matrices $M, N$ satisfying the above equation. To find the dimension, select a $d_X \times d_X$ matrix $P$ arbitrarily, and let $N$ be

$$N = -\mathcal{K}\left(A_0, B_0\right)^{-1} \int_0^\infty e^{D_0^\top t} \left[P^\top \mathcal{K}\left(A_0, B_0\right) + \mathcal{K}\left(A_0, B_0\right) P\right] e^{D_0 t} B_0 \mathrm{d}t = 0. \tag{52}$$

Note that since $\boldsymbol{\lambda}_{\min}\left(Q\right) > 0$, the inverse $\mathcal{K}\left(A_0, B_0\right)^{-1}$ exists. Then, solve for $M$ according to $M + N\mathcal{L}\left(A_0, B_0\right) = P$. Therefore, the matrices $M, N$ satisfy in $N^\top \mathcal{K}\left(A_0, B_0\right) + B_0^\top \Delta = 0$, and so correspond to a member of $\mathcal{M}_0$. Conversely, every matrices $M, N$ in the tangent space of $\mathcal{M}_0$ provide a $d_X \times d_X$ matrix $P = M + N\mathcal{L}\left(A_0, B_0\right)$ such that $N^\top \mathcal{K}\left(A_0, B_0\right) + B_0^\top \Delta = 0$. Thus, $\mathcal{M}_0$ is of dimension $d_X^2$, which is the desired result. ■