Thompson Sampling Efficiently Learns to Control Diffusion Processes

Anonymous Author(s)
Affiliation
Address
email

Abstract

Diffusion processes that evolve according to linear stochastic differential equations are an important family of continuous-time dynamic decision-making models. Optimal policies are well-studied for them, under full certainty about the drift matrices. However, little is known about data-driven control of diffusion processes with uncertain drift matrices as conventional discrete-time analysis techniques are not applicable. In addition, while the task can be viewed as a reinforcement learning problem involving exploration and exploitation trade-off, ensuring system stability is a fundamental component of designing optimal policies. We establish that the popular Thompson sampling algorithm learns optimal actions fast, incurring only a square-root of time regret, and also stabilizes the system in a short time period. To the best of our knowledge, this is the first such result for Thompson sampling in a diffusion process control problem. We validate our theoretical results through empirical simulations with real parameter matrices from two settings of airplane and blood glucose control. Moreover, we observe that Thompson sampling significantly improves (worst-case) regret, compared to the state-of-the-art algorithms, suggesting Thompson sampling explores in a more guarded fashion. Our theoretical analysis involves characterization of a certain *optimality manifold* that ties the local geometry of the drift parameters to the optimal control of the diffusion process. We expect this technique to be of broader interest.

1 Introduction

One of the most natural reinforcement learning (RL) algorithms for controlling a diffusion process with unknown parameters is based on Thompson sampling (TS) [1]: a Bayesian posterior for the model is calculated based on its time evolution, and a control policy is then designed by treating a sampled model from the posterior as the truth. Despite its simplicity, guaranteeing efficiency and whether sampling the actions from the posterior could lead to unbounded future trajectories is unknown. In fact, the only known such theoretical result for control of a diffusion process is for an epsilon-greedy type policy that requires selecting purely random actions at a certain rate [2].

In this work, we consider a p dimensional state signal $\{\boldsymbol{x}_t\}_{t\geq 0}$ that obeys the (Ito) stochastic differential equation (SDE)

$$\boldsymbol{dx}_t = (A_0\boldsymbol{x}_t + B_0\boldsymbol{u}_t)\,\boldsymbol{d}t + \boldsymbol{d}\mathbb{W}_t\,, \tag{1}$$

where the *drift matrices* A_0 and B_0 are unknown, $\boldsymbol{u}_t \in \mathbb{R}^q$ is the control action at any time $t \geq 0$, and it is designed based on values of $\boldsymbol{x}_s$ for $s \in [0, t]$. The matrix $B_0 \in \mathbb{R}^{p\times q}$ models the influence of the control action on the state evolution over time, while $A_0 \in \mathbb{R}^{p\times p}$ is the (open-loop) transition matrix reflecting interactions between the coordinates of the state vector $\boldsymbol{x}_t$. The diffusion term in (1) consists of a non-standard Wiener process $\mathbb{W}_t$ that will be defined in the next section. The goal is to

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

study efficient RL policies that can design $\boldsymbol{u}_t$ to minimize a quadratic cost function, defined in the next section, subject to uncertainties around A_0 and B_0.

At a first glance, this problem is similar to most RL problems since the optimal policy must balance between the two objectives of learning the unknown matrices A_0 and B_0 (exploration) and optimally selecting the control signals $\boldsymbol{u}_t$ to minimize the cost (exploitation). However, unlike most RL problems that have finite or bounded-support state space, ensuring *stability*, that $\boldsymbol{x}_t$ stays bounded, is a crucial part of designing optimal policies. For example, in the discrete-time version of the problem, robust exploration is used to protect against unpredictably unstable trajectories [3–6].

Related literature. The existing literature studies efficiency of TS for learning optimal decisions in finite action spaces [7–12]. In this stream of research, it is shown that, over time, the posterior distribution concentrates around low-cost actions [13–15]. TS is also studied in further discrete-time settings with the environment represented by parameters that belong to a continuum, and Bayesian and frequentist regret bounds are shown for linear-quadratic regulators [16–19]. However, effectiveness of TS in highly noisy environments that are modeled by diffusion processes remains unexplored to date, due to technical challenges that will be described below.

For continuous-time linear time invariant dynamical systems, infinite-time consistency results are shown under a variety of technical assumptions, followed by alternating policies that cause (small) linear regrets [20–24]. From a computational viewpoint, pure exploration algorithms for computing optimal policies based on multiple trajectories of the state and action data are studied as well [25–27], for which a useful survey is available [28]. However, papers that study exploration versus exploitation, and provide non-asymptotic estimation rates or regret bounds are limited to a few recent work about offline RL or stabilized processes [29, 2, 30].

Contributions. This work, first establishes that TS learns to stabilize the diffusion process (1). Specifically, in Theorem 1 of Section 3, we provide the first theoretical stabilization guarantee for diffusion processes, showing that the probability of preventing the state process from growing unbounded grows to 1, at an exponential rate that depends on square-root of the time length devoted to stabilization. As mentioned above, for RL problems with finite state spaces, the process is by definition stabilized, regardless of the policy. However, for the Euclidean state space of $\boldsymbol{x}_t$ in (1), stabilization is necessary to ensure that the state and the cost do not grow unbounded.

Then, efficiency of TS in balancing exploration versus exploitation for minimizing a cost function that has a quadratic form of both the state and the control action is shown. Indeed, we establish in Theorem 2 of Section 4 that the regret TS incurs, grows as the *square-root of time*, while the squared estimation error decays with the same rate. It is also shown that both the above quantities grow quadratically with the dimension. To the authors' knowledge, the presented results are the first theoretical analyses of TS for learning to control diffusion processes.

Additionally, through extensive simulations we illustrate that TS enjoys smaller average regret and substantially lower worst-case regret than the existing RL policies, thanks to its informed exploration.

It is important to highlight that theoretical analysis of RL policies for diffusion processes is highly non-trivial. Specifically, the conventional discrete-time RL technical tools are not applicable, due to uncountable cardinality of the random variables involved in a diffusion process, the unavoidable dependence between them, and the high level of processing and estimation noise. To address these, we make four main contributions. First, non-asymptotic and uniform upper bounds for continuous-time martingales and for Ito integrals are required to quantify the estimation accuracy. For that purpose, we establish concentration inequalities and show sub-exponential tail bounds for *double stochastic integrals*. Second, one needs sharp bounds for the impact of estimation errors on eigenvalues of certain non-linear matrices of the drift parameters that determine actions taken by TS policy. To tackle that, we perform a novel and tight *eigenvalue perturbation-analysis* based on the approximation error, dimension, and spectrum of the matrices. We also establish *Lipschitz continuity* of the control policy with respect to the drift matrices, by developing new techniques based on matrix-valued curves. Third, to capture evaluation of both immediate and long-term effects of sub-optimal actions, we employ *Ito calculus* to bound the stochastic regret and specify effects of all problem parameters. Finally, to study learning from data trajectories that the condition number of their information matrix grows unbounded, we develop stochastic inequalities for *self-normalized continuous-time martingales*, and *spectral analysis* of non-linear functions of random matrices.

Organization. The organization of the subsequent sections is as follows. We formulate the problem in Section 2, while Algorithm 1 that utilizes TS for learning to stabilize the process and its high-probability performance guarantee are presented in Section 3. Then, in Section 4, TS is considered for learning to minimize a quadratic cost function, and the rates of estimation and regret are established. Next, theoretical analysis are provided in Section 5, followed by real-world numerical results of Section 6. Detailed proofs and auxiliary lemmas are delegated to the appendices.

Notation. The smallest (the largest) eigenvalue of matrix M, in magnitude, is denoted by $\underline{\lambda}(M)$ ($\overline{\lambda}(M)$). For a vector a, $\|a\|$ is the ℓ_2 norm, and for a matrix M, $\|M\|$ is the operator norm that is the supremum of $\|Ma\|$ for a on the unit sphere. $\boldsymbol{N}(\mu, \Sigma)$ is Gaussian distribution with mean μ and covariance Σ. If μ is a matrix (instead of vector), then $\boldsymbol{N}(\mu, \Sigma)$ denotes a distribution on matrices of the same dimension as μ, such that all columns are independent and share the covariance matrix Σ. In this paper, transition matrices $A \in \mathbb{R}^{p\times p}$ together with input matrices $B \in \mathbb{R}^{p\times q}$ are jointly denoted by the $(p+q)\times p$ parameter matrix $\boldsymbol{\theta} = [A, B]^\top$. We employ $\vee$ ($\wedge$) for maximum (minimum). Finally, $a \lesssim b$ expresses that $a \leq \alpha_0 b$, for some fixed constant α_0.

2 Problem Statement

We study the problem of designing provably efficient reinforcement learning policies for minimizing a quadratic cost function in an uncertain linear diffusion process. To proceed, fix the complete probability space $(\Omega, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$, where Ω is the sample space, $\{\mathcal{F}_t\}_{t\geq 0}$ is a continuous-time filtration (i.e., increasing sigma-fields), and $\mathbb{P}$ is the probability measure defined on $\mathcal{F}_\infty$.

The state comprises the diffusion process $\boldsymbol{x}_t$ in (1), where $\boldsymbol{\theta_0} = [A_0, B_0]^\top \in \mathbb{R}^{(p+q)\times p}$ is the unknown drift parameter. The diffusion term in (1) follows infinitesimal variations of the p dimensional Wiener process $\{\mathbb{W}_t\}_{t\geq 0}$. That is, $\{\mathbb{W}_t\}_{t\geq 0}$ is a multivariate Gaussian process with independent increments and with the stationary covariance matrix $\Sigma_{\mathbb{W}}$, such that for all $0 \leq s_1 \leq s_2 \leq t_1 \leq t_2$,

$$\begin{bmatrix} \mathbb{W}_{t_2} - \mathbb{W}_{t_1} \\ \mathbb{W}_{s_2} - \mathbb{W}_{s_1} \end{bmatrix} \sim \boldsymbol{N}\left(\begin{bmatrix} 0_p \\ 0_p \end{bmatrix}, \begin{bmatrix} (t_2 - t_1)\Sigma_{\mathbb{W}} & 0_{p\times p} \\ 0_{p\times p} & (s_2 - s_1)\Sigma_{\mathbb{W}} \end{bmatrix} \right). \tag{2}$$

Existence, construction, continuity, and non-differentiability of Wiener processes are well-known [31]. It is standard to assume that $\Sigma_{\mathbb{W}}$ is positive definite, which is a common condition in learning-based control [28, 29, 2, 30] to ensure accurate estimation over time.

The RL policy designs the action $\{\boldsymbol{u}_t\}_{t\geq 0}$, based on the observed system state by the time, as well as the previously applied actions, to minimize the long-run average cost

$$\limsup_{T\to\infty} \frac{1}{T} \int_0^T \left[\boldsymbol{x}_t^\top, \boldsymbol{u}_t^\top\right] Q \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{u}_t \end{bmatrix} \boldsymbol{d}t, \quad \text{for} \quad Q = \begin{bmatrix} Q_x & Q_{xu} \\ Q_{xu}^\top & Q_u \end{bmatrix}. \tag{3}$$

Above, the cost is determined by the positive definite matrix Q, where $Q_x \in \mathbb{R}^{p\times p}$, $Q_u \in \mathbb{R}^{q\times q}$, $Q_{xu} \in \mathbb{R}^{p\times q}$. In fact, Q determines the weights of different coordinates of $\boldsymbol{x}_t, \boldsymbol{u}_t$ in the cost function, so that the policy aims to make the states small, by deploying small actions. The cost matrix Q is assumed known to the policy. Formally, the problem is to minimize (3) by the policy

$$\boldsymbol{u}_t = \widehat{\boldsymbol{\pi}}\left(Q, \{\boldsymbol{x}_s\}_{0\leq s\leq t}, \{\boldsymbol{u}_s\}_{0\leq s<t}\right). \tag{4}$$

Without loss of generality, and for the ease of presentation, we follow the canonical formulation that sets $Q_{xu} = 0$; one can simply convert the case $Q_{xu} \neq 0$ to the canonical form, by employing a rotation to $\boldsymbol{x}_t, \boldsymbol{u}_t$ [32–35]. It is well-known that if, hypothetically, the truth $\boldsymbol{\theta_0}$ was known, an optimal policy $\boldsymbol{\pi_{opt}}$ could be explicitly found by solving the continuous-time algebraic Riccati equation. That is, for a generic drift matrix $\boldsymbol{\theta} = [A, B]^\top$, finding the symmetric $p \times p$ matrix $P(\boldsymbol{\theta})$ that satisfies

$$A^\top P(\boldsymbol{\theta}) + P(\boldsymbol{\theta}) A - P(\boldsymbol{\theta}) B Q_u^{-1} B^\top P(\boldsymbol{\theta}) + Q_x = 0. \tag{5}$$

This means, for the true parameter $\boldsymbol{\theta_0} = [A_0, B_0]^\top$, we can let $P(\boldsymbol{\theta_0})$ solve the above equation, and define the policy

$$\boldsymbol{\pi_{opt}}: \quad \boldsymbol{u}_t = -Q_u^{-1} B_0^\top P(\boldsymbol{\theta_0})\, \boldsymbol{x}_t, \quad \forall t \geq 0. \tag{6}$$

It is known that the linear time-invariant policy $\boldsymbol{\pi_{opt}}$ minimizes the average cost in (3) [32–35].

Definition 1 *The process in* (1) *is stabilizable, if all eigenvalues of* $\overline{A} = A_0 + B_0 K$ *have negative real-parts, for a matrix* K*. Such* $K, \overline{A}$ *are called a stabilizer and the stable closed-loop matrix.*

We assume that the process (1) with the drift parameter $\boldsymbol{\theta_0}$ is stabilizable. Therefore, $P(\boldsymbol{\theta_0})$ exists, is unique, and can be computed using continuous-time Riccati differential equations similar to (5), except that the zero matrix on the right-hand side will be replaced by the derivative of $P(\boldsymbol{\theta})$ [32–35]. Furthermore, it is known that real-parts of all eigenvalues of $\overline{A}_0 = A_0 - B_0 Q_u^{-1} B_0^\top P(\boldsymbol{\theta_0})$ are negative, i.e., $\left|\overline{\lambda}\left(\exp\left(\overline{A}_0 t\right)\right)\right| < 1$, which means the matrix $\exp\left(\overline{A}_0 t\right)$ decays exponentially fast as t grows [32–35]. In the sequel, we use (5) and refer to the solution $P(\boldsymbol{\theta})$ for different stabilizable $\boldsymbol{\theta}$. More details about the above optimal feedback policy can be found in the aforementioned references.

In absence of exact knowledge of $\boldsymbol{\theta_0}$, a policy $\widehat{\boldsymbol{\pi}}$ collects data and leverages it to approximate $\boldsymbol{\pi_{\mathbf{opt}}}$ in (6). Therefore, at all (finite) times, there is a gap between the cost of $\widehat{\boldsymbol{\pi}}$, compared to that of $\boldsymbol{\pi_{\mathbf{opt}}}$. The cumulative performance degradation due to this gap is the *regret* of the policy $\widehat{\boldsymbol{\pi}}$, that we aim to minimize. Technically, whenever the control action $\boldsymbol{u}_t$ is designed by the policy $\widehat{\boldsymbol{\pi}}$ according to (4), concatenate the resulting state and input signals to get the observation $\boldsymbol{z}_t(\widehat{\boldsymbol{\pi}}) = \left[\boldsymbol{x}_t^\top, \boldsymbol{u}_t^\top\right]^\top$. If it is clear from the context, we drop $\widehat{\boldsymbol{\pi}}$. Similarly, $\boldsymbol{z}_t(\boldsymbol{\pi_{\mathbf{opt}}})$ denotes the observation signal of $\boldsymbol{\pi_{\mathbf{opt}}}$. Now, the regret at time T is defined by:

$$\mathrm{Reg}_{\widehat{\boldsymbol{\pi}}}(T) = \int_0^T \left(\left\| Q^{1/2} \boldsymbol{z}_t(\widehat{\boldsymbol{\pi}}) \right\|^2 - \left\| Q^{1/2} \boldsymbol{z}_t(\boldsymbol{\pi_{\mathbf{opt}}}) \right\|^2 \right) \boldsymbol{d}t.$$

A secondary objective is the learning accuracy of $\boldsymbol{\theta_0}$ from the single trajectory of the data generated by $\widehat{\boldsymbol{\pi}}$. Letting $\widehat{\boldsymbol{\theta}}_{\boldsymbol{t}}$ be the parameter estimate at time t, we are interested in scaling of $\left\| \widehat{\boldsymbol{\theta}}_{\boldsymbol{t}} - \boldsymbol{\theta_0} \right\|$ with respect to t, p, and q.

3 Stabilizing the Diffusion Process

This section focuses on establishing that Thompson sampling (TS) learns to stabilize the diffusion process (1). First, let us intuitively discuss the problem of stabilizing unknown diffusion processes. Given that the optimal policy in (6) stabilizes the process in (1), a natural candidate to obtain a stable process under uncertainty of the drift matrices A_0, B_0, is a linear feedback of the form $\boldsymbol{u}_t = K\boldsymbol{x}_t$. So, letting $\overline{A} = A_0 + B_0 K$, the solution of (1) is the Ornstein–Uhlenbeck process $\boldsymbol{x}_t = e^{\overline{A}t}\boldsymbol{x}_0 + \int_0^t e^{\overline{A}(t-s)} \boldsymbol{d}\mathbb{W}_s$ [31]. Thus, if real-part of an eigenvalue of $\overline{A}$ is non-negative, then the magnitude of $\boldsymbol{x}_t$ grows unbounded with t [31]. Therefore, addressing instabilities of this form is important, *prior* to minimizing the cost. Otherwise, the regret grows (super) linearly with time. In particular, if A_0 has some eigenvalue(s) with non-negative real-part(s), then it is necessary to employ feedback to preclude instabilities.

In addition to minimizing the cost, the algebraic Riccati equation in (5) provides a reliable and widely-used framework for stabilization, as discussed after (6). Accordingly, due to uncertainty about $\boldsymbol{\theta_0}$, one can solve (5) and find $P\left(\widehat{\boldsymbol{\theta}}\right)$, only for an approximation $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta_0}$. Then, we expect to stabilize the system in (1) by applying a linear feedback that is designed for the approximate drift matrix $\widehat{\boldsymbol{\theta}}$. Technically, we need to ensure that all eigenvalues of $A_0 - B_0 Q_u^{-1} \widehat{B}^\top P\left(\widehat{\boldsymbol{\theta}}\right)$ lie in the open left half-plane. To ensure that these requirements are met in a sustainable manner, the main challenges are

(i) fast and accurate learning of $\boldsymbol{\theta_0}$ so that after a short time period, a small error $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}$ is guaranteed,

(ii) specifying the effect of the error $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}$, on stability of $A_0 - B_0 Q_u^{-1} \widehat{B}^\top P\left(\widehat{\boldsymbol{\theta}}\right)$, and

(iii) devising a remedy for the case that the stabilization procedure fails.

Note that the last challenge is unavoidable, since learning from finite data can never be perfectly accurate, and so any finite-time stabilization procedure has a (possibly small) positive failure probability.

Algorithm 1 addresses the above challenges by applying additionally randomized control actions, and using them to provide a posterior belief $\mathcal{D}$ about $\boldsymbol{\theta_0}$. Note that the posterior is *not* concentrated at $\boldsymbol{\theta_0}$, and a sample $\widehat{\boldsymbol{\theta}}$ from $\mathcal{D}$ approximates $\boldsymbol{\theta_0}$, crudely. Still, the theoretical analysis of Theorem 1

indicates that the failure probability of Algorithm 1 decays exponentially fast with the length of the time interval it is executed. Importantly, this small failure probability can shrink further by repeating the procedure of sampling from $\mathcal{D}$. So, stabilization under uncertainty is guaranteed, after a limited time of interacting with the environment.

To proceed, let $\{w_n\}_{n=0}^{\boldsymbol{\kappa}}$ be a sequence of independent Gaussian vectors with the distribution $w_n \sim \boldsymbol{N}\left(0, \sigma_w^2 I_q\right)$, for some fixed constant σ_w. Suppose that we aim to devote the time length $\boldsymbol{\tau}$ to collect observations for learning to stabilize. Note that since stabilization is performed before moving forward to the main objective of minimizing the cost functions, the stabilization time length $\boldsymbol{\tau}$ is desired to be as short as possible. We divide this time interval of length $\boldsymbol{\tau}$ to $\boldsymbol{\kappa}$ sub-intervals of equal length, and randomize an initial linear feedback policy by adding $\{w_n\}_{n=0}^{\boldsymbol{\kappa}}$. That is, for $n = 0, 1, \cdots, \boldsymbol{\kappa} - 1$, Algorithm 1 employs the control action

$$\boldsymbol{u}_t = K\boldsymbol{x}_t + w_n, \qquad \text{for} \qquad \frac{n\boldsymbol{\tau}}{\boldsymbol{\kappa}} \leq t < \frac{(n+1)\boldsymbol{\tau}}{\boldsymbol{\kappa}}, \tag{7}$$

where K is an initial stabilizing feedback so that all eigenvalues of $A_0 + B_0 K$ lie in the open left half-plane. In practice, such K is easily found using physical knowledge of the model, e.g., via conservative control sequence for an airplane [36, 37]. However, note that such actions are sub-optimal involving large regrets. Therefore, they are only temporarily applied, for the sake of data collection. Then, the data collected during the time interval $0 \leq t \leq \boldsymbol{\tau}$ will be utilized by the algorithm to determine the posterior belief $\mathcal{D}_{\boldsymbol{\tau}}$, as follows. Recalling the notation $\boldsymbol{z}_t^\top = \left[\boldsymbol{x}_t^\top, \boldsymbol{u}_t^\top\right]$, let $\widehat{\mu}_0, \widehat{\Sigma}_0$ be the mean and the precision matrix of a prior normal distribution on $\boldsymbol{\theta_0}$ (using the notation defined in Section 1 for random matrices). Nonetheless, if there is no such prior, we simply let $\widehat{\mu}_0 = 0_{(p+q)\times p}$ and $\widehat{\Sigma}_0 = I_{p+q}$. Then, define

$$\widehat{\Sigma}_{\boldsymbol{\tau}} = \widehat{\Sigma}_0 + \int_0^{\boldsymbol{\tau}} \boldsymbol{z}_s \boldsymbol{z}_s^\top \boldsymbol{d}s, \qquad \widehat{\mu}_{\boldsymbol{\tau}} = \widehat{\Sigma}_{\boldsymbol{\tau}}^{-1} \left(\widehat{\Sigma}_0 \widehat{\mu}_0 + \int_0^{\boldsymbol{\tau}} \boldsymbol{z}_s \boldsymbol{d}\boldsymbol{x}_s^\top \right). \tag{8}$$

Using $\widehat{\Sigma}_{\boldsymbol{\tau}} \in \mathbb{R}^{(p+q)\times(p+q)}$ together with the mean matrix $\widehat{\mu}_{\boldsymbol{\tau}}$, Algorithm 1 forms the posterior belief

$$\mathcal{D}_{\boldsymbol{\tau}} = \boldsymbol{N}\left(\widehat{\mu}_{\boldsymbol{\tau}}, \widehat{\Sigma}_{\boldsymbol{\tau}}^{-1}\right), \tag{9}$$

about the drift parameter $\boldsymbol{\theta_0}$. So, as defined in the notation, the posterior distribution of every column $i = 1, \cdots, p$ of $\boldsymbol{\theta_0}$, is an independent multivariate normal with the covariance matrix $\widehat{\Sigma}_{\boldsymbol{\tau}}^{-1}$, while the mean is the column i of $\widehat{\mu}_{\boldsymbol{\tau}}$. The final step of Algorithm 1 is to output a sample $\widehat{\boldsymbol{\theta}}$ from $\mathcal{D}_{\boldsymbol{\tau}}$.

Algorithm 1 : Stabilization under Uncertainty

Inputs: initial feedback K, stabilization time length $\boldsymbol{\tau}$
for $n = 0, 1, \cdots, \boldsymbol{\kappa} - 1$ **do**
  while $n\boldsymbol{\tau}\boldsymbol{\kappa}^{-1} \leq t < (n+1)\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}$ **do**
    Apply control action $\boldsymbol{u}_t$ in (7)
  end while
end for
Calculate $\widehat{\Sigma}_{\boldsymbol{\tau}}, \widehat{\mu}_{\boldsymbol{\tau}}$ according to (8)
Return sample $\widehat{\boldsymbol{\theta}}$ from the distribution $\mathcal{D}_{\boldsymbol{\tau}}$ in (9)

Next, to establish performance guarantees for Algorithm 1, let us quantify the *ideal* stability by

$$\boldsymbol{\zeta_0} = -\log \overline{\lambda}\left(\exp\left[A_0 - B_0 Q_u^{-1} B_0^\top P\left(\boldsymbol{\theta_0}\right)\right]\right). \tag{10}$$

By definition, $\boldsymbol{\zeta_0}$ is positive. In fact, it is the smallest distance between the imaginary axis in the complex-plane, and the eigenvalues of the transition matrix $\overline{A}_0 = A_0 - B_0 Q_u^{-1} B_0^\top P\left(\boldsymbol{\theta_0}\right)$, under the optimal policy in (6). Since $\boldsymbol{\theta_0}$ is unavailable, it is *not* realistic to expect that after applying a policy based on $\widehat{\boldsymbol{\theta}}$ given by Algorithm 1, real-parts of all eigenvalues of the resulting matrix $A_0 - B_0 Q_u^{-1} \widehat{B}^\top P\left(\widehat{\boldsymbol{\theta}}\right)$ are at most $-\boldsymbol{\zeta_0}$. However, $\boldsymbol{\zeta_0}$ is crucial in studying stabilization, such that stabilizing controllers for systems with larger $\boldsymbol{\zeta_0}$ can be learned faster. The exact effect of this quantity, as well as those of other properties of the diffusion process, are formally established in the following result. Informally, the failure probability of Algorithm 1 decays exponentially with $\boldsymbol{\tau}^{1/2}$.

Theorem 1 (Stabilization Guarantee) *For the sample $\widehat{\boldsymbol{\theta}}$ given by Algorithm 1, let $\mathcal{E}_{\boldsymbol{\tau}}$ be the failure event that $A_0 - B_0 Q_u^{-1} \widehat{B}^\top P\left(\widehat{\boldsymbol{\theta}}\right)$ has an eigenvalue in the closed right half-plane. Then, if $\boldsymbol{\kappa} \gtrsim \boldsymbol{\tau}^2$, we have*

$$\log \mathbb{P}(\mathcal{E}_{\boldsymbol{\tau}}) \lesssim - \ \frac{\underline{\lambda}\left(\Sigma_{\mathbb{W}}\right) \wedge \sigma_w^2}{\overline{\lambda}\left(\Sigma_{\mathbb{W}}\right) \vee \sigma_w^2} \ \frac{1 \wedge \boldsymbol{\zeta_0}^p}{1 \vee \|K\|^3} \ \sqrt{\frac{\boldsymbol{\tau}}{p^3 q}}. \tag{11}$$

The above result indicates that more heterogeneity in coordinates of the Wiener noise renders stabilization harder. Moreover, using (10), the term $1 \wedge \boldsymbol{\zeta_0}^p$ reflects that less stable diffusion processes with smaller $\boldsymbol{\zeta_0}$, are significantly harder to stabilize under uncertainty. Also as one can expect, larger dimensions make learning to stabilize harder. This is contributed by higher number of parameters to learn, as well as higher sensitivity of eigenvalues for processes of larger dimensions. Finally, the failure probability decays as $\boldsymbol{\tau}^{1/2}$, mainly because continuous-time martingales have sub-exponential distributions, unlike sub-Gaussianity of discrete-time counterparts [38–40].

4 Thompson Sampling for Efficient Control: Algorithm and Theory

In this section, we proceed towards analysis of Thompson sampling (TS) for minimizing the quadratic cost in (3), and show that it efficiently learns the optimal control actions. That is, TS balances the exploration versus exploitation, such that its regret grows with (nearly) the square-root rate, as time grows. In the sequel, we introduce Algorithm 2 and discuss the conceptual and technical frameworks it relies on. Then, we establish efficiency by showing regret bounds in terms of different problem parameters and provide the rates of estimating the unknown drift matrices.

In Algorithm 2, first the learning-based stabilization Algorithm 1 is run during the time period $0 \leq t < \boldsymbol{\tau}_0$. So, according to Theorem 1, the optimal feedback of $\widehat{\boldsymbol{\theta}}_{\mathbf{0}}$ stabilizes the system with a high probability, as long as $\boldsymbol{\tau}_0$ is sufficiently large. Note that if growth of the state indicates that Algorithm 1 failed to stabilize, one can repeat sampling from $\mathcal{D}_{\boldsymbol{\tau}_0}$. So, we can assume that the evolution of the controlled diffusion process remains stable when Algorithm 2 is being executed. On the other hand, the other benefit of running Algorithm 1 at the beginning is that it performs an initial exploration phase that will be utilized by Algorithm 2 to minimize the regret.

Then, in order to learn the optimal policy $\boldsymbol{\pi}_{\mathbf{opt}}$ with minimal sub-optimality, RL algorithms need to cope with a fundamental challenge, commonly known as the exploration-exploitation dilemma. To see that, first note that an acceptable policy that aims to have sub-linear regret, needs to take near-optimal control actions in a long run; $\boldsymbol{u}_t \approx -Q_u^{-1} B_0^\top P\left(\boldsymbol{\theta_0}\right) \boldsymbol{x}_t$. Although such policies exploit well and their control actions are close to that of $\boldsymbol{\pi}_{\mathbf{opt}}$, their regret grows large since they fail to explore. Technically, the trajectory of observations $\{\boldsymbol{z}_t\}_{t \geq 0}$ is not rich enough to provide accurate estimations, since in $\boldsymbol{z}_t^\top = \left[\boldsymbol{x}_t^\top, \boldsymbol{u}_t^\top\right]$, the signal $\boldsymbol{u}_t$ is (almost) a linear function of the state signal $\boldsymbol{x}_t$, and so does not contribute towards gathering information about the unknown parameter $\boldsymbol{\theta_0}$. Conversely, for sufficient explorations, RL policies need to take actions that deviate from those of $\boldsymbol{\pi}_{\mathbf{opt}}$, which imposes large regret (as quantified in Lemma 7). Accordingly, the above trade-off needs to be delicately balanced; what we show that TS does.

Algorithm 2 is episodic; the parameter estimates $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ are updated only at the end of the episodes at times $\{\boldsymbol{\tau}_n\}_{n=0}^{\infty}$, while during every episode, actions are taken *as if* $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} = \left[\widehat{A}_n, \widehat{B}_n\right]^\top$ is the unknown truth $\boldsymbol{\theta_0}$. That is, for $\boldsymbol{\tau}_{n-1} \leq t < \boldsymbol{\tau}_n$, using $P\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}\right)$ in (5), we let $\boldsymbol{u}_t = -Q_u^{-1} \widehat{B}_n^\top P\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}\right) \boldsymbol{x}_t$. Then, for each $n = 1, 2, \cdots$, at time $\boldsymbol{\tau}_n$, we use all the observations collected so far, to find $\widehat{\Sigma}_{\boldsymbol{\tau}_n}, \widehat{\mu}_{\boldsymbol{\tau}_n}$ according to (8). Next, we use them to sample $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ from the posterior $\mathcal{D}_{\boldsymbol{\tau}_n}$ in (9).

The episodes in Algorithm 2 are chosen such that their end points satisfy

$$0 < \underline{\alpha} \leq \inf_{n \geq 0} \frac{\boldsymbol{\tau}_{n+1} - \boldsymbol{\tau}_n}{\boldsymbol{\tau}_n} \leq \sup_{n \geq 0} \frac{\boldsymbol{\tau}_{n+1} - \boldsymbol{\tau}_n}{\boldsymbol{\tau}_n} \leq \overline{\alpha} < \infty, \tag{12}$$

for some fixed constants $\underline{\alpha}, \overline{\alpha}$. Broadly speaking, (12) lets the episode lengths of Algorithm 2 scale properly to avoid unnecessary updates of parameter estimates, while at the same time performing sufficient exploration. To see that, first note that since $\widehat{\Sigma}_{\boldsymbol{\tau}}$ grows with $\boldsymbol{\tau}$, the estimation error $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta_0}$ decays (at best polynomially fast) with $\boldsymbol{\tau}_n$. So, until ensuring that updating the posterior yields to

significantly better approximations, it will not be beneficial to update it, sample from it, and solve (5). So, the period $\boldsymbol{\tau}_{n+1} - \boldsymbol{\tau}_n$ that the data up to time $\boldsymbol{\tau}_n$ is utilized, is set to be as long as $\underline{\alpha}\boldsymbol{\tau}_n$. On the other hand, the above period cannot be too long, since we aim to improve the parameter estimates after collecting enough new observations; $\boldsymbol{\tau}_{n+1} \leq (1+\overline{\alpha})\,\boldsymbol{\tau}_n$. A simple setting is to let $\underline{\alpha} = \overline{\alpha}$, which yields to exponential episodes $\boldsymbol{\tau}_n = \boldsymbol{\tau}_0\,(1+\overline{\alpha})^n$. Note that for TS in continuous time, posterior updates should be limited to sufficiently-apart time points. Otherwise, repetitive updates are computationally impractical, and also can degrade the performance by preventing control actions from having enough time to effectively influence.

Algorithm 2 : Thompson Sampling for Efficient Control of Diffusion Processes

Inputs: stabilization time $\boldsymbol{\tau}_0$
Calculate sample $\widehat{\boldsymbol{\theta}}_{\mathbf{0}}$ by running Algorithm 1 for time $\boldsymbol{\tau}_0$
for $n = 1, 2, \cdots$ **do**
  while $\boldsymbol{\tau}_{n-1} \leq t < \boldsymbol{\tau}_n$ **do**
    Apply control action $\boldsymbol{u}_t = -Q_u^{-1}\widehat{B}_{n-1}^{\top} P\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n-1}}\right)\boldsymbol{x}_t$
  end while
  Letting $\widehat{\Sigma}_{\boldsymbol{\tau}_n}, \widehat{\mu}_{\boldsymbol{\tau}_n}$ be as (8), sample $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ from $\mathcal{D}_{\boldsymbol{\tau}_n}$ given in (9)
end for

We show next that Algorithm 2 addresses the exploration-exploitation trade-off efficiently. To see the intuition, consider the sequence of posteriors $\mathcal{D}_{\boldsymbol{\tau}_n}$. The explorations Algorithm 2 performs by sampling $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ from $\mathcal{D}_{\boldsymbol{\tau}_n}$, depends on $\widehat{\Sigma}_{\boldsymbol{\tau}_n}$. Now, if hypothetically $\underline{\lambda}\left(\widehat{\Sigma}_{\boldsymbol{\tau}_n}\right)$ is not large enough, then $\mathcal{D}_{\boldsymbol{\tau}_n}$ does not sufficiently concentrate around $\widehat{\mu}_{\boldsymbol{\tau}_n}$ and so $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ will probably deviate from the previous samples $\left\{\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}\right\}_{i=1}^{n-1}$. So, the algorithm explores more and obtains richer data $\boldsymbol{z}_t$ by diversifying the control signal $\boldsymbol{u}_t$. This renders the next mean $\widehat{\mu}_{\boldsymbol{\tau}_{n+1}}$ a more accurate approximation of $\boldsymbol{\theta}_{\mathbf{0}}$, and also makes $\underline{\lambda}\left(\widehat{\Sigma}_{\boldsymbol{\tau}_{n+1}}\right)$ grow faster than before. Thus, the next posterior $\mathcal{D}_{\boldsymbol{\tau}_{n+1}}$ provides a better sample with smaller estimation error $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n+1}} - \boldsymbol{\theta}_{\mathbf{0}}$. Similarly, if a posterior is excessively concentrated, in a few episodes the posteriors adjust accordingly to the proper level of exploration. Hence, TS eventually balances the exploration versus the exploitation. This is formalized below.

Theorem 2 (Regret and Estimation Rates) *Parameter estimates and regret of Algorithm 2, satisfy*

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta}_{\mathbf{0}}\right\|^2 &\lesssim \frac{\overline{\lambda}\left(\Sigma_{\mathbb{W}}\right)}{\underline{\lambda}\left(\Sigma_{\mathbb{W}}\right)}\log\left(1+\overline{\alpha}\right) \quad (p+q)\,p \quad \boldsymbol{\tau}_n^{-1/2}\log\boldsymbol{\tau}_n\,, \\
\mathrm{Reg}\,(T) &\lesssim \left(\overline{\lambda}\left(\Sigma_{\mathbb{W}}\right)+\sigma_w^2\right)\boldsymbol{\tau}_0 + \frac{\overline{\lambda}\left(\Sigma_{\mathbb{W}}\right)^2}{\underline{\lambda}\left(\Sigma_{\mathbb{W}}\right)}\frac{\overline{\alpha}\|P\left(\boldsymbol{\theta}_{\mathbf{0}}\right)\|^6}{\log(\underline{\alpha}+1)\underline{\lambda}\left(Q\right)^6} \quad (p+q)\,p \quad T^{1/2}\log T\,.
\end{aligned}$$

In the above regret and estimation rates, and similar to Theorem 1, $\overline{\lambda}\left(\Sigma_{\mathbb{W}}\right)/\underline{\lambda}\left(\Sigma_{\mathbb{W}}\right)$ reflects the impact of heterogeneity in coordinates of $\mathbb{W}_t$ on the quality of learning. Also, larger $\log(1+\overline{\alpha})$ corresponds to longer episodes which compromises the estimation. Further, $p(p+q)$ shows that larger number of parameters linearly worsens the learning accuracy. In the regret bound, $\|P\left(\boldsymbol{\theta}_{\mathbf{0}}\right)\|/\underline{\lambda}\left(Q\right)$ indicates effect of the true problem parameters $\boldsymbol{\theta}_{\mathbf{0}}, Q$. Finally, $\left(\overline{\lambda}\left(\Sigma_{\mathbb{W}}\right)+\sigma_w^2\right)\boldsymbol{\tau}_0$ captures the initial phase that Algorithm 1 is run for stabilization, which takes sub-optimal control actions as in (7).

5 Intuition and Summary of the Analysis

The goal of this section is to provide a high-level roadmap of the proofs of Theorems 1 and 2, and convey the main intuition behind the analysis. Complete proofs and the technical lemmas are provided in Appendices A and B, respectively.

Summary of the Proof of Theorem 1. The main steps involve analyzing the estimation (Lemma 4), studying its effect on the solutions of (5) (Lemma 12), and characterizing impact of errors in entries of parameter matrices on their eigenvalues (Lemma 5). Next, we elaborate on these steps.

We show that the error satisfies $\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta_0}\right\| \lesssim p(p+q)^{1/2}\boldsymbol{\tau}^{-1/2}$ (Lemma 4). More precisely, the error depends mainly on total strength of the observation signals $\boldsymbol{z}_t$, which are captured in the precision matrix $\widehat{\Sigma}_{\boldsymbol{\tau}}$, as well as total interactions between the signal $\boldsymbol{z}_t$ and the noise $\mathbb{W}_t$ in the form of the stochastic integral matrix $\int_0^{\boldsymbol{\tau}} \boldsymbol{z}_t \boldsymbol{d}\mathbb{W}_t^\top$. However, we establish an upper bound $\overline{\lambda}\left(\widehat{\Sigma}_{\boldsymbol{\tau}}^{-1}\right) \lesssim \boldsymbol{\tau}^{-1}$, that indicates the concentration rate of the posterior $\mathcal{D}_{\boldsymbol{\tau}}$ (Lemma 3). Similarly, thanks to the randomization signal w_n, the signals $\boldsymbol{z}_t$ are diverse enough to effectively explore the set of matrices $\boldsymbol{\theta} = [A, B]^\top$, leading to accurate approximation of $\boldsymbol{\theta_0}$ by the posterior mean matrix $\widehat{\mu}_{\boldsymbol{\tau}}$. Then, to bound the error terms caused by the Wiener noise $\mathbb{W}_t$, we establish the rate $p(p+q)^{1/2}\boldsymbol{\tau}^{1/2}$ (Lemma 2). Indeed, we show that the entries of this error matrix are continuous-time martingales, and use exponential inequalities for quadratic forms and double stochastic integrals [39, 38] to establish that they have a sub-exponential distribution.

Moreover, the error rate of the feedback satisfies a similar property; $\left\|\widehat{B}^\top P\left(\widehat{\boldsymbol{\theta}}\right) - B_0^\top P\left(\boldsymbol{\theta_0}\right)\right\| \lesssim p(p+q)^{1/2}\boldsymbol{\tau}^{-1/2}$ (Lemma 12). So, letting $\overline{A} = A_0 - B_0 Q_u^{-1}\widehat{B}^\top P\left(\widehat{\boldsymbol{\theta}}\right)$ and $\overline{A}_0 = A_0 - B_0 Q_u^{-1} B_0^\top P\left(\boldsymbol{\theta_0}\right)$, it holds that $\left\|\overline{A} - \overline{A}_0\right\| \lesssim p(p+q)^{1/2}\boldsymbol{\tau}^{-1/2}$. Next, to consider the effect of the errors on the eigenvalues of $\overline{A}$, we compare them to the eigenvalues of $\overline{A}_0$, which are bounded by $-\boldsymbol{\zeta_0}$ in (10). To that end, we establish a novel and tight perturbation analysis for eigenvalues of matrices, with respect to their entries and spectral properties (Lemma 5). Using that, we show that the difference between the eigenvalues of $\overline{A}$ and $\overline{A}_0$ scales as $\left(1 \vee \boldsymbol{r}^{1/2}\left\|\overline{A} - \overline{A}_0\right\|\right)^{1/\boldsymbol{r}}$, where $\boldsymbol{r}$ is the size of the largest block in the Jordan block-diagonalization of $\overline{A}_0$. Therefore, for stability of $\overline{A}$, we need $\left\|\overline{A} - \overline{A}_0\right\| \lesssim p^{-1/2}\left(1 \wedge {\boldsymbol{\zeta_0}}^p\right)$, since $\boldsymbol{r} \leq p$. Note that if $\overline{A}_0$ is diagonalizable, $\boldsymbol{r} = 1$ implies that we can replace the above upper bound by $1 \wedge \boldsymbol{\zeta_0}$. Putting this stability result together with the estimation error in the previous paragraph, we obtain (11).

Summary of the Proof of Theorem 2. To establish the estimation rates, we develop multiple intermediate lemmas quantifying the exact amount of exploration Algorithm 2 performs. First, we utilize the fact that the bias of the posterior distribution $\mathcal{D}_{\boldsymbol{\tau}_n}$ depends on its covariance matrix $\widehat{\Sigma}_{\boldsymbol{\tau}_n}$, as well as a self-normalized continuous-time matrix-valued martingale. For the effect of the former, i.e., $\overline{\lambda}\left(\widehat{\Sigma}_{\boldsymbol{\tau}_n}^{-1/2}\right)$, we show an upper-bound of the order $\boldsymbol{\tau}_n^{-1/4}$ (Lemma 9). To that end, the local geometry of the optimality manifolds that contain drift parameters $\boldsymbol{\theta}$ that has the same optimal feedback as that of the unknown truth $\boldsymbol{\theta_0}$ in (6) are fully specified (Lemma 6), and spectral properties of non-linear functions of random matrices are studied. Then, we establish a stochastic inequality for the self-normalized martingale, indicating that its scaling is of the order $p(p+q)\log \boldsymbol{\tau}_n$ (Lemma 8). Therefore, utilizing the fact that $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \widehat{\mu}_{\boldsymbol{\tau}_n}$ has the same scaling as the bias matrix $\widehat{\mu}_{\boldsymbol{\tau}_n} - \boldsymbol{\theta_0}$, we obtain the estimation rates of Theorem 2.

Next, to prove the presented regret bound, we establish a delicate and tight analysis for the dominant effect of the control signal $\boldsymbol{u}_t$ on the regret Algorithm 2 incurs. Technically, by carefully examining the infinitesimal influences of the control actions at every time on the cost, we show that it suffices to integrate the squared deviations $\left\|\boldsymbol{u}_t + Q_u^{-1}\widehat{B}_n^\top P\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}\right)\boldsymbol{x}_t\right\|^2$ to obtain $\mathrm{Reg}\,(T)$ (Lemma 7). We proceed toward specifying the effect of the exploration Algorithm 2 performs on its exploitation performance by proving the Lipschitz continuity of the solutions of the Riccati equation (5) with respect to the drift parameters: $\left\|P\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}\right) - P\left(\boldsymbol{\theta_0}\right)\right\| \lesssim \left\|\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta_0}\right\|$ (Lemma 12). This result is a very important property of (5) that lets the rates of deviations from the optimal action scale the same as the estimation error, and is proven by careful analysis of integration along matrix-valued curves in the space of drift matrices, as well as spectral analysis for approximate solutions of a Lyapunov equation (Lemma 10). Thus, the regret bound is achieved, using the estimation error result in Theorem 2.

6 Numerical Analysis

We empirically evaluate the theoretical results of Theorems 1 and 2 under three control problems. The first two are for the flight control of X-29A airplane at 2000 ft [36] and for Boeing 747 [37]. The third

simulation is for blood glucose control [41]. We present the results for X-29A airplane in this section, and defer the other two examples to the appendix. The true drift matrices of the X-29A airplane are $A_0 = \begin{bmatrix} -0.16 & 0.07 & -1.00 & 0.04 \\ -15.20 & -2.60 & 1.11 & 0.00 \\ 6.84 & -0.10 & -0.06 & 0.00 \\ 0.00 & 1.00 & 0.07 & 0.00 \end{bmatrix}$, $B_0 = \begin{bmatrix} -0.0006 & 0.0007 \\ 1.3430 & 0.2345 \\ 0.0897 & -0.0710 \\ 0.0000 & 0.0000 \end{bmatrix}$. Further, we let $\Sigma_{\mathbb{W}} = 0.5\, I_p$, $Q_x = I_p$, and $Q_u = 0.1\, I_q$ where I_n is the n by n identity matrix. To update the diffusion process $\boldsymbol{x}_t$ in (1), time-steps of length 10^{-3} are employed. Then, in Algorithm 1, we let $\sigma_w = 5, \boldsymbol{\kappa} = \lfloor \boldsymbol{\tau}^{3/2} \rfloor$, while $\boldsymbol{\tau}$ varies from 4 to 20 seconds. The initial feedback K is generated randomly. The results for 1000 repetitions are depicted on the left plot of Figure 1, confirming Theorem 1 that the failure probability of stabilization, decreases exponentially in $\boldsymbol{\tau}$.

On the right hand side of Figure 1, Algorithm 2 is executed for 600 second, for $\boldsymbol{\tau}_n = 20 \times 1.1^n$. We compare TS with the *Randomized Estimate* algorithm [2] for 100 different repetitions. Average- and worst-case values of the estimation error and the regret are reported, both normalized by their scaling with time and dimension, as in Theorem 2. The graphs show that (especially the worst-case) regret of TS substantially outperforms, suggesting that TS explores in a more robust fashion. Simulations for Boeing 747 and for the blood glucose control, in the appendix, corroborate the above findings.

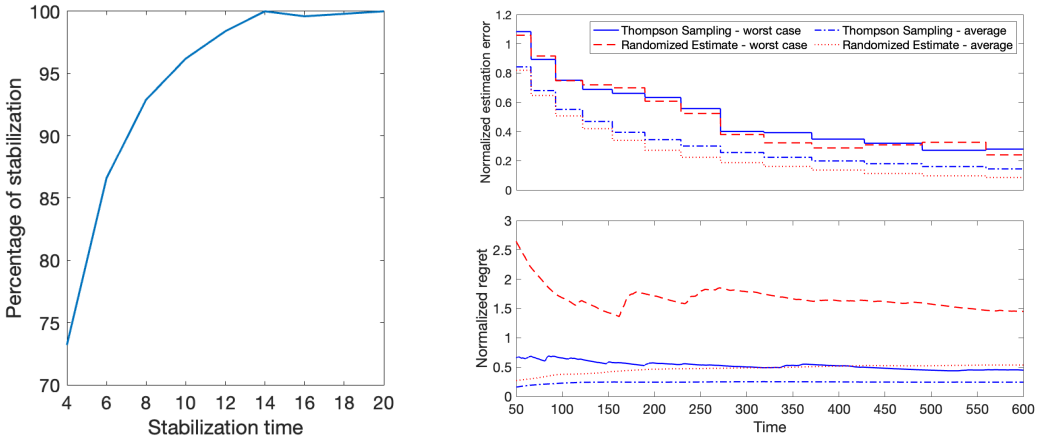

Figure 1: For the X-29A flight control problem, percentage of stabilization for 1000 runs of Algorithm 1 is plotted on the left. The graphs on the right depict the performance of Algorithm 2 (blue) compared to Randomized Estimate policy (red) [2]. The top graph plots the normalized squared estimation error, $\left\| \widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta}_{\mathbf{0}} \right\|^2$ divided by $p(p+q)\boldsymbol{\tau}_n^{-1/2} \log \boldsymbol{\tau}_n$, versus time, while the lower one showcases the regret $\mathrm{Reg}\,(T)$, normalized by $p(p+q)T^{1/2} \log T$. Curves for the worst-case among 100 replications are provided for both quantities, as well as for the averages over all replicates.

7 Concluding Remarks and Future Work

We studied Thompson sampling (TS) RL policies to control a diffusion process with unknown drift matrices. First, we proposed a stabilization algorithm for linear diffusion processes, and established that its failure probability decays exponentially with time. Further, efficiency of TS in balancing exploration versus exploitation for minimizing a quadratic cost function is shown. More precisely, regret bounds growing as square-root of time and square of dimensions are established for Algorithm 2. Empirical studies showcasing superiority of TS over state-of-the-art are provided as well.

As the first theoretical analysis of TS for control of a continuous-time model, this work implies multiple important future directions. Establishing minimax regret lower-bounds for diffusion process control problem is yet unanswered. Moreover, studying the performance of TS for robust control of the diffusion processes aiming to simultaneously minimize the cost function for a family of drift matrices, is also an interesting direction for further investigation. Another problem of interest is efficiency of TS for learning to control under partial observation where the state is not observed and instead a noisy linear function of the state is available as the output signal.

References

[1] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933. (Cited on page 1)

[2] M. K. S. Faradonbeh and M. S. S. Faradonbeh, "Efficient estimation and control of unknown stochastic differential equations," *arXiv preprint arXiv:2109.07630*, 2021. (Cited on page 1, 2, 3, 9, 42, 43, 44)

[3] P. A. Ioannou and J. Sun, *Robust adaptive control*. PTR Prentice-Hall Upper Saddle River, NJ, 1996, vol. 1. (Cited on page 2)

[4] F. L. Lewis, L. Xie, and D. Popa, *Optimal and robust estimation: with an introduction to stochastic control theory*. CRC press, 2017.

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

[6] J. Umenberger, M. Ferizbegovic, T. B. Schön, and H. Hjalmarsson, "Robust exploration in linear quadratic reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019. (Cited on page 2)

[7] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *Conference on learning theory*. JMLR Workshop and Conference Proceedings, 2012, pp. 39–1. (Cited on page 2)

[8] ——, "Further optimal regret bounds for thompson sampling," in *Artificial intelligence and statistics*. PMLR, 2013, pp. 99–107.

[9] A. Gopalan and S. Mannor, "Thompson sampling for learning parameterized markov decision processes," in *Conference on Learning Theory*. PMLR, 2015, pp. 861–898.

[10] M. J. Kim, "Thompson sampling for stochastic control: The finite parameter case," *IEEE Transactions on Automatic Control*, vol. 62, no. 12, pp. 6415–6422, 2017.

[11] M. Abeille and A. Lazaric, "Linear thompson sampling revisited," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 176–184.

[12] N. Hamidi and M. Bayati, "On worst-case regret of linear thompson sampling," *arXiv preprint arXiv:2006.06790*, 2020. (Cited on page 2)

[13] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," *Mathematics of Operations Research*, vol. 39, no. 4, pp. 1221–1243, 2014. (Cited on page 2)

[14] ——, "An information-theoretic analysis of thompson sampling," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2442–2471, 2016.

[15] D. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen, "A tutorial on thompson sampling," *arXiv preprint arXiv:1707.02038*, 2017. (Cited on page 2)

[16] M. Abeille and A. Lazaric, "Improved regret bounds for thompson sampling in linear quadratic control problems," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1–9. (Cited on page 2)

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

[18] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "On adaptive linear–quadratic regulators," *Automatica*, vol. 117, p. 108982, 2020.

[19] S. Sudhakara, A. Mahajan, A. Nayyar, and Y. Ouyang, "Scalable regret for learning to control network-coupled subsystems with unknown dynamics," *arXiv preprint arXiv:2108.07970*, 2021. (Cited on page 2)

[20] P. Mandl, "Consistency of estimators in controlled systems," in *Stochastic Differential Systems*. Springer, 1989, pp. 227–234. (Cited on page 2)

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

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

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

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

[25] S. A. A. Rizvi and Z. Lin, "Output feedback reinforcement learning control for the continuous-time linear quadratic regulator problem," in *2018 Annual American Control Conference (ACC)*. IEEE, 2018, pp. 3417–3422. (Cited on page 2)

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

[27] H. Wang, T. Zariphopoulou, and X. Y. Zhou, "Reinforcement learning in continuous time and space: A stochastic control approach." *J. Mach. Learn. Res.*, vol. 21, pp. 198–1, 2020. (Cited on page 2)

[28] Z.-P. Jiang, T. Bian, and W. Gao, "Learning-based control: A tutorial and some recent results," *Foundations and Trends® in Systems and Control*, vol. 8, no. 3, 2020. (Cited on page 2, 3)

[29] M. Basei, X. Guo, A. Hu, and Y. Zhang, "Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon," *Available at SSRN 3848428*, 2021. (Cited on page 2, 3)

[30] L. Szpruch, T. Treetanthiploet, and Y. Zhang, "Exploration-exploitation trade-off for continuous-time episodic reinforcement learning with linear-convex models," *arXiv preprint arXiv:2112.10264*, 2021. (Cited on page 2, 3)

[31] I. Karatzas and S. Shreve, *Brownian motion and stochastic calculus*. Springer Science & Business Media, 2012, vol. 113. (Cited on page 3, 4, 16, 17, 19, 20, 21, 30, 32)

[32] G. Chen, G. Chen, and S.-H. Hsu, *Linear stochastic control systems*. CRC press, 1995, vol. 3. (Cited on page 3, 4)

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

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

[35] S. P. Bhattacharyya and L. H. Keel, *Linear Multivariable Control Systems*. Cambridge University Press, 2022. (Cited on page 3, 4)

[36] J. T. Bosworth, *Linearized aerodynamic and control law models of the X-29A airplane and comparison with flight data*. National Aeronautics and Space Administration, Office of Management . . . , 1992, vol. 4356. (Cited on page 5, 8)

[37] T. Ishihara, H.-J. Guo, and H. Takeda, "A design of discrete-time integral controllers with computation delays via loop transfer recovery," *Automatica*, vol. 28, no. 3, pp. 599–603, 1992. (Cited on page 5, 8, 42)

[38] P. Cheridito, H. M. Soner, and N. Touzi, "Small time path behavior of double stochastic integrals and applications to stochastic control," *The Annals of Applied Probability*, vol. 15, no. 4, pp. 2472–2495, 2005. (Cited on page 6, 8)

[39] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Annals of Statistics*, pp. 1302–1338, 2000. (Cited on page 8, 17, 20, 21)

[40] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of computational mathematics*, vol. 12, no. 4, pp. 389–434, 2012. (Cited on page 6, 18)

[41] T. Zhou, J. L. Dickson, and J. Geoffrey Chase, "Autoregressive modeling of drift and random error to characterize a continuous intravascular glucose monitoring sensor," *Journal of Diabetes Science and Technology*, vol. 12, no. 1, pp. 90–104, 2018. (Cited on page 9, 42)

[42] P. Hartman and A. Wintner, "The spectra of toeplitz's matrices," *American Journal of Mathematics*, vol. 76, no. 4, pp. 867–882, 1954. (Cited on page 17)

[43] L. Reichel and L. N. Trefethen, "Eigenvalues and pseudo-eigenvalues of toeplitz matrices," *Linear algebra and its applications*, vol. 162, pp. 153–185, 1992. (Cited on page 17)

[44] S. I. Resnick, *Adventures in stochastic processes*. Springer Science & Business Media, 1992.

[45] P. Billingsley, "Convergence of probability measures," *INC, New York*, vol. 2, no. 2.4, 1999.

[46] R. Durrett, *Probability: theory and examples*. Cambridge university press, 2019, vol. 49.

[47] R. Gondhalekar, E. Dassau, and F. J. Doyle III, "Periodic zone-mpc with asymmetric costs for outpatient-ready safety of an artificial pancreas to treat type 1 diabetes," *Automatica*, vol. 71, pp. 237–246, 2016. (Cited on page 42)

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] ; See Sections 3 and 4.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] ; See Theorems 1 and 2, as well as their discussions.
   (b) Did you include complete proofs of all theoretical results? [Yes] ; Proofs are provided as appendices.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] ; See Section 6
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] ; The reported curves reflect the worst-case analysis and so no error bars are needed.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] .
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Organization of Appendices

This paper has four appendices. First, we prove Theorem 1 in Appendix A, together with multiple lemmas that the proof of the theorem relies on, and their statements and proofs are provided in Appendix A as well. Similarly, the proof for Theorem 2 together with intermediate steps, all are presented in Appendix B. Then, Appendix C consists of statements and proofs of other results that are used for establishing both theorems. Finally, empirical simulations beyond those presented in Section 6 are presented in Appendix D.

A Proof of Theorem 1

For analyzing the estimation error, we establish Lemma 4, which under the condition $\boldsymbol{\kappa} \gtrsim \boldsymbol{\tau}^2$ provides that with probability at least $1-\delta$,

$$\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}\right\| \lesssim \frac{p\,(p+q)^{1/2}}{\boldsymbol{\tau}^{1/2}} \frac{\overline{\lambda}\,(\Sigma_{\mathbb{W}}) \vee \sigma_w^2}{\underline{\lambda}\,(\Sigma_{\mathbb{W}}) \wedge \sigma_w^2} \left(1+\|K\|\right)^3 \log\left(\frac{pq\boldsymbol{\kappa}}{\delta}\right).$$

Note that in the proof of Lemma 4, results of Lemmas 1, 2, and 3 are used.

Therefore, Lemma 12 implies that for solutions of (5), with probability at least $1-\delta$, it holds that

$$\left\|P\left(\widehat{\boldsymbol{\theta}}\right) - P\left(\boldsymbol{\theta_0}\right)\right\| \lesssim \frac{p\,(p+q)^{1/2}}{\boldsymbol{\tau}^{1/2}} \frac{\overline{\lambda}\,(\Sigma_{\mathbb{W}}) \vee \sigma_w^2}{\underline{\lambda}\,(\Sigma_{\mathbb{W}}) \wedge \sigma_w^2} \left(1+\|K\|\right)^3 \log\left(\frac{pq\boldsymbol{\kappa}}{\delta}\right).$$

Note that we get the same expression as the right-hand-side above, as an upper bound for $\left\|\widehat{B}^\top P\left(\widehat{\boldsymbol{\theta}}\right) - B_0^\top P\left(\boldsymbol{\theta_0}\right)\right\|$. So, letting

$$\overline{A} = A_0 - B_0 Q_u^{-1} \widehat{B}^\top P\left(\widehat{\boldsymbol{\theta}}\right), \qquad \overline{A}_0 = A_0 - B_0 Q_u^{-1} B_0^\top P\left(\boldsymbol{\theta_0}\right),$$

we obtain

$$\left\|\overline{A} - \overline{A}_0\right\| \lesssim \sqrt{\frac{p^2 q}{\boldsymbol{\tau}}} \frac{\overline{\lambda}\,(\Sigma_{\mathbb{W}}) \vee \sigma_w^2}{\underline{\lambda}\,(\Sigma_{\mathbb{W}}) \wedge \sigma_w^2} \left(1+\|K\|\right)^3 \log\left(\frac{pq\boldsymbol{\kappa}}{\delta}\right), \tag{13}$$

with probability at least $1-\delta$.

Next, to consider the effect of the above errors on the eigenvalues of $\overline{A}$, we compare them to that of $\overline{A}_0$. Note that real-parts of all eigenvalues of $\overline{A}_0$ are at most $-\boldsymbol{\zeta_0}$, as defined in (10). So, using the result and the notation of Lemma 5, for all eigenvalues of $\overline{A}$ being on the open left half-plane it suffices to have $\Delta_{\overline{A}_0}\left(\overline{A} - \overline{A}_0\right) \leq \boldsymbol{\zeta_0}$. Also, in lights of Lemma 5, suppose that $\boldsymbol{r}$ is the size of the largest block in the Jordan block-diagonalization of $\overline{A}_0$. So, (36) implies that if

$$\left\|\overline{A} - \overline{A}_0\right\| \lesssim \boldsymbol{r}^{-1/2}\left(1 \wedge \boldsymbol{\zeta_0}^{\boldsymbol{r}}\right),$$

then Algorithm 1 successfully stabilizes the diffusion process in (1). Thus, (13) shows that the failure probability of Algorithm 1; $\mathbb{P}(\mathcal{E}_{\boldsymbol{\tau}})$, satisfies

$$\sqrt{\frac{p^2 q}{\boldsymbol{\tau}}} \frac{\overline{\lambda}\,(\Sigma_{\mathbb{W}}) \vee \sigma_w^2}{\underline{\lambda}\,(\Sigma_{\mathbb{W}}) \wedge \sigma_w^2} \left(1+\|K\|\right)^3 \log\left(\frac{pq\boldsymbol{\kappa}}{\mathbb{P}(\mathcal{E}_{\boldsymbol{\tau}})}\right) \gtrsim \boldsymbol{r}^{-1/2}\left(1 \wedge \boldsymbol{\zeta_0}^{\boldsymbol{r}}\right).$$

Finally, $\boldsymbol{r} \leq p$ together with $\log\left(pq\boldsymbol{\kappa}\right) \lesssim \boldsymbol{\tau}^{1/2}$, lead to the desired result.

In the remainder of this section, technical lemmas above are stated and their proofs will be provided.

A.1 Bounding cross products of state and randomization

Definition 2 *For a set $\mathcal{S}$, let $\mathbb{1}\{\mathcal{S}\}$ be the indicator function that is 1 on $\mathcal{S}$, and vanishes outside of $\mathcal{S}$.*

Lemma 1 *In Algorithm 1, for $t \geq 0$, define the piecewise-constant signal $v(t)$ below according to the randomization sequence w_n:*

$$v(t) = \sum_{n=0}^{\boldsymbol{\kappa}-1} \mathbb{1}\left\{\frac{n\boldsymbol{\tau}}{\boldsymbol{\kappa}} \leq t < \frac{(n+1)\boldsymbol{\tau}}{\boldsymbol{\kappa}}\right\} w_n. \tag{14}$$

Then, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\left\| \int_0^{\boldsymbol{\tau}} \boldsymbol{x}_s v(s)^\top \boldsymbol{d}s - \frac{\boldsymbol{\tau}^2}{2\boldsymbol{\kappa}^2} B_0 \sum_{n=0}^{\boldsymbol{\kappa}-1} w_n w_n^\top \right\| \\
\lesssim\ & \left(\sigma_w^2 + \overline{\lambda}\left(\Sigma_{\mathbb{W}}\right)\right) \left(1 + \int_0^{\boldsymbol{\tau}} \left\| e^{\overline{A}t} \right\| \boldsymbol{d}t\right) \left(pq^{1/2} \boldsymbol{\tau}^{1/2} \log\frac{pq}{\delta} + q\left[1 + \frac{\boldsymbol{\tau}^2}{\boldsymbol{\kappa}^2}\right] \frac{\boldsymbol{\tau}}{\boldsymbol{\kappa}^{1/2}} \log^{3/2}\frac{\boldsymbol{\kappa} q}{\delta} \right).
\end{aligned}$$

Proof. First, after plugging the control signal $\boldsymbol{u}_t$ in (1) and solving the resulting stochastic differential equation, we obtain

$$\boldsymbol{x}_t = e^{\overline{A}t}\boldsymbol{x}_0 + \int_0^t e^{\overline{A}(t-s)} \boldsymbol{d}\mathbb{W}_s + \int_0^t e^{\overline{A}(t-s)} B_0 v(s) \boldsymbol{d}s. \tag{15}$$

This implies that

$$\int_0^{\boldsymbol{\tau}} \boldsymbol{x}_t v(t)^\top \boldsymbol{d}t = \Phi_1 + \Phi_2 + \Phi_3,$$

where

$$\begin{aligned}
\Phi_1 &= \int_0^{\boldsymbol{\tau}} e^{\overline{A}t} \boldsymbol{x}_0 v(t)^\top \boldsymbol{d}t = \sum_{n=0}^{\boldsymbol{\kappa}-1} \left(\int_{n\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}}^{(n+1)\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}} e^{\overline{A}t} \boldsymbol{d}t \right) \boldsymbol{x}_0 w_n^\top, \\
\Phi_2 &= \int_0^{\boldsymbol{\tau}} \int_0^t e^{\overline{A}(t-s)} \boldsymbol{d}\mathbb{W}_s v(t)^\top \boldsymbol{d}t = \sum_{n=0}^{\boldsymbol{\kappa}-1} \left(\int_{n\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}}^{(n+1)\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}} \int_0^t e^{\overline{A}(t-s)} \boldsymbol{d}\mathbb{W}_s \boldsymbol{d}t \right) w_n^\top, \\
\Phi_3 &= \int_0^{\boldsymbol{\tau}} \int_0^t e^{\overline{A}(t-s)} B_0 v(s) \boldsymbol{d}s v(t)^\top \boldsymbol{d}t.
\end{aligned}$$

To analyze Φ_1, we use the fact that every entry of Φ_1 is a normal random variable with mean zero and variance at most

$$\sigma_w^2 \sum_{n=0}^{\boldsymbol{\kappa}-1} \left\| \int_{n\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}}^{(n+1)\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}} e^{\overline{A}t} \boldsymbol{d}t \right\|^2 \|\boldsymbol{x}_0\|^2 \le \sigma_w^2 \left(\int_0^{\boldsymbol{\tau}} \left\| e^{\overline{A}t} \right\| \boldsymbol{d}t \right)^2 \|\boldsymbol{x}_0\|^2.$$

Therefore, with probability at least $1-\delta$, it holds that

$$\|\Phi_1\| \lesssim \sigma_w \left(\int_0^{\boldsymbol{\tau}} \left\| e^{\overline{A}t} \right\| \boldsymbol{d}t \right) \|\boldsymbol{x}_0\| \sqrt{pq \log\left(\frac{pq}{\delta}\right)}. \tag{16}$$

Furthermore, to study Φ_2, Fubini Theorem [31] gives

$$\begin{aligned}
\int_{n\boldsymbol{\tau\kappa}^{-1}}^{(n+1)\boldsymbol{\tau\kappa}^{-1}} \int_0^t e^{\overline{A}(t-s)}\boldsymbol{d}\mathbb{W}_s\boldsymbol{d}t &= \int_0^{n\boldsymbol{\tau\kappa}^{-1}} \left(\int_{n\boldsymbol{\tau\kappa}^{-1}}^{(n+1)\boldsymbol{\tau\kappa}^{-1}} e^{\overline{A}(t-n\boldsymbol{\tau\kappa}^{-1})}\boldsymbol{d}t \right) e^{\overline{A}(n\boldsymbol{\tau\kappa}^{-1}-s)}\boldsymbol{d}\mathbb{W}_s \\
&+ \int_{n\boldsymbol{\tau\kappa}^{-1}}^{(n+1)\boldsymbol{\tau\kappa}^{-1}} \left(\int_s^{(n+1)\boldsymbol{\tau\kappa}^{-1}} e^{\overline{A}(t-s)}\boldsymbol{d}t \right) \boldsymbol{d}\mathbb{W}_s \\
&= F\sum_{m=1}^{n} e^{\overline{A}(n-m)\boldsymbol{\tau\kappa}^{-1}} \int_{(m-1)\boldsymbol{\tau\kappa}^{-1}}^{m\boldsymbol{\tau\kappa}^{-1}} e^{\overline{A}(m\boldsymbol{\tau\kappa}^{-1}-s)}\boldsymbol{d}\mathbb{W}_s \\
&+ \int_{n\boldsymbol{\tau\kappa}^{-1}}^{(n+1)\boldsymbol{\tau\kappa}^{-1}} G_s\boldsymbol{d}\mathbb{W}_s
\end{aligned}$$

where the matrix $F = \int_{n\boldsymbol{\tau\kappa}^{-1}}^{(n+1)\boldsymbol{\tau\kappa}^{-1}} e^{\overline{A}(t-n\boldsymbol{\tau\kappa}^{-1})}\boldsymbol{d}t$, does not depend on s or n, and the matrix $G_s = \int_s^{(n+1)\boldsymbol{\tau\kappa}^{-1}} e^{\overline{A}(t-s)}\boldsymbol{d}t$, does not depend on n, since $n\boldsymbol{\tau\kappa}^{-1} \le s \le (n+1)\boldsymbol{\tau\kappa}^{-1}$.

So, letting $\mathrm{e}_i, i = 1, \cdots, p$, be the standard basis of the Euclidean space, conditioned on the Wiener process $\{\mathbb{W}_s\}_{s\ge 0}$, for every $j = 1, \cdots, q$, the coordinate j of $\mathrm{e}_i^\top \Phi_2$ is a mean zero normal random variable. Thus, given $\{\mathbb{W}_s\}_{s\ge 0}$, with probability at least $1-\delta$, it holds that

$$\left(\mathrm{e}_i^\top \Phi_2 \mathrm{e}_j\right)^2 \lesssim \mathrm{var}\left(\mathrm{e}_\mathrm{i}^\top \Phi_2 \mathrm{e}_\mathrm{j} \middle| \mathcal{F}\left(\mathbb{W}_{0:\boldsymbol{\tau}}\right)\right) \log\frac{1}{\delta}.$$

Now, to calculate the conditional variance, we can write

$$\frac{\mathrm{var}\left(\mathrm{e}_\mathrm{i}^\top \Phi_2 \mathrm{e}_\mathrm{j} \middle| \mathcal{F}\left(\mathbb{W}_{0:\boldsymbol{\tau}}\right)\right)}{\sigma_w^2} = \sum_{n=0}^{\boldsymbol{\kappa}-1} \left[\mathrm{e}_i^\top \int_{n\boldsymbol{\tau\kappa}^{-1}}^{(n+1)\boldsymbol{\tau\kappa}^{-1}} \int_0^t e^{\overline{A}(t-s)}\boldsymbol{d}\mathbb{W}_s\boldsymbol{d}t \right]^2 \lesssim \sum_{n=1}^{\boldsymbol{\kappa}-1} \left[\left(\sum_{m=1}^{n} \beta_{m,n} \right)^2 + \alpha_n^2 \right],$$

where

$$\begin{aligned}
\beta_{m,n} &= \mathrm{e}_i^\top F e^{\overline{A}(n-m)\boldsymbol{\tau\kappa}^{-1}} \int_{(m-1)\boldsymbol{\tau\kappa}^{-1}}^{m\boldsymbol{\tau\kappa}^{-1}} e^{\overline{A}(m\boldsymbol{\tau\kappa}^{-1}-s)}\boldsymbol{d}\mathbb{W}_s, \\
\alpha_n &= \mathrm{e}_i^\top \int_{n\boldsymbol{\tau\kappa}^{-1}}^{(n+1)\boldsymbol{\tau\kappa}^{-1}} G_s\boldsymbol{d}\mathbb{W}_s.
\end{aligned}$$

To proceed, define the matrix $H = [H_{n,m}]$, where for $1 \le m, n \le \boldsymbol{\kappa}-1$, every block $H_{n,m} \in \mathbb{R}^{1\times p}$ is

$$H_{n,m} = \mathrm{e}_i^\top F e^{\overline{A}(n-m)\boldsymbol{\tau\kappa}^{-1}},$$

for $m \le n$, and is 0 for $m > n$. Then, denote

$$\Gamma = \begin{bmatrix} \int_0^{\boldsymbol{\tau\kappa}^{-1}} e^{\overline{A}(\boldsymbol{\tau\kappa}^{-1}-s)}\boldsymbol{d}\mathbb{W}_s \\ \int_{\boldsymbol{\tau\kappa}^{-1}}^{2\boldsymbol{\tau\kappa}^{-1}} e^{\overline{A}(2\boldsymbol{\tau\kappa}^{-1}-s)}\boldsymbol{d}\mathbb{W}_s \\ \vdots \\ \int_{(\boldsymbol{\kappa}-1)\boldsymbol{\tau\kappa}^{-1}}^{\boldsymbol{\tau}} e^{\overline{A}(\boldsymbol{\tau}-s)}\boldsymbol{d}\mathbb{W}_s \end{bmatrix} \in \mathbb{R}^{p(\boldsymbol{\kappa}-1)\times 1},$$

to get

$$\sum_{n=0}^{\boldsymbol{\kappa}-1}\left(\sum_{m=1}^{n}\beta_{m,n}^2\right)=\|H\Gamma\|^2\leq\overline{\lambda}\left(H^\top H\right)\|\Gamma\|^2.$$

Now, for the matrix H, we have [42, 43]:

$$\overline{\lambda}\left(H^\top H\right)\lesssim\left(\sum_{n=1}^{\boldsymbol{\kappa}-1}\|H_{n,1}\|\right)^2\lesssim\left(\boldsymbol{\tau\kappa}^{-1}\sum_{n=1}^{\boldsymbol{\kappa}}e^{\overline{A}n\boldsymbol{\tau\kappa}^{-1}}\right)^2\lesssim\left(\int_0^{\boldsymbol{\tau}}\left\|e^{\overline{A}t}\right\|\boldsymbol{d}t\right)^2.$$

Note that thanks to the independent increments of the Wiener process, the blocks of Γ are statistically independent. Further, by Ito Isometry [31], every block of Γ is a mean-zero normally distributed vector with the covariance matrix

$$\int_0^{\boldsymbol{\tau\kappa}^{-1}}e^{\overline{A}(\boldsymbol{\tau\kappa}^{-1}-s)}\Sigma_{\mathbb{W}}e^{\overline{A}^\top(\boldsymbol{\tau\kappa}^{-1}-s)}\boldsymbol{d}s.$$

So, according to the exponential inequalities for quadratic forms of normally distributed random variables [39], it holds with probability at least $1-\delta$, that

$$\|\Gamma\|^2\lesssim p\boldsymbol{\kappa}\overline{\lambda}\left(\Sigma_{\mathbb{W}}\right)\left(\boldsymbol{\tau\kappa}^{-1}\right)\log\frac{1}{\delta}.$$

Thus, with probability at least $1-\delta$, we have

$$\sum_{n=0}^{\boldsymbol{\kappa}-1}\left(\sum_{m=1}^{n}\beta_{m,n}^2\right)\lesssim\left(\int_0^{\boldsymbol{\tau}}\left\|e^{\overline{A}t}\right\|\boldsymbol{d}t\right)^2p\overline{\lambda}\left(\Sigma_{\mathbb{W}}\right)\boldsymbol{\tau}\log\frac{1}{\delta}.$$

Similarly, the bound above can be shown for $\sum_{n=1}^{\boldsymbol{\kappa}-1}\alpha_n^2$. Hence, we obtain the corresponding high probability bound for a single entry $\mathrm{e}_i^\top\Phi_2\mathrm{e}_j$ of Φ_2, which together with a union bound, implies that

$$\|\Phi_2\|\lesssim\sigma_w pq^{1/2}\left(\int_0^{\boldsymbol{\tau}}\left\|e^{\overline{A}t}\right\|\boldsymbol{d}t\right)\overline{\lambda}\left(\Sigma_{\mathbb{W}}\right)^{1/2}\boldsymbol{\tau}^{1/2}\log\left(\frac{pq}{\delta}\right),\tag{17}$$

with probability at least $1-\delta$.

Next, according to Fubini Theorem, Φ_3 can also be written as

$$\Phi_3=\int_0^{\boldsymbol{\tau}}\int_0^{s}e^{\overline{A}(s-t)}B_0v(t)v(s)^\top\boldsymbol{d}t\boldsymbol{d}s=\int_0^{\boldsymbol{\tau}}\int_t^{\boldsymbol{\tau}}e^{\overline{A}(s-t)}B_0v(t)v(s)^\top\boldsymbol{d}s\boldsymbol{d}t.$$

Thus, we have

$$2\Phi_3=\int_0^{\boldsymbol{\tau}}\int_0^{\boldsymbol{\tau}}e^{\overline{A}|t-s|}B_0v(t\wedge s)v(s\vee t)^\top\boldsymbol{d}t\boldsymbol{d}s.$$

Recall that the signal $v(t)$ in (14) is piecewise-constant, with values determined by the randomization sequence w_n. So, the above double integral can be written as a double sum

$$\begin{aligned}2\Phi_3&=\sum_{n=0}^{\boldsymbol{\kappa}-1}\sum_{m=0}^{\boldsymbol{\kappa}-1}\left(\int_{n\boldsymbol{\tau\kappa}^{-1}}^{(n+1)\boldsymbol{\tau\kappa}^{-1}}\int_{m\boldsymbol{\tau\kappa}^{-1}}^{(m+1)\boldsymbol{\tau\kappa}^{-1}}e^{\overline{A}|t-s|}\boldsymbol{d}s\boldsymbol{d}t\right)B_0w_{m\wedge n}w_{m\vee n}^\top\\&=\sum_{n=0}^{\boldsymbol{\kappa}-1}\sum_{m=0}^{\boldsymbol{\kappa}-1}\left(e^{\overline{A}|m-n|\boldsymbol{\tau\kappa}^{-1}}\int_0^{\boldsymbol{\tau\kappa}^{-1}}\int_0^{\boldsymbol{\tau\kappa}^{-1}}e^{\overline{A}|t-s|}\boldsymbol{d}s\boldsymbol{d}t\right)B_0w_{m\wedge n}w_{m\vee n}^\top.\end{aligned}$$

Thus, we have

$$2\Phi_3 - \frac{\boldsymbol{\tau}^2}{\boldsymbol{\kappa}^2} B_0 \sum_{n=0}^{\boldsymbol{\kappa}-1} w_n w_n^\top = \Phi_4 + \Phi_5, \tag{18}$$

for

$$\begin{aligned}
\Phi_4 &= \left(\int_0^{\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}} \int_0^{\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}} e^{\overline{A}|t-s|} \boldsymbol{d}s\boldsymbol{d}t - \boldsymbol{\tau}^2 \boldsymbol{\kappa}^{-2} I_q \right) B_0 \sum_{n=0}^{\boldsymbol{\kappa}-1} w_n w_n^\top, \\
\Phi_5 &= 2 \left(\int_0^{\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}} \int_0^{\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}} e^{\overline{A}|t-s|} \boldsymbol{d}s\boldsymbol{d}t \right) \sum_{n=0}^{\boldsymbol{\kappa}-1} \sum_{m=n+1}^{\boldsymbol{\kappa}-1} \left(e^{\overline{A}(m-n)\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}} B_0 w_n w_m^\top \right).
\end{aligned}$$

To proceed, we use the following concentration inequality for random matrices with martingale difference structures, titled as Matrix Azuma inequality [40].

Theorem 3 *Let $\{\Psi_n\}_{n=1}^k$ be a $d_1 \times d_2$ martingale difference sequence. That is, for some filtration $\{\mathcal{F}_n\}_{n=0}^k$, the matrix Ψ_n is $\mathcal{F}_n$-measurable, and $\mathbb{E}\left[\Psi_n \middle| \mathcal{F}_{n-1}\right] = 0$. Suppose that $\|\Psi_n\| \leq \sigma_n$, for some fixed sequence $\{\sigma_n\}_{n=1}^k$. Then, with probability at least $1-\delta$, we have*

$$\left\| \sum_{n=1}^k \Psi_n \right\|^2 \lesssim \left(\sum_{n=1}^k \sigma_n^2 \right) \log \frac{d_1+d_2}{\delta}.$$

So, to study Φ_4, we apply Theorem 3 to the random matrices $\Psi_n = w_n w_n^\top - \sigma_w^2 I_q$, using the trivial filtration and the high probability upper-bounds for $\|\Psi_n\| \leq \|w_n\|^2 + \sigma_w^2$;

$$\|\Psi_n\| \leq \sigma_n = \sigma_w^2 \left(1 + q \log \frac{q\boldsymbol{\kappa}}{\delta}\right),$$

as well as the fact

$$\left\| \int_0^{\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}} \int_0^{\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}} \left(e^{\overline{A}|t-s|} - I_q \right) \boldsymbol{d}s\boldsymbol{d}t \right\| \lesssim \boldsymbol{\tau}^3 \boldsymbol{\kappa}^{-3},$$

to obtain the following bound, which holds with probability at least $1-\delta$:

$$\|\Phi_4\| \lesssim \|B_0\| \sigma_w^2 \boldsymbol{\tau}^3 \boldsymbol{\kappa}^{-2} \left(1 + \frac{q}{\boldsymbol{\kappa}^{1/2}} \log^{3/2} \frac{\boldsymbol{\kappa} q}{\delta} \right). \tag{19}$$

On the other hand, to establish an upper-bound for Φ_5, consider the random matrices

$$\Psi_n = \sum_{m=n+1}^{\boldsymbol{\kappa}-1} \left(e^{\overline{A}(m-n)\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}} B_0 w_n w_m^\top \right),$$

subject to the natural filtration they generate, and apply Theorem 3, using the bounds

$$\|\Psi_n\| \leq \sigma_n \lesssim \boldsymbol{\tau}^{-1} \boldsymbol{\kappa} \left(\int_0^{\boldsymbol{\tau}} \left\| e^{\overline{A}t} \right\| \boldsymbol{d}t \right) \|B_0\| \sigma_w^2 q \log \frac{\boldsymbol{\kappa} q}{\delta},$$

together with

$$\left\| \int_0^{\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}} \int_0^{\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}} e^{\overline{A}|t-s|} \boldsymbol{d}s\boldsymbol{d}t \right\| \lesssim \boldsymbol{\tau}^2 \boldsymbol{\kappa}^{-2}.$$

Therefore, Theorem 3 indicates that with probability at least $1-\delta$, it holds that

$$\Phi_5 \lesssim \frac{\boldsymbol{\tau}}{\boldsymbol{\kappa}^{1/2}} \left(\int_0^{\boldsymbol{\tau}} \left\| e^{\overline{A}t} \right\| \boldsymbol{d}t \right) \|B_0\| \sigma_w^2 q \log^{3/2} \frac{\boldsymbol{\kappa} q}{\delta}. \tag{20}$$

Finally, put (16), (17), (18), (19), and (20) together, to get the desired result.

■

A.2 Bounding cross products of state and Wiener process

Lemma 2 *In Algorithm 1, with probability at least* $1-\delta$*, we have*

$$\left\| \int_0^t \boldsymbol{x}_s \boldsymbol{d}\mathbb{W}_s^\top \right\| \lesssim \left(\int_0^{\boldsymbol{\tau}} \left\| e^{\overline{A}t} \right\| \boldsymbol{d}t \right) \left(\overline{\lambda}\left(\Sigma_{\mathbb{W}}\right) + \sigma_w^2 \right) p\left(p+q\right)^{1/2} \boldsymbol{\tau}^{1/2} \log\left(\frac{pq}{\delta}\right).$$

Proof. First, according to (15), we can write

$$\int_0^{\boldsymbol{\tau}} \boldsymbol{x}_t \boldsymbol{d}\mathbb{W}_t^\top = \Phi_1 + \Phi_2 + \Phi_3,$$

where

$$\Phi_1 = \int_0^{\boldsymbol{\tau}} e^{\overline{A}t} \boldsymbol{x}_0 \boldsymbol{d}\mathbb{W}_t^\top, \tag{21}$$

$$\Phi_2 = \int_0^{\boldsymbol{\tau}} \int_0^t e^{\overline{A}(t-s)} B_0 v(s) \boldsymbol{d}s \boldsymbol{d}\mathbb{W}_t^\top, \tag{22}$$

$$\Phi_3 = \int_0^{\boldsymbol{\tau}} \int_0^t e^{\overline{A}(t-s)} \boldsymbol{d}\mathbb{W}_s \boldsymbol{d}\mathbb{W}_t^\top. \tag{23}$$

Now, according to Ito Isometry [31], similar to (16), we have

$$\|\Phi_1\| \lesssim \overline{\lambda}\left(\Sigma_{\mathbb{W}}\right)^{1/2} \left(\int_0^{\boldsymbol{\tau}} \left\| e^{\overline{A}t} \right\| \boldsymbol{d}t \right) \|\boldsymbol{x}_0\| \sqrt{pq \log\left(\frac{pq}{\delta}\right)}, \tag{24}$$

with probability at least $1-\delta$. Moreover, in a procedure similar to the one that lead to (17), one can show that with probability at least $1-\delta$, it holds that

$$\|\Phi_2\| \lesssim \left(\int_0^{\boldsymbol{\tau}} \left\| e^{\overline{A}t} \right\| \boldsymbol{d}t \right) \overline{\lambda}\left(\Sigma_{\mathbb{W}}\right)^{1/2} \sigma_w pq^{1/2} \boldsymbol{\tau}^{1/2} \log\left(\frac{pq}{\delta}\right). \tag{25}$$

Therefore, we need to find a similar upper-bound for Φ_3. To that end, Ito formula provides

$$\boldsymbol{d}\left(e^{-\overline{A}s}\mathbb{W}_s\right) = -\overline{A}e^{-\overline{A}s}\mathbb{W}_s \boldsymbol{d}s + e^{-\overline{A}s}\boldsymbol{d}\mathbb{W}_s.$$

Therefore, integration gives

$$\int_0^t e^{-\overline{A}s}\boldsymbol{d}\mathbb{W}_s = e^{-\overline{A}t}\mathbb{W}_t + \overline{A}\int_0^t e^{-\overline{A}s}\mathbb{W}_s \boldsymbol{d}s,$$

which after rearranging and letting $\Psi_t = \int_0^t e^{\overline{A}(t-s)}\boldsymbol{d}\mathbb{W}_s$, leads to

$$\Psi_t \mathbb{W}_t^\top = \left(\int_0^t e^{\overline{A}(t-s)} \boldsymbol{d}\mathbb{W}_s \right) \mathbb{W}_t^\top = \mathbb{W}_t \mathbb{W}_t^\top + \overline{A} \left(\int_0^t e^{\overline{A}(t-s)} \mathbb{W}_s \boldsymbol{d}s \right) \mathbb{W}_t^\top.$$

Now, since $\boldsymbol{d}\Psi_t = \boldsymbol{d}\mathbb{W}_t$, Ito Isometry [31] implies that $\boldsymbol{d}\Psi_t \boldsymbol{d}\mathbb{W}_t^\top = \Sigma_{\mathbb{W}} \boldsymbol{d}t$. So, apply integration by part and use the above equation to get

$$\begin{aligned}
\Phi_3 = \int_0^{\boldsymbol{\tau}} \Psi_t \boldsymbol{d}\mathbb{W}_t^\top &= \int_0^{\boldsymbol{\tau}} \boldsymbol{d}\big(\Psi_t \mathbb{W}_t^\top\big) - \left(\int_0^{\boldsymbol{\tau}} \mathbb{W}_t \boldsymbol{d}\Psi_t^\top\right)^\top - \int_0^{\boldsymbol{\tau}} \boldsymbol{d}\Psi_t \boldsymbol{d}\mathbb{W}_t^\top \\
&= \Psi_{\boldsymbol{\tau}} \mathbb{W}_{\boldsymbol{\tau}}^\top - \left(\int_0^{\boldsymbol{\tau}} \mathbb{W}_t \boldsymbol{d}\mathbb{W}_t^\top\right)^\top - \Sigma_{\mathbb{W}} \boldsymbol{\tau} \\
&= \mathbb{W}_{\boldsymbol{\tau}} \mathbb{W}_{\boldsymbol{\tau}}^\top + \overline{A}\left(\int_0^{\boldsymbol{\tau}} e^{\overline{A}(\boldsymbol{\tau}-s)} \mathbb{W}_s \boldsymbol{d}s\right) \mathbb{W}_{\boldsymbol{\tau}}^\top - \left(\int_0^{\boldsymbol{\tau}} \mathbb{W}_t \boldsymbol{d}\mathbb{W}_t^\top\right)^\top - \Sigma_{\mathbb{W}} \boldsymbol{\tau}.
\end{aligned}$$

Therefore, every entry of Φ_3 is a quadratic function of the normally distributed random vectors $\mathbb{W}_{\boldsymbol{\tau}}, \int_0^{\boldsymbol{\tau}} e^{\overline{A}(\boldsymbol{\tau}-s)} \mathbb{W}_s \boldsymbol{d}s$. Note that we used the fact that

$$\mathbb{W}_{\boldsymbol{\tau}} \mathbb{W}_{\boldsymbol{\tau}}^\top = \int_0^{\boldsymbol{\tau}} \boldsymbol{d}\big(\mathbb{W}_t \mathbb{W}_t^\top\big) = \left(\int_0^{\boldsymbol{\tau}} \mathbb{W}_t \boldsymbol{d}\mathbb{W}_t^\top\right)^\top + \left(\int_0^{\boldsymbol{\tau}} \mathbb{W}_t \boldsymbol{d}\mathbb{W}_t^\top\right) + \Sigma_{\mathbb{W}} \boldsymbol{\tau}.$$

Thus, exponential inequalities for quadratic forms of normal random vectors [39] imply that for all $i, j = 1, \cdots, p$, it holds that

$$\big(\mathrm{e}_i^\top \Phi_3 \mathrm{e}_j\big)^2 \lesssim p\mathbb{E}\left[\big(\mathrm{e}_i^\top \Phi_3 \mathrm{e}_j\big)^2\right] \log^2 \frac{1}{\delta}, \tag{26}$$

since $\mathbb{E}\left[\mathrm{e}_i^\top \Phi_3 \mathrm{e}_j\right] = 0$. So, it suffices to find the expectation in (26). For that purpose, we use Ito Isometry [31] to obtain:

$$\begin{aligned}
\mathbb{E}\left[\big(\mathrm{e}_i^\top \Phi_3 \mathrm{e}_j\big)^2\right] &= \mathbb{E}\left[\left(\int_0^{\boldsymbol{\tau}} \mathrm{e}_i^\top \Psi_t \mathrm{e}_j^\top \Sigma_{\mathbb{W}}^{1/2} \boldsymbol{d}\Big(\Sigma_{\mathbb{W}}^{-1/2} \mathbb{W}_t\Big)\right)^2\right] = \mathbb{E}\left[\int_0^{\boldsymbol{\tau}} \left\|\mathrm{e}_i^\top \Psi_t \Sigma_{\mathbb{W}}^{1/2} \mathrm{e}_j\right\|^2 \boldsymbol{d}t\right] \\
&\leq \mathrm{e}_j^\top \Sigma_{\mathbb{W}} \mathrm{e}_j \mathbb{E}\left[\int_0^{\boldsymbol{\tau}} \big(\mathrm{e}_i^\top \Psi_t\big)^2 \boldsymbol{d}t\right] = \mathrm{e}_j^\top \Sigma_{\mathbb{W}} \mathrm{e}_j \mathbb{E}\left[\int_0^{\boldsymbol{\tau}} \left(\mathrm{e}_i^\top \int_0^t e^{\overline{A}(t-s)} \boldsymbol{d}\mathbb{W}_s\right)^2 \boldsymbol{d}t\right].
\end{aligned}$$

To proceed with the above expression, apply Fubini Theorem [31] to interchange the expected value with the integral, and then use Ito Isometry again:

$$\begin{aligned}
\mathbb{E}\left[\int_0^{\boldsymbol{\tau}} \left(\mathrm{e}_i^\top \int_0^t e^{\overline{A}(t-s)} \boldsymbol{d}\mathbb{W}_s\right)^2 \boldsymbol{d}t\right] &= \int_0^{\boldsymbol{\tau}} \mathbb{E}\left[\left(\mathrm{e}_i^\top \int_0^t e^{\overline{A}(t-s)} \Sigma_{\mathbb{W}}^{1/2} \boldsymbol{d}\Big(\Sigma_{\mathbb{W}}^{-1/2} \mathbb{W}_s\Big)\right)^2\right] \boldsymbol{d}t \\
&= \int_0^{\boldsymbol{\tau}} \mathrm{e}_i^\top \left(\int_0^t e^{\overline{A}(t-s)} \Sigma_{\mathbb{W}} e^{\overline{A}^\top (t-s)} \boldsymbol{d}s\right) \mathrm{e}_i \boldsymbol{d}t \\
&\leq \mathrm{e}_i^\top \left(\int_0^{\boldsymbol{\tau}} e^{\overline{A}s} \Sigma_{\mathbb{W}} e^{\overline{A}^\top s} \boldsymbol{d}s\right) \mathrm{e}_i \boldsymbol{\tau}.
\end{aligned}$$

Therefore, (26) yields to

$$\begin{aligned}\|\Phi_3\|^2 \leq \sum_{i,j=1}^{p} \left(\mathrm{e}_i^\top \Phi_3 \mathrm{e}_j\right)^2 &\lesssim \sum_{i,j=1}^{p} \left[\mathrm{e}_j^\top \Sigma_{\mathbb{W}} \mathrm{e}_j \mathrm{e}_i^\top \left(\int_0^{\boldsymbol{\tau}} e^{\overline{A}s} \Sigma_{\mathbb{W}} e^{\overline{A}^\top s} \boldsymbol{d}s\right) \mathrm{e}_i\right] \boldsymbol{\tau} p \log^2 \frac{p}{\delta} \\ &= \mathbf{tr}\left(\Sigma_{\mathbb{W}}\right) \mathbf{tr}\left(\int_0^{\boldsymbol{\tau}} e^{\overline{A}s} \Sigma_{\mathbb{W}} e^{\overline{A}^\top s} \boldsymbol{d}s\right) p\boldsymbol{\tau} \log^2 \frac{p}{\delta} \\ &\lesssim \mathbf{tr}\left(\Sigma_{\mathbb{W}}\right)^2 \left(\int_0^{\boldsymbol{\tau}} \left\|e^{\overline{A}s}\right\| \boldsymbol{d}s\right)^2 p\boldsymbol{\tau} \log^2 \frac{p}{\delta}. \end{aligned} \tag{27}$$

Finally, putting (24), (25), and (27) together, we obtain the desired result.

■

A.3 Concentration of normal posterior distribution in Algorithm 1

Lemma 3 *In Algorithm 1, letting $\overline{A} = A_0 + B_0 K$, suppose that*

$$\boldsymbol{\tau} \gtrsim \left(\int_0^{\boldsymbol{\tau}} \left\|\exp(\overline{A}s)\right\|^2 \boldsymbol{d}s\right) \left(\overline{\lambda}\left(\Sigma_{\mathbb{W}}\right) + \sigma_w^2 \|B_0\|^2\right) (p+q) \log \frac{1}{\delta}, \tag{28}$$

$$\frac{\boldsymbol{\kappa}}{\boldsymbol{\tau}} \gtrsim \frac{\sigma_w^2}{\sigma_w^2 \wedge \underline{\lambda}\left(\Sigma_{\mathbb{W}}\right)} \|B_0\| \left(1 \vee \|K\|\right) q \log \frac{\boldsymbol{\kappa} q}{\delta}. \tag{29}$$

Then, for the matrix $\widehat{\Sigma}_{\boldsymbol{\tau}}$ in (8)*, with probability at least $1-\delta$ we have*

$$\underline{\lambda}\left(\widehat{\Sigma}_{\boldsymbol{\tau}}\right) \gtrsim \boldsymbol{\tau} \left(\underline{\lambda}\left(\Sigma_{\mathbb{W}}\right) \wedge \sigma_w^2\right) \left(1 + \|K\|^2\right)^{-1}.$$

Proof. First, we can write the control action in (7) as $\boldsymbol{u}_t = K\boldsymbol{x}_t + v(t)$, for the piecewise-constant signal $v(t)$ in (14). Then, the dynamics in (1) provides

$$\boldsymbol{d}\boldsymbol{x}_t = \left(\overline{A}\boldsymbol{x}_t + B_0 v(t)\right) \boldsymbol{d}t + \boldsymbol{d}\mathbb{W}_t.$$

Therefore, similar to (15), one can solve the above stochastic differential equation to get

$$\boldsymbol{x}_t = e^{\overline{A}t}\boldsymbol{x}_0 + \int_0^t e^{\overline{A}(t-s)} \boldsymbol{d}\mathbb{W}_s + \int_0^t e^{\overline{A}(t-s)} B_0 v(s) \boldsymbol{d}s.$$

So, using the exponential inequalities for quadratic forms [39], with probability at least $1-\delta$, it holds that

$$\left\|\boldsymbol{x}_{\boldsymbol{\tau}} - e^{\overline{A}\boldsymbol{\tau}}\boldsymbol{x}_0\right\|^2 \lesssim \overline{\lambda}\left(\int_0^{\boldsymbol{\tau}} e^{\overline{A}s} \Sigma_{\mathbb{W}} e^{\overline{A}^\top s} \boldsymbol{d}s + \sigma_w^2 \sum_{n=0}^{\boldsymbol{\kappa}-1} J_n B_0 B_0^\top J_n^\top\right) \left(p + p^{1/2} \log \frac{1}{\delta}\right), \tag{30}$$

where

$$J_n = \int_{n\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}}^{(n+1)\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}} e^{\overline{A}s} \boldsymbol{d}s.$$

Furthermore, an application of Ito calculus [31] leads to $\boldsymbol{d}\boldsymbol{x}_t \boldsymbol{d}\boldsymbol{x}_t^\top = \boldsymbol{d}\mathbb{W}_t \boldsymbol{d}\mathbb{W}_t^\top = \Sigma_{\mathbb{W}} \boldsymbol{d}t$. Now, by defining the matrix valued processes

$$\Phi_t = \int_0^t \boldsymbol{x}_s \boldsymbol{x}_s^\top \boldsymbol{d}s, \qquad M_t = \int_0^t \boldsymbol{x}_s \boldsymbol{d}\mathbb{W}_s^\top + \int_0^t \boldsymbol{x}_s v(s)^\top B_0^\top \boldsymbol{d}s,$$

we obtain

$$\begin{aligned}\boldsymbol{d}\left(\boldsymbol{x}_t\boldsymbol{x}_t^\top\right) &= \boldsymbol{x}_t\boldsymbol{d}\boldsymbol{x}_t^\top + \boldsymbol{d}\boldsymbol{x}_t\boldsymbol{x}_t^\top + \boldsymbol{d}\boldsymbol{x}_t\boldsymbol{d}\boldsymbol{x}_t^\top \\ &= \boldsymbol{x}_t\left(\left(\overline{A}\boldsymbol{x}_t + B_0 v(t)\right)\boldsymbol{d}t + \boldsymbol{d}\mathbb{W}_t\right)^\top \\ &+ \left(\left(\overline{A}\boldsymbol{x}_t + B_0 v(t)\right)\boldsymbol{d}t + \boldsymbol{d}\mathbb{W}_t\right)\boldsymbol{x}_t^\top + \Sigma_{\mathbb{W}}\boldsymbol{d}t \\ &= \boldsymbol{d}\Phi_t\overline{A}^\top + \overline{A}\boldsymbol{d}\Phi_t + \boldsymbol{d}M_t + \boldsymbol{d}M_t^\top + \Sigma_{\mathbb{W}}\boldsymbol{d}t.\end{aligned}$$

Thus, after integrating both sides of the above equality, we obtain

$$\Phi_t\overline{A}^\top + \overline{A}\Phi_t + M_t + M_t^\top + t\Sigma_{\mathbb{W}} + \boldsymbol{x}_0\boldsymbol{x}_0^\top - \boldsymbol{x}_t\boldsymbol{x}_t^\top = 0.$$

Because all eigenvalues of $\overline{A}$ are in the open left half-plane, we can solve the above equation for Φ_t, to get

$$\Phi_t = \int_0^\infty \exp\left(\overline{A}s\right)\left[M_t + M_t^\top + t\Sigma_{\mathbb{W}} + \boldsymbol{x}_0\boldsymbol{x}_0^\top - \boldsymbol{x}_t\boldsymbol{x}_t^\top\right]\exp\left(\overline{A}^\top s\right)\boldsymbol{d}s. \tag{31}$$

Next, putting Lemma 1, Lemma 2, and (30) together, as long as (28) holds, with probability at least $1-\delta$ we have

$$\underline{\lambda}\left(M_{\boldsymbol{\tau}} + M_{\boldsymbol{\tau}}^\top + \boldsymbol{\tau}\Sigma_{\mathbb{W}} + \boldsymbol{x}_0\boldsymbol{x}_0^\top - \boldsymbol{x}_{\boldsymbol{\tau}}\boldsymbol{x}_{\boldsymbol{\tau}}^\top\right) \gtrsim \boldsymbol{\tau}\underline{\lambda}\left(\Sigma_{\mathbb{W}}\right).$$

Thus, (31) implies that $\underline{\lambda}\left(\Phi_{\boldsymbol{\tau}}\right) \gtrsim \boldsymbol{\tau}\underline{\lambda}\left(\Sigma_{\mathbb{W}}\right)$. To proceed, consider the matrix $\widehat{\Sigma}_{\boldsymbol{\tau}}$ in (8), which comprises two signals $\boldsymbol{x}_t, v(t)$. The empirical covariance matrix of the state signal is studied above, while for the piecewise-constant randomization signal $v(t)$ in (14), we have

$$\int_0^t v(s)v(s)^\top\boldsymbol{d}s = \sum_{n=0}^{\boldsymbol{\kappa}-1}\int_{n\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}}^{(n+1)\boldsymbol{\tau}\boldsymbol{\kappa}^{-1}} w_n w_n^\top\boldsymbol{d}s = \boldsymbol{\tau}\boldsymbol{\kappa}^{-1}\sum_{n=0}^{\boldsymbol{\kappa}-1} w_n w_n^\top.$$

Thus, according to Theorem 3, similar to (19) we have

$$\left\|\sum_{n=0}^{\boldsymbol{\kappa}-1} w_n w_n^\top - \boldsymbol{\kappa}\sigma_w^2 I_q\right\| \lesssim \boldsymbol{\kappa}^{1/2}\sigma_w^2 q\log^{3/2}\frac{\boldsymbol{\kappa}q}{\delta},$$

with probability at least $1-\delta$, which for

$$H_{\boldsymbol{\tau}} = \int_0^{\boldsymbol{\tau}}\begin{bmatrix}0_p\\ v(s)\end{bmatrix}\begin{bmatrix}0_p\\ v(s)\end{bmatrix}^\top\boldsymbol{d}s - \boldsymbol{\tau}\sigma_w^2\begin{bmatrix}0_{p\times p} & 0_{p\times q}\\ 0_{q\times p} & I_q\end{bmatrix},$$

leads to

$$\|H_{\boldsymbol{\tau}}\| \lesssim \sigma_w^2 q\log^{3/2}\frac{\boldsymbol{\kappa}q}{\delta}, \tag{32}$$

because $\boldsymbol{\kappa} \gtrsim \boldsymbol{\tau}^2$.

Next, using $\boldsymbol{z}_s = \left[\boldsymbol{x}_s^\top, \boldsymbol{x}_s^\top K^\top + v(s)^\top\right]^\top$, the matrix $\widehat{\Sigma}_{\boldsymbol{\tau}}$ can be written as

$$\widehat{\Sigma}_{\boldsymbol{\tau}} = \begin{bmatrix}I_p\\ K\end{bmatrix}\Phi_{\boldsymbol{\tau}}\begin{bmatrix}I_p\\ K\end{bmatrix}^\top + \boldsymbol{\tau}\sigma_w^2\begin{bmatrix}0_{p\times p} & 0_{p\times q}\\ 0_{q\times p} & I_q\end{bmatrix} + F_{\boldsymbol{\tau}} + H_{\boldsymbol{\tau}}, \tag{33}$$

where

$$F_{\boldsymbol{\tau}} = \int_0^{\boldsymbol{\tau}}\left(\begin{bmatrix}I_p\\ K\end{bmatrix}\boldsymbol{x}_s\begin{bmatrix}0_p\\ v(s)\end{bmatrix}^\top + \begin{bmatrix}0_p\\ v(s)\end{bmatrix}\boldsymbol{x}_s^\top\begin{bmatrix}I_p\\ K\end{bmatrix}^\top\right)\boldsymbol{d}s.$$

However, Lemma 1 and $\boldsymbol{\kappa} \gtrsim \boldsymbol{\tau}^2$ give a high probability upper-bound for the above matrix:

$$\|F_{\boldsymbol{\tau}}\| \lesssim (1\vee\|K\|)\left(\overline{\lambda}\left(\Sigma_{\mathbb{W}}\right) + \sigma_w^2\right)\left(pq^{1/2}\boldsymbol{\tau}^{1/2}\log\frac{pq}{\delta} + q\log^{3/2}\frac{\boldsymbol{\kappa}q}{\delta}\right). \tag{34}$$

In the sequel, we show that with probability at least $1-\delta$, it holds that

$$\underline{\lambda}\left(\begin{bmatrix} I_p \\ K \end{bmatrix} \Phi_{\boldsymbol{\tau}} \begin{bmatrix} I_p \\ K \end{bmatrix}^\top + \boldsymbol{\tau}\sigma_w^2 \begin{bmatrix} 0_{p\times p} & 0_{p\times q} \\ 0_{q\times p} & I_q \end{bmatrix}\right) \gtrsim \boldsymbol{\tau}\left(\underline{\lambda}\left(\Sigma_{\mathbb{W}}\right) \wedge \sigma_w^2\right)\left(1+\|K\|^2\right)^{-1},$$

which, according to (32), (33), and (34), implies the desired result. To show the above least eigenvalue inequality, we use $\underline{\lambda}\left(\Phi_{\boldsymbol{\tau}}\right) \gtrsim \boldsymbol{\tau}\underline{\lambda}\left(\Sigma_{\mathbb{W}}\right)$ to obtain

$$\underline{\lambda}\left(\begin{bmatrix} I_p \\ K \end{bmatrix} \Phi_{\boldsymbol{\tau}} \begin{bmatrix} I_p \\ K \end{bmatrix}^\top + \boldsymbol{\tau}\sigma_w^2 \begin{bmatrix} 0_{p\times p} & 0_{p\times q} \\ 0_{q\times p} & I_q \end{bmatrix}\right) \gtrsim \boldsymbol{\tau}\left(\underline{\lambda}\left(\Sigma_{\mathbb{W}}\right) \wedge \sigma_w^2\right)\underline{\lambda}\left(\begin{bmatrix} I_p & K^\top \\ K & KK^\top + I_q \end{bmatrix}\right).$$

However, block matrix inversion gives

$$\underline{\lambda}\left(\begin{bmatrix} I_p & K^\top \\ K & KK^\top + I_q \end{bmatrix}\right) = \overline{\lambda}\left(\begin{bmatrix} I_p & K^\top \\ K & KK^\top + I_q \end{bmatrix}^{-1}\right)^{-1} = \overline{\lambda}\left(\begin{bmatrix} K^\top K + I_p & -K^\top \\ -K & I_q \end{bmatrix}\right)^{-1},$$

that is clearly at least $\left(1+\|K\|^2\right)^{-1}$, apart from a constant factor. Therefore, we get the desired result. ■

A.4 Approximation of true drift parameter by Algorithm 1

Lemma 4 *Suppose that $\widehat{\boldsymbol{\theta}}$ is given by Algorithm 1. Then, with probability at least $1-\delta$, we have*

$$\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathbf{0}}\right\| \lesssim \frac{p\left(p+q\right)^{1/2}}{\boldsymbol{\tau}^{1/2}} \frac{\overline{\lambda}\left(\Sigma_{\mathbb{W}}\right) \vee \sigma_w^2}{\underline{\lambda}\left(\Sigma_{\mathbb{W}}\right) \wedge \sigma_w^2} \left(1+\|K\|\right)^3 \log\left(\frac{pq\boldsymbol{\kappa}}{\delta}\right). \tag{35}$$

Proof. First, consider the mean matrix of the Gaussian posterior distribution. Using the data generation mechanism $\boldsymbol{dx}_t = \boldsymbol{\theta_0}^\top \boldsymbol{z}_t \boldsymbol{dt} + \boldsymbol{d}\mathbb{W}_t$, we have

$$\widehat{\mu}_{\boldsymbol{\tau}} = \widehat{\Sigma}_{\boldsymbol{\tau}}^{-1} \int_0^{\boldsymbol{\tau}} \boldsymbol{z}_s \boldsymbol{dx}_s^\top = \widehat{\Sigma}_{\boldsymbol{\tau}}^{-1} \left(\int_0^{\boldsymbol{\tau}} \boldsymbol{z}_s \boldsymbol{z}_s^\top \boldsymbol{ds}\boldsymbol{\theta_0} + \int_0^{\boldsymbol{\tau}} \boldsymbol{z}_s \boldsymbol{d}\mathbb{W}_s^\top\right) = \boldsymbol{\theta_0} - \widehat{\Sigma}_{\boldsymbol{\tau}}^{-1}\left(\boldsymbol{\theta_0} - \int_0^{\boldsymbol{\tau}} \boldsymbol{z}_s \boldsymbol{d}\mathbb{W}_s^\top\right),$$

where we used the definition of $\widehat{\Sigma}_\tau$ in (8). Now, the sample $\widehat{\boldsymbol{\theta}}$ from $\mathcal{D}_{\boldsymbol{\tau}}$ can be written as $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\tau}} = \widehat{\mu}_{\boldsymbol{\tau}} + \widehat{\Sigma}_{\boldsymbol{\tau}}^{-1/2}\Phi$, where $\Phi \sim \boldsymbol{N}\left(0_{(p+q)\times p}, I_{p+q}\right)$ is a standard normal random matrix, as defined in the notation. So, for the error matrix, it holds that

$$\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathbf{0}}\right\| \leq \left\|\widehat{\Sigma}_{\boldsymbol{\tau}}^{-1}\right\| \left(\|\boldsymbol{\theta_0}\| + \left\|\int_0^{\boldsymbol{\tau}} \boldsymbol{z}_s \boldsymbol{d}\mathbb{W}_s^\top\right\|\right) + \left\|\widehat{\Sigma}_{\boldsymbol{\tau}}^{-1}\right\|^{1/2} \|\Phi\|.$$

To proceed towards bounding the above error matrix, use

$$\int_0^{\boldsymbol{\tau}} \boldsymbol{z}_s \boldsymbol{d}\mathbb{W}_s^\top = \int_0^{\boldsymbol{\tau}} \left(\begin{bmatrix} I_p \\ K \end{bmatrix} \boldsymbol{x}_s + \begin{bmatrix} 0 \\ v(s) \end{bmatrix}\right) \boldsymbol{d}\mathbb{W}_s^\top,$$

to obtain

$$\left\|\int_0^{\boldsymbol{\tau}} \boldsymbol{z}_s \boldsymbol{d}\mathbb{W}_s^\top\right\| \leq \left(1 \vee \|K\|\right) \left\|\int_0^{\boldsymbol{\tau}} \boldsymbol{x}_s \boldsymbol{d}\mathbb{W}_s^\top\right\| + \left\|\int_0^{\boldsymbol{\tau}} v(s) \boldsymbol{d}\mathbb{W}_s^\top\right\|,$$

To proceed, note that with probability at least $1-\delta$, we have

$$\|\Phi\|^2 \lesssim p(p+q) \log \frac{p(p+q)}{\delta}.$$

Now, by putting this together with the results of Lemma 2, Lemma 3, and (25), we get the desired result.

■

A.5 Eigenvalue ratio bound for sum of two matrices

Lemma 5 *Suppose that M, E are $p \times p$ matrices, and let $M = \Gamma^{-1}\Lambda\Gamma$ be the Jordan diagonalization of M. So, for some positive integer k, we have $\Lambda \in \mathbb{C}^{p\times p} = \mathrm{diag}\,(\Lambda_1, \cdots, \Lambda_k)$, where the blocks $\Lambda_1, \cdots, \Lambda_k$ are Jordan matrices of the form*

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

Then, let $\boldsymbol{r} = \max\limits_{1\le i\le k} \boldsymbol{r}_i \le p$, and define $\Delta_M(E)$ as the difference between the largest real-part of the eigenvalues of $M + E$ and that of M. Then, it holds that

$$\Delta_M(E) \le \left(1 \vee \boldsymbol{r}^{1/2}\|E\|\mathrm{cond}\,(\Gamma)\right)^{1/\boldsymbol{r}}, \tag{36}$$

where $\mathrm{cond}\,(\Gamma)$ is the condition number of Γ: $\mathrm{cond}\,(\Gamma) = \overline{\lambda}\left(\Gamma^\top\Gamma\right)^{1/2}\underline{\lambda}\left(\Gamma^\top\Gamma\right)^{-1/2}$.

Proof. Since the expression on the right-hand-side of (36) is positive, it is enough to consider an eigenvalue λ of $M + E$ which is not an eigenvalue of M, and show that $\Re\,(\lambda) - \log\overline{\lambda}\,(\exp\,(M))$ is less than the expression on the RHS of (36). So, for such λ, the matrix $M - \lambda I_p$ is non-singular, while $M + E - \lambda I_p$ is singular. Let the vector $v \ne 0$ be such that $(M + E - \lambda I_p)\,v = 0$, which by Jordan diagonalization above implies that

$$v = -\Gamma^{-1}\,(\Lambda - \lambda I)^{-1}\,\Gamma E v. \tag{37}$$

Then, $\Lambda = \mathrm{diag}\,(\Lambda_1, \cdots, \Lambda_k)$ indicates that $\Lambda - \lambda I$ and $(\Lambda - \lambda I)$ are block diagonal, the latter consisting of the blocks $\mathrm{diag}\left((\Lambda_1 - \lambda I_{\boldsymbol{r}_1})^{-1}, \cdots, (\Lambda_k - \lambda I_{\boldsymbol{r}_k})^{-1}\right)$.

Now, multiplications show that

$$(\Lambda_i - \lambda I_{\boldsymbol{r}_i})^{-1} = -\begin{bmatrix} (\lambda - \lambda_i)^{-1} & (\lambda - \lambda_i)^{-2} & \cdots & (\lambda - \lambda_i)^{-\boldsymbol{r}_i} \\ 0 & (\lambda - \lambda_i)^{-1} & \cdots & (\lambda - \lambda_i)^{-\boldsymbol{r}_i+1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & (\lambda - \lambda_i)^{-1} \end{bmatrix}.$$

Therefore, according to the definition of matrix operator norms in Section 1, we obtain

$$\left\|(\Lambda_i - \lambda I_{\boldsymbol{r}_i})^{-1}\right\|^2 \le \boldsymbol{r}\left(1 \vee |\lambda - \lambda_i|^{-\boldsymbol{r}}\right)^2.$$

Putting these bounds for the blocks of $(\Lambda - \lambda I)^{-1}$ together, (37) leads to

$$\begin{aligned} 1 &\le \left\|(\Lambda - \lambda I)^{-1}\right\|\|\Gamma\|\left\|\Gamma^{-1}\right\|\|E\| \\ &\le \boldsymbol{r}^{1/2}\mathrm{cond}\,(\Gamma)\,\|E\|\max_{1\le i\le k}\left(1 \wedge |\lambda - \lambda_i|^{\boldsymbol{r}}\right)^{-1} \\ &\le \boldsymbol{r}^{1/2}\mathrm{cond}\,(\Gamma)\,\|E\|\left(1 \wedge \left(\Re(\lambda) - \log\overline{\lambda}\,(\exp\,(M))\right)^{\boldsymbol{r}}\right)^{-1}. \end{aligned}$$

To see the last inequality above, note that if $\Re(\lambda) - \log\overline{\lambda}\,(\exp\,(M))$ is positive, then it is larger than all the terms $|\lambda - \lambda_i|$, for $i = 1, \cdots, k$. Thus, for

$$\Re\,(\lambda) = \log\overline{\lambda}\,(\exp\,(M + E)),$$

we obtain (36). ■

B Proof of Theorem 2

To establish the rates of exploration Algorithm 2 performs, we utilize Lemma 8, which indicates that

$$\|\widehat{\mu}_{\boldsymbol{\tau}_n} - \boldsymbol{\theta_0}\| \lesssim \left\|\widehat{\Sigma}_{\boldsymbol{\tau}_n}^{-1/2}\right\| \left\|\widehat{\Sigma}_{\boldsymbol{\tau}_n}^{-1/2} \int_0^{\boldsymbol{\tau}_n} \boldsymbol{x}_t d\mathbb{W}_t^\top\right\| \lesssim \underline{\lambda}\left(\widehat{\Sigma}_{\boldsymbol{\tau}_n}\right)^{-1/2} \left(p(p+q)\overline{\lambda}\left(\Sigma_{\mathbb{W}}\right) \log \overline{\lambda}\left(\widehat{\Sigma}_{\boldsymbol{\tau}_n}\right)\right)^{1/2}.$$

Now, (51) gives $\log \overline{\lambda}\left(\widehat{\Sigma}_{\boldsymbol{\tau}_n}\right) \lesssim \log \boldsymbol{\tau}_n$, while Lemma 9 provides $\underline{\lambda}\left(\widehat{\Sigma}_{\boldsymbol{\tau}_n}\right) \gtrsim \boldsymbol{\tau}_n^{1/2} \underline{\lambda}\left(\Sigma_{\mathbb{W}}\right)$. Moreover, since $\widehat{\Sigma}_{\boldsymbol{\tau}_n}^{1/2}\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \widehat{\mu}_{\boldsymbol{\tau}_n}\right)$ is a standard normal $(p+q) \times p$ matrix, we have

$$\left\|\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \widehat{\mu}_{\boldsymbol{\tau}_n}\right\| \lesssim \boldsymbol{\tau}_n^{-1/4} \underline{\lambda}\left(\Sigma_{\mathbb{W}}\right)^{-1/2} \left(p(p+q)\log(pq)\right)^{1/2}.$$

Thus, we obtain the desired result for the estimation error.

To proceed toward establishing the regret bound, Lemma 7 shows that we need to integrate $\left\|\boldsymbol{u}_t + Q_u^{-1}\widehat{B}_n^\top P\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}\right)\boldsymbol{x}_t\right\|^2$ over the stabilized period of Algorithm 2: $\boldsymbol{\tau}_0 \leq t \leq T$:

$$\operatorname{Reg}(T) \lesssim \left(\overline{\lambda}\left(\Sigma_{\mathbb{W}}\right) + \sigma_w^2\right)\boldsymbol{\tau}_0 + \int_{\boldsymbol{\tau}_0}^{T} \left\|\boldsymbol{u}_t + Q_u^{-1}B_0^\top P\left(\boldsymbol{\theta_0}\right)\boldsymbol{x}_t\right\|^2 dt.$$

Further, according to (51), for $\boldsymbol{\tau}_{n-1} < T \leq \boldsymbol{\tau}_n$, we have

$$\int_{\boldsymbol{\tau}_0}^{T} \left\|\boldsymbol{u}_t + Q_u^{-1}B_0^\top P\left(\boldsymbol{\theta_0}\right)\boldsymbol{x}_t\right\|^2 dt \lesssim \overline{\lambda}\left(\Sigma_{\mathbb{W}}\right) \sum_{i=0}^{n-1} \left(\boldsymbol{\tau}_{i+1} - \boldsymbol{\tau}_i\right) \left\|K\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}\right) - K\left(\boldsymbol{\theta_0}\right)\right\|^2.$$

On the other hand, Lemma 12 implies that

$$\left\|K\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}\right) - K\left(\boldsymbol{\theta_0}\right)\right\| \lesssim \frac{\left\|P\left(\boldsymbol{\theta_0}\right)\right\|^3}{\underline{\lambda}\left(Q_x\right)\underline{\lambda}\left(Q_u\right)^2} \left\|\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}} - \boldsymbol{\theta_0}\right\|.$$

Thus, we have

$$\operatorname{Reg}(T) \lesssim \left(\overline{\lambda}\left(\Sigma_{\mathbb{W}}\right) + \sigma_w^2\right)\boldsymbol{\tau}_0 + \frac{\overline{\lambda}\left(\Sigma_{\mathbb{W}}\right)^2}{\underline{\lambda}\left(\Sigma_{\mathbb{W}}\right)} \frac{\left\|P\left(\boldsymbol{\theta_0}\right)\right\|^6}{\underline{\lambda}\left(Q_x\right)^2 \underline{\lambda}\left(Q_u\right)^4} p(p+q) \sum_{i=0}^{n-1} \left(\boldsymbol{\tau}_{i+1} - \boldsymbol{\tau}_i\right) \frac{\log \boldsymbol{\tau}_i}{\boldsymbol{\tau}_i^{1/2}}.$$

Thus, according to (12), we obtain the desired regret bound result in Theorem 2.

B.1 Geometry of drift parameters and optimal policies

Lemma 6 *For the drift parameter $\boldsymbol{\theta_1}$, and for $X \in \mathbb{R}^{p\times p}, Y \in \mathbb{R}^{p\times q}$, define*

$$\Delta_{\boldsymbol{\theta_1}}(X,Y) = P\left(\boldsymbol{\theta_1}\right)Y + \int_0^{\infty} e^{\overline{A}_1^\top t}\left[M\left(X,Y\right)^\top P\left(\boldsymbol{\theta_1}\right) + P\left(\boldsymbol{\theta_1}\right)M\left(X,Y\right)\right] e^{\overline{A}_1 t} B_1 dt,$$

where $\overline{A}_1 = A_1 - B_1 Q_u^{-1} B_1^\top P\left(\boldsymbol{\theta_1}\right)$ and $M\left(X,Y\right) = X - YQ_u^{-1}B_1^\top P\left(\boldsymbol{\theta_1}\right)$. Then, $\Delta_{\boldsymbol{\theta_1}}(X,Y)$ is the directional derivative of $B^\top P\left(\boldsymbol{\theta}\right)$ at $\boldsymbol{\theta_1}$ in the direction $[X,Y]$. Importantly, the tangent space of the manifold of matrices $\boldsymbol{\theta} \in \mathbb{R}^{p\times(p+q)}$ that satisfy $B^\top P\left(\boldsymbol{\theta}\right) = B_1^\top P\left(\boldsymbol{\theta_1}\right)$ at $\boldsymbol{\theta_1}$ contains all matrices X, Y that $\Delta_{\boldsymbol{\theta_1}}(X,Y) = 0$.

Proof. First, note that according to the Lipschitz continuity of $P\left(\boldsymbol{\theta}\right)$ in Lemma 12, the directional derivative exists and is well-defined, as long as $\left\|P\left(\boldsymbol{\theta_1}\right)\right\| < \infty$. However, Lemma 11 provides that $P\left(\boldsymbol{\theta_1}\right)$ is finite in a neighborhood of $\boldsymbol{\theta_0}$, and so the required condition holds. Below, we start by establishing the second result to identify the tangent space, and then prove the general result on the directional derivative.

To proceed, let $\boldsymbol{\theta} = \boldsymbol{\theta_1} + \epsilon [X, Y]^\top$ be such that $B^\top P(\boldsymbol{\theta}) = B_1^\top P(\boldsymbol{\theta_1})$, and denote $K(\boldsymbol{\theta_1}) = -Q_u^{-1} B_1^\top P(\boldsymbol{\theta_1})$. So, the directional derivative of $P(\boldsymbol{\theta_1})$ along the matrix $[X, Y]^\top$ can be found as follows. First, denoting the closed-loop transition matrix by $\overline{A} = A - BQ_u^{-1} B^\top P(\boldsymbol{\theta})$, since

$$\overline{A}^\top P(\boldsymbol{\theta}) + P(\boldsymbol{\theta}) \overline{A} + Q_x + K(\boldsymbol{\theta})^\top Q_u K(\boldsymbol{\theta}) = 0,$$

we have

$$\begin{aligned} & \left(\overline{A}_1 + \epsilon X + \epsilon Y K(\boldsymbol{\theta_1})\right)^\top P(\boldsymbol{\theta}) + P(\boldsymbol{\theta}) \left(\overline{A}_1 + \epsilon X + \epsilon Y K(\boldsymbol{\theta_1})\right) \\ = \quad & -Q_x - K(\boldsymbol{\theta_1})^\top Q_u K(\boldsymbol{\theta_1}) = \overline{A}_1^\top P(\boldsymbol{\theta_1}) + P(\boldsymbol{\theta_1}) \overline{A}_1. \end{aligned}$$

For the matrix $E = \lim_{\epsilon \to 0} \epsilon^{-1} (P(\boldsymbol{\theta}) - P(\boldsymbol{\theta_1}))$, the latter result implies that

$$\overline{A}_1^\top E + E \overline{A}_1 + (X + Y K(\boldsymbol{\theta_1}))^\top P(\boldsymbol{\theta_1}) + P(\boldsymbol{\theta_1}) (X + Y K(\boldsymbol{\theta_1})) = 0.$$

Then, since all eigenvalues of $\overline{A}_1$ are in the open left half-plane, the above Lyapunov equation for E leads to the integral form

$$E = \int_0^\infty e^{\overline{A}_1^\top t} \left((X + Y K(\boldsymbol{\theta_1}))^\top P(\boldsymbol{\theta_1}) + P(\boldsymbol{\theta_1}) (X + Y K(\boldsymbol{\theta_1})) \right) e^{\overline{A}_1 t} \boldsymbol{d}t.$$

On the other hand, $K(\boldsymbol{\theta}) = -Q_u^{-1} B^\top P(\boldsymbol{\theta})$ gives

$$0 = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(B^\top P(\boldsymbol{\theta}) - B_1^\top P(\boldsymbol{\theta_1}) \right) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[\left(B^\top - B_1^\top \right) P(\boldsymbol{\theta}) - B_1^\top \left(P(\boldsymbol{\theta_1}) - P(\boldsymbol{\theta}) \right) \right],$$

which, according to the definitions of $E, M(X, Y)$, implies the desired result about the tangent space of the manifold under consideration.

Next, to establish the more general result on the directional derivative, we use the directional derivative of $P(\boldsymbol{\theta})$ in (65):

$$\int_0^\infty e^{\overline{A}_1^\top t} \left(P(\boldsymbol{\theta_1}) [X + Y K(\boldsymbol{\theta_1})] + [X + Y K(\boldsymbol{\theta_1})]^\top P(\boldsymbol{\theta_1}) \right) e^{\overline{A}_1 t} \boldsymbol{d}t.$$

Finally, since the directional derivative for $B^\top$ is Y, for $B^\top P(\boldsymbol{\theta})$, by the product rule it is $\Delta_{\boldsymbol{\theta_1}}(X, Y)$, which finishes the proof. ■

B.2 Regret bounds in terms of deviations in control actions

Lemma 7 *Let $\boldsymbol{u}_t$ be the action that Algorithm 2 takes at time t. Then, for the regret of Algorithm 2, it holds that*

$$\begin{aligned} \mathrm{Reg}(T) \quad \lesssim \quad & \left(\overline{\lambda}(\Sigma_{\mathbb{W}}) + \sigma_w^2\right) \boldsymbol{\tau}_0 \left\| K + Q_u^{-1} B_0^\top P(\boldsymbol{\theta_0}) \right\|^2 \\ + \quad & \int_{\boldsymbol{\tau}_0}^{T} \left\| \boldsymbol{u}_t + Q_u^{-1} B_0^\top P(\boldsymbol{\theta_0}) \boldsymbol{x}_t \right\|^2 \boldsymbol{d}t + {x_T^*}^\top P(\boldsymbol{\theta_0}) x_T^*, \end{aligned}$$

where x_T^* is the terminal state under the optimal trajectory $\boldsymbol{\pi}_{\mathbf{opt}}$ in (6)*.*

Proof. First, denote the optimal linear feedback of $\boldsymbol{\pi}_{\mathbf{opt}}$ in (6) by $\boldsymbol{u}_t = K(\boldsymbol{\theta_0}) \boldsymbol{x}_t$, where $K(\boldsymbol{\theta_0}) = -Q_u^{-1} B_0 P(\boldsymbol{\theta_0})$. According to the episodic structure of Algorithm 2, for $\boldsymbol{\tau}_n \le t < \boldsymbol{\tau}_{n+1}$, denote

$$K_t = -Q_u^{-1} \widehat{B}_n^\top P\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}\right).$$

We first consider the regret of Algorithm 2 after finishing stabilization by running Algorithm 1; i.e., for $\boldsymbol{\tau}_0 \le t \le T$. Fix some small $\epsilon > 0$, that we will let decay later. We proceed by finding an approximation of the regret through sampling at times $\boldsymbol{\tau}_0 + k\epsilon$, for non-negative integers k. To do that, denote $N = \lceil (T - \boldsymbol{\tau}_0)/\epsilon \rceil$, and define the sequence of policies $\{\widehat{\boldsymbol{\pi}}_i\}_{i=0}^N$ according to

$$\widehat{\boldsymbol{\pi}}_i = \begin{cases} \boldsymbol{u}_t = K_t \boldsymbol{x}_t & t < \boldsymbol{\tau}_0 + i\epsilon \\ \boldsymbol{u}_t = K(\boldsymbol{\theta_0}) \boldsymbol{x}_t & t \ge \boldsymbol{\tau}_0 + i\epsilon \end{cases}.$$

That is, the policy $\widehat{\boldsymbol{\pi}}_i$ switches to the optimal feedback at time $\boldsymbol{\tau}_0 + i\epsilon$. So, the zeroth policy $\widehat{\boldsymbol{\pi}}_0$ corresponds to applying the optimal policy $\boldsymbol{\pi}_{\mathbf{opt}}$ after stabilization at time $\boldsymbol{\tau}_0$, while the last one $\widehat{\boldsymbol{\pi}}_N$ is nothing but the one in Algorithm 2, that we denote by $\widehat{\boldsymbol{\pi}}$, for the sake of brevity. As such, we have $\mathrm{Reg}_{\widehat{\boldsymbol{\pi}}_0}(T) = 0$, and the telescopic summation below holds true:

$$\mathrm{Reg}_{\widehat{\boldsymbol{\pi}}}(T) = \sum_{i=0}^{N-1} \Big(\mathrm{Reg}_{\widehat{\boldsymbol{\pi}}_{i+1}}(T) - \mathrm{Reg}_{\widehat{\boldsymbol{\pi}}_i}(T) \Big). \tag{38}$$

Now, to consider the difference $\mathrm{Reg}_{\widehat{\boldsymbol{\pi}}_{i+1}}(T) - \mathrm{Reg}_{\widehat{\boldsymbol{\pi}}_i}(T)$, for a fixed i in the range $0 \leq i < N$, denote $t_1 = \boldsymbol{\tau}_0 + i\epsilon$ and let $\boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_i}, \boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_{i+1}}$ be the state trajectories under $\widehat{\boldsymbol{\pi}}_i, \widehat{\boldsymbol{\pi}}_{i+1}$, respectively. By definition, we have $\boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_i} = \boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_{i+1}}$, for all $t \leq t_1$. So, we drop the policy superscript and use $\boldsymbol{x}_{t_1}$ to refer to the states of both of them at time t_1. Therefore, as long as $t_1 \leq t < t_1 + \epsilon$, similar to (15), the solutions of the stochastic differential equation are

$$\begin{aligned} \boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_i} &= e^{\overline{A}_0(t-t_1)}\boldsymbol{x}_{t_1} + \int_{t_1}^{t} e^{\overline{A}_0(t-s)}\boldsymbol{d}\mathbb{W}_s, \\ \boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_{i+1}} &= e^{\overline{A}(t-t_1)}\boldsymbol{x}_{t_1} + \int_{t_1}^{t} e^{\overline{A}(t-s)}\boldsymbol{d}\mathbb{W}_s, \end{aligned}$$

where $\overline{A}_0 = A_0 + B_0 K(\boldsymbol{\theta}_{\mathbf{0}})$ and $\overline{A} = A_0 + B_0 K_{t_1}$ are the closed-loop transition matrices under $\widehat{\boldsymbol{\pi}}_i$ and $\widehat{\boldsymbol{\pi}}_{i+1}$, respectively. To work with the above two state trajectories, we define some notations for convenience:

$$\begin{aligned} M_0 &= Q_x + K(\boldsymbol{\theta}_{\mathbf{0}})^\top Q_u K(\boldsymbol{\theta}_{\mathbf{0}}), \\ M_1 &= Q_x + K_{t_1} Q_u K_{t_1}, \\ y_t &= \boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_{i+1}} - \boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_i}, \\ E_t &= e^{\overline{A}(t-t_1)} - e^{\overline{A}_0(t-t_1)}. \end{aligned}$$

Thus, letting

$$Z_t = \int_{t_1}^{t} \Big[e^{\overline{A}(t-s)} - e^{\overline{A}_0(t-s)} \Big] \boldsymbol{d}\mathbb{W}_s,$$

it holds that $y_t = E_t \boldsymbol{x}_{t_1} + Z_t + O(\epsilon^2)$. Further, for the observation signal $\boldsymbol{z}_t$ and the cost matrix Q defined in Section 2, we have

$$\begin{aligned} & \int_{t_1}^{t_1+\epsilon} \Big(\big\| Q^{1/2} \boldsymbol{z}_t(\widehat{\boldsymbol{\pi}}_{i+1}) \big\|^2 - \big\| Q^{1/2} \boldsymbol{z}_t(\widehat{\boldsymbol{\pi}}_i) \big\|^2 \Big) \boldsymbol{d}t \\ = & \int_{t_1}^{t_1+\epsilon} \Big[\Big(\boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_i} + y_t \Big)^\top M_1 \Big(\boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_i} + y_t \Big) - {\boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_i}}^\top M_0 \boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_i} \Big] \boldsymbol{d}t \\ = & \int_{t_1}^{t_1+\epsilon} \Big[{\boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_i}}^\top S \boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_i} + 2 y_t^\top M_1 \boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_i} + y_t^\top M_1 y_t \Big] \boldsymbol{d}t, \end{aligned} \tag{39}$$

where $S = M_1 - M_0 = K_{t_1}^\top Q_u K_{t_1} - K(\boldsymbol{\theta}_{\mathbf{0}})^\top Q_u K(\boldsymbol{\theta}_{\mathbf{0}})$.

On the other hand, for $t \geq t_1 + \epsilon$, the evolutions of the state vectors are the same for the two policies and we have

$$\boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_i} = e^{\overline{A}_0(t-t_1-\epsilon)} \boldsymbol{x}_{t_1+\epsilon}^{\widehat{\boldsymbol{\pi}}_i} + \int_{t_1+\epsilon}^{t} e^{\overline{A}_0(t-s)} \boldsymbol{d}\mathbb{W}_s.$$

Therefore, the difference signal becomes

$$y_t = e^{\overline{A}_0(t-t_1+\epsilon)}\left[\boldsymbol{x}_{t_1+\epsilon}^{\widehat{\boldsymbol{\pi}}_{i+1}} - \boldsymbol{x}_{t_1+\epsilon}^{\widehat{\boldsymbol{\pi}}_i}\right] = e^{\overline{A}_0(t-t_1+\epsilon)} y_{t_1+\epsilon} = e^{\overline{A}_0(t-t_1+\epsilon)}\left[E_{t_1+\epsilon}\boldsymbol{x}_{t_1} + Z_{t_1+\epsilon}\right],$$

and we obtain

$$\begin{aligned}
&\int_{t_1+\epsilon}^{T}\left(\left\|Q^{1/2}\boldsymbol{z}_t\left(\widehat{\boldsymbol{\pi}}_{i+1}\right)\right\|^2 - \left\|Q^{1/2}\boldsymbol{z}_t\left(\widehat{\boldsymbol{\pi}}_i\right)\right\|^2\right)\boldsymbol{d}t \\
= &\int_{t_1+\epsilon}^{T}\left[\left(\boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_i} + y_t\right)^\top M_0\left(\boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_i} + y_t\right) - \boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_i\top} M_0 \boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_i}\right]\boldsymbol{d}t \\
= &\int_{t_1+\epsilon}^{T}\left[2y_t^\top M_0 \boldsymbol{x}_t^{\widehat{\boldsymbol{\pi}}_i} + y_t^\top M_0 y_t\right]\boldsymbol{d}t. \qquad (40)
\end{aligned}$$

Now, after doing some algebra, the expressions in (39) and (40) lead ro the following for small ϵ:

$$\mathrm{Reg}_{\widehat{\boldsymbol{\pi}}_{i+1}}(T) - \mathrm{Reg}_{\widehat{\boldsymbol{\pi}}_i}(T) = \left(\boldsymbol{x}_{t_1}^\top F_{t_1}\boldsymbol{x}_{t_1} + 2\boldsymbol{x}_{t_1}^\top g_{t_1}\right)\epsilon + O\left(\epsilon^2\right),$$

where

$$\begin{aligned}
F_{t_1} &= S_t + \int_{t_1}^{T}\left(2H_{t_1}^\top e^{\overline{A}_0^\top(s-t_1)}\left(Q_x + K\left(\boldsymbol{\theta_0}\right)^\top Q_u K\left(\boldsymbol{\theta_0}\right)\right)e^{\overline{A}_0(s-t_1)}\right)\boldsymbol{d}s + O\left(\epsilon\right), \\
g_{t_1} &= \int_{t_1}^{T}\left(H_{t_1}^\top e^{\overline{A}_0^\top(s-t_1)}\left(Q_x + K\left(\boldsymbol{\theta_0}\right)^\top Q_u K\left(\boldsymbol{\theta_0}\right)\right)\int_{t_1}^{s} e^{\overline{A}_0(s-u)}\boldsymbol{d}\mathbb{W}_u\right)\boldsymbol{d}s + O\left(\epsilon\right), \\
S_{t_1} &= K_{t_1}^\top Q_u K_{t_1} - K\left(\boldsymbol{\theta_0}\right)^\top Q_u K\left(\boldsymbol{\theta_0}\right), \\
H_{t_1} &= B_0\left(K_{t_1} - K\left(\boldsymbol{\theta_0}\right)\right).
\end{aligned}$$

Thus, as ϵ tends to zero, by (38), we have

$$\mathrm{Reg}_{\widehat{\boldsymbol{\pi}}}(T) - \mathrm{Reg}_{\widehat{\boldsymbol{\pi}}}\left(\boldsymbol{\tau}_0\right) = \int_{\boldsymbol{\tau}_0}^{T}\left(\boldsymbol{x}_t^\top F_t\boldsymbol{x}_t + 2\boldsymbol{x}_t^\top g_t\right)\boldsymbol{d}t, \qquad (41)$$

where F_t, g_t are the above expressions, without the $O\left(\epsilon\right)$ terms.

Next, by (61), the quadratic expression in terms of the matrix F_t can be equivalently written with

$$F_t = S_t + H_t^\top P\left(\boldsymbol{\theta_0}\right) + P\left(\boldsymbol{\theta_0}\right)H_t - H_t^\top E_t - E_t H_t,$$

where

$$E_t = \int_{T}^{\infty} e^{\overline{A}_0^\top(s-t)} M_0 e^{\overline{A}_0(s-t)}\boldsymbol{d}s = e^{\overline{A}_0^\top(T-t)} P\left(\boldsymbol{\theta_0}\right) e^{\overline{A}_0(T-t)}.$$

Note that in the last equality above, we again used (61). Now, after doing some algebra similar to the expression in (59), we have

$$S_t + H_t^\top P\left(\boldsymbol{\theta_0}\right) + P\left(\boldsymbol{\theta_0}\right)H_t = \left(K_t - K\left(\boldsymbol{\theta_0}\right)\right)^\top Q_u\left(K_t - K\left(\boldsymbol{\theta_0}\right)\right),$$

which in turn implies that

$$\int_{\boldsymbol{\tau}_0}^{T}\boldsymbol{x}_t^\top F_t\boldsymbol{x}_t\boldsymbol{d}t = \int_{\boldsymbol{\tau}_0}^{T}\left\|Q_u^{1/2}\left(K_t - K\left(\boldsymbol{\theta_0}\right)\right)\boldsymbol{x}_t\right\|^2\boldsymbol{d}t - 2\int_{\boldsymbol{\tau}_0}^{T}\boldsymbol{x}_t^\top E_t H_t\boldsymbol{x}_t\boldsymbol{d}t. \qquad (42)$$

To study the latter integral, suppose that x_t^* is the state trajectory under the optimal policy $\boldsymbol{\pi_{opt}}$ in (6), and define $\xi_t = \boldsymbol{x}_t - x_t^*$.Note that (1) gives $\boldsymbol{d}\boldsymbol{x}_t = \left(\overline{A}_0 + H_t\right)\boldsymbol{x}_t\boldsymbol{d}t + \boldsymbol{d}\mathbb{W}_t$, as well as

$\boldsymbol{d}x_t^* = \overline{A}_0 x_t^* \boldsymbol{d}t + \boldsymbol{d}\mathbb{W}_t$. Thus, we get $\boldsymbol{d}\xi_t = H_t \boldsymbol{x}_t \boldsymbol{d}t + \overline{A}_0 \xi_t \boldsymbol{d}t$, using which, we have the following for $\varphi_t = e^{-\overline{A}_0 t}\xi_t$:

$$\boldsymbol{d}\varphi_t = \boldsymbol{d}\left(e^{-\overline{A}_0 t}\xi_t\right) = e^{-\overline{A}_0 t}\boldsymbol{d}\xi_t - \overline{A}_0 e^{-\overline{A}_0 t}\xi_t \boldsymbol{d}t = e^{-\overline{A}_0 t} H_t \boldsymbol{x}_t \boldsymbol{d}t.$$

Above, we used the fact that the matrices $e^{-\overline{A}_0 t}, \overline{A}_0$ commute. So, it holds that

$$\begin{aligned}
\boldsymbol{x}_t^\top E_t H_t \boldsymbol{x}_t \boldsymbol{d}t &= \boldsymbol{x}_t^\top e^{\overline{A}_0^\top (T-t)} P\left(\boldsymbol{\theta_0}\right) e^{\overline{A}_0 T} \boldsymbol{d}\varphi_t \\
&= x_t^{*\top} e^{\overline{A}_0^\top (T-t)} P\left(\boldsymbol{\theta_0}\right) e^{\overline{A}_0 T} \boldsymbol{d}\varphi_t + \varphi_t^\top e^{\overline{A}_0^\top T} P\left(\boldsymbol{\theta_0}\right) e^{\overline{A}_0 T} \boldsymbol{d}\varphi_t \\
&= x_t^{*\top} e^{\overline{A}_0^\top (T-t)} P\left(\boldsymbol{\theta_0}\right) e^{\overline{A}_0 T} \boldsymbol{d}\varphi_t + \frac{1}{2}\boldsymbol{d}\left[\varphi_t^\top e^{\overline{A}_0^\top T} P\left(\boldsymbol{\theta_0}\right) e^{\overline{A}_0 T}\varphi_t\right].
\end{aligned}$$

In the above expression, writing the solution of the stochastic differential equation as in (15), we have

$$e^{\overline{A}_0 (T-t)} x_t^* = x_T^* - \int_t^T e^{\overline{A}_0 (T-s)} \boldsymbol{d}\mathbb{W}_s,$$

which gives

$$\begin{aligned}
2\boldsymbol{x}_t^\top E_t H_t \boldsymbol{x}_t \boldsymbol{d}t &= 2x_t^{*\top} e^{\overline{A}_0^\top (T-t)} P\left(\boldsymbol{\theta_0}\right) e^{\overline{A}_0 T} \boldsymbol{d}\varphi_t + \boldsymbol{d}\left[\varphi_t^\top e^{\overline{A}_0^\top T} P\left(\boldsymbol{\theta_0}\right) e^{\overline{A}_0 T}\varphi_t\right] \\
&= -2\left(\int_t^T e^{\overline{A}_0 (T-s)} \boldsymbol{d}\mathbb{W}_s\right)^\top P\left(\boldsymbol{\theta_0}\right) e^{\overline{A}_0 T} \boldsymbol{d}\varphi_t \\
&+ 2x_T^{*\top} P\left(\boldsymbol{\theta_0}\right) e^{\overline{A}_0 T} \boldsymbol{d}\varphi_t + \boldsymbol{d}\left[\varphi_t^\top e^{\overline{A}_0^\top T} P\left(\boldsymbol{\theta_0}\right) e^{\overline{A}_0 T}\varphi_t\right] \\
&= -2\left(\int_t^T e^{\overline{A}_0 (T-s)} \boldsymbol{d}\mathbb{W}_s\right)^\top P\left(\boldsymbol{\theta_0}\right) e^{\overline{A}_0 T} \boldsymbol{d}\varphi_t \\
&+ \boldsymbol{d}\left[\left(x_T^* + e^{\overline{A}_0 T}\varphi_t\right)^\top P\left(\boldsymbol{\theta_0}\right)\left(x_T^* + e^{\overline{A}_0 T}\varphi_t\right)\right],
\end{aligned}$$

where the latest equality holds since the differential of the constant term $x_T^* P\left(\boldsymbol{\theta_0}\right) x_T^*$ is zero. Next, integration by part yields to

$$\begin{aligned}
\int_{\boldsymbol{\tau}_0}^T \left(\int_t^T e^{\overline{A}_0 (T-s)} \boldsymbol{d}\mathbb{W}_s\right)^\top P\left(\boldsymbol{\theta_0}\right) e^{\overline{A}_0 T} \boldsymbol{d}\varphi_t &= -\left(\int_{\boldsymbol{\tau}_0}^T e^{\overline{A}_0 (T-s)} \boldsymbol{d}\mathbb{W}_s\right)^\top P\left(\boldsymbol{\theta_0}\right) e^{\overline{A}_0 T}\varphi_{\boldsymbol{\tau}_0} \\
&+ \int_{\boldsymbol{\tau}_0}^T \varphi_t^\top e^{\overline{A}_0^\top T} P\left(\boldsymbol{\theta_0}\right) e^{\overline{A}_0 (T-t)} \boldsymbol{d}\mathbb{W}_t
\end{aligned}$$

Now, note the following simplifying expressions: First, by definition, we have $x_T^* + e^{\overline{A}_0 T}\varphi_T = x_T^* + \xi_T = \boldsymbol{x}_T$ and

$$x_T^* + e^{\overline{A}_0 T}\varphi_{\boldsymbol{\tau}_0} = x_T^* + e^{\overline{A}_0 (T-t)}(\boldsymbol{x}_{\boldsymbol{\tau}_0} - x_{\boldsymbol{\tau}_0}^*) = e^{\overline{A}_0 (T-t)}\boldsymbol{x}_{\boldsymbol{\tau}_0} + \int_{\boldsymbol{\tau}_0}^T e^{\overline{A}_0 (T-s)} \boldsymbol{d}\mathbb{W}_s,$$

is the terminal state vector under the policy $\widehat{\boldsymbol{\pi}}_0$ that switches to the optimal policy $\boldsymbol{\pi_{\mathbf{opt}}}$ after the time $\boldsymbol{\tau}_0$, because $\int_{\boldsymbol{\tau}_0}^T e^{\overline{A}_0 (T-s)} \boldsymbol{d}\mathbb{W}_s = x_T^* - e^{\overline{A}_0 (T-\boldsymbol{\tau}_0)} x_{\boldsymbol{\tau}_0}^*$. Finally, according to Lemma 8, we have

$$\left\| \int_{\boldsymbol{\tau}_0}^T \varphi_t^\top e^{\overline{A}_0^\top T} P\left(\boldsymbol{\theta_0}\right) e^{\overline{A}_0 (T-t)} \boldsymbol{d}\mathbb{W}_t \right\| \lesssim \left(\int_{\boldsymbol{\tau}_0}^T \left\| e^{\overline{A}_0 (T-t)}\xi_t \right\|^2 \boldsymbol{d}t\right)^{1/2} \log \int_{\boldsymbol{\tau}_0}^T \left\| e^{\overline{A}_0 (T-t)}\xi_t \right\|^2 \boldsymbol{d}t.$$

Putting the above bounds together, we obtain

$$-2\int_{\boldsymbol{\tau}_0}^{T} \boldsymbol{x}_t^\top E_t H_t \boldsymbol{x}_t \boldsymbol{d}t - {x_T^*}^\top P\left(\boldsymbol{\theta_0}\right) x_T^* \lesssim \int_{\boldsymbol{\tau}_0}^{T} \left\| e^{\overline{A}_0 (T-t)} \left(\boldsymbol{x}_t\right) \right\|^2 \boldsymbol{d}t. \tag{43}$$

To proceed toward working with the integration of $\boldsymbol{x}_t^\top g_t$, employ Fubini Theorem [31] to obtain

$$\begin{aligned}
\int_{\boldsymbol{\tau}_0}^{T} \boldsymbol{x}_t^\top \widetilde{g}_t \boldsymbol{d}t &= \int_{\boldsymbol{\tau}_0}^{T} \int_{t}^{T} \int_{t}^{s} \left(\boldsymbol{x}_t^\top H_t^\top e^{\overline{A}_0^\top (s-t)} M e^{\overline{A}_0 (s-u)} \right) \boldsymbol{d}\mathbb{W}_u \boldsymbol{d}s \boldsymbol{d}t \\
&= \int_{\boldsymbol{\tau}_0}^{T} \int_{\boldsymbol{\tau}_0}^{u} \int_{u}^{T} \left(\boldsymbol{x}_t^\top H_t^\top e^{\overline{A}_0^\top (s-t)} M e^{\overline{A}_0 (s-u)} \right) \boldsymbol{d}s \boldsymbol{d}t \boldsymbol{d}\mathbb{W}_u.
\end{aligned}$$

Now, denote the inner double integral by $y_u^\top$:

$$y_u^\top = \int_{\boldsymbol{\tau}_0}^{u} \int_{u}^{T} \left(\boldsymbol{x}_t^\top H_t^\top e^{\overline{A}_0^\top (s-t)} M_0 e^{\overline{A}_0 (s-u)} \right) \boldsymbol{d}s \boldsymbol{d}t = \int_{0}^{u} \left(\boldsymbol{x}_t^\top \left(K_t - K\left(\boldsymbol{\theta_0}\right)\right)^\top P_{t,u}^\top \right) \boldsymbol{d}t,$$

where

$$P_{t,u}^\top = B_0^\top \int_{u}^{T} e^{\overline{A}_0^\top (s-t)} M_0 e^{\overline{A}_0 (s-u)} \boldsymbol{d}s.$$

Now, let $\beta_T = \int_{\boldsymbol{\tau}_0}^{T} \|y_u\|^2 \boldsymbol{d}u$, and employ Lemma 8 to get

$$\int_{\boldsymbol{\tau}_0}^{T} \boldsymbol{x}_t^\top \widetilde{g}_t \boldsymbol{d}t = \int_{\boldsymbol{\tau}_0}^{T} y_u^\top \boldsymbol{d}\mathbb{W}_u = O\left(\beta_T^{1/2} \log^{1/2} \beta_T \right). \tag{44}$$

Thus, we can work with β_T to bound the portion of the regret the integral of $\boldsymbol{x}_t^\top g_t$ captures. For that purpose, the triangle inequality and Fubini Theorem [31] lead to

$$\begin{aligned}
\beta_T &\leq \int_{0}^{T} \int_{0}^{u} \left\| P_{t,u} \left(K_t - K\left(\boldsymbol{\theta_0}\right)\right) \boldsymbol{x}_t \right\|^2 \boldsymbol{d}t \boldsymbol{d}u \\
&= \int_{0}^{T} \left(\boldsymbol{x}_t^\top \left(K_t - K\left(\boldsymbol{\theta_0}\right)\right)^\top \left[\int_{t}^{T} P_{t,u}^\top P_{t,u} \boldsymbol{d}u \right] \left(K_t - K\left(\boldsymbol{\theta_0}\right)\right) \boldsymbol{x}_t \right) \boldsymbol{d}t \\
&\leq \overline{\lambda} \left(\int_{t}^{T} P_{t,u}^\top P_{t,u} \boldsymbol{d}u \right) \int_{0}^{T} \left\| \left(K_t - K\left(\boldsymbol{\theta_0}\right)\right) \boldsymbol{x}_t \right\|^2 \boldsymbol{d}t.
\end{aligned}$$

We can show that $\overline{\lambda}\left(\int_t^T P_{t,u}^\top P_{t,u} \boldsymbol{d}u\right) \lesssim 1$:

$$\begin{aligned}
\overline{\lambda}\left(\int_t^T P_{t,u}^\top P_{t,u} \boldsymbol{d}u\right) &\leq \int_t^T \left\|P_{t,u}^\top\right\|^2 \boldsymbol{d}u \\
&\lesssim \int_t^T \left\| \int_u^T e^{\overline{A}_0^\top (s-t)} M_0 e^{\overline{A}_0 (s-u)} \boldsymbol{d}s \right\|^2 \boldsymbol{d}u \\
&\leq \int_t^T \left\|e^{\overline{A}_0^\top (u-t)}\right\|^2 \left\| \int_u^T e^{\overline{A}_0^\top (s-u)} M_0 e^{\overline{A}_0 (s-u)} \boldsymbol{d}s \right\|^2 \boldsymbol{d}u \\
&\leq \|P(\boldsymbol{\theta_0})\|^2 \int_t^T \left\|e^{\overline{A}_0^\top (u-t)}\right\|^2 \boldsymbol{d}u \lesssim 1.
\end{aligned}$$

Above, in the last inequality we use (61). Note that the last expression is a bounded constant, since all eigenvalues of $\overline{A}_0$ are in the open left half-plane.

Thus, according to (44), it is enough to consider

$$\beta_T \lesssim \int_0^T \|(K_t - K(\boldsymbol{\theta_0}))\, \boldsymbol{x}_t\|^2 \boldsymbol{d}t, \tag{45}$$

in order to bound the portion of the regret that the integration of $\boldsymbol{x}_t^\top g_t$ contributes.

While the above discussions apply to the regret during the time interval $\boldsymbol{\tau}_0 \leq t \leq T$, we can similarly bound the regret during the stabilization period $0 \leq t \leq \boldsymbol{\tau}_0$. The difference is in the randomization sequence $w_n, n = 0, 1, \cdots$, which is reflected through the piece-wise constant signal $v(t)$ in (14). Therefore, it suffices to add the effect of $v(t)$ to the one of the Wiener process $\mathbb{W}_t$, and so $\Sigma_{\mathbb{W}}$ will be replaced with $\left(\Sigma_{\mathbb{W}} + \sigma_w^2\right)$:

$$\mathrm{Reg}_{\widehat{\boldsymbol{\pi}}}(\boldsymbol{\tau}_0) \leq \left(\Sigma_{\mathbb{W}} + \sigma_w^2\right) \boldsymbol{\tau}_0 \|K - K(\boldsymbol{\theta_0})\|^2. \tag{46}$$

Finally, (41), (42), (43), (44), (45), and (46) together, we get the desired result. ■

B.3 Stochastic inequality for continuous-time self-normalized martingales

Lemma 8 *Let $\boldsymbol{z}_t = \left[\boldsymbol{x}_t^\top, \boldsymbol{u}_t^\top\right]^\top$ be the observation signal and $\widehat{\Sigma}_t$ be as in* (8)*. Then, for the stochastic integral $\Phi_t = \int_0^t \boldsymbol{z}_s \boldsymbol{d}\mathbb{W}_s^\top$, we have*

$$\overline{\lambda}\left(\Phi_t^\top \widehat{\Sigma}_t^{-1} \Phi_t\right) \lesssim p\overline{\lambda}(\Sigma_{\mathbb{W}}) \left[\log\det\widehat{\Sigma}_t - \log\det\widehat{\Sigma}_0\right]. \tag{47}$$

Proof. We approximate the integrals over the interval $[0, t]$ through n equally distanced points in the interval, and then let $n \to \infty$. So, let $\epsilon = \lfloor t/n \rfloor$, and for $k = 0, 1, \cdots, n-1$, consider the matrix

$$M_k = \frac{1}{\epsilon}\widehat{\Sigma}_0 + \sum_{i=0}^{k} \boldsymbol{z}_{i\epsilon} \boldsymbol{z}_{i\epsilon}^\top.$$

Using the above matrices, for $k = 1, \cdots, n$, define $\alpha_k = \boldsymbol{z}_{k\epsilon}^\top M_{k-1}^{-1} \boldsymbol{z}_{k\epsilon}$. Thus, we have

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

Now, $M_{k-1}^{-1} \boldsymbol{z}_{k\epsilon} \boldsymbol{z}_{k\epsilon}^\top$ is a rank-one matrix, and so $p + q - 1$ eigenvalues of $I + M_{k-1}^{-1} \boldsymbol{z}_{k\epsilon} \boldsymbol{z}_{k\epsilon}^\top$ except one are 1, and one eigenvalue is $1 + \alpha_k$. So, it holds that

$$\frac{\det M_k}{\det M_{k-1}} = 1 + \alpha_k.$$

Next, it is straightforward to show that

$$M_k^{-1} = \left(M_{k-1} + \boldsymbol{z}_{k\epsilon}\boldsymbol{z}_{k\epsilon}^\top\right)^{-1} = M_{k-1}^{-1} - \frac{M_{k-1}^{-1}\boldsymbol{z}_{k\epsilon}\boldsymbol{z}_{k\epsilon}^\top M_{k-1}^{-1}}{1+\boldsymbol{z}_{k\epsilon}^\top M_{k-1}^{-1}\boldsymbol{z}_{k\epsilon}}.$$

Therefore, we have

$$\boldsymbol{z}_{k\epsilon}^\top M_k^{-1}\boldsymbol{z}_{k\epsilon} = \boldsymbol{z}_{k\epsilon}^\top \left(M_{k-1} + \boldsymbol{z}_{k\epsilon}\boldsymbol{z}_{k\epsilon}^\top\right)^{-1}\boldsymbol{z}_{k\epsilon} = \alpha_k - \frac{\alpha_k^2}{1+\alpha_k} = \frac{\det M_k - \det M_{k-1}}{\det M_k}.$$

However, since for all $\alpha \in \mathbb{R}$ we have $1+\alpha \leq e^\alpha$, we obtain

$$\boldsymbol{z}_{k\epsilon}^\top M_k^{-1}\boldsymbol{z}_{k\epsilon} \leq \log\det M_k - \log\det M_{k-1}. \tag{48}$$

To proceed, let $\mathcal{F}_k$ be the sigma-field generated by the Wiener process up to time $k\epsilon$:

$$\mathcal{F}_k = \mathcal{F}\left(\mathbb{W}_s, 0 \leq s \leq k\epsilon\right).$$

Further, define $L_k = \sum_{i=0}^{k} \boldsymbol{z}_{i\epsilon}\left(\mathbb{W}_{(i+1)\epsilon} - \mathbb{W}_{i\epsilon}\right)^\top$.

So, we have

$$\mathbb{E}\left[L_k^\top M_k^{-1} L_k\right] = \mathbb{E}\left[\mathbb{E}\left[L_k^\top M_k^{-1} L_k \middle| \mathcal{F}_k\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\Psi_k^\top M_k^{-1}\Psi_k \middle| \mathcal{F}_k\right]\right],$$

where $\Psi_k = L_{k-1} + \boldsymbol{z}_{k\epsilon}\left(\mathbb{W}_{(k+1)\epsilon} - \mathbb{W}_{k\epsilon}\right)^\top$. Since L_{k-1} is $\mathcal{F}_k$–measurable, we get

$$\begin{aligned}
& \mathbb{E}\left[L_k^\top M_k^{-1} L_k\right] \\
= \;& \mathbb{E}\left[L_{k-1}^\top M_k^{-1} L_{k-1} + \mathbb{E}\left[\left(\mathbb{W}_{(k+1)\epsilon} - \mathbb{W}_{k\epsilon}\right)\boldsymbol{z}_{k\epsilon}^\top M_k^{-1}\boldsymbol{z}_{k\epsilon}\left(\mathbb{W}_{(k+1)\epsilon} - \mathbb{W}_{k\epsilon}\right)^\top \middle| \mathcal{F}_k\right]\right] \\
= \;& \mathbb{E}\left[L_{k-1}^\top M_k^{-1} L_{k-1} + \left(\boldsymbol{z}_{k\epsilon}^\top M_k^{-1}\boldsymbol{z}_{k\epsilon}\right)\epsilon\Sigma_{\mathbb{W}}\right],
\end{aligned}$$

where in the last line above we used $\mathcal{F}_k$–measurability of $\boldsymbol{z}_{k\epsilon}, M_k$, as well as the independent increments property and the covariance matrix of the Wiener process. So, (48) implies that

$$\overline{\lambda}\left(\mathbb{E}\left[L_k^\top M_k^{-1} L_k\right]\right) - \overline{\lambda}\left(\mathbb{E}\left[L_{k-1}^\top M_{k-1}^{-1} L_{k-1}\right]\right) \leq \epsilon\overline{\lambda}\left(\Sigma_{\mathbb{W}}\right)\left(\log\det\left(\epsilon M_k\right) - \log\det\left(\epsilon M_{k-1}\right)\right).$$

Thus, summing over $k = 1, \cdots, n$, we get

$$\overline{\lambda}\left(\mathbb{E}\left[L_n^\top\left(\epsilon M_n\right)^{-1} L_n\right]\right) \leq \overline{\lambda}\left(\Sigma_{\mathbb{W}}\right)\left(\log\det\left(\epsilon M_n\right) - \log\det\left(\epsilon M_0\right)\right).$$

Now, consider $\overline{\lambda}\left(L_n^\top\left(\epsilon M_n\right)^{-1} L_n\right)$. Since $L_n^\top\left(\epsilon M_n\right)^{-1} L_n$ is positive semidefinite, its largest eigenvalue can be upper-bounded by its trace, which implies that

$$\begin{aligned}
\mathbb{E}\left[\overline{\lambda}\left(L_n^\top\left(\epsilon M_n\right)^{-1} L_n\right)\right] &\leq \mathbb{E}\left[\mathbf{tr}\left(L_n^\top\left(\epsilon M_n\right)^{-1} L_n\right)\right] \\
&= \mathbf{tr}\left(\mathbb{E}\left[L_n^\top\left(\epsilon M_n\right)^{-1} L_n\right]\right) \\
&\leq p\overline{\lambda}\left(\mathbb{E}\left[L_n^\top\left(\epsilon M_n\right)^{-1} L_n\right]\right) \\
&\leq p\overline{\lambda}\left(\Sigma_{\mathbb{W}}\right)\left(\log\det\left(\epsilon M_n\right) - \log\det\left(\epsilon M_0\right)\right),
\end{aligned}$$

where we used the fact that the linear operators of trace and expected value interchange.

Thus, Martingale Convergence Theorem [31] implies that

$$\overline{\lambda}\left(L_n^\top\left(\epsilon M_n\right)^{-1} L_n\right) \lesssim p\overline{\lambda}\left(\Sigma_{\mathbb{W}}\right)\left(\log\det\left(\epsilon M_n\right) - \log\det\left(\epsilon M_0\right)\right)$$

Finally, as n tends to infinity, ϵ shrinks and we obtain the desired result. ■

B.4 Anti-concentration of the posterior precision matrix in Algorithm 2

Lemma 9 *In Algorithm 2, we have the following for the matrix $\widehat{\Sigma}_{\boldsymbol{\tau}_n}$ that is defined in* (8)*:*

$$\liminf_{n\to\infty} \boldsymbol{\tau}_n^{-1/2} \underline{\lambda}\left(\widehat{\Sigma}_{\boldsymbol{\tau}_n}\right) \gtrsim \underline{\lambda}\left(\Sigma_{\mathbb{W}}\right).$$

Proof. First, we define some notation. Recall that during the time interval $\boldsymbol{\tau}_i \le t < \boldsymbol{\tau}_{i+1}$ corresponding to episode i, Algorithm 2 uses a single parameter estimate $\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}$. So, for $i = 0, 1, \cdots$, we use $\Phi_i, K_i, \overline{A}_i$ to denote the sample covariance matrix of the state vectors of episode i, and the feedback and closed-loop matrices during episode i:

$$\begin{aligned}
\Phi_i &= \int_{\boldsymbol{\tau}_i}^{\boldsymbol{\tau}_{i+1}} \boldsymbol{x}_t \boldsymbol{x}_t^\top \boldsymbol{dt}, \\
K_i &= -Q_u^{-1} \widehat{B}_i^\top P\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}\right), \\
\overline{A}_i &= A_0 + B_0 K_i.
\end{aligned}$$

So, it holds that

$$\widehat{\Sigma}_{\boldsymbol{\tau}_n} = \widehat{\Sigma}_{\boldsymbol{\tau}_0} + \sum_{i=0}^{n-1} L_i \Phi_i L_i^\top, \tag{49}$$

where $L_i = \begin{bmatrix} I_p \\ K_i \end{bmatrix}$.

Now, consider the matrix Φ_i. Note that according to the bounded grows rates of the episode (from both above and below) in (12), both $\boldsymbol{\tau}_{i+1} - \boldsymbol{\tau}_i$ and $\boldsymbol{\tau}_i$ tend to infinity as i grows. Thus, in the sequel, we suppose that the indices n, i, j, k that are used for denoting the episodes, are large enough. Similar to (31), we have

$$\Phi_i = \int_0^\infty e^{\overline{A}_i s} \left[(\boldsymbol{\tau}_{i+1} - \boldsymbol{\tau}_i) \Sigma_{\mathbb{W}} + M_i + M_i^\top + \boldsymbol{x}_{\boldsymbol{\tau}_i} \boldsymbol{x}_{\boldsymbol{\tau}_i}^\top - \boldsymbol{x}_{\boldsymbol{\tau}_{i+1}} \boldsymbol{x}_{\boldsymbol{\tau}_{i+1}}^\top \right] e^{\overline{A}_i^\top s} \boldsymbol{ds},$$

where

$$M_i = \int_{\boldsymbol{\tau}_i}^{\boldsymbol{\tau}_{i+1}} \boldsymbol{x}_t \boldsymbol{d}\mathbb{W}_t^\top.$$

So, using the fact that the real-parts of all eigenvalues of $\overline{A}_i$ are negative and so $\boldsymbol{x}_{\boldsymbol{\tau}_{i+1}}$ can be bounded with $\exp\left(\overline{A}_i(\boldsymbol{\tau}_{i+1} - \boldsymbol{\tau}_i)\right) \boldsymbol{x}_{\boldsymbol{\tau}_i}$ similar to (30), as well as Lemma 2, we obtain the following bounds for the largest and smallest eigenvalues of Φ_i

$$\underline{\lambda}\left(\Phi_i\right) \gtrsim (\boldsymbol{\tau}_{i+1} - \boldsymbol{\tau}_i) \underline{\lambda}\left(\Sigma_{\mathbb{W}}\right) \underline{\lambda}\left(\int_0^\infty e^{\overline{A}_i s} e^{\overline{A}_i^\top s} \boldsymbol{ds} \right), \tag{50}$$

$$\overline{\lambda}\left(\Phi_i\right) \lesssim (\boldsymbol{\tau}_{i+1} - \boldsymbol{\tau}_i) \overline{\lambda}\left(\Sigma_{\mathbb{W}}\right) \int_0^\infty \left\| e^{\overline{A}_i s} \right\|^2 \boldsymbol{ds}. \tag{51}$$

On the other hand, for the parameter estimates at the end of episodes, similar to (35), we have

$$\widehat{\Sigma}_{\boldsymbol{\tau}_i}^{1/2} \left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}} - \boldsymbol{\theta_0}\right) = \widehat{\Sigma}_{\boldsymbol{\tau}_i}^{1/2} \left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}} - \widehat{\mu}_{\boldsymbol{\tau}_i}\right) + \widehat{\Sigma}_{\boldsymbol{\tau}_i}^{-1/2} \left(-\boldsymbol{\theta_0} + \int_0^{\boldsymbol{\tau}_i} \boldsymbol{z}_s \boldsymbol{d}\mathbb{W}_s^\top \right).$$

Note that by the construction of the posterior $\mathcal{D}_{\boldsymbol{\tau}_i}$ in (9), for the first term we have $\widehat{\Sigma}_{\boldsymbol{\tau}_i}^{1/2} \left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}} - \widehat{\mu}_{\boldsymbol{\tau}_i}\right) \sim \boldsymbol{N}\left(0, I_{p+q}\right)$. Further, for the second term, Lemma 8 together with (51) lead to

$$\left\| \widehat{\Sigma}_{\boldsymbol{\tau}_i}^{-1/2} \left(-\boldsymbol{\theta_0} + \int_0^{\boldsymbol{\tau}_i} \boldsymbol{z}_s \boldsymbol{d}\mathbb{W}_s^\top \right) \right\| \lesssim (p+q) \log^{1/2} \boldsymbol{\tau}_i.$$

Therefore, we have

$$\left\|\widehat{\Sigma}_{\boldsymbol{\tau}_i}^{1/2}\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}-\boldsymbol{\theta}_{\mathbf{0}}\right)\right\| \lesssim (p+q)\log^{1/2}\boldsymbol{\tau}_i.$$

However, using the relationship between $\widehat{\Sigma}_{\boldsymbol{\tau}_i}$ and $\Phi_0,\cdots,\Phi_{i-1}$ in (49), we can write

$$(p+q)^2\log\boldsymbol{\tau}_i \gtrsim \left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}-\boldsymbol{\theta}_{\mathbf{0}}\right)^\top \widehat{\Sigma}_{\boldsymbol{\tau}_i}\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}-\boldsymbol{\theta}_{\mathbf{0}}\right) \geq \left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}-\boldsymbol{\theta}_{\mathbf{0}}\right)^\top \left[\sum_{j=0}^{i-1} L_j\Phi_j L_j^\top\right]\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}-\boldsymbol{\theta}_{\mathbf{0}}\right),$$

which according to the bound in (50) implies that

$$\underline{\lambda}\left(\Sigma_{\mathbb{W}}\right)\sum_{j=0}^{i-1}(\boldsymbol{\tau}_{j+1}-\boldsymbol{\tau}_j)\left\|L_j^\top\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}-\boldsymbol{\theta}_{\mathbf{0}}\right)\right\|^2 \lesssim (p+q)^2\log\boldsymbol{\tau}_i.$$

Clearly, the above result indicates that for $j<i$, it holds that

$$\left\|\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}-\boldsymbol{\theta}_{\mathbf{0}}\right)^\top L_j\right\|^2 \lesssim \frac{(p+q)^2\log\boldsymbol{\tau}_i}{\underline{\lambda}\left(\Sigma_{\mathbb{W}}\right)\left(\boldsymbol{\tau}_{j+1}-\boldsymbol{\tau}_j\right)}. \tag{52}$$

Next, we employ Lemma 6 to study hoe Algorithm 2 utilizes Thompson sampling to diversify the matrices $L_1, L_2, \cdots$. To do so, we consider the randomization the posterior $\mathcal{D}_{\boldsymbol{\tau}_i}$ applies to the sub-matrix of the parameter estimate corresponding to the input matrix $\widehat{B}_i$. That is, we aim to find the distribution of the random $p\times q$ matrix $\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}-\widehat{\mu}_{\boldsymbol{\tau}_i}\right)^\top\begin{bmatrix}0_{p\times q}\\ I_q\end{bmatrix}$. Since $\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}-\widehat{\mu}_{\boldsymbol{\tau}_i}\sim\boldsymbol{N}\left(0,\widehat{\Sigma}_{\boldsymbol{\tau}_i}^{-1}\right)$, we have

$$E_i=\begin{bmatrix}0_{p\times q}\\ I_q\end{bmatrix}^\top\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}-\widehat{\mu}_{\boldsymbol{\tau}_i}\right)\sim\boldsymbol{N}\left(0,\begin{bmatrix}0_{p\times q}\\ I_q\end{bmatrix}^\top\widehat{\Sigma}_{\boldsymbol{\tau}_i}^{-1}\begin{bmatrix}0_{p\times q}\\ I_q\end{bmatrix}\right)=\boldsymbol{N}\left(0,\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}^{-1}\right]_{22}\right), \tag{53}$$

where $\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}^{-1}\right]_{22}$ is the $q\times q$ lower-left block in $\widehat{\Sigma}_{\boldsymbol{\tau}_i}^{-1}$:

$$\widehat{\Sigma}_{\boldsymbol{\tau}_i}^{-1}=\begin{bmatrix}\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}^{-1}\right]_{11} & \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}^{-1}\right]_{12}\\ \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}^{-1}\right]_{21} & \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}^{-1}\right]_{22}\end{bmatrix}.$$

Note that $\widehat{\Sigma}_{\boldsymbol{\tau}_0}$ is a positive semi-definite matrix. Therefore, it suffices to show the desired result for $\widehat{\Sigma}_{\boldsymbol{\tau}_n}-\widehat{\Sigma}_{\boldsymbol{\tau}_0}$, and so in the sequel we remove the effect of $\widehat{\Sigma}_{\boldsymbol{\tau}_0}$ by treating $\boldsymbol{\tau}_0$ as 0. So, to calculate the inverse $\widehat{\Sigma}_{\boldsymbol{\tau}_i}^{-1}$, we apply block matrix inversion to

$$\widehat{\Sigma}_{\boldsymbol{\tau}_i}=\begin{bmatrix}\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11} & \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12}\\ \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{21} & \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{22}\end{bmatrix}=\begin{bmatrix}\sum_{j=0}^{i-1}\Phi_j & \sum_{j=0}^{i-1}\Phi_j K_j^\top\\ \sum_{j=0}^{i-1}K_j\Phi_j & \sum_{j=0}^{i-1}K_j\Phi_j K_j^\top\end{bmatrix},$$

to obtain

$$\begin{aligned}
\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}^{-1}\right]_{11} &= \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1}+\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1}\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12}\Omega_i^{-1}\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{21}\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1},\\
\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}^{-1}\right]_{12} &= -\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1}\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12}\Omega_i^{-1},\\
\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}^{-1}\right]_{22} &= \Omega_i^{-1},\\
\Omega_i &= \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{22}-\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{21}\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1}\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12}.
\end{aligned}$$

The smallest eigenvalue of $\widehat{\Sigma}_{\boldsymbol{\tau}_i}$ is related to that of Ω_i. On one hand, since Ω_i^{-1} is a sub-matrix of $\widehat{\Sigma}_{\boldsymbol{\tau}_i}^{-1}$; i.e., $\overline{\lambda}\left(\Omega_i^{-1}\right)\leq\overline{\lambda}\left(\widehat{\Sigma}_{\boldsymbol{\tau}_i}^{-1}\right)$, which implies that $\underline{\lambda}\left(\Omega_i\right)\geq\underline{\lambda}\left(\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right)$. Now, we show that the

inequality holds in the opposite direction as well, modulo a constant factor. Suppose that $\nu \in \mathbb{R}^{p+q}$ is a unit vector, $\nu = [\nu_1^\top, \nu_2^\top]^\top$, $\nu_1 \in \mathbb{R}^p$, and $\nu_2 \in \mathbb{R}^q$. So, after doing some algebra as follows, we have

$$\begin{aligned}
\nu^\top \widehat{\Sigma}_{\boldsymbol{\tau}_i} \nu &= \nu_1^\top \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11} \nu_1 + 2\nu_1^\top \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12} \nu_2 + \nu_2^\top \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{22} \nu_2 \\
&= \nu_1^\top \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11} \nu_1 + 2\nu_1^\top \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12} \nu_2 \\
&+ \nu_2^\top \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{21} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12} \nu_2 \\
&+ \nu_2^\top \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{22} \nu_2 - \nu_2^\top \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{21} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12} \nu_2 \\
&= \left(\nu_1 + \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12} \nu_2\right)^\top \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11} \left(\nu_1 + \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12} \nu_2\right) \\
&+ \nu_2 \Omega_i \nu_2.
\end{aligned}$$

For the matrix $\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11} = \sum_{j=0}^{i-1} \Phi_j$, the smallest eigenvalue lower bounds in (50) lead to $\underline{\lambda}\left(\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}\right) \gtrsim \boldsymbol{\tau}_i \underline{\lambda}\left(\Sigma_{\mathbb{W}}\right)$. Thus, in order to show the desired smallest eigenvalue result for $\widehat{\Sigma}_{\boldsymbol{\tau}_n}$, it suffices to consider unit vectors ν for which $\left\|\nu_1 + \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12} \nu_2\right\| \lesssim \boldsymbol{\tau}_i^{-1/4}$ holds. For such unit vectors ν, the expressions $\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11} = \sum_{j=0}^{i-1} \Phi_j$ and $\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12} = \sum_{j=0}^{i-1} \Phi_j K_j^\top$, as well as Lemma 11 that indicates that the matrices K_j are bounded, $\|\nu_2\|$ needs to be bounded away from zero since $\|\nu_1\|^2 + \|\nu_2\|^2 = \|\nu\|^2 = 1$. Thus, we have

$$\underline{\lambda}\left(\Omega_i\right) \geq \underline{\lambda}\left(\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right) \gtrsim \underline{\lambda}\left(\Omega_i\right). \tag{54}$$

Otherwise, the desired result about the eigenvalue of $\widehat{\Sigma}_{\boldsymbol{\tau}_n}$ holds true.

By simplifying the following expression, we get

$$\begin{aligned}
&\sum_{j=0}^{i-1} \left(K_j^\top - \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12}\right)^\top \Phi_j \left(K_j^\top - \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12}\right) \\
= &\sum_{j=0}^{i-1} K_j \Phi_j K_j^\top - \sum_{j=0}^{i-1} K_j \Phi_j \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12} \\
- &\sum_{j=0}^{i-1} \left(\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12}\right)^\top \Phi_j K_j^\top + \sum_{j=0}^{i-1} \left(\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12}\right)^\top \Phi_j \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12} \\
= &\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{22} - \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{21} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12} - \left(\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12}\right)^\top \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{21} \\
+ &\left(\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12}\right)^\top \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12} \\
= &\Omega_i.
\end{aligned}$$

However, we have

$$K_j^\top - \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{12} = \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \left(\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11} K_j^\top - \sum_{k=0}^{i-1} \Phi_k K_k^\top\right) = \left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \sum_{k=0}^{i-1} \Phi_k \left(K_j - K_k\right)^\top,$$

i.e.,

$$\Omega_i = \sum_{j=0}^{i-1} \left(\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \sum_{k=0}^{i-1} \Phi_k \left(K_j - K_k\right)^\top\right)^\top \Phi_j \left(\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \sum_{k=0}^{i-1} \Phi_k \left(K_j - K_k\right)^\top\right). \tag{55}$$

We use the above expression to relate the matrices $\Omega_0, \Omega_1, \cdots$ to each others. First, let $\Psi_0, \Psi_1, \cdots$ be a sequence of independent random $q \times p$ matrices with standard normal distribution

$$\Psi_i \sim \boldsymbol{N}\left(0_{q\times p}, I_q\right). \tag{56}$$

Then, since $\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}^{-1}\right]_{22} = \Omega_i^{-1}$ and (53), we can let $E_i = \Omega_i^{-1/2}\Psi_i$. Further, for $j, k = 0, 1, \cdots$, denote the B-part of the differences $\widehat{\mu}_k - \widehat{\mu}_j$ by

$$H_{kj} = [0_{q\times p}, I_q]\left(\widehat{\mu}_k - \widehat{\mu}_j\right).$$

Note that the above result together with (53) give

$$[0_{q\times p}, I_q]\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{k}} - \widehat{\boldsymbol{\theta}}_{\boldsymbol{j}}\right) = H_{kj} + \Omega_k^{-1/2}\Psi_k - \Omega_j^{-1/2}\Psi_j.$$

We will show in the sequel that the above normally distributed random matrices are the effective randomizations that Thompson sampling Algorithm 2 applies for exploration. For that purpose, using the directional derivatives and the optimality manifolds in Lemma 6, we calculate $K_k - K_j$ according to $H_{kj} + \Omega_k^{1/2}\Psi_k - \Omega_j^{1/2}\Psi_j$. Plugging (52) in the expression for $\Delta_{\boldsymbol{\theta_1}}(X, Y)$ in Lemma 6 for

$$[X, Y] = \widehat{\boldsymbol{\theta}}_{\boldsymbol{k}}^{\top} - \widehat{\boldsymbol{\theta}}_{\boldsymbol{j}}^{\top},$$

we have

$$\left\| \int_0^\infty e^{\overline{A}_j^\top t}\left[L_j^\top\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{k}} - \widehat{\boldsymbol{\theta}}_{\boldsymbol{j}}\right) P\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{j}}\right) + P\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{j}}\right)\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{k}} - \widehat{\boldsymbol{\theta}}_{\boldsymbol{j}}\right)^\top L_j\right] e^{\overline{A}_j t}\boldsymbol{dt}\right\|^2 \lesssim \frac{(p+q)^2 \log \boldsymbol{\tau}_k}{\underline{\lambda}\left(\Sigma_{\mathbb{W}}\right)\left(\boldsymbol{\tau}_{j+1} - \boldsymbol{\tau}_j\right)},$$

and

$$P\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{j}}\right) Y = P\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{j}}\right)\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{k}} - \widehat{\boldsymbol{\theta}}_{\boldsymbol{j}}\right)^\top \begin{bmatrix} 0_{p\times q} \\ I_q \end{bmatrix} = P\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{j}}\right)\left(H_{kj} + \Omega_k^{-1/2}\Psi_k - \Omega_j^{-1/2}\Psi_j\right)^\top.$$

Putting the above two portions of $\Delta_{\boldsymbol{\theta_1}}(X, Y)$ together, since Ψ_k, Ψ_j are independent and standard normal random matrices, (54) implies that the latter portion of $\Delta_{\boldsymbol{\theta_1}}(X, Y)$ is the dominant one. Thus, according to Lemma 6 and the expression for the optimal feedbacks in (6), we can approximate $K_k - K_j$ in (55) by

$$-Q_u^{-1}\left(H_{kj} + \Omega_k^{-1/2}\Psi_k - \Omega_j^{-1/2}\Psi_j\right) P\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{j}}\right).$$

We use the above approximation for the matrix $\left[\widehat{\Sigma}_{\boldsymbol{\tau}_i}\right]_{11}^{-1} \sum_{k=0}^{i-1} \Phi_k \left(K_j - K_k\right)^\top$ in (55), letting the episode number i grow. So, the following expression captures the limit behavior of the least eigenvalue of $\widehat{\Sigma}_{\boldsymbol{\tau}_n}$ in Algorithm 2:

$$\begin{aligned} \lim_{n\to\infty} \quad & \frac{Q_u \Omega_n Q_u}{\boldsymbol{\tau}_n^{1/2}\underline{\lambda}\left(\Sigma_{\mathbb{W}}\right)} = \lim_{n\to\infty} \sum_{j=0}^{n-1}\left(\sum_{k=0}^{n-1} \widetilde{\Phi_k} P\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{j}}\right)\left(\frac{H_{kj} + \Omega_k^{-1/2}\Psi_k - \Omega_j^{-1/2}\Psi_j}{\boldsymbol{\tau}_n^{-1/4}}\right)^\top\right)^\top \\ & \frac{\boldsymbol{\tau}_{j+1} - \boldsymbol{\tau}_j}{\boldsymbol{\tau}_n \underline{\lambda}\left(\Sigma_{\mathbb{W}}\right)} \frac{\Phi_j}{\boldsymbol{\tau}_{j+1} - \boldsymbol{\tau}_j}\left(\sum_{k=0}^{n-1} \widetilde{\Phi_k} P\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{j}}\right)\left(\frac{H_{kj} + \Omega_k^{-1/2}\Psi_k - \Omega_j^{-1/2}\Psi_j}{\boldsymbol{\tau}_n^{-1/4}}\right)^\top\right), \end{aligned} \tag{57}$$

where

$$\widetilde{\Phi_k} = \left[\sum_{i=0}^{n-1} \Phi_i\right]^{-1} \Phi_k.$$

The equation in (57) provides the limit behavior of the randomized exploration Algorithm 2 performs for learning to control the diffusion process. More precisely, it shows the roles of the random samples from the posteriors through the random matrices $\Omega_k^{-1/2}\Psi_k$, for $k = 0, \cdots, n-1$, which render the limit matrix in (57) a positive definite one, as describe below.

Note that since $\sum_{i=0}^{n-1} \widetilde{\Phi_i} = I_p$, the expression

$$\sum_{k=0}^{n-1} \widetilde{\Phi_k} P\left(\widehat{\boldsymbol{\theta}_{\boldsymbol{j}}}\right) \left(\frac{H_{kj} + \Omega_k^{-1/2} \Psi_k - \Omega_j^{-1/2} \Psi_j}{\boldsymbol{\tau}_n^{-1/4}} \right)^\top$$

is a weighted average of the random matrices $\boldsymbol{\tau}_n^{1/4} \left(H_{kj} + \Omega_k^{-1/2} \Psi_k - \Omega_j^{-1/2} \Psi_j \right)^\top$. Moreover, according to the discussions leading to (50) and (51), the matrix $(\boldsymbol{\tau}_{j+1} - \boldsymbol{\tau}_j)^{-1} \Phi_j$ converge as j grows to a positive definite matrix, for which all eigenvalues are larger than $\underline{\lambda}(\Sigma_{\mathbb{W}})$, modulo a constant factor. On the other hand, because the lengths of the episodes satisfies the bounded growth rates in (12), the ratios $\boldsymbol{\tau}_n^{-1}(\boldsymbol{\tau}_{j+1} - \boldsymbol{\tau}_j)$ are bounded from above and below by $\underline{\alpha}^{n-j}$ and $\overline{\alpha}(\overline{\alpha}+1)^{n-j-1}$, and their sum over $j = 0, \cdots, n-1$ is 1. A similar property of boundedness from above and below applies to $\boldsymbol{\tau}_j^{-1/4} \boldsymbol{\tau}_n^{-1/4}$. So, the expression on the right-hand-side of (57) is in fact a weighted average of

$$\sum_{k=0}^{n-1} \widetilde{\Phi_k} P\left(\widehat{\boldsymbol{\theta}_{\boldsymbol{j}}}\right) \left(\frac{H_{kj} + \Omega_k^{-1/2} \Psi_k - \Omega_j^{-1/2} \Psi_j}{\boldsymbol{\tau}_n^{-1/4}} \right)^\top ,$$

for $j = 0, \cdots, n-1$.

Note that by the distribution of the random matrices in (56), all rows of $\boldsymbol{\tau}_n^{1/4} \left(H_{kj} + \Omega_k^{-1/2} \Psi_k - \Omega_j^{-1/2} \Psi_j \right)^\top$ are independent normal random vectors, implying that these random matrices are almost surely full-rank. Therefore, $\boldsymbol{\tau}_n^{-1/2} \Omega_n$ converges to a positive definite random matrix, which according to (54) implies the desired result.

■

C Auxiliary Lemmas

C.1 Behaviors of diffusion processes under non-optimal feedback

Lemma 10 *Let $\widehat{A}, \widehat{B}$ be an arbitrary pair of stabilizable system matrices. Suppose that for the closed-loop matrix $\overline{A} = \widehat{A} + \widehat{B}K$, we have $\overline{\lambda}\left(\exp\left(\overline{A}\right)\right) < 1$, and P satisfies*

$$\overline{A}^\top P + P\overline{A} + Q_x + K^\top Q_u K = 0.$$

Then, it holds that

$$P = P\left(\widehat{\boldsymbol{\theta}}\right) + \int_0^\infty e^{\overline{A}^\top t}\left(K + Q_u^{-1}\widehat{B}^\top P\left(\widehat{\boldsymbol{\theta}}\right)\right)^\top \left(Q_u K + \widehat{B}^\top P\left(\widehat{\boldsymbol{\theta}}\right)\right) e^{\overline{A}t}\boldsymbol{dt}.$$

Proof. Denote $K\left(\widehat{\boldsymbol{\theta}}\right) = -Q_u^{-1}\widehat{B}^\top P\left(\widehat{\boldsymbol{\theta}}\right)$ and $\widehat{\overline{A}} = \widehat{A} + \widehat{B}K\left(\widehat{\boldsymbol{\theta}}\right)$. So, after doing some algebra, it is easy to show that the algebraic Riccati equation in (5) gives

$$\widehat{\overline{A}}^\top P\left(\widehat{\boldsymbol{\theta}}\right) + P\left(\widehat{\boldsymbol{\theta}}\right)\widehat{\overline{A}} + Q_x + K\left(\widehat{\boldsymbol{\theta}}\right)^\top Q_u K\left(\widehat{\boldsymbol{\theta}}\right).$$

Now, let $\Phi = K^\top Q_u K - K\left(\widehat{\boldsymbol{\theta}}\right)^\top Q_u K\left(\widehat{\boldsymbol{\theta}}\right)$, and subtract the above equation that $P\left(\widehat{\boldsymbol{\theta}}\right)$ solves, from the similar one in the statement of the lemma that P satisfies, to get

$$\left(\overline{A} - \widehat{\overline{A}}_1\right)^\top P\left(\widehat{\boldsymbol{\theta}}\right) + P\left(\widehat{\boldsymbol{\theta}}\right)\left(\overline{A} - \widehat{\overline{A}}_1\right) + \overline{A}^\top\left(P - P\left(\widehat{\boldsymbol{\theta}}\right)\right) + \left(P - P\left(\widehat{\boldsymbol{\theta}}\right)\right)\overline{A} + \Phi = 0. \quad (58)$$

Because $\overline{\lambda}\left(\exp\left(\overline{A}\right)\right) < 1$, by solving (58) for $P - P\left(\widehat{\boldsymbol{\theta}}\right)$, we have

$$P - P\left(\widehat{\boldsymbol{\theta}}\right) = \int_0^\infty e^{\overline{A}^\top t}\left(\Phi + \left[K - K\left(\widehat{\boldsymbol{\theta}}\right)\right]^\top \widehat{B}^\top P\left(\widehat{\boldsymbol{\theta}}\right) + P\left(\widehat{\boldsymbol{\theta}}\right)\widehat{B}\left[K - K\left(\widehat{\boldsymbol{\theta}}\right)\right]\right) e^{\overline{A}t}\boldsymbol{dt},$$

where the fact $\overline{A} - \widehat{\overline{A}} = \widehat{B}\left[K - K\left(\widehat{\boldsymbol{\theta}}\right)\right]$ is used above. Then, using $\widehat{B}^\top P\left(\widehat{\boldsymbol{\theta}}\right) = -Q_u K\left(\widehat{\boldsymbol{\theta}}\right)$, it is straightforward to see

$$\begin{aligned} &\left(K - K\left(\widehat{\boldsymbol{\theta}}\right)\right)^\top Q_u\left(K - K\left(\widehat{\boldsymbol{\theta}}\right)\right) \\ = \;& \Phi + \left[K - K\left(\widehat{\boldsymbol{\theta}}\right)\right]^\top \widehat{B}^\top P\left(\widehat{\boldsymbol{\theta}}\right) + P\left(\widehat{\boldsymbol{\theta}}\right)\widehat{B}\left[K - K\left(\widehat{\boldsymbol{\theta}}\right)\right], \end{aligned} \quad (59)$$

which leads to the desired result. ■

C.2 Behaviors of diffusion processes in a neighborhood of the truth

Lemma 11 *Letting $\boldsymbol{\zeta_0}$ be as defined in* (10)*, assume that*

$$\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}\right\| \lesssim \frac{\underline{\lambda}\left(Q_u\right)}{\|B_0\|\|P\left(\boldsymbol{\theta_0}\right)\|}\left(\frac{\left[\boldsymbol{\zeta_0} \wedge 1\right]^p}{p^{1/2}} \wedge \frac{\underline{\lambda}\left(Q_x\right)}{\|P\left(\boldsymbol{\theta_0}\right)\|}\right). \quad (60)$$

Then, for the Riccati equation in (5) *which is denoted by $P\left(\boldsymbol{\theta}\right)$, we have $\left\|P\left(\widehat{\boldsymbol{\theta}}\right)\right\| \lesssim \|P\left(\boldsymbol{\theta_0}\right)\|$. Furthermore, for any eigenvalue λ of $\widehat{A} - \widehat{B}Q_u^{-1}\widehat{B}^\top P\left(\widehat{\boldsymbol{\theta}}\right)$, it holds that $\Re\left(\lambda\right) \lesssim -\underline{\lambda}\left(Q_x\right)\|P\left(\boldsymbol{\theta_0}\right)\|^{-1}$.*

Proof. First, let us write $\overline{A}_0 = A_0 - B_0 Q_u^{-1} B_0^\top P\left(\boldsymbol{\theta_0}\right)$ and

$$\overline{A}_1 = \widehat{A} - \widehat{B}Q_u^{-1}B_0^\top P\left(\boldsymbol{\theta_0}\right) = \overline{A}_0 + E_1,$$

where $E_1 = \widehat{A} - A_0 - \left(\widehat{B} - B_0\right)Q_u^{-1}B_0^\top P\left(\boldsymbol{\theta_0}\right)$. Since $\boldsymbol{r} \leq p$, (60) implies that E_1 satisfies

$$\|E_1\| \lesssim \boldsymbol{r}^{-1/2}\left[\boldsymbol{\zeta_0} \wedge 1\right]^{\boldsymbol{r}}.$$

So, letting $M = \overline{A}_0$ in (36), Lemma 5 leads to the fact that all eigenvalues of $\exp\left(\overline{A}_1\right)$ are inside the unit-circle. Therefore, all eigenvalues of $\overline{A}_1$ are on the open left half-plane of the complex plane. Now, in Lemma 10, let $K = -Q_u^{-1}B_0^\top P\left(\boldsymbol{\theta_0}\right)$ and $\overline{A} = \overline{A}_1$, to obtain the matrix denoted by P in the lemma. Since P satisfies

$$Q_x + K^\top Q_u K = -\overline{A}_1^\top P - P\overline{A}_1 = -\overline{A}_0^\top P - P\overline{A}_0 - E_1^\top P - PE_1,$$

writing Lemma 10 for $\overline{A} = \overline{A}_0$, but replacing Q_x with $Q_x + E_1^\top P + PE_1$, we have

$$P \;\; = \;\; \int_0^\infty e^{\overline{A}_0^\top t}\left[Q_x + K^\top Q_u K + E_1^\top P + PE_1\right] e^{\overline{A}_0 t}\boldsymbol{d}t.$$

However, according to (5), we have

$$P\left(\boldsymbol{\theta_0}\right) \;\; = \;\; \int_0^\infty e^{\overline{A}_0^\top t}\left[Q_x + K^\top Q_u K\right] e^{\overline{A}_0 t}\boldsymbol{d}t. \tag{61}$$

Thus, it holds that

$$P = P\left(\boldsymbol{\theta_0}\right) + \int_0^\infty e^{\overline{A}_0^\top t}\left[E_1^\top P + PE_1\right] e^{\overline{A}_0 t}\boldsymbol{d}t,$$

which leads to

$$\|P\| \leq \|P\left(\boldsymbol{\theta_0}\right)\| + 2\|E_1\|\|P\| \int_0^\infty \left\|e^{\overline{A}_0 t}\right\|^2 \boldsymbol{d}t.$$

We will shortly show that $2\|E_1\| \int_0^\infty \left\|e^{\overline{A}_0 t}\right\|^2 \boldsymbol{d}t < 1$. So, by Lemma 10, we have $\left\|P\left(\widehat{\boldsymbol{\theta}}\right)\right\| \leq \|P\| \lesssim \|P\left(\boldsymbol{\theta_0}\right)\|$, which is the desired result.

To proceed, denote the closed-loop matrix by $\widehat{\overline{A}} = \widehat{A} - \widehat{B}Q_u^{-1}\widehat{B}^\top P\left(\widehat{\boldsymbol{\theta}}\right)$, and let the p dimensional unit vector ν attain the maximum of $\left\|\exp(\widehat{\overline{A}})\nu\right\|$, i.e., $\left\|\exp(\widehat{\overline{A}})\nu\right\| = \left\|\exp(\widehat{\overline{A}})\right\|$. Then, (61) for $\widehat{\boldsymbol{\theta}}$ (instead of $\boldsymbol{\theta_0}$) implies that

$$\left\|P\left(\widehat{\boldsymbol{\theta}}\right)\right\| \;\; \geq \;\; \nu^\top P\left(\widehat{\boldsymbol{\theta}}\right)\nu = \int_0^\infty \nu^\top e^{\widehat{\overline{A}}^\top t}\left[Q_x + P\left(\widehat{\boldsymbol{\theta}}\right)^\top \widehat{B}Q_u^{-1}\widehat{B}^\top P\left(\widehat{\boldsymbol{\theta}}\right)\right] e^{\widehat{\overline{A}} t}\nu\boldsymbol{d}t.$$

Therefore, $\underline{\lambda}\left(Q_x + P\left(\widehat{\boldsymbol{\theta}}\right)^\top \widehat{B}Q_u^{-1}\widehat{B}^\top P\left(\widehat{\boldsymbol{\theta}}\right)\right) \geq \underline{\lambda}\left(Q_x\right)$, together with the fact that the magnitudes of all eigenvalues are smaller than the operator norm, imply that for an arbitrary eigenvalue λ of $\widehat{\overline{A}}$, we have

$$\|P\left(\boldsymbol{\theta_0}\right)\| \gtrsim \left\|P\left(\widehat{\boldsymbol{\theta}}\right)\right\| \geq \underline{\lambda}\left(Q_x\right) \int_0^\infty e^{2\Re(\lambda)t}\boldsymbol{d}t, \tag{62}$$

which leads to the second desired result of the lemma. To complete the proof, we need to establish that $\|E_1\| \int_0^\infty \left\|e^{\overline{A}_0 t}\right\|^2 \boldsymbol{d}t < 1/2$. For that purpose, if we write (62) for $\boldsymbol{\theta_0}$ instead of $\widehat{\boldsymbol{\theta}}$, the condition in (60) implies the above bound. ■

C.3 Perturbation analysis for algebraic Riccati equation in (5)

Lemma 12 *Assume that* (60) *holds. Then, we have*

$$\left\|P\left(\widehat{\boldsymbol{\theta}}\right) - P\left(\boldsymbol{\theta_0}\right)\right\| \lesssim \frac{\|P\left(\boldsymbol{\theta_0}\right)\|^2}{\underline{\lambda}\left(Q_x\right)}\left(1 \vee \left\|Q_u^{-1}B_0^\top P\left(\boldsymbol{\theta_0}\right)\right\|\right)\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}\right\|.$$

Proof. First, fix the dynamics matrix $\widehat{\boldsymbol{\theta}}$, and let $\mathcal{C}$ be a linear segment connecting $\boldsymbol{\theta_0}$ and $\widehat{\boldsymbol{\theta}}$:

$$\mathcal{C} = \left\{(1-\alpha)\boldsymbol{\theta_0} + \alpha\widehat{\boldsymbol{\theta}}\right\}_{0\leq\alpha\leq 1}.$$

Let $\boldsymbol{\theta_1} \in \mathcal{C}$ be arbitrary. Then, the derivative of $P(\boldsymbol{\theta})$ at $\boldsymbol{\theta_1}$ in the direction of $\mathcal{C}$ can be found by using the difference matrices $E_A = \widehat{A} - A_0$, $E_B = \widehat{B} - B_0$. Denote $E = [E_A, E_B]^\top$. Then, we find $P(\boldsymbol{\theta_2})$, where $\boldsymbol{\theta_2} = \boldsymbol{\theta_1} + \epsilon E$, for an infinitesimal value of ϵ. So, we have

$$\begin{aligned}
& P(\boldsymbol{\theta_1}) B_2 Q_u^{-1} B_2^\top P(\boldsymbol{\theta_1}) \\
= \;& \epsilon P(\boldsymbol{\theta_1}) E_B Q_u^{-1} B_2^\top P(\boldsymbol{\theta_1}) + P(\boldsymbol{\theta_1}) B_1 Q_u^{-1} B_2^\top P(\boldsymbol{\theta_1}) \\
= \;& O\left(\epsilon^2\right) + \epsilon P(\boldsymbol{\theta_1}) E_B Q_u^{-1} B_1^\top P(\boldsymbol{\theta_1}) + \epsilon P(\boldsymbol{\theta_1}) B_1 Q_u^{-1} E_B^\top P(\boldsymbol{\theta_1}) + P(\boldsymbol{\theta_1}) B_1 Q_u^{-1} B_1^\top P(\boldsymbol{\theta_1}).
\end{aligned}$$

Therefore, we can calculate $P(\boldsymbol{\theta_2}) B_2 Q_u^{-1} B_2^\top P(\boldsymbol{\theta_2})$. To that end, let $P = P(\boldsymbol{\theta_2}) - P(\boldsymbol{\theta_1})$, write $P(\boldsymbol{\theta_2})$ in terms of $P, P(\boldsymbol{\theta_1})$, and use the above result to get

$$\begin{aligned}
& P(\boldsymbol{\theta_2}) B_2 Q_u^{-1} B_2^\top P(\boldsymbol{\theta_2}) \\
= \;& P(\boldsymbol{\theta_2}) B_2 Q_u^{-1} B_2^\top P + P(\boldsymbol{\theta_2}) B_2 Q_u^{-1} B_2^\top P(\boldsymbol{\theta_1}) \\
= \;& O\left(\|P\|^2\right) + P(\boldsymbol{\theta_1}) B_2 Q_u^{-1} B_2^\top P + P B_2 Q_u^{-1} B_2^\top P(\boldsymbol{\theta_1}) + P(\boldsymbol{\theta_1}) B_2 Q_u^{-1} B_2^\top P(\boldsymbol{\theta_1}) \\
= \;& O\left(\|P\|^2\right) + P(\boldsymbol{\theta_1}) B_2 Q_u^{-1} B_2^\top P + P B_2 Q_u^{-1} B_2^\top P(\boldsymbol{\theta_1}) \\
+ \;& O\left(\epsilon^2\right) + \epsilon P(\boldsymbol{\theta_1}) E_B Q_u^{-1} B_1^\top P(\boldsymbol{\theta_1}) + \epsilon P(\boldsymbol{\theta_1}) B_1 Q_u^{-1} E_B^\top P(\boldsymbol{\theta_1}) \\
+ \;& P(\boldsymbol{\theta_1}) B_1 Q_u^{-1} B_1^\top P(\boldsymbol{\theta_1}).
\end{aligned} \tag{63}$$

Again, expanding $A_2 = A_1 + E_A$ and $P(\boldsymbol{\theta_2}) = P(\boldsymbol{\theta_1}) + P$, it yields to

$$\begin{aligned}
& A_2^\top P(\boldsymbol{\theta_2}) + P(\boldsymbol{\theta_2}) A_2 \\
= \;& A_2^\top P(\boldsymbol{\theta_1}) + A_2^\top P + P(\boldsymbol{\theta_1}) A_2 + P A_2 \\
= \;& A_1^\top P(\boldsymbol{\theta_1}) + \epsilon E_A^\top P(\boldsymbol{\theta_1}) + A_2^\top P \\
+ \;& P(\boldsymbol{\theta_1}) A_1 + \epsilon P(\boldsymbol{\theta_1}) E_A + P A_2.
\end{aligned}$$

To proceed, plug in the continuous-time algebraic Riccati equation in (5) for $\boldsymbol{\theta_1}, \boldsymbol{\theta_2}$ below in the above expression:

$$\begin{aligned}
A_2^\top P(\boldsymbol{\theta_2}) + P(\boldsymbol{\theta_2}) A_2 &= P(\boldsymbol{\theta_2}) B_2 Q_u^{-1} B_2^\top P(\boldsymbol{\theta_2}) + Q_x, \\
A_1^\top P(\boldsymbol{\theta_1}) + P(\boldsymbol{\theta_1}) A_1 &= P(\boldsymbol{\theta_1}) B_1 Q_u^{-1} B_1^\top P(\boldsymbol{\theta_1}) + Q_x.
\end{aligned}$$

So, we obtain

$$\begin{aligned}
& A_2^\top P(\boldsymbol{\theta_2}) + P(\boldsymbol{\theta_2}) A_2 - A_1^\top P(\boldsymbol{\theta_1}) - P(\boldsymbol{\theta_1}) A_1 \\
= \;& \epsilon E_A^\top P(\boldsymbol{\theta_1}) + \epsilon P(\boldsymbol{\theta_1}) E_A + P A_2 + A_2^\top P \\
= \;& P(\boldsymbol{\theta_2}) B_2 Q_u^{-1} B_2^\top P(\boldsymbol{\theta_2}) - P(\boldsymbol{\theta_1}) B_1 Q_u^{-1} B_1^\top P(\boldsymbol{\theta_1}) \\
= \;& O\left(\|P\|^2\right) + P(\boldsymbol{\theta_1}) B_2 Q_u^{-1} B_2^\top P + P B_2 Q_u^{-1} B_2^\top P(\boldsymbol{\theta_1}) \\
+ \;& O\left(\epsilon^2\right) + \epsilon P(\boldsymbol{\theta_1}) E_B Q_u^{-1} B_1^\top P(\boldsymbol{\theta_1}) + \epsilon P(\boldsymbol{\theta_1}) B_1 Q_u^{-1} E_B^\top P(\boldsymbol{\theta_1}),
\end{aligned}$$

where in the last equality above, we used (63). Now, rearrange the terms in the above statement to get an equation that does not contain any expression in term of $\boldsymbol{\theta_2}$. So, it becomes

$$\begin{aligned}
0 \;=\; & \left[A_2^\top - P(\boldsymbol{\theta_1}) B_2 Q_u^{-1} B_2^\top\right] P + P\left[A_2 - B_2 Q_u^{-1} B_2^\top P(\boldsymbol{\theta_1})\right] - O\left(\|P\|^2\right) - O\left(\epsilon^2\right) \\
+ \;& \epsilon E_A^\top P(\boldsymbol{\theta_1}) + \epsilon P(\boldsymbol{\theta_1}) E_A - \epsilon P(\boldsymbol{\theta_1}) E_B Q_u^{-1} B_1^\top P(\boldsymbol{\theta_1}) - \epsilon P(\boldsymbol{\theta_1}) B_1 Q_u^{-1} E_B^\top P(\boldsymbol{\theta_1}).
\end{aligned}$$

Next, to simplify the above equality, define the followings:

$$\begin{aligned}
D &= A_2 - B_2 Q_u^{-1} B_2^\top P(\boldsymbol{\theta_1}), \\
K(\boldsymbol{\theta_1}) &= -Q_u^{-1} B_1^\top P(\boldsymbol{\theta_1}), \\
R &= \epsilon P(\boldsymbol{\theta_1}) \left[E_A + E_B K(\boldsymbol{\theta_1})\right] + \epsilon\left[K(\boldsymbol{\theta_1})^\top E_B^\top + E_A^\top\right] P(\boldsymbol{\theta_1}) - O\left(\epsilon^2\right).
\end{aligned}$$

So, writing our equation in terms of $D, K(\boldsymbol{\theta_1}), R$, it gives

$$0 = D^\top P + PD - O\left(\|P\|^2\right) + R. \tag{64}$$

The discussion after (6) states that all eigenvalues of $\overline{A}_1 = A_1 - B_1 Q_u^{-1} B_1^\top P(\boldsymbol{\theta_1})$ lie in the open left half-plane. Therefore, if ϵ is small enough, real-parts of all eigenvalues of D are negative, according to Lemma 5. Therefore, (64) implies that

$$\overline{\lambda}(P) \leq \overline{\lambda}\left(\int_0^\infty e^{D^\top t} R e^{Dt} \boldsymbol{d}t\right) \leq \|R\| \int_0^\infty \left\|e^{Dt}\right\|^2 \boldsymbol{d}t.$$

So, as ϵ decays, R vanishes, which by the above inequality shows that P shrinks as ϵ tends to zero. Further, as ϵ decays, D converges to $\overline{A}_1$. Thus, by (64), we have

$$\lim_{\epsilon \to 0} \epsilon^{-1} P = \int_0^\infty e^{\overline{A}_1^\top t} \left(P(\boldsymbol{\theta_1}) [E_A + E_B K(\boldsymbol{\theta_1})] + [E_A + E_B K(\boldsymbol{\theta_1})]^\top P(\boldsymbol{\theta_1})\right) e^{\overline{A}_1 t} \boldsymbol{d}t. \tag{65}$$

Recall that the above expression is the derivative of $P(\boldsymbol{\theta})$ at $\boldsymbol{\theta_1}$, along the linear segment $\mathcal{C}$. Thus, integrating along $\mathcal{C}$, (65) and Cauchy-Schwarz Inequality imply that

$$\left\|P\left(\widehat{\boldsymbol{\theta}}\right) - P(\boldsymbol{\theta_0})\right\| \lesssim \left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}\right\| \sup_{\boldsymbol{\theta_1} \in \mathcal{C}} \|P(\boldsymbol{\theta_1})\| (1 \vee \|K(\boldsymbol{\theta_1})\|) \int_0^\infty \left\|e^{\overline{A}_1 t}\right\|^2 \boldsymbol{d}t.$$

Finally, using Lemma 11, (61), and (62), we obtain the desired result. ■

D Numerical Results

In this section, we provide further empirical results illustrating the performance of Algorithm 2 in the settings of flight control, as well and blood glucose control. First, we provide box plots depicting the distribution of the normalized squared estimation error and the normalized regret of Algorithm 2 for X-29A airplane. Note that the corresponding worst- and average-case curves are presented in Figure 1. Then, Figures 3 and 4 provide the corresponding curves of estimation and regret versus time as well as the box-plots, for Boeing 747. Finally, we present similar empirical result for learning to control blood glucose level. As shown in the presented figures, Thompson sampling Algorithm 2 clearly outperforms the competing reinforcement learning policy.

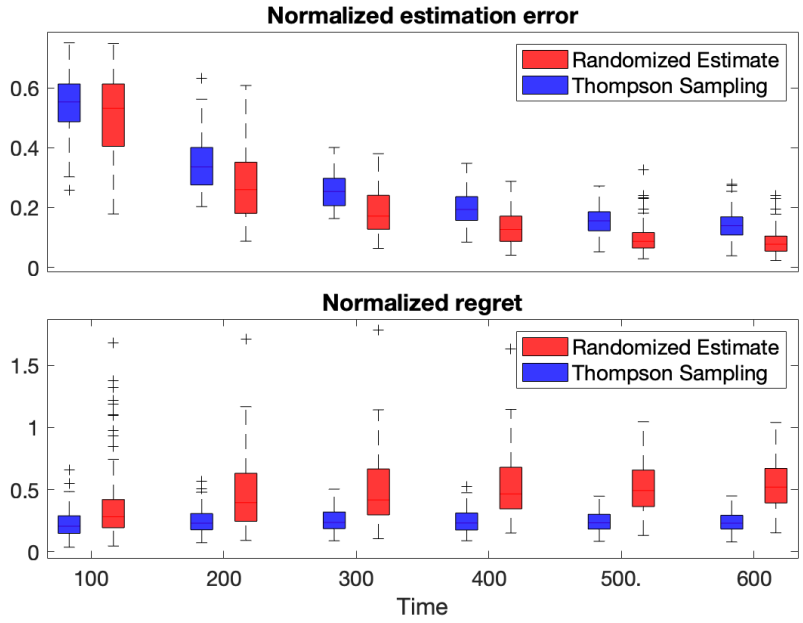

Figure 2: The performance of Algorithm 2 (blue) compared to Randomized Estimate policy (red) [2] for flight control of X-29A airplane. The top box-plots are for the normalized squared estimation error, $\left\|\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta}_{\mathbf{0}}\right\|^2$ divided by $p(p+q)\boldsymbol{\tau}_n^{-1/2}\log \boldsymbol{\tau}_n$, at times 100, 200, 300, 400, 500, and 600 for 100 replications. Similarly, the lower graph showcases the distribution of the regret $\mathrm{Reg}\,(T)$, normalized by $p(p+q)T^{1/2}\log T$.

Figure 2 depicts the box plot corresponding to Figure 1 that is for the flight control of X-29A airplane at 2000 ft. In the following experiments, we keep the setting given in Section 6 for the cost and noise covariance matrices, and compare Algorithm 2 to Randomized Estimate policy [2].

Next, the empirical results of the flight control problem in Boeing 747 airplane at 20000 ft altitude are provided [37]. The true drift matrices of the Boeing 747 are

$$A_0 = \begin{bmatrix} -0.199 & 0.003 & -0.980 & 0.038 \\ -3.868 & -0.929 & 0.471 & -0.008 \\ 1.591 & -0.015 & -0.309 & 0.003 \\ -0.198 & 0.958 & 0.021 & 0.000 \end{bmatrix}, \quad B_0 = \begin{bmatrix} -0.001 & 0.058 \\ 0.296 & 0.153 \\ 0.012 & -0.908 \\ 0.015 & 0.008 \end{bmatrix}.$$

Then, the blood glucose control problem is studied [41, 47]. The true drift matrices are

$$A_0 = \begin{bmatrix} 1.91 & -2.82 & 0.91 \\ 1.00 & -1.00 & 0.00 \\ 0.00 & 1.00 & -1.00 \end{bmatrix}, \quad B_0 = \begin{bmatrix} -0.0992 \\ 0.0000 \\ 0.0000 \end{bmatrix}.$$

Note that from a practical point of view, worst-case behavior are of crucial importance in this problem.

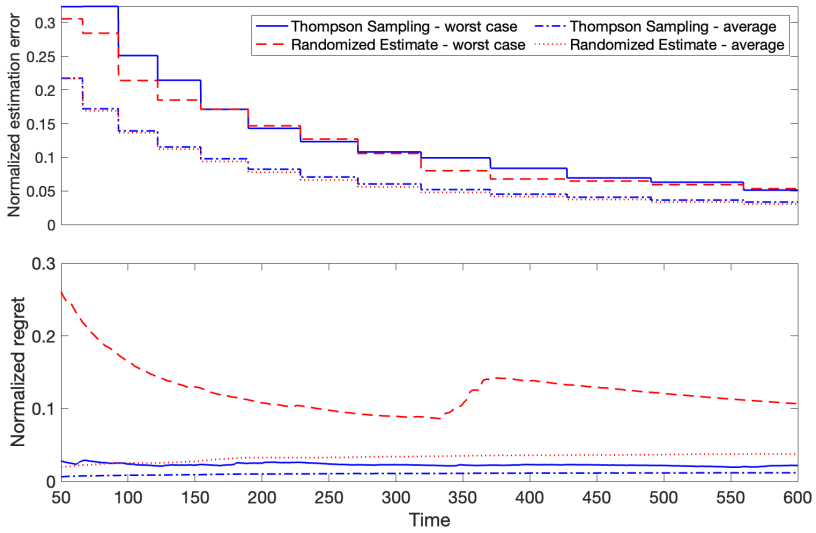

Figure 3: The performance of Algorithm 2 (blue) compared to Randomized Estimate policy (red) [2] for the flight control of Boeing 747 airplane. The top graph plots the normalized squared estimation error, $\left\|\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta_0}\right\|^2$ divided by $p(p+q)\boldsymbol{\tau}_n^{-1/2}\log \boldsymbol{\tau}_n$, for 100 replications. Similarly, the lower graph showcases the regret $\mathrm{Reg}\,(T)$, normalized by $p(p+q)T^{1/2}\log T$.

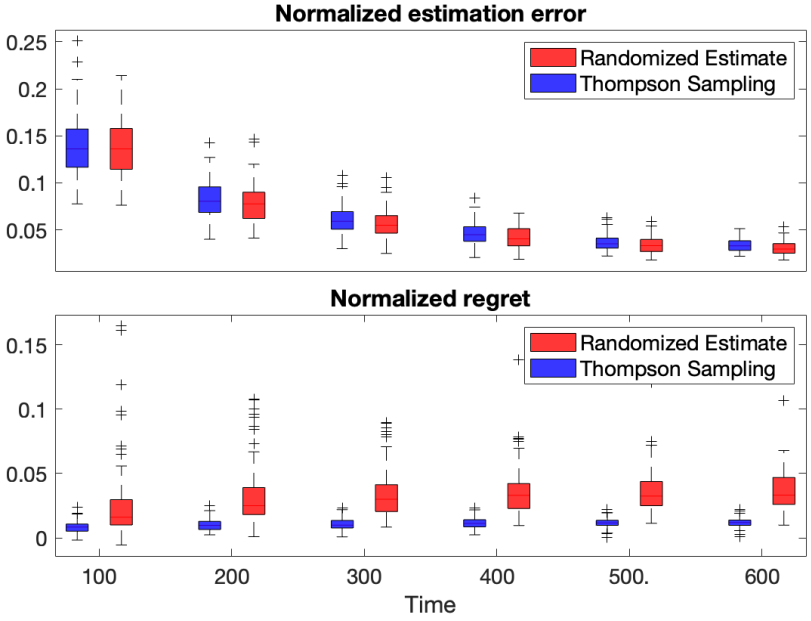

Figure 4: The performance of Algorithm 2 (blue) compared to Randomized Estimate policy (red) [2] for the flight control of Boeing 747 airplane. The top graph plots the normalized squared estimation error, $\left\|\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta_0}\right\|^2$ divided by $p(p+q)\boldsymbol{\tau}_n^{-1/2}\log \boldsymbol{\tau}_n$, at times 100, 200, 300, 400, 500, and 600 for 100 replications. Similarly, the lower graph showcases the regret $\mathrm{Reg}\,(T)$, normalized by $p(p+q)T^{1/2}\log T$.

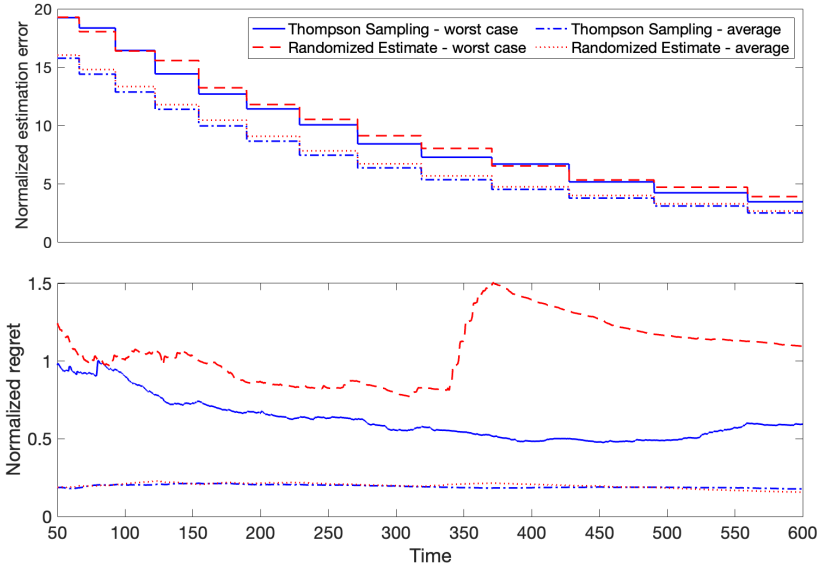

Figure 5: The performance of Algorithm 2 (blue) compared to Randomized Estimate policy (red) [2] for the the blood glucose control. The top graph plots the normalized squared estimation error, $\left\|\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta}_{\mathbf{0}}\right\|^2$ divided by $p(p+q)\boldsymbol{\tau}_n^{-1/2}\log\boldsymbol{\tau}_n$, for 100 replications. Similarly, the lower graph showcases the regret $\mathrm{Reg}\,(T)$, normalized by $p(p+q)T^{1/2}\log T$.

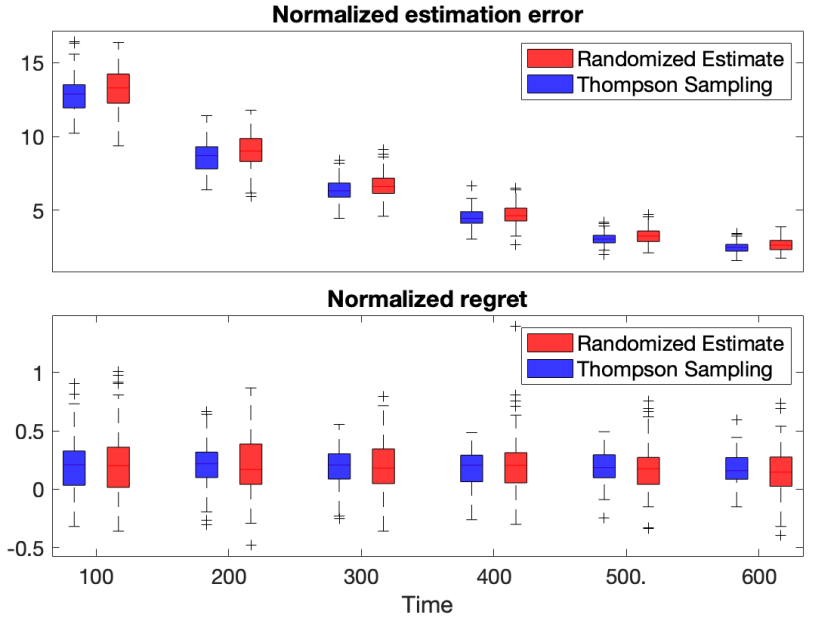

Figure 6: The performance of Algorithm 2 (blue) compared to Randomized Estimate policy (red) [2] for the the blood glucose control. The top graph plots the normalized squared estimation error, $\left\|\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta}_{\mathbf{0}}\right\|^2$ divided by $p(p+q)\boldsymbol{\tau}_n^{-1/2}\log\boldsymbol{\tau}_n$, at times 100, 200, 300, 400, 500, and 600 for 100 replications. Similarly, the lower graph showcases the regret $\mathrm{Reg}\,(T)$, normalized by $p(p+q)T^{1/2}\log T$.