Approximate Thompson Sampling for Learning Linear Quadratic Regulators with $O(\sqrt{T})$ Regret

Yeoneung Kim KIMYEONEUNG@GMAIL.COM
Department of Applied Artificial Intelligence, SeoulTech, Seoul 01811, Korea

Gihun Kim HOON2680@SNU.AC.KR
Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, Korea

Jiwhan Park JIWHANPARK@SNU.AC.KR
Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, Korea

Insoon Yang INSOONYANG@SNU.AC.KR
Department of Electrical and Computer Engineering, ASRI, Seoul National University, Seoul 08826, Korea

Editors: N. Ozay, L. Balzano, D. Panagou, A. Abate

Abstract

We propose a novel Thompson sampling algorithm that learns linear quadratic regulators (LQR) with a Bayesian regret bound of $O(\sqrt{T})$. Our method leverages Langevin dynamics with a carefully designed preconditioner and incorporates a simple excitation mechanism. We show that the excitation signal drives the minimum eigenvalue of the preconditioner to grow over time, thereby accelerating the approximate posterior sampling process. Furthermore, we establish nontrivial concentration properties of the approximate posteriors generated by our algorithm. These properties enable us to bound the moments of the system state and attain an $O(\sqrt{T})$ regret bound without relying on the restrictive assumptions that are often used in the literature.

Keywords: Online learning, reinforcement learning, linear systems, MCMC.

1. Introduction

Balancing the exploration–exploitation trade-off is a fundamental challenge in reinforcement learning (RL). This dilemma has been systematically addressed by two principal approaches: *optimism in the face of uncertainty* (OFU) and *Thompson sampling* (TS). OFU-based methods construct confidence sets for the environment or model parameters using the data observed thus far. An optimistic or reward-maximizing set of parameters is then selected from within this confidence set, and a corresponding optimal policy is executed (Lai and Robbins, 1985). Algorithms based on OFU have been shown to provide strong theoretical guarantees, particularly in the context of bandit problems (Kearns and Singh, 2002). On the other hand, TS is a Bayesian method in which the environment or model parameters are sampled from a posterior distribution that is updated over time using observed data and a prior (Thompson, 1933). An optimal policy with respect to the sampled parameters is then constructed and executed. TS is often more computationally tractable than OFU, as OFU typically requires solving a nonconvex optimization problem over a confidence set in each episode. TS has demonstrated effectiveness in online learning across a wide range of sequential decision-making problems, including multi-armed bandits (Agrawal and Goyal, 2012, 2013; Kaufmann et al., 2012), Markov decision processes (Osband et al., 2013; Osband and Van Roy, 2016; Gopalan and Mannor, 2015), and LQR problems (Ouyang et al., 2019; Osband and Van Roy, 2016; Abbasi-Yadkori and Szepesvári, 2015; Abeille and Lazaric, 2017; Faradonbeh et al., 2020a).

In TS-based online learning, posterior sampling becomes challenging in high-dimensional settings. It is also computationally intractable when the posterior distribution lacks a closed-form expression, which occurs when the noise and prior distributions are not conjugate. To address this, Markov Chain Monte Carlo (MCMC) methods—particularly Langevin MCMC—have been proposed (Gilks et al., 1995; Roberts and Tweedie, 1996; Durmus and Moulines, 2016; Welling and Teh, 2011). Nevertheless, Langevin MCMC is computationally intensive. To mitigate this issue, various acceleration techniques have been studied (see (Welling and Teh, 2011; Li et al., 2019; Mou et al., 2019; Ding et al., 2021; Lu et al., 2019) and references therein). In particular, preconditioning has been shown to be effective for improving sampling efficiency (Welling and Teh, 2011; Girolami and Calderhead, 2011; Dalalyan, 2017; Dwivedi et al., 2018). Motivated by these findings, we incorporate preconditioned Langevin MCMC into TS for LQR problems.

In this paper, we propose a computationally efficient approximate Thompson sampling algorithm for learning linear quadratic regulators (LQR) with a Bayesian regret bound of $O(\sqrt{T})$. Our algorithm is based on carefully designed Langevin dynamics that achieve an improved convergence rate. The regret analysis is conducted under the assumption that the system noise follows a strongly log-concave distribution—a relaxation of the Gaussian noise assumption commonly adopted in prior works. To the best of our knowledge, our method achieves the tightest known Bayesian regret bound for online LQR learning, improving upon the existing $\tilde{O}(\sqrt{T})$ bounds[1] in the literature (Abeille and Lazaric, 2018; Ouyang et al., 2019; Gagrani et al., 2022).

It is worth noting that in (Ouyang et al., 2019; Gagrani et al., 2022), the system noise is assumed to follow independent and identically distributed Gaussian. Moreover, the columns of the system parameter matrix are assumed to be mutually independent and Gaussian in the prior, which is key to both the tractability of their regret analysis and the simplification of posterior updates. In contrast, our work not only achieves a tighter regret bound but also relaxes these restrictive assumptions. While we adopt the assumption on system parameters from (Abeille and Lazaric, 2018), we go beyond their analysis by establishing a regret bound that holds for multi-dimensional systems.

The two key components of our method are: (i) a preconditioned unadjusted Langevin algorithm (ULA) for approximate Thompson sampling, and (ii) a simple excitation mechanism. The proposed excitation mechanism injects a noise signal into the control input at the end of each episode, which causes the minimum eigenvalue of the preconditioner to increase over time, thereby accelerating the posterior sampling process. We identify appropriate step sizes and iteration counts for the preconditioned Langevin MCMC and demonstrate both an accelerated convergence rate for approximate Thompson sampling and improved learning performance. Specifically, we show that the sampled system parameters converge to the true parameters at a rate of $\tilde{O}(t^{-\frac{1}{4}})$. This improvement yields a tighter bound on the system state norm, which in turn contributes to achieving the improved regret bound of $O(\sqrt{T})$. Due to space limitations, the proofs of our theoretical results are provided in Appendix A of the extended version (Kim et al., 2024).

2. Related Work

Certainty Equivalence (CE): The certainty equivalence principle (Landau et al., 1998) has been widely adopted for learning dynamical systems with unknown transitions, wherein the optimal policy is designed under the assumption that the estimated system parameters accurately represent the true parameters. The performance of CE-based methods has been extensively studied across various settings, including online learning (Simchowitz and Foster, 2020; Dean et al., 2018; Mania

[1] Here, $\tilde{O}(\cdot)$ hides logarithmic factors.

et al., 2019; Jedra and Proutiere, 2022), sample complexity analysis (Dean et al., 2020), finite-time stabilization (Faradonbeh et al., 2018), and asymptotic regret bounds (Faradonbeh et al., 2020a).

Optimism in the Face of Uncertainty (OFU): Abbasi-Yadkori and Szepesvári (2011); Ibrahimi et al. (2012) proposed OFU-based learning algorithms that iteratively select high-performing control actions while constructing confidence sets. These methods achieve a frequentist regret bound of $\tilde{O}(\sqrt{T})$, but are often computationally impractical due to the complexity of the resulting constraints. To address this issue, subsequent works (Cohen et al., 2019; Abeille and Lazaric, 2020) translated the nonconvex optimization problem inherent in OFU into a semidefinite programming (SDP) formulation, attaining the same $\tilde{O}(\sqrt{T})$ regret bound with high probability. Alternatively, Faradonbeh et al. (2020a,b) introduced randomized control actions to avoid constructing confidence sets, while still achieving an asymptotic regret bound of $\tilde{O}(\sqrt{T})$. More recently, Lale et al. (2022) proposed an algorithm that rapidly stabilizes the system and attains a $\tilde{O}(\sqrt{T})$ frequentist regret bound without requiring a stabilizing control gain matrix.

Thompson Sampling (TS): It has been shown that the upper bound for the frequentist regret under Gaussian noise can be as large as $\tilde{O}(T^{2/3})$ (Abeille and Lazaric, 2017), which was later improved to $\tilde{O}(\sqrt{T})$ in (Abeille and Lazaric, 2018) using a TS-based approach; however, this result is limited to *scalar* systems. Subsequently, Kargin et al. (2022) extended the analysis to multidimensional systems, achieving a $\tilde{O}(\sqrt{T})$ frequentist regret bound. Nonetheless, the Gaussian noise assumption remains essential for establishing these guarantees. For Bayesian regret, prior results (Ouyang et al., 2019; Gagrani et al., 2022) demonstrate the potential of TS-based algorithms to achieve a $\tilde{O}(\sqrt{T})$ Bayesian regret bound. However, these methods are subject to several limitations. Specifically, both the noise and the prior distribution over system parameters are assumed to be Gaussian, ensuring conjugacy between the prior and posterior. Additionally, the columns of the system parameter matrix are assumed to be mutually independent.

Comparison with Mazumdar et al. (2020): Our work builds on the ideas introduced in Mazumdar et al. (2020), which focuses on multi-armed bandits. However, key differences arise due to the fundamentally different nature of LQR problems. For example, in the bandit setting, the strong log-concavity of the reward function ensures linear growth of the likelihood function as more data is collected. This property plays a crucial role in their analysis. In contrast, such growth does not occur in LQR problems, prompting us to introduce an adaptive preconditioner to improve computational efficiency. Moreover, the Lipschitz smoothness of the log-reward function in Mazumdar et al. (2020) facilitates the analysis of the gap between exact and approximate posteriors—a simplification that does not hold in the LQR setting.

3. Preliminaries

3.1. Linear-quadratic regulators

Consider a linear stochastic system of the form

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad t = 1, 2, \ldots, \tag{1}$$

where $x_t \in \mathbb{R}^n$ is the system input, and $u_t \in \mathbb{R}^{n_u}$ is the control input. The disturbance $w_t \in \mathbb{R}^n$ is an independent and identically distributed (i.i.d.) zero-mean random vector with covariance matrix $\mathbf{W}$. Let I_n denote the n by n identity matrix and let $|v|_P := \sqrt{v^\top P v}$ be the weighted 2-norm of a vector v with respect to a positive semidefinite matrix P, and $|v|$ indicate the Euclidean norm.

Assumption 1 *For every $t = 1, 2, \ldots$, the random vector w_t satisfies the following properties:*

1. *The probability density function (pdf) of the noise $p_w(\cdot)$ is known and twice differentiable. Additionally, $\underline{m}I_n \preceq -\nabla^2 \log p_w(\cdot) \preceq \overline{m}I_n$. for some $\underline{m}, \overline{m} > 0$.*[2]
2. $\mathbb{E}[w_t] = 0$ *and* $\mathbb{E}[w_t w_t^\top] = \mathbf{W}$, *where* $\mathbf{W}$ *is positive definite.*

Our paper deals with a broader class of disturbances compared to existing methods (Abeille and Lazaric, 2018; Ouyang et al., 2019; Gagrani et al., 2022), as any multivariate Gaussian distribution satisfies the assumption.

Let $d := n + n_u$ and Θ be the system parameter matrix defined by $\Theta := \begin{bmatrix}\Theta(1) & \cdots & \Theta(n)\end{bmatrix} := \begin{bmatrix}A & B\end{bmatrix}^\top \in \mathbb{R}^{d\times n}$, where $\Theta(i) \in \mathbb{R}^d$ is the ith column of Θ. We also let $\theta := \text{vec}(\Theta) := (\Theta(1), \Theta(2), \dots, \Theta(n)) \in \mathbb{R}^{dn}$ denote the vectorized version of Θ. We often refer to θ as the parameter vector. Let $h_t := (x_1, u_1, \dots, x_{t-1}, u_{t-1}, x_t)$ be the *history* of observations made up to time t, and let H_t denote the collection of such histories at stage t. A (deterministic) policy π_t maps history h_t to action u_t, i.e., $\pi_t(h_t) = u_t$. The set of admissible policies is defined as $\Pi := \{\pi = (\pi_1, \pi_2, \dots) \mid \pi_t : H_t \to \mathbb{R}^{n_u} \text{ is measurable } \forall t\}$.

The stage-wise cost is chosen to be a quadratic function of the form $c(x_t, u_t) := x_t^\top Q x_t + u_t^\top R u_t$, where $Q \in \mathbb{R}^{n\times n}$ is symmetric positive semidefinite and $R \in \mathbb{R}^{n_u \times n_u}$ is symmetric positive definite. The cost matrices Q and R are assumed to be known. [3] We consider the infinite-horizon average cost LQ setting with the following cost function: $J_\pi(\theta) := \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_\pi\big[\sum_{t=1}^T c(x_t, u_t)\big]$. Given $\theta \in \mathbb{R}^{dn}$, $\pi_*(x;\theta)$ denotes an optimal policy if it exists, and the corresponding optimal cost is given by $J(\theta) = \inf_{\pi\in\Pi} J_\pi(\theta)$. Assuming (A, B) is stabilizable, and $(A, Q^{1/2})$ is observable, the following algebraic Riccati equation (ARE) has a unique positive definite solution $P^*(\theta)$:

$$P^*(\theta) = Q + A^\top P^*(\theta)A - A^\top P^*(\theta)B(R + B^\top P^*(\theta)B)^{-1}B^\top P^*(\theta)A. \tag{2}$$

Furthermore, $J(\theta) = \text{tr}(\mathbf{W}P^*(\theta))$, which is continuously differentiable with respect to θ, and the optimal policy is uniquely obtained as $\pi_*(x;\theta) = K(\theta)x$, where the control gain matrix $K(\theta)$ is given by $K(\theta) := -(R + B^\top P^*(\theta)B)^{-1}B^\top P^*(\theta)A$.

3.2. Online learning of LQR

The theory of LQR is applicable when the true system parameters $\theta_* := \text{vec}(\Theta_*) := \text{vec}(\begin{bmatrix}A_* & B_*\end{bmatrix}^\top)$ are fully known and stabilizable. However, we consider the case where the true parameter vector θ_* is unknown. Online learning is a popular approach to addressing this case. The performance of an online learning algorithm is typically measured by regret. In particular, we consider the Bayesian setting, where the prior distribution p_1 of the true system parameter random variable $\bar{\theta}_*$ is assumed to be given, and define the Bayesian regret over T stages as:

$$R(T) := \mathbb{E}\Bigg[\sum_{t=1}^{T}\big(c(x_t, u_t) - J(\bar{\theta}_*)\big)\Bigg]. \tag{3}$$

The expectation is taken with respect to the distributions of the system noise $(w_1, w_2, \dots, w_T)$, the internal randomness of the learning algorithm, and the prior distribution.

[2] The density of a multivariate normal distribution whose covariance Σ lies between $\underline{m}$ and $\overline{m}$ satisfies this assumption.

[3] This assumption is common in the literature (Abbasi-Yadkori and Szepesvári, 2011; Faradonbeh et al., 2020a; Abeille and Lazaric, 2020; Jedra and Proutiere, 2022; Kargin et al., 2022; Dean et al., 2020).

3.3. The unadjusted Langevin algorithm (ULA)

Consider the problem of sampling from a probability distribution with density $p(x) \propto e^{-U(x)}$, where the potential $U : \mathbb{R}^{n_x} \to \mathbb{R}$ is twice differentiable. The Langevin dynamics take the form

$$dX_\tau = -\nabla U(X_\tau)\, d\tau + \sqrt{2}\, dB_\tau, \tag{4}$$

where B_τ is standard Brownian motion in $\mathbb{R}^{n_x}$. It is known that the density of X_ξ converges to $p(x)$ as $\xi \to \infty$ (Mou et al., 2019; Pavliotis, 2014). To approximate X_τ, we apply the Euler–Maruyama discretization, yielding the *unadjusted Langevin algorithm* (ULA):

$$X_{j+1} = X_j - \gamma_j \nabla U(X_j) + \sqrt{2\gamma_j} W_j, \tag{5}$$

where $(W_j)_{j\geq 1}$ are i.i.d. standard Gaussian random vectors, and $(\gamma_j)_{j\geq 1}$ are step sizes. While Metropolis–Hastings corrections are often used to mitigate discretization error (Roberts and Tweedie, 1996; Bou-Rabee and Hairer, 2013), small step sizes can eliminate the need for such adjustments. In this work, we propose adaptive step sizes and iteration counts that ensure improved concentration properties, as discussed in Section 4.2.

The condition number of the Hessian of the potential is a key factor in determining the rate of convergence. More precisely, the following concentration property of ULA holds, which is a modification of Theorem 5 in Mazumdar et al. (2020).

Theorem 2 *Suppose that the pdf $p(x) \propto e^{-U(x)}$ is strongly log-concave and $\lambda_{\min} I \preceq \nabla^2 U(x) \preceq \lambda_{\max} I$ for all x, where $\lambda_{\max}, \lambda_{\min} > 0$. Let the stepsize be given by $\gamma_j \equiv \gamma = O\big(\frac{\lambda_{\min}}{\lambda_{\max}^2}\big)$ and the number of iterations N satisfy $N = \Omega\big((\frac{\lambda_{\max}}{\lambda_{\min}})^2\big)$.[4] Given $X_0 \in \arg\min U(x)$, let p_N denote the pdf of X_N obtained by iterating (5). Then, $\mathbb{E}_{x\sim p, \tilde{x}\sim p_N}\big[|x - \tilde{x}|^2\big]^{\frac{1}{2}} \leq O\Big(\sqrt{\frac{1}{\lambda_{\min}}}\Big)$, where $x = x_{\gamma N}$ is a solution to (4) with $X_0 \sim e^{-U(x)}$ and the joint probability distribution of $x \sim p$ and $\tilde{x} \sim p_N$ is obtained via the shared Brownian motion.*

4. Online Learning Algorithm

The naive TS approach for learning LQR has two main weaknesses. The first arises from the potential selection of a destabilizing controller, which can cause the system state to grow exponentially and lead to unbounded regret. To address this issue, we control the probability of the state exhibiting excessively large norms. The second weakness stems from inefficiencies in the sampling process when the system noise and prior distribution are not conjugate. In such cases, ULA offers an alternative for posterior approximation, but it is often extremely slow. To accelerate the sampling process, we introduce a preconditioning technique.

4.1. Preconditioned ULA for approximate posterior sampling

One of the key components of our learning algorithm is approximate posterior sampling via preconditioned Langevin dynamics. The potential in ULA is chosen as $U_t(\theta) := -\log p(\theta|h_t)$, where $p(\theta|h_t)$ denotes the posterior distribution of the true system parameter given the history up to t. Unfortunately, a direct implementation of ULA to TS for LQR is inefficient as it requires a large number of iterations. To accelerate the convergence of Langevin dynamics, we propose a preconditioning technique.[5]

[4] $a_n = O(b_n)$ means $\limsup_{n\to\infty} |a_n/b_n| < \infty$, and $a_n = \Omega(b_n)$ indicates $\liminf_{n\to\infty} |a_n/b_n| > 0$.

[5] Preconditioning techniques have been used for Langevin algorithms in different contexts; see, e.g., Li et al. (2016); Lu et al. (2020); Bras (2022).

To describe the preconditioned Langevin dynamics, we choose a positive definite matrix P, referred to as a *preconditioner*. The change of variables $\theta' = P^{\frac{1}{2}}\theta$ yields $\mathrm{d}\theta_\tau = -P^{-1}\nabla U_t(\theta_\tau)\,\mathrm{d}\tau + \sqrt{2P^{-1}}\,\mathrm{d}B_\tau$. Applying the Euler–Maruyama discretization with constant stepsize γ yields the preconditioned ULA:

$$\theta_{j+1} = \theta_j - \gamma P^{-1}\nabla U_t(\theta_j) + \sqrt{2\gamma P^{-1}}\,W_j, \tag{6}$$

where $(W_j)_{j\geq 1}$ is an i.i.d. sequence of standard dn-dimensional Gaussian random vectors.

Given the data $z_t = (x_t, u_t)$ collected, the preconditioner in our setting is defined as[6]

$$P_t := \lambda I_{dn} + \sum_{s=1}^{t-1} \operatorname{blkdiag}\{z_s z_s^\top\}_{i=1}^n, \tag{7}$$

where $\operatorname{blkdiag}\{A_i\}_{i=1}^n \in \mathbb{R}^{dn\times dn}$ denotes the block diagonal matrix of the A_i, and $\lambda > 0$ is a constant determined by the prior. Then, the following lemma holds:

Lemma 3 *Suppose Assumption 1 holds and the potential of the prior satisfies* $\nabla_\theta^2 U_1(\cdot) = \lambda I_{dn}$ *for some* $\lambda > 0$. *Then, for all* θ *and* t, *we have* $mI_{dn} \preceq P_t^{-\frac{1}{2}}\nabla^2 U_t(\theta)P_t^{-\frac{1}{2}} \preceq MI_{dn}$, *where* $m = \min\{\underline{m}, 1\}$ *and* $M = \max\{\overline{m}, 1\}$.

It follows from Lemma 3 and Theorem 2 that we can rescale the number of iterations required for the convergence of ULA while ensuring improved accuracy in the concentration of the sampled system parameter. In fact, we show later that the number of required iterations scales only with n. To demonstrate the effect of preconditioning, note that Lemma 3 implies $m\lambda_{\min}(P_t)I_{dn} \preceq \nabla^2 U_t \preceq M\lambda_{\max}(P_t)I_{dn}$. Theorem 2 then implies that $O\big((\lambda_{\max}(P_t)/\lambda_{\min}(P_t))^2\big)$ iterations are needed to achieve an error bound of $O\big(1/\sqrt{\lambda_{\min}(P_t)}\big)$. Our algorithm improves this bound to $O\big(1/\sqrt{\max\{\lambda_{\min}(P_t), t\}}\big)$. Throughout the paper, we use the notation $\mathbf{U}_k := U_{t_k}$ to explicitly indicate the dependence on the current episode k.

4.2. Algorithm

We begin by introducing the following log-concavity condition on the prior, centered arbitrarily. This condition is a slight relaxation of the assumption in Ouyang et al. (2019).

Assumption 4 *The prior* p_1 *satisfies* $\nabla_\theta^2 U_1(\cdot) = \lambda I_{dn}$ *for* $U_1(\cdot) := -\log p_1(\cdot)$ *and some* $\lambda \geq 1$.

The initialization of the preconditioner P_t plays a crucial role in the efficiency of the sampling process. If P_0 is too small, the algorithm may suffer from slow exploration due to small step sizes in the Langevin dynamics. Conversely, if P_0 is too large, the algorithm may place excessive trust in the prior, potentially slowing adaptation to the true system parameters. Our choice of $P_0 = \lambda I$ with a moderate λ ensures a balance between these effects. For mathematical convenience, it suffices to set $\lambda > 0$, but we assume $\lambda \geq 1$ to simplify the analysis.

Following Abeille and Lazaric (2018), we consider an admissible set of parameters defined as $\mathcal{C} := \{\theta \in \mathbb{R}^{dn} : |\theta| \leq S,\ |A + BK(\theta)| \leq \rho < 1,\ J(\theta) \leq M_J\}$ for some constants $S, \rho, M_J > 0$, where $\theta = \operatorname{vec}(\begin{bmatrix} A & B\end{bmatrix}^\top)$. To sample from the posterior distribution, we restrict the sample to lie within $\mathcal{C}$ via rejection sampling. This ensures that for any sampled system parameter $\theta \in \mathcal{C}$, there exists a positive constant M_{P^*} such that $|P^*(\theta)| \leq M_{P^*}$ (Abeille and Lazaric, 2017). Consequently, $|[I \quad K(\theta)^\top]| \leq M_K$ for some $M_K > 1$, and therefore, $|A_* + B_*K(\theta)| \leq M_\rho$ for some $M_\rho \geq 1$.

[6] Our preconditioner can be viewed as an adaptive scaling mechanism, analogous to the Fisher information matrix in natural policy gradient methods. This connection arises because the empirical covariance matrix captures the local curvature of the posterior distribution, effectively conditioning the Langevin dynamics for more efficient sampling.

Algorithm 1 Thompson sampling with Langevin dynamics for LQR

```
Input: $p_1$;
Initialization: $t \leftarrow 1$, $t_0 \leftarrow 0$, $x_1 \leftarrow 0$, $\mathcal{D} \leftarrow \emptyset$, $\mathbf{U}_0 \leftarrow U_1$, $\tilde{\theta}_0 \leftarrow \arg\min U_1(\theta)$, $\theta_{\min,0} \leftarrow \tilde{\theta}_0$;
for Episode $k = 1, 2, \dots$ do
    $T_k \leftarrow k + 1$, and $t_k \leftarrow t$;
    $\mathbf{U}_k(\cdot) \leftarrow \mathbf{U}_{k-1}(\cdot) - \sum_{(z_t, x_{t+1}) \in \mathcal{D}} \log p_w(x_{t+1} - \Theta^\top z_t)$;
    $\theta_{\min,k} \in \arg\min \mathbf{U}_k(\theta)$ and $\mathcal{D} \leftarrow \emptyset$;
    Compute the preconditioner $\tilde{P}_k$, the step size $\tilde{\gamma}_k$, and the number of iterations $\tilde{N}_k$ as (8);
    while True do
        $\theta_0 \leftarrow \theta_{\min,k}$;
        for Step $j = 0, 1, \dots, \tilde{N}_k - 1$ do
            Sample $\theta_{j+1} \sim \mathcal{N}\big(\theta_j - \tilde{\gamma}_k \tilde{P}_k^{-1} \nabla \mathbf{U}_k(\theta_j),\, 2\tilde{\gamma}_k \tilde{P}_k^{-1}\big)$;
        end
        if $\theta_{\tilde{N}_k} \in \mathcal{C}$ then
            $\tilde{\theta}_k \leftarrow \theta_{\tilde{N}_k}$;
            break;
        end
    end
    Compute the gain matrix $K_k \leftarrow K(\tilde{\theta}_k)$;
    while $t \leq t_k + T_k - 1$ do
        Execute control $u_t \leftarrow K_k x_t + \nu_t$ for $\nu_t$ satisfying Assumption 5;
        Observe new state $x_{t+1}$, and update $\mathcal{D} \leftarrow \mathcal{D} \cup \{(z_t, x_{t+1})\}$; $t \leftarrow t + 1$;
    end
end
```

Our proposed algorithm is presented in Algorithm 1. We employ dynamic episode scheduling, as it has been shown to be effective in the literature Abbasi-Yadkori and Szepesvári (2011); Ouyang et al. (2019); Abeille and Lazaric (2017). In the algorithm, t_k and T_k denote the start time and the length of episode k, respectively. By definition, $t_1 = 1$ and $t_{k+1} = t_k + T_k$. The episode length is chosen as $T_k = k + 1$. To update the posterior—or equivalently, the potential—at episode k, we use the dataset $\mathcal{D} := \{(z_t, x_{t+1})\}_{t_{k-1} \leq t \leq t_k - 1}$ collected during the previous episode. Approximate TS is then performed using the preconditioned ULA with the preconditioner, step size, and number of iterations chosen as $\tilde{P}_k := P_{t_k}$, $\tilde{\gamma}_k := \gamma_{t_k}$, and $\tilde{N}_k := \max(1, \lceil N_{t_k} \rceil)$, where

$$P_t := \lambda I_{dn} + \sum_{s=1}^{t-1} \text{blkdiag}\{z_s z_s^\top\}_{i=1}^n, \; \gamma_t := \frac{m\lambda_{\min,t}}{16M^2 \max\{\lambda_{\min,t}, t\}}, \; N_t := \frac{4\log_2\left(\frac{\max\{\lambda_{\min,t}, t\}}{\lambda_{\min,t}}\right)}{m\gamma_t}. \tag{8}$$

Here, $\lambda_{\min,t}$ and $\lambda_{\max,t}$ denote the minimum and maximum eigenvalues of P_t. This choice is based on a detailed analysis of the concentration properties of ULA, as established in Proposition 6. The additional operations on N_{t_k} ensure $\tilde{N}_k \in \mathbb{N}$, avoiding the possibility of infinite rejection when $\tilde{N}_k = 0$. After performing the preconditioned ULA update $\tilde{N}_k$ times, we check whether $\theta_{\tilde{N}_k} \in \mathcal{C}$. If so, the sampled parameter is accepted, and the corresponding control gain matrix is computed via ARE (2). To ensure that the rejection step ends in a finite number of iterations, we assume that there exists a small positive constant ϵ such that, for each episode k, $\Pr(\tilde{\theta}_k \in \mathcal{C}) \geq 1 - \epsilon$ under the posterior distribution. Although this assumption may appear restrictive, it has been empirically validated in all of our examples, as shown in Appendix C of the extended version (Kim et al., 2024).

A novel component of our algorithm is the injection of a noise signal into the control input u_t at the end of each episode. The external noise signal is assumed to satisfy the following:

Assumption 5 *The random variable $\nu_s \in \mathbb{R}^{n_u}$ is $\bar{L}_\nu$-sub-Gaussian,*[7] *and satisfies $\nu_s = 0$ if $s \in [t_j, t_{j+1} - 2]$ for $j \geq 2$. Moreover, $\mathbb{E}[\nu_s] = 0$, and $\mathbf{W}' := \mathbb{E}[\nu_s \nu_s^\top]$ is a positive definite matrix whose maximum and minimum eigenvalues are identical to those of $\mathbf{W}$.*[8]

Since our algorithm does not rely on a predefined stabilizing set of parameters, one may be concerned that the control policies generated during the early learning phase could exhibit instability due to limited data. To address this issue, our excitation mechanism ensures that the preconditioner matrix grows over time, thereby improving the concentration properties of the sampled system parameters, as shown in the following section.

5. Concentration Properties

To show that Algorithm 1 achieves an $O(\sqrt{T})$ regret bound, we first examine the concentration properties of the exact and approximate posterior distributions given the history up to a fixed time t, for the potential $U_t(\theta) = U_1(\theta) - \sum_{s=1}^{t-1} \log p_w(x_{s+1} - \Theta^\top z_s)$. When t is chosen as t_k, we recover the case corresponding to Algorithm 1. The concentration results established in this section enable us to bound the moments of the system state, which is essential for attaining the desired regret bound.

5.1. Comparing exact and approximate posteriors

Let μ_t denote the *exact posterior* distribution defined by $\mu_t \propto \exp(-U_t)$.[9] For the approximate posterior, recall the preconditioned ULA that generates $\theta_{j+1} \sim \mathcal{N}\left(\theta_j - \gamma_t P_t^{-1} \nabla U_t(\theta_j),\, 2\gamma_t P_t^{-1}\right)$ starting from $\theta_0 \in \arg\min U_t(\cdot)$. After repeating this update for N_t steps, we obtain θ_{N_t}. We let $\tilde{\mu}_t$ denote the *approximate posterior*, defined as the distribution of θ_{N_t}. The joint distribution between $\theta_t \sim \mu_t$ and $\tilde{\theta}_t \sim \tilde{\mu}_t$ is characterized via a shared Brownian path driving both the continuous Langevin diffusion and the discrete ULA dynamics with the preconditioner (see Remark 2.3 and Appendix A.1 of the extended version (Kim et al., 2024)).

Proposition 6 *Suppose Assumptions 1 and 4 hold. Then, the exact posterior μ_t and the approximate posterior $\tilde{\mu}_t$ obtained via preconditioned ULA satisfy $\mathbb{E}_{\theta_t \sim \mu_t, \tilde{\theta}_t \sim \tilde{\mu}_t}\left[|\theta_t - \tilde{\theta}_t|_{P_t}^p \mid h_t\right] \leq D_p$ for all $p \geq 2$, where $D_p = \left(\frac{pdn}{m}\right)^{\frac{p}{2}}\left(2^{2p+1} + 5^p\right)$. When $p = 2$, we further have*

$$\mathbb{E}_{\theta_t \sim \mu_t, \tilde{\theta}_t \sim \tilde{\mu}_t}\left[|\theta_t - \tilde{\theta}_t|^2 \mid h_t\right]^{\frac{1}{2}} \leq \sqrt{\frac{D}{\max\{\lambda_{\min,t},\, t\}}}, \tag{9}$$

where $D = 114\frac{dn}{m}$ and $\lambda_{\min,t}$ denotes the minimum eigenvalue of P_t.

This result shows that the use of our preconditioner improves the convergence rate from $O(1/\sqrt{\lambda_{\min,t}})$ to $O(1/\sqrt{\max\{\lambda_{\min,t},\, t\}})$ (see Theorem 2).

[7] A distribution is L_ν-sub-Gaussian if $\Pr(|\nu| > y) < C \exp(-\frac{1}{2L_\nu^2} y^2)$ for some $C > 0$.

[8] The assumption on the minimum eigenvalue of $\mathbf{W}'$ is made for simplicity in the proof of Proposition 9, which concerns the growth of $\lambda_{\min}(P_t)$.

[9] Throughout this subsection, in the definition of the potential U_t, we let $(z_s)_{s\geq 1}$ be an $\mathbb{R}^d$-valued stochastic process adapted to a filtration $(\mathcal{F}_t)_{t\geq 0}$, where each z_s is assumed to be $\mathcal{F}_{s-1}$-measurable for all $s \geq 1$.

Another important observation is a concentration bound for the exact posterior. This concentration property is essential for characterizing a confidence set used in the proof of Theorem 8.

Proposition 7 *Suppose Assumptions 1 and 4 hold. Then, the following inequality*

$$\mathbb{E}_{\theta_t \sim \mu_t}\left[|\theta_t - \theta_*|_{P_t}^p \mid h_t\right]^{\frac{1}{p}} \leq 2p\sqrt{\frac{8nM^2}{m^3}\log\left(\frac{n}{\delta}\left(\frac{\lambda_{\max,t}}{\lambda}\right)^{\frac{d}{2}}\right)} + C, \quad t > 0 \tag{10}$$

holds with probability at least $1-\delta$ for any $0 < \delta < 1$ and $p \geq 2$, where the constant $C > 0$ depends only on p, m, n, d, and λ, and $\lambda_{\max,t}$ denotes the maximum eigenvalue of P_t.[10]

5.2. Bounding expected state norms by a polynomial of time

A key result we derive from Propositions 6 and 7 is that the system state grows at most polynomially in expectation over time. To show this property, we modify the confidence set construction and self-normalization technique developed for the OFU approach (Abbasi-Yadkori et al., 2011; Abbasi-Yadkori and Szepesvári, 2011). Our key idea is to construct a set that contains the system parameters sampled via ULA with high probability. The higher-moment bounds from Propositions 6 and 7 are crucial to our analysis as Markov-type inequalities can be exploited for any p. We then partition the probability space of the stochastic process into two sets, "good" and "bad," as in the OFU approach.

Theorem 8 *Suppose Assumptions 1, 4, and 5 hold. For $T > 0$, $p \geq 2$, and a random trajectory $(x_s)_{s=1}^T$ generated by Algorithm 1, we have $\mathbb{E}\left[\max_{j \leq t} |x_j|^p\right] \leq Ct^{\frac{7}{2}p(d+1)}$ for $t \geq 1$, where the constant $C > 0$ depends only on p, m, n, n_u, $\mathbf{W}$, M_ρ, and λ.*

5.3. Concentration of exact and approximate posteriors

We begin by characterizing the growth of the minimum eigenvalue of the preconditioner, which results from injecting a random noise signal ν_s to perturb the action at the end of each episode. To derive this result, we decompose the preconditioner in each episode into two parts—a random matrix and a self-normalized matrix-valued process—as in Jedra and Proutiere (2022):

$$\sum z_s z_s^\top = \underbrace{\sum (L_s\psi_s)(L_s\psi_s)^\top}_{\text{random matrix part}} - \underbrace{\left(\sum y_s (L_s\psi_s)^\top\right)^\top \left(\sum y_s y_s^\top + I_d\right)^{-1} \left(\sum y_s (L_s\psi_s)^\top\right)}_{\text{self-normalization}} - I_d,$$

where $y_s := \begin{bmatrix} A_* x_{s-1} + B_* u_{s-1} \\ K_j(A_* x_{s-1} + B_* u_{s-1}) \end{bmatrix}$, $L_s := \begin{bmatrix} I_n & 0 \\ K_j & I_{n'} \end{bmatrix}$, $\psi_s := \begin{bmatrix} w_{s-1} \\ \nu_s \end{bmatrix}$, and K_j is the control gain matrix used in the jth episode. The random matrix part contributes to the growth of the minimum eigenvalue of the preconditioner with high probability. By Theorem 8, the self-normalization term is bounded by $O(\log T)$ with high probability. More precisely, the following proposition holds:

Proposition 9 *Suppose Assumptions 1–5 hold. For $k \geq k_0(m, n, n_u, \lambda, M_K, M_\rho, \mathbf{W})$, we have $\mathbb{E}\left[\lambda_{\min,t_{k+1}}^{-p}\right] \leq Ck^{-p}$, for $p \geq 2$, where t_{k+1} is the start time of episode $k+1$ in Algorithm 1, $\lambda_{\min,t_{k+1}}$ denotes the minimum eigenvalue of $\tilde{P}_{k+1} := P_{t_{k+1}}$, and the constant $C > 0$ depends only on p, n, n_u, $\mathbf{W}$, M_K, and λ.*

[10] Here, the probability $1-\delta$ is with respect to the randomness of the trajectory $(z_s)_{s\geq 1}$.

Recalling the probabilistic bound for $|\theta_t - \theta_*|_{P_t}$ from Proposition 7, we observe that $|\theta_t - \theta_*|$ is controlled by $1/\sqrt{\lambda_{\min,t}}$ and the self-normalization term. Using Theorem 8, we can show that the latter is dominated by the former, which grows at most polynomially in time due to Proposition 9. Consequently, the following improved concentration bound holds for the exact posterior.

Theorem 10 *Suppose Assumptions 1–5 hold. Then, the exact posterior μ_t and the approximate posterior $\tilde{\mu}_t$ realized from the shared Brownian motion satisfy*

$$\mathbb{E}\big[\mathbb{E}_{\theta_t\sim\mu_t}[|\theta_t-\theta_*|^p \mid h_t]\big] \leq C\left(t^{-\frac{1}{4}}\sqrt{\log t}\right)^p, \textit{ and } \mathbb{E}\big[\mathbb{E}_{\tilde{\theta}_t\sim\tilde{\mu}_t}[|\tilde{\theta}_t-\theta_*|^p \mid h_t]\big] \leq C\left(t^{-\frac{1}{4}}\sqrt{\log t}\right)^p$$

for all $t \geq 1$ and $p \geq 2$, where the outer expectation is taken over all histories, and the constant $C > 0$ depends only on p, n, n_u, $\mathbf{W}$, M_K, M_ρ, and λ.

6. Regret Bound

To further improve the bound in Theorem 8, we decompose the moment of the system state into two parts based on the following cases: $|\tilde{\theta}_t - \theta_*| \leq \epsilon_0$ and $|\tilde{\theta}_t - \theta_*| > \epsilon_0$, where ϵ_0 is a positive constant. When ϵ_0 is sufficiently small, we have $|A_* + B_*K(\tilde{\theta}_t)| < 1$, and thus the first part can be handled easily. For the second part, we invoke the Markov inequality to balance the growth of the state with the tail probability by choosing an appropriate value of p. This intuitive argument can be made rigorous using Theorems 8 and 10, leading to the following result.

Theorem 11 *Suppose that Assumptions 1–5 hold. For any $T > 0$ and a random trajectory $(x_s)_{s=1}^T$ generated by Algorithm 1, we have $\mathbb{E}[|x_t|^q] < C$ for $q = 2, 4$, where the constant $C > 0$ depends only on $p, n, n_u, \mathbf{W}, M_K, M_\rho, \epsilon_0$, and λ. Here, ϵ_0 is a positive constant such that $|\theta - \theta_*| \leq \epsilon_0$ implies $|A_* + B_*K(\theta)| < 1$.*

Finally, we establish our main result: Algorithm 1 achieves an $O(\sqrt{T})$ Bayesian regret bound.[11]

Theorem 12 *Suppose Assumptions 1–5 hold. Then, the Bayesian regret (3) of Algorithm 1 is bounded as follows: $R(T) \leq O(\sqrt{T})$.*

7. Concluding Remarks

We proposed a novel approximate Thompson sampling algorithm for learning LQR with an improved $O(\sqrt{T})$ regret bound. Our method does not require the noise to be Gaussian or the columns of Θ to be independent. This relaxation of restrictive assumptions is enabled by a carefully designed preconditioned ULA and the use of perturbed control actions only at the end of each episode.

As a future research direction, it may be possible to extend our algorithm to settings with noise distributions having non-log-concave potentials. In our work, the log-concavity of the posterior potential is preserved under the considered noise models, which enables acceleration of the sampling process through preconditioning. To handle more general classes of noise, alternative techniques beyond the current ULA framework may be necessary. Recently, Cheng et al. (2018) derived sharp non-asymptotic convergence rates for Langevin dynamics in nonconvex settings. We plan to investigate the incorporation of such results into our framework.

[11] The regret bound is also empirically verified by the results of our experiments, which are presented in Appendix C of the extended version (Kim et al., 2024).

Acknowledgments

The authors would like to thank Prof. Rahul Jain at the University of Southern California for his insightful discussion. The first two authors contributed equally. This work was supported in part by the Information and Communications Technology Planning and Evaluation (IITP) grants funded by MSIT No. 2022-0-00124, No. 2022-0-00480 and No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University). Yeoneung Kim was supported in part by the National Research Foundation of Korea (NRF) grant funded by MSIT (RS-2023-00219980, RS-2023-00211503).

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 19.1–26. PMLR, 2011.

Yasin Abbasi-Yadkori and Csaba Szepesvári. Bayesian optimal control of smoothly parameterized systems. In *Proceedings of 31st Conference on Uncertainty in Artificial Intelligence*, pages 1–11. Citeseer, 2015.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24:2312–2320, 2011.

Marc Abeille and Alessandro Lazaric. Thompson sampling for linear-quadratic control problems. In *Artificial Intelligence and Statistics*, pages 1246–1254. PMLR, 2017.

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

Marc Abeille and Alessandro Lazaric. Efficient optimistic exploration in linear-quadratic regulators via Lagrangian relaxation. In *International Conference on Machine Learning*, pages 23–31. PMLR, 2020.

Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, pages 39.1–26. PMLR, 2012.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135. PMLR, 2013.

Nawaf Bou-Rabee and Martin Hairer. Nonasymptotic mixing of the MALA algorithm. *IMA Journal of Numerical Analysis*, 33(1):80–110, 2013.

Pierre Bras. Langevin algorithms for very deep neural networks with application to image classification. *arXiv preprint arXiv:2212.14718*, 2022.

Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter L Bartlett, and Michael I Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018.

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pages 1300–1309. PMLR, 2019.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3): 651–676, 2017.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Zhiyan Ding, Qin Li, Jianfeng Lu, and Stephen J Wright. Random coordinate Langevin Monte Carlo. In *Conference on Learning Theory*, pages 1683–1710. PMLR, 2021.

Alain Durmus and Eric Moulines. Sampling from a strongly log-concave distribution with the unadjusted Langevin algorithm. 2016.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Conference on learning theory*, pages 793–797. PMLR, 2018.

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

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

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

Mukul Gagrani, Sagar Sudhakara, Aditya Mahajan, Ashutosh Nayyar, and Yi Ouyang. A modified Thompson sampling-based learning algorithm for unknown linear systems. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 6658–6665. IEEE, 2022.

Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov Chain Monte Carlo in practice*. CRC press, 1995.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2): 123–214, 2011.

Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized Markov decision processes. In *Proceedings of The 28th Conference on Learning Theory*, pages 861–898. PMLR, 2015.

Morteza Ibrahimi, Adel Javanmard, and Benjamin Roy. Efficient reinforcement learning for high dimensional linear quadratic systems. *Advances in Neural Information Processing Systems*, 25, 2012.

Yassir Jedra and Alexandre Proutiere. Minimal expected regret in linear quadratic control. In *International Conference on Artificial Intelligence and Statistics*, pages 10234–10321. PMLR, 2022.

Taylan Kargin, Sahin Lale, Kamyar Azizzadenesheli, Animashree Anandkumar, and Babak Hassibi. Thompson sampling achieves $\tilde{O}(\sqrt{T})$ regret in linear quadratic control. In *Conference on Learning Theory*, pages 3235–3284. PMLR, 2022.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pages 199–213. Springer, 2012.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2):209–232, 2002.

Yeoneung Kim, Gihun Kim, Jiwhan Park, and Insoon Yang. Approximate Thompson sampling for learning linear quadratic regulators with $O(\sqrt{T})$ regret. *arXiv preprint arXiv:2405.19380*, 2024.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Animashree Anandkumar. Reinforcement learning with fast stabilization in linear dynamical systems. In *International Conference on Artificial Intelligence and Statistics*, pages 5354–5390. PMLR, 2022.

Ioan Doré Landau, Rogelio Lozano, Mohammed M'Saad, et al. *Adaptive control*, volume 51. Springer New York, 1998.

Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *30th AAAI Conference on Artificial Intelligence*, 2016.

Xuechen Li, Denny Wu, Lester Mackey, and Murat A Erdogdu. Stochastic Runge-Kutta accelerates Langevin Monte Carlo and beyond. *arXiv preprint arXiv:1906.07868*, 2019.

Jianfeng Lu, Yulong Lu, and Zhennan Zhou. Continuum limit and preconditioned Langevin sampling of the path integral molecular dynamics. *Journal of Computational Physics*, 423:109788, 2020.

Yulong Lu, Jianfeng Lu, and James Nolen. Accelerating Langevin sampling with birth-death. *arXiv preprint arXiv:1905.09863*, 2019.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. *Advances in Neural Information Processing Systems*, 32, 2019.

Eric Mazumdar, Aldo Pacchiano, Yi-an Ma, Peter L Bartlett, and Michael I Jordan. On Thompson sampling with Langevin algorithms. *arXiv preprint arXiv:2002.10002*, 2020.

Wenlong Mou, Yi-An Ma, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. High-order Langevin diffusion yields an accelerated MCMC algorithm. *arXiv preprint arXiv:1908.10859*, 2019.

Ian Osband and Benjamin Van Roy. Posterior sampling for reinforcement learning without episodes. *arXiv preprint arXiv:1608.02731*, 2016.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (More) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

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

Grigorios A Pavliotis. *Stochastic processes and applications: Diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer, 2014.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

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

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, pages 681–688, 2011.