M.S. THESIS

Learning Linear-Quadratic Regulators via Thompson Sampling with Preconditioned Langevin Dynamics

사전 조건화된 랑주뱅 동역학을 결합한 톰슨 샘플링을 통한 선형 2차 제어기 학습

BY

Gihun Kim

AUGUST 2023

DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

M.S. THESIS

Learning Linear-Quadratic Regulators via Thompson Sampling with Preconditioned Langevin Dynamics

사전 조건화된 랑주뱅 동역학을 결합한 톰슨 샘플링을 통한 선형 2차 제어기 학습

BY

Gihun Kim

AUGUST 2023

DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Learning Linear-Quadratic Regulators via Thompson Sampling with Preconditioned Langevin Dynamics

사전 조건화된 랑주뱅 동역학을 결합한 톰슨 샘플링을 통한 선형 2차 제어기 학습

지도교수 양 인 순

이 논문을 공학석사 학위논문으로 제출함

2023년 8월

서울대학교 대학원

전기 컴퓨터 공학부

김 기 훈

김기훈의 공학석사 학위 논문을 인준함

2023년 8월

위 원 장: _____
부위원장: _____
위 원: _____

Abstract

Thompson sampling (TS) is a widely used approach for addressing the exploration-exploitation trade-off in online learning problems, including reinforcement learning for linear quadratic regulators (LQR). However, in TS for learning LQR, its theoretical analysis is often limited to the case of Gaussian noises. The sampling can be performed directly when we further assume that the unknown system parameters lie in a prespecified compact set as in [1], which is seemingly restrictive. We propose a new TS algorithm for LQR, exploiting Langevin dynamics to handle a larger class of problems including those with non-Gaussian noises. The notion of the preconditioner is introduced to generate samples from non-conjugate posterior distributions. Our algorithm is capable of sampling parameters from approximate posteriors quickly. It attains $O(\sqrt{T})$ expected regret bound slightly improving the result of [1] under the minimal assumption on the prior distribution and admissible set requiring neither a particular initialization technique nor information on the true system parameter. Our regret analysis leverages a nontrivial concentration inequality for the preconditioned Langevin algorithm together with self-normalization techniques. The performance of our algorithm has been demonstrated through numerical experiments as well.

keywords: Linear quadratic regulator, Thompson sampling, Langevin dynamics, preconditioning technique

student number: 2020-26137

Contents

List of Figures

Chapter 1

Introduction

Balancing the exploration-exploitation trade-off is a fundamental dilemma in reinforcement learning (RL) because it is mostly unclear to choose between acting to learn about an unknown environment ('exploration') or making a reward-maximizing decision given the information gathered so far ('exploitation'). This issue has been systemically addressed in two main approaches, namely optimism in the face of uncertainty (OFU) and Thompson sampling (TS). The methods using OFU first construct confidence sets for the environment or model parameters given the samples observed so far. After finding the reward-maximizing or optimistic parameters within the confidence set, an optimal policy with respect to the parameters is constructed and executed [3]. Various algorithms using OFU are shown to have strong theoretical guarantees in bandits [4].

On the other hand, TS is a Bayesian method in which environment or model parameters are sampled from the posterior that is updated along the process using samples and a prior, and an optimal policy with respect to the sampled parameter is constructed and executed [5]. In terms of computational tractability, TS has an advantage over OFU that requires an optimal solution to a nontrivial optimization problem over a confidence set in each episode. Furthermore, TS has been successfully used in online learning for various sequential decision-making problems such as multi-armed bandit problems

[6, 7, 8], Markov Decision Process (MDP) [9, 10, 11] and LQR [1, 10, 12, 13, 14], among others.

A fundamental step in TS-based learning is to sample from a distribution. Unfortunately, posterior sampling is generally challenging as well-known sampling techniques do not scale to high dimensional spaces. To overcome the limitation Markov Chain Monte Carlo (MCMC) based sampling methods are proposed [15]. In particular, Langevin MCMC is one of the most widely used sampling techniques in the field [16, 17, 18]. The convergence is also studied extensively as found in literature [16, 17, 19, 20]. Thanks to its advantages over existing sampling methods it has been applied to various learning problems such as Bayesian learning [18] and inverse reinforcement learning [21]. Yet tractable even in high dimensional spaces, sampling via Langevin MCMC still suffers from the curse of dimensionality requiring a tremendous amount of computation. To alleviate the issues various acceleration methods are studied (see [18, 22, 23, 24, 25] and references therein). In particular, [18, 26, 27] introduced a preconditioner from which our new algorithm and analysis are motivated.

1.1 Contributions

We propose a new computationally tractable TS-based algorithm achieving the state-of-art regret $O(\sqrt{T})$ for learning LQR as well as the exact rate of convergence of the sampled system parameter. Our algorithm features that a wide class of system noises can be used and no a priori information on the admissible is needed. Central to enhancing computational efficiency is introducing preconditioned Langevin dynamics for sampling, which enables us to achieve $O(\sqrt{T})$ Bayesian regret bound for learning LQR problems.

- *Preconditioned ULA*: We introduce the preconditioned Langevin MCMC for the acceleration of sampling process. The improved convergence rate between the exact and approximate posterior is obtained, which results in achieving $O(\sqrt{T})$

Bayesian regret bound.

- *Rate of convergence around true system parameter*: The sampled system parameter obtained via our new algorithm concentrates around the true system parameter with the rate $\tilde{O}(t^{-\frac{1}{4}})$. The action is perturbed only one time at the end of each episode for efficient exploration. Thanks to this, we can improve the polynomial-in-time state bound to constant and achieve the better regret as above.

- Above all, we simply assume that the admissible set is bounded to achieve the aforementioned results. It is the first work for achieving $O(\sqrt{T})$ Bayesian regret bound with non-Gaussian noise under such a mild assumption.

1.2 Related work

There is a rich body of literature regarding the estimation of system parameters and synthesis of a control gain matrix for LQR problems, which can be categorized as followings.

Optimism in the Face of Uncertainty (OFU): [28, 29] propose an OFU-based learning algorithm that iteratively selects the best-performing control actions while constructing the confidence sets. It is shown that the $\tilde{O}(\sqrt{T})$ is regret bound yet computationally unfavorable due to the complex constraint. To circumvent there is an attempt to translate the original nonconvex optimization problem arising in the OFU approach into semidefinite programming [30, 31], which obtains the same regret $\tilde{O}(\sqrt{T})$ with high probability. On the other hand, in[14, 32], randomized actions are employed to avoid constructing confidence sets and address asymptotic regret bound $\tilde{O}(\sqrt{T})$. Recently, [33] proposes an algorithm that quickly stabilizes the system and obtains $\tilde{O}(\sqrt{T})$ regret bound without using stabilizing control gain matrix.

Thompson sampling (TS): It is shown that the upper bound for the frequentist regret can be as worse as $\tilde{O}(T^{2/3})$ [13] and it is improved to $\tilde{O}(\sqrt{T})$ [34] based on

TS. However, both of them assume that the noise follows the Gaussian distribution and deals with one-dimensional only. Later on, [35] extends the previous work to the multi-dimensional case under the Gaussian noise. For Bayesian regret, previous results [1, 2] open up the possibility of applying TS based algorithm with provable $\tilde{O}(\sqrt{T})$ Bayesian regret bound yet the result suffers from some limitations. In their works both noise and the prior distribution of system parameters are assumed to follow the Gaussian, which allows updated posteriors to have the same structural properties and log-concave potential thanks to its conjugacy. In their work, it is crucial to assume that system parameters lie in a compact set that is defined via the true parameter itself. The following work [2] relaxes the technical assumption but the admissible set is not identified explicitly as well. Additionally, the columns of the system parameter matrix are assumed to be independent.

Chapter 2

Preliminaries

2.1 Linear-Quadratic Regulators

Consider a linear stochastic system of the form

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad t = 1, 2, \ldots, \tag{2.1}$$

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 $\mathbf{W}$. Throughout the paper, I_n represents n by n identity matrix and, we define the norm as $|v|_P := \sqrt{v^\top P v}$ for a positive semidefinite matrix P and a vector v.

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

1. *The probability density function (pdf) of noise $p_w(\cdot)$ is known, smooth and twice differentiable. Additionally, the following inequalities hold:*

$$\underline{m} I_n \preceq -\nabla^2_{w_t} \log p_w(w_t) \preceq \overline{m} I_n$$

for some $\underline{m}, \overline{m} > 0$,

2. $\mathbb{E}[w_t] = 0$ *and* $\mathbb{E}[w_t w_t^\top] = \mathbf{W}$*, where* $\mathbf{W}$ *is positive definite.*

Let $d := n + n_u$ and $\Theta \in \mathbb{R}^{d \times n}$ be the system parameter matrix defined by $\Theta := \begin{bmatrix} \Theta(1) & \cdots & \Theta(n) \end{bmatrix} := \begin{bmatrix} A & B \end{bmatrix}^\top$, where $\Theta(i) \in \mathbb{R}^d$ denotes the ith column of Θ. Here, the columns are not assumed to be Gaussian or independent as in [2, 1].

We also let $\theta := \mathrm{vec}(\Theta) := (\Theta(1), \Theta(2), \ldots, \Theta(n)) \in \mathbb{R}^{dn}$ denote the vectorized version of Θ. We often refer to θ as the parameter vector.

Let $h_t := (x_1, u_1, \ldots, x_{t-1}, u_{t-1}, x_t)$ 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

$$\Pi := \{\pi = (\pi_1, \pi_2, \ldots) \mid \pi_t : H_t \to \mathbb{R}^{n_u} \text{ is measurable } \forall t\}.$$

The stage-wise cost is chosen to be quadratic and is given by $c(x_t, u_t) := x_t^\top Q x_t + u_t^\top R u_t$ 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.[1] We consider the infinite-horizon average cost LQ setting with the following cost function:

$$J_\pi(\theta) := \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}_\pi \left[\sum_{t=1}^{T} c(x_t, u_t) \right].$$

Given $\theta \in \mathbb{R}^{dn}$, $\pi_*(x; \theta)$ denotes an optimal policy if it exists, and the corresponding optimal cost is given by

$$J(\theta) = \inf_\pi J_\pi(\theta).$$

It is well known that the optimal policy and cost can be obtained using the Riccati equation under the standard stabilizability and observability assumptions [38].

Theorem 1. *Suppose that* (A, B) *is stabilizable, and* $(A, Q^{1/2})$ *is observable. Then, the following algebraic Riccati equation (ARE) has a unique positive definite solution*

[1] This assumption is common in the literature [28, 14, 31, 36, 35, 37]

$P(\theta)$:

$$P(\theta) = Q + A^\top P(\theta)A - A^\top P(\theta)B(R + B^\top P(\theta)B)^{-1}B^\top P(\theta)A. \tag{2.2}$$

Furthermore, the optimal cost function is given by

$$J(\theta) = \mathrm{tr}(\mathbf{W}P(\theta)),$$

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 $K(\theta) := -(R+B^\top P(\theta)B)^{-1}B^\top P(\theta)A$.

The optimal policy called the linear-quadratic regulator (LQR) is an asymptotically stabilizing controller: it drives the closed-loop system state to the origin, that is, the spectrum of $A + BK(\theta)$ is contained in the interior of a unit circle.

2.2 Online learning of LQR

The theory of LQR is useful when the true system parameters $\theta_* := \mathrm{vec}(\Theta_*) := \mathrm{vec}(\begin{bmatrix} A_* & B_* \end{bmatrix}^\top)$ are fully known and stabilizable, which is not common. When the true parameter vector θ_* is unknown, online learning is a popular approach as pioneered in [28]. At each stage t, given the history h_t of observations, the learner executes a control action u_t and observes the resulting cost $c(x_t, u_t)$. Then, the system evolves according to the true linear dynamics $x_{t+1} = A_*x_t + B_*u_t + w_t$. Through such interactions between the learner and the system, the parameter vector and the policy are updated online. The performance of a learning algorithm is measured by regret. In particular, we consider the Bayesian setting, where the prior distribution μ_1 (with density p_1) of θ_* is assumed to be given, and use the following expected regret over T stages:

$$R(T) := \mathbb{E}\left[\sum_{t=1}^{T}(c(x_t, u_t) - J(\theta_*))\right]. \tag{2.3}$$

Here, θ_* is considered as a random variable of true parameter, and the expectation is taken with respect to the prior of θ_*, the probability distribution of noise $(w_1, w_2, \ldots, w_T)$ and the randomness of the learning algorithm. It is desirable for a learning algorithm to have a sublinear regret bound so that $R(T)/T \to 0$ as $T \to \infty$. When $\rho(A_* + B_* K(\theta)) > 1$ where $\rho(X)$ denotes the spectral radius of the matrix X, it is pessimistic to obtain the sublinear regret bound. To cope with this problem, [1, 2] assume that a compact stabilizing set whose element θ satisfies $\rho(A_* + B_* K(\theta)) < 1$ is given, and the system parameter can be sampled from the set. However, this assumption is unrealistic since one cannot tell if $\rho(A_* + B_* K(\theta)) < 1$ without knowing true system parameters A_* and B_*.

2.3 Thompson sampling

Thompson sampling (TS) or posterior sampling has been used in a large class of online learning problems [39]. The description of the naive TS algorithm for learning LQR is as follows. It starts with sampling a system parameter from the posterior μ_k at the beginning of episode k. Regarding this sample parameter as true, the control gain matrix $K(\theta_k)$ is computed by solving the ARE (2.2). During the episode, the control gain matrix is used to produce control action $u_t = K(\theta_k) x_t$, where x_t is the system state observed at time t. Along the way, the data $\mathcal{D}$ is collected and the posterior is updated.

The posterior update is performed using Bayes' rule and it preserves the log-concavity of distributions. To see this we let $z_t := (x_t, u_t) \in \mathbb{R}^d$ and write $p(x_{t+1}|z_t, \theta) = p_w(x_{t+1} - \Theta^\top z_t)$ which is log-concave in θ under Assumption 1. Hence, the posterior at stage t is given as

$$p(\theta|h_{t+1}) \propto p(x_{t+1}|z_t, \theta) p(\theta|h_t) = p_w(x_{t+1} - \Theta^\top z_t) p(\theta|h_t) \tag{2.4}$$

and it is log-concave as long as $p(\theta|h_t)$ is log-concave.

Bayesian learning always involves sampling from posterior distributions. However, sampling is computationally intractable particularly when the distributions at hand do not have conjugacy. Without conjugacy, posterior distribution does not have a closed-form expression, hence, a novel numerical method has to be developed. To sample from general distributions, a Markov chain Monte Carlo (MCMC) type algorithm needs to be introduced, which is of interest in the following subsection.

2.4 The Unadjusted Langevin algorithm (ULA)

To relax the decomposable Gaussian assumption in [1, 2] and handle a larger class of distributions, it is necessary to introduce an approximate posterior sampling method. To this end, we propose exploiting the unadjusted Langevin Algorithm (ULA), an MCMC method which generates samples approximately from a target distribution. We briefly go over the notion of Langevin algorithms together with the rate of convergence.

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 continuously differentiable. The Langevin dynamics takes the form of

$$dX_\xi = -\nabla U(X_\xi)\mathrm{d}\xi + \sqrt{2}\mathrm{d}B_\xi,$$

where B_ξ denotes the standard Brownian motion in $\mathbb{R}^{n_x}$. It is well-known that given an arbitrary X_0, the pdf of X_ξ converges to the target pdf $p(x)$ as $\xi \to \infty$ [23, 40].

To solve for X_ξ numerically, we apply the Euler–Maruyama discretization to the Langevin diffusion and obtain the following *unadjusted Langevin algorithm* (ULA):

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

where $(W_j)_{j\geq 1}$ is an i.i.d. sequence of standard n_x-dimensional Gaussian random vectors, and $(\gamma_j)_{j\geq 1}$ is a sequence of step sizes. Due to the discretization error, the Metropolis–Hasting algorithm that corrects the error is used together in general [16,

41, 42]. However, when the stepsize is small enough, such an adjustment can be omitted.

The condition number of the Hessian of the potential is an important factor in determining the rate of convergence. More precisely, we can show the following concentration property of ULA, which is a modification of Theorem 5 in [43]. For the sake of completeness, we present the proof in Appendix A.

Theorem 2. *Suppose that pdf* $p(x) \propto e^{-U(x)}$ *is strongly log-concave and* $U(x)$ *is Lipschitz smooth with respect to* x*, i.e.,* $\lambda_{\min} \preceq \nabla^2 U(x) \preceq \lambda_{\max}$ *for some* $\lambda_{\max}, \lambda_{\min} > 0$*. Let the stepsize is given by* $\gamma_j \equiv \gamma = O\left(\frac{\lambda_{\min}}{\lambda_{\max}^2}\right)$ *and the number of iterations* N *satisfy* $N \geq O\left((\frac{\lambda_{\max}}{\lambda_{\min}})^2\right)$*. Given* $X_0 = \arg\min U(x)$*, let* p_N *denote the pdf of* X_N*. Then, the following inequality holds:*

$$\mathbb{E}_{x\sim p, \tilde{x}\sim p_N}\left[|x-\tilde{x}|^2\right]^{\frac{1}{2}} \leq O\left(\sqrt{\frac{1}{\lambda_{\min}}}\right).$$

Chapter 3

Learning algorithm

The naive TS for learning LQR has two weaknesses. One of them arises in choosing a destabilizing controller which makes the state grow exponentially and causes the regret to blow up. To handle this problem, [1, 2] introduce an admissible set that allows us to select only a stabilizing controller. However, verification of such a set is impossible in general without knowing the true system parameter. We show that no additional assumption beyond compactness is needed to achieve $O(\sqrt{T})$ Bayesian regret bound. This means that the state grows exponentially with low probability and can be quantified. The other comes from inefficiency in the sampling process when system noises and the prior are not conjugate distributions. In such cases, ULA is an alternative but it is often extremely slow. To speed up, we introduce a preconditioning technique, which is indeed a simple change variable but results in faster convergence.

3.1 Preconditioned ULA for approximate posterior sampling

One key component of our approach 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 step iterations. To accelerate, we propose a preconditioning technique that has been used for Langevin algorithms in different contexts, e.g., see [44, 45, 46].

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

$$\theta_{j+1} = \theta_j - \gamma P^{-1}\nabla U(\theta_j) + \sqrt{2\gamma P^{-1}}W_j, \tag{3.1}$$

where $(W_j)_{j\geq 1}$ is an i.i.d. sequence of standard n_x-dimensional Gaussian random vectors. With the data $z_t = (x_t, u_t)$ collected, the preconditioner in our problem is defined as

$$P_t := \lambda I_{dn} + \sum_{s=1}^{t-1} blkdiag\{z_s z_s^\top\}_{i=1}^n, \tag{3.2}$$

where $blkdiag\{A_i\}_{i=1}^n \in \mathbb{R}^{dn\times dn}$ denotes the block diagonal matrix consisting of A_i's, and $\lambda > 0$ is determined by the prior. Our preconditioner is designed in a way to reduce the number of step iterations, thereby guaranteeing a faster convergence for general noise and prior distributions. We now propose the following lemma which implies that the curvature of the Hessian of the potential is bounded when scaled along the spectrum of the preconditioner.

Lemma 1. *Under Assumption 1, for all θ and t,*

$$m \preceq P_t^{-\frac{1}{2}}\nabla^2 U_t(\theta)P_t^{-\frac{1}{2}} \preceq M,$$

where $m = \min\{\underline{m}, 1\}$, $M = \max\{\overline{m}, 1\}$, $P_t = \lambda I_{dn} + \sum_{s=1}^{t-1} blkdiag(\{z_s z_s^\top\}_{i=1}^n)$ and the potential of the posterior $U_t(\theta) = -\log p(\theta|h_t)$ where U_1 satisfies $\nabla_\theta^2 U_1(\cdot) = \lambda I_{dn}$ for some $\lambda > 0$.

The proof is given in Appendix B. It follows from this lemma that we can rescale the number of iterations needed for the convergence of ULA while ensuring a better

level of accuracy for the concentration of the sampled system parameter. Throughout the paper, we use $U_k := U_{t_k}$ to explicitly show their dependency on the current episode k.

3.2 Main Algorithm

Before illustrating the main algorithm, let us first specify the admissible set for prior avoiding the unrealistic prespecified compact set of stabilizing parameters as in [1, 2]. In [1], their algorithms assume that $\{\theta : |A_* + B_* K(\theta) \leq \delta < 1\}$ is available, which is not verifiable when the true parameters (A_*, B_*) are unknown. In the following work [2], authors assume existence of the confidence set Ω_1 as follows: for any $\theta, \phi \in \Omega_1$ and $0 < \delta < 1$, $\rho(A_\theta + B_\theta K(\phi)) \leq \delta$. However, the construction of such a set is still mysterious. To alleviate this issue, they bypass the explicit construction of such a set leveraging the result [47].

We emphasize that even with the stabilization a technique that exploits random control gain matrices to identify the stabilizable set, one can only obtain a probabilistic guarantee. Furthermore, such an implementation is performed before the algorithm begins. We instead introduce a simple bounded set whose element θ is assumed to be stabilizable and to induce finite infinite-horizon cost J. The verification of this condition is indeed straightforward as no information on the true system parameter is needed, which is the major difference from the existing approach in the Bayesian setup.

Let us introduce an admissible set used for the algorithm as suggested in [34].

Definition 1. $\mathcal{S} := \{\theta \in \mathbb{R}^{dn} : |\theta| \leq S, |A + BK(\theta)| \leq \rho < 1, J(\theta) \leq M_J\}$ *for some* $S, \rho, M_J > 0$ *and* $\theta = \mathrm{vec}(\begin{bmatrix} A & B \end{bmatrix}^\top)$.

For $\theta \in \mathcal{S}$, there exists $M_P > 0$ such that $|P(\theta)| \leq M_P$ as found in [13]. Therefore, $|[I \quad K(\theta)^\top]| \leq M_K$ for some $M_K > 1$ and a direct implication of this result is that $|A_* + B_* K(\theta)| \leq M_\rho$ for some $M_\rho > 0$. Here, $K(\theta)$ denotes the control gain

matrix associated with θ.

Definition 2. *$|P(\theta)| \leq M_P$, $|[I \quad K(\theta)^\top]| \leq M_K$, and $|A_* + B_* K(\theta)| \leq M_\rho$ for some $M_P, M_K, M_\rho > 0$ when $\theta \in \mathcal{S}$. We further assume that $M_\rho \geq 1$.*

Assumption 2. *For $\lambda \geq 1$, let the prior p_1 satisfy that $\nabla^2_\theta U_1(\cdot) = \lambda I_{dn}$ for potential $U_1(\theta) = -\log p_1(\cdot)$ and $supp(p_1(\cdot)) \subset \mathcal{S}$.*

Remark 1. *For instance, the projection of multivariate normal distribution with covariance $\frac{1}{\lambda} I_{dn}$ on $\mathcal{S}$ yields the prior satisfying the Assumption 2.*

Furthermore, once constants S, ρ, and M_J are specified, one easily rejects sampled system parameters if it is not contained in $\mathcal{S}$, which is one of the major differences from [1] as no miracle stabilizing set is needed.

We next state our main algorithm. Let t_k and T_k denote the start time and the length of episode k respectively. By the definition, $t_1 = 1$ and $t_{k+1} = t_k + T_k$. The length of episode k is chosen as $T_k = k + 1$.

To update the posterior, or equivalently, its potential at episode k, we use the transition dataset $\mathcal{D} := \{(z_t, x_{t+1})\}_{t_{k-1} \leq t \leq t_k - 1}$ collected during the previous episode. It follows from (2.4) that the potential can be updated using the observations as

$$U_k(\theta) = U_{k-1}(\theta) - \sum_{(z_t, x_{t+1}) \in \mathcal{D}} \log p_w(x_{t+1} - \Theta^\top z_t),$$

where U_0 is set to be U_1, the potential of the prior.

Having the posterior updated, approximate posterior sampling is performed using the preconditioned ULA. To begin, we choose the preconditioner, stepsize, and number of iterations as $P_k = P_{t_k}$, $\gamma_k = \gamma_{t_k}$ and $N_k = N_{t_k}$ for $P_t := \lambda I_{dn} + \sum_{s=1}^{t-1} blkdiag\{z_s z_s^\top\}_{i=1}^n$, $\gamma_t := \frac{m\lambda_{\min,t}}{16 M^2 \max\{\lambda_{\min,t}, t\}}$ and $N_t := \frac{4 \log_2(\max\{\lambda_{\min,t}, t\}/\lambda_{\min,t})}{m\gamma_t}$ where $\theta_{\min,t}$ is a minimizer of the potential U_t, and $\lambda_{\min,t}$, $\lambda_{\max,t}$ are minimum, maximum eigenvalues of P_t.

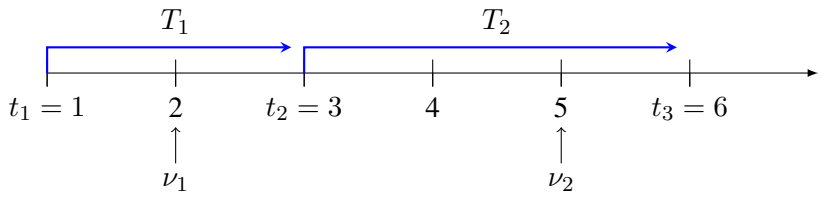

Figure 3.1: Infusing noise for better exploration

With γ_k and N_k defined above, the update rule (3.1) for the preconditioned ULA is expressed as

$$\theta_{j+1} \sim \mathcal{N}(\theta_j - \gamma_k P_k^{-1} \nabla U_k(\theta_j), 2\gamma_k P_k^{-1}). \tag{3.3}$$

Given $S > 0$, $0 < \rho < 1$ and $M_J > 0$, we check whether θ_{N_k} achieved from performing the update above N_k times is contained in $\mathcal{S}$. If so, set $\tilde{\theta}_k = \theta_{N_k}$. Finally, the controller $K_k = K(\tilde{\theta}_k)$ for kth episode is computed using Theorem 1 with the sampled system parameter $\tilde{\theta}_k$. As soon as observing the current state x_t, the control action u_t is executed to the system at time t. Accordingly, the dataset $\mathcal{D}$ is constructed collecting (z_t, x_{t+1}) for all $t \in [t_k, t_{k+1} - 1]$. One notable feature of our algorithm is that the action u_t is perturbed by random vector $(\nu_s)_{s \geq 1}$ right before the end of each episode. Precisely, the action $u_t = K_k x_t$ is applied when $t = [t_k, t_{k+1} - 2]$ and $u_t = K_k x_t + \nu_t$ is executed when $t = t_{k+1} - 1$ for additional random noise ν_t satisfying the assumption below. This perturbation enhances the exploration. The external noise signal contributes to the effect of persistence excitation Proposition 3 which states that the minimum eigenvalue of the preconditioner grows in time. The schemetic diagram is provided in Figure 3.2.

Assumption 3. *The sequence of $\bar{L}_\nu$-sub-Gaussian[1] random variable $\nu_s \in \mathbb{R}^{n_u}$ satisfies $\nu_s = 0$ if $s \in [t_j, t_{j+1} - 2]$ for all $j \geq 2$. For $s \notin [t_j, t_{j+1} - 2]$, let $\mathbb{E}[\nu_s] = 0$ and $\mathbf{W}' := \mathbb{E}[\nu_s \nu_s^\top]$ is a positive definite matrix whose maximum and minimum eigen-*

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

values are identical to those of $\mathbf{W}$*, the covariance of system noise. Without loss of generality we may assume* $\nu_1 = \nu_2 = 0$*.*

Remark 2. *The assumption on the minimum eigenvalue of* $\mathbf{W}'$ *is needed just for simplicity in the proof of Proposition 3 which is about the growth of* $\lambda_{\min}(P_t)$*.*

We end this section by discussing in detail why the proposed preconditioner P_k is useful. Recalling Lemma 1, we see that $m\lambda_{\min,k} I_{dn} \preceq \nabla^2 U_k \preceq M\lambda_{\max,k} I_{dn}$. It follows from Theorem 2 that $N_k = O((\frac{\lambda_{\max,k}}{\lambda_{\min,k}})^2)$ iterations is required for $\frac{1}{\sqrt{\lambda_{\min,k}}}$ error bound.

On the other if we can show that $|x_t| < C$ for some C, the trace inequality would yield that $\lambda_{\max,k} = O(t_k)$ since $\lambda_{\max,k} \leq \mathrm{tr}(P_t) \leq Ct$ for different constant C. If we further have $\lambda_{\min,k} = O(\sqrt{t_k})$, then $N_k = O(\sqrt{t_k})$ by our choice of γ_k and N_k. We will show in the following section that this particular choice allows us to achieve $\frac{1}{\sqrt{t_k}}$ rate of convergence rather than $\frac{1}{\sqrt{\lambda_{\min,k}}}$.

Algorithm 1 Thompson sampling with Langevin dynamics for LQR

Given p_1;

Initialization: $t \leftarrow 1$, $t_0 \leftarrow 0$, $x_1 \leftarrow 0$, $\mathcal{D} \leftarrow \emptyset$,

$U_0 \leftarrow U_1$, $\tilde{\theta}_0 \leftarrow \arg\min U_1(\theta)$, $\theta_{\min,0} \leftarrow \tilde{\theta}_0$;

for Episode $k = 1, 2, \ldots$ **do**

  $T_k \leftarrow k + 1$;

  $t_k \leftarrow t$;

  $U_k(\cdot) := U_{k-1}(\cdot) - \sum_{(z_t, x_{t+1}) \in \mathcal{D}} \log p(x_{t+1}|z_t, \cdot)$;

  $\mathcal{D} \leftarrow \emptyset$;

  $\theta_{\min,k} \in \arg\min U_k(\theta)$;

  Compute the preconditioner P_k, the step size γ_k, and the number of iterations N_k;

  while True **do**

    $\theta_0 \leftarrow \theta_{\min,k}$;

    for Step $j = 0, 1, \ldots, N_k - 1$ **do**

      Sample θ_{j+1} according to (3.3);

    end for

    if $\theta_{N_k} \in \mathcal{S}$ **then**

      $\tilde{\theta}_k \leftarrow \theta_{N_k}$

      Break;

    end if

  end while

  Compute the gain matrix $K_k := K(\tilde{\theta}_k)$;

  while $t \leq t_k + T_k - 1$ **do**

    Apply control $u_t = K_k x_t + \nu_t$ for ν_t satisfying Assumption 3;

    Observe new state x_{t+1};

    Update $\mathcal{D} \leftarrow \mathcal{D} \cup \{(z_t, x_{t+1})\}$;

    $t \leftarrow t + 1$;

  end while

end for

Chapter 4

Concentration properties

In this section, the concentration properties as well as the growth of the state trajectory are discussed.

Let us define the exact and approximate posterior distributions concerned with the potential U_t. We call the probability measure $\mu_t \sim exp(-U_t)$ *exact posterior*. For the approximate posterior, let us recall the preconditioned ULA,

$$\theta_{j+1} \sim \mathcal{N}(\theta_j - \gamma_t P_t^{-1} \nabla U_t(\theta_j), 2\gamma_t P_t^{-1}),$$

for $\theta_0 = \theta_{\min,t}$ and P_t, γ_t, N_t defined in Section 3.2. Here, $\theta_{\min,t}$ is a minimizer of U_t. We call the distribution of θ_{N_t} *approximate posterior* and denote it by $\tilde{\mu}_t$. Throughout the section, we denote the random variable following μ_t and $\tilde{\mu}_t$ by θ_t and $\tilde{\theta}_t$ respectively. Unless stated otherwise, we continue to use the following previously introduced notations to state results; λ satisfying Assumption 2, ρ, M_K, M_ρ, S from Definition 2 and 1, $\bar{L}_\nu$ and $\mathbf{W}$ from Assumption 3, $\bar{L} = \frac{1}{\sqrt{2m}}$ with m defined in Lemma 1.

4.1 Comparing exact and approximate posteriors

We begin by introducing a concentration result for the distribution of approximate system parameters and exact posterior. It is one of the essential parts for obtaining a

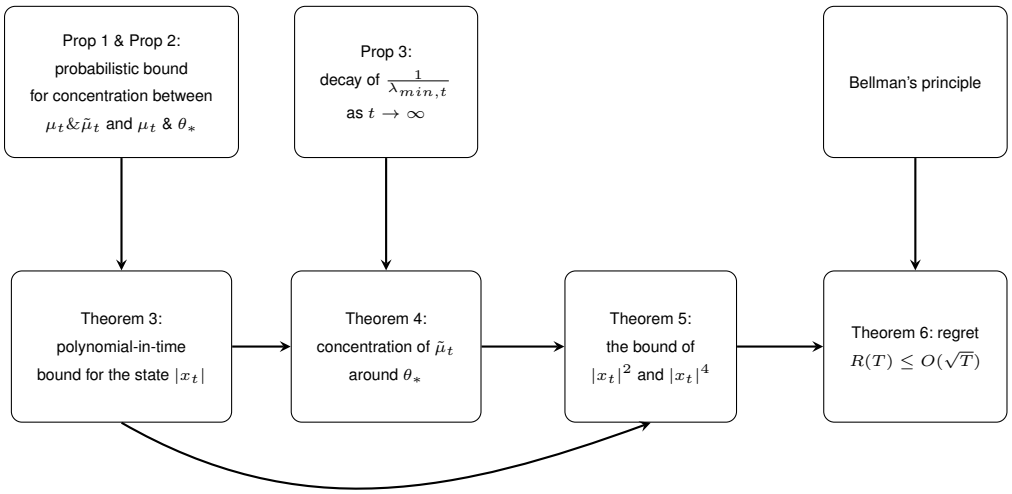

Figure 4.1: Outline of the proofs

non-asymptotic guarantee of the improved regret $O(\sqrt{T})$ dropping $\log T$ as it will be noted in Section 5.

The following proposition gives us the concentration between μ_t and $\tilde{\mu}_t$. The result quantifies the concentration depending on the moment p. The higher moment bound for $p > 2$ is used to characterize a set of system parameters where the state does not grow exponentially as illustrated in the following subsection while the bound for $p = 2$ is necessary for the regret analysis.

Proposition 1. *Suppose Assumption 1 and 2 hold. For any $t > 0$ and trajectory $(z_s)_{s\geq 1}$, the exact posterior μ_t and the approximate posterior $\tilde{\mu}_t$ obtained by pre-conditioned ULA satisfy*

$$\mathbb{E}_{\theta_t\sim\mu_t,\tilde{\theta}_t\sim\tilde{\mu}_t}\left[|\theta_t - \tilde{\theta}_t|_{P_t}^p \mid h_t\right] \leq D_p,$$

where $D_p = \left(\frac{pdn}{m}\right)^{\frac{p}{2}}\left(2^{2p+1} + 5^p\right)$ for $p \geq 2$. When $p = 2$, we further have

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

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

Proof. See Appendix C. □

Without the preconditioner, Theorem 2 would yield that we are only able to get $O(\frac{1}{\sqrt{\lambda_{\min,t}}})$ rate of convergence, which is an LQR version of Theorem 5 in [43]. To improve the concentration, we infuse the timestep t into the stepsize required for ULA so that the right-hand side decreases as the episode proceeds. The relationship $\max\{\lambda_{\min,t}, t\} \geq \lambda_{\min,t}$ contributes to achieving the better concentration.

Another important result we need is the probabilistic bound for the distance between the exact posterior and the true system parameter θ_*, which is essential in characterizing a confidence set relevant for TS-based learning.

Proposition 2. *Let Assumption 1 and 2 be enforced. Given a trajectory $(z_s)_{s\geq 1}$, define $P_t = \lambda I_{dn} + \sum_{s=1}^{t-1} blkdiag\{z_s z_s^\top\}_{j=1}^n$. Then for any $\delta > 0$ and $p \geq 2$,*

$$\mathbb{E}_{\theta_t \sim \mu_t}\left[|\theta_t - \theta_*|_{P_t}^p \mid h_t\right]^{\frac{1}{p}} \leq 2p\sqrt{\frac{8nM^2}{m^3}\log\left(\frac{n}{\delta}\left(\frac{\lambda_{\max,t}}{\lambda}\right)^{\frac{d}{2}}\right) + C} \tag{4.1}$$

holds with probability at least $1-\delta$ for some constant $C = C(d, \lambda, m, n) > 0$. Here, $\lambda_{\max,t}$ denotes the maximum eigenvalue of P_t.

4.2 Bounding expected state norms by a polynomial of time

A nontrivial result we can derive from Proposition 1 and 2 is that the system state has a polynomial-time growth in expectation. To justify this property we modify the confidence set and self-normalization technique developed for OFU approach [28, 48]. Our idea is to construct a set containing sampled system parameters obtained by ULA with high probability. The higher moment bound from Proposition 1 and 2 is crucial for the analysis as Markov type inequalities can be exploited for any power p. We then split the probability space of the stochastic process into two sets, good and bad as in standard OFU approaches.

Theorem 3. *Suppose that Assumption 1,2 and 3 hold. For $T > 0$, $p \geq 2$ and any trajectory $(x_s)_{s=1}^T$ generated by Algorithm 1, we have*

$$\mathbb{E}[\max_{j \leq t} |x_j|^p] \leq C t^{\frac{7}{2}p(d+1)}$$

for some constant $C(d, \lambda, m, p, \rho, \bar{L}_\nu, M_\rho, S) > 0$.

Remark 3. *In the next section, we will further improve the bound to constant, which is one of the main contributions of this work.*

4.2.1 Concentration of μ_t and $\tilde{\mu}_t$ around θ_*

Leveraging the previous results on the concentration and the expected state norms, we can deduce that the minimum eigenvalue of the preconditioner actually grows in time which is given in Proposition 3. With this property as well as Theorem 3 on hand, the concentration property of exact posterior follows. Finally, the triangle inequality yields the result desired, the concentration of approximate posterior around the true system parameter.

Let us begin with the observation that $\lambda_{\min}(P_t)$ grows at least $\sqrt{t}$ with high probability, which is motivated by [36]. The high-level description is as follows. To analyze the minimum eigenvalue of $\sum_s z_s z_s^\top$, we recall the decomposition

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

Here, for $j \leq k$ and s denoting the timestep, and

$$y_s := \begin{bmatrix} A_* x_{s-1} + B_* u_{s-1} \\ K_j (A_* x_{s-1} + B_* u_{s-1}) \end{bmatrix},$$

where K_j denotes the control gain matrix computed at the beginning of jth episode. We also let

$$L_s := \begin{bmatrix} I_n & 0 \\ K_j & I_{n_u} \end{bmatrix}, \quad \text{and} \quad \psi_s := \begin{bmatrix} w_{s-1} \\ \nu_s \end{bmatrix}.$$

The random matrix part is indeed a sum of random matrices and it is shown that they accumulate the minimum eigenvalue high probability.

The self-normalization term must be minimized to guarantee the growth of minimum eigenvalue. Thanks to Theorem 3, it is bounded by $O(\log T)$ with high probability. Since the random matrix part has $\Omega(\sqrt{T})$ growth rate, we obtain our desired result.

Proposition 3. *Suppose that Assumption 1,2 and 3 hold. Given $p \geq 3$ and $k \geq k_0(d, \lambda, m, p, \rho, \bar{L}_\nu, M_K, M_\rho, S, \mathbf{W})$, we have*

$$\mathbb{E}\left[\frac{1}{\lambda_{\min,k+1}^p}\right] \leq Ck^{-p}$$

for some constant $C(d, \lambda, m, p, \rho, \bar{L}_\nu, M_K, M_\rho, S, \mathbf{W}) > 0$. Here, $\lambda_{\min,k+1}$ denotes the smallest eigenvalue of $\lambda I_d + \sum_{s=1}^{t_{k+1}-1} z_s z_s^\top$ where $(z_s)_{s\geq 1}$ is obtained via our main algorithm.

Remark 4. *In fact, $\lambda_{\min,k}$ is same as that of our preconditioner P_k.*

A direct consequence of the proposition above is that $\mathbb{E}\left[\frac{1}{\lambda_{\min,t}^p}\right] \leq Ct^{-\frac{p}{2}}$ as $\lambda_{\min,t}$ increases as t grows.

Recalling the probabilistic bound for $|\theta_t - \theta_*|_{P_t}$ from Proposition 2, one can see that $|\theta_t - \theta_*|$ is controlled in terms of $\frac{1}{\lambda_{\min,t}}$ and self-normalization term. Thanks to Theorem 3, the latter is dominated by the former that has polynomial-time growth as seen in Proposition 3. Consequently, we claim the concentration result on the exact posterior μ_t.

Proposition 4. *Suppose that Assumption 1,2 and 3 hold. Given $p \geq 3$ and $t \geq t_0(d, \lambda, m, p, \rho, \bar{L}_\nu, M_K, M_\rho, S, \mathbf{W})$, the exact posterior μ_t obtained by Algorithm 1 and the true system parameter θ_* satisfy*

$$\mathbb{E}[\mathbb{E}_{\theta_t \sim \mu_t}[|\theta_t - \theta_*|^p h_t]] \leq C\left(t^{-\frac{1}{4}}\sqrt{\log t}\right)^p$$

for some constant $C(d, \lambda, m, n, p, \rho, \bar{L}_\nu, M_K, M_\rho, S, \mathbf{W}) > 0$.

Combining the result above with Proposition 1 through triangle inequality, we obtain the following concentration property of the approximate posterior.

Theorem 4. *Suppose that Assumption 1,2 and 3 hold. Given $p \geq 3$, $t \geq t_0(d, \lambda, m, p, \rho, \bar{L}_\nu, M_K, M_\rho, S, \mathbf{W})$, the true parameter θ_* and the approximate posterior $\tilde{\mu}_t$ satisfy*

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

for some constant $C(d, \lambda, m, n, p, \rho, \bar{L}_\nu, M_K, M_\rho, S, \mathbf{W}) > 0$. Here, $\tilde{\mu}_t$ denotes the approximate posterior corresponding to the posterior μ_t obtained by our algorithm.

Proof. By Jensen's inequality,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}_{\tilde{\theta}_t \sim \tilde{\mu}_t}\left[|\tilde{\theta}_t - \theta_*|^p | h_t\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}_{\theta_t \sim \mu_t, \tilde{\theta}_t \sim \tilde{\mu}_t}\left[|\tilde{\theta}_t - \theta_*|^p \middle| h_t\right]\right] \\
&\leq 2^{p-1}\mathbb{E}\left[\mathbb{E}_{\theta_t \sim \mu_t, \tilde{\theta}_t \sim \tilde{\mu}_t}\left[|\theta_t - \tilde{\theta}_t|^p | h_t\right]\right] + 2^{p-1}\mathbb{E}\left[\mathbb{E}_{\theta_t \sim \mu_t, \tilde{\theta}_t \sim \tilde{\mu}_t}\left[|\theta_t - \theta_*|^p | h_t\right]\right] \\
&\leq 2^{p-1}\mathbb{E}\left[\frac{D_p}{(\sqrt{\lambda_{\min,t}})^p}\right] + 2^{p-1}C\left(t^{-\frac{1}{4}}\sqrt{\log t}\right)^p \\
&\leq C\left(t^{-\frac{1}{4}}\sqrt{\log t}\right)^p,
\end{aligned}$$

where the second inequality comes from Proposition 1 and 4. □

This result is surprising in the sense that the learner disregards the possibility of choosing a destabilizing $\tilde{\theta}$ for t large even if we use a general admissible set $\mathcal{S}$ instead of miracle sets [1, 2]. Furthermore, the result provides a hint on the sample complexity for the quantification of the posterior around the true system parameter.

Chapter 5

Main result

We finally present that the Algorithm 1 indeed achieves $O(\sqrt{T})$ regret bound which is a slight improvement from $O(\sqrt{T}\log T)$ while handling a broader class of system noises under the minimal assumption on the admissible set. One of key components for obtaining this result is a uniform bound of the moment of state improving the polynomial-in-time bound.

5.1 Improved state bound for $\mathbb{E}[|x_t|^2]$ and $\mathbb{E}[|x_t|^4]$

As noted, we further improve the result 3 to constant bound. To do so we decompose the state moment into two parts: $|\tilde{\theta}_t - \theta_*| \leq \epsilon_0$ and $|\theta_t - \theta_*| > \epsilon_0$ for some $\epsilon_0 > 0$. When ϵ_0 is small enough, $|A_* + B_* K(\tilde{\theta}_t)| < 1$, hence, the state bound is obtained easily. To deal with the second part, we invoke Markov inequality to balance out the growth of the state and the tail probability by choosing an appropriate p. Such an analysis is available thanks to Theorem 3 and Theorem 4.

Theorem 5. *Suppose that Assumption 1,2 and 3 hold. For any $T > 0$ and trajectory $(x_s)_{s=1}^T$ generated by Algorithm 1, we have*

$$\mathbb{E}[|x_t|^q] < C, \quad q = 2, 4,$$

for some constant $C(\epsilon_0, d, \lambda, m, n, \rho, \bar{L}_\nu, M_K, M_\rho, S, \mathbf{W}) > 0$. *Furthermore,* ϵ_0 *is a number such that* $|\theta - \theta_*| \leq \epsilon_0$ *implies that* $|A_* + B_* K(\theta)| < 1$.

5.2 Regret bound

Finally, we present our main result which states that $O(\sqrt{T})$ regret bound is achieved. Since we consider the Bayesian regime, we write the regret in the form of

$$R(T) = \mathbb{E}\bigg[\sum_{k=1}^{n_T}\sum_{t=1}^{T}(c(x_t, u_t) - J(\bar{\theta}_*)\bigg],$$

where $\bar{\theta}_*$ denotes the random variable for the true system parameter. Here, Thanks to the astonishing result by Bellman [49], we have the recursive relation for the cumulative cost

$$\begin{aligned}
&J(\tilde{\theta}_k) + x_t^\top \tilde{P}_k x_t \\
&= x_t^\top Q x_t + \tilde{u}_t^\top R \tilde{u}_t + \mathbb{E}[(\tilde{A}_k x_t + \tilde{B}_k \tilde{u}_t + w_t)^\top \tilde{P}_k (\tilde{A}_k x_t + \tilde{B}_k \tilde{u}_t + w_t) \mid h_t] \\
&= x_t^\top Q x_t + \tilde{u}_t^\top R \tilde{u}_t + (\tilde{A}_k x_t + \tilde{B}_k \tilde{u}_t)^\top \tilde{P}_k (\tilde{A}_k x_t + \tilde{B}_k \tilde{u}_t) + \mathbb{E}[w_t^\top \tilde{P}_k w_t \mid h_t],
\end{aligned}$$

where $t \in [t_k, t_{k+1})$, $\tilde{\theta}_k$ sampled at the beginning of the kth episode, $\tilde{u}_t = K(\tilde{\theta}_k)$, and $\tilde{P}_k = P(\tilde{\theta}_k)$. One should note that there is a small gap between controllers $\tilde{u}_t$ and the u_t that we use for the algorithm since we infuse noise ν_t once in each episode. However, the contribution of this perturbation to the regret is as low as $\sqrt{T}$ since it is executed at most $\sqrt{T}$ times.

For the rest, we follow the argument provided in [1]. The difference is that we use the Proposition 1 to control the part containing $|\tilde{\theta}_t - \theta|$ whereas [1] deals with such terms using the explicit structure of distributions, hence, our concentration result provides a novel way of reducing the regret even when ULA is exploited for sampling. Furthermore, the use of Theorem 5 contributes to dropping the term $\log T$. In regret analysis based on Bellan's principle, we estimate the second and fourth power of the state by invoking Cauchy-Schwartz inequality to handle terms such as $|x_t||\tilde{\theta} - \theta|$.

When the higher moment is available, we can effectively estimate quantities involving such terms.

Theorem 6. *Let our prior p_1 satisfy Assumption 2. Then, under Assumption 1 and 3, the expected cumulative regret* (2.3) *of Algorithm 1 satisfies*

$$R(T) \leq O(\sqrt{T}).$$

To our best knowledge, all Bayesian regret bounds obtained in the aforementioned literature contain polylogarithmic terms in time horizon T while ours only includes constants. The presence decreasing gap between the exact and approximate posterior as shown in Proposition 1 contributes to obtaining the improved regret while the concentration property is not taken into account in [1]. We are able to achieve such a concentration property thanks to the unique characteristic of our preconditioned ULA, which results in an effective exploration of the true system parameter while learning.

Chapter 6

Experiment

To test our algorithm, we plot the expected cumulative regret in various dimensions considering Gaussian mixture and asymmetric noises which are non-Gaussian. In addition to that, we take the comparison experiment with [1] using Gaussian distrurbance. Finally, we experimentally show that our preconditioner method is computationally efficient.

6.1 Experimental setup

For the true system parameter Θ_*, we use

$$A_* = \begin{bmatrix} 0.3 & 0.1 & 0.2 \\ 0.1 & 0.4 & 0 \\ 0 & 0.7 & 0.6 \end{bmatrix}, \quad B_* = \begin{bmatrix} 0.5 & 0.4 & 0.5 \\ 0.6 & 0.3 & 0 \\ 0.3 & 0 & 0.2 \end{bmatrix},$$

and

$$A_* = \begin{bmatrix} 0.3 & 0.6 & 0.2 & 0.3 & 0.1 \\ 0 & 0.1 & 0.4 & 0 & 0.6 \\ 0.1 & 0.5 & 0.3 & 0 & 0.2 \\ 0.4 & 0 & 0.3 & 0.3 & 0 \\ 0.3 & 0.3 & 0.1 & 0.4 & 0.4 \end{bmatrix}$$

$$B_* = \begin{bmatrix} 0.5 & 0.4 & 0.2 & 0.5 & 0.4 \\ 0.6 & 0 & 0.3 & 0.1 & 0.3 \\ 0.5 & 0 & 0 & 0.1 & 0.2 \\ 0.1 & 0.5 & 0 & 0.2 & 0.4 \\ 0.2 & 0.1 & 0.6 & 0 & 0 \end{bmatrix}$$

and $Q = 2I_n, \quad R = I_n$. for $n = n_u = 3$ and 5 repsectively.

6.1.1 Gaussian mixture noise

In this section, we consider a Gaussian mixture noise which is given by

$$p_w(w_t) = \frac{1}{2(2\pi)^{3/2}}(e^{\frac{-(w_t-a)^2}{2}} + e^{\frac{-(w_t+a)^2}{2}}),$$

where $a = [\frac{1}{2}, \frac{1}{2}, \frac{1}{2}]^\top$ and $[\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}]^\top$ for $n = 3$ and 5. Taking gradients,

$$-\nabla \log p_w(w_t) = w_t - a + \frac{2a}{1 + e^{2w_t^\top a}},$$

and

$$\begin{aligned} -\nabla^2 \log p_w(w_t) &= I_n - 4aa^\top \frac{e^{2w_t^\top a}}{(1 + e^{2w_t^\top a})^2} \\ &\geq I_n - aa^\top \\ &\geq (1 - |a|^2) I_n. \end{aligned}$$

Therefore, the first condition in Assumption 1 is satisfied for $n = 3$ and 5.:

$$\frac{1}{4} I_3 \leq -\nabla^2 \log p_w(w_t) \leq I_3,$$

$$\frac{11}{16} I_5 \leq -\nabla^2 \log p_w(w_t) \leq I_5.$$

Executing Algorithm 1 under the Gaussian noise, Figure 6.1 shows that the trajectory oscillates around the origin.

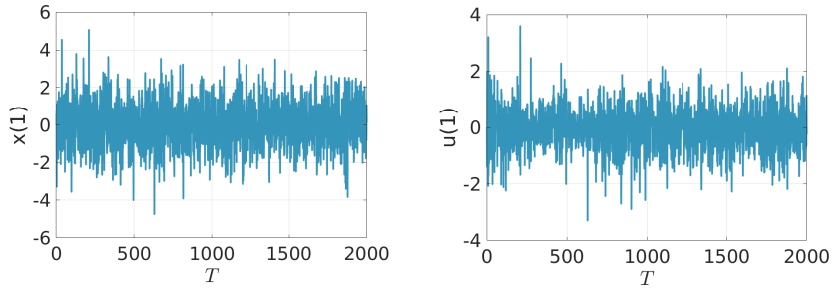

Figure 6.1: First component of state $x(1)$ and control $u(1)$

6.1.2 Asymmetric noise

We also consider asymmetric noise. To proceed we begin by constructing a noise as follows. Let all components of w_t be independent and its components $w_t(1), w_t(2), \ldots,$ $w_t(n-1)$ follow the standard Gaussian distribution where $w_t(i)$ denotes ith component of w_t. We set the Hessian of $\log w_t(n)$ to be piecewise linear, namely,

$$
\begin{aligned}
&-\frac{\partial^2 \log p(w_t)}{\partial w_t(n)^2} \\
&= \begin{cases} m & \text{if } w_t(n) < \alpha, \\ \frac{M-m}{\beta-\alpha} w_t(n) + m - \frac{(M-m)\alpha}{\beta-\alpha} & \text{if } \alpha \le w_t(n) < \beta, \\ M & \text{if } \beta \le w_t(n) \end{cases}
\end{aligned}
$$

for $\alpha < \beta$ which are chosen carefully to satisfy Assumption 1. We choose $m = 1$ and $M = 10$ for the experiment. The comparison with the standard Gaussian distribution using various values for M and $m = 1$ is demonstrated in Figure 6.2. We first generate a sequence of noises following the prescribed distribution offline through ULA. With the sample, the covariance is estimated.

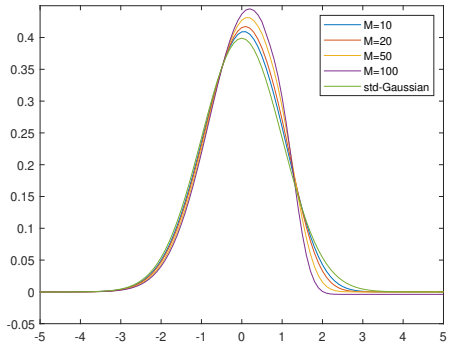

Figure 6.2: Comparison between $w_t(n)$ and standard Gaussian noise

6.2 Performance of our algorithm

We test our algorithm with a Gaussian mixture and asymmetric noises. We also consider the Gaussian disturbance to make a comparison with [1] as their method is only applicable to this particular noise.

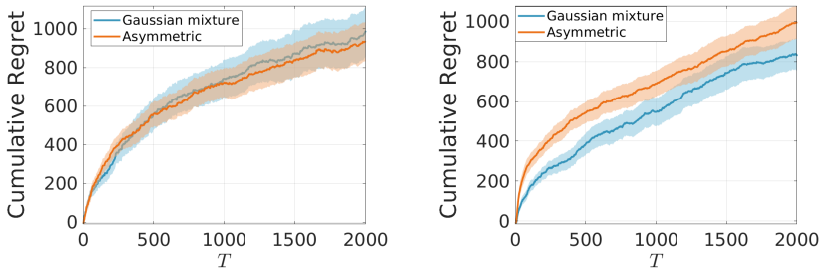

Figure 6.3: Expected cumulative regret $R(T)$ over a time horizon T using Gaussian mixture noise and asymmetric noise for $n = n_u = 3$ (left), for $n = n_u = 5$ (right).

We verify the effectiveness of our algorithm for various dimensions, $n = n_u = 3$ and 5. The simulation result is preseneted in Figure 6.3. For the experiment, we set true system parameters (A_*, B_*) to satisfy $\rho(A_* + B_*K) = 0.3365$ for $n = n_u = 3$ and 0.3187 for $n = n_u = 5$ where K is the control gain matrix associated with (A_*, B_*). The explicit numbers are demonstrated in Section 6.1. For the admissible set $\mathcal{S}$, we choose $S = 20$, $M_J = 20000$ and $\rho = 0.99$ for both cases regardless of the type

noises. We also sample action perturbation ν_s from $\mathcal{N}(0, \frac{1}{10000} I_{n_u})$ at the end of each episode. For all experiments, a prior is set to be Gaussian distribution where $\lambda = 5$ and the mean of each is 0.5. The details for pathological noises we use for the experiment is illustrated in Section 6.1 as well.

As shown in Figure 6.3, our algorithm effectively achieves $\sqrt{T}$ expected regret bound in all dimensions 3 and 5 with different type of noises.

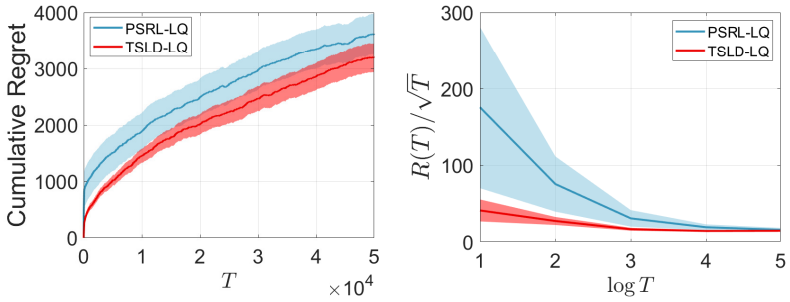

Figure 6.4: The comparison of expected cumulative regret $R(T)$ (left) and ratio over a time horizon T in comparison with PSRL-LQ [1] for $n = n_u = 3$ (right).

We also provide experimental evidence to emphasize the benefit of our algorithm. For this sake, the regret achieved by our algorithm is compared with obtained in [1], which is referred as PSRL-LQ. A critical assumption needed for this experiment is that the system noise follows from the Gaussian distribution as by no means the latter algorithm can be applied yet ours can handle general noises. For PSRL-LQ, the distribution of $\Theta(i)$ is assumed to be independent where $\Theta = \begin{bmatrix} \Theta(1) \cdots \Theta(n) \end{bmatrix}$ and $|\Theta| \leq S$ for some $S > 0$ so that each column is updated independently as PSRL-LQ algorithm proceeds. However, such a restriction is not required for our algorithm as long as $\nabla^2_\theta U_1(\theta) \succeq \lambda I_{dn}$ for $\lambda \geq 1$. It is also worth noting that PSRL-LQ requires that sampled system parameter θ be rejected based on the condition $|A_* + B_* K(\theta)| \leq \rho < 1$ for the true system parameters A_* and B_* whereas ours only imposes the condition $\theta \in \mathcal{S}$. For a fair comparison, we replace the rejection step in PSRL-LQ by the condition $\theta \in \mathcal{S}$. The figures in Figure 6.4 shows the superiority of our method compared to

PSRL-LQ as it always achieves lower regret. Furthermore, the ratio $\frac{R(T)}{\sqrt{T}}$ is maintained to be a constant as T increases.

6.3 Effect of preconditioner on number iterations

The computational advantage of our new method is corroborated empirically as seen in Figure 6.5. For naive ULA, one chooses the stepsize and number of iterations from Theorem 2 while preconditioned ULA chooses those based on Algorithm 1. We utilize the system parameter chosen at the beginning of this section and use the standard Gaussian distribution.

We observe a significant reduction in the number of iterations needed for the sampling process when preconditioned ULA is implemented compared to the naive ULA.

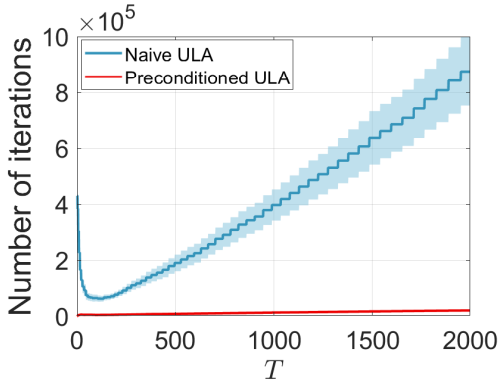

Figure 6.5: Comparison for the number of iterations over time horizon T between TSLD-LQ with naive ULA and preconditioned ULA. For naive ULA, we use the stepsize and the number of iterations in Theorem 2.

As shown in the figure, the number of iterations of naive ULA increases almost linearly and is remarkably greater than that of preconditioned ULA. Altogether, it is verified empirically that our algorithm achieves improved regret while using fewer computational resources.

Chapter 7

Conclusion

We propose a state-of-art computationally tractable Thompson sampling-based algorithm for learning LQR problems with the various classes of disturbance achieving $O(\sqrt{T})$ Bayesian regret bound. A salient feature of our method is that we not only drop the stabilizing compact set assumption but also the independence of columns of Θ by introducing preconditioned ULA and executing a perturbed control action only at the end of each episode. Several directions for future research can be proposed. Extending our algorithm to noises with non-convex potential is an important subject of study. As the log-concavity of the potential of posteriors is preserved even for noises we consider, acceleration of the sampling process was available. To handle more general noises, some different aspects of ULA should be explored. Additionally, we also address an open question on characterizing control gain matrices which induce more efficient learning of LQR problems.

Bibliography

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

[2] M. Gagrani, S. Sudhakara, A. Mahajan, A. Nayyar, and Y. Ouyang, "A modified Thompson sampling-based learning algorithm for unknown linear systems," in 2022 *IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 6658–6665.

[3] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[4] M. Kearns and S. Singh, "Near-optimal reinforcement learning in polynomial time," *Machine Learning*, vol. 49, no. 2, pp. 209–232, 2002.

[5] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3-4, pp. 285–294, 1933.

[6] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," in *Proceedings of the 25th Annual Conference on Learning Theory*. PMLR, 2012, pp. 39.1–26.

[7] S. Agrawal and N. Goyal,"Thompson sampling for contextual bandits with linear payoffs," in *International Confer ence on Machine Learning*. PMLR, 2013, pp. 127–135.

[8] E. Kaufmann, N. Korda, and R. Munos, "Thompson sampling: An asymptotically optimal finite-time analysis," in *International Conference on Algorithmic Learning Theory*. Springer, 2012, pp. 199–213.

[9] I. Osband, D. Russo, and B. Van Roy, "(More) efficient reinforcement learning via posterior sampling," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[10] I. Osband and B. Van Roy, "Posterior sampling for reinforcement learning without episodes," *arXiv preprint arXiv:1608.02731*, 2016.

[11] A. Gopalan and S. Mannor, "Thompson sampling for learning parameterized Markov decision processes," in *Proceedings of The 28th Conference on Learning Theory*. PMLR, 2015, pp. 861–898.

[12] Y. Abbasi-Yadkori and C. Szepesvári, "Bayesian optimal control of smoothly parameterized systems." in*Proceedings of 31st Conference on Uncertainty in Artificial Intelligence*. Citeseer, 2015, pp. 1–11.

[13] M. Abeille and A. Lazaric, "Thompson sampling for linear-quadratic control problems," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1246–1254.

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

[15] W. R. Gilks, S. Richardson, and D. Spiegelhalter, Markov Chain Monte Carlo in practice. *CRC press*, 1995.

[16] G. O. Roberts and R. L. Tweedie, "Exponential convergence of Langevin distributions and their discrete approximations," *Bernoulli*, pp. 341–363, 1996.

[17] A. Durmus and E. Moulines, "Sampling from a strongly log-concave distribution with the unadjusted Langevin algorithm," 2016.

[18] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient Langevin dynamics," in *International Conference on Machine Learning*. Citeseer, 2011, pp. 681–688.

[19] A. Durmus and E. Moulines, "Nonasymptotic convergence analysis for the unadjusted Langevin algorithm," *The Annals of Applied Probability*, vol. 27, no. 3, pp. 1551–1587, 2017

[20] W. Mou, N. Ho, M. J.Wainwright, P. Bartlett, and M. I. Jordan, "A diffusion process perspective on posterior contraction rates for parameters," *arXiv preprint arXiv:1909.00966*, 2019.

[21] V. Krishnamurthy and G. Yin, "Langevin dynamics for adaptive inverse reinforcement learning of stochastic gradient algorithms," *Journal of Machine Learning Research*, vol. 22, no. 121, pp. 1–49, 2021.

[22] X. Li, D. Wu, L. Mackey, and M. A. Erdogdu, "Stochastic Runge-Kutta accelerates Langevin Monte Carlo and beyond," *arXiv preprint arXiv:1906.07868*, 2019.

[23] W. Mou, Y.-A. Ma, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan, "High-order Langevin diffusion yields an accelerated MCMC algorithm," *arXiv preprint arXiv:1908.10859*, 2019.

[24] Z. Ding, Q. Li, J. Lu, and S. J. Wright, "Random coordinate Langevin Monte Carlo," in *Conference on Learning Theory*. PMLR, 2021, pp. 1683–1710.

[25] Y. Lu, J. Lu, and J. Nolen, "Accelerating Langevin sampling with birth-death," *arXiv preprint arXiv:1905.09863*, 2019.

[26] M. Girolami and B. Calderhead, "Riemann manifold Langevin and Hamiltonian Monte Carlo methods," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 2, pp. 123–214, 2011.

[27] A. S. Dalalyan, "Theoretical guarantees for approximate sampling from smooth and log-concave densities," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 79, no. 3, pp. 651–676, 2017.

[28] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems," in *Proceedings of the 24th Annual Conference on Learning Theory*. PMLR, 2011, pp. 19.1–26.

[29] M. Ibrahimi, A. Javanmard, and B. Roy, "Efficient reinforcement learning for high dimensional linear quadratic systems," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

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

[31] M. Abeille and A. Lazaric, "Efficient optimistic exploration in linear-quadratic regulators via 1025 Lagrangian relaxation," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1026 23–31.

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

[33] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, "Reinforcement learning with fast stabilization in linear dynamical systems," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 5354–5390.

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

[35] T. Kargin, S. Lale, K. Azizzadenesheli, A. Anandkumar, and B. Hassibi, "Thompson sampling achieves ˜O($\sqrt{T}$) regret in linear quadratic control," in *Conference on Learning Theory*. PMLR, 2022, pp. 3235–3284.

[36] Y. Jedra and A. Proutiere, "Minimal expected regret in linear quadratic control," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 10 234–10 321.

[37] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *Foundations of Computational Mathematics*, vol. 20, no. 4, pp. 633–679, 2020.

[38] D. Bertsekas, *Dynamic programming and optimal control: Volume II*. Athena Scientific, 2011.

[39] D. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z.Wen, "A tutorial on Thompson sampling," *arXiv preprint arXiv:1707.02038*, 2017.

[40] G. A. Pavliotis, *Stochastic processes and applications: Diffusion processes, the Fokker-Planck and Langevin equations*. Springer, 2014, vol. 60.

[41] G. O. Roberts and O. Stramer, "Langevin diffusions and Metropolis-Hastings algorithms," *Methodology and Computing in Applied Probability*, vol. 4, no. 4, pp. 337–357, 2002.

[42] N. Bou-Rabee and M. Hairer, "Nonasymptotic mixing of the MALA algorithm," *IMA Journal of Numerical Analysis*, vol. 33, no. 1, pp. 80–110, 2013.

[43] E. Mazumdar, A. Pacchiano, Y.-a. Ma, P. L. Bartlett, and M. I. Jordan, "On Thompson sampling with Langevin algorithms," *arXiv preprint arXiv:2002.10002*, 2020.

[44] C. Li, C. Chen, D. Carlson, and L. Carin, "Preconditioned stochastic gradient Langevin dynamics for deep neural networks," in *30th AAAI Conference on Artificial Intelligence*, 2016.

[45] J. Lu, Y. Lu, and Z. Zhou, "Continuum limit and preconditioned Langevin sampling of the path integral molecular dynamics," *Journal of Computational Physics*, vol. 423, p. 109788, 2020.

[46] P. Bras, "Langevin algorithms for very deep neural networks with application to image classifi cation,"*arXiv preprint arXiv:2212.14718*, 2022.

[47] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite time identification in unstable linear systems," *Automatica*, vol. 96, pp. 342–353, 2018.

[48] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," *Advances in Neural Information Processing Systems*, vol. 24, pp. 2312–2320, 2011.

[49] R. E. Bellman, *Dynamic programming*. Princeton university press, 2010.

[50] Y.-F. Ren, "On the Burkholder–Davis–Gundy inequalities for continuous martingales," *Statistics & Probability Letters*, vol. 78, no. 17, pp. 3034–3039, 2008.

[51] L. Lovász and S. Vempala, "Logconcave functions: Geometry and efficient sampling algorithms," in *44th Annual IEEE Symposium on Foundations of Computer Science*, 2003. Proceedings. IEEE, 2003, pp. 640–649.

[52] M. Ledoux, *The concentration of measure phenomenon*. American Mathematical Soc., 2001, no. 89.

[53] M. Ledoux, "Concentration of measure and logarithmic Sobolev inequalities," in *Seminaire de probabilites XXXIII*. Springer, 1999, pp. 120–216.

[54] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018, vol. 47.

Appendix A

Proof of Theorem 2

Lemma 2. *Suppose Assumptions 1 holds. Let $X \in \mathbb{R}^{n_x}$ be a random variable with probability density function $p(x) \propto e^{-U(x)}$ where $\lambda_{\min} I_{n_x} \preceq \nabla^2 U \preceq \lambda_{\max} I_{n_x}$ for $\lambda_{\max}, \lambda_{\min} > 0$. Set $\{Y_j\}$, $Y_j \in \mathbb{R}^{n_x}$ be generated by the ULA as*

$$Y_{j+1} = Y_j - \gamma \nabla U(Y_j) + \sqrt{2\gamma} W_j,$$

where Y_0 is a random variable with an arbitrary density function, $\gamma \leq \frac{\lambda_{\min}}{16\lambda_{\max}^2}$. Then, we have

$$\mathbb{E}[|Y_j - X|^2] < 2^{-\frac{\lambda_{\min}\gamma j}{4}} \mathbb{E}[|Y_0 - X|^2] + 2^8 \frac{n_x \lambda_{\max}^2}{\lambda_{\min}^2} \gamma,$$

where

Proof. Let $\{Z_\xi\}_{\xi \geq 0}$ be a continuous interpolation of $\{Y_j\}$, defined by

$$\begin{cases} \mathrm{d}Z_\xi = -\nabla U(Y_j)\mathrm{d}\xi + \sqrt{2}\mathrm{d}B_\xi & \text{for } \xi \in [j\gamma, (j+1)\gamma) \\ Z_\xi = Y_j & \text{for } \xi = j\gamma. \end{cases} \tag{A.1}$$

Note that $\lim_{\xi \nearrow j\gamma} Z_\xi = Y_j = \lim_{\xi \searrow j\gamma} Z_\xi$ for each j, and thus $\{Z_\xi\}$ is a continuous process. We introduce another stochastic process $\{X_\xi\}$, defined by

$$\mathrm{d}X_\xi = -\nabla U(X_\xi)\mathrm{d}\xi + \sqrt{2}\mathrm{d}B_\xi,$$

where X_0 is a random variable with pdf $p(x) \propto e^{-U(x)}$. By Lemma 3, X_ξ has the same pdf $p(x)$ for all ξ. We use the same Brownian motion B_ξ to define both $\{Z_\xi\}$ and

$\{X_\xi\}$. Fix an arbitrary j. Differentiating $|Z_\xi - X_\xi|^2$ with respect to $\xi \in [j\gamma, (j+1)\gamma)$, we have

$$\frac{\mathrm{d}|Z_\xi - X_\xi|^2}{\mathrm{d}\xi} = 2(Z_\xi - X_\xi)^\top \left(\frac{\mathrm{d}Z_\xi}{\mathrm{d}\xi} - \frac{\mathrm{d}X_\xi}{\mathrm{d}\xi}\right)$$
$$= 2(Z_\xi - X_\xi)^\top(-\nabla U(Y_j) + \nabla U(Z_\xi)) + 2(Z_\xi - X_\xi)^\top(-\nabla U(Z_\xi) + \nabla U(X_\xi)).$$

Therefore,

$$2(Z_\xi - X_\xi)^\top(-\nabla U(Y_j) + \nabla U(Z_\xi)) + 2(Z_\xi - X_\xi)^\top(-\nabla U(Z_\xi) + \nabla U(X_\xi))$$
$$\leq 2(Z_\xi - X_\xi)^\top(-\nabla U(Y_j) + \nabla U(Z_\xi)) - 2\lambda_{\min}(Z_\xi - X_\xi)^\top(Z_\xi - X_\xi)$$
$$= 2|Z_\xi - X_\xi|\nabla U(Z_\xi)) - \nabla U(Y_j)| - 2\lambda_{\min}|Z_\xi - X_\xi|^2.$$

where the second inequality follows from the strong log-concavity.

Using Young's inequality, we have

$$|Z_\xi - X_\xi||\nabla U(Z_\xi)) - \nabla U(Y_j)| \leq \frac{\lambda_{\min}|Z_\xi - X_\xi|_P^2}{2} + \frac{|\nabla U(Z_\xi)) - \nabla U(Y_j)|^2}{2\lambda_{\min}}.$$

It follows from the equality and the inequalities above that

$$\frac{\mathrm{d}|Z_\xi - X_\xi|^2}{\mathrm{d}\xi} \leq -\lambda_{\min}|Z_\xi - X_\xi|^2 + \frac{1}{\lambda_{\min}}|\nabla U(Z_\xi)) - \nabla U(Y_j)|^2,$$

which implies

$$\frac{\mathrm{d}}{\mathrm{d}\xi}(e^{\lambda_{\min}\xi}|Z_\xi - X_\xi|^2) \leq \frac{e^{\lambda_{\min}\xi}}{\lambda_{\min}}|\nabla U(Z_\xi)) - \nabla U(Y_j)|^2.$$

Integrating both sides from $j\gamma$ to $(j+1)\gamma$ and multiplying $e^{-\lambda_{\min}(j+1)\gamma}$, we have

$$|Z_{(j+1)\gamma} - X_{(j+1)\gamma}|^2 \leq e^{-\lambda_{\min}\gamma}|Z_{j\gamma} - X_{j\gamma}|^2 + \frac{1}{\lambda_{\min}}\int_{j\gamma}^{(j+1)\gamma} e^{-\lambda_{\min}((j+1)\gamma - s)}|\nabla U(Z_s)) - \nabla U(Y_j)|^2 \mathrm{d}s.$$

Since X_t and X have the same pdf by Lemma 3, we have

$$\mathbb{E}[|Z_{(j+1)\gamma} - X|^2] \tag{A.2}$$
$$\leq e^{-\lambda_{\min}\gamma}\mathbb{E}[|Z_{j\gamma} - X|^2] + \frac{1}{\lambda_{\min}}\int_{j\gamma}^{(j+1)\gamma}\mathbb{E}[|\nabla U(Z_s)) - \nabla U(Y_j)|^2]\mathrm{d}s$$
$$\leq e^{-\lambda_{\min}\gamma}\mathbb{E}[|Z_{j\gamma} - X|^2] + \frac{\lambda_{\max}^2}{\lambda_{\min}}\int_{j\gamma}^{(j+1)\gamma}\mathbb{E}[|(Z_s - Y_j)|^2]\mathrm{d}s, \tag{A.3}$$

where the first inequality follows from $e^{-\lambda_{\min}((j+1)\gamma - s)} \leq 1$ and the second inequality follows from the Lipschitz smoothness.

To bound (A.3), we handle the first and second terms separately. Regarding the second term, we first integrate the SDE (A.1) from $j\gamma$ to $s \in [j\gamma, (j+1)\gamma)$ to obtain

$$Z_s - Y_j = -(s - j\gamma)\nabla U(Y_j) + \sqrt{2}(B_s - B_{j\gamma}). \tag{A.4}$$

The second term of (A.3) can then be bounded by

$$\begin{aligned}
&\int_{j\gamma}^{(j+1)\gamma} \mathbb{E}[|Z_s - Y_j|^2]\mathrm{d}s = \int_{j\gamma}^{(j+1)\gamma} \mathbb{E}[|-(s-j\gamma)\nabla U(Y_j) + \sqrt{2}(B_s - B_{j\gamma})|^2]\mathrm{d}s \\
&\leq 2\bigg[\int_{j\gamma}^{(j+1)\gamma} \mathbb{E}[|(s-j\gamma)\nabla U(Y_j)|^2]\mathrm{d}s + 2\int_{j\gamma}^{(j+1)\gamma} \mathbb{E}[|B_s - B_{j\gamma}|^2]\mathrm{d}s\bigg].
\end{aligned} \tag{A.5}$$

For $s \in [j\gamma, (j+1)\gamma)$, we note that $|s - j\gamma| \leq \gamma$, and thus

$$\begin{aligned}
\int_{j\gamma}^{(j+1)\gamma} \mathbb{E}[|(s-j\gamma)\nabla U(Y_j)|^2]\mathrm{d}s &\leq \gamma^2 \int_{j\gamma}^{(j+1)\gamma} \mathbb{E}[|\nabla U(Y_j)|^2] \\
&= \gamma^3 \mathbb{E}[|\nabla U(Y_j)|^2] \\
&= \gamma^3 \mathbb{E}[|\nabla U(Y_j) - \nabla U(x_{\min})|^2] \\
&\leq \gamma^3 \lambda_{\max}^2 \mathbb{E}[|Y_j - x_{\min}|^2],
\end{aligned} \tag{A.6}$$

where $x_{\min}$ is a minimizer of potential U.

Then,

$$\begin{aligned}
\mathbb{E}[|Y_j - x_{\min}|^2] &\leq (\mathbb{E}[|Y_j - X|^2]^{\frac{1}{2}} + \mathbb{E}[|X - x_{\min}|^2]^{\frac{1}{2}})^2 \\
&\leq 2(\mathbb{E}[|Y_j - X|^2] + \mathbb{E}[|\tilde{X} - \tilde{x}_{\min}|^2]).
\end{aligned} \tag{A.7}$$

Applying Lemma 10 in [43],

$$\mathbb{E}[|Y_j - x_{\min}|^2] \leq 2\mathbb{E}[|Y_j - X|^2] + 10^2 \frac{n_x}{\lambda_{\min}}. \tag{A.8}$$

On the other hand, Lemma 8 in [43] yields

$$\int_{j\gamma}^{(j+1)\gamma} \mathbb{E}[|B_s - B_{j\gamma}|^2]\mathrm{d}s \leq \frac{4n_x}{e}\gamma^2. \tag{A.9}$$

Combining (A.5)–(A.9), we obtain that

$$\int_{j\gamma}^{(j+1)\gamma} \mathbb{E}[|Z_s - Y_j|^2]\mathrm{d}s \leq 2^2\lambda_{\max}^2\gamma^3\mathbb{E}[|Y_j - X|^2] + 2(10\lambda_{\max})^2\gamma^3\frac{n_x}{\lambda_{\min}} + \frac{16n_x}{e}\gamma^2$$
$$\leq 2^2\lambda_{\max}^2\gamma^3\mathbb{E}[|Y_j - X|^2] + 2^5 n_x\gamma^2, \tag{A.10}$$

where the second inequality follows from $\gamma \leq \frac{\lambda_{\min}}{16\lambda_{\max}^2}$.

Applying the result above to (A.3), we have

$$\mathbb{E}[|Z_{(j+1)\gamma} - X|^2] < e^{-\lambda_{\min}\gamma}\mathbb{E}[|Z_{j\gamma} - X|^2] + 2^2\frac{\lambda_{\max}^4}{\lambda_{\min}}\gamma^3\mathbb{E}[|Y_j - X|^2] + 2^5 n_x\frac{\lambda_{\max}^2}{\lambda_{\min}}\gamma^2$$
$$\leq \left(1 - \frac{\lambda_{\min}}{4}\gamma\right)^2\mathbb{E}[|Y_j - X|^2]$$
$$+ 2^2\frac{\lambda_{\max}^4}{\lambda_{\min}}\gamma^3\mathbb{E}[|Y_j - X|^2] + 2^5 n_x\frac{\lambda_{\max}^2}{\lambda_{\min}}\gamma^2, \tag{A.11}$$

where the second inequality follows from the fact that $e^{-x} \leq 1 - \frac{x}{2}$ for $x \in [0, 1]$. To further simplify the upper-bound, the following inequalities are needed:

$$2^2\frac{\lambda_{\max}^4}{\lambda_{\min}}\gamma^3 = \frac{\lambda_{\min}}{64}\left(\frac{16\lambda_{\max}^2}{\lambda_{\min}}\right)^2\gamma^3 \leq \frac{\lambda_{\min}}{64}\gamma,$$

and

$$\left(1 - \frac{\lambda_{\min}}{4}\gamma\right)^2 + \frac{\lambda_{\min}}{64}\gamma \leq \left(1 - \frac{\lambda_{\min}}{8}\gamma\right)^2.$$

Consequently, $\mathbb{E}[|Z_{(j+1)\gamma} - X|^2]$ is bounded as

$$\mathbb{E}[|Z_{(j+1)\gamma} - X|^2] < \left(1 - \frac{\lambda_{\min}}{8}\gamma\right)^2\mathbb{E}[|Y_j - X|^2] + 2^5 n_x\frac{\lambda_{\max}^2}{\lambda_{\min}}\gamma^2.$$

Invoking the bound repeatedly, we obtain

$$\begin{aligned}
&\mathbb{E}[|Z_{(j+1)\gamma} - X|^2] \\
&< \left(1 - \frac{\lambda_{\min}}{8}\gamma\right)^{2(j+1)} \mathbb{E}[|Y_0 - X|^2] + \sum_{i=0}^{j} \left(1 - \frac{\lambda_{\min}}{8}\gamma\right)^{2i} 2^5 n_x \frac{\lambda_{\max}^2}{\lambda_{\min}}\gamma^2 \\
&< \left(1 - \frac{\lambda_{\min}}{8}\gamma\right)^{2(j+1)} \mathbb{E}[|Y_0 - X|^2] + \frac{1}{1 - (1 - \frac{\lambda_{\min}}{8}\gamma)} 2^5 n_x \frac{\lambda_{\max}^2}{\lambda_{\min}}\gamma^2 \qquad \text{(A.12)} \\
&= \left(1 - \frac{\lambda_{\min}}{8}\gamma\right)^{2(j+1)} \mathbb{E}[|Y_0 - X|^2] + 2^8 n_x \frac{\lambda_{\max}^2}{\lambda_{\min}^2}\gamma \\
&\leq \left(1 - \frac{\lambda_{\min}}{8}\gamma\right)^{2(j+1)} \mathbb{E}[|Y_0 - X|^2] + 2^8 n_x \frac{\lambda_{\max}^2}{\lambda_{\min}^2}\gamma.
\end{aligned}$$

Since $(1 - \frac{\lambda_{\min}}{8}\gamma) \leq (\frac{1}{2})^{\frac{\lambda_{\min}}{8}\gamma}$, $Z_{(j+1)\gamma} = Y_{j+1}$, we conclude that

$$\mathbb{E}[|Y_{j+1} - X|^2] = \mathbb{E}[|Z_{(j+1)\gamma} - X|^2] < \left(\frac{1}{2}\right)^{\frac{\lambda_{\min}\gamma(j+1)}{4}} \mathbb{E}[|Y_0 - X|^2] + 2^8 \frac{n_x \lambda_{\max}^2}{\lambda_{\min}^2}\gamma.$$

Replacing $j + 1$ with j, the result follows. □

Proof of Theorem 2. We now prove Theorem 2. It follows from Lemma 10 in [43] that

$$\mathbb{E}_{x\sim p}\left[|x - x_{\min}|^2\right]^{\frac{1}{2}} \leq 5\sqrt{\frac{2n_x}{\lambda_{\min}}}, \qquad \text{(A.13)}$$

where $x_{\min}$ is a minimizer of U. Using Lemma 2 in with $n_x = dn$ and the initial distribution $X_0 \sim \delta(x_{\min})$ to obtain that

$$\mathbb{E}_{x\sim p, \tilde{x}\sim p_N}\left[|x - \tilde{x}|^2\right] < 2^{-\frac{\lambda_{\min}\gamma N}{4}} \mathbb{E}_{x\sim p}\left[|x - x_{\min}|^2\right] + 2^8 \frac{n_x \lambda_{\max}^2}{\lambda_{\min}^2}\gamma.$$

Taking the stepsize and the number of steps as $\gamma = \frac{\lambda_{\min}}{16\lambda_{\max}^2}$ and $N = \frac{64\lambda_{\max}^2}{\lambda_{\min}^2}$, respectively, the first and second terms in the inequality above is bounded as

$$\begin{aligned}
2^{-\frac{\lambda_{\min}\gamma N}{4}} \mathbb{E}_{x\sim p}\left[|x - x_{\min}|^2\right] &= \frac{1}{2}\mathbb{E}_{x\sim p}\left[|x - x_{\min}|^2\right] \\
&\leq 25\frac{n_x}{\lambda_{\min}},
\end{aligned}$$

and

$$2^8 \frac{n_x \lambda_{\max}^2}{\lambda_{\min}^2} \gamma \leq 2^4 \frac{n_x}{\lambda_{\min}},$$

respectively. Therefore, we have

$$\mathbb{E}_{x \sim p, \tilde{x} \sim p_N} \left[|x - \tilde{x}|^2 \right]^{\frac{1}{2}} < \sqrt{41 \frac{n_x}{\lambda_{\min}}} = O(\sqrt{\frac{1}{\lambda_{\min}}}).$$

□

Appendix B

Proof of Lemma 1

Proof. By direct calculation, the following holds:

$$\nabla_\theta^2 \log p_w(x_{s+1} - \Theta^\top z_s) = \nabla_{w_s}^2 \log p_w(x_{s+1} - \Theta^\top z_s) \otimes z_s z_s^\top,$$

where $\otimes$ denotes Kronecker product. Then, $\nabla_\theta^2 U_t$ is given by

$$\nabla_\theta^2 U_t = \lambda I_{dn} - \sum_{s=1}^{t-1} \nabla_{w_s}^2 \log p_w(x_{s+1} - \Theta^\top z_s) \otimes z_s z_s^\top.$$

By Assumption 1, for any state action pair $z_s = (x_s, u_s)$,

$$\underline{m}\text{blkdiag}(\{z_s z_s^\top\}_{i=1}^n) \preceq \nabla_{w_s}^2 \log p_w(x_{s+1} - \Theta^\top z_s) \otimes z_s z_s^\top \preceq \overline{m}\text{blkdiag}(\{z_s z_s^\top\}_{i=1}^n).$$

Then, we have

$$\min\{\underline{m}, 1\}\left(\lambda I_{dn} + \sum_{s=1}^{t-1} \text{blkdiag}(\{z_s z_s^\top\}_{i=1}^n)\right) \preceq \nabla_\theta^2 U_t,$$

and

$$\nabla_\theta^2 U_t \preceq \max\{\overline{m}, 1\}\left(\lambda I_{dn} + \sum_{s=1}^{t-1} \text{blkdiag}(\{z_s z_s^\top\}_{i=1}^n)\right).$$

Finally, letting the preconditioner $P_t = \lambda I_{dn} + \sum_{s=1}^{t-1} \text{blkdiag}(\{z_s z_s^\top\}_{i=1}^n)$, we obtain

$$m \preceq P_t^{-\frac{1}{2}} \nabla^2 U_t(\theta) P_t^{-\frac{1}{2}} \preceq M.$$

□

Appendix C

Details for Section 4.1

C.1 Proof of Proposition 1

To prove Proposition 1, we first introduce the following two lemmas regarding the stationarity of the preconditioned Langevin diffusion and the non-asymptotic behavior of the preconditioned ULA.

Lemma 3. *Suppose that Assumptions 1 holds. Let $X_\xi \in \mathbb{R}^{n_x}$ denote the solution of the preconditioned Langevin equation*

$$\mathrm{d}X_\xi = -P^{-1}\nabla U(X_\xi)\mathrm{d}\xi + \sqrt{2}P^{-\frac{1}{2}}\mathrm{d}B_\xi,$$

where X_0 is distributed according to $p(x) \propto e^{-U(x)}$, and $P \in \mathbb{R}^{n_x \times n_x}$ is an arbitrary positive definite matrix. Then, X_ξ has the same probability density $p(x)$ for all $\xi \geq 0$.

Proof. Consider the following Fokker-Planck equation associated with the preconditioned Langevin equation:

$$\frac{\partial q(x,\xi)}{\partial \xi} = -\sum_{i=1}^{n_x} \frac{\partial}{\partial x_i}([P^{-1}\nabla \log p(x)]_i q(x,\xi)) + \sum_{i=1}^{n_x}\sum_{j=1}^{n_x} \frac{\partial^2}{\partial x_i \partial x_j}([P^{-1}]_{ij} q(x,\xi)). \tag{C.1}$$

Then, it is well known that $q(x,\xi)$ is the probability density function of X_ξ. We can check that $p(x)$ is a solution of the Fokker-Planck equation by plugging $q(x,\xi) = p(x)$

into (C.1). Specifically,

$$\begin{aligned}
&-\sum_{i=1}^{n_x}\frac{\partial}{\partial x_i}([P^{-1}\nabla\log p(x)]_i p(x))+\sum_{i=1}^{n_x}\sum_{j=1}^{n_x}\frac{\partial^2}{\partial x_i\partial x_j}([P^{-1}]_{ij}p(x))\\
&=-\sum_{i=1}^{n_x}\frac{\partial}{\partial x_i}\left(\sum_{j=1}^{n_x}[P^{-1}]_{ij}\frac{\partial}{\partial x_j}p(x)\right)+\sum_{i=1}^{n_x}\sum_{j=1}^{n_x}\frac{\partial^2}{\partial x_i\partial x_j}([P^{-1}]_{ij}p(x)) \qquad \text{(C.2)}\\
&=0=\frac{\partial p(x)}{\partial\xi}.
\end{aligned}$$

Since the Fokker-Planck equation has a unique smooth solution [40], we conclude that $q(x,t)\equiv p(x)$ for all t, and the result follows. □

Lemma 4. *Suppose Assumptions 1 holds. Let $X\in\mathbb{R}^{n_x}$ be a random variable with probability density function $p(x)\propto e^{-U(x)}$, and $\{Y_j\}$, $Y_j\in\mathbb{R}^{n_x}$ be generated by the preconditioned ULA as*

$$Y_{j+1}=Y_j-\gamma P^{-1}\nabla U(Y_j)+\sqrt{2\gamma P^{-1}}W_j,$$

where Y_0 is a random variable with an arbitrary density function, $\gamma\le\frac{m\lambda_{\min}}{16M^2\max\{\lambda_{\min},t\}}$, and $P\in\mathbb{R}^{n_x^2}$ is a positive definite matrix such that $mI_{n_x}\preceq P^{-\frac{1}{2}}\nabla^2UP^{-\frac{1}{2}}\preceq MI_{n_x}$ and $\lambda_{\min}I_{n_x}\preceq P\preceq\lambda_{\max}I_{n_x}$. Then, we have

$$\mathbb{E}[|Y_j-X|_P^2]<2^{-\frac{m\gamma j}{4}}\mathbb{E}[|Y_0-X|_P^2]+2^8\frac{n_xM^2}{m^2}\gamma.$$

Proof. Let $\{Z_\xi\}_{\xi\ge0}$ be a continuous interpolation of $\{Y_j\}$, defined by

$$\begin{cases}\mathrm{d}Z_\xi=-P^{-1}\nabla U(Y_j)\mathrm{d}\xi+\sqrt{2P^{-1}}\mathrm{d}B_\xi & \text{for } \xi\in[j\gamma,(j+1)\gamma)\\ Z_\xi=Y_j & \text{for } \xi=j\gamma.\end{cases} \qquad \text{(C.3)}$$

Note that $\lim_{\xi\nearrow j\gamma}Z_\xi=Y_j=\lim_{\xi\searrow j\gamma}Z_\xi$ for each j, and thus $\{Z_\xi\}$ is a continuous process. We introduce another stochastic process $\{X_\xi\}$, defined by

$$\mathrm{d}X_\xi=-P^{-1}\nabla U(X_\xi)\mathrm{d}\xi+\sqrt{2}P^{-\frac{1}{2}}\mathrm{d}B_\xi,$$

where X_0 is a random variable with pdf $p(x)\propto e^{-U(x)}$. By Lemma 3, X_ξ has the same pdf $p(x)$ for all ξ. We use the same Brownian motion B_ξ to define both $\{Z_\xi\}$ and

$\{X_\xi\}$. Fix an arbitrary j. Differentiating $|Z_\xi - X_\xi|_P^p = |P^{\frac{1}{2}}(Z_\xi - X_\xi)|^p$ with respect to $\xi \in [j\gamma, (j+1)\gamma)$, we have

$$\begin{aligned}\frac{\mathrm{d}|Z_\xi - X_\xi|_P^p}{\mathrm{d}\xi} &= p|P^{\frac{1}{2}}(Z_\xi - X_\xi)|^{p-2}(Z_\xi - X_\xi)^\top P\left(\frac{\mathrm{d}Z_\xi}{\mathrm{d}\xi} - \frac{\mathrm{d}X_\xi}{\mathrm{d}\xi}\right)\\ &= p|P^{\frac{1}{2}}(Z_\xi - X_\xi)|^{p-2}(Z_\xi - X_\xi)^\top(-\nabla U(Y_j) + \nabla U(Z_\xi))\\ &\quad + p|P^{\frac{1}{2}}(Z_\xi - X_\xi)|^{p-2}(Z_\xi - X_\xi)^\top(-\nabla U(Z_\xi) + \nabla U(X_\xi)).\end{aligned}$$

Noting that $mI_{n_x} \preceq P^{-\frac{1}{2}}\nabla^2 U P^{-\frac{1}{2}} \preceq MI_{n_x}$, it follows that

$$\begin{aligned}&p|P^{\frac{1}{2}}(Z_\xi - X_\xi)|^{p-2}(Z_\xi - X_\xi)^\top(-\nabla U(Y_j) + \nabla U(Z_\xi))\\ &+ p|P^{\frac{1}{2}}(Z_\xi - X_\xi)|^{p-2}(Z_\xi - X_\xi)^\top(-\nabla U(Z_\xi) + \nabla U(X_\xi))\\ &\le p|P^{\frac{1}{2}}(Z_\xi - X_\xi)|^{p-2}(Z_\xi - X_\xi)^\top P^{\frac{1}{2}}P^{-\frac{1}{2}}(-\nabla U(Y_j) + \nabla U(Z_\xi))\\ &- pm|P^{\frac{1}{2}}(Z_\xi - X_\xi)|^{p-2}(Z_\xi - X_\xi)^\top P(Z_\xi - X_\xi)\\ &= p|P^{\frac{1}{2}}(Z_\xi - X_\xi)|^{p-2}\left(|Z_\xi - X_\xi|_P|P^{-\frac{1}{2}}\nabla U(Z_\xi)) - P^{-\frac{1}{2}}\nabla U(Y_j)| - m|Z_\xi - X_\xi|_P^2\right).\end{aligned}$$

where the first inequality follows from the mean value theorem.

Recall the generalized Young's inequality stating $ab \le \frac{s^\alpha a^\alpha}{\alpha} + \frac{s^{-\beta}b^\beta}{\beta}$ for $a, b, \alpha, \beta > 0$ and $s > 0$ and $\frac{1}{\alpha} + \frac{1}{\beta} = 1$. Choosing $s = (\frac{pm}{2(p-1)})^{(p-1)/p}$, $\alpha = \frac{p}{p-1}$, and $\beta = p$ we further derive that

$$\begin{aligned}&|Z_\xi - X_\xi|_P^{p-1}|P^{-\frac{1}{2}}\nabla U(Z_\xi)) - P^{-\frac{1}{2}}\nabla U(Y_j)|\\ &\le \frac{p-1}{p}\frac{pm}{2(p-1)}|Z_\xi - X_\xi|_P^p + \frac{1}{p}\frac{1}{(\frac{pm}{2(p-1)})^{p-1}}|P^{-\frac{1}{2}}\nabla U(Z_\xi)) - P^{-\frac{1}{2}}\nabla U(Y_j)|^p.\end{aligned}$$

Hence,

$$\frac{\mathrm{d}|Z_\xi - X_\xi|_P^p}{\mathrm{d}t} \le -\frac{pm}{2}|Z_\xi - X_\xi|_P^p + \frac{2^{p-1}}{m^{p-1}}|P^{-\frac{1}{2}}\nabla U(Z_\xi)) - P^{-\frac{1}{2}}\nabla U(Y_j)|^p,$$

as $\frac{pm}{2(p-1)} \ge \frac{m}{2}$. As a result,

$$\frac{\mathrm{d}}{\mathrm{d}\xi}(e^{\frac{pm}{2}\xi}|Z_\xi - X_\xi|_P^p) \le e^{\frac{pm}{2}\xi}\frac{2^{p-1}}{m^{p-1}}|P^{-\frac{1}{2}}\nabla U(Z_\xi)) - P^{-\frac{1}{2}}\nabla U(Y_j)|^p.$$

Integrating both sides from $j\gamma$ to $(j+1)\gamma$ and multiplying both sides by $e^{-\frac{pm}{2}(j+1)\gamma}$, we have

$$
\begin{aligned}
&|Z_{(j+1)\gamma} - X_{(j+1)\gamma}|_P^p \\
&\leq e^{-\frac{pm}{2}\gamma}|Z_{j\gamma} - X_{j\gamma}|_P^p \\
&+ \frac{2^{p-1}}{m^{p-1}} \int_{j\gamma}^{(j+1)\gamma} e^{-\frac{pm}{2}((j+1)\gamma - s)} |P^{-\frac{1}{2}}\nabla U(Z_s)) - P^{-\frac{1}{2}}\nabla U(Y_j)|^p \mathrm{d}s.
\end{aligned}
$$

Since X_ξ and X have the same pdf by Lemma 3, we have

$$
\begin{aligned}
&\mathbb{E}[|Z_{(j+1)\gamma} - X|_P^p] \\
&\leq e^{-\frac{pm}{2}\gamma}\mathbb{E}[|Z_{j\gamma} - X|_P^p] + \frac{2^{p-1}}{m^{p-1}} \int_{j\gamma}^{(j+1)\gamma} \mathbb{E}[|P^{-\frac{1}{2}}\nabla U(Z_s)) - P^{-\frac{1}{2}}\nabla U(Y_j)|^p]\mathrm{d}s \\
&= e^{-\frac{pm}{2}\gamma}\mathbb{E}[|Z_{j\gamma} - X|_P^p] \qquad \text{(C.4)} \\
&+ \frac{2^{p-1}}{m^{p-1}} \int_{j\gamma}^{(j+1)\gamma} \mathbb{E}[|P^{-\frac{1}{2}}(\int_0^1 \nabla^2 U(Y_j + t(Y_j - Z_s))\mathrm{d}t)(Z_s - Y_j)|^p]\mathrm{d}s \\
&\leq e^{-\frac{pm}{2}\gamma}\mathbb{E}[|Z_{j\gamma} - X|_P^p] \qquad \text{(C.5)} \\
&+ \frac{2^{p-1}}{m^{p-1}} \int_{j\gamma}^{(j+1)\gamma} \mathbb{E}[|P^{-\frac{1}{2}}(\int_0^1 \nabla^2 U(Y_j + t(Y_j - Z_s))\mathrm{d}t)P^{-\frac{1}{2}}|^p |P^{\frac{1}{2}}(Z_s - Y_j)|^p]\mathrm{d}s \\
&\leq e^{-\frac{pm}{2}\gamma}\mathbb{E}[|Z_{j\gamma} - X|_P^p] + \frac{2^{p-1}M^p}{m^{p-1}} \int_{j\gamma}^{(j+1)\gamma} \mathbb{E}[|P^{\frac{1}{2}}(Z_s - Y_j)|^p]\mathrm{d}s, \qquad \text{(C.6)}
\end{aligned}
$$

where the first inequality follows from $e^{-m((j+1)\gamma - s)} \leq 1$ and the second inequality follows from the mean value theorem and the last inequality follows from the assumption in the lemma. To bound (C.6), we handle the first and second terms separately.

For the second term, we integrate (C.3) from $j\gamma$ to $s \in [j\gamma, (j+1)\gamma)$ to obtain

$$
Z_s - Y_j = -(s - j\gamma)P^{-1}\nabla U(Y_j) + \sqrt{2P^{-1}}(B_s - B_{j\gamma}). \qquad \text{(C.7)}
$$

Ignoring the constant coefficient, the second term of (C.6) is then bounded by

$$\begin{aligned}
&\int_{j\gamma}^{(j+1)\gamma} \mathbb{E}[|P^{\frac{1}{2}}(Z_s - Y_j)|^p]\mathrm{d}s \\
&= \int_{j\gamma}^{(j+1)\gamma} \mathbb{E}[|-(s-j\gamma)P^{-\frac{1}{2}}\nabla U(Y_j) + \sqrt{2}(B_s - B_{j\gamma})|^p]\mathrm{d}s \\
&\leq 2^{p-1}\Bigg[\int_{j\gamma}^{(j+1)\gamma} \mathbb{E}[|(s-j\gamma)P^{-\frac{1}{2}}\nabla U(Y_j)|^p]\mathrm{d}s + 2^{p/2}\int_{j\gamma}^{(j+1)\gamma} \mathbb{E}[|B_s - B_{j\gamma}|^p]\mathrm{d}s\Bigg].
\end{aligned} \tag{C.8}$$

For $s \in [j\gamma, (j+1)\gamma)$, we note that $|s - j\gamma| \leq \gamma$, and thus

$$\begin{aligned}
\int_{j\gamma}^{(j+1)\gamma} \mathbb{E}[|(s-j\gamma)P^{-\frac{1}{2}}\nabla U(Y_j)|^p]\mathrm{d}s &\leq \gamma^p \int_{j\gamma}^{(j+1)\gamma} \mathbb{E}[|P^{-\frac{1}{2}}\nabla U(Y_j)|^p] \\
&= \gamma^{p+1}\mathbb{E}[|P^{-\frac{1}{2}}\nabla U(Y_j)|^p] \\
&= \gamma^{p+1}\mathbb{E}[|P^{-\frac{1}{2}}\nabla U(Y_j) - P^{-\frac{1}{2}}\nabla U(x_{\min})|^p] \\
&\leq \gamma^{p+1}M^p\mathbb{E}[|Y_j - x_{\min}|_P^p],
\end{aligned} \tag{C.9}$$

where $x_{\min}$ is a minimizer of potential U.

Let $\tilde{X} = P^{\frac{1}{2}}X$ and denote the distribution of $\tilde{X}$ by $\tilde{p}(\tilde{x})$, i.e., $\tilde{X} \sim \tilde{p}(\tilde{x})$,

$$\mathbb{E}[|Y_j - x_{\min}|_P^p] \leq 2^{p-1}(\mathbb{E}[|Y_j - X|_P^p] + \mathbb{E}[|\tilde{X} - \tilde{x}_{\min}|^p]). \tag{C.10}$$

Note first that $\tilde{p}(\tilde{x}) = \det(P^{-\frac{1}{2}})p(P^{-\frac{1}{2}}\tilde{x})$.

Hence, $-\nabla_{\tilde{x}}^2 \log \tilde{p}(\tilde{x}) = -P^{-\frac{1}{2}}\nabla_x^2 \log p(P^{-\frac{1}{2}}\tilde{x})P^{-\frac{1}{2}}$ which is m-strongly convex with respect to $\tilde{x}$, one can apply Lemma 10 in [43]. As a consequence,

$$\mathbb{E}[|Y_j - x_{\min}|_P^p] \leq 2^{p-1}\mathbb{E}[|Y_j - X|_P^p] + \frac{10^p}{2}\big(\frac{pn_x}{m}\big)^{p/2}. \tag{C.11}$$

On the other hand, Lemma 8 in [43] yields that

$$\int_{j\gamma}^{(j+1)\gamma} \mathbb{E}[|B_s - B_{j\gamma}|^p]\mathrm{d}s \leq 2\big(\frac{pn_x}{e}\big)^{p/2}\gamma^{p/2+1}. \tag{C.12}$$

Combining (C.8)–(C.12), we obtain that

$$\begin{aligned}
&\int_{j\gamma}^{(j+1)\gamma} \mathbb{E}[|Z_s - Y_j|_P^p]\mathrm{d}s \\
&\leq 2^{2p-2} M^p \gamma^{p+1} \mathbb{E}[|Y_j - X|_P^p] + 2^{p-2}(10M)^p \gamma^{p+1} \big(\frac{pn_x}{m}\big)^{p/2} \\
&+ 2^{3p/2} \big(\frac{pn_x}{e}\big)^{p/2} \gamma^{p/2+1} \\
&\leq 2^{2p-2} M^p \gamma^{p+1} \mathbb{E}[|Y_j - X|_P^p] + 2^{3p} (pn_x)^{p/2} \gamma^{p/2+1},
\end{aligned} \tag{C.13}$$

where the second inequality follows from $\gamma \leq \frac{m\lambda_{\min}}{16M^2 \max\{\lambda_{\min}, t\}} \leq \frac{m}{16M^2}$.

Consequently, applying the result above to (C.6), we have

$$\begin{aligned}
&\mathbb{E}[|Z_{(j+1)\gamma} - X|_P^p] \\
&\leq e^{-\frac{pm}{2}\gamma} \mathbb{E}[|Z_{j\gamma} - X|_P^p] + 2^{3p-3} \frac{M^{2p}}{m^{p-1}} \gamma^{p+1} \mathbb{E}[|Y_j - X|_P^p] \\
&+ 2^{4p-1} (pn_x)^{p/2} \frac{M^p}{m^{p-1}} \gamma^{p/2+1}.
\end{aligned}$$

To further simplify the bound, we modify the coefficient as

$$2^{3p-3} \frac{M^{2p}}{m^{p-1}} \gamma^{p+1} = \frac{m}{2^{p+3}} \left(\frac{16M^2 \max\{\lambda_{\min}, t\}}{m\lambda_{\min}} \right)^p \left(\frac{\lambda_{\min}}{\max\{\lambda_{\min}, t\}} \right)^p \gamma^{p+1} \leq \frac{m}{32}\gamma,$$

and

$$\begin{aligned}
e^{-\frac{pm}{2}\gamma} + \frac{m}{32}\gamma &\leq e^{-m\gamma} + \frac{m}{32}\gamma \leq 1 - \frac{m}{2}\gamma + \frac{m}{32}\gamma \\
&< 1 - \frac{m}{4}\gamma,
\end{aligned}$$

where the second inequality follows from the fact that $e^{-x} \leq 1 - \frac{x}{2}$ for $x \in [0, 1]$. Consequently, $\mathbb{E}[|Z_{(j+1)\gamma} - X|_P^p]$ is bounded as

$$\mathbb{E}[|Z_{(j+1)\gamma} - X|_P^p] < \left(1 - \frac{m}{4}\gamma\right) \mathbb{E}[|Y_j - X|_P^p] + 2^{4p-1} (pn_x)^{p/2} \frac{M^p}{m^{p-1}} \gamma^{p/2+1}.$$

Invoking the bound repeatedly, we obtain that

$$\begin{aligned}
&\mathbb{E}[|Z_{(j+1)\gamma} - X|_P^p] \\
&< \left(1 - \frac{m}{4}\gamma\right)^{(j+1)} \mathbb{E}[|Y_0 - X|_P^p] + \sum_{i=0}^{j} \left(1 - \frac{m}{4}\gamma\right)^{i} 2^{4p-1}(pn_x)^{p/2} \frac{M^p}{m^{p-1}} \gamma^{p/2+1} \\
&< \left(1 - \frac{m}{4}\gamma\right)^{(j+1)} \mathbb{E}[|Y_0 - X|_P^p] + \frac{1}{1 - (1 - \frac{m}{4}\gamma)} 2^{4p-1}(pn_x)^{p/2} \frac{M^p}{m^{p-1}} \gamma^{p/2+1} \\
&= \left(1 - \frac{m}{4}\gamma\right)^{(j+1)} \mathbb{E}[|Y_0 - X|_P^p] + 2^{4p+1}(pn_x)^{p/2} \frac{M^p}{m^p} \gamma^{p/2}.
\end{aligned}$$

Since $(1 - \frac{m}{4}\gamma) \leq (\frac{1}{2})^{\frac{m}{4}\gamma}$, $Z_{(j+1)\gamma} = Y_{j+1}$, we conclude that

$$\begin{aligned}
&\mathbb{E}[|Y_{j+1} - X|_P^p] \\
&= \mathbb{E}[|Z_{(j+1)\gamma} - X|_P^p] \\
&< \left(\frac{1}{2}\right)^{\frac{m\gamma(j+1)}{4}} \mathbb{E}[|Y_0 - X|_P^p] + 2^{4p+1}(pn_x)^{p/2} \frac{M^p}{m^p} \gamma^{p/2}.
\end{aligned}$$

Replacing $j + 1$ with j, the result follows. □

Proof of Proposition 1. We now prove Proposition 1. For simplicity, we use the following notation throughout the proof. For a positive definite matrix P,

$$E_P^p(\mu, \tilde{\mu}|h) := \mathbb{E}_{x\sim\mu, \tilde{x}\sim\tilde{\mu}}[|x - \tilde{x}|_P^p|h],$$

and define $\lambda_{\max,t}$, $\lambda_{\min,t}$ be the maximum, minimum eigenvalues of P_t.

Once again it follows from Lemma 10 in [43] that

$$E_{P_t}^p(\mu_t, \delta(\theta_{\min,t})|h_t) \leq 5^p \left(\frac{pdn}{m}\right)^{\frac{p}{2}} \tag{C.14}$$

for all t since μ_t's are m-strongly log-concave. Here $\theta_{\min,t}$ is a minimizer of U_t.

Then, we use Lemma 4 with $n_x = dn$ and the initial distribution $\theta_0 \sim \delta(\theta_{\min,t})$ in Algorithm 1 to obtain that

$$E_{P_t}^p(\mu_t, \tilde{\mu}_t|h_t) < 2^{-\frac{m\gamma_t N_t}{4}} E_{P_t}^p(\mu_t, \delta(\theta_{\min,t})|h_t) + 2^{4p+1}(pn_x)^{p/2} \frac{M^p}{m^p} \gamma_t^{p/2}.$$

In Algorithm 1, the stepsize and number of iterations are chosen to be $\gamma_t = \frac{m\lambda_{\min,t}}{16M^2 \max\{\lambda_{\min,t},t\}}$ and $N_t = \frac{4\log_2(\max\{\lambda_{\min,t},t\}/\lambda_{\min,t})}{m\gamma_t}$. Thus, the first and second term in the inequality above are bounded as

$$2^{-\frac{\gamma_t m N_t}{4}} E^p_{P_t}(\mu_t, \delta(\theta_{\min,t})|h_t) = 2^{-\log_2(\max\{\lambda_{\min,t},t\}/\lambda_{\min,t})} E^p_{P_k}(\mu_t, \delta(\theta_{\min,t})|h_t)$$
$$\leq 5^p \left(\frac{pdn}{m}\right)^{p/2} \left(\frac{\lambda_{\min,t}}{\max\{\lambda_{\min,t},t\}}\right),$$

and

$$2^{4p+1}(pn_x)^{p/2}\frac{M^p}{m^p}\gamma_t^{p/2} \leq 2^{2p+1}\frac{(pdn)^{p/2}}{m^{\frac{p}{2}}}\left(\frac{\lambda_{\min,t}}{\max\{\lambda_{\min,t},t\}}\right)^{\frac{p}{2}}.$$

Therefore, we have

$$E^p_{P_k}(\mu_t, \tilde{\mu}_t|h_t) < \left(\frac{pdn}{m}\right)^{\frac{p}{2}}\left(5^p\frac{\lambda_{\min,t}}{\max\{\lambda_{\min,t},t\}} + 2^{2p+1}\left(\frac{\lambda_{\min,t}}{\max\{\lambda_{\min,t},t\}}\right)^{\frac{p}{2}}\right).$$

Finally,

$$\begin{aligned}
&\mathbb{E}_{\theta\sim\mu_t,\theta'\sim\tilde{\mu}_t}[|\theta-\tilde{\theta}|^p_{P_t} \mid h_t] \\
&\leq \left(\frac{pdn}{m}\right)^{\frac{p}{2}}\left(5^p\frac{\lambda_{\min,t}}{\max\{\lambda_{\min,t},t\}} + 2^{2p+1}\left(\frac{\lambda_{\min,t}}{\max\{\lambda_{\min,t},t\}}\right)^{\frac{p}{2}}\right) \\
&\leq \left(\frac{pdn}{m}\right)^{\frac{p}{2}}\left(2^{2p+1}+5^p\right).
\end{aligned}$$

For a special case when $p = 2$, a simpler bound is achieved. Noting that

$$\lambda_{\min,t}\mathbb{E}_{\theta_t\sim\mu_t,\tilde{\theta}_t\sim\tilde{\mu}_t}[|\theta_t-\tilde{\theta}_t|^2 \mid h_t] \leq E^2_{P_t}(\mu_t, \tilde{\mu}_t \mid h_t),$$

one can deduce that

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

where $D = 114\frac{dn}{m}$. □

C.2 Proof of Proposition 2

Proof of Proposition 2. Let $\theta_\xi \in \mathbb{R}^{dn}$ denote the solution of the following stochastic differential equation:

$$d\theta_\xi = -P_t^{-1}\nabla U_t(\theta_\xi)d\xi + \sqrt{2}P_t^{-\frac{1}{2}}dB_\xi,$$

where $P_t = \lambda I_{dn} + \sum_{s=1}^{t-1} blkdiag(\{z_s z_s^\top\}_{i=1}^n)$ and $U_t = U_1 + U_t'$ for $U_t' = \sum_{s=1}^{t-1} \log p_w(x_{s+1} - \Theta^\top z_s | z_s, \theta)$. Define $V(\theta_\xi)$ as

$$V(\theta_\xi) = \frac{1}{2} e^{\alpha\xi} |\theta_\xi - \theta_*|_{P_t}^2,$$

for $\alpha > 0$ fixed. Applying Ito's lemma to $V(\theta_\xi)$, we have

$$V(\theta_\xi) = F_1 + F_2 + F_3,$$

where

$$\begin{aligned}
F_1 &= \int_0^\xi e^{\alpha\eta} \nabla_\theta U_t(\theta_\eta)^\top (\theta_* - \theta_\eta) d\eta + \frac{\alpha}{2} \int_0^t e^{\alpha\eta} |\theta_\eta - \theta_*|_{P_t}^2 d\eta, \\
F_2 &= \frac{dn}{2} \int_0^\xi e^{\alpha\eta} d\eta, \\
F_3 &= \int_0^\xi e^{\alpha\eta} (\theta_\eta - \theta_*)^\top P_t^{\frac{1}{2}} dB_\eta.
\end{aligned}$$

To bound F_1, we expand as following.

$$\begin{aligned}
F_1 &= \frac{1}{2} \int_0^\xi e^{\alpha\eta} \nabla_\theta U_t(\theta_\eta)^\top (\theta_* - \theta_\eta) d\eta + \frac{\alpha}{2} \int_0^\xi e^{\alpha\eta} |\theta_\eta - \theta_*|_{P_t}^2 d\eta \\
&= -\frac{1}{2} \int_0^\xi e^{\alpha\eta} (\nabla_\theta U_t(\theta_\eta) - \nabla_\theta U_t(\theta_*))^\top (\theta_\eta - \theta_*) d\eta + \frac{\alpha}{2} \int_0^\xi e^{\alpha\eta} |\theta_\eta - \theta_*|_{P_t}^2 d\eta \\
&\quad + \frac{1}{2} \int_0^\xi e^{\alpha\eta} \nabla_\theta U_1(\theta_*)^\top (\theta_* - \theta_\eta) d\eta + \frac{1}{2} \int_0^\xi e^{\alpha\eta} \nabla_\theta U_t'(\theta_*)^\top (\theta_* - \theta_\eta) d\eta \\
&\leq -\frac{m}{2} \int_0^\xi e^{\alpha\eta} (\theta_\eta - \theta_*)^\top P_t (\theta_\eta - \theta_*) d\eta + \frac{\alpha}{2} \int_0^\xi e^{\alpha\eta} |\theta_\eta - \theta_*|_P^2 d\eta \\
&\quad + \frac{1}{2} \int_0^\xi e^{\alpha\eta} \nabla_\theta U_1(\theta_*)^\top (\theta_* - \theta_\eta) d\eta + \frac{1}{2} \int_0^\xi e^{\alpha\eta} \nabla_\theta U_t'(\theta_*)^\top (\theta_* - \theta_\eta) d\eta \\
&\leq \frac{\alpha - m}{2} \int_0^\xi e^{\alpha\eta} |\theta_\eta - \theta_*|_{P_t}^2 d\eta + \frac{1}{2} \int_0^\xi e^{\alpha\eta} \nabla_\theta U_1(\theta_*)^\top (\theta_* - \theta_\eta) d\eta \\
&+ \frac{1}{2} \int_0^\xi e^{\alpha\eta} \nabla_\theta U_t'(\theta_*)^\top (\theta_* - \theta_\eta) d\eta,
\end{aligned}$$

To bound the second and third terms on the right-hand side, we invoke Young's in-

equality, which yields that

$$\begin{aligned}
&\int_0^\xi e^{\alpha\eta}\nabla_\theta U_1(\theta_*)^\top(\theta_*-\theta_\eta)d\eta \\
&\le \int_0^\xi e^{\alpha\eta}|P_t^{-\frac{1}{2}}\nabla_\theta U_1(\theta_*)||P_t^{\frac{1}{2}}(\theta_*-\theta_\eta)|d\eta \\
&\le \frac{1}{m}\int_0^\xi e^{\alpha\eta}|P_t^{-\frac{1}{2}}\nabla_\theta U_1(\theta_*)|^2 d\eta + \frac{m}{4}\int_0^\xi e^{\alpha\eta}|\theta_*-\theta_\eta|^2_{P_t}d\eta.
\end{aligned}$$

and

$$\begin{aligned}
&\int_0^\xi e^{\alpha\eta}\nabla_\theta U_t'(\theta_*)^\top(\theta_*-\theta_\eta)d\eta \\
&\le \int_0^\xi e^{\alpha\eta}|P_t^{-\frac{1}{2}}\nabla_\theta U_t'(\theta_*)||P_t^{\frac{1}{2}}(\theta_*-\theta_\eta)|d\eta \\
&\le \frac{1}{m}\int_0^\xi e^{\alpha\eta}|P_t^{-\frac{1}{2}}\nabla_\theta U_t'(\theta_*)|^2 d\eta + \frac{m}{4}\int_0^\xi e^{\alpha\eta}|\theta_*-\theta_\eta|^2_{P_t}d\eta.
\end{aligned}$$

Putting altogether,

$$\begin{aligned}
F_1 \le{}& \frac{2\alpha-m}{4}\int_0^\xi e^{\alpha\eta}|\theta_\eta-\theta_*|^2_{P_t}d\eta + \frac{1}{2m}\int_0^\xi e^{\alpha\eta}|P_t^{-\frac{1}{2}}\nabla_\theta U_1(\theta_*)|^2 d\eta \\
&+ \frac{1}{2m}\int_0^\xi e^{\alpha\eta}|P_t^{-\frac{1}{2}}\nabla_\theta U_t'(\theta_*)|^2 d\eta.
\end{aligned}$$

Choosing $\alpha = \frac{m}{2}$, we obtain

$$\begin{aligned}
F_1 &\le \frac{1}{2m}\int_0^\xi e^{\alpha\eta}|P_t^{-\frac{1}{2}}\nabla_\theta U_1(\theta_*)|^2 d\eta + \frac{1}{2m}\int_0^\xi e^{\alpha\eta}|P_t^{-\frac{1}{2}}\nabla_\theta U_t'(\theta_*)|^2 d\eta \\
&\le Ce^{\alpha\xi} + \frac{1}{2m}\int_0^\xi e^{\alpha\eta}|P_t^{-\frac{1}{2}}\nabla_\theta U_t'(\theta_*)|^2 d\eta.
\end{aligned}$$

On the other hand, F_2 is bounded as

$$F_2 = \frac{dn}{2}\int_0^\xi e^{\alpha\eta}d\eta = \frac{dn}{2\alpha}(e^{\alpha\xi}-1) \le \frac{dn}{2\alpha}e^{\alpha\xi} = \frac{dn}{m}e^{\alpha\xi}.$$

The last term F_3 is bounded as follows. By Burkholder-Davis-Gundy inequality

[50], for $\Xi > 0$ fixed,

$$\begin{aligned}
\mathbb{E}[\sup_{0\le\xi\le\Xi}|F_3|] &\le 2\sqrt{2}\mathbb{E}\left[\left(\int_0^{\Xi} e^{2\alpha\eta}|\theta_\eta-\theta_*|^2_{P_t}d\eta\right)^{\frac{1}{2}}\right] \\
&\le 2\sqrt{2}\mathbb{E}\left[\left(\sup_{0\le\xi\le\Xi} e^{\alpha\xi}|\theta_\xi-\theta_*|^2_{P_t}\int_0^{\Xi} e^{\alpha\eta}d\xi\right)^{\frac{1}{2}}\right] \\
&= 2\sqrt{2}\mathbb{E}\left[\left(\sup_{0\le\xi\le\Xi} e^{\alpha\xi}|\theta_\xi-\theta_*|^2_{P_t}\left(\frac{e^{\alpha\Xi}-1}{\alpha}\right)\right)^{\frac{1}{2}}\right] \\
&\le \mathbb{E}\left[\left(\frac{8e^{\alpha\Xi}}{\alpha}\right)^{\frac{1}{2}}(\sup_{0\le\xi\le\Xi} e^{\alpha\xi}|\theta_\xi-\theta_*|^2_{P_t})^{\frac{1}{2}}\right],
\end{aligned}$$

where the expectation is taken with respect to θ_ξ. By Young's inequality,

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{8e^{\alpha\Xi}}{\alpha}\right)^{\frac{1}{2}}(\sup_{0\le\xi\le\Xi} e^{\alpha\xi}|\theta_\xi-\theta_*|^2_{P_t})^{\frac{1}{2}}\right] &\le \mathbb{E}\left[\frac{8e^{\alpha\Xi}}{\alpha}+\frac{1}{4}(\sup_{0\le\xi\le\Xi} e^{\alpha\xi}|\theta_\xi-\theta_*|^2_{P_t})\right] \\
&= 16me^{\alpha\Xi}+\frac{1}{2}\mathbb{E}\left[\sup_{0\le\xi\le\Xi} V(\theta_\xi)\right].
\end{aligned}$$

Finally,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{0\le\xi\le\Xi} V(\theta_\xi)\right] \\
&= \mathbb{E}\left[\sup_{0\le\xi\le\Xi}(F_1+F_2+F_3)\right] \\
&\le \mathbb{E}\left[\sup_{0\le\xi\le\Xi} F_1\right]+\mathbb{E}\left[\sup_{0\le t\le\Xi} F_2\right]+\mathbb{E}\left[\sup_{0\le t\le\Xi} F_3\right] \\
&\le \mathbb{E}\left[\left(C+\frac{1}{m^2}|P_t^{-\frac{1}{2}}\nabla_\theta U_t'(\theta_*)|^2+\frac{dn}{m}+16m\right)\right]e^{\alpha\Xi}+\frac{1}{2}\mathbb{E}\left[\sup_{0\le t\le\Xi} V(\theta_\xi)\right].
\end{aligned} \tag{C.15}$$

Here, we use different C whenever it appears but it only depends on m, d, n and the prior U_1. Rearranging with respect to $\mathbb{E}\left[\sup_{0\le\xi\le\Xi} V(\theta_\xi)\right]$,

$$\mathbb{E}\left[\sup_{0\le\xi\le\Xi} V(\theta_\xi)\right] \le 2\left(C+\frac{1}{m^2}|P_t^{-\frac{1}{2}}\nabla_\theta U_t'(\theta_*)|^2+\frac{dn}{m}+16m\right)e^{\alpha\Xi}.$$

Then,

$$\begin{aligned}\mathbb{E}[|\theta_\Xi - \theta_*|_{P_t}|h_t] &= \mathbb{E}[\sqrt{2}e^{-\frac{1}{2}\alpha\Xi}V(\theta_\Xi)^{\frac{1}{2}}] \\ &\leq \sqrt{2}e^{-\frac{1}{2}\alpha\Xi}\left(\mathbb{E}[\sup_{0\leq\xi\leq\Xi} V(\theta_{\Xi)}]\right)^{\frac{1}{2}} \\ &\leq 2\sqrt{\left(C + \frac{1}{m^2}|P_t^{-\frac{1}{2}}\nabla_\theta U_t'(\theta_*)|^2 + \frac{dn}{m} + 16m\right)}.\end{aligned}$$

Taking the limit $\Xi \to \infty$ and using Fatou's Lemma, we have

$$\mathbb{E}_{\theta_t\sim\mu_t}[|\theta_t - \theta_*|_{P_t}|h_t] \leq 2\sqrt{\left(C + \frac{1}{m^2}|P_t^{-\frac{1}{2}}\nabla_\theta U_t'(\theta_*)|^2 + \frac{dn}{m} + 16m\right)}.$$

For a random vector X following log-concave distribution, Theorem 5.22 in [51] yields that

$$\mathbb{E}[|X|^p]^{\frac{1}{p}} \leq 2p\mathbb{E}[|X|]$$

for any $p > 0$. Observe that $y := P_t^{\frac{1}{2}}(\theta_t - \theta_*)$ is a random vector from a log-concave distribution since its potential $U_t(P_t^{-\frac{1}{2}}y + \theta_*)$ is convex. Therefore, it follows that

$$\begin{aligned}\mathbb{E}_{\theta_t\sim\mu_t}[|\theta_t - \theta_*|_{P_t}^p|h_t] &\leq (2p)^p\mathbb{E}_{\theta_t\sim\mu_t}[|\theta_t - \theta_*|_{P_t}|h_t]^p \\ &\leq (2p)^p\left(C + \frac{4}{m^2}|P_t^{-\frac{1}{2}}\nabla_\theta U_t'(\theta_*)|^2 + \frac{4dn}{m} + 64m\right)^{\frac{p}{2}}.\end{aligned} \tag{C.16}$$

To proceed let us define $Z := \begin{bmatrix} z_1 & \dots & z_t \end{bmatrix}^\top$. Noting that $\frac{\partial U_t'(\theta_*)}{\partial \Theta_{ij}} = -\sum_{t=1}^T Z_{ti}\frac{\partial \log p_w(w_t)}{\partial w_t(j)}$ where the j-th component of noise w_t is denoted by $w_t(j)$. Therefore, P_t can be written as $P_t = \lambda I_{dn} + blkdiag\{Z^\top Z\}_{i=1}^n = I_n \otimes (Z^\top Z + \lambda I_d)$, and it is straightforward to verify that $P_t^{-1} = I_n \otimes (Z^\top Z + \lambda I_d)^{-1}$.

Denoting by $\theta_\ell := \Theta_{ij}$ for $\ell = (j-1)d + i$, we deduce that

$$
\begin{aligned}
|P_t^{-\frac{1}{2}} \nabla_\theta U_t'(\theta_*)|^2 &= \sum_{\ell,k=1}^{dn} \frac{\partial U_t'(\theta_*)}{\partial \theta_\ell} (P_t)^{-1}_{\ell k} \frac{\partial U_t'(\theta_*)}{\partial \theta_k} \\
&= \sum_{i',i=1}^{d} \sum_{j',j=1}^{n} \frac{\partial U_t'(\theta_*)}{\partial \Theta_{i'j'}} P^{-1}_{(j'-1)d+i',(j-1)d+i} \frac{\partial U_t'(\theta_*)}{\partial \Theta_{ij}} \\
&\leq \sum_{j=1}^{n} \sum_{s',s=1}^{t-1} \frac{\partial \log p_w(w_{s'})}{\partial w_{s'}(j)} (Z(Z^\top Z + \lambda I_d)^{-1} Z^\top)_{s's} \frac{\partial \log p_w(w_s)}{\partial w_s(j)}.
\end{aligned}
$$

Recall first that $v^\top \nabla_w \log p_w(w_t)$ is a $\frac{M}{\sqrt{m}}$-sub-Gaussian random variable (Proposition 2.18 in [52]).

We are now ready to leverage the self-normalization technique, Lemma 6. For each j fixed, we take $X_s = z_s$ and $V_t = \lambda I_d + \sum_{s=1}^{t-1} z_s z_s^\top$, $S_t = \sum_{s=1}^{t-1} \frac{\partial \log p_w(w_s)}{\partial w_s(j)} z_s$ and the probability bound δ to be $\frac{\delta}{n}$ in the statement of the lemma. Consequently, we derive that

$$
\begin{aligned}
&\sum_{s,s'=1}^{t-1} \frac{\partial \log p_w(w_{s'})}{\partial w_{s'}(j)} (Z(Z^\top Z + \lambda I_d)^{-1} Z^\top)_{s's} \frac{\partial \log p_w(w_s)}{\partial w_s(j)} \\
&\leq 2\frac{M^2}{m} \log\left(\frac{n}{\delta} \left(\frac{\sqrt[n]{\det(P_t)}}{\det(\lambda I_{dn})} \right)^{\frac{1}{2}} \right)
\end{aligned}
$$

holds with probability at least 1-$\frac{\delta}{n}$ for each j fixed.

Plugging all into (C.16) and taking the union bound, with probability $1-\delta$

$$
\begin{aligned}
&\mathbb{E}_{\theta_t \sim \mu_t}[|\theta_t - \theta_*|^p_{P_t} | h_t] \\
&\leq (2p)^p \left(\left(\sum_{j=1}^{n} \frac{8M^2}{m^3} \log\left(\frac{n}{\delta} \left(\frac{\sqrt[n]{\det(P_t)}}{\det(\lambda I_d)} \right)^{\frac{1}{2}} \right) \right) + \frac{4dn}{m} + 64m + C \right)^{\frac{p}{2}} \\
&\leq (2p)^p \left(8\frac{nM^2}{m^3} \log\left(\frac{n}{\delta} \left(\frac{\lambda_{\max,t}}{\lambda} \right)^{\frac{d}{2}} \right) + C \right)^{\frac{p}{2}},
\end{aligned}
$$

which finishes the proof. □

Appendix D

Details for Theorem 3

To get a uniform bound for the state, we start by showing that $\mathbb{E}[|x_t|^p]$ has a polynomial-in-time bound for any p where the expectation is taken over all noises and the randomized algorithm. A key idea is to decompose an event into a good set and a bad set as proposed in [28]. Let us first define Ω to be the probability space representing all randomness incurred from noises and preconditioned ULA. Then we define the event $E_{t'}$ and $F_{t'}$ as

$$E_{t'} = \{w \in \Omega : \forall t \leq t', |\tilde{\theta}_t - \theta_*|_{P_t} \leq \beta_t(\delta)\}$$

with the constant $C > 0$ from Proposition 2, and

$$F_{t'} = \{w \in \Omega : \forall t \leq t', |x_t| \leq \alpha_t\},$$

where

$$\begin{aligned}
&\beta_t(\delta) \\
&:= e(t(t+1))^{-1/\log\delta} \\
&\times \left(10\sqrt{\frac{dn}{m}\log\left(\frac{1}{\delta}\right)} + 2\log\left(\frac{1}{\delta}\right)\sqrt{\frac{8M^2n}{m^3}\log\left(\frac{nt(t+1)}{\delta}\left(\frac{\lambda_{\max,t}}{\lambda}\right)^{\frac{d}{2}}\right)} + C\right),
\end{aligned}$$

and

$$\alpha_t := \frac{1}{1-\rho}\left(\frac{M_\rho}{\rho}\right)^d \\ \times \left(G(\max_{j\le t}|z_j|)^{\frac{d}{d+1}}\beta_t(\delta)^{\frac{1}{2(d+1)}} + d(\bar{L}+S\bar{L}_\nu)\sqrt{2\log\left(\frac{2t^2(t+1)}{\delta}\right)}\right),$$

where $\bar{L} = \frac{1}{\sqrt{2m}}$ denotes the subgaussianity of our system noise obtained from Herbst argument in [53] and G is a positive constant defined in Lemma 5. Let us briefly describe how the proof proceeds. First, we examine the distance between the exact posterior and true system parameter θ_*, which is given in Proposition 2 below. This quantification in turn allows us to estimate $|\tilde{\theta}_t - \theta_*|$ with high probability with respect to E_t and F_t. Finally, one achieves the polynomial bound for the state combining all together with Proposition 1 as given in Theorem 3. Our result is an extension of [28] to the TS framework.

The next proposition asserts that the event F_t defined at the beginning of the section happens with high probability. Thanks to this result, we can integrate the OFU-based approach with the Bayesian approach where Thompson sampling is exploited. We provide some details of the proof for the sake of completeness focusing on the part which is different from [28].

Proposition 5. *Suppose Assumption 1, 2 and 3 hold. Then for any $\delta > 0$ such that $log(\frac{1}{\delta}) \geq 2$ and $t' \in [1, T]$, we have*

$$Pr(E_{t'} \cap F_{t'}) \geq 1 - 4\delta.$$

Before proving the proposition, let us introduce some auxiliary results on the behavior of $M_t := \tilde{\Theta}_t - \Theta_*$. One of the fundamental ideas is to identify critical columns of M_t representing the column space of M_t where $\tilde{\Theta}_t$ is a matrix whose vectorization is $\tilde{\theta}_t \in \mathbb{R}^{dn}$. We follow the argument presented in Appendix D of [28].

For $\mathcal{B} \subset \mathbb{R}^d$, $v \in \mathbb{R}^d$ and $M \in \mathbb{R}^{d\times n}$, let us define $\pi(v, \mathcal{B})$ be projection of the vector v onto the space $\mathcal{B}$. Similarly, we define $\pi(M, \mathcal{B})$ to be a column-wise

projection of M onto $\mathcal{B}$. We then define a sequence of subspaces $\mathcal{B}_t$ for $t = T, ..., 1$ with $\mathcal{B}_{T+1} = \emptyset$ in the following way. Let $\epsilon > 0$ be given and set $\mathcal{B}_t = \mathcal{B}_{t+1}$. If $|\pi(M_t, \mathcal{B}^\perp)|_F > d\epsilon$ where $|\cdot|_F$ denotes the Frobenius norm, we pick a column v from M_t satisfying $\pi(v, \mathcal{B}^\perp) > \epsilon$ and update $\mathcal{B}_t \leftarrow \mathcal{B}_t \oplus \{v\}$. Therefore, we have

$$|\pi(M_t, \mathcal{B}_t^\perp)| \le |\pi(M_t, \mathcal{B}_t^\perp)|_F \le d\epsilon,$$

after this process ends.

Definition 3. *Let $\mathcal{T}_T$ be the set of timesteps at which subspace $\mathcal{B}_t$ expand. Clearly, $m := |\mathcal{T}_T| \le d$ since M_t has d columns. Let us denote the timesteps by $t_1 > t_2 > ... > t_m$ and define $i(t) := \max\{i \le m : t_i \ge t\}$.*

A key insight of this procedure is to discover a sequence of subspaces $\mathcal{B}_t$ supporting M_t's. That way we derive the following estimate for the projection of any vector x onto $\mathcal{B}_t$:

$$U\epsilon^{2d} \le |\pi(x, \mathcal{B}_t)|^2 \le \sum_{i=1}^{i(t)} |M_{t_i}^\top x|^2,$$

where $U = \frac{U_0}{H}$. Here, $U_0 = \frac{1}{16^{d-2}\max\{1, S^{2(d-2)}\}}$, and H is chosen to be a positive number strictly larger than $\max\{16, \frac{4S^2\tilde{M}^2}{dU_0}\}$ and $\tilde{M} = \sup_{Y \ge 0} \frac{\left(n\bar{L}\sqrt{d\log\left(\frac{1+TY/\lambda}{\delta}\right)} + \sqrt{\lambda}S\right)}{Y}$ where $S > 0$ is from Definition 1, λ satisfies Assumption 2, and T denotes the time horizon.

Using this relation, we have the following result. For the proof, we only highlight the part which is different from [28].

Lemma 5. *For any $t \in [1, T]$, on the event E_t,*

$$\max_{s \le t, s \notin \mathcal{T}_t} |M_s^\top z_s| \le G Z_t^{\frac{d}{d+1}} \beta_t(\delta)^{\frac{1}{2(d+1)}},$$

where $G = 2\left(\frac{2Sd^{d+0.5}}{\sqrt{U}}\right)^{\frac{1}{d+1}}$ and $Z_t = \max_{s \le t} |z_s|$.

Proof. For $M_t = \tilde{\Theta}_t - \Theta_*$, we note that the following holds on the event E_t:

$$\begin{aligned}
\beta_t(\delta) &\geq |\tilde{\theta}_t - \theta_*|_{P_t} \\
&= \sum_{i,i'=1}^{d} \sum_{j,j'=1}^{n} (\tilde{\theta}_t - \theta_*)_{d(j-1)+i} P_{d(j-1)+i,d(j'-1)+i'} (\tilde{\theta}_t - \theta_*)_{d(j'-1)+i'} \\
&= \sum_{i,i'=1}^{d} \sum_{j,j'=1}^{n} (\tilde{\Theta}_t - \Theta_*)_{ij} (I_n)_{jj'} (\sum_{s=1}^{t-1} z_s z_s^\top + \lambda I_d)_{ii'} (\tilde{\Theta}_t - \Theta_*)_{i'j'} \\
&= \sum_{i,i'=1}^{d} \sum_{j=1}^{n} (\tilde{\Theta}_t - \Theta_*)_{ji}^\top (\sum_{s=1}^{t-1} z_s z_s^\top + \lambda I_d)_{ii'} (\tilde{\Theta}_t - \Theta_*)_{i'j} \\
&= \mathbf{tr}(M_t^\top (\sum_{s=1}^{t-1} z_s z_s^\top + \lambda I_d) M_t) \\
&\geq \max_{1 \leq s < t} |M_t^\top z_s|^2.
\end{aligned}$$

Therefore, $\max_{1 \leq s < t} |M_t^\top z_s|^2 \leq \beta_t(\delta)$ so that we can follow the same lines in Lemma 18 [28]. □

We are now ready to prove Proposition 5. Roughly, we combine Proposition 1 and 2 to show the event E_t happens with high probability, which gives us an estimate for $|\tilde{\theta}_t - \theta_*|$. Once established, one can control the event on which the state norm is bounded above by the state norm with lower power, i.e., $|x_t| \leq C|x_t|^{\frac{d}{d+1}}$ for all t.

Proof of Proposition 5. By Proposition 2, we first see that

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

holds with probability $1 - \frac{\delta}{t(t+1)}$. Recalling Proposition 1 and using Minkowski inequality, it holds that

$$\begin{aligned}
&\mathbb{E}_{\tilde{\theta}_t \sim \tilde{\mu}_t} \left[|\tilde{\theta}_t - \theta_*|_{P_t}^p \mid h_t \right]^{\frac{1}{p}} \\
&\leq \mathbb{E}_{\theta_t \sim \mu_t, \tilde{\theta}_t \sim \tilde{\mu}_t} \left[|\tilde{\theta}_t - \theta_t|_{P_t}^p \mid h_t \right]^{\frac{1}{p}} + \mathbb{E}_{\theta_t \sim \mu_t} \left[|\theta - \theta_*|_{P_t}^p \mid h_t \right]^{\frac{1}{p}} \\
&\leq 10 \sqrt{\frac{pdn}{m}} + 2p \sqrt{\frac{8M^2 n}{m^3} \log \left(\frac{nt(t+1)}{\delta} \left(\frac{\lambda_{\max,t}}{\lambda} \right)^{\frac{d}{2}} \right)} + C,
\end{aligned}$$

with probability $1 - \frac{\delta}{t(t+1)}$ for $p \geq 2$. By Markov inequality, for any $\epsilon > 0$,

$$\begin{aligned}
&Pr(|\tilde{\theta}_t - \theta_*|_{P_t} > \epsilon \mid h_t) \\
&\leq \frac{\mathbb{E}_{\tilde{\theta}\sim\tilde{\mu}_t}\left[|\tilde{\theta} - \theta_*|_{P_t}^p \mid h_t\right]}{\epsilon^p} \\
&\leq \frac{1}{\epsilon^p}\left(10\sqrt{\frac{pdn}{m}} + 2p\sqrt{\frac{8M^2n}{m^3}\log\left(\frac{nt(t+1)}{\delta}\left(\frac{\lambda_{\max,t}}{\lambda}\right)^{\frac{d}{2}}\right)} + C\right)^p.
\end{aligned}$$

We choose $p = \log\left(\frac{1}{\delta}\right)$ and

$$\epsilon = e(t(t+1))^{1/p}\left(10\sqrt{\frac{pdn}{m}} + 2p\sqrt{\frac{8M^2n}{m^3}\log\left(\frac{nt(t+1)}{\delta}\left(\frac{\lambda_{\max,t}}{\lambda}\right)^{\frac{d}{2}}\right)} + C\right).$$

Then,

$$Pr\left(|\tilde{\theta}_t - \theta_*|_{P_t} > \beta_t(\delta) \mid h_t\right) \leq \frac{\delta}{t(t+1)},$$

which reads that $Pr(|\theta_t - \theta_*|_t \leq \beta_t(\delta)|h_t) \geq 1 - \frac{\delta}{t(t+1)}$. Noticing that

$$\begin{aligned}
Pr(|\theta_t - \theta_*|_t \leq \beta_t(\delta)) &= \mathbb{E}[\mathbb{E}[\mathbb{1}_{|\theta_t - \theta_*|_t \leq \beta_t(\delta)}|h_t]] \\
&= \mathbb{E}[Pr(|\theta_t - \theta_*|_t \leq \beta_t(\delta)|h_t)] \\
&\geq (1 - \frac{\delta}{t(t+1)})^2,
\end{aligned}$$

we derive that $Pr(|\theta_t - \theta_*|_t \leq \beta_t(\delta)) \geq 1 - \frac{2\delta}{t(t+1)}$ for any $t \geq 1$. Set $S_t = \{w \in \Omega : |\theta_t - \theta_*| \leq \beta_t(\delta)\}$, then $Pr(S_t^c) \leq \frac{2\delta}{t(t+1)}$.

$$Pr(\cap_{t=1}^{t'} S_t) = 1 - Pr(\cup_{t=1}^{t'} S_t^c) \geq 1 - \sum_{t=1}^{t'} Pr(S_t^c) \geq 1 - 2\delta.$$

Therefore,

$$Pr(E_{t'}) \geq 1 - 2\delta$$

for any $t' \leq T$.

Let $t \leq T$ be given. We rewrite the system equation as

$$x_{s+1} = \Gamma_s x_s + r_s,$$

where

$$\Gamma_s = \begin{cases} \tilde{\Theta}_s^\top \tilde{K}(\tilde{\theta}_s) & s \notin \mathcal{T}_t, \\ \Theta_*^\top \tilde{K}(\tilde{\theta}_s) & s \in \mathcal{T}_t, \end{cases}$$

and

$$r_s = \begin{cases} (\tilde{\Theta}_s - \Theta_*)^\top z_s + B_* \nu_s + w_s & s \notin \mathcal{T}_t, \\ B_* \nu_s + w_s & s \in \mathcal{T}_t. \end{cases}$$

Here, $\tilde{K}(\theta)^\top = \begin{bmatrix} I_n & K(\theta)^\top \end{bmatrix}$. Then,

$$\begin{aligned} x_t &= \Gamma_{t-1} x_{t-1} + r_{t-1} \\ &= \Gamma_{t-1}(\Gamma_{t-2} x_{t-2} + r_{t-2}) + r_{t-1} \\ &= \Gamma_{t-1}\Gamma_{t-2} x_{t-2} + \Gamma_{t-1} r_{t-2} + r_{t-1} \\ &= \Gamma_{t-1}\Gamma_{t-2}\Gamma_{t-3} x_{t-3} + \Gamma_{t-1}\Gamma_{t-2} r_{t-3} + \Gamma_{t-1} r_{t-2} + r_{t-1} \\ &= \Gamma_{t-1}\Gamma_{t-2}\ldots\Gamma_2 r_1 + \cdots + \Gamma_{t-1}\Gamma_{t-2} r_{t-3} + \Gamma_{t-1} r_{t-2} + r_{t-1} \\ &= \sum_{j=1}^{t-2} \Big(\prod_{s=j+1}^{t-1} \Gamma_s \Big) r_j + r_{t-1}. \end{aligned}$$

We know that

$$|\tilde{\Theta}_t^\top \tilde{K}(\tilde{\theta}_t)| \leq \rho < 1,$$

and

$$|\Theta_*^\top \tilde{K}(\tilde{\theta}_t)| \leq M_\rho,$$

as the prior has compact support (Assumption 2). Since $|\mathcal{T}_t| \leq d$,

$$\prod_{s=j+1}^{t-1} |\Gamma_s| \leq M_\rho^d \rho^{t-d-j-1}.$$

Hence, we obtain that

$$|x_t| = \left(\frac{M_\rho}{\rho}\right)^d \sum_{j=1}^{t-2} \rho^{t-j-1} |r_j| + |r_{t-1}| \leq \frac{1}{1-\rho} \left(\frac{M_\rho}{\rho}\right)^d \max_{j \leq t} |r_j|.$$

By the definition of r_t,

$$\max_{j\le t}|r_j| \le \max_{j\le t, j\notin\mathcal{T}_t}|(\tilde{\Theta}_j-\Theta_*)^\top z_j| + S\max_{j\le t}|\nu_j| + \max_{j\le t}|w_j|,$$

and from Lemma 5, on E_t, we have

$$\max_{j\le t, j\notin\mathcal{T}_t}|(\tilde{\Theta}_j-\Theta_*)^\top z_j| \le G(\max_{j\le t}|z_j|)^{\frac{d}{d+1}}\beta_t(\delta)^{\frac{1}{2(d+1)}}$$

with probability $1-2\delta$ since $Pr(E_t)\ge Pr(E_T)\ge 1-2\delta$.

Noticing our system noise is $\bar{L}$-sub-Gaussian random vector where $\bar{L}=\frac{1}{\sqrt{2m}}$ by Herbst argument in [53], we have

$$\max_{j\le t}|w_j| \le d\bar{L}\sqrt{2\log\left(\frac{2t^2(t+1)}{\delta}\right)} \tag{D.1}$$

with probability $1-\frac{\delta}{t(t+1)}$. Similarly, since ν_j is $\bar{L}_\nu$-sub-Gaussian random vector, we also have

$$\max_{j\le t}|\nu_j| \le d\bar{L}_\nu\sqrt{2\log\left(\frac{2t^2(t+1)}{\delta}\right)} \tag{D.2}$$

with probability $1-\frac{\delta}{t(t+1)}$. Let us denote the events satisfying (D.1) and (D.2) by $\hat{E}_{w,t}, \hat{E}_{\nu,t}$ respectively. Then, on the event $\hat{E}_{w,t}\cap\hat{E}_{\nu,t}$, we obtain that

$$\begin{aligned}
&|x_t| \\
&\le \frac{1}{1-\rho}\left(\frac{M_\rho}{\rho}\right)^d\left(G(\max_{j\le t}|z_j|)^{\frac{d}{d+1}}\beta_t(\delta)^{\frac{1}{2(d+1)}} + d(\bar{L}+S\bar{L}_\nu)\sqrt{2\log\left(\frac{2t^2(t+1)}{\delta}\right)}\right) \\
&= \alpha_t.
\end{aligned}$$

For $H_{t'} := \cap_{t=1}^{t'}(\hat{E}_{w,t}\cap\hat{E}_{\nu,t})$, we can see that

$$H_{t'}\cap E_{t'} \subset F_{t'}.$$

By the union bound argument,

$$Pr(H_{t'}\cap E_{t'}) \ge 1 - Pr(\cup_{t=1}^{t'}(\hat{E}_{w,t}^c\cup\hat{E}_{\nu,t}^c)) - Pr(E_{t'}^c) \ge 1-4\delta$$

since $Pr(\hat{E}_{w,t}^c)\le\frac{\delta}{t(t+1)}$, $Pr(\hat{E}_{\nu,t}^c)\le\frac{\delta}{t(t+1)}$ and $Pr(E_{t'}^c)\le 2\delta$.

Consequently, we deduce that

$$Pr(E_{t'} \cap F_{t'}) \geq Pr(H_{t'} \cap E_{t'} \cap F_{t'}) = Pr(H_{t'} \cap E_{t'}) \geq 1 - 4\delta.$$

□

Proof of Theorem 3. One can decompose $\mathbb{E}[\max_{j\leq t} |x_t|^p]$ as

$$\mathbb{E}[\max_{j\leq t} |x_t|^p] = \mathbb{E}[\max_{j\leq t} |x_t|^p \mathbb{1}_{F_t}] + \mathbb{E}[\max_{j\leq t} |x_t|^p \mathbb{1}_{F_t^c}]. \tag{D.3}$$

Using Cauchy-Schwartz inequality and the fact that $Pr(F_t^c) \leq 4\delta$, the second term is estimated as

$$\mathbb{E}[\max_{j\leq t} |x_t|^p \mathbb{1}_{F_t^c}] \leq \sqrt{\mathbb{E}[\mathbb{1}_{F_t^c}]}\sqrt{\mathbb{E}[\max_{j\leq t} |x_t|^{2p}]} \leq \sqrt{4\delta}\sqrt{\mathbb{E}[\max_{j\leq t} |x_t|^{2p}]}.$$

Letting $D_t = \Theta_*^\top \tilde{K}(\tilde{\theta}_t)$ and $r_t = B_*\nu_t + w_t$,

$$\begin{aligned}
x_t &= D_{t-1}x_{t-1} + r_{t-1} = D_{t-1}(D_{t-2}x_{t-2} + r_{t-2}) + r_{t-1} \\
&= D_{t-1}D_{t-2}D_{t-3}x_{t-3} + D_{t-1}D_{t-2}r_{t-3} + D_{t-1}r_{t-2} + r_{t-1} \\
&= D_{t-1}D_{t-2}\dots D_2 r_1 + \cdots + D_{t-1}D_{t-2}r_{t-3} + D_{t-1}r_{t-2} + r_{t-1} \\
&= \sum_{j=1}^{t-2}\Big(\prod_{s=j+1}^{t-1} D_s\Big)r_j + r_{t-1}.
\end{aligned}$$

Since $|D_t| \leq M_\rho$,

$$\begin{aligned}
\mathbb{E}[|x_t|^{2p}] &= \mathbb{E}[|\sum_{j=1}^{t-2}\Big(\prod_{s=j+1}^{t-1} D_s\Big)r_j + r_{t-1}|^{2p}] \\
&\leq (t-1)^{2p-1}\mathbb{E}[\sum_{j=1}^{t-2}|\Big(\prod_{s=j+1}^{t-1} D_s\Big)r_j|^{2p} + |r_{t-1}|^{2p}] \\
&\leq (t-1)^{2p-1}\mathbb{E}[\sum_{j=1}^{t-1} M_\rho^{2p(t-j-1)}|r_j|^{2p}] \\
&\leq (t-1)^{2p-1}\mathbb{E}[|r_t|^{2p}]\frac{(M_\rho^{2p(t-1)} - 1)}{M_\rho^{2p} - 1} \\
&\leq (t-1)^{2p-1}\mathbb{E}[|r_t|^{2p}]M_\rho^{2pt},
\end{aligned}$$

where the second inequality holds from Jensen's inequality.

Using Lemma 7 with $\delta = \frac{1}{t^{2p} M_\rho^{2pt}} \leq \frac{1}{t}$, the first term of (D.3) is estimated as

$$\begin{aligned}\mathbb{E}[\max_{j\leq t} |x_t|^p \mathbb{1}_{F_t}] &\leq \mathbb{E}\left[C\left(\log\left(\frac{1}{\delta}\right)^2 \sqrt{\log\left(\frac{t}{\delta}\right)}\right)^{p(d+1)} \mathbb{1}_{F_t}\right] \\ &\leq C\left(\log\left(\frac{1}{\delta}\right)^2 \sqrt{\log\left(\frac{t}{\delta}\right)}\right)^{p(d+1)}.\end{aligned}$$

Finally,

$$\begin{aligned}&\mathbb{E}[\max_{j\leq t} |x_t|^p] \\ &\leq C\left(\log\left(\frac{1}{\delta}\right)^2 \sqrt{\log\left(\frac{t}{\delta}\right)}\right)^{p(d+1)} + \sqrt{4\delta}\sqrt{\mathbb{E}[|x_t|^{2p}]} \\ &\leq C\left(\log\left(t^{2p} M_\rho^{2pt}\right)^2 \sqrt{\log\left(t^{2p+1} M_\rho^{2pt}\right)}\right)^{p(d+1)} + \sqrt{\mathbb{E}[|r_t|^{2p}]} \\ &\leq C t^{\frac{5}{2}p(d+1)} + \sqrt{\mathbb{E}[|r_t|^{2p}]}.\end{aligned}$$

By Jensen's inequality and the subgaussianity of ν_t and w_t,

$$\begin{aligned}\mathbb{E}[|r_t|^{2p}] &\leq 2^{p-1}(S^{2p}\mathbb{E}[|\nu_t|^{2p}] + \mathbb{E}[|w_t|^{2p}]) \\ &\leq 2^{p-1} p!(S^{2p}(4\bar{L}_\nu^2)^p + (\frac{2}{m})^p).\end{aligned}$$

Hence, the result follows. □

Appendix E

Details for Section 4.2.1

Proof of Proposition 3. Given $j \in [1, k]$, let A_*, B_* be the true system parameters and $s \in (t_j, t_{j+1}) := \mathcal{I}_j$. We first define the following quantities for $s \in \mathcal{I}_j$:

$$y_s := \begin{bmatrix} A_* x_{s-1} + B_* u_{s-1} \\ K_j(A_* x_{s-1} + B_* u_{s-1}) \end{bmatrix},$$

where K_j denotes the control gain matrix computed at the beginning of jth episode.

Writing

$$L_s := \begin{bmatrix} I_n & 0 \\ K_j & I_{n_u} \end{bmatrix}, \quad \text{and} \quad \psi_s := \begin{bmatrix} w_{s-1} \\ \nu_s \end{bmatrix},$$

we can decompose z_s as $z_s = y_s + L_s \psi_s$ by the construction of the algorithm.

For a trajectory $(z_s)_{s \geq 1}$, let us introduce a sequence of random variables up to time s, which is denoted by

$$\tilde{h}_s := (x_1, W_1, \nu_1, ..., x_s, W_s, \nu_s),$$

where W_s denotes randomness incurred by the ULA when triggered, hence, $W_s = 0$ if $s \neq t_j$ for some j. Defining the index set

$$\mathcal{J}_k := \{s \in \mathcal{I}_j : j \in [1, k]\},$$

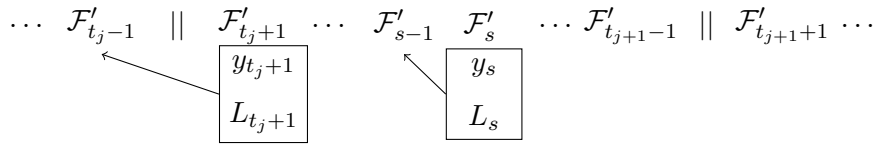

Figure E.1: Filtration and the measurability of (y_s) and (L_s).

we consider the modified filtration

$$\mathcal{F}'_s := \begin{cases} \sigma(\cup_{j\le s}\tilde{h}_j) \quad \text{for} \quad s \in \mathcal{J}_k - \{t_2-1, t_3-1, ..., t_k-1\}, \\ \sigma(\cup_{j\le s+1}\tilde{h}_j) \quad \text{for} \quad s \in \{t_2-1, t_3-1, ..., t_k-1\}. \end{cases}$$

This way we can incorporate the information observed at $s = t_j$ with that made up to $s = t_j - 1$ as seen in Figure E.1.

Yet simple but important observation is that for $\mathcal{J}_k = \{n_i : n_1 < n_2 < ... < n_{\frac{k(k+1)}{2}}\}$ both stochastic processes (L_{n_s}), (y_{n_s}) are $\mathcal{F}'_{n_{s-1}}$-measurable and (ψ_{n_s}) is $\mathcal{F}'_{n_s}$-measurable.

To proceed we first notice that

$$\lambda_{\min}(\lambda I_d + \sum_{s=1}^{t_{k+1}-1} z_s z_s^\top) \succeq \lambda_{\min}(\lambda I_d + \sum_{s\in\mathcal{J}_k} z_s z_s^\top).$$

Invoking Lemma 8 with $\epsilon = \tilde{\lambda} = 1$ and $\xi_s = L_s\psi_s$, it follows that

$$\begin{aligned} &\sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j} z_s z_s^\top \\ &\quad\succeq \sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j} (L_s\psi_s)(L_s\psi_s)^\top \\ &\quad\underbrace{- (\sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j} y_s(L_s\psi_s)^\top)^\top (I_d + \sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j} y_s y_s^\top)^{-1} (\sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j} y_s(L_s\psi_s)^\top)}_{(*)} - I_d. \end{aligned} \tag{E.1}$$

Our goal is to find a lower bound of (E.1).

To begin with, define $\psi_{1,s} = \begin{bmatrix} w_{s-1} \\ 0 \end{bmatrix}$ and $\psi_{2,s} = \begin{bmatrix} 0 \\ \nu_s \end{bmatrix}$ for $s \geq 1$ setting $w_0 = 0$ for simplicity. Noting that $L_s\psi_s = L_s\psi_{1,s} + \psi_{2,s}$, we apply Lemma 8 with $\epsilon = \frac{1}{2}$, $\tilde{\lambda} = 1$ to obtain

$$\begin{aligned}
&\sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j}(L_s\psi_s)(L_s\psi_s)^\top \\
&= \sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j}(L_s\psi_{1,s})(L_s\psi_{1,s})^\top + \frac{1}{2}\sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j}\psi_{2,s}\psi_{2,s}^\top \\
&\underbrace{- 2\,(\sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j}\psi_{2,s}(L_s\psi_{1,s})^\top)^\top(I_d + \sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j}\psi_{2,s}\psi_{2,s}^\top)^{-1}(\sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j}\psi_{2,s}(L_s\psi_{1,s})^\top)}_{(**)} - \frac{1}{2}I_d.
\end{aligned} \tag{E.2}$$

The first term of (E.2) is written as

$$\begin{aligned}
\sum_{s\in\mathcal{J}_k}(L_s\psi_{1,s})(L_s\psi_{1,s})^\top &= \sum_{s\in\mathcal{J}_k}\begin{bmatrix} w_{s-1}w_{s-1}^\top & w_{s-1}(K_{v(s)}w_{s-1})^\top \\ (K_{v(s)}w_{s-1})w_{s-1}^\top & (K_{v(s)}w_{s-1})(K_{v(s)}w_{s-1})^\top \end{bmatrix} \\
&=: \begin{bmatrix} X^\top X & X^\top Y \\ Y^\top X & Y^\top Y \end{bmatrix},
\end{aligned}$$

where $v(s)$ is indicates the episode number such that $s \in \mathcal{I}_{v(s)}$. By Lemma 9, we conclude that

$$\sum_{s\in\mathcal{J}_k}(L_s\psi_{1,s})(L_s\psi_{1,s})^\top = \begin{bmatrix} X^\top X & X^\top Y \\ Y^\top X & Y^\top Y \end{bmatrix} \succeq \begin{bmatrix} \frac{\bar{\lambda}}{|Y|^2+\bar{\lambda}}X^\top X & 0 \\ 0 & -\bar{\lambda}I_{n_u} \end{bmatrix} \tag{E.3}$$

for any $\bar{\lambda} > 0$ where $X^\top X = \sum_{s\in\mathcal{J}_k} w_{s-1}w_{s-1}^\top$ and $Y^\top Y = (K_{v(s)}w_{s-1})(K_{v(s)}w_{s-1})^\top$.

Next, we invoke Lemma 11 with $\epsilon = \frac{1}{2}\lambda_{\min}(\mathbf{W})$ for $\psi_s = w_{s-1}$, $\psi_s = \nu_s$ respectively to characterize good noise sets. Choosing $\rho = \log\frac{2}{\delta}$ in Lemma 11, there exists $C > 0$ such that for any $\delta > 0$ and $k \geq C\sqrt{2\log(\frac{2}{\delta}) + 5d}$, the following events hold

with probability at least $1-\delta$:

$$E_{1,k} = \{w \in \Omega : \frac{1}{4}\lambda_{\min}(\mathbf{W})k(k-1)I_n \preceq \sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j} w_{s-1}w_{s-1}^\top \preceq \frac{1}{2}\lambda_{\mathbf{W}}k(k-1)I_n\},$$

$$E_{2,k} = \{\nu \in \Omega_\nu : \frac{1}{2}\lambda_{\min}(\mathbf{W})kI_{n_u} \preceq \sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j} \nu_s\nu_s^\top \preceq \lambda_{\mathbf{W}}kI_{n_u}\},$$

where $\lambda_{\mathbf{W}} = \lambda_{\max}(\mathbf{W}) + \frac{1}{2}\lambda_{\min}(\mathbf{W})$, $\Omega_\nu \subset \Omega$ denotes the probability spaces associated with the noise sequence $(\nu_s)_{s\geq 1}$ and Ω is the probability space representing all randomness in the algorithm as defined in the previous subsection.

Furthermore, from the observation,

$$\begin{aligned}\mathrm{tr}(\sum_{s\in\mathcal{J}_k}(K_{v(s)}w_{s-1})(K_{v(s)}w_{s-1})^\top) &\leq \sum_{s\in\mathcal{J}_k}\mathrm{tr}((K_{v(s)}w_{s-1})(K_{v(s)}w_{s-1})^\top)\\ &\leq M_K^2\sum_{s\in\mathcal{J}_k}|w_{s-1}|^2\\ &= M_K^2\mathrm{tr}(\sum_{s\in\mathcal{J}_k}w_{s-1}w_{s-1}^\top),\end{aligned}$$

we also have the following event is a subevent of $E_{1,k}$:

$$E_{3,k} = \{w \in \Omega : \sum_{s\in\mathcal{J}_k}(K_{v(s)}w_{s-1})(K_{v(s)}w_{s-1})^\top \preceq \frac{nM_K^2}{2}\lambda_{\mathbf{W}}k(k-1)I_{n_u}\}.$$

To proceed we choose $\bar{\lambda} = \frac{1}{8}\lambda_{\min}(\mathbf{W})k$ in (E.3) and recall that $|Y|^2 = \lambda_{\max}(Y^\top Y)$. On the event $E_{1,k}\cap E_{2,k}\cap E_{3,k}$, first two terms on the right-hand side of (E.2) is lower

bounded as

$$
\begin{aligned}
&\sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j}(L_s\psi_{1,s})(L_s\psi_{1,s})^\top+\frac{1}{2}\sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j}\psi_{2,s}\psi_{2,s}^\top\\
&\succeq\begin{bmatrix}\frac{\bar{\lambda}}{|Y|^2+\bar{\lambda}}X^\top X & 0\\ 0 & -\bar{\lambda}I_{n_u}\end{bmatrix}+\frac{1}{2}\sum_{s\in\mathcal{J}_k}\begin{bmatrix}0\\ \nu_s\end{bmatrix}\begin{bmatrix}0 & \nu_s^\top\end{bmatrix}\\
&\succeq\begin{bmatrix}\frac{\frac{1}{32}\lambda_{\min}^2(\mathbf{W})k^2(k-1)}{\frac{1}{2}nM_K^2\lambda_{\mathbf{W}}k(k-1)+\frac{1}{8}\lambda_{\min}(\mathbf{W})}I_n & 0\\ 0 & -\frac{1}{8}\lambda_{\min}(\mathbf{W})kI_{n_u}\end{bmatrix}+\begin{bmatrix}0 & 0\\ 0 & \frac{1}{4}\lambda_{\min}(\mathbf{W})kI_{n_u}\end{bmatrix}\\
&=k\begin{bmatrix}\frac{\lambda_{\min}^2(\mathbf{W})k(k-1)}{16M_K^2\lambda_{\mathbf{W}}k(k-1)+4\lambda_{\min}(\mathbf{W})}I_n & 0\\ 0 & \frac{1}{8}\lambda_{\min}(\mathbf{W})I_{n_u}\end{bmatrix}\\
&\succeq CkI_d
\end{aligned}
$$

for some $C>0$.

We next deal with $(*)$ in (E.1) and the $(**)$ in (E.2) together as they have the same structure. Let us begin by defining

$$
\begin{aligned}
&S_k(\psi_2,L\psi_1)\\
&:=(\sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j}\psi_{2,s}(L_s\psi_{1,s})^\top)^\top(I_d+\sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j}\psi_{2,s}\psi_{2,s}^\top)^{-1}(\sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j}\psi_{2,s}(L_s\psi_{1,s})^\top).
\end{aligned}
$$

Similarly,

$$
S_k(y,L\psi):=(\sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j}y_s(L_s\psi_s)^\top)^\top(I_d+\sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j}y_sy_s^\top)^{-1}(\sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j}y_s(L_s\psi_s)^\top).
$$

Applying Lemma 12 with $\rho=\log(\frac{1}{\delta})$ to the stochastic processes $(\psi_s)_{s\in\mathcal{I}_j,\forall j}$ and $(y_s)_{s\in\mathcal{I}_j,\forall j}$, the following holds with probability at least $1-\delta$:

$$
\begin{aligned}
E_{4,k}&=\{w\in\Omega,\nu\in\Omega_\nu:|S_k(\psi_2,L\psi_1)|\leq 7\bar{L}_\nu^2\sqrt{M_K^2+2}\log\left(\frac{e^d\Psi}{\delta}\right)\},\\
E_{5,k}&=\{w\in\Omega,\nu\in\Omega_\nu:|S_k(y,L\psi)|\leq 7\bar{L}^2\sqrt{M_K^2+2}\log\left(\frac{e^dY}{\delta}\right)\},
\end{aligned}
$$

where $\Psi = \det(I_d + \sum_{j=1}^{k} \sum_{s \in \mathcal{I}_j} \psi_{2,s} \psi_{2,s}^\top)$ and $Y = \det(I_d + \sum_{j=1}^{k} \sum_{s \in \mathcal{I}_j} y_s y_s^\top)$. Due to $\max_{s \le t} |L_s| \le \sqrt{M_K^2 + 2}$, this result holds. To verify it, we recall that $|L_s| = \sqrt{\lambda_{\max}(L_s L_s^\top)}$. For

$$L_s L_s^\top = \begin{bmatrix} I_n & K_j^\top \\ K_j & K_j K_j^\top + I_{n_u} \end{bmatrix}$$

and any $v = [x^\top, y^\top]^\top$ with $|v| = 1$ where $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^{n_u}$, we have

$$\begin{aligned} v^\top \begin{bmatrix} I_n & K_j^\top \\ K_j & K_j K_j^\top + I_{n_u} \end{bmatrix} v &\le |x|^2 + 2x^\top K_j^\top y + M_K^2 |y|^2 + |y|^2 \\ &\le (M_K^2 + 1)(x^2 + y^2) + |y|^2 \\ &\le M_K^2 + 2. \end{aligned}$$

- Bound of $S_k(\psi_2, L\psi_1)$ on $E_{2,k} \cap E_{4,k}$:

  On $E_{2,k}$,

$$\begin{aligned} \det\left(I_d + \sum_{s \in \mathcal{J}_k} \psi_{2,s} \psi_{2,s}^\top\right)^{\frac{1}{d}} &\le \frac{1}{d}(d + \sum_{s \in \mathcal{J}_k} \psi_{2,s}^\top \psi_{2,s}) \\ &= \frac{1}{d}(d + \sum_{s \in \mathcal{J}_k} |\nu_s|^2) \\ &\le \frac{n_u}{d} \lambda_{\mathbf{W}} k + 1 \\ &\le Ck \end{aligned}$$

  for some $C > 0$ where the second inequality follows by

$$\begin{aligned} \sum_{s \in \mathcal{J}} |\nu_s|^2 = \mathrm{tr}(\sum_{s \in \mathcal{J}_k} \nu_s \nu_s^\top) &\le n_u \lambda_{\max}(\sum_{s \in \mathcal{J}_k} \nu_s \nu_s^\top) \\ &\le n_u \lambda_{\mathbf{W}} k. \end{aligned}$$

Altogether, on the event $E_{2,k} \cap E_{4,k}$,

$$\begin{aligned}
&S_k(\psi_2, L\psi_1) \\
&= |(\sum_{s\in\mathcal{J}_k} \psi_{2,s}(L_s\psi_{1,s})^\top)^\top (I_d + \sum_{s\in\mathcal{J}_k} \psi_{2,s}\psi_{2,s}^\top)^{-1} (\sum_{s\in\mathcal{J}_k} \psi_{2,s}(L_s\psi_{1,s})^\top)| \\
&\le 7\bar{L}_\nu^2 \sqrt{M_K^2+2}\log\left(\frac{Ce^d k^d}{\delta}\right).
\end{aligned}$$

- Bound of $S_k(y, L\psi)$ on $F_{t_{k+1}} \cap E_{1,k} \cap E_{5,k}$:

  On $E_{1,k}$,

$$\begin{aligned}
&\det\left(I_d + \sum_{s\in\mathcal{J}_k} y_s y_s^\top\right)^{\frac{1}{d}} \\
&\le \frac{1}{d}\left(d + \sum_{s\in\mathcal{J}_k} |y_s|^2\right) \\
&= \frac{1}{d}\left(d + \sum_{s\in\mathcal{J}_k} (\underbrace{|x_s - w_{s-1}|^2}_{\le 2|x_s|^2+2|w_{s-1}|^2} + \underbrace{|K_{v(s)}(x_s - w_{s-1})|^2}_{\le 2M_K^2|x_s|^2+2M_K^2|w_{s-1}|^2})\right) \\
&\le \frac{1}{d}\left(d + \sum_{s\in\mathcal{J}_k} ((2+2M_K^2)|x_s|^2 + (2+2M_K^2)|w_{s-1}|^2)\right) \\
&\le \frac{(M_K^2+1)}{d}\Bigg(\underbrace{2\sum_{s\in\mathcal{J}_k}|x_s|^2}_{\text{(a)}} + \underbrace{n(\lambda_{\max}(\mathbf{W}) + \frac{1}{2}\lambda_{\min}(\mathbf{W}))k(k-1)}_{\text{by taking trace in } E_{1,k}}\Bigg) + 1,
\end{aligned}$$

  where the last inequality follows from

$$\begin{aligned}
\sum_{s\in\mathcal{J}} |w_{s-1}|^2 = \mathrm{tr}(\sum_{s\in\mathcal{J}_k} w_{s-1}w_{s-1}^\top) &\le n\lambda_{\max}(\sum_{s\in\mathcal{J}_k} w_{s-1}w_{s-1}^\top) \\
&\le \frac{n}{2}(\lambda_{\max}(\mathbf{W}) + \frac{1}{2}\lambda_{\min}(\mathbf{W}))k(k-1).
\end{aligned}$$

  To bound (a) above, let us observe that $t_{k+1} = \frac{(k+1)(k+2)}{2} \le k^p$ for any $p \ge 3$ and consider the event $F_{t_{k+1}} \cap E_{1,k}$. Applying Lemma 7 with $\delta = k^{-p} \le t_{k+1}^{-1}$,

we deduce that

$$\begin{aligned}\sum_{s\in\mathcal{J}_k}|x_s|^2=\sum_{s\in\mathcal{J}_k}|x_s|^2 &\le t_{k+1}\max_{s\le t_{k+1}}|x_s|^2\\ &\le t_{k+1}\left(C(\log k)^3\sqrt{\log k}\right)^{2(d+1)}\\ &\le Ck^2\left(k\sqrt{\log k}\right)^{2(d+1)}\\ &\le Ck^{3d+5}\end{aligned}$$

for some $C>0$ depending on $p\ge 3$ and the constant from Lemma 7.

Therefore, on the event $F_{t_{k+1}}\cap E_{1,k}\cap E_{5,k}$, we have

$$\begin{aligned}&\det\left(I_d+\sum_{s\in\mathcal{J}_k}y_sy_s^\top\right)^{\frac{1}{d}}\\ &\le (M_K^2+1)\left(\frac{2C}{d}k^{3d+5}+\lambda_{\mathbf{W}}k(k-1)\right)+1\\ &\le Ck^{3d+5}\end{aligned}$$

for some constant $C>0$. As a result,

$$\begin{aligned}S_k(y,L\psi)&=|(\sum_{s\in\mathcal{J}_k}y_s(L_s\psi_s)^\top)^\top(I_d+\sum_{s\in\mathcal{J}_k}y_sy_s^\top)^{-1}(\sum_{s\in\mathcal{J}_k}y_s(L_s\psi_s)^\top)|\\ &\le 7\bar{L}^2\sqrt{M_K^2+2}\log\left(\frac{Ce^dk^{d(3d+5)}}{\delta}\right).\end{aligned}$$

Combining altogether and plugging them into (E.1), on the event $F_{t_{k+1}}\cap E_{1,k}\cap E_{2,k}\cap E_{3,k}\cap E_{4,k}\cap E_{5,k}$, one can show that

$$\begin{aligned}\lambda_{\min}(\lambda I_d+\sum_{j=1}^{k}\sum_{s\in\mathcal{I}_j}z_sz_s^\top)&\ge\lambda+C_1k-C_2\log k+C_3\log(\delta)-C_4\\ &\ge Ck\end{aligned}$$

for some $C_i, C > 0$ with $\delta = k^{-p}$ and k large enough. Moreover, for such a k,

$$\begin{aligned}
&Pr\left(\lambda_{\min}(\lambda I_d + \sum_{s=1}^{t} z_s z_s^\top) \geq Ck\right) \\
&\geq 1 - Pr(F_{t_{k+1}}^c \cup E_{1,k}^c \cup E_{2,k}^c \cup E_{3,k}^c \cup E_{4,k}^c \cup E_{5,k}^c) \\
&\geq 1 - 9\delta.
\end{aligned}$$

Finally, defining the event $\bar{F}_{k+1} := F_{t_{k+1}} \cap E_{1,k} \cap E_{2,k} \cap E_{3,k} \cap E_{4,k} \cap E_{5,k}$,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{\lambda_{\min,k+1}^p}\right] &= \mathbb{E}\left[\frac{1}{\lambda_{\min,k+1}^p}\mathbb{1}_{\bar{F}_{k+1}}\right] + \mathbb{E}\left[\frac{1}{\lambda_{\min,k+1}^p}\mathbb{1}_{\bar{F}_{k+1}^c}\right] \\
&\leq C\mathbb{E}\left[k^{-p}\mathbb{1}_{\bar{F}_{k+1}}\right] + \mathbb{E}\left[\mathbb{1}_{\bar{F}_{k+1}^c}\right] \\
&\leq Ck^{-p} + 9\delta \leq Ck^{-p} \qquad \text{(E.4)}
\end{aligned}$$

where second inequality holds from $\lambda_{\min,t} \geq \lambda \geq 1$. □

Proof of Proposition 4. From (C.16) in Proposition 2 in Appendix D, we know that

$$\mathbb{E}_{\theta_t \sim \mu_t}[|\theta_t - \theta_*|_{P_t}^p | h_t] \leq (2p)^p \left(\frac{4}{m^2}|P_t^{-\frac{1}{2}} \nabla_\theta U_t'(\theta_*)|^2 + \frac{4dn}{m} + 64m + C\right)^{\frac{p}{2}}.$$

where $U_t'(\theta) = \sum_{s=1}^{t-1} \log p_w(x_{s+1} - \Theta^\top z_s)$. Since $\lambda_{\min,t}^{\frac{p}{2}} \mathbb{E}[|\theta_t - \theta_*|^p] \leq \mathbb{E}[|\theta_t - \theta_*|_{P_t}^p]$, it follows that

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}_{\theta \sim \mu_t}[|\theta - \theta_*|^p | h_t]] \\
&\leq (2p)^p \sqrt{\mathbb{E}\left[\frac{1}{\lambda_{\min,t}^p}\right]} \sqrt{\mathbb{E}\left[\left(\frac{4}{m^2}|P_t^{-\frac{1}{2}} \nabla_\theta U_t'(\theta_*)|^2 + \frac{4dn}{m} + 64m + C\right)^p\right]} \\
&\leq (2p)^p \sqrt{\mathbb{E}\left[\frac{1}{\lambda_{\min,t}^p}\right]} \sqrt{2^{p-1}\left(\frac{4^p}{m^{2p}}\mathbb{E}\left[|P_t^{-\frac{1}{2}} \nabla_\theta U_t'(\theta_*)|^{2p}\right] + \left(\frac{4dn}{m} + 64m + C\right)^p\right)}, \\
&\qquad \text{(E.5)}
\end{aligned}$$

where the second inequality holds by Jensen's inequality. To bound $\mathbb{E}\left[|P_t^{-\frac{1}{2}} \nabla_\theta U_t'(\theta_*)|^{2p}\right]$, let us first define $Z := \begin{bmatrix} z_1 & \cdots & z_{t-1} \end{bmatrix}^\top$ and denote the jth component of noise w_t

by $w_t(j)$. A naive bound is achieved as

$$
\begin{aligned}
|P_t^{-\frac{1}{2}}\nabla_\theta U_t'(\theta_*)|^2 &= \sum_{j=1}^{n}\sum_{s',s=1}^{t-1}\frac{\partial \log p_w(w_{s'})}{\partial w_{s'}(j)}(Z(Z^\top Z+\lambda I_d)^{-1}Z^\top)_{s's}\frac{\partial \log p_w(w_s)}{\partial w_s(j)} \\
&\le \sum_{j=1}^{n}\sum_{s',s=1}^{t-1}\frac{\partial \log p_w(w_{s'})}{\partial w_{s'}(j)}(Z(Z^\top Z)^{-1}Z^\top)_{s's}\frac{\partial \log p_w(w_s)}{\partial w_s(j)} \\
&\le \sum_{j=1}^{n}\sum_{s=1}^{t-1}\left(\frac{\partial \log p_w(w_s)}{\partial w_s(j)}\right)^2 \\
&= \sum_{s=1}^{t-1}|\nabla_w \log p_w(w_s)|^2, \qquad \text{(E.6)}
\end{aligned}
$$

where the second inequality follows from the fact that $Z(Z^\top Z)^{-1}Z^\top$ is a projection matrix.

We now claim that $\mathbb{E}\left[|P_t^{-\frac{1}{2}}\nabla_\theta U_t'(\theta_*)|^{2p}\right]$ has a better bound compared to the naive one with high probability. For $s \ge 0$, let us consider the natural filtration

$$\mathcal{F}_s = \sigma((z_1, ..., z_{s+1}))$$

where $z_s = (x_s, u_s)$. Clearly, for $s \ge 1$, z_s is $\mathcal{F}_{s-1}$-measurable and the random vector $\nabla_w \log p_w(w_s)$ is $\mathcal{F}_s$-measurable. Then for each $j \in [1, n]$, we set $\eta_s = \frac{\partial \log p_w(w_s)}{\partial w_s(j)}$, $X_s = z_s$, $S_t = \sum_{s=1}^{t-1}\eta_s X_s = \sum_{s=1}^{t-1}\frac{\partial \log p_w(w_s)}{\partial w_s(j)}z_s$. Here, η_s is a $\frac{M}{\sqrt{m}}$-sub-Gaussian random variable since $v^\top \nabla_w \log p_w(w_t)$ is $\frac{M}{\sqrt{m}}$-sub-Gaussian random variable for any $v \in \mathbb{R}^n$ given when w_t is sub-Gaussian (Proposition 2.18 in [52]). We then invoke Lemma 6 yielding that

$$\lambda I_d + \sum_{s=1}^{t-1} X_s X_s^\top = \lambda I_d + Z^\top Z.$$

Consequently,

$$
\begin{aligned}
&(\sum_{s=1}^{t-1}\eta_s X_s)^\top(\lambda I_d+\sum_{s=1}^{t-1}X_sX_s^\top)^{-1}(\sum_{s=1}^{t-1}\eta_s X_s)\\
&=\sum_{s,s'=1}^{t-1}\frac{\partial\log p_w(w_{s'})}{\partial w_{s'}(j)}(Z(Z^\top Z+\lambda I_d)^{-1}Z^\top)_{s's}\frac{\partial\log p_w(w_s)}{\partial w_s(j)}\\
&\le 2\frac{M^2}{m}\log\left(\frac{n}{\delta}\left(\frac{\sqrt[n]{\det(P_t)}}{\det(\lambda I_d)}\right)^{\frac{1}{2}}\right),
\end{aligned}
$$

holds with probability at least $1-\frac{\delta}{n}$. Here, we use the fact that $\det(\lambda I_d+Z^\top Z)=\sqrt[n]{\det(\lambda I_{dn}+\sum_{s=1}^{t-1}blkdiag\{z_sz_s^\top\}_{i=1}^n)}=\sqrt[n]{\det(P_t)}$.

By the union bound argument,

$$
\begin{aligned}
|P_t^{-\frac{1}{2}}\nabla_\theta U_t'(\theta_*)|^2&=\sum_{j=1}^n\sum_{s,s'=1}^{t-1}\frac{\partial\log p_w(w_{s'})}{\partial w_{s'}(j)}(Z(Z^\top Z+\lambda I_d)^{-1}Z^\top)_{s's}\frac{\partial\log p_w(w_s)}{\partial w_s(j)}\\
&\le 2\frac{nM^2}{m}\log\left(\frac{n}{\delta}\left(\frac{\sqrt[n]{\det(P_t)}}{\det(\lambda I_d)}\right)^{\frac{1}{2}}\right),
\end{aligned}
\tag{E.7}
$$

with probability at least $1-\delta$ for any $\delta>0$. Let us denote this event as $\tilde{E}$ so that $Pr(\tilde{E})\ge 1-\delta$.

Combining the naive bound (E.6) and improved bound (E.7),

$$
\begin{aligned}
&\mathbb{E}\left[|P_t^{-\frac{1}{2}}\nabla_\theta U_t'(\theta_*)|^{2p}\right]\\
&=\mathbb{E}\left[\mathbb{1}_{\tilde{E}}|P_t^{-\frac{1}{2}}\nabla_\theta U_t'(\theta_*)|^{2p}\right]+\mathbb{E}\left[\mathbb{1}_{\tilde{E}^c}|P_t^{-\frac{1}{2}}\nabla_\theta U_t'(\theta_*)|^{2p}\right]\\
&\le\underbrace{\mathbb{E}\left[\left(2\frac{nM^2}{m}\log\left(\frac{n}{\delta}\left(\frac{\sqrt[n]{\det(P_t)}}{\det(\lambda I_d)}\right)^{\frac{1}{2}}\right)\right)^p\right]}_{\text{by (E.7)}}+\sqrt{\mathbb{E}\left[\mathbb{1}_{\tilde{E}^c}\right]}\sqrt{\mathbb{E}\left[|P_t^{-\frac{1}{2}}\nabla_\theta U_t'(\theta_*)|^{4p}\right]}\\
&\le\mathbb{E}\left[\left(2\frac{nM^2}{m}\log\left(\frac{n}{\delta}\left(\frac{\lambda_{\max,t}}{\lambda}\right)^{\frac{d}{2}}\right)\right)^p\right]+\sqrt{\delta}\underbrace{\sqrt{\mathbb{E}\left[\left(\sum_{s=1}^{t-1}|\nabla_w\log p_w(w_s)|^2\right)^{2p}\right]}}_{\text{by (E.6)}}.
\end{aligned}
\tag{E.8}
$$

We handle two terms on the right hand side separately. Recall that $g:x\to(\log x)^p$ is concave on $x\ge 1$ whenever $p>0$. By the Jensen's inequality, the first term is

bounded as

$$
\begin{aligned}
&\mathbb{E}\left[\left(2\frac{nM^2}{m}\log\left(\frac{n}{\delta}\left(\frac{\lambda_{\max,t}}{\lambda}\right)^{\frac{d}{2}}\right)\right)^p\right] \\
&= \mathbb{E}\left[\left(\frac{dnM^2}{m}\log\left(\frac{n}{\delta^{\frac{2}{d}}}\frac{\lambda_{\max,t}}{\lambda}\right)\right)^p\right] \\
&\leq \left(\frac{dnM^2}{m}\right)^p \log\left(\frac{n}{\lambda\delta^{\frac{2}{d}}}\mathbb{E}[\lambda_{\max,t}]\right)^p \\
&\leq \left(\frac{dnM^2}{m}\right)^p \log\left(\frac{n}{\lambda\delta^{\frac{2}{d}}}\mathbb{E}[\frac{1}{n}\mathrm{tr}(P_t)]\right)^p \\
&\leq \left(\frac{dnM^2}{m}\right)^p \log\left(\frac{n}{\lambda\delta^{\frac{2}{d}}}\mathbb{E}[d\lambda + \sum_{s=1}^{t-1}|z_s|^2]\right)^p \\
&\leq \left(\frac{dnM^2}{m}\right)^p \log\left(\frac{n}{\lambda\delta^{\frac{2}{d}}}\left(d\lambda + M_K^2 t\mathbb{E}[\max_{j\leq t-1}|x_j|^2]\right)\right)^p \\
&\leq \left(\frac{dnM^2}{m}\right)^p \log\left(\frac{n}{\lambda\delta^{\frac{2}{d}}}\left(d\lambda + CM_K^2 t^{7d+8}\right)\right)^p,
\end{aligned}
$$

where the last inequality holds from the Theorem 3.

On the other hand, the second term of (E.8) can be handled similarly. Recalling the Jensen's inequality,

$$
(\frac{\sum_i^n a_i}{n})^{2p} \leq \frac{\sum_{i=1}^n a_i^{2p}}{n}
$$

for $a_i \in \mathbb{R}$ and $p \geq 1$, we have that

$$
\begin{aligned}
\sqrt{\delta}\sqrt{\mathbb{E}\left[\left(\sum_{s=1}^{t-1}|\nabla_w \log p_w(w_s)|^2\right)^{2p}\right]} &\leq \sqrt{\delta}\sqrt{t^{2p-1}\mathbb{E}\left[\sum_{s=1}^{t-1}|\nabla_w \log p_w(w_s)|^{4p}\right]} \\
&\leq \sqrt{\delta}t^p\sqrt{\mathbb{E}\left[|\nabla_w \log p_w(w_t)|^{4p}\right]} \\
&\leq \sqrt{\delta}t^p\sqrt{\left(\frac{4M^2}{m}\right)^{2p}(2p)!} \\
&\leq 8^p\frac{M^{2p}}{m^p}p^p\sqrt{\delta}t^p,
\end{aligned}
$$

where the third inequality comes from well-known fact that any $\bar{L}$-sub-Gaussian random vector X satisfies $\mathbb{E}[X^{2q}] \leq q!(4\bar{L}^2)^q$ for any $q > 0$.

Choosing $\delta = \frac{1}{t^{2p}}$ and combining two bounds,

$$\begin{aligned}
&\mathbb{E}\left[|P_t^{-\frac{1}{2}}\nabla_\theta U_t'(\theta_*)|^{2p}\right] \\
&\leq \left(\frac{dnM^2}{m}\right)^p \log\left(\frac{n}{\lambda\delta^{\frac{2}{d}}}\left(d\lambda + CM_K^2 t^{7d+8}\right)\right)^p + 8^p\frac{M^{2p}}{m^p}p^p\sqrt{\delta}t^p \\
&\leq \left(\frac{dnM^2}{m}\right)^p \log\left(nt^{\frac{4p}{d}}\left(d + \frac{CM_K^2}{\lambda}t^{7d+8}\right)\right)^p + 8^p\frac{M^{2p}}{m^p}p^p.
\end{aligned}$$

Finally, plugging (E.8) and the result of Proposition 3 to (E.5),

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}_{\theta_t\sim\mu_t}[|\theta_t - \theta_*|^p|h_t]] \\
&\leq (2p)^p\sqrt{\mathbb{E}\left[\frac{1}{\lambda_{\min,t}^p}\right]}\sqrt{2^{p-1}\left(\frac{4^p}{m^{2p}}\mathbb{E}\left[|P_t^{-\frac{1}{2}}\nabla_\theta U_t'(\theta_*)|^{2p}\right] + \left(\frac{4dn}{m} + 64m + C\right)^p\right)} \\
&\leq O(t^{-\frac{1}{4}}\sqrt{\log t})^p.
\end{aligned}$$

□

Appendix F

Miscellaneous Lemmas

Lemma 6 (Theorem 1 [48], Self-Normalized Bound for Vector-Valued Martingale)**.** *Let $(\mathcal{F}_s)_{s=1}^{\infty}$ be a filtration. Let $(\eta_s)_{s=1}^{\infty}$ be a real-valued stochastic process such that η_s is $\mathcal{F}_s$-measurable and η_s is conditionally R-sub-Gaussian for some $R > 0$. Let $(X_s)_{s=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that X_s is $\mathcal{F}_{s-1}$-measurable. For any $t \geq 1$, define*

$$V_t = \lambda I_d + \sum_{s=1}^{t} X_s X_s^{\top}, \quad S_t = \sum_{s=1}^{t} \eta_s X_s,$$

where $\lambda > 0$ is given constant. Then, for any $\delta > 0$, with probability $1 - \delta$, for all $t \geq 1$, one has

$$|S_t|^2_{V_t^{-1}} \leq 2R^2 \log\left(\frac{1}{\delta}\sqrt{\frac{\det(V_t)}{\det(\lambda I_d)}}\right).$$

Lemma 7 (Lemma 5 in [28])**.** *Let $t > 1$ be given. For some $C(d, m, \rho, M_\rho, \bar{L}_\nu, S) > 0$ and any $\delta \leq \frac{1}{t}$,*

$$\mathbb{1}_{F_t} \max_{j \leq t} |x_j| \leq C\left(\log\left(\frac{1}{\delta}\right)^2 \sqrt{\log\left(\frac{t}{\delta}\right)}\right)^{d+1}.$$

Proof. On the event F_t, define $X_t := \max_{j \leq t} |x_j| \leq \alpha_t$. Here, we may assume that $X_t \geq 1$ as the result above holds with some $C > 0$ large enough when $X_t < 1$. We

observe that

$$
\begin{aligned}
&|x_t| \\
&\leq \frac{1}{1-\rho}\left(\frac{M_\rho}{\rho}\right)^d \left(G(\max_{j\leq t}|z_j|)^{\frac{d}{d+1}}\beta_t(\delta)^{\frac{1}{2(d+1)}} + d(\bar{L}+S\bar{L}_\nu)\sqrt{2\log\left(\frac{2t^2(t+1)}{\delta}\right)}\right) \\
&= \alpha_t,
\end{aligned}
$$

and α_t is monotone increasing in F_t. From

$$
X_t = \max_{j\leq t}|x_j| \leq \alpha_t,
$$

in F_t, we derive that

$$
X_t \leq G_1\beta_t(\delta)X_t^{\frac{d}{d+1}} + G_2\sqrt{\log\left(\frac{t}{\delta}\right)} \tag{F.1}
$$

by choosing constants G_i's appropriately. Let us recall $\beta_t(\delta)$ which is given as

$$
\begin{aligned}
&\beta_t(\delta) \\
&= e(t(t+1))^{-1/\log\delta} \\
&\times \left(10\sqrt{\frac{dn}{m}\log\left(\frac{1}{\delta}\right)} + 2\log\left(\frac{1}{\delta}\right)\sqrt{\frac{8M^2n}{m^3}\log\left(\frac{nt(t+1)}{\delta}\left(\frac{\lambda_{\max,t}}{\lambda}\right)^{\frac{d}{2}}\right) + C}\right).
\end{aligned}
$$

For $\delta \leq \frac{1}{t}$,

$$
\begin{aligned}
(t(t+1))^{-1/\log\delta} &\leq (t(t+1))^{1/\log t} \\
&\leq (2t^2)^{1/\log t} \\
&= 2^{1/\log t}t^{2/\log t} \\
&\leq e^3.
\end{aligned}
$$

As a result,

$$
\begin{aligned}
&\beta_t(\delta) \\
&\leq e^4\left(10\sqrt{\frac{dn}{m}\log\left(\frac{1}{\delta}\right)} + 2\log\left(\frac{1}{\delta}\right)\sqrt{\frac{8M^2n}{m^3}\log\left(\frac{nt(t+1)}{\delta}\left(\frac{\lambda_{\max,t}}{\lambda}\right)^{\frac{d}{2}}\right) + C}\right) \\
&=: \beta_t'(\delta).
\end{aligned}
$$

In turn, (F.1) implies that

$$X_t \leq G_1 \beta_t'(\delta) X_t^{\frac{d}{d+1}} + G_2 \sqrt{\log\left(\frac{t}{\delta}\right)}.$$

We now claim that one further has

$$X_t \leq \left(G_1 \beta_t'(\delta) + G_2 \sqrt{\log\left(\frac{t}{\delta}\right)} \right)^{d+1}, \tag{F.2}$$

when $G_1 \beta_t'(\delta) + G_2 \sqrt{\log\left(\frac{t}{\delta}\right)} \geq 1$. To see this, let us set

$$f(x) = x - \alpha x^{\frac{d}{d+1}} - \beta$$

with $\alpha = G_1 \beta_t'(\delta)$ and $\beta = G_2 \sqrt{\log\left(\frac{t}{\delta}\right)}$. Here, we may assume that $\alpha + \beta \geq 1$ by adjusting the constants. Clearly, $f(x)$ is increasing when $x > \left(\frac{\alpha d}{d+1}\right)^{\frac{1}{d+1}}$. Noting that

$$f((\alpha+\beta)^{d+1}) = \beta(\alpha+\beta)^d - \beta \geq 0,$$

since $\alpha + \beta \geq 1$, it follows that $x \leq (\alpha+\beta)^{d+1}$ whenever $f(x) \leq 0$. Therfore, the claim follows.

To proceed let us handle the term $\beta_t'(\delta)$. We first see that the preconditioner P_t satisfies that

$$\lambda_{\max,t} \leq \frac{1}{n}\mathrm{tr}(P_t) = d\lambda + \sum_{s=1}^{t-1} |z_s|^2 \leq d\lambda + M_K^2 t X_t^2, \tag{F.3}$$

where M_K is from Definition 2. Using this relation, one derives that

$$\begin{aligned}
&\beta_t'(\delta) \\
&= G_1 \sqrt{\log\left(\frac{1}{\delta}\right)} + G_2 \log\left(\frac{1}{\delta}\right) \sqrt{G_3 \log X_t + G_4 \log\left(\frac{t}{\delta}\right) + C} \\
&\leq G_1 \sqrt{\log\left(\frac{1}{\delta}\right)} + G_2 \log\left(\frac{1}{\delta}\right) \sqrt{\log X_t} + G_3 \log\left(\frac{1}{\delta}\right) \sqrt{\log\left(\frac{t}{\delta}\right)} + G_4 \log\left(\frac{1}{\delta}\right).
\end{aligned} \tag{F.4}$$

for appropriately chosen $G_i > 0$. Here, G_i's represent different constants whenever it appears for brevity.

Define $a_t := X_t^{\frac{1}{d+1}} \geq 1$. Combining (F.2) and (F.4), for $\delta > 0$ small enough,

$$a_t \leq G_1 \log\left(\frac{1}{\delta}\right)\sqrt{\log a_t} + G_2 \log\left(\frac{1}{\delta}\right)\sqrt{\log\left(\frac{t}{\delta}\right)}.$$

To finish the proof, we claim the following.

Claim] Given $c_1, c_2 \geq 1$, when $x \geq 1$ satisfies that

$$x \leq c_1\sqrt{\log x} + c_2,$$

then, $x \leq Cc_1^2c_2$ where C is independent of c_1 and c_2.

Proof of the claim. Let

$$f(x) = x - c_1\sqrt{\log x} - c_2.$$

From

$$f(x) \geq x - c_1\sqrt{x} - c_2 = (\sqrt{x} - \frac{c_1 + \sqrt{c_1^2 + 4c_2}}{2})(\sqrt{x} - \frac{c_1 - \sqrt{c_1^2 + 4c_2}}{2}),$$

$f(x) \leq 0$ implies that $x \leq Cc_1c_2$ from some $C > 0$ which is independent of c_1 and c_2. □

Finally, setting

$$c_1 = G_1 \log\left(\frac{1}{\delta}\right) \text{ and } c_2 = \log\left(\frac{1}{\delta}\right)\sqrt{\log\left(\frac{t}{\delta}\right)},$$

we deduce that

$$a_t \leq G_1 \log\left(\frac{1}{\delta}\right)^3\sqrt{\log\left(\frac{t}{\delta}\right)}.$$

□

Lemma 8 (Lemma 10 in [36])**.** *Let $(z_s)_{s=1}^{\infty}$, $(y_s)_{s=1}^{\infty}$ and $(\psi_s)_{s=1}^{\infty}$ be three sequences of vectors in $\mathbb{R}^d$, satisfying the linear relation $z_s = y_s + \psi_s$ for all $s \geq 0$. Then, for all*

$\tilde{\lambda} > 0$, *all* $t \geq 1$ *and all* $\epsilon \in (0, 1]$, *we have*

$$\sum_{s=1}^{t} z_s z_s^\top \succeq \sum_{s=1}^{t} \psi_s \psi_s^\top + (1-\epsilon) \sum_{s=1}^{t} y_s y_s^\top - \frac{1}{\epsilon} (\sum_{s=1}^{t} y_s \psi_s^\top)^\top (\tilde{\lambda} I_d + \sum_{s=1}^{t} y_s y_s^\top)^{-1} (\sum_{s=1}^{t} y_s \psi_s^\top) - \epsilon \tilde{\lambda} I_d.$$

Lemma 9 (Lemma 12 in [36]). *For two matrices* X *and* Y *and any* $\bar{\lambda} > 0$, *we have*

$$\begin{bmatrix} X^\top X & X^\top Y \\ Y^\top X & Y^\top Y \end{bmatrix} \preceq \begin{bmatrix} \frac{\bar{\lambda}}{|Y|^2 + \bar{\lambda}} X^\top X & 0 \\ 0 & -\bar{\lambda} I \end{bmatrix}.$$

Lemma 10 ([54]). *Let* $W \in \mathbb{R}^{d \times d}$ *be a random matrix and* $\epsilon \in (0, \frac{1}{2})$ *and* $\mathcal{N}$ *be* ϵ*-net in* S^{d-1} *with minimal cardinality. Then, for any* $\rho > 0$,

$$Pr(|W| > \rho) \leq (\frac{2}{\epsilon} + 1)^d \max_{x \in \mathcal{N}} Pr(|x^\top W x| > (1 - 2\epsilon)\rho).$$

Lemma 11 (Modification of Proposition 8 in [36]). *Let* $(\psi_s)_{s=1}^{\infty}$ *be a sequence of independent, zero mean,* $\bar{L}$*-sub-Gaussian and* $\mathcal{F}_s$*-measurable random vector* $\in \mathbb{R}^d$. *Then, for all* $\rho > 0$, $0 < \epsilon < 1$ *and* $t \geq \min\{\frac{32^2 \bar{L}^4}{\epsilon^2}, \frac{32 \bar{L}^2}{\epsilon}\}(2\rho + 5d)$,

$$Pr\left((\lambda_{\min}(\mathbb{E}[\psi_t \psi_t^\top]) - \epsilon) t I_d \preceq \sum_{s=1}^{t} \psi_s \psi_s^\top \preceq (\lambda_{\max}(\mathbb{E}[\psi_t \psi_t^\top]) + \epsilon) t I_d \right) \geq 1 - 2e^{-\rho}.$$

Proof. Recalling the proof of Proposition 8 in [36] with $M_s = I_d$ and $\xi_s = \psi_s$, it is straightforward to obtain

$$Pr\left(\left| \sum_{s=1}^{t} (x^\top \psi_s)^2 - \sum_{s=1}^{t} \mathbb{E}[(x^\top \psi_s)^2] \right| > \rho \right) \leq 2 \exp\left(-\frac{1}{2} \min\{\frac{\rho^2}{(4\bar{L})^4 t}, \frac{\rho}{(4\bar{L})^2}\} \right).$$

Now we apply Lemma 10 with $\epsilon = \frac{1}{4}$ and $W = \sum_{s=1}^{t} (\psi_s \psi_s^\top - \mathbb{E}[\psi_s \psi_s^\top])$, we get

$$Pr\left(\left| \sum_{s=1}^{t} \psi_s \psi_s^\top - \sum_{s=1}^{t} \mathbb{E}[\psi_s \psi_s^\top] \right| > \rho \right) \leq 2 \cdot 9^d \exp\left(-\frac{1}{2} \min\{\frac{\rho^2}{4(4\bar{L})^4 t}, \frac{\rho}{2(4\bar{L})^2}\} \right).$$

Reparameterizing, we further obtain

$$Pr\left(\left| \sum_{s=1}^{t} \psi_s \psi_s^\top - \sum_{s=1}^{t} \mathbb{E}[\psi_s \psi_s^\top] \right| > 32 \bar{L}^2 t \max\{\sqrt{\frac{2\rho + 5d}{t}}, \frac{2\rho + 5d}{t}\} \right) \leq 2e^{-\rho}.$$

For $t \geq \min\{\frac{32^2\bar{L}^4}{\epsilon^2}, \frac{32\bar{L}^2}{\epsilon}\}(2\rho + 5d)$,

$$Pr\left(\left|\sum_{s=1}^{t}\psi_s\psi_s^\top - \sum_{s=1}^{t}\mathbb{E}[\psi_s\psi_s^\top]\right| > \epsilon t\right) \leq 2e^{-\rho}.$$

Since $\psi_s\psi_s^\top$ is a symmetric matrix, the inequality $\left|\sum_{s=1}^{t}\psi_s\psi_s^\top - \sum_{s=1}^{t}\mathbb{E}[\psi_s\psi_s^\top]\right| \leq \epsilon t$ implies that

$$\lambda_{\max}^2(\sum_{s=1}^{t}\psi_s\psi_s^\top - \sum_{s=1}^{t}\mathbb{E}[\psi_s\psi_s^\top]) \leq \epsilon^2 t^2$$

and

$$\lambda_{\min}^2(\sum_{s=1}^{t}\psi_s\psi_s^\top - \sum_{s=1}^{t}\mathbb{E}[\psi_s\psi_s^\top]) \leq \epsilon^2 t^2.$$

As a result,

$$\begin{aligned}(\lambda_{\min}(\mathbb{E}[\psi_t\psi_t^\top]) - \epsilon)tI_d &\preceq \sum_{s=1}^{t}\mathbb{E}[\psi_s\psi_s^\top] - \epsilon t I_d \\ &\preceq \sum_{s=1}^{t}\psi_s\psi_s^\top \\ &\preceq \sum_{s=1}^{t}\mathbb{E}[\psi_s\psi_s^\top] + \epsilon t I_d \\ &\preceq (\lambda_{\max}(\mathbb{E}[\psi_t\psi_t^\top]) + \epsilon)tI_d.\end{aligned}$$

□

Lemma 12 (Proposition 9 in [36])**.** *Let $\mathcal{F}_s$ be a filtration and $(\psi_s)_{s=1}^{\infty}$ be a sequence of independent, zero mean, $\bar{L}$-sub-Gaussian and $\mathcal{F}_s$ measurable random vectors in $\mathbb{R}^d$. Let $(L_s)_{s=1}^{\infty}$ be a sequence of random matrices in $\mathbb{R}^{d\times d}$ such that $\mathcal{F}_{s-1}$ measurable and $|L_s| < \infty$. Let $(y_s)_{s=1}^{\infty}$ be a sequence of $\mathcal{F}_{s-1}$ measurable random variables in $\mathbb{R}^d$. Then for all positive definite matrix $V \succ 0$, the following self-normalized matrix process defined by*

$$S_t(y, L\psi) = (\sum_{s=1}^{t} y_s(L_s\psi_s)^\top)^\top(V + \sum_{s=1}^{t} y_s y_s^\top)^{-1}(\sum_{s=1}^{t} y_s(L_s\psi_s)^\top)$$

satisfies

$$Pr\Bigg[|S_t(y, L\psi)| > \bar{L}^2(\max_{1\leq s\leq t}|L_s|)\big((2\log\Bigg(\det\Bigg(I_d + V^{-1}\sum_{s=1}^{t} y_s y_s^\top\Bigg)\Bigg) + 4\rho + 7d\Bigg)\Bigg]$$
$$\leq e^{-\rho}.$$

for all $\rho, t \geq 1$.

Appendix G

Details for Section 5

G.0.1 Proof of Theorem 5

At kth episode, for timestep $t \in [t_k, t_{k+1})$, x_t is written as

$$x_{t+1} = (A_* + B_* K(\tilde{\theta}_t))x_t + r_t. \tag{G.1}$$

where $r_t = B_*\nu_t + w_t$. Squaring and taking expectations on both sides of the equation above with respect to noises, the prior and randomized actions,

$$\mathbb{E}[|x_{t+1}|^2] \leq \mathbb{E}[|D_t|^2|x_t|^2] + \mathbb{E}[|r_t|^2], \tag{G.2}$$

where $D_t = A_* + B_* K(\tilde{\theta}_t)$.

Since θ_* is stabilizable, it is clear to see that there exists $\epsilon_0 > 0$ small for which $|\theta - \theta_*| \leq \epsilon_0$ implies that $|A_* + B_* K(\theta)| \leq \Delta < 1$ for some $\Delta > 0$. Splitting $\mathbb{E}[|D_t|^2|x_t|^2]$ around the true system parameter θ_*,

$$\mathbb{E}[|D_t|^2|x_t|^2] = \underbrace{\mathbb{E}[|D_t|^2|x_t|^2 \mathbb{1}_{|\tilde{\theta}_t - \theta_*| \leq \epsilon_0}]}_{(i)} + \underbrace{\mathbb{E}[|D_t|^2|x_t|^2 \mathbb{1}_{|\tilde{\theta}_t - \theta_*| > \epsilon_0}]}_{(ii)}.$$

One can see that (i) is bounded by $\Delta^2 \mathbb{E}[|x_t|^2]$ by the construction. For (ii), we note that $|D_t| \leq M_\rho$ by Assumption 2. Using Cauchy-Schwartz inequality, (ii) is bounded as

$$\mathbb{E}[|D_t|^2|x_t|^2 \mathbb{1}_{|\tilde{\theta}_t - \theta_*| > \epsilon_0}]] \leq M_\rho^2 \sqrt{Pr(|\tilde{\theta}_t - \theta_*| > \epsilon_0)} \sqrt{\mathbb{E}[|x_t|^4]}. \tag{G.3}$$

By Markov's inequality,

$$\begin{aligned} Pr(|\tilde{\theta}_t - \theta_*| > \epsilon_0) &\leq \frac{\mathbb{E}[|\tilde{\theta}_t - \theta_*|^p]}{\epsilon_0^p} \\ &\leq C\left(t^{-\frac{1}{4}}\sqrt{\log t}\right)^p, \end{aligned}$$

where the last inequality holds for $t \geq t_0$ thanks to Theorem 4, and C is a positive constant depending only on p and ϵ_0. Taking p large enough to satisfy $p > 28(d+1)$, Theorem 3 yields that

$$M_\rho^2\sqrt{Pr(|\tilde{\theta}_t - \theta_*| > \epsilon_0)}\sqrt{\mathbb{E}[|x_t|^4]} \leq M_\rho^2 C\left(t^{-\frac{1}{4}}\sqrt{\log t}\right)^p t^{7(d+1)} < C$$

for some $C > 0$.

Therefore, $\mathbb{E}[|x_{t+1}|^2]$ is estimated as

$$\mathbb{E}[|x_{t+1}|^2] \leq \Delta^2\mathbb{E}[|x_t|^2] + C + \mathbb{E}[|r_t|^2].$$

As r_t is sub-Gaussian, we also have $\mathbb{E}[|r_t|^2]$ is bounded, and hence,

$$\mathbb{E}[|x_t|^2] < C$$

for all $t \in [1, T]$ and $C > 0$ by the recursive relation.

To handle the fourth moment, we take the fourth power on both sides and expectation to (G.1) to get

$$\begin{aligned} &\mathbb{E}[|x_{t+1}|^4] \\ &\leq \mathbb{E}[|D_t x_t|^4] + \underbrace{4\mathbb{E}[|D_t x_t|^2 (D_t x_t)^\top w_t]}_{=0} + 6\mathbb{E}[|D_t x_t|^2|r_t|^2] + 4\mathbb{E}[|D_t x_t||r_t|^3] + \mathbb{E}[|r_t|^4] \\ &\leq [|D_t|^4|x_t|^4 \mathbb{1}_{|\tilde{\theta}_t - \theta_*| \leq \epsilon_0}] + \mathbb{E}[|D_t|^4|x_t|^4 \mathbb{1}_{|\tilde{\theta}_t - \theta_*| \geq \epsilon_0}] \\ &\quad + \underbrace{6M_\rho^2\mathbb{E}[|r_t|^2]\mathbb{E}[|x_t|^2] + 4M_\rho\mathbb{E}[|r_t|^3]\mathbb{E}[|x_t|] + \mathbb{E}[|r_t|^4]}_{<C} \\ &\leq \Delta^4\mathbb{E}[|x_t|^4] + M_\rho^4\sqrt{Pr(|\tilde{\theta}_t - \theta_*| \geq \epsilon_0)}\sqrt{\mathbb{E}[|x_t|^8]} + C, \end{aligned}$$

since $\mathbb{E}[|x_t|^2] \leq C$. We recall Theorem 3 once again with p satisfying $p > 56(d+1)$ to deduces that

$$M_\rho^2 \sqrt{Pr(|\tilde{\theta}_t - \theta_*| > \epsilon_0)} \sqrt{\mathbb{E}[|x_t|^8]} \leq M_\rho^2 C \left(t^{-\frac{1}{4}} \sqrt{\log t} \right)^p t^{14(d+1)} \leq C$$

for some $C > 0$.

Hence,

$$\mathbb{E}[|x_{t+1}|^4] \leq \Delta^4 \mathbb{E}[|x_t|^4] + C,$$

and, one can conclude that

$$\mathbb{E}[|x_t|^4] < C$$

for some $C > 0$.

G.0.2 Proof of Theorem 6

It follows from [13] that J is Lipschitz continuous on Ω with a Lipschitz constant $L_J > 0$. We then estimate one of the key components of regret.

Lemma 13. *Suppose that Assumption 1,2 and 3 hold. Recall that $\bar{\Theta}_* \in \mathbb{R}^{d \times n}$ denote the matrix of the true parameter random variables, $\tilde{\Theta}_k \in \mathbb{R}^{d \times n}$ is the matrix of the parameters sampled in episode k, and $z_t := (x_t, u_t) \in \mathbb{R}^d$. Then, the following inequality holds:*

$$\begin{aligned} R_1 &:= \mathbb{E}\left[\sum_{k=1}^{n_T} \sum_{t=t_k}^{t_{k+1}-1} z_t^\top [\bar{\Theta}_* \tilde{P}_k \bar{\Theta}_*^\top - \tilde{\Theta}_k \tilde{P}_k \tilde{\Theta}_k^\top] z_t \right] \\ &\leq \sqrt{D}(2\sqrt{C} S M_P M_K^2 + \mathrm{tr}(\mathbf{W})) n_T, \end{aligned}$$

where $\tilde{P}_k := P(\tilde{\theta}_k)$ is the symmetric positive definite solution of the ARE (2.2) with $\theta := \tilde{\theta}_k$.

Proof of Lemma 13. We first observe that for any θ which satisfies $|\theta| \leq S$,

$$|z_t| = |(x_t, u_t)| = |(x_t, K(\theta)x_t + \nu_t)| = \left| \begin{bmatrix} I_n \\ K(\theta) \end{bmatrix} x_t + \nu_t \right| \leq M_K |x_t| + |\nu_t|,$$

and

$$|\tilde{P}_k^{1/2}\Theta^\top z_t| \le M_P^{1/2} S|z_t|,$$

where M_P is from Definition 2. We then consider

$$\begin{aligned}
&|\tilde{P}_k^{1/2}\bar{\Theta}_*^\top z_t|^2 - |\tilde{P}_k^{1/2}\tilde{\Theta}_k^\top z_t|^2 \\
&= (|\tilde{P}_k^{1/2}\bar{\Theta}_*^\top z_t| + |\tilde{P}_k^{1/2}\tilde{\Theta}_k^\top z_t|)(|\tilde{P}_k^{1/2}\bar{\Theta}_*^\top z_t| - |\tilde{P}_k^{1/2}\tilde{\Theta}_k^\top z_t|) \\
&\le (|\tilde{P}_k^{1/2}\bar{\Theta}_*^\top z_t| + |\tilde{P}_k^{1/2}\tilde{\Theta}_k^\top z_t|)|\tilde{P}_k^{1/2}(\bar{\Theta}_* - \tilde{\Theta}_k)^\top z_t| \\
&\le 2M_P S|z_t||(\bar{\Theta}_* - \tilde{\Theta}_k)^\top z_t|.
\end{aligned} \tag{G.4}$$

Note that

$$\Theta^\top z_t = \begin{bmatrix}\Theta(1) & \cdots & \Theta(d)\end{bmatrix}^\top z_t \in \mathbb{R}^n.$$

Thus, with $< x, y >$ denoting the inner product of two vectors $x, y \in \mathbb{R}^d$,

$$\begin{aligned}
|(\bar{\Theta}_* - \tilde{\Theta}_k)^\top z_t|^2 &= \sum_{i=1}^{d} | < (\bar{\Theta}_* - \tilde{\Theta}_k)(i), z_t > |^2 \\
&\le \sum_{i=1}^{d} |(\bar{\Theta}_* - \tilde{\Theta}_k)(i)|^2 |z_t|^2 \\
&\le |z_t|^2 \sum_{i=1}^{d} |(\bar{\Theta}_* - \tilde{\Theta}_k)(i)|^2 \\
&= |z_t|^2 |\bar{\theta}_* - \tilde{\theta}_k|^2.
\end{aligned} \tag{G.5}$$

Combining (G.4) and (G.5) yields that

$$\begin{aligned}
R_1 &\le 2M_P S\mathbb{E}\Big[\sum_{k=1}^{n_T}\sum_{t=t_k}^{t_{k+1}-1} |z_t|^2|\bar{\theta}_* - \tilde{\theta}_k|\Big] \\
&\le 2M_P S\Big(M_K^2\mathbb{E}\Big[\sum_{k=1}^{n_T}\sum_{t=t_k}^{t_{k+1}-1} |x_t|^2|\bar{\theta}_* - \tilde{\theta}_k|\Big] + \mathbb{E}\Big[\sum_{k=1}^{n_T}\sum_{t=t_k}^{t_{k+1}-1} |\nu_t|^2|\bar{\theta}_* - \tilde{\theta}_k|\Big]\Big).
\end{aligned} \tag{G.6}$$

Invoking the Cauchy-Schwarz inequality, we have

$$\mathbb{E}[|x_t|^2|\bar{\theta}_* - \tilde{\theta}_k|] \le \sqrt{\mathbb{E}[|x_t|^4]\mathbb{E}[|\bar{\theta}_* - \tilde{\theta}_k|^2]}.$$

It follows from the tower rule together with Proposition 1 that

$$\sqrt{\mathbb{E}[|\bar{\theta}_* - \tilde{\theta}_k|^2]} = \sqrt{\mathbb{E}[\mathbb{E}_{\bar{\theta}_* \sim \mu_k, \tilde{\theta}_k \sim \tilde{\mu}_k}[|\bar{\theta}_* - \tilde{\theta}_k|^2 | h_{t_k}]]} \leq \sqrt{\frac{D}{\max\{\lambda_{\min,k}, t_k\}}} \leq \sqrt{\frac{D}{t_k}},$$

where $D = 66\frac{dn}{m}$. Similarly, second term of (G.6) is bounded as

$$\begin{aligned}
\mathbb{E}\Bigg[\sum_{k=1}^{n_T} \sum_{t=t_k}^{t_{k+1}-1} |\nu_t|^2 |\bar{\theta}_* - \tilde{\theta}_k|\Bigg] &\leq \operatorname{tr}(\mathbf{W}) \sum_{k=1}^{n_T} \sum_{t=t_k}^{t_{k+1}-1} \mathbb{E}[|\bar{\theta}_* - \tilde{\theta}_k|] \\
&\leq \operatorname{tr}(\mathbf{W}) \sum_{k=1}^{n_T} \sum_{t=t_k}^{t_{k+1}-1} \sqrt{\mathbb{E}[|\bar{\theta}_* - \tilde{\theta}_k|^2]} \\
&\leq \operatorname{tr}(\mathbf{W}) \sqrt{D} \sum_{k=1}^{n_T} \sum_{t=t_k}^{t_{k+1}-1} \frac{1}{\sqrt{t_k}}.
\end{aligned}$$

Now putting these together with Lemma 5, we obtain

$$R_1 \leq \sqrt{D}(2CM_P S M_K^2 + \operatorname{tr}(\mathbf{W})) \sum_{k=1}^{n_T} \frac{T_k}{\sqrt{t_k}}. \tag{G.7}$$

Finally, to bound $\sum_{k=1}^{n_T} \frac{T_k}{\sqrt{t_k}}$, we recall that $T_k = k + 1$ and $t_k = t_{k-1} + T_{k-1}$. Thus, $t_k = \frac{T_k(T_k+1)}{2}$. Then, the sum $\sum_{k=1}^{n_T} \frac{T_k}{\sqrt{t_k}}$ is bounded as follows:

$$\sum_{k=1}^{n_T} \frac{T_k}{\sqrt{t_k}} \leq \sum_{k=1}^{n_T} \frac{\sqrt{2}T_k}{\sqrt{T_k(T_k+1)}} \leq \sum_{k=1}^{n_T} \sqrt{2} = \sqrt{2} n_T. \tag{G.8}$$

Therefore, the result follows. □

Combining Proposition 5 and Lemma 13, we finally prove Theorem 6, which yields the $O(\sqrt{T})$ regret bound. Recall that the system parameter sampled in Algorithm 1 is denoted by $\tilde{\theta}_k$, which is used in obtaining the control gain matrix $K_k = K(\tilde{\theta}_k)$ for $t \in [t_k, t_{k+1})$. Let $\tilde{P}_k := P(\tilde{\theta}_k)$ for brevity and $\tilde{u}_t = K_k x_t$ be an optimal action for $\tilde{\theta}_k$. Fix an arbitrary $t \in [t_k, t_{k+1})$. Then, the Bellman equation for t in episode k is given by

$$\begin{aligned}
&J(\tilde{\theta}_k) + x_t^\top \tilde{P}_k x_t \\
&= x_t^\top Q x_t + \tilde{u}_t^\top R \tilde{u}_t + \mathbb{E}[(\tilde{A}_k x_t + \tilde{B}_k \tilde{u}_t + w_t)^\top \tilde{P}_k (\tilde{A}_k x_t + \tilde{B}_k \tilde{u}_t + w_t) \mid h_t] \\
&= x_t^\top Q x_t + \tilde{u}_t^\top R \tilde{u}_t + (\tilde{A}_k x_t + \tilde{B}_k \tilde{u}_t)^\top \tilde{P}_k (\tilde{A}_k x_t + \tilde{B}_k \tilde{u}_t) + \mathbb{E}[w_t^\top \tilde{P}_k w_t \mid h_t],
\end{aligned} \tag{G.9}$$

where the expectation is taken with respect to w_t, and the second inequality holds because the mean of w_t is zero. On the other hand, the observed next state is expressed as

$$x_{t+1} = \bar{\Theta}_*^\top z_t + w_t,$$

where $\bar{\Theta}_* \in \mathbb{R}^{d \times n}$ is the matrix of the true parameter random variables. We then notice that

$$\begin{aligned} &\mathbb{E}[w_t^\top \tilde{P}_k w_t \mid h_t] \\ &= \mathbb{E}[x_{t+1}^\top \tilde{P}_k x_{t+1} \mid h_t] - (\bar{\Theta}_*^\top z_t)^\top \tilde{P}_k (\bar{\Theta}_*^\top z_t). \end{aligned} \tag{G.10}$$

Plugging (G.10) into (G.9) and rearranging it,

$$\begin{aligned} x_t^\top Q x_t + \tilde{u}_t^\top R \tilde{u}_t = J(\tilde{\theta}_k) + x_t^\top \tilde{P}_k x_t - \mathbb{E}[x_{t+1}^\top \tilde{P}_k x_{t+1} \mid h_t] \\ + (\bar{\Theta}_*^\top z_t)^\top \tilde{P}_k (\bar{\Theta}_*^\top z_t) - (\tilde{A}_k x_t + \tilde{B}_k \tilde{u}_t)^\top \tilde{P}_k (\tilde{A}_k x_t + \tilde{B}_k \tilde{u}_t). \end{aligned} \tag{G.11}$$

Since $\tilde{u}_t = u_t - \nu_t$, we derive that

$$\tilde{u}_t^\top R \tilde{u}_t = u_t^\top R u_t - \nu_t^\top R \tilde{u}_t - \tilde{u}_t^\top R \nu_t - \nu_t^\top R \nu_t, \tag{G.12}$$

and

$$\begin{aligned} &(\tilde{A}_k x_t + \tilde{B}_k \tilde{u}_t)^\top \tilde{P}_k (\tilde{A}_k x_t + \tilde{B}_k \tilde{u}_t) \\ &= (\bar{\Theta}_k^\top z_t)^\top \tilde{P}_k (\bar{\Theta}_k^\top z_t) - (\tilde{B}_k \nu_t)^\top \tilde{P}_k (\tilde{A}_k x_t) - (\tilde{A}_k x_t)^\top \tilde{P}_k (\tilde{B}_k \nu_t) \\ &\quad - (\tilde{B}_k \nu_t)^\top \tilde{P}_k (\tilde{B}_k \tilde{u}_t) - (\tilde{B}_k \tilde{u}_t) \tilde{P}_k (\tilde{B}_k \nu_t) - \nu_t^\top \tilde{B}_k^\top \tilde{P}_k \tilde{B}_k \nu_t. \end{aligned} \tag{G.13}$$

Combining (G.11), (G.12) and (G.13), we conclude that

$$\begin{aligned} &\mathbb{E}[c(x_t, u_t)] \\ &= J(\tilde{\theta}_k) + x_t^\top \tilde{P}_k x_t - \mathbb{E}[x_{t+1}^\top \tilde{P}_k x_{t+1} \mid h_t] \\ &\quad + (\bar{\Theta}_*^\top z_t)^\top \tilde{P}_k (\bar{\Theta}_*^\top z_t) - (\bar{\Theta}_k^\top z_t)^\top \tilde{P}_k (\bar{\Theta}_k^\top z_t) + \mathbb{E}[\nu_t^\top \tilde{B}_k^\top \tilde{P}_k \tilde{B}_k \nu_t] + \mathbb{E}[\nu_t^\top R \nu_t], \end{aligned}$$

where the expectation is taken with respect to w_t and ν_t.

Using this expression and observing $t_{n_T} \leq T \leq t_{n_T+1} - 1$, the expected regret of Algorithm 1 is decomposed as

$$\begin{aligned} R(T) &= \mathbb{E}\Bigg[\sum_{k=1}^{n_T}\sum_{t=t_k}^{t_{k+1}-1}(c(x_t,u_t) - J(\bar{\theta}_*))\Bigg] - \mathbb{E}\Bigg[\sum_{t=T+1}^{t_{n_T+1}-1}(c(x_t,u_t) - J(\bar{\theta}_*))\Bigg] \\ &:= R_1 + R_2 + R_3 + R_4 + R_5, \end{aligned}$$

where

$$\begin{aligned} R_1 &= \mathbb{E}\Bigg[\sum_{k=1}^{n_T}\sum_{t=t_k}^{t_{k+1}-1} z_t^\top[\bar{\Theta}_*\tilde{P}_k\bar{\Theta}_*^\top - \tilde{\Theta}_k\tilde{P}_k\tilde{\Theta}_k^\top]z_t\Bigg], \\ R_2 &= \mathbb{E}\Bigg[\sum_{k=1}^{n_T}\sum_{t=t_k}^{t_{k+1}-1}(x_t^\top\tilde{P}_k x_t - \mathbb{E}[x_{t+1}^\top\tilde{P}_k x_{t+1}|h_t])\Bigg], \\ R_3 &= \mathbb{E}\Bigg[\sum_{k=1}^{n_T} T_k(J(\tilde{\theta}_k) - J(\bar{\theta}_*))\Bigg], \\ R_4 &= \mathbb{E}\Bigg[\sum_{k=1}^{n_T}\sum_{t=t_k}^{t_{k+1}-1}(\nu_t^\top\tilde{B}_k^\top\tilde{P}_k\tilde{B}_k\nu_t + \nu_t^\top R\nu_t)\Bigg], \\ R_5 &= \mathbb{E}\Bigg[\sum_{t=T+1}^{t_{n_T+1}-1}(J(\bar{\theta}_*) - c(x_t,u_t))\Bigg]. \end{aligned}$$

To obtain the exact regret bound, we include R_5 which is not considered in [1]. By Lemma 13, R_1 is bounded as

$$R_1 \leq \sqrt{D}(2CSM_PM_K^2 + \mathrm{tr}(\mathbf{W}))n_T.$$

Since $T_k = k+1$, we have

$$T \geq 1 + \sum_{k=1}^{n_T-1} T_k = \frac{n_T(n_T+1)}{2} \geq \frac{n_T^2}{2},$$

which implies that

$$n_T \leq \sqrt{2T}. \tag{G.14}$$

Therefore, we conclude that

$$R_1 \leq \sqrt{2D}(2CSM_PM_K^2 + \mathrm{tr}(\mathbf{W}))\sqrt{T}.$$

Regarding R_2, we use the tower rule $\mathbb{E}[\mathbb{E}[X_t|h_t]] = \mathbb{E}[X_t]$ to obtain

$$\begin{aligned}
R_2 &= \mathbb{E}\Bigg[\sum_{k=1}^{n_T}\sum_{t=t_k}^{t_{k+1}-1}(x_t^\top \tilde{P}_k x_t - x_{t+1}^\top \tilde{P}_k x_{t+1})\Bigg] \\
&= \mathbb{E}\Bigg[\sum_{k=1}^{n_T}(x_{t_k}^\top \tilde{P}_k x_{t_k} - x_{t_{k+1}}^\top \tilde{P}_k x_{t_{k+1}})\Bigg] \\
&\leq \mathbb{E}\Bigg[\sum_{k=1}^{n_T} x_{t_k}^\top \tilde{P}_k x_{t_k}\Bigg] \\
&\leq \mathbb{E}\Bigg[\sum_{k=1}^{n_T} M_P |x_{t_k}|^2\Bigg] \\
&\leq M_P C n_T \quad (\because \text{Theorem 5}) \\
&\leq M_P C\sqrt{2T},
\end{aligned}$$

where the last inequality follows from (G.14).

We also need to deal with R_3 carefully. What is different from the analysis presented in [1], the term simply vanishes using the intrinsic property of probability matching of Thompson sampling as exact posterior distributions are used. However, in our analysis, approximate posterior is considered instead so a different approach is required. To cope with this problem, we adopt the notion of Lipschitz continuity of J for estimation. Specifically,

$$\begin{aligned}
R_3 &\leq \mathbb{E}\Bigg[\sum_{k=1}^{n_T} T_k |J(\tilde{\theta}_k) - J(\bar{\theta}_*)|\Bigg] \\
&\leq \mathbb{E}\Bigg[\sum_{k=1}^{n_T} T_k L_J |\tilde{\theta}_k - \bar{\theta}_*|\Bigg] \\
&= \sum_{k=1}^{n_T} T_k L_J \mathbb{E}\big[\mathbb{E}[|\tilde{\theta}_k - \bar{\theta}_*||h_{t_k}]\big] \\
&\leq \sum_{k=1}^{n_T} T_k L_J \mathbb{E}\big[\mathbb{E}[|\tilde{\theta}_k - \bar{\theta}_*|^2|h_{t_k}]^{\frac{1}{2}}\big] \\
&\leq \sum_{k=1}^{n_T} L_J\sqrt{D} T_k \frac{1}{\sqrt{t_k}},
\end{aligned}$$

where L_J is a Lipschitz constant of J and the last inequality follows from Proposition 1

with $D = 66\frac{dn}{m}$.

Using the bound (G.8) of $\sum_{k=1}^{n_T} \frac{T_k}{\sqrt{t_k}}$ in the proof of Lemma 13, we have

$$\begin{aligned} R_3 &\leq \sqrt{2} L_J \sqrt{D} n_T \\ &\leq 2 L_J \sqrt{D} \sqrt{T}. \end{aligned}$$

By the definition of ν_t, R_4 is bounded as

$$\begin{aligned} R_4 &= \mathbb{E}\left[\sum_{k=1}^{n_T} \sum_{t=t_k}^{t_{k+1}-1} (\nu_t^\top \tilde{B}_k^\top \tilde{P}_k \tilde{B}_k \nu_t + \nu_t^\top R \nu_t) \right] \\ &\leq \mathbb{E}\left[\sum_{k=1}^{n_T} \sum_{t=t_k}^{t_{k+1}-1} (S^2 M_P + |R|) |\nu_t|^2 \right] \\ &\leq \sum_{k=1}^{n_T} (S^2 M_P + |R|) \mathrm{tr}(\mathbf{W}) \\ &\leq (S^2 M_P + |R|) \mathrm{tr}(\mathbf{W}) n_T \\ &\leq (S^2 M_P + |R|) \mathrm{tr}(\mathbf{W}) \sqrt{2T}, \end{aligned}$$

where M_P is from Definition 2. Lastly, R_5 is bounded as

$$\begin{aligned} R_5 &= \mathbb{E}\left[\sum_{t=T+1}^{t_{n_T+1}-1} (J(\bar{\theta}_*) - c(x_t, u_t)) \right] \\ &\leq \mathbb{E}\left[\sum_{t=T+1}^{t_{n_T+1}-1} J(\bar{\theta}_*) \right] \\ &\leq (t_{n_T+1} - T - 1) M_J \\ &\leq (T_{n_T} - 1) M_J \quad (\because t_{n_T} \leq T \leq t_{n_T+1} - 1) \\ &\leq M_J n_T \\ &\leq M_J \sqrt{2T}, \end{aligned}$$

where M_J is from Definition 2. Putting all the bounds together, we conclude that

$$R(T) \leq C\sqrt{T},$$

and thus the result follows. One novelty in our analysis is that the concentration of approximate posterior is naturally embedded into the analysis, which eventually drops the $\log T$ term in the resulting regret.

초 록

톰슨 샘플링(Thompson sampling)은 온라인 학습 문제에서 탐색과 활용 사이의 균형을 맞추는 데 널리 사용되는 방법으로, 이에는 선형 이차 제어기 (Linear Quadratic Regulator)를 위한 강화 학습을 포함한다. 그러나 선형 이차 제어기 학습에 사용되는 톰슨 샘플링의 이론적 분석은 종종 가우시안 잡음의 경우에만 제한되는 경우가 많다. 또한, 우리는 알려진 시스템 파라미터가 미리 지정된 한정된 집합에 속한다는 가정을 더할 때 샘플링을 직접 수행할 수 있으며, 이는 제한적인 것으로 보인다[1]. 이에 우리는 선형 이차 제어기를 위한 새로운 톰슨 샘플링 알고리즘을 제안하며, 비가우시안 잡음을 포함한 더 넓은 범위의 문제를 다루기 위해 랑주뱅 동역학(Langevin dynamics)를 활용하려 한다. 또한, 특정 초기화 방법이나 실제 시스템 파라미터에 대한 정보를 필요로 하지 않으면서도, 사전 분포와 허용 가능한 집합에 대한 최소한의 가정만으로 우리의 알고리즘은 근사 사후 분포로부터 빠르게 샘플링할 수 있다. 우리 알고리즘은 $O(\sqrt{T})$의 기대 후회(regret) 상한을 가지며, 이는 [1]의 알고리즘 성능보다 개선된 결과이다. 또한, 우리의 알고리즘 성능 분석은 자기 정규화 기법과 함께 사전 조건화된 랑주뱅 동역학의 수렴 부등식을 활용한다. 우리 알고리즘의 성능은 수치 실험을 통해 입증되었다.

주요어: 선형 2차 제어기, 톰슨 샘플링, 랑주뱅 동역학

학번: 2020-26137

ACKNOWLEGEMENT

First and foremost, I would like to express my deepest gratitude to Professor Insoon Yang, my advisor, for his invaluable guidance and continuous support throughout the completion of this paper. His constant assistance has allowed me to establish the proper direction for the research and together we have navigated through challenges by finding suitable solutions. Furthermore, his rigorous and meticulous feedback has enabled me to conduct more refined research.

I would also like to extend my gratitude to Professor Yeoneung Kim, who collaborated with me on this paper and provided valuable assistance. With his guidance, I was able to articulate the paper logically and present a more rigorous description of the mathematical proofs.

Lastly, I want to express my heartfelt appreciation to my parents and my sister. Throughout my graduate studies and the process of writing this paper, they have shown unwavering dedication and love, providing me with the support I needed to persevere.