A modified Thompson sampling-based learning algorithm for unknown linear systems

Mukul Gagrani, Sagar Sudhakara, Aditya Mahajan, Ashutosh Nayyar, and Yi Ouyang

Abstract— We revisit the Thompson sampling-based learning algorithm for controlling an unknown linear system with quadratic cost proposed in [1]. This algorithm operates in episodes of dynamic length and it is shown to have a regret bound of $\tilde{\mathcal{O}}(\sqrt{T})$, where T is the time-horizon. The regret bound of this algorithm is obtained under a technical assumption on the induced norm of the closed loop system. We propose a variation of this algorithm that enforces a lower bound $T_{\min}$ on the episode length. We show that a careful choice of $T_{\min}$ (that depends on the uncertainty about the system model) allows us to recover the $\tilde{\mathcal{O}}(\sqrt{T})$ regret bound under a milder technical condition about the closed loop system.

I. INTRODUCTION

The problem of learning an optimal policy for a system with linear dynamics and quadratic cost with unknown parameters has received considerable attention in the literature. Historically, the focus has been on developing algorithms which asymptotically learn optimal policies using techniques from adaptive control and reinforcement learning [2]–[5]. In recent years, the emphasis has shifted towards developing algorithms with finite-time regret guarantees.

Broadly speaking, three classes of learning algorithms have been considered in the literature: optimism in the face of uncertainty [6]–[9], certainty equivalence [10]–[14], and Thompson sampling [14]–[17]. These algorithms provide two kinds of regret guarantees: frequentist and Bayesian. In the frequentist setting, it is established that the regret for the unknown system is bounded with high probability (with respect to the distribution of the process noise and the randomness introduced by the algorithm). In the Bayesian setting, it is assumed that there is a prior on the unknown system parameters and it is established that the expected regret is bounded (where the expectation is with respect to the prior, the distribution of the process noise, and the randomness introduced by the algorithm). These two notions of regret are different and, since the per-step cost is not bounded, one form of the regret does not imply the other.

Mukul Gagrani is with Qualcomm AI research, San Diego. (email: mgagrani@qti.qualcomm.com)

Sagar Sudhakara and Ashutosh Nayyar are with the Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, USA. (email: sagarsud@usc.edu, ashutosn@usc.edu)

Aditya Mahajan is with the department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada. (email: aditya.mahajan@mcgill.ca)

Yi Ouyang is with Preferred Networks America, Burlingame, CA, USA (email: ouyangyi@preferred-america.com)

In this paper, we revisit a recently proposed algorithm for establishing Bayesian regret called Thompson sampling with dynamic episodes (TSDE) [1]. The main result of [1] is to show that the Bayesian regret of TSDE accumulated up to time T is bounded by $\tilde{\mathcal{O}}(\sqrt{T})$, where the $\tilde{\mathcal{O}}(\cdot)$ notation hides constants and poly-logarithmic factors. This result was derived under a technical assumption on the induced norm of the closed loop system. In this paper, we present a variation of the TSDE algorithm and obtain a $\tilde{\mathcal{O}}(\sqrt{T})$ bound on the Bayesian regret by imposing a much milder technical assumption.

II. MODEL AND PROBLEM FORMULATION

We consider the same model as [1]. For the sake of completeness, we present the model below.

Consider a linear system with state $x_t \in \mathbb{R}^n$, control input $u_t \in \mathbb{R}^m$, and disturbance $w_t \in \mathbb{R}^n$. For the ease of exposition, we assume that the system starts from an initial state $x_1 = 0$. The state evolves over time according to

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad t \geq 1, \tag{1}$$

where $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m}$ are the system dynamics matrices. The noise $\{w_t\}_{t \geq 1}$ is an independent and identically distributed Gaussian process with $w_t \sim \mathcal{N}(0, \sigma_w^2 I)$.

Remark 1 In [1], it was assumed that $\sigma_w^2 = 1$. Using a general $\sigma_w^2 > 0$ does not fundamentally change any of the results or the proof arguments.

At each time t, the system incurs a per-step cost given by

$$c(x_t, u_t) = x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t, \tag{2}$$

where Q and R are positive definite matrices.

Let $\theta^\mathsf{T} = [A, B]$ denote the parameters of the system. $\theta \in \mathbb{R}^{d \times n}$, where $d = n + m$. The performance of any policy $\pi = (\pi_1, \pi_2, \dots)$ is measured by the long-term average cost given by

$$J(\pi; \theta) = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}^\pi \Big[\sum_{t=1}^{T} c(x_t, u_t) \Big]. \tag{3}$$

Let $J(\theta)$ denote the minimum of $J(\pi; \theta)$ over all policies. It is well known [18] that if the pair (A, B) is stabilizable, then $J(\theta)$ is given by

$$J(\theta) = \sigma_w^2 \operatorname{Tr}(S(\theta)),$$

where $S(\theta)$ is the unique positive semi-definite solution of the following Riccati equation:

$$S(\theta) = Q + A^\mathsf{T} S(\theta) A$$
$$- A^\mathsf{T} S(\theta) B (R + B^\mathsf{T} S(\theta) B)^{-1} B^\mathsf{T} S(\theta) A. \tag{4}$$

Furthermore, the optimal control policy is given by

$$u_t = G(\theta)x_t, \tag{5}$$

where the gain matrix $G(\theta)$ is given by

$$G(\theta) = -(R + B^\mathsf{T} S(\theta)B)^{-1} B^\mathsf{T} S(\theta)A. \tag{6}$$

As in [1], we are interested in the setting where the system parameters are unknown. We denote the unknown parameters by a random variable θ_1 and assume that there is a prior distribution on θ_1. The Bayesian regret of a policy π operating for horizon T is defined by

$$R(T; \pi) = \mathbb{E}^\pi \left[\sum_{t=1}^{T} c(x_t, u_t) - T J(\theta_1) \right], \tag{7}$$

where the expectation is with respect to the prior on θ_1, the noise processes, the initial conditions, and the potential randomizations done by the policy π.

III. THOMSON SAMPLING BASED LEARNING ALGORITHM

As in [1], we assume that the unknown model parameters θ lie in a compact subset Ω_1 of $\mathbb{R}^{d \times n}$. We use $p|\Omega$ to denote the restriction of probability distribution p on the set Ω. We assume that there is a prior μ_1 on Ω_1 which satisfies the following assumption.

Assumption 1 *There exist* $\hat\theta_1(i) \in \mathbb{R}^d$ *for* $i \in \{1, \ldots, n\}$ *and a positive definite matrix* $\Sigma_1 \in \mathbb{R}^{d \times d}$ *such that for any* $\theta \in \mathbb{R}^{d \times n}$, $\mu_1 = \bar\mu_1\big|_{\Omega_1}$, *where*

$$\bar\mu_1(\theta) = \prod_{i=1}^{n} \bar\mu_1(\theta(i)) \quad and \quad \bar\mu_1(\theta(i)) = \mathcal{N}(\hat\theta_1(i), \Sigma_1).$$

We maintain a posterior distribution μ_t on Ω_1 based on the history $(x_{1:t-1}, u_{1:t-1})$ of the observations until time t. From standard results in linear Gaussian regression [19], we know that the posterior is a truncated Gaussian distribution

$$\mu_t(\theta) = \left[\prod_{i=1}^{n} \bar\mu_t(\theta(i)) \right]\Bigg|_{\Omega_1}$$

where $\bar\mu_t(\theta(i)) = \mathcal{N}(\hat\theta_t(i), \Sigma_t)$ and $\{\hat\theta_t(i)\}_{i=1}^{n}$ and Σ_t can be updated recursively as follows:

$$\hat\theta_{t+1}(i) = \hat\theta_t(i) + \frac{\Sigma_t z_t (x_{t+1}(i) - \hat\theta_t(i)^\mathsf{T} z_t)}{\sigma_w^2 + z_t^\mathsf{T} \Sigma_t z_t}, \tag{8}$$

$$\Sigma_{t+1}^{-1} = \Sigma_t^{-1} + \frac{1}{\sigma_w^2} z_t z_t^\mathsf{T}, \tag{9}$$

where $z_t = [x_t^\mathsf{T}, u_t^\mathsf{T}]^\mathsf{T}$.

A. Thompson sampling with dynamic episodes algorithm

We now present a variation of the Thompson sampling with dynamic episodes (TSDE) algorithm of [1]. As the name suggests, the algorithm operates in episodes of dynamic length. The key difference from [1] is that we enforce that each episode is of a minimum length $T_{\min}$. The choice of $T_{\min}$ will be explained later.

Let t_k and T_k denote the start time and the length of episode k, respectively. Episode k has a minimum length of

---

Algorithm 1 `TSDE`

1: **input:** $\Omega_1, \hat\theta_1, \Sigma_1$
2: **initialization:** $t \leftarrow 1$, $t_0 \leftarrow -T_{\min}$, $T_{-1} \leftarrow T_{\min}$, $k \leftarrow 0$.
3: **for** $t = 1, 2, \ldots$ **do**
4:     observe x_t
5:     update $\bar\mu_t$ according to (8)–(9)
6:     **if** $(t - t_k > T_{\min})$ and
7:         $\big((t - t_k > T_{k-1})$ or $(\det \Sigma_t < \frac{1}{2} \det \Sigma_{t_k})\big)$
8:     **then**
9:         $T_k \leftarrow t - t_k$, $k \leftarrow k+1$, $t_k \leftarrow t$
10:        sample $\bar\theta_k \sim \mu_t$
11:    **end if**
12:    Apply control $u_t = G(\bar\theta_k)x_t$
13: **end for**

$T_{\min}$ and ends when the length of the episode is strictly larger than the length of the previous episode (i.e., $t - t_k > T_{k-1}$) or at the first time after $t_k + T_{\min}$ when the determinant of the covariance Σ_t falls below half of its value at time t_k, i.e., $\det \Sigma_t < \frac{1}{2} \det \Sigma_{t_k}$. Thus,

$$t_{k+1} = \min \left\{ t > t_k + T_{\min} \,\middle|\, \begin{matrix} t - t_k > T_{k-1} \text{ or} \\ \det \Sigma_t < \frac{1}{2} \det \Sigma_{t_k} \end{matrix} \right\}. \tag{10}$$

Note that the stopping condition (10) implies that

$$T_{\min} + 1 \leq T_k \leq T_{k-1} + 1, \quad \forall k \tag{11}$$

If we select $T_{\min} = 0$ in the above algorithm, we recover the stopping condition of [1].

The TSDE algorithm works as follows. At the beginning of episode k, a parameter $\bar\theta_k$ is sampled from the posterior distribution μ_{t_k}. During the episode, the control inputs are generated using the sampled parameters $\bar\theta_k$, i.e.,

$$u_t = G(\bar\theta_k)x_t, \quad t_k \leq t < t_{k+1}. \tag{12}$$

The complete algorithm is presented in Algorithm 1.

B. A technical assumption and the choice of minimum episode length

For each $\theta^\mathsf{T} = [A, B]$, we define a 4-dimensional row-vector $\eta(\theta)$ as follows:

$$\eta(\theta) := (\|\theta\|, \mathrm{Tr}(S(\theta)), \|S(\theta)\|, \|[I, G(\theta)^\mathsf{T}]^\mathsf{T}\|), \tag{13}$$

where $S(\theta)$ is the solution of the Riccati equation in (4) and $G(\theta)$ is the optimal gain matrix defined in (6).

Definition 1 *Let* $M_\theta, M_J, M_S, M_G, \alpha, \delta$ *be positive constants such that* $\alpha \geq 1$ *and* $0 < \delta < 1$. *We say that the uncertainty set* Ω_1 *is of Type* $(M_\theta, M_J, M_S, M_G, \alpha, \delta)$ *if the following conditions hold:*

1) *For all* $\theta \in \Omega_1$,

$$\eta(\theta) \leq (M_\theta, M_J, M_S, M_G) \tag{14}$$

where the inequality is component-wise.

2) For any $\theta, \phi \in \Omega_1$ with $\theta^\mathsf{T} = [A_\theta, B_\theta]$ and for any integer $t \geq 1$,

$$\|(A_\theta + B_\theta G(\phi))^t\| \leq \alpha\delta^t.$$

Assumption 2 *We assume that the uncertainty set Ω_1 is of Type $(M_\theta, M_J, M_S, M_G, \alpha, \delta)$, where M_θ, M_J, M_S, M_G, α, δ are positive constants such that $\alpha \geq 1$ and $0 < \delta < 1$.*

The following simple observation plays a critical role in analyzing the regret of TSDE.

Lemma 1 *Suppose Assumption 2 is true. Define*

$$T_{\min}^* = \left\lceil \frac{\log \alpha}{-\log \delta} \right\rceil. \tag{15}$$

Then, for $\theta, \phi \in \Omega_1$ with $\theta^\mathsf{T} = [A_\theta, B_\theta]$, we have

$$\|(A_\theta + B_\theta G(\phi))^{T_{\min}^*+1}\| < 1 \tag{16}$$

PROOF The proof follows immediately from Assumption 2 and the definition of $T_{\min}^*$. ∎

Before presenting our regret analysis under Assumption 2, we present two special cases of this assumption. The first case is identical to the assumption made in [1] about the uncertainty set.

Assumption 3 *Let M_θ, M_J, M_S, M_G, α, δ be positive constants such that $\alpha \geq 1$ and $0 < \delta < 1$. Assume that the uncertainty set Ω_1 satisfies the following conditions:*

1) *Equation (14) holds for all $\theta \in \Omega_1$.*
2) *For any $\theta, \phi \in \Omega_1$ with $\theta^\mathsf{T} = [A_\theta, B_\theta]$,*

$$\|(A_\theta + B_\theta G(\phi))\| \leq \delta.$$

In [1], part 2) of Assumption 3 was stated explicitly. In addition, the uncertainty set was assumed to be compact, which ensures part 1) of Assumption 3.

An uncertainty set that satisfies Assumption 3 also satisfies Assumption 2 with $\alpha = 1$. This is because if $\|(A_\theta + B_\theta G(\phi))\| \leq \delta$, then

$$\|(A_\theta + B_\theta G(\phi))^t\| \leq (\|(A_\theta + B_\theta G(\phi))\|)^t \leq \delta^t. \tag{17}$$

Thus, Assumption 2 is weaker than Assumption 3 used in [1].

For a square matrix A, let $\rho(A)$ denote the spectral radius of matrix A. The next assumption can also be viewed as a special case of Assumption 2.

Assumption 4 *Let M_θ, M_J, M_S, M_G, α, $\tilde{\delta}$ be positive constants such that $\alpha \geq 1$ and $0 < \tilde{\delta} < 1$. Assume that the uncertainty set Ω_1 satisfies the following conditions:*

1) *Equation (14) holds for all $\theta \in \Omega_1$.*
2) *For any $\theta, \phi \in \Omega_1$ with $\theta^\mathsf{T} = [A_\theta, B_\theta]$,*

$$\rho(A_\theta + B_\theta G(\phi)) \leq \tilde{\delta}.$$

We note that Assumption 4 is weaker than Assumption 3 (since $\rho(A) \leq \|A\|$ for any matrix A). Consider, for example, a family of matrices $A_q = \begin{bmatrix} \tilde{\delta} & q \\ 0 & \tilde{\delta} \end{bmatrix}$, where $q \in \mathbb{N}$ and $0 <$ $\tilde{\delta} < 1$. For each q, the spectral radius of A_q is $\tilde{\delta}$ while its norm is at least q. Thus, each A_q satisfies Assumption 4 but not Assumption 3.

The following lemma shows that an uncertainty set that satisfies Assumption 4 also satisfies Assumption 2 for some constants $\alpha \geq 1$ and $\tilde{\delta} < \delta < 1$.

Lemma 2 *Suppose Assumption 4 is true. Then, there exist $\alpha \geq 1$ and $\tilde{\delta} < \delta < 1$ such that for any $\theta, \phi \in \Omega_1$ with $\theta^\mathsf{T} = [A_\theta, B_\theta]$ and for any integer $t \geq 1$,*

$$\|(A_\theta + B_\theta G(\phi))^t\| \leq \alpha\delta^t.$$

PROOF Define $\varepsilon = \delta - \tilde{\delta}$. Let $\mathcal{L} = \{A_\theta + B_\theta G(\phi) : \theta, \phi \in \Omega_1\}$. Since Ω_1 is compact, so is $\mathcal{L}$. Now for any $L \in \mathcal{L}$, there exists a norm (call it norm_L) such that $\mathrm{norm}_L(L) < \rho(L) + \varepsilon \leq \tilde{\delta} + \varepsilon = \delta$.

Since norms are continuous, there is an open ball centered at L (let's call this ball_L) such that for any $H \in \mathrm{ball}_L$, we have $\mathrm{norm}_L(H) < \delta$. Consider the collection of open balls $\{\mathrm{ball}_L : L \in \mathcal{L}\}$. This is an open cover of compact set $\mathcal{L}$. So, there is a finite sub-cover. Let's denote this sub-cover by $\mathrm{ball}_{L_1}, \ldots, \mathrm{ball}_{L_\ell}$. By equivalence of norms, there is a finite constant α_k such that $\|A\| \leq \alpha_{L_k} \mathrm{norm}_{L_k}(A)$ for any matrix A, for all $k \in \{1, \ldots, \ell\}$. Let $\alpha = \max(1, \max_k \alpha_{L_k})$.

Now consider an arbitrary $H \in \mathcal{L}$. It belongs to ball_{L_k} for some $k \in \{1, \ldots, \ell\}$. Therefore, $\mathrm{norm}_{L_k}(H) < \delta$. Hence, for any integer t, the above inequalities and the submulitplicity of norms give that $\|H^t\| \leq \alpha_{L_k} \mathrm{norm}_{L_k}(H^t) \leq \alpha(\mathrm{norm}_{L_k}(H))^t < \alpha(\delta)^t$. ∎

Remark 2 Condition 2 in Definition 1 states that $A_\theta + B_\theta G(\phi)$ is *uniformly* exponentially stable for all all $\theta, \phi \in \Omega_1$. Condition 2 of Assumption 4 states that $A_\theta + B_\theta G(\phi)$ is *uniformly* asymptotically stable for all $\theta, \phi \in \Omega_1$. For linear systems, asymptotic stability implies exponential stability. In Lemma 2, we are effectively showing that when the uncertainty set is compact, uniform asymptotic stability implies uniform exponential stability.

C. Regret bounds

The following result provides an upper bound on the regret of the proposed algorithm.

Theorem 1 *Under Assumptions 1 and 2 and with $T_{\min} \geq T_{\min}^*$, the regret of TSDE is upper bounded by*

$$R(T; \mathtt{TSDE}) \leq \tilde{\mathcal{O}}(\sigma_w^2(n+m)\sqrt{nT}). \tag{18}$$

The proof is presented in the next section.

Remark 3 The constants hidden in the $\tilde{\mathcal{O}}(\cdot)$ notation in Theorem 1 depend only on the type $(M_\theta, M_J, M_S, M_G, \alpha, \delta)$ of the uncertainty set Ω_1. In particular, these constants do not depend on n, m, and T.

IV. REGRET ANALYSIS

For the ease of notation, we use $R(T)$ instead of $R(T; \mathtt{TSDE})$ in this section. Let K_T denote the number of

episodes until horizon T. Following the exact same steps as [1], we can show that

$$R(T) = R_0(T) + R_1(T) + R_2(T) \tag{19}$$

where

$$R_0(T) = \mathbb{E}\left[\sum_{k=1}^{K_T} T_k J(\bar{\theta}_k)\right] - T\mathbb{E}[J(\theta_1)], \tag{20}$$

$$R_1(T) = \mathbb{E}\left[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left[x_t^\mathsf{T} S(\bar{\theta}_k) x_t - x_{t+1}^\mathsf{T} S(\bar{\theta}_k) x_{t+1}\right]\right] \tag{21}$$

$$R_2(T) = \mathbb{E}\left[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left[(\theta_1^\mathsf{T} z_t)^\mathsf{T} S(\bar{\theta}_k)\theta_1^\mathsf{T} z_t - (\theta_k^\mathsf{T} z_t)^\mathsf{T} S(\bar{\theta}_k)\theta_k^\mathsf{T} z_t\right]\right] \tag{22}$$

We establish the bound on $R(T)$ by individually bounding $R_0(T)$, $R_1(T)$, and $R_2(T)$.

Lemma 3 *The terms in* (19) *are bounded as follows:*
  1) $R_0(T) \leq \tilde{\mathcal{O}}(\sigma_w^2\sqrt{(n+m)T})$.
  2) $R_1(T) \leq \tilde{\mathcal{O}}(\sigma_w^2\sqrt{(n+m)T})$.
  3) $R_2(T) \leq \tilde{\mathcal{O}}(\sigma_w^2(n+m)\sqrt{nT})$.

Combining Lemma 3 with equation (19) establishes Theorem 1. Before presenting the proof of Lemma 3, we establish some preliminary results.

A. Preliminary results

Let $X_T = \sigma_w + \max_{1\leq t\leq T}\|x_t\|$ denote the maximum of the norm of the state plus the noise standard deviation.

Lemma 4 *For any $q \geq 1$ and any $T \geq 1$,*

$$\mathbb{E}\left[\frac{X_T^q}{\sigma_w^q}\right] \leq \mathcal{O}((\log T)^{q/2}).$$

See Appendix II for proof.

Lemma 5 *For any $q \geq 1$, we have*

$$\mathbb{E}\left[\frac{X_T^q}{\sigma_w^q}\log\left(\frac{X_T^2}{\sigma_w^2}\right)\right] \leq \tilde{\mathcal{O}}(1).$$

See Appendix III for proof.

Lemma 6 *The number of episodes is bounded by*

$$K_T \leq \mathcal{O}\left(\sqrt{(n+m)T\log\left(T\frac{X_T^2}{\sigma_w^2}\right)}\right).$$

See Appendix IV for proof.

Remark 4 The statement of Lemmas 4 and 6 are the same as that of the corresponding lemmas in [1]. The proof of Lemma 4 in [1] relied on Assumption 3. Since we impose a weaker assumption, our proof is more involved. The proof of Lemma 6 is similar to the proof of [1, Lemma 3]. However, since our TSDE algorithm is different from that in [1], some of the details of the proof are different.

B. Proof of Lemma 3

We now prove each part of Lemma 3 separately.

1) Proof of bound on $R_0(T)$: Following exactly the same argument as the proof of [1, Lemma 5], we can show that

$$R_0(T) \leq \mathcal{O}(\sigma_w^2\mathbb{E}[K_T]). \tag{23}$$

Substituting the result of Lemma 6, we get

$$R_0(T) \leq \mathcal{O}\left(\sigma_w^2\mathbb{E}\left[\sqrt{(n+m)T\log(TX_T^2/\sigma_w^2)}\right]\right)$$
$$\overset{(a)}{\leq} \mathcal{O}\left(\sigma_w^2\sqrt{(n+m)T\log(T\mathbb{E}[X_T^2/\sigma_w^2])}\right)$$
$$\overset{(b)}{\leq} \tilde{\mathcal{O}}\left(\sigma_w^2\sqrt{(n+m)T}\right)$$

where (a) follows from Jensen's inequality and (b) follows from Lemma 4.

2) Proof of bound on $R_1(T)$: Following exactly the same argument as in the proof of [1, Lemma 6], we can show that

$$R_1(T) \leq \mathcal{O}(\mathbb{E}[K_T X_T^2]) \tag{24}$$

Substituting the result of Lemma 6, we get

$$R_1(T) \leq \mathcal{O}\left(\sqrt{(n+m)T}\,\mathbb{E}\left[X_T^2\sqrt{\log(TX_T^2/\sigma_w^2)}\right]\right) \tag{25}$$

Now, consider the term

$$\mathbb{E}\left[X_T^2\sqrt{\log(TX_T^2/\sigma_w^2)}\right] \overset{(a)}{\leq} \sqrt{\mathbb{E}\left[X_T^4\right]\mathbb{E}\left[\log(TX_T^2/\sigma_w^2)\right]}$$
$$\overset{(b)}{\leq} \sqrt{\mathbb{E}\left[X_T^4\right]\log(T\mathbb{E}[X_T^2/\sigma_w^2])}$$
$$\overset{(c)}{\leq} \tilde{\mathcal{O}}(\sigma_w^2) \tag{26}$$

where (a) follows from Cauchy-Schwartz inequality, (b) follows from Jensen's inequality, and (c) follows from Lemma 4.

Substituting (26) in (25), we get the bound on $R_1(T)$.

3) Proof of bound on $R_2(T)$: As in [1], we can bound the inner summand in $R_2(T)$ as

$$\|S(\bar{\theta}_k)^{0.5}\theta_1^\mathsf{T} z_t\|^2 - \|S(\bar{\theta}_k)^{0.5}\theta_k^\mathsf{T} z_t\|^2 \leq \mathcal{O}(X_T\|(\theta_1-\bar{\theta}_k)^\mathsf{T} z_t\|).$$

Therefore,

$$R_2(T) \leq \mathcal{O}\left(\mathbb{E}\left[X_T\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\|(\theta_1-\bar{\theta}_k)^\mathsf{T} z_t\|\right]\right),$$

which is same as [1, Eq. (45)]. Now, by simplifying the term inside $\mathcal{O}(\cdot)$ using Cauchy-Schwartz inequality, we get

$$\mathbb{E}\left[X_T\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\|(\theta_1-\bar{\theta}_k)^\mathsf{T} z_t\|\right]$$
$$\leq \sqrt{\mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\|\Sigma_{t_k}^{-0.5}(\theta_1-\bar{\theta}_k)\|^2\right]}$$
$$\times \sqrt{\mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}X_T^2\|\Sigma_{t_k}^{0.5}z_t\|^2\right]} \tag{27}$$

Note that (27) is slightly different than the simplification of [1, Eq. (45)] using Cauchy-Schwartz inequality presented in [1, Eq. (46)], which used Σ_t in each term in the right hand side instead of Σ_{t_k}.

We bound each term of (27) separately as follows.

Lemma 7 *We have the following inequality*

$$\mathbb{E}\left[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \|\Sigma_{t_k}^{-0.5}(\theta_1 - \bar{\theta}_k)\|^2 \right]$$
$$\leq \mathcal{O}(n(n+m)(T + \mathbb{E}[K_T])) \leq \mathcal{O}(n(n+m)T).$$

See Appendix V for a proof.

Lemma 8 *We have the following inequality*

$$\mathbb{E}\left[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} X_T^2 \|\Sigma_{t_k}^{0.5} z_t\|^2 \right] \leq \tilde{\mathcal{O}}\big((n+m)\sigma_w^4\big)$$

See Appendix VI for a proof.

We get the bound on $R_2(T)$ by substituting the result of Lemmas 7 and 8 in (27).

V. DISCUSSION AND CONCLUSION

In this paper, we present a variation of the TSDE algorithm of [1] and show that its Bayesian regret up to time T is bounded by $\tilde{\mathcal{O}}(\sqrt{T})$ under a milder technical assumption than [1]. The result in [1] was derived under the assumption that there exists a $\delta < 1$ such that for any $\theta, \phi \in \Omega_1$, $\|A_\theta + B_\theta G(\phi)\| \leq \delta$. For our analysis, we impose a different assumption for the closed loop gain when the system dynamics are θ and the controller is chosen according to ϕ. We show that the assumption of [1] implies our assumption. Our assumption is also implied by $\rho(A_\theta + B_\theta G(\phi)) \leq \delta$.

The key technical result in [1] as well as our paper is Lemma 4, which shows that for any $q \geq 1$, $\mathbb{E}[X_T^q/\sigma_w^q] \leq \tilde{\mathcal{O}}(\log T)$. The proof argument in both [1] as well as our paper is to show that there is some constant α_0 such that $X_T \leq \sigma_w + \alpha_0 W_T$. Under the stronger assumption in [1], one can show that for all t, $\|x_{t+1}\| \leq \delta\|x_t\| + \|w_t\|$, which directly implies that $X_T \leq \sigma_w + W_T/(1-\delta)$. Under the weaker assumption in this paper, the argument is more subtle. The basic intuition is that in each episode, the system is asymptotically stable and, being a linear system, also exponentially stable (in the sense of Lemma 2). So, if the episode length is sufficiently long, then we can ensure that $\|x_{t_{k+1}}\| \leq \beta\|x_{t_k}\| + \bar{\alpha} W_T$, where $\beta < 1$ and $\bar{\alpha}$ is a constant. This is sufficient to ensure that $X_T \leq \sigma_w + \alpha_0 W_T$ for an appropriately defined α_0.

The fact that each episode must be of length $T_{\min}$ implies that the second triggering condition is not triggered for the first $T_{\min}$ steps in an episode. Therefore, in this interval, the determinant of the covariance can be smaller than half of its value at the beginning of the episode. Consequently, we cannot use the same proof argument as [1] to bound $R_2(T)$ because that proof relied on the fact that for any $t \in \{t_k, \ldots, t_{k+1} - 1\}$, $\det \Sigma_t^{-1}/\det \Sigma_{t_k}^{-1} \leq 2$. So, we provide a variation of that proof argument, where we use a coarser bound on $\det \Sigma_t^{-1}/\det \Sigma_{t_k}^{-1}$ given by Lemma 11.

We conclude by observing that the milder technical assumption imposed in this paper may not be necessary. Numerical experiments indicate that the regret of the TSDE algorithm shows $\tilde{\mathcal{O}}(\sqrt{T})$ behavior even when the uncertainty set Ω_1 does not satisfy Assumption 4 (as was also reported in [1]). This suggests that it might be possible to further relax Assumption 4 and still establish an $\tilde{\mathcal{O}}(\sqrt{T})$ regret bound.

APPENDIX I
AN AUXILIARY RESULT

Lemma 9 *For any $q > 0$, we have*

$$\mathbb{E}\left[\max_{1 \leq t \leq T} \frac{\|w_t\|^q}{\sigma_w^q} \right] \leq \mathcal{O}((\log T)^{q/2}) \qquad (28)$$

PROOF For ease of notation, define random variables $\bar{w}_t^{(q)} = \|w_t\|^q/\sigma_w^q$ and $\bar{w}_t^* = (\bar{w}_t^{(q)})^{2/q}$. Note that $\bar{w}_t^*$ has a χ^2-distribution with n-degrees of freedom. Therefore, $\bar{w}_t^*$ has a moment generating function $\mathbb{E}[e^{s\bar{w}_t^*}] = (1 - 2s)^{-n/2}$ for $s < 1/2$. Pick a $\lambda \in (0, 1/2)$. By Chernoff bound, we have

$$\mathbb{P}(\bar{w}_t^* > z) \leq \frac{\mathbb{E}[e^{\lambda \bar{w}_t^*}]}{e^{\lambda z}} = C_\lambda e^{-\lambda z},$$

where $C_\lambda = (1 - 2\lambda)^{-n/2}$. Therefore, the complementary CDF of $\bar{w}_t^{(q)}$ is bounded by

$$1 - F_{\bar{w}_t^{(q)}}(z) = \mathbb{P}(\bar{w}_t^{(q)} > z) = \mathbb{P}(\bar{w}_t^* > z^{q/2}) \leq C_\lambda e^{-\lambda z^{q/2}}.$$

Now, pick $\lambda' \in (0, \lambda)$ and consider an i.i.d. process $\{\xi_t\}_{t \geq 1}$, where ξ_t has a CDF $F_{\xi_t}(z) = 1 - e^{-\lambda' z^{q/2}}$. We let $\xi_{(T)}$ denote $\max_{1 \leq t \leq T} \xi_t$ and use a similar notation for $\bar{w}_{(T)}^{(q)}$. Note that ξ_t has a Weibull distribution with shape $q/2$. Therefore, (see e.g., [20, Eq. (3)])

$$\mathbb{E}[\xi_{(T)}] = \mathcal{O}((\log T)^{q/2}). \qquad (29)$$

Now we present a bound on $\mathbb{E}[\bar{w}_{(T)}^{(q)}]$ in terms of $\mathbb{E}[\xi_{(T)}]$. Since $0 < \lambda' < \lambda$, there exists a $z^\circ > 0$ such that for all $z > z^\circ$, $F_{\bar{w}_t^{(q)}}(z) > F_{\xi_t}(z)$. Thus, $\bar{w}_t^{(q)}$ is stochastically dominated by ξ_t in the *weak* stochastic order[1], as defined in [20]. Therefore, by [20, Theorem 2.1], there exists a constant $c > 0$ such that $\bar{w}_t^{(q)}$ is stochastically dominated by $c\xi_t$ in the convex order.[2] Consequently, by [20, Theorem 3.1(1)] (or [21, Theorem 2.2(1)]), we have that $\mathbb{E}[\bar{w}_{(T)}^{(q)}] \leq c\mathbb{E}[\xi_{(T)}]$. Substituting (29) establishes the result of the Lemma. ∎

[1] A random variable x is said to be dominated by a random variable y in the weak stochastic order if for all increasing functions f supported sufficiently away from 0, $\mathbb{E}[f(x)] \leq \mathbb{E}[f(y)]$.

[2] A random variable x is said to be dominated by a random variable y in the convex order if for all increasing and convex functions f, $\mathbb{E}[f(x)] \leq \mathbb{E}[f(y)]$.

APPENDIX II
PROOF OF LEMMA 4

For the ease of notation, let $\bar{\alpha} = \alpha/(1-\delta)$, and $\beta = \alpha\delta^{T_{\min}+1}$. In addition, define $W_T = \max_{1 \le t \le T} \|w_t\|$, $\bar{X}_k = \max_{t_k < t \le t_{k+1}} \|x_t\|$, $Y_k = \|x_{t_k}\|$, and $H_k = A + BG(\bar{\theta}_k)$ where A and B are the true parameters.

From the system dynamics under the TSDE algorithm, we know that for any time $t \in \{t_k + 1, \ldots, t_{k+1}\}$, we have

$$x_t = H_k^{t-t_k} x_{t_k} + \sum_{j=t_k}^{t-1} H_k^{t-1-j} w_j.$$

Thus, from triangle inequality and Assumption 2, we get

$$\|x_t\| \le \alpha\delta^{t-t_k} Y_k + \left[\sum_{j=t_k}^{t-1} \alpha\delta^{t-1-j}\right] W_T$$

$$\le \alpha\delta^{t-t_k} Y_k + \underbrace{\left[\frac{\alpha}{1-\delta}\right]}_{=: \bar{\alpha}} W_T. \qquad (30)$$

Now at time $t = t_{k+1}$, we have

$$Y_{k+1} = \|x_{t_{k+1}}\| \le \alpha\delta^{T_k} Y_k + \bar{\alpha} W_T.$$
$$\le \beta Y_k + \bar{\alpha} W_T \qquad (31)$$

where the second inequality follows from (11), which implies $\alpha\delta^{T_k} \le \alpha\delta^{T_{\min}+1} =: \beta$. From Lemma 1, $\beta < 1$. Recursively expanding (31), we get

$$Y_k \le \bar{\alpha} W_T + \beta\bar{\alpha} W_T + \cdots + \beta^{k-2}\bar{\alpha} W_T$$
$$\le \frac{\bar{\alpha}}{1-\beta} W_T =: \bar{\beta} W_T. \qquad (32)$$

Substituting (32) is (30), we get that for any $t \in \{t_k + 1, \ldots, t_{k+1}\}$, we have

$$\|x_t\| \le \alpha\delta^{t-t_k} \bar{\beta} W_T + \bar{\alpha} W_T \le \underbrace{[\alpha\bar{\beta} + \bar{\alpha}]}_{=: \alpha_0} W_T$$

where in the last inequality, we have used the fact that $\delta \in (0,1)$. Thus, for any episode k, we have

$$\bar{X}_k = \max_{t_k < t \le t_{k+1}} \|x_t\| \le \alpha_0 W_T.$$

Hence,

$$X_T \le \sigma_w + \max\{\bar{X}_1, \ldots, \bar{X}_{K_T}\} \le \sigma_w + \alpha_0 W_T.$$

Therefore, for any $q \ge 1$, we have

$$\mathbb{E}[X_T^q] \le \sum_{p=0}^{q} \binom{q}{p} \sigma_w^{q-p} \alpha_0^p \mathbb{E}[W_T^p] \qquad (33)$$

From Lemma 9, we have that

$$\sigma_w^{q-p} \mathbb{E}[W_T^p] = \sigma_w^{q-p} \mathbb{E}\left[\max_{1 \le t \le T} \|w_t\|^p\right] \le \sigma_w^q \mathcal{O}((\log T)^{p/2}).$$

Substituting this is (33), we obtain the result of the lemma.

APPENDIX III
PROOF OF LEMMA 5

Since log is an increasing function, $\log x \le \log \max(e, x)$ for any $x > 0$. Therefore,

$$\mathbb{E}\left[\frac{X_T^q}{\sigma_w^q} \log\left(\frac{X_T^2}{\sigma_w^2}\right)\right] \le \mathbb{E}\left[\frac{X_T^q}{\sigma_w^q} \log\max(e, X_T^2/\sigma_w^2)\right]$$

$$\le \sqrt{\mathbb{E}\left[\frac{X_T^{2q}}{\sigma_w^{2q}}\right] \mathbb{E}\left[\left(\log\max(e, X_T^2/\sigma_w^2)\right)^2\right]} \qquad (34)$$

where the last inequality follows from Cauchy-Schwartz inequality. Since $(\log x)^2$ is concave for $x \ge e$, we can use Jensen's inequality to write

$$\mathbb{E}\left[\left(\log\max(e, X_T^2/\sigma_w^2)\right)^2\right] \le \left(\log(\mathbb{E}[\max(e, X_T^2/\sigma_w^2)])\right)^2$$
$$\le \left(\log(e + \mathbb{E}[X_T^2/\sigma_w^2])\right)^2$$
$$\overset{(a)}{\le} \left(\log(e + \mathcal{O}(\log T))\right)^2$$
$$\le \tilde{\mathcal{O}}(1) \qquad (35)$$

where (a) uses Lemma 4. Substituting (35) in (34) and using Lemma 4 for bounding $\mathbb{E}[X_T^{2q}/\sigma_w^{2q}]$, we get

$$\mathbb{E}\left[\frac{X_T^q}{\sigma_w^q} \log\left(\frac{X_T^2}{\sigma_w^2}\right)\right] \le \sqrt{\mathbb{E}\left[\frac{X_T^{2q}}{\sigma_w^{2q}}\right] \tilde{\mathcal{O}}(1)} \le \tilde{\mathcal{O}}(1).$$

APPENDIX IV
PROOF OF LEMMA 6

The high-level idea of the proof is same as that of [1, Lemma 3]. Define macro episodes with start times t_{n_i}, $i \in \mathbb{N}_{>0}$, where $n_1 = 1$ and for $i \ge 1$,

$$n_{i+1} = \min\left\{k > n_i \,\big|\, \det\Sigma_{t_k} < \tfrac{1}{2}\det\Sigma_{t_{k-1}}\right\}.$$

Thus, a new macro-episode starts whenever an episode ends due to the second stopping criterion. Let M denote the number of macro-episodes until time T and define $n_{M+1} = K_T + 1$. Let $\bar{T}_i$ denote the length of the i-th macro-episode. Within a macro-episode, all but the last episode must be triggered by the first stopping criterion. Thus, for $k \in \{n_i, n_i + 1, \ldots, n_{i+1} - 2\}$,

$$T_k = \max\{T_{k-1} + 1, T_{\min} + 1\} = T_{k-1} + 1$$

where the last equality follows from (11). Hence, by following exactly the same argument as [1], we have

$$n_{i+1} - n_i \le \sqrt{2\bar{T}_i}$$

and therefore following [1, Eq. (40)], we have

$$K_T \le \sqrt{2MT} \qquad (36)$$

which is same as [1, Eq. (41)].

Now, observe that

$$\det\Sigma_T^{-1} \overset{(a)}{\ge} \det\Sigma_{t_{n_M}}^{-1} \overset{(b)}{\ge} 2\det\Sigma_{t_{n_{M-1}}}^{-1}$$
$$\ge \cdots \ge 2^{M-1}\det\Sigma_1^{-1}, \qquad (37)$$

where (a) follows because $\{\det\Sigma_t^{-1}\}_{t \ge 1}$ is a non-decreasing sequence (because $\Sigma_1^{-1} \le \Sigma_2^{-1} \ldots$) and (b) and subsequent

inequalities follow from the definition of the macro episode and the second triggering condition.

Then following the same idea as the rest of the proof in [1], we get

$$M \leq \mathcal{O}((n+m)\log(TX_T^2/\sigma_w^2)). \qquad (38)$$

Substituting (38) in (36), we obtain the result of the lemma.

APPENDIX V
PROOF OF LEMMA 7

Observe that the summand is constant for each episode. Therefore,

$$\mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\left[\|\Sigma_{t_k}^{-0.5}(\theta_1 - \bar{\theta}_k)^{\mathsf{T}}\|^2\right]\right]$$

$$= \mathbb{E}\left[\sum_{k=1}^{K_T}\left[T_k\|\Sigma_{t_k}^{-0.5}(\theta_1 - \bar{\theta}_k)^{\mathsf{T}}\|^2\right]\right]$$

$$\overset{(a)}{\leq} \mathbb{E}\left[\sum_{k=1}^{K_T}\left[(T_{k-1}+1)\|\Sigma_{t_k}^{-0.5}(\theta_1 - \bar{\theta}_k)^{\mathsf{T}}\|^2\right]\right]$$

$$= \sum_{k=1}^{\infty}\mathbb{E}\left[\mathbb{1}_{\{t_k \leq T\}}(T_{k-1}+1)\|\Sigma_{t_k}^{-0.5}(\theta_1 - \bar{\theta}_k)^{\mathsf{T}}\|^2\right]$$

$$= \sum_{k=1}^{\infty}\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{t_k \leq T\}}(T_{k-1}+1)\|\Sigma_{t_k}^{-0.5}(\theta_1 - \bar{\theta}_k)^{\mathsf{T}}\|^2 \mid h_{t_k}\right]\right]$$

$$\overset{(b)}{=} \sum_{k=1}^{\infty}\mathbb{E}\left[\mathbb{1}_{\{t_k \leq T\}}(T_{k-1}+1)\mathbb{E}\left[\|\Sigma_{t_k}^{-0.5}(\theta_1 - \bar{\theta}_k)^{\mathsf{T}}\|^2 \mid h_{t_k}\right]\right]$$

$$\overset{(c)}{\leq} \sum_{k=1}^{\infty}\mathbb{E}\left[\mathbb{1}_{\{t_k \leq T\}}(T_{k-1}+1)2(n+m)n\right]$$

$$\leq 2(n+m)n(T + \mathbb{E}[K_T]), \qquad (39)$$

where (a) follows from (11), (b) follows from the fact that $\mathbb{1}_{\{t_k < T\}}(T_{k-1}+1)$ is $\sigma(h_{t_k})$ measurable, and (c) hold because conditioned on h_{t_k} each column of $\|\Sigma_{t_k}^{-0.5}(\theta_1 - \bar{\theta}_k)^{\mathsf{T}}\|^2$ is the difference of two i.i.d. vectors $\sim \mathcal{N}(0, I)$.

Eq. (39) proves the first part of the Lemma. The second part follows from the fact that $K_T \leq T$.

APPENDIX VI
PROOF OF LEMMA 8

For any $s < t$. Eq. (9) implies that $\Sigma_s^{-1} \preceq \Sigma_t^{-1}$ and consequently Σ_t^{-1} is positive definite. Therefore, we have the following:

Lemma 10 *Let $\lambda_{\min}$ be the smallest eigenvalue of Σ_1^{-1}. Then, each eigenvalue of Σ_t^{-1} is no less than $\lambda_{\min}$. Therefore, each eigenvalue of Σ_t is no more than $1/\lambda_{\min}$.*

An immediate implication of Lemma 10 is the following: For any t and s,

$$z_t^{\mathsf{T}}\Sigma_s z_t \leq \frac{1}{\lambda_{\min}}\|z_t\|^2 \leq \frac{1}{\lambda_{\min}}M_G^2 X_T^2, \qquad (40)$$

where $M_G = \sup_{\theta \in \Omega_1}\|[I, G(\theta)^{\mathsf{T}}]^{\mathsf{T}}\|$.

For any $s < t$, $\Sigma_s^{-1} \preceq \Sigma_t^{-1}$ implies that $\Sigma_s \succeq \Sigma_t$. Therefore, from [7, Lemma 11], we get that for any $V \neq 0$ (of appropriate dimensions),

$$\frac{\|V^{\mathsf{T}}\Sigma_s V\|}{\|V^{\mathsf{T}}\Sigma_t V\|} \leq \frac{\det \Sigma_s}{\det \Sigma_t} = \frac{\det \Sigma_t^{-1}}{\det \Sigma_s^{-1}}. \qquad (41)$$

Eq. (41) implies that for any $t \in \{t_k, \ldots, t_{k+1}-1\}$, we have

$$\|\Sigma_{t_k}^{0.5}z_t\|^2 = z_t^{\mathsf{T}}\Sigma_{t_k}z_t \leq \frac{\det \Sigma_t^{-1}}{\det \Sigma_{t_k}^{-1}}z_t^{\mathsf{T}}\Sigma_t z_t \qquad (42)$$

For the ease of notation, let $\tau_k = t_k + T_{\min}$. Then we have the following bound on $\det \Sigma_t^{-1}/\det \Sigma_{t_k}^{-1}$.

Lemma 11 *The following inequalities hold:*

1) *For $t \in \{t_k, \ldots, \tau_k\}$, we have*

$$\frac{\det \Sigma_t^{-1}}{\det \Sigma_{t_k}^{-1}} \leq \left(1 + \frac{1}{\lambda_{\min}\sigma_w^2}M_G^2 X_T^2\right)^{T_{\min}}.$$

2) *For $t \in \{\tau_k+1, \ldots, t_{k+1}-1\}$, we have*

$$\frac{\det \Sigma_t^{-1}}{\det \Sigma_{t_k}^{-1}} \leq 2.$$

Consequently, for all $t \in \{t_k, \ldots, t_{k+1}-1\}$, we have

$$\frac{\det \Sigma_t^{-1}}{\det \Sigma_{t_k}^{-1}} \leq \left(2 + \frac{M_G^2 X_T^2}{\lambda_{\min}\sigma_w^2}\right)^{T_{\min}\vee 1}. \qquad (43)$$

PROOF The second relationship follows from the second stopping criterion. We now prove the first relationship. Eq. (9) implies that

$$\Sigma_{t+1}^{-1} = \Sigma_t^{-1}\left(I + \frac{1}{\sigma_w^2}\Sigma_t z_t z_t^{\mathsf{T}}\right).$$

Therefore,

$$\frac{\det \Sigma_{t+1}^{-1}}{\det \Sigma_t^{-1}} = \det\left(I + \frac{1}{\sigma_w^2}\Sigma_t z_t z_t^{\mathsf{T}}\right) = 1 + \frac{1}{\sigma_w^2}z_t^{\mathsf{T}}\Sigma_t z_t$$

$$\leq 1 + \frac{1}{\lambda_{\min}\sigma_w^2}M_G^2 X_T^2, \qquad (44)$$

where the last inequality follows from (40). Thus, for any $t \in \{t_k, \ldots, \tau_k\}$, we have

$$\frac{\det \Sigma_t^{-1}}{\det \Sigma_{t_k}^{-1}} \leq \left(1 + \frac{1}{\lambda_{\min}\sigma_w^2}M_G^2 X_T^2\right)^{t-t_k}$$

$$\leq \left(1 + \frac{1}{\lambda_{\min}\sigma_w^2}M_G^2 X_T^2\right)^{T_{\min}} \qquad (45)$$

where the first inequality follows by repeatedly applying (44) as a telescopic product.

Let $\bar{M} = M_G^2 X_T^2/\lambda_{\min}\sigma_w^2$. Then, (43) follows by observing that $(1 + \bar{M})^{T_{\min}} \leq (2 + \bar{M})^{T_{\min}\vee 1}$ and $2 < (2 + \bar{M})^{T_{\min}\vee 1}$. ∎

Using Lemma 11 and (42), we get

$$\sum_{t=t_k}^{t_{k+1}-1} \|\Sigma_{t_k}^{0.5} z_t\|^2 \le \sum_{t=t_k}^{t_{k+1}-1} \frac{\det \Sigma_t^{-1}}{\det \Sigma_{t_k}^{-1}} z_t^{\mathsf{T}} \Sigma_t z_t$$

$$\le \left(2 + \frac{M_G^2 X_T^2}{\lambda_{\min}\sigma_w^2}\right)^{T_{\min}\vee 1} \sum_{t=t_k}^{t_{k+1}-1} z_t^{\mathsf{T}} \Sigma_t z_t \tag{46}$$

where the first inequality follows from (42) and the second inequality follows from Lemma 11. Therefore,

$$\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} X_T^2 \|\Sigma_{t_k}^{0.5} z_t\|^2$$

$$\le \left(2 + \frac{M_G^2 X_T^2}{\lambda_{\min}\sigma_w^2}\right)^{T_{\min}\vee 1} X_T^2 \sum_{t=1}^{T} z_t^{\mathsf{T}} \Sigma_t z_t \tag{47}$$

From (40) for $s = t$, we get that

$$z_t^{\mathsf{T}} \Sigma_t z_t \le \max\left(\sigma_w^2, \frac{M_G^2 X_T^2}{\lambda_{\min}}\right) \min\left(1, \frac{z_t^{\mathsf{T}} \Sigma_t z_t}{\sigma_w^2}\right). \tag{48}$$

Hence

$$\sum_{t=1}^{T} z_t^{\mathsf{T}} \Sigma_t z_t \le \left(\sigma_w^2 + \frac{M_G^2 X_T^2}{\lambda_{\min}}\right) \sum_{t=1}^{T} \min\left(1, \frac{z_t^{\mathsf{T}} \Sigma_t z_t}{\sigma_w^2}\right) \tag{49}$$

Using (9) and the intermediate step of the proof of [22, Lemma 6], we have

$$\sum_{t=1}^{T} \min\left(1, \frac{z_t^{\mathsf{T}} \Sigma_t z_t}{\sigma_w^2}\right) = \sum_{t=1}^{T} \min\left(1, \left\|\frac{\Sigma_t^{0.5} z_t z_t^{\mathsf{T}} \Sigma_t^{0.5}}{\sigma_w^2}\right\|\right)$$

$$\le 2(n + m) \log\left(\frac{\mathrm{Tr}(\Sigma_{T+1}^{-1})}{(n + m)}\right) - \log \det \Sigma_1^{-1}. \tag{50}$$

Now, from (9), we get that

$$\mathrm{Tr}(\Sigma_{T+1}^{-1}) = \mathrm{Tr}(\Sigma_1^{-1}) + \sum_{t=1}^{T} \frac{1}{\sigma_w^2} \mathrm{Tr}(z_t z_t^{\mathsf{T}})$$

$$\le \mathrm{Tr}(\Sigma_1^{-1}) + \frac{T}{\sigma_w^2} M_G^2 X_T^2, \tag{51}$$

where the last inequality uses the fact that $\mathrm{Tr}(z_t z_t^{\mathsf{T}}) = \mathrm{Tr}(z_t^{\mathsf{T}} z_t) = \|z_t\|^2 \le M_G^2 X_T^2$. Combining (49) with (50) and (51), we get

$$\sum_{t=1}^{T} z_t^{\mathsf{T}} \Sigma_t z_t \le \mathcal{O}\big((n+m)(\sigma_w^2 + X_T^2)\log(TX_T^2/\sigma_w^2)\big). \tag{52}$$

Therefore, we can bound the expectation of the right hand side of (47) as

$$\mathbb{E}\left[\left(2 + \frac{M_G^2 X_T^2}{\lambda_{\min}\sigma_w^2}\right)^{T_{\min}\vee 1} X_T^2 \sum_{t=1}^{T} z_t^{\mathsf{T}} \Sigma_t z_t\right]$$

$$\le \mathcal{O}\big(\sigma_w^4(n + m)\mathbb{E}[F(X_T)]\big)$$

$$\le \tilde{\mathcal{O}}(\sigma_w^4(n + m)), \tag{53}$$

where the first inequality follows from (52) with $F(X_T) = \left(2 + \frac{M_G^2 X_T^2}{\lambda_{\min}\sigma_w^2}\right)^{T_{\min}\vee 1} \left(\frac{X_T^2}{\sigma_w^2} + \frac{X_T^4}{\sigma_w^4}\right)\log(TX_T^2/\sigma_w^2)$, and the last

inequality follows from Lemma 5 by noting that $F(X_T)$ is a polynomial of X_T/σ_w multiplied by a poly-log term.

The result follows from (47) and (53).

REFERENCES

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

[2] K. J. Astrom and B. Wittenmark, *Adaptive Control*. Addison-Wesley Longman Publishing Co., Inc., 1994.

[3] P. E. Caines, *Linear Stochastic Systems*. John Wiley, 1988. Republished by *Society of Industrial Applied Mathematics,* 2018.

[4] S. J. Bradtke, "Reinforcement learning applied to linear quadratic regulation," in *Neural Information Processing Systems*, pp. 295–302, 1993.

[5] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, "Adaptive linear quadratic control using policy iteration," in *Proceedings of American Control Conference*, vol. 3, pp. 3475–3479, IEEE, 1994.

[6] M. C. Campi and P. Kumar, "Adaptive linear quadratic Gaussian control: the cost-biased approach revisited," *SIAM Journal on Control and Optimization*, vol. 36, no. 6, pp. 1890–1907, 1998.

[7] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems," in *Annual Conference on Learning Theory*, pp. 1–26, 2011.

[8] A. Cohen, T. Koren, and Y. Mansour, "Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret," in *International Conference on Machine Learning*, pp. 1300–1309, PMLR, 2019.

[9] M. Abeille and A. Lazaric, "Efficient optimistic exploration in linear-quadratic regulators via lagrangian relaxation," in *International Conference on Machine Learning*, pp. 23–31, PMLR, 2020.

[10] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "Regret bounds for robust adaptive control of the linear quadratic regulator," in *Neural Information Processing Systems*, pp. 4192–4201, 2018.

[11] H. Mania, S. Tu, and B. Recht, "Certainty equivalent control of LQR is efficient." arXiv:1902.07826, 2019.

[12] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Input perturbations for adaptive control and learning," *Automatica*, vol. 117, p. 108950, 2020.

[13] M. Simchowitz and D. Foster, "Naive exploration is optimal for online lqr," in *International Conference on Machine Learning*, pp. 8937–8948, PMLR, 2020.

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

[15] Y. Ouyang, M. Gagrani, and R. Jain, "Control of unknown linear systems with thompson sampling," in *Allerton Conference on Communication, Control, and Computing*, pp. 1198–1205, 2017.

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

[17] M. Gagrani, S. Sudhakara, A. Mahajan, A. Nayyar, and Y. Ouyang, "Thompson sampling for linear quadratic mean-field teams," in *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 720–727, IEEE, 2021.

[18] K. J. Astrom, *Introduction to stochastic control theory*. Academic Press New York, 1970.

[19] J. Sternby, "On consistency for the method of least squares using martingale theory," *IEEE Trans. Autom. Control*, vol. 22, no. 3, pp. 346–352, 1977.

[20] P. J. Downey and R. S. Maier, "Stochastic orderings and the growth of expected extremes," tech. rep., Tech. Report 90-9, Department of Computer Science, University of Arizona, 1990.

[21] P. J. Downey and R. S. Maier, "Orderings arising from expected extremes, with an application," in *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, pp. 66–75, Hayward, CA: Institute of Mathematical Statistics, 1992.

[22] Y. Abbasi-Yadkori and C. Szepesvari, "Bayesian optimal control of smoothly parameterized systems: The lazy posterior sampling algorithm." arXiv preprint arXiv:1406.3926, 2014.