A relaxed technical assumption for posterior sampling-based reinforcement learning for control of unknown linear systems

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

***Abstract*—We revisit the Thompson sampling algorithm to control an unknown linear quadratic (LQ) system recently proposed by Ouyang et al. [1]. The regret bound of the algorithm was derived under a technical assumption on the induced norm of the closed loop system. In this technical note, we show that by making a minor modification in the algorithm (in particular, ensuring that an episode does not end too soon), this technical assumption on the induced norm can be replaced by a milder assumption in terms of the spectral radius of the closed loop system. The modified algorithm has the same Bayesian regret of $\tilde{\mathcal{O}}(\sqrt{T})$, where T is the time-horizon and the $\tilde{\mathcal{O}}(\cdot)$ notation hides logarithmic terms in T.**

I. Introduction

In a recent paper, Ouyang et al. [1] presented a Thompson sampling (also called posterior sampling) algorithm to control a linear quadratic (LQ) system with unknown parameters. Their algorithm is 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 logarithmic factors. This result was derived under a technical assumption on the induced norm of the closed loop system. In this technical note, we present a minor 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, which is in terms of the spectral radius of the closed loop system (rather than the induced norm, as was assumed in [1]).

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 quadratic system with state $x_t \in \mathbb{R}^n$, control input $u_t \in \mathbb{R}^m$, and disturbance $w_t \in \mathbb{R}^n$. We assume that the system starts from an initial state $x_1 = 0$ and 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

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

where Q and R are positive definite matrices.

Let $\theta^\intercal = [A, B]$ 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 [4] 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:

$$\begin{aligned} S(\theta) = Q + A^\intercal S(\theta) A \\ - A^\intercal S(\theta) B (R + B^\intercal S(\theta) B)^{-1} B^\intercal S(\theta) A. \end{aligned} \tag{4}$$

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

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

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

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

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)

[1] In [1], the algorithm is called posterior sampling reinforcement learning (PSRL-LQ). We use the term TSDE as it was used in [2], which was the conference version of [1], and is also used in other variations of the algorithm [3].

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 assume that there is a prior μ_1 on Ω_1, which satisfies the following.

Assumption 1 *There exist $\hat{\theta}_1(i) \in \mathbb{R}^d$ for $i \in \{1,\dots,n\}$ 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*

$$\bar{\mu}_1(\theta) = \prod_{i=1}^{n} \bar{\mu}_1(\theta(i)) \quad \textit{and} \quad \bar{\mu}_1(\theta(i)) = \mathcal{N}(\hat{\theta}_1(i), \Sigma_1).$$

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 [5], we know that the posterior is a truncated Gaussian distribution

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

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:

$$\hat{\theta}_{t+1}(i) = \hat{\theta}_t(i) + \frac{\Sigma_t z_t (x_{t+1}(i) - \hat{\theta}_t(i)^\intercal z_t)}{\sigma_w^2 + z_t^\intercal \Sigma_t z_t}, \tag{8}$$

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

where $z_t = [x_t^\intercal, u_t^\intercal]^\intercal$.

A. Thompson sampling with dynamic episodes algorithm

We now present a minor 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 $T_{\min}$ 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

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

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.,

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

The complete algorithm is presented in Algorithm 1.

Algorithm 1 TSDE

input: Ω_1, $\hat{\theta}_1$, Σ_1
initialization: $t \leftarrow 1$, $t_0 \leftarrow -T_{\min}$, $T_{-1} \leftarrow T_{\min}$, $k \leftarrow 0$.
for $t = 1, 2, \dots$ **do**
  observe x_t
  update $\bar{\mu}_t$ according to (8)–(9)
  if $(t - t_k > T_{\min})$ and
    $\big((t - t_k > T_{k-1})$ or $(\det \Sigma_t < \frac{1}{2} \det \Sigma_{t_k})\big)$
  then
    $T_k \leftarrow t - t_k$, $k \leftarrow k + 1$, $t_k \leftarrow t$
    sample $\bar{\theta}_k \sim \mu_t$
  end if
  Apply control $u_t = G(\bar{\theta}_k) x_t$
end for

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

We make the following assumption on the support of the prior distribution.

Assumption 2 *There exists a positive number $\delta < 1$ such that for any $\theta, \phi \in \Omega_1$, where $\theta^\intercal = [A_\theta, B_\theta]$,*

$$\rho(A_\theta + B_\theta G(\phi)) \le \delta.$$

Assumption 2 is a weaker form of the following assumption imposed in [1] (since $\rho(A) \le ||A||$ for any matrix A).

Assumption 3 *There exists a positive number $\delta < 1$ such that for any $\theta, \phi \in \Omega_1$, where $\theta^\intercal = [A_\theta, B_\theta]$,*

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

Note that it is much easier to satisfy Assumption 2 than Assumption 3. For example, consider a family of matrices $A_q = \begin{bmatrix} 1-\delta & q \\ 0 & 1-\delta \end{bmatrix}$, where $q \in \mathbb{N}$ and $0 < \delta < 1$. For each q, the sprectral radius of A_q is $1 - \delta$ while its norm is at least q. Thus, each A_q satisfies Assumption 2 but not Assumption 3.

Lemma 1 *Assumption 2 implies that for any $\varepsilon \in (0, 1-\delta)$, there exists an $\alpha \ge 1$ such that for any $\theta, \phi \in \Omega_1$ with $\theta^\intercal = [A_\theta, B_\theta]$ and for any integer $t \ge 1$,*

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

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

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 $\mathrm{norm}_L(H) < \delta + \varepsilon$. 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

$\mathrm{ball}_{L_1}, \dots, \mathrm{ball}_{L_\ell}$. By equivalence of norms, there is a finite constant α_k such that $\|A\| \le \alpha_{L_k} \mathrm{norm}_{L_k}(A)$ for any matrix A, for all $k \in \{1, \dots, \ell\}$. 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 $k \in \{1, \dots, \ell\}$. Therefore, $\mathrm{norm}_{L_k}(H) < \delta + \varepsilon$. Hence, for any integer t, the above inequalities and the submulitplicity of norms give that $\|H^t\| \le \alpha_{L_k} \mathrm{norm}_{L_k}(H^t) \le \alpha(\mathrm{norm}_{L_k}(H))^t < \alpha(\delta+\varepsilon)^t$. ∎

A key implication of Lemma 1 is the following, which plays a critical role in analyzing the regret of TSDE.

Lemma 2 *Fix an $\varepsilon \in (0, 1-\delta)$ and let α be as given by Lemma 1. Define*

$$T^*_{\min} = \left\lceil \frac{\log \alpha}{-\log(\varepsilon + \delta)} \right\rceil. \tag{13}$$

Then, for $\theta, \phi \in \Omega_1$ with $\theta^\intercal = [A_\theta, B_\theta]$ and $\tau \ge T^*_{\min}$, we have

$$\|(A_\theta + B_\theta G(\phi))^\tau\| \le 1 \tag{14}$$

PROOF The proof follows immediately from the choice of $T^*_{\min}$, Lemma 1 and (11). ∎

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 $T_{\min} \ge T^*_{\min}$, the regret of TSDE is upper bounded by*

$$R(T; \texttt{TSDE}) \le \tilde{\mathcal{O}}(\sigma_w^2 (n+m)\sqrt{nT}). \tag{15}$$

The proof is presented in the next section.

IV. Regret analysis

For the ease of notation, we use $R(T)$ instead of $R(T; \texttt{TSDE})$ in this section. Following the exact same steps as [1], we can show that

$$R(T) = R_0(T) + R_1(T) + R_2(T) \tag{16}$$

where

$$R_0(T) = \mathbb{E}\left[\sum_{k=1}^{K_T} T_k J(\bar\theta_k)\right] - T\mathbb{E}[J(\theta_1)], \tag{17}$$

$$R_1(T) = \mathbb{E}\left[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left[x_t^\intercal S(\bar\theta_k) x_t - x_{t+1}^\intercal S(\bar\theta_k) x_{t+1}\right]\right] \tag{18}$$

$$R_2(T) = \mathbb{E}\left[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left[(\theta_1^\intercal z_t)^\intercal S(\bar\theta_k) \theta_1^\intercal z_t - (\theta_k^\intercal z_t)^\intercal S(\bar\theta_k) \theta_k^\intercal z_t\right]\right] \tag{19}$$

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* (16) *are bounded as follows:*

1) $R_0(T) \le \tilde{\mathcal{O}}(\sigma_w^2 \sqrt{(n+m)T})$.
2) $R_1(T) \le \tilde{\mathcal{O}}(\sigma_w^2 \sqrt{(n+m)T})$.
3) $R_2(T) \le \tilde{\mathcal{O}}(\sigma_w^2 (n+m)\sqrt{nT})$.

Before presenting the proof of this lemma, we establish some preliminary results.

A. Preliminary results

Let $X_T = \max_{1 \le t \le T} \|x_t\|$ denote the maximum of the norm of the state and K_T denote the number of episodes until horizon T.

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

$$\mathbb{E}[X_T^q] \le \mathcal{O}(\sigma_w^q \log T).$$

See Appendix A for proof.

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

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

See Appendix B for proof.

Lemma 6 *The number of episodes is bounded by*

$$K_T \le \mathcal{O}\Big(\sqrt{(n+m)T \log(TX_T^2)}\Big).$$

See Appendix C for proof.

Remark 2 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

$$R_0(T) \le \mathcal{O}(\sigma_w^2 \mathbb{E}[K_T]). \tag{20}$$

Substituting the result of Lemma 6, we get

$$\begin{aligned} R_0(T) &\le \mathcal{O}\Big(\sigma_w^2 \mathbb{E}\Big[\sqrt{(n+m)T \log(TX_T^2)}\Big]\Big) \\ &\overset{(a)}{\le} \mathcal{O}\Big(\sigma_w^2 \sqrt{(n+m)T \log(T\mathbb{E}[X_T^2])}\Big) \\ &\overset{(b)}{\le} \tilde{\mathcal{O}}\big(\sigma_w^2 \sqrt{(n+m)T}\big) \end{aligned}$$

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

$$R_1(T) \le \mathcal{O}(\mathbb{E}[K_T X_T^2]) \tag{21}$$

Substituting the result of Lemma 6, we get

$$R_1(T) \le \mathcal{O}\Big(\sqrt{(n+m)T}\,\mathbb{E}\big[X_T^2\sqrt{\log(TX_T^2)}\big]\Big) \tag{22}$$

Now, consider the term

$$\begin{aligned}\mathbb{E}\big[X_T^2\sqrt{\log(TX_T^2)}\big] &\overset{(a)}{\le} \sqrt{\mathbb{E}\big[X_T^4\big]\mathbb{E}\big[\log(TX_T^2)\big]} \\ &\overset{(b)}{\le} \sqrt{\mathbb{E}\big[X_T^4\big]\log(T\mathbb{E}[X_T^2])} \\ &\overset{(c)}{\le} \tilde{\mathcal{O}}(\sigma_w^2)\end{aligned} \tag{23}$$

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

Substituting (23) in (22), 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

$$\|S(\bar\theta_k)^{0.5}\theta_1^\intercal z_t\|^2 - \|S(\bar\theta_k)^{0.5}\theta_k^\intercal z_t\|^2 \le \mathcal{O}(X_T\|(\theta_1-\bar\theta_k)^\intercal z_t\|).$$

Therefore,

$$R_2(T) \le \mathcal{O}\Bigg(\mathbb{E}\Bigg[X_T\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\|(\theta_1-\bar\theta_k)^\intercal z_t\|\Bigg]\Bigg),$$

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

$$\begin{aligned}&\mathbb{E}\Bigg[X_T\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\|(\theta_1-\bar\theta_k)^\intercal z_t\|\Bigg] \\ &\quad\le \sqrt{\mathbb{E}\Bigg[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\|\Sigma_{t_k}^{-0.5}(\theta_1-\bar\theta_k)\|^2\Bigg]} \\ &\qquad\times\sqrt{\mathbb{E}\Bigg[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}X_T^2\|\Sigma_{t_k}^{0.5}z_t\|^2\Bigg]}\end{aligned} \tag{24}$$

Note that (24) 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 (24) separately as follows.

Lemma 7 *We have the following inequality*

$$\begin{aligned}&\mathbb{E}\Bigg[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\|\Sigma_{t_k}^{-0.5}(\theta_1-\bar\theta_k)\|^2\Bigg] \\ &\quad\le \mathcal{O}(n(n+m)(T+\mathbb{E}[K_T])) \le \mathcal{O}(n(n+m)T).\end{aligned}$$

See Appendix D for a proof.

Lemma 8 *We have the following inequality*

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

See Appendix E for a proof.

We get the bound on $R_2(T)$ by substituting the result of Lemmas 7 and 8 in (24).

V. Discussion and Conclusion

In this paper, we present a minor 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 that [1]. The result in [1] was derived under the assumption that there exists a $\delta < 1$ such that for any $\theta, \phi \in \Omega_1$, $\|A_\theta + B_\theta G(\phi)\| \le \delta$. We require that $\rho(A_\theta + B_\theta G(\phi)) \le \delta$. Our assumption on the sprectral radius of the closed loop system is milder and, in some sense, more natural than the assumption on the induced norm of the closed loop system.

The key technical result in [1] as well as our paper is Lemma 4, which shows that for any $q \ge 1$, $\mathbb{E}[X_T^q] \le \tilde{\mathcal{O}}(\sigma_w^2 \log T)$. The proof argument in both [1] as well as our paper is to show that there is some constant α_0 such that $X_T \le \alpha_0 W_T$. Under the stronger assumption in [1], one can show that for all t, $\|x_{t+1}\| \le \delta\|x_t\| + \|w_t\|$, which directly implies that $X_T \le W_T/(1-\delta)$. 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 1). So, if the episode length is sufficiently long, then we can ensure that $\|x_{t_{k+1}}\| \le \beta\|x_{t_k}\| + \bar\alpha W_T$, where $\beta < 1$ and $\bar\alpha$ is a constant. This is sufficient to ensure that $X_T \le \alpha_0 W_T$ 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 $t \in \{t_k, \dots, t_{k+1}-1\}$, $\det\Sigma_t^{-1}/\det\Sigma_{t_k}^{-1} \le 2$. 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 10.

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 2 (as was also reported in [1]). This suggests that it might be possible to further relax Assumption 2 and still establish an $\tilde{\mathcal{O}}(\sqrt{T})$ regret bound.

Appendix A
Proof of Lemma 4

For the ease of notation, let $\bar\delta = \delta + \varepsilon$, $\bar\alpha = \alpha/(1-\bar\delta)$, and $\beta = \alpha\bar\delta^{T_{\min}+1}$. 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 $t \in \{t_k+1, \dots, t_{k+1}\}$, 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 Lemma 1, we get

$$\begin{aligned}\|x_t\| &\le \alpha\bar{\delta}^{t-t_k}Y_k + \left[\sum_{j=t_k}^{t-1}\alpha\bar{\delta}^{t-1-j}\right]W_T \\ &\le \alpha\bar{\delta}^{t-t_k}Y_k + \underbrace{\left[\frac{\alpha}{1-\bar{\delta}}\right]}_{=:\bar{\alpha}} W_T. \end{aligned} \tag{25}$$

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

$$\begin{aligned}Y_{k+1} = \|x_{t_{k+1}}\| &\le \alpha\bar{\delta}^{T_k}Y_k + \bar{\alpha}W_T. \\ &\le \beta Y_k + \bar{\alpha}W_T \end{aligned} \tag{26}$$

where the second inequality follows from (11), which implies $\alpha\bar{\delta}^{T_k} \le \alpha\bar{\delta}^{T_{\min}+1} =: \beta$. Recursively expanding (26), we get

$$\begin{aligned}Y_k &\le \bar{\alpha}W_T + \beta\bar{\alpha}W_T + \cdots + \beta^{k-2}\bar{\alpha}W_T \\ &\le \frac{\bar{\alpha}}{1-\beta}W_T =: \bar{\beta}W_T. \end{aligned} \tag{27}$$

Substituting (27) is (25), we get that for any $t \in \{t_k+1, \ldots, t_{k+1}\}$, we have

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

where in the last inequality, we have used the fact that $\bar{\delta} \in (0,1)$. 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,

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

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

$$\mathbb{E}[X_T^q] \le \alpha_0^q \mathbb{E}[W_T^q] = \alpha_0^q \mathbb{E}\Big[\max_{1\le t\le T}\|w_t\|^q\Big] \tag{28}$$

From [1, Eq. (39)], we have that

$$\mathbb{E}\Big[\max_{1\le t\le T}\|w_t\|^q\Big] \le \sigma_w^q \mathcal{O}(\log T).$$

Substituting this is (28), we obtain the result of the lemma.

Appendix B
Proof of Lemma 5

Since log is an increasing function, $\log X_T^2 \le \log\max(e, X_T^2)$. Therefore,

$$\begin{aligned}\mathbb{E}[X_T^q \log X_T^2] &\le \mathbb{E}[X_T^q \log\max(e, X_T^2)] \\ &\le \sqrt{\mathbb{E}[X_T^{2q}]\ \mathbb{E}[(\log\max(e, X_T^2))^2]} \end{aligned} \tag{29}$$

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

$$\begin{aligned}\mathbb{E}[(\log\max(e, X_T^2))^2] &\le \big(\log(\mathbb{E}[\max(e, X_T^2)])\big)^2 \\ &\le \big(\log(e + \mathbb{E}[X_T^2])\big)^2 \\ &\overset{(a)}{\le} \big(\log(e + \mathcal{O}(\sigma_w^2 \log T))\big)^2 \\ &\le \tilde{\mathcal{O}}(1) \end{aligned} \tag{30}$$

where (a) uses Lemma 4. Substituting (30) in (29) and using Lemma 4 for bounding $\mathbb{E}[X_T^{2q}]$, we get

$$\begin{aligned}\mathbb{E}[X_T^q \log X_T^2] &\le \sqrt{\mathbb{E}[X_T^{2q}]\ \mathbb{E}[(\log\max(e, X_T^2))^2]} \\ &\le \sigma_w^q \tilde{\mathcal{O}}(1). \end{aligned}$$

Appendix C
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 $i \ge 1$,

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

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

$$K_T \le \sqrt{2MT} \tag{31}$$

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

Now, observe that

$$\begin{aligned}\det\Sigma_T^{-1} &\overset{(a)}{\ge} \det\Sigma_{t_{n_M}}^{-1} \overset{(b)}{\ge} 2\det\Sigma_{t_{n_{M-1}}}^{-1} \\ &\ge \cdots \ge 2^{M-1}\det\Sigma_1^{-1}, \end{aligned} \tag{32}$$

where (a) follows because $\{\det\Sigma_t^{-1}\}_{t\ge1}$ 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

$$M \le \mathcal{O}((n+m)\log(TX_T^2)). \tag{33}$$

Substituting (33) in (31), we obtain the result of the lemma.

Appendix D
Proof of Lemma 7

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

$$\begin{aligned}&\mathbb{E}\Bigg[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\big[\|\Sigma_{t_k}^{-0.5}(\theta_1 - \bar{\theta}_k)^\intercal\|^2\big]\Bigg] \\ &= \mathbb{E}\Bigg[\sum_{k=1}^{K_T}\big[T_k\|\Sigma_{t_k}^{-0.5}(\theta_1 - \bar{\theta}_k)^\intercal\|^2\big]\Bigg] \\ &\overset{(a)}{\le} \mathbb{E}\Bigg[\sum_{k=1}^{K_T}\big[(T_{k-1}+1)\|\Sigma_{t_k}^{-0.5}(\theta_1 - \bar{\theta}_k)^\intercal\|^2\big]\Bigg] \end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{\infty} \mathbb{E}\big[\mathbb{1}_{\{t_k \le T\}}(T_{k-1}+1)\|\Sigma_{t_k}^{-0.5}(\theta_1-\bar{\theta}_k)^{\intercal}\|^2\big] \\
&= \sum_{k=1}^{\infty} \mathbb{E}\Big[\mathbb{E}\big[\mathbb{1}_{\{t_k \le T\}}(T_{k-1}+1)\|\Sigma_{t_k}^{-0.5}(\theta_1-\bar{\theta}_k)^{\intercal}\|^2 \mid h_{t_k}\big]\Big] \\
&\overset{(b)}{=} \sum_{k=1}^{\infty} \mathbb{E}\Big[\mathbb{1}_{\{t_k \le T\}}(T_{k-1}+1)\mathbb{E}\big[\|\Sigma_{t_k}^{-0.5}(\theta_1-\bar{\theta}_k)^{\intercal}\|^2 \mid h_{t_k}\big]\Big] \\
&\overset{(c)}{\le} \sum_{k=1}^{\infty} \mathbb{E}\big[\mathbb{1}_{\{t_k \le T\}}(T_{k-1}+1)2(n+m)n\big] \\
&\le 2(n+m)n(T+\mathbb{E}[K_T]),
\end{aligned} \tag{34}
$$

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 $\|\Sigma_{t_k}^{-0.5}(\theta_1-\bar{\theta}_k)^{\intercal}\|^2$ is the difference of two i.i.d. vectors $\sim \mathcal{N}(0,I)$.

Eq. (34) proves the first part of the Lemma. The second part follows from the fact that $K_T \le T$.

Appendix E
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 9 *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 9 is the following: For any t and s,

$$
z_t^{\intercal}\Sigma_s z_t \le \frac{1}{\lambda_{\min}}\|z_t\|^2 \le \frac{1}{\lambda_{\min}} M_G^2 X_T^2, \tag{35}
$$

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

For any $s<t$, $\Sigma_s^{-1} \preceq \Sigma_t^{-1}$ implies that $\Sigma_s \succeq \Sigma_t$. Therefore, from [6, Lemma 11], we get that for any $V \neq 0$ (of appropriate dimensions),

$$
\frac{\|V^{\intercal}\Sigma_s V\|}{\|V^{\intercal}\Sigma_t V\|} \le \frac{\det\Sigma_s}{\det\Sigma_t} = \frac{\det\Sigma_t^{-1}}{\det\Sigma_s^{-1}}. \tag{36}
$$

Eq. (36) implies that for any $t \in \{t_k, \dots, t_{k+1}-1\}$, we have

$$
\|\Sigma_{t_k}^{0.5} z_t\|^2 = z_t^{\intercal}\Sigma_{t_k} z_t \le \frac{\det\Sigma_t^{-1}}{\det\Sigma_{t_k}^{-1}} z_t^{\intercal}\Sigma_t z_t \tag{37}
$$

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 10 *The following inequalities hold:*

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

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

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

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

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

$$
\frac{\det\Sigma_t^{-1}}{\det\Sigma_{t_k}^{-1}} \le \left(2 + \frac{M_G^2 X_T^2}{\lambda_{\min}\sigma_w^2}\right)^{T_{\min}\vee 1}. \tag{38}
$$

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

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

Therefore,

$$
\begin{aligned}
\frac{\det\Sigma_{t+1}^{-1}}{\det\Sigma_t^{-1}} &= \det\left(I + \frac{1}{\sigma_w^2}\Sigma_t z_t z_t^{\intercal}\right) = 1 + \frac{1}{\sigma_w^2} z_t^{\intercal}\Sigma_t z_t \\
&\le 1 + \frac{1}{\lambda_{\min}\sigma_w^2} M_G^2 X_T^2,
\end{aligned} \tag{39}
$$

where the last inequality follows from (35). Thus, for any $t \in \{t_k, \dots, \tau_k\}$, we have

$$
\begin{aligned}
\frac{\det\Sigma_t^{-1}}{\det\Sigma_{t_k}^{-1}} &\le \left(1 + \frac{1}{\lambda_{\min}\sigma_w^2} M_G^2 X_T^2\right)^{t-t_k} \\
&\le \left(1 + \frac{1}{\lambda_{\min}\sigma_w^2} M_G^2 X_T^2\right)^{T_{\min}}
\end{aligned} \tag{40}
$$

where the first inequality follows by repeatedly applying (39) as a telescopic product.

Let $\bar{M} = M_G^2 X_T^2/\lambda_{\min}\sigma_w^2$. Then, (38) follows by observing that $(1+\bar{M})^{T_{\min}} \le (2+\bar{M})^{T_{\min}\vee 1}$ and $2 < (2+\bar{M})^{T_{\min}\vee 1}$.■

Using Lemma 10 and (37), we get

$$
\begin{aligned}
\sum_{t=t_k}^{t_{k+1}-1} \|\Sigma_{t_k}^{0.5} z_t\|^2 &\le \sum_{t=t_k}^{t_{k+1}-1} \frac{\det\Sigma_t^{-1}}{\det\Sigma_{t_k}^{-1}} z_t^{\intercal}\Sigma_t z_t \\
&\le \left(2 + \frac{M_G^2 X_T^2}{\lambda_{\min}\sigma_w^2}\right)^{T_{\min}\vee 1} \sum_{t=t_k}^{t_{k+1}-1} z_t^{\intercal}\Sigma_t z_t
\end{aligned} \tag{41}
$$

where the first inequality follows from (37) and the second inequality follows from Lemma 10. Therefore,

$$
\begin{aligned}
&\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} X_T^2 \|\Sigma_{t_k}^{0.5} z_t\|^2 \\
&\quad\le \left(2 + \frac{M_G^2 X_T^2}{\lambda_{\min}\sigma_w^2}\right)^{T_{\min}\vee 1} X_T^2 \sum_{t=1}^{T} z_t^{\intercal}\Sigma_t z_t
\end{aligned} \tag{42}
$$

From (35) for $s=t$, we get that

$$
z_t^{\intercal}\Sigma_t z_t \le \max\left(\sigma_w^2, \frac{M_G^2 X_T^2}{\lambda_{\min}}\right)\min\left(1, \frac{z_t^{\intercal}\Sigma_t z_t}{\sigma_w^2}\right). \tag{43}
$$

Hence

$$
\sum_{t=1}^{T} z_t^{\intercal}\Sigma_t z_t \le \left(\sigma_w^2 + \frac{M_G^2 X_T^2}{\lambda_{\min}}\right)\sum_{t=1}^{T}\min\left(1, \frac{z_t^{\intercal}\Sigma_t z_t}{\sigma_w^2}\right) \tag{44}
$$

Using (9) and the intermediate step of the proof of [7, Lemma 6], we have

$$\sum_{t=1}^{T} \min\left(1, \frac{z_t^\intercal \Sigma_t z_t}{\sigma_w^2}\right) = \sum_{t=1}^{T} \min\left(1, \left\|\frac{\Sigma_t^{0.5} z_t z_t^\intercal \Sigma_t^{0.5}}{\sigma_w^2}\right\|\right)$$
$$\leq 2(n+m)\log\left(\frac{\mathrm{Tr}(\Sigma_{T+1}^{-1})}{(n+m)}\right) - \log\det\Sigma_1^{-1}. \quad (45)$$

Now, from (9), we get that

$$\mathrm{Tr}(\Sigma_{T+1}^{-1}) = \mathrm{Tr}(\Sigma_1^{-1}) + \sum_{t=1}^{T} \frac{1}{\sigma_w^2}\mathrm{Tr}(z_t z_t^\intercal)$$
$$\leq \mathrm{Tr}(\Sigma_1^{-1}) + \frac{T}{\sigma_w^2} M_G^2 X_T^2, \quad (46)$$

where the last inequality uses the fact that $\mathrm{Tr}(z_t z_t^\intercal) = \mathrm{Tr}(z_t^\intercal z_t) = \|z_t\|^2 \leq M_G^2 X_T^2$. Combining (44) with (45) and (46), we get

$$\sum_{t=1}^{T} z_t^\intercal \Sigma_t z_t \leq \mathcal{O}\big((n+m)(\sigma_w^2 + X_T^2)\log(TX_T^2)\big). \quad (47)$$

Therefore, we can bound the expectation of the right hand side of (42) as

$$\mathbb{E}\left[\left(2 + \frac{M_G^2 X_T^2}{\lambda_{\min}\sigma_w^2}\right)^{T_{\min}\vee 1} X_T^2 \sum_{t=1}^{T} z_t^\intercal \Sigma_t z_t\right]$$
$$\leq \mathcal{O}\big(\sigma_w^4 (n+m)\mathbb{E}[F(X_T)]\big)$$
$$\leq \tilde{\mathcal{O}}(\sigma_w^4 (n+m)), \quad (48)$$

where the first inequality follows from (47) with $F(X_T) = \big(2 + \frac{M_G^2 X_T^2}{\lambda_{\min}\sigma_w^2}\big)^{T_{\min}\vee 1}(\frac{X_T^2}{\sigma_w^2} + \frac{X_T^4}{\sigma_w^4})\log(TX_T^2)$, 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 log term.

The result follows from (42) and (48).

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. 6, pp. 3600–3607, 2020.

[2] ——, "Control of unknown linear systems with Thompson sampling," in *Annual Allerton Conference on Communication, Control, and Computing*, 2017, pp. 1198–1205.

[3] Y. Ouyang, M. Gagrani, A. Nayyar, and R. Jain, "Learning unknown Markov decision processes: A Thompson sampling approach," in *Advances in Neural Information Processing Systems*, 2017, pp. 1333–1342.

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

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

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

[7] ——, "Bayesian optimal control of smoothly parameterized systems: The lazy posterior sampling algorithm," 2014, arXiv preprint arXiv:1406.3926.