Langevin Monte Carlo for Contextual Bandits

Pan Xu [1] Hongkai Zheng [1] Eric Mazumdar [1] Kamyar Azizzadenesheli [2] Anima Anandkumar [1]

Abstract

We study the efficiency of Thompson sampling for contextual bandits. Existing Thompson sampling-based algorithms need to construct a Laplace approximation (i.e., a Gaussian distribution) of the posterior distribution, which is inefficient to sample in high dimensional applications for general covariance matrices. Moreover, the Gaussian approximation may not be a good surrogate for the posterior distribution for general reward generating functions. We propose an efficient posterior sampling algorithm, viz., Langevin Monte Carlo Thompson Sampling (LMC-TS), that uses Markov Chain Monte Carlo (MCMC) methods to directly sample from the posterior distribution in contextual bandits. Our method is computationally efficient since it only needs to perform noisy gradient descent updates without constructing the Laplace approximation of the posterior distribution. We prove that the proposed algorithm achieves the same sublinear regret bound as the best Thompson sampling algorithms for a special case of contextual bandits, viz., linear contextual bandits. We conduct experiments on both synthetic data and real-world datasets on different contextual bandit models, which demonstrates that directly sampling from the posterior is both computationally efficient and competitive in performance.

1. Introduction

A bandit problem is a sequential decision-making problem wherein an agent, in each round, observes an action set, chooses an action (or arm) from the set, and then observes a reward from the environment. A bandit learning algorithm aims to learn a policy for the agent to maximize its cumulative rewards based on its historical observations of actions and rewards. In the vast majority of real-world applications, each arm is usually associated with side information in the form of a feature or context vector that describes the arm. The mean reward for an arm is expressed as some unknown function of the arm's feature vector and an unknown weight parameter that is shared across different arms. This setting— known as the contextual bandit problem— has been extensively studied in the literature (Langford & Zhang, 2007; Chu et al., 2011; Abbasi-Yadkori et al., 2011; Agrawal & Goyal, 2013; Filippi et al., 2010; Li et al., 2017; Lale et al., 2019; Kveton et al., 2020).

The main challenge in contextual bandit problems is addressing the well known exploitation versus exploration trade-off, which requires a careful balance between choosing the myopically better arm and choosing an under-sampled worse arm. Existing algorithms for maximizing the cumulative reward in bandit problems mainly follow either one of the following two algorithmic frameworks. The first framework follows the principle of optimism in the face of uncertainty (OFU), and algorithms designed using such ideas have been widely applied to both finite armed bandits, also known as multi-armed bandits (MAB) (Auer et al., 2002; Ménard & Garivier, 2017), and contextual bandits (Chu et al., 2011; Abbasi-Yadkori et al., 2011; Li et al., 2017; Zhou et al., 2020; Xu et al., 2022). The second dominant category of bandit algorithm makes use of the idea of Thompson or posterior sampling (Thompson, 1933). Such algorithms have been widely used in practice due to their ease of implementation and impressive empirical performance, and have only recently started to be well understood theoretically in multi-armed bandits (Agrawal & Goyal, 2012; Kaufmann et al., 2012; Russo & Van Roy, 2014; Jin et al., 2021) and contextual bandits (Chapelle & Li, 2011; Agrawal & Goyal, 2013; Riquelme et al., 2018; Wang & Zhou, 2020; Zhang et al., 2021).

One crucial area in which the two types of algorithms differ is in their ease of implementation. In contextual bandit problems, algorithms based on the OFU principle usually need to solve a bi-linear optimization problem, making them computationally expensive to implement outside of simple problems despite coming with stronger theoretical guarantees. In contrast, Thompson sampling algorithms only need to solve a linear optimization problem over the arm set since

[1] Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA [2] Department of Computer Science, Purdue University, West Lafayette, IN, USA. Correspondence to: Pan Xu <pan.xu@duke.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

the uncertainty in the posterior distribution automatically accounts for the exploration in the parameter space. Furthermore, Thompson sampling has been observed to be empirically competitive to—or sometimes even better—than OFU based algorithms (Chapelle & Li, 2011).

Most existing Thompson sampling algorithms first construct a Laplace approximation (which is essentially a Gaussian distribution) (Chapelle & Li, 2011) of the underlying posterior distribution on the data and then sample from the Gaussian distribution to explore in the parameter space. The Laplace approximation in Thompson sampling usually leads to a non-isotropic covariance matrix. It is well known that sampling from a Gaussian distribution with a general covariance matrix is usually computationally expensive in high dimensional applications. Moreover, when the reward generating function is nonlinear with respect to the weight parameter such as in generalized linear bandits (Kveton et al., 2020) and neural contextual bandits (Riquelme et al., 2018; Zhang et al., 2021), the true posterior distribution may not be well approximated by a Gaussian distribution and thus the Laplace approximation could be a poor surrogate for the posterior distribution.

Our approach: In this paper, we propose an algorithm, viz., Langevin Monte Carlo Thompson sampling (LMC-TS), which directly samples from the data posterior distribution instead of a Laplace approximation in contextual bandits. In particular, by incorporating Langevin Monte Carlo (Bakry et al., 2014), our algorithm only needs to perform noisy gradient descent updates, which can generate samples that provide a good approximation of the posterior with arbitrary accuracy if it is run for sufficiently many steps. This is contrast with Laplace approximation for Thompson sampling (Chapelle & Li, 2011), which has a fixed approximation error for the posterior distribution and thus the covariance matrix needs to be carefully redesigned in different contextual bandit problems to achieve reasonable performance (Chapelle & Li, 2011; Kveton et al., 2020; Riquelme et al., 2018; Zhang et al., 2021). Moreover, due to the simplicity of noisy gradient descent updates, the proposed algorithm is directly applicable to many bandit problems where deep neural network function classes are used.

Contributions of this paper are summarized as follows:

- We propose a practical and efficient bandit algorithm LMC-TS, which only needs to perform noisy gradient descent updates to approximately sample from the data posterior distribution. LMC-TS is easily implementable and scalable to large-scale and high dimensional problems including deep learning applications. It also works simultaneously for a large class of contextual bandit models including linear contextual bandits, generalized linear bandits, and neural contextual bandits.
- We theoretically prove that LMC-TS achieves a $\widetilde{O}(d\sqrt{dT})$ regret for linear contextual bandits, where d is the dimension of the problem and T is the time horizon. This result matches the best regret bound for Thompson sampling algorithms in linear contextual bandits (Agrawal & Goyal, 2013).
- We further conduct thorough experiments on both synthetic datasets and real-world datasets (UCI machine learning datasets and a high dimensional image dataset CIFAR10) to show that one algorithm is enough for learning many different complex bandit models by comparing it with different baseline algorithms in linear contextual bandits, generalized bandits, and neural contextual bandits respectively.

Notation We use $[k]$ to denote a set $\{1, \ldots, k\}$, $k \in \mathbb{N}^+$. $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^\top \mathbf{x}}$ is the Euclidean norm of a vector $\mathbf{x} \in \mathbb{R}^d$. For a matrix $\mathbf{V} \in \mathbb{R}^{m\times n}$, we denote by $\|\mathbf{V}\|_2$ and $\|\mathbf{V}\|_F$ its operator norm and Frobenius norm respectively. For a semi-positive definite matrix $\mathbf{V} \in \mathbb{R}^{d\times d}$ and a vector $\mathbf{x} \in \mathbb{R}^d$, we denote the Mahalanobis norm as $\|\mathbf{x}\|_{\mathbf{V}} = \sqrt{\mathbf{x}^\top \mathbf{V}\mathbf{x}}$. For an event E on a probability space, we denote E^c as its complement event such that $\mathbb{P}(E) + \mathbb{P}(E^c) = 1$. For a function $f(T)$, we use the common big O notation $O(f(T))$ to hide constant factors with respect to T and use $\widetilde{O}(f(T))$ to omit the logarithmic dependence on T.

2. Preliminary

Contextual Bandits Contextual bandits are a wide class of sequential decision problems, where the player makes the decision based on an observation of an action set consisting of feature vectors as contexts for different actions. In particular, at round t, the player observes an action set $\mathcal{X}_t \subseteq \mathbb{R}^d$, and chooses an arm or action which is represented by a feature vector $\mathbf{x}_t \in \mathcal{X}_t$. Note that in this paper we do not assume the action set is finite nor is fixed in each round. Then a reward r_t is immediately revealed to the agent by the environment. In contextual bandit problems, it is often assumed that the mean reward of an action with feature $\mathbf{x} \in \mathbb{R}^d$ is given by a reward generating function $f(\mathbf{x}, \boldsymbol{\theta}^*)$ and the observed reward is $r(\mathbf{x}) = f(\mathbf{x}, \boldsymbol{\theta}^*) + \xi$, where $\boldsymbol{\theta}^* \in \mathbb{R}^{d'}$ is an unknown weight parameter that is shared across all arms, and ξ is a random noise incurred in the observation. For instance, in linear contextual bandits (Chu et al., 2011; Abbasi-Yadkori et al., 2011; Agrawal & Goyal, 2013), we have $\boldsymbol{\theta}^* \in \mathbb{R}^d$ and $f(\mathbf{x}, \boldsymbol{\theta}^*) = \mathbf{x}^\top \boldsymbol{\theta}^*$; in generalized linear bandits (Filippi et al., 2010; Li et al., 2017; Kveton et al., 2020; Ding et al., 2021), we have $f(\mathbf{x}, \boldsymbol{\theta}^*) = \mu(\mathbf{x}^\top \boldsymbol{\theta}^*)$ for some link function $\mu(\cdot)$; and for neural contextual bandits (Riquelme et al., 2018; Zhou et al., 2020; Zhang et al., 2021; Xu et al., 2022), $f(\mathbf{x}, \boldsymbol{\theta}^*)$ is a neural network, where $\boldsymbol{\theta}^*$ is the concatenation of all weight parameters and $\mathbf{x}$ is the input.

The objective of a bandit algorithm is to maximize the cumulative rewards over a time horizon T, which is equivalent to minimizing the following pseudo regret (Lattimore & Szepesvári, 2020):

$$R(T) = \mathbb{E}\left[\sum_{t=1}^{T}(r(\mathbf{x}_t^*) - r(\mathbf{x}_t))\right], \qquad (2.1)$$

where $\mathbf{x}_t \in \mathcal{X}_t$ is the arm chosen by the bandit algorithm at round t, and $\mathbf{x}_t^* = \mathrm{argmax}_{\mathbf{x}\in\mathcal{X}_t} \mathbb{E}[r(\mathbf{x})]$ is the arm with the maximum expected reward at round t. Note that this definition of regret is based on the best oracle arm $\mathbf{x}_t^*$, which is more general than the definition based on the best policy achievable within a predefined policy class in adversarial bandits (Bubeck & Cesa-Bianchi, 2012).

Laplace Approximation Thompson Sampling Among the most popular bandit algorithms, Thompson sampling (Thompson, 1933; Chapelle & Li, 2011; Russo et al., 2018) is known to be simple and efficient in practice, which uses a Laplace approximation to approximate the posterior distribution of the data. After $t-1$ rounds of the bandit problem, assume we have collected data $\{\mathbf{x}_1, r_1, \mathbf{x}_2, r_2, \ldots, \mathbf{x}_{t-1}, r_{t-1}\}$. Define the following quantities based on historical data.

$$\mathbf{V}_t = \lambda\mathbf{I} + \sum_{s=1}^{t-1}\mathbf{x}_s\mathbf{x}_s^\top, \quad \mathbf{b}_t = \sum_{s=1}^{t-1} r_s\mathbf{x}_s, \qquad (2.2)$$

where $\lambda > 0$ is a regularization parameter. Denote $\widehat{\boldsymbol{\theta}}_t = \mathbf{V}_t^{-1}\mathbf{b}_t$. At round t, the agent receives an action set $\mathcal{X}_t \subseteq \mathbb{R}^d$ which consists of feature vectors of candidate actions at round t. Then linear Thompson sampling (LinTS) (Agrawal & Goyal, 2013) samples a parameter $\widetilde{\boldsymbol{\theta}}_t$ from distribution $\mathcal{N}(\widehat{\boldsymbol{\theta}}_t, v_t\mathbf{V}_t^{-1})$ and then chooses the arm as follows $\mathbf{x}_t = \mathrm{argmax}_{\mathbf{x}\in\mathcal{X}_t} \mathbf{x}^\top\widetilde{\boldsymbol{\theta}}_t$. After that, it observes the reward r_t for round t. Based on newly collected action feature $\mathbf{x}_t$ and reward r_t, the quantities in (2.2) can be updated and the learning process proceeds to the next round.

Approximating the posterior distribution using a Gaussian distribution is also called Laplace Thompson sampling (Chapelle & Li, 2011). Note that sampling from $\mathcal{N}(\widehat{\boldsymbol{\theta}}_t, v_t\mathbf{V}_t^{-1})$ is usually implemented as $\widetilde{\boldsymbol{\theta}}_t = \widehat{\boldsymbol{\theta}}_t + \sqrt{v_t}\mathbf{V}_t^{-1/2}\boldsymbol{\zeta}$ in practice, where $\boldsymbol{\zeta}$ is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $v_t > 0$ is a scaling parameter. The computation complexity of calculating $\mathbf{V}^{-1/2}$ is at least $O(d^3)$ with Cholesky decomposition, which is prohibitively high, especially for high-dimensional machine learning problems. On the other hand, the Gaussian distribution used in Thompson sampling might not be a good approximation of the posterior distribution for general bandit models with more complicated structures than linear contextual bandits.

3. Langevin Monte Carlo Thompson Sampling

Algorithm 1 Langevin Monte Carlo Thompson Sampling (LMC-TS)

```
Input: step sizes $\{\eta_t > 0\}_{t\geq 1}$, inverse temperature parameters $\{\beta_t\}_{t\geq 1}$, loss function $L_t(\boldsymbol{\theta})$, and reward model function $f(\mathbf{x}, \boldsymbol{\theta})$. $\boldsymbol{\theta}_{1,0} = \mathbf{0}$, $K_0 = 0$.
for $t = 1, 2, \ldots$ do
    $\boldsymbol{\theta}_{t,0} = \boldsymbol{\theta}_{t-1,K_{t-1}}$
    for $k = 1, \ldots, K_t$ do
        sample a standard normal vector $\boldsymbol{\epsilon}_{t,k} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
        $\boldsymbol{\theta}_{t,k} = \boldsymbol{\theta}_{t,k-1} - \eta_t\nabla L_t(\boldsymbol{\theta}_{t,k-1}) + \sqrt{2\eta_t\beta_t^{-1}}\boldsymbol{\epsilon}_{t,k}$
    end for
    Play arm $\mathbf{x}_t = \mathrm{argmax}_{\mathbf{x}\in\mathcal{X}_t} f(\mathbf{x}, \boldsymbol{\theta}_{t,K_t})$ and observe reward $r_t$
end for
```

In this paper, we propose the Langevin Monte Carlo Thompson Sampling (LMC-TS) algorithm, which is presented in Algorithm 1. Unlike existing work that use Laplace Approximation (Chapelle & Li, 2011; Agrawal & Goyal, 2013; Kveton et al., 2020; Zhang et al., 2021), which is essentially a Gaussian distribution, to approximate the unknown posterior distribution, we use Langevin Monte Carlo (Roberts & Tweedie, 1996; Bakry et al., 2014) to learn the exact posterior distribution of parameter $\boldsymbol{\theta}^*$ up to a high precision. One closely related work to ours is Mazumdar et al. (2020) which proposed to combine LMC and SGLD with Thompson sampling algorithms in finite-armed bandit problems without any contextual features. Their analysis heavily depends on their assumption on prior distributions and is hard to extend to contextual bandits (with potentially infinite arms) even for the simplest linear contextual bandits, which we will discuss in further details in the next section.

In specific, Algorithm 1 works as follows. At the t-th round of the algorithm, we run the following subroutine for K_t steps. For each $k = 1, \ldots, K_t$, we have

$$\boldsymbol{\theta}_{t,k} = \boldsymbol{\theta}_{t,k-1} - \eta\nabla L_t(\boldsymbol{\theta}_{t,k-1}) + \sqrt{2\eta_t\beta_t^{-1}}\boldsymbol{\epsilon}_{t,k}, \qquad (3.1)$$

where $\boldsymbol{\epsilon}_{t,k}$ is an isotropic Gaussian random vector in $\mathbb{R}^d$, $\eta > 0$ is a step size parameter, β_t is the inverse temperature parameter, and $L_t(\boldsymbol{\theta})$ is loss function between the observed rewards $\{r_i\}_{i=1,\ldots,t-1}$ and estimated rewards $\{f(\mathbf{x}, \boldsymbol{\theta})\}$ that is specified by the user. (3.1) is called the Langevin Monte Carlo (LMC) method in the approximate sampling literature (Roberts & Tweedie, 1996; Bakry et al., 2014; Dalalyan, 2017b;a), which could be viewed as the Euler-Maruyama discretization of the following stochastic differential equation from physics called Langevin dynamics (Langevin, 1908):

$$\mathrm{d}\boldsymbol{\theta}(s) = -\nabla L_t\big(\boldsymbol{\theta}(s)\big)\mathrm{d}s + \sqrt{2\beta_t^{-1}}\mathrm{d}\boldsymbol{B}(s), \tag{3.2}$$

where $s > 0$ is a continuous time index, $\beta > 0$ is the inverse temperature parameter and $\boldsymbol{B}(t) \in \mathbb{R}^d$ is a Brownian motion. It has been showed that under certain conditions on the drift term $-\nabla L(\boldsymbol{\theta}(t))$, Langevin dynamics will converge to a unique stationary distribution $\pi(\mathrm{d}\mathbf{x}) \propto e^{-\beta L(\mathbf{x})}\mathrm{d}\mathbf{x}$. Therefore, one can use (1) to approximately sample from an arbitrary distribution $\pi_t \propto \exp(-\beta_t L_t(\boldsymbol{\theta}))$.

Note that Algorithm 1 is applicable to various different contextual bandit settings if we choose the corresponding reward model $f(\mathbf{x}, \boldsymbol{\theta})$ and log-density function $L_t(\boldsymbol{\theta})$. Another advantage of our algorithm is that only noisy gradient descent update is performed in order to do proper exploration in different bandit problems. Thus our LMC-TS is both flexible in design and is easy to implement in practice. In the next few subsections, we show that how we can instantiate our algorithm for linear contextual bandits, generalized linear bandits, and neural contextual bandits respectively.

3.1. Implication to Linear Contextual Bandits

In linear contextual bandits, it is assumed that the reward generating function is $f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\theta}^*$ for all $\mathbf{x} \in \mathcal{X}$. Define the following loss function

$$L_t(\boldsymbol{\theta}) = \sum_{i=1}^{t-1} \big(\mathbf{x}_i^\top \boldsymbol{\theta} - r_i\big)^2 + \lambda \|\boldsymbol{\theta}\|^2, \tag{3.3}$$

where $\lambda > 0$ is a regularization parameter. Then we have the gradient of $L_t(\boldsymbol{\theta})$ as $\nabla L_t(\boldsymbol{\theta}) = 2(\mathbf{V}_t \boldsymbol{\theta} - \mathbf{b}_t)$, where $\mathbf{V}_t$ and $\mathbf{b}_t$ are defined in the same way as in (2.2).

Based on the linear bandit model and the loss function chosen in (3.3), we can show that the inner loop of Algorithm 1 generates samples approximately from the Gaussian posterior distribution.

Proposition 3.1. *If the epoch length K_t in Algorithm 1 is sufficiently large, the distribution of K_t converges to Gaussian distribution $\mathcal{N}(\mathbf{V}_t^{-1}\mathbf{b}_t, \beta_t^{-1}\mathbf{V}_t^{-1})$ up to an arbitrary accuracy.*

Note that LMC does not converge exactly to the posterior distribution but instead converges to it with an arbitrarily pre-chosen prevision parameter for large enough K_t. In the next section, we will show this will be sufficient for the proposed bandit algorithm to achieve a sublinear regret.

Proof. According to Roberts & Tweedie (1996); Bakry et al. (2014), we know the Markov chain generated by Langevin dynamics (3.2) converges to a stationary distribution π_t, which is defined as $\pi_t(\boldsymbol{\theta}) = Z^{-1}\exp(-\beta_t L_t(\boldsymbol{\theta}))$, where $Z = \int \exp(-\beta_t L_t(\boldsymbol{\theta}))\mathrm{d}\boldsymbol{\theta}$ is the normalization term. In the inner loop of Algorithm 1, we apply the LMC update defined in (3.1) which is a discretization of (3.2) and thus obtain another Markov chain $\{\boldsymbol{\theta}_{t,k}\}_{k=0,1,\ldots}$. A recent line of non-asymptotic analyses show that the Markov chain $\{\boldsymbol{\theta}_{t,k}\}_{k=0,1,\ldots}$ generated by LMC converges to π_t up to an arbitrary accuracy for (strongly)-convex L_t (Dalalyan, 2017b) and nonconvex L_t (Vempala & Wibisono, 2019) respectively, as long as the epoch length K_t of Algorithm 1 is large enough.

In what follows, we show that π_t is the Gaussian distribution we need in linear contextual bandits. By the definition in (3.3), we have

$$\begin{aligned}
L_t(\boldsymbol{\theta}) &= \sum_{i=1}^{t-1} \big(\mathbf{x}_i^\top \boldsymbol{\theta} - r_i\big)^2 + \lambda\|\boldsymbol{\theta}\|^2 \\
&= \sum_{i=1}^{t-1} \big(\boldsymbol{\theta}^\top \mathbf{x}_i \mathbf{x}_i^\top \boldsymbol{\theta} - \langle \boldsymbol{\theta}, 2 r_i \mathbf{x}_i\rangle + r_i^2\big) + \lambda\|\boldsymbol{\theta}\|^2 \\
&= \boldsymbol{\theta}^\top \Big[\lambda \mathbf{I} + \sum_{i=1}^{t-1} \mathbf{x}_i \mathbf{x}_i^\top\Big]\boldsymbol{\theta} - 2\Big\langle \boldsymbol{\theta}, \sum_{i=1}^{t-1} r_i \mathbf{x}_{a_i}\Big\rangle + \sum_{i=1}^{t-1} r_i^2 \\
&= \boldsymbol{\theta}^\top \mathbf{V}_t \boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \mathbf{b}_t + \sum_{i=1}^{t-1} r_i^2,
\end{aligned}$$

where the last equality is due to (2.2). We denote $\widehat{\boldsymbol{\theta}}_t = \mathrm{argmin}_{\boldsymbol{\theta}}\, L_t(\boldsymbol{\theta})$ which is the solution of ridge regression (Hoerl & Kennard, 1970). It is easy to verify that the solution of the ridge regression problem has the form $\widehat{\boldsymbol{\theta}}_t = \mathbf{V}_t^{-1}\mathbf{b}_t$. Then, we have

$$\begin{aligned}
\big(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\big)^\top \mathbf{V}_t \big(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\big) &= \boldsymbol{\theta}^\top \mathbf{V}_t \boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \mathbf{V}_t \widehat{\boldsymbol{\theta}}_t + \widehat{\boldsymbol{\theta}}_t^\top \mathbf{V}_t \widehat{\boldsymbol{\theta}}_t \\
&= \boldsymbol{\theta}^\top \mathbf{V}_t \boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \mathbf{b}_t + \widehat{\boldsymbol{\theta}}_t^\top \mathbf{V}_t \widehat{\boldsymbol{\theta}}_t,
\end{aligned}$$

which immediately implies that

$$\begin{aligned}
\pi_t(\boldsymbol{\theta}) &\propto \exp(-\beta_t L_t(\boldsymbol{\theta})) \\
&\propto \exp\big(-\beta_t\big(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\big)^\top \mathbf{V}_t \big(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\big)\big).
\end{aligned}$$

Therefore, we can conclude that the distribution of $\boldsymbol{\theta}_{t,K_t}$ converges to Gaussian distribution $\mathcal{N}(\widehat{\boldsymbol{\theta}}_t, \beta_t^{-1}\mathbf{V}_t^{-1})$. □

Although the update in Algorithm 1 is presented as a full gradient descent step plus an isotropic noise, one can also replace the full gradient $\nabla L_t(\boldsymbol{\theta}_{t,k-1})$ with a stochastic gradient or a variance reduced stochastic gradient of the loss function $L_t(\boldsymbol{\theta}_{t,k-1})$ calculated from a mini-batch of data, which leads to Stochastic Gradient Langevin Dynamics (SGLD) algorithm (Welling & Teh, 2011) and Stochastic Variance Reduced Gradient Langevin Dynamics (SVRG-LD) (Dubey et al., 2016; Xu et al., 2018). And similar results to Proposition 3.1 can be obtained by following the proofs in Dalalyan & Karagulyan (2019); Xu et al. (2018); Zou et al. (2021).

Now we have shown that our Algorithm 1 is similar to the Thompson sampling algorithm derived from Laplace approximation of the posterior distribution in linear contextual bandits. Nevertheless, our Algorithm 1 is more preferable than Thompson sampling in practice since at each iteration of LMC-TS we only need the computation of first order information, which is much more computationally efficient than computing the Cholesky decomposition for sampling from a general multivariate normal distribution in Thompson sampling when the feature dimension is high as we discussed in Section 2.

3.2. Implication to Generalized Linear Bandits

In generalized linear bandits (GLB), the true reward r for arm $\mathbf{x} \in \mathcal{X}_t$ at round t is assumed to be from a generalized linear model (GLM) (McCullagh & Nelder, 2019). Specifically, conditional on feature vector $\mathbf{x}$, r follows an exponential family distribution with mean $\mu(\mathbf{x}^\top \boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^*$ is an unknown weight parameter that is shared across all arms, and $\mu(\cdot)$ is called the link function. It is worth noting that generalized linear bandits cover a class of common bandit models used in practice. For instance, when $\mu(z) = z$ is the identity function, it reduces to linear contextual bandits; when $\mu(z) = 1/(1+e^{-z})$ is the sigmoid function, it reduces to the logistic bandits (Dong et al., 2019).

Based on samples $\{\mathbf{x}_1, r_1, \ldots, \mathbf{x}_{t-1}, r_{t-1}\}$, the negative log-likelihood function is defined as $\tilde{L}_t(\boldsymbol{\theta}) = \sum_{i=1}^{t-1}(m(\mathbf{x}_{a_i}^\top \boldsymbol{\theta}) - r_i \mathbf{x}_{a_i}^\top \boldsymbol{\theta})$, where $m(z)$ is twice differentiable and $m'(z) = \mu(z)$ is the link function defined in the previous paragraph. Existing Thompson sampling based algorithms on generalized linear bandits (Filippi et al., 2010; Kveton et al., 2020; Ding et al., 2021) usually first solve the following MLE estimator

$$\hat{\boldsymbol{\theta}}_t = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{t-1} (r_i \mathbf{x}_{a_i}^\top \boldsymbol{\theta} - m(\mathbf{x}_{a_i}^\top \boldsymbol{\theta})), \tag{3.4}$$

and then construct a Laplace approximation of the underlying posterior distribution, which is given by Gaussian distribution $\mathcal{N}(\widehat{\boldsymbol{\theta}}_t, a^2(\nabla^2 \tilde{L}_t(\widehat{\boldsymbol{\theta}}_t))^{-1})$, where $a > 0$ is a scaling parameter. Similar to Thompson sampling for linear contextual bandits discussed in Section 2, sampling from $\mathcal{N}(\widehat{\boldsymbol{\theta}}_t, a^2(\nabla^2 \tilde{L}_t(\widehat{\boldsymbol{\theta}}_t))^{-1})$ is computationally inefficient in high dimensional applications. Moreover, due to the existence of the link function, the posterior itself is not necessary a Gaussian distribution, and thus the Laplace approximation might cause a fixed approximation error.

In contrast, our LMC-TS algorithm can be easily applied to GLB by choosing the reward model as $f(\mathbf{x}, \boldsymbol{\theta}^*) = \mu(\mathbf{x}^\top \boldsymbol{\theta}^*)$ and the loss function $L_t(\boldsymbol{\theta})$ as the following regularized negative log-likelihood function

$$L_t(\boldsymbol{\theta}) = \sum_{i=1}^{t-1} (m(\mathbf{x}_{a_i}^\top \boldsymbol{\theta}) - r_i \mathbf{x}_{a_i}^\top \boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_2^2, \tag{3.5}$$

where $\lambda > 0$ is a tuning parameter. Under some standard conditions on the link function in GLB, it could be shown that the posterior density $\pi_t \propto \exp(-\beta_t L_t(\boldsymbol{\theta}))$ is strongly log-concave and log-smooth. Similar to the proof of Proposition 3.1, by recalling results in the study of Langevin Monte Carlo (Dalalyan, 2017b), we can show that the distribution of iterates $\boldsymbol{\theta}_{t,K_t}$ in Algorithm 1 converges to the true posterior distribution π_t if the epoch length K_t of the inner loop is sufficiently large. In addition, due to the simplicity of our algorithm, we only need to perform gradient descent based updates, which is computationally more efficient than Laplace approximation based Thompson sampling for generalized linear bandits (Kveton et al., 2020).

3.3. Implication to Neural Contextual Bandits

Our algorithm is also applicable to more general contextual bandit problems, where the reward function $f(\mathbf{x}, \boldsymbol{\theta}^*)$ is a neural network with $\mathbf{x}$ as its input and $\boldsymbol{\theta}^*$ as the collection of all weight matrices. This type of bandit model is usually referred to as the neural contextual bandits in the literature (Riquelme et al., 2018; Zhou et al., 2020; Zhang et al., 2021; Xu et al., 2022). One possible choice of the function $L_t(\boldsymbol{\theta})$ used in Algorithm 1 is the squared loss:

$$L_t(\boldsymbol{\theta}) = \sum_{i=1}^{t-1} \big(f(\mathbf{x}_i, \boldsymbol{\theta}) - r_i\big)^2 + \lambda \|\boldsymbol{\theta}\|^2, \tag{3.6}$$

where $\lambda > 0$. Due to the flexibility in the choice of loss functions in our method, one can always choose another loss function $L_t(\boldsymbol{\theta})$ based on the belief in the prior and posterior distributions in specific applications to boost the empirical performance. This makes our method directly applicable to complicated deep learning applications.

4. Theoretical Analysis of LMC-TS for Linear Contextual Bandits

In this section, we provide the regret analysis of our proposed LMC-TS algorithm when we apply it to a specific contextual bandit problem, viz., the linear contextual bandit. We first state the assumption on the details of the model.

Assumption 4.1. There is an unknown parameter $\boldsymbol{\theta}^* \in \mathbb{R}^d$ such that for any arm $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$, the reward is $r(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\theta}^* + \xi$, where ξ is assumed to be a R-subGaussian random variable for some constant $R > 0$.

The following theorem states the regret bound of LMC-TS.

Theorem 4.2. *Under Assumption 4.1, we choose a linear reward model $f(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{x}^\top \boldsymbol{\theta}$ in Algorithm 1. Let $\delta \in (0, 1)$.*

For any $j = 1, 2, \ldots$, let the step size $\eta_j = 1/(4\lambda_{\max}(\mathbf{V}_t))$, the epoch length $K_j = \kappa_j \log(3R\sqrt{2dT\log(T^3/\delta)})$, and the inverse temperature $\beta_j^{-1} = 4(R\sqrt{d\log(T^3/\delta)})$, where $\kappa_j = \lambda_{\max}(\mathbf{V}_j)/\lambda_{\min}(\mathbf{V}_j)$ is the condition number of $\mathbf{V}_j$. Then with probability $1-\delta$, it holds that

$$R(T) \le CRd\log(1/\delta)\sqrt{dT\log^3(1+T/(\lambda d))},$$

where $C > 0$ is an absolute constant that is independent of the problem.

Note that the regret upper bound in Theorem 4.2 is in the order of $\widetilde{O}(d\sqrt{dT})$ which matches the best result for Thompson sampling based algorithm in linear contextual bandits with infinite arms (Agrawal & Goyal, 2013). Moreover, if the arm set $|\mathcal{X}_t| \le N$ is finite in each round for some integer N, following a similar proof as in Agrawal & Goyal (2013), this regret bound can be improved to $\widetilde{O}(d\sqrt{T})$, where a logarithmic dependence on the number of arms N replaces the additional term $O(\sqrt{d})$ and is omitted in the $\widetilde{O}(\cdot)$ notation. This shows that LMC-TS is theoretically comparable to Laplace approximation based Thompson sampling in linear contextual bandits. Nevertheless, due to the fact that we add multiple noises in Algorithm 1 at different time steps, the coupling of these noises makes the regret analysis of existing Laplace approximation Thompson sampling not directly applicable to our case. In specific, it has been shown by Phan et al. (2019) that the approximation error caused by the posterior sampling step for Thompson sampling can yield a linear regret in general and thus we have to develop nontrivial proof techniques to achieve a sublinear regret for LMC-TS.

Remark 4.3. The only existing work that study the sublinear regret of TS with approximate sampling is given by Mazumdar et al. (2020). Compared with their work which combines Langevin algorithms with Thompson sampling for multi-armed bandits, our analysis applies to a more general class of bandit problems which covers MAB as a special case. Moreover, we do not assume that the reward distribution of a single data point is strongly log-concave, which is hard to be justified in contextual bandits. Specifically, they assume that the reward r conditional on context $\mathbf{x}_i$ has a distribution $p(r|\mathbf{x}, \boldsymbol{\theta}^*)$ which is strongly log-concave w.r.t. both $\mathbf{x}$ and $\boldsymbol{\theta}^*$. However, this assumption is not satisfied even for Gaussian rewards. For instance, if r is a Gaussian reward with mean $\mathbf{x}^\top\boldsymbol{\theta}^*$ and unit variance, then $-\log p(r|\mathbf{x}, \boldsymbol{\theta}^*)$ is not strongly convex with respect to $\mathbf{x} \in \mathbb{R}^d$ when the dimension d is larger than 1. In contrast, we only assume that the reward is subGaussian, which is among the most common assumptions in linear contextual bandits (Abbasi-Yadkori et al., 2011; Agrawal & Goyal, 2013).

Remark 4.4. We note that the epoch length of the inner loop of Algorithm 1 depends on the condition number $\kappa_j = \lambda_{\max}(\mathbf{V}_j)/\lambda_{\min}(\mathbf{V}_j)$ which could be $O(j)$ in the worst case. This is due to that the optimization loss function $L_t(\boldsymbol{\theta})$ is the sum of t squared loss functions and we do not assume each loss function is strongly convex. Moreover, under certain assumption on the diversity of the arm set as is studied in Hamidi & Bayati (2020); Wu et al. (2020), the condition number will become $O(1)$. On the other hand, if we apply Newton's method to minimize the loss function $L_t(\boldsymbol{\theta})$ in each round, we could get rid of the condition number κ_j in the dependence of the epoch length of the inner loop K_j. Lastly, we observe from our empirical study that LMC-TS often requires a small number of iterations to achieve a good performance.

5. Empirical Evaluation of LMC-TS

In this section, we conduct experiments on both synthetic datasets and real-world datasets to show that the proposed algorithm achieves the best performance in terms of regret minimization and also is scalable to large-scale and high-dimensional problems. All experiments are conducted on Amazon EC2 P3 instances with NVIDIA V100 GPUs and Broadwell E5-2686 v4 processors. Our implementation can be found at `https://github.com/devzhk/LMCTS`.

Benchmarks and baseline algorithms As we discussed in Section 3, our LMC-TS algorithm is applicable to many different contextual bandit problems. Hence we compare it with baseline algorithms in different bandit settings including linear contextual bandits, logistic bandits, quadratic bandits, and neural bandits (also known as deep bandits) respectively. For linear bandit problems, we compare our algorithm with baseline linear bandit algorithms such as LinUCB (Chu et al., 2011), LinTS (Agrawal & Goyal, 2013), and the ϵ-greedy algorithm. For logistic bandit problems, we compare our algorithm with existing state-of-the-art algorithms for generalized linear bandits including UCB-GLM (Li et al., 2017), GLM-TSL (Kveton et al., 2020), SGD-TS (Ding et al., 2021), and ϵ-greedy. For quadratic bandits and neural bandits, we compare our algorithm additionally with NeuralUCB (Zhou et al., 2020), NeuralTS (Zhang et al., 2021), Neural-LinUCB (Xu et al., 2022), and Neural ϵ-greedy (Riquelme et al., 2018) which applies ϵ-greedy exploration to a neural network reward model trained by SGD.

5.1. Simulation Study on Linear, Logistic, and Quadratic Contextual Bandits

We first compare our algorithm with baseline methods on simulated bandit problems presented in Section 2 including linear bandits, logistic bandits, and quadratic bandits, where the true reward model $f(\mathbf{x}, \boldsymbol{\theta}^*)$ is known but the weight parameter $\boldsymbol{\theta}^* \in \mathbb{R}^d$ is unknown. Throughout this subsection, the context feature dimension is $d = 20$, the size of the arm set at round t is $|\mathcal{X}_t| = 50$, and the time horizon of all learning algorithms is $T = 10000$.

Linear contextual bandits: We first generate a linear contextual dataset following the problem setup in Section 2. To simulate the bandit environment, we first generate $\boldsymbol{\theta}^* \in \mathbb{R}^d$ with each coordinate randomly sampled from $\mathcal{N}(0,1)$ and then scale $\boldsymbol{\theta}^*$ to unit norm. We consider the following two settings: (1) we have a fixed arm set $\mathcal{X} \subseteq \mathbb{R}^d$ that remains the same during the whole learning process; (2) we receive a new arm set $\mathcal{X}_t \subseteq \mathbb{R}^d$ at each round t. For the feature vectors, we generate $\mathbf{x} \in \mathbb{R}^d$ with each coordinate randomly sampled from $\mathcal{N}(0,1)$ and then scale each vector to unit norm. The true reward for an arm $\mathbf{x}$ is then generated by $r(\mathbf{x}) = \boldsymbol{\theta}^{*\top}\mathbf{x} + \xi$, where the noise ξ is sampled from $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 0.5$.

Logistic bandits: In the logistic bandit experiment, we follow the setting in Kveton et al. (2020) and consider the fixed arm setting where $\mathcal{X} \subseteq \mathbb{R}^d$ with the context dimension $d = 20$ and the size of arm set $|\mathcal{X}| = 50$. Each contextual vector is randomly generated from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and scaled to unit norm. The reward for arm $\mathbf{x} \in \mathcal{X}$ is generated by a Bernoulli distribution, viz., $r(\mathbf{x}) \sim \text{Ber}\left(\mu(\boldsymbol{\theta}^{*\top}\mathbf{x})\right)$, where $\boldsymbol{\theta}^* \in \mathbb{R}^d$ is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and scaled to unit norm, and $\mu(v) = 1/(1 + \exp(-v))$ is the logistic function.

Quadratic bandits: Following the setting in Zhou et al. (2020), we generate a quadratic contextual bandit problem. We generate a changing arm set $\mathcal{X}_t \subseteq \mathbb{R}^d$ at each round. The context dimension $d = 20$. The size of arm set $|\mathcal{X}_t| = 50$ at each round. Each contextual vector $\mathbf{x} \in \mathcal{X}_t$ is randomly generated from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and scaled to unit norm. At round t, the reward function for the chosen arm $\mathbf{x}_t \in \mathcal{X}_t$ is given by $r_t = 10\left(\boldsymbol{\theta}^{*\top}\mathbf{x}_t\right)^2 + \xi$, where $\boldsymbol{\theta}^* \in \mathbb{R}^d$ is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and scaled to unit norm, and the noise $\xi \sim \mathcal{N}(0, 1)$.

Implementation details: For linear bandits, a linear reward model $f(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{x}^\top\boldsymbol{\theta}$ is used in all algorithms. For LinUCB, we set the UCB bonus parameter as $\nu_t = c\sqrt{d \log t}$ following Li et al. (2010) and find the best parameter c by performing a grid search. For LinTS, we set the variance parameter as $\nu = c\sqrt{d \log T}$ following the theory in Agrawal & Goyal (2013), and pick the best hyperparameter constant c from a grid search. For ϵ-greedy, the exploration rate is $\frac{c}{\sqrt{t}}$ at time t, where c is selected by a gird search. For LMC-TS, we set the step size $\eta_t = \frac{\eta_0}{t}$ as suggested in our theory and do a grid search for the constant η_0 and the temperature parameter β^{-1}. We fix the epoch length for the inner loop of our algorithm as $K_t = 100$ for all t.

For logistic bandits, a generalized linear reward model $f(\mathbf{x}, \boldsymbol{\theta}) = \mu(\mathbf{x}^\top\boldsymbol{\theta})$ is used, where $\mu(v) = 1/(1+\exp(-v))$. The MLE estimator in (3.4) is solved via SGD for UCB-GLM, GLM-TSL, and ϵ-greedy. For UCB-GLM, and GLM-TSL, we follow the same parameter setting proposed in Kveton et al. (2020). For LMC-TS, we use the loss function defined in (3.5), and the step size and temperature parameters are tuned in the same way as in linear bandits.

For quadratic bandits, a 4-layer fully-connected neural network with width 20 is used in NeuralUCB, NeuralTS, Neural-LinUCB, Neural ϵ-greedy, and LMC-TS. We try both ReLU and LeakyReLU and pick the best activation function for each algorithm. Neural networks are all updated by 100 gradient descent steps every round. Following the original implementation of NeuralTS and NeuralUCB, the matrix inverse is approximated by the inverse of diagonal. We perform grid search for the regularization parameter λ and variance parameter ν in their work.

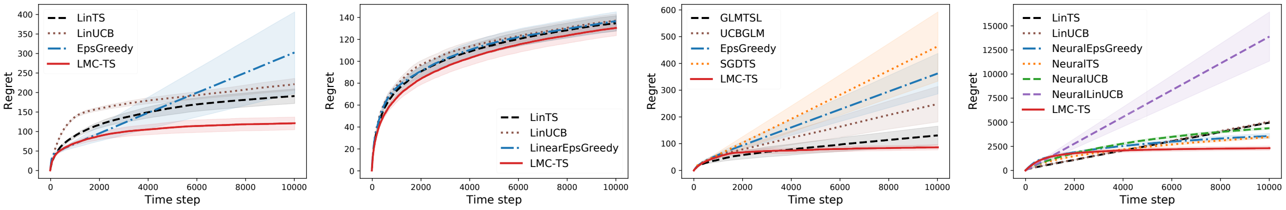
To make a fair comparison, we first perform grid searches for the parameters of all algorithms. We then fix the best hyperparameters, and shuffle the order of the dataset and repeat experiments for 10 times with different random seeds.

Results: We report the mean and the standard error of the accumulative regret of different algorithms over 10 runs on all simulated bandit problems in Figure 1. The results on linear contextual bandits are shown in Figure 1(a) with a fixed arm set and in Figure 1(b) with time-changing arm sets. Our method LMC-TS achieves the best performance in both settings, and the performance gain in the changing arm setting is slightly lower since it is a more challenging problem. The results for logistic bandits are shown in Figure 1(c), and LMC-TS again outperforms baseline methods. The results for quadratic bandits are shown in Figure 1(d). In this setting, LinUCB and LinTS works poorly due to their dependence on the linear bandit structure. All the other algorithms use a neural network to model the reward, and our method achieves significant lower regret. Neural-LinUCB performs much worse than other baseline methods, possibly due to its insufficient exploration in this setting.

5.2. Real-World Datasets

In this subsection, we consider UCI machine learning datasets (Dua & Graff, 2019) including *Shuttle*, *MagicTelescope*, *Mushroom*, and *Covertype*, and a high dimension image dataset *CIFAR10* (Krizhevsky et al., 2009). The specifications of these datasets are summarized in Table 1. To use these N-class classification datasets for contextual bandit problems, we follow Riquelme et al. (2018); Kveton et al. (2020) to construct context vectors for different arms in the following way: given a data feature $\mathbf{x} \in \mathbb{R}^d$, we transform it into N contextual vectors $\mathbf{x}^{(1)} = (\mathbf{x}, \mathbf{0}, \ldots, \mathbf{0}), \ldots, \mathbf{x}^{(N)} = (\mathbf{0}, \ldots, \mathbf{0}, \mathbf{x}) \in \mathbb{R}^{Nd}$. Only the arm $\mathbf{x}^{(j)}$ where j matches the correct class of this data has reward 1 and all other arms have reward 0.

Implementation details: For all methods using neural networks, we choose the best architecture between a two-layer neural network with width 100 and a four-layer neural net-

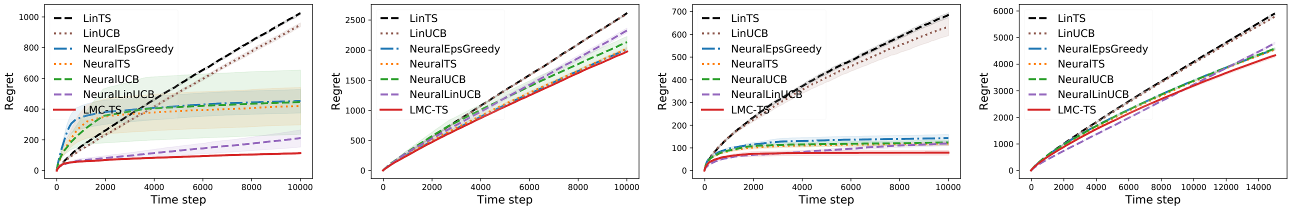


(a) Linear bandit (fixed arm set) (b) Linear bandit (c) Logistic bandit (d) Quadratic bandit

Figure 1. Regret comparison on simulated bandit problems. The mean and standard error are reported over 10 runs.

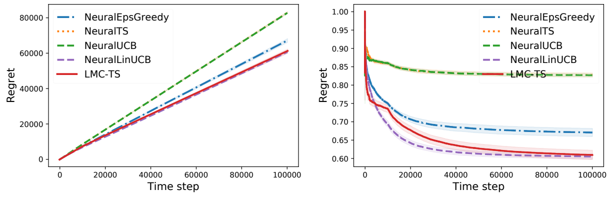
Table 1. Specifications of real-world datasets used in this paper.

	Shuttle	*MagicTelescope*	*Mushroom*	*Covertype*	*CIFAR10*
Number of attributes	9	10	22	54	$3 \times 32 \times 32$
Number of arms	7	2	2	7	10
Dimension of context feature	63	20	48	378	30720
Number of instances	58,000	19,020	8124	581,012	10,000



(a) Shuttle (b) MagicTelescope (c) Mushroom (d) Covertype

Figure 2. Regret comparison on UCI datasets. The mean and standard error are reported over 10 runs.



(a) Cumulative regret (b) Averaged regret

Figure 3. Regret comparison on CIFAR10. The mean and standard error are reported over 5 runs.

work with width 50. Activation function is also selected by grid search over ReLU and LeakyReLU. All baseline algorithms update the neural network by 100 gradient descent steps every round. The parameters of all methods are tuned in the same way as in Section 5.1 for quadratic bandits.

Results: We plot the average regret with std over 10 repeats for all algorithms on UCI datasets in Figure 2. It can be seen that our algorithm LMC-TS consistently outperforms other neural network based algorithms. The results for linear bandit based approaches are the worst on all datasets. The results on CIFAR10 is displayed in Figure 3. Note that the dimension of this dataset is much higher than those of UCI datasets that are used in existing neural contextual bandit papers (Riquelme et al., 2018; Zhou et al., 2020; Zhang et al., 2021; Xu et al., 2022). We plot the cumulative regret in Figure 3(a) and the average regret in Figure 3(b), which shows that LMC-TS achieves a sublinear cumulative regret and significantly outperforms all baseline methods except Neural-LinUCB which converges almost as fast as LMC-TS. This is due to the shallow exploration of Neural-LinUCB which only perform UCB exploration on the last layer parameter of the neural network. However, it also leads to a higher asymptotic regret than LMC-TS accoding to Figure 3(b), which explores over the whole parameter space. This shows the potential of our algorithm in high dimensional bandit problems that use deep neural networks.

We also compare the runtime of our algorithm with NeuralTS, which is a Laplace approximation based Thompson sampling algorithm, to demonstrate the computational efficiency of LMC-TS. The runtime for running 1000 rounds is shown in Table 2. We present both the total runtime of these rounds and the time only for arm selection. It can be seen that LMC-TS is more computationally efficient than NeuralTS, which is due to the fact that we do not need to calculate the inverse of a high dimensional matrix or sample from a nonisotropic Gaussian distribution.

Table 2. Runtime (seconds) on the CIFAR10 dataset for running 1000 rounds.

	TIME FOR ARM SELECTION	TOTAL TIME
NEURALTS	174.1	776.0
LMC-TS	9.0	662.7

6. Conclusions

In this paper, we proposed the Langevin Monte Carlo Thompson Sampling (LMC-TS) algorithm, which uses Langevin Monte Carlo to approximately sample the model parameter in contextual bandit problems from the posterior distribution. Unlike existing Thompson sampling based bandit algorithms that need to construct a Laplace approximation of the posterior which has a fixed approximation error, our method can directly sample from the posterior distribution with an arbitrarily small approximation error. Moreover, our algorithm only requires to perform noisy gradient descent updates which is more computationally efficient than sampling from a high dimensional non-isotropic Gaussian distribution. It is also worth noting that the proposed LMC-TS algorithm works for a large class of contextual bandit problems where the sampling distribution may not necessarily be a conjugate posterior and the reward models are general functions.

As a sanity check, we proved that LMC-TS enjoys the same regret upper bound as best Thompson sampling algorithms for linear contextual bandits, which theoretically demonstrates the competitiveness of LMC-TS in terms of regret minimization. Due to the efficiency and flexibility of our method, it would be an interesting and promising future direction to further investigate its theoretical performance in more general contextual bandits including generalized linear bandits and neural bandits.

We also conducted numerical experiments on both synthetic datasets and real-world datasets to empirically evaluate the performance of our method. We compared our method LMC-TS with various baseline algorithms in linear contextual bandits, generalized linear bandits, and neural contextual bandits. We observed that our method consistently outperforms existing baseline algorithms in these settings respectively. Moreover, our method is much more scalable to high dimensional problems where deep neural networks are used owing to its simplicity and efficiency.

Acknowledgements

The authors would like to thank the anonymous reviewers for their invaluable comments. PX is supported by PIMCO Postdoctoral Fellowship. AA is partially supported by Bren Named Chair Professorship at Caltech.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

Abramowitz, M. and Stegun, I. A. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1964.

Agrawal, S. and Goyal, N. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pp. 39–1. JMLR Workshop and Conference Proceedings, 2012.

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

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Bakry, D., Gentil, I., Ledoux, M., et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Machine Learning*, 5(1):1–122, 2012.

Chapelle, O. and Li, L. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24:2249–2257, 2011.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Dalalyan, A. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. In *Conference on Learning Theory*, pp. 678–689. PMLR, 2017a.

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

Dalalyan, A. S. and Karagulyan, A. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12): 5278–5311, 2019.

Ding, Q., Hsieh, C.-J., and Sharpnack, J. An efficient algorithm for generalized linear bandit: Online stochastic gradient descent and thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pp. 1585–1593. PMLR, 2021.

Dong, S., Ma, T., and Van Roy, B. On the performance of thompson sampling on logistic bandits. In *Conference on Learning Theory*, pp. 1158–1160. PMLR, 2019.

Dua, D. and Graff, C. UCI machine learning repository, 2019. URL http://archive.ics.uci.edu/ml.

Dubey, K. A., J Reddi, S., Williamson, S. A., Poczos, B., Smola, A. J., and Xing, E. P. Variance reduction in stochastic gradient langevin dynamics. *Advances in neural information processing systems*, 29, 2016.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. Parametric bandits: The generalized linear case. In *NIPS*, volume 23, pp. 586–594, 2010.

Hamidi, N. and Bayati, M. On worst-case regret of linear thompson sampling. *arXiv preprint arXiv:2006.06790*, 2020.

Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Jin, T., Xu, P., Shi, J., Xiao, X., and Gu, Q. Mots: Minimax optimal thompson sampling. In *International Conference on Machine Learning*, pp. 5074–5083. PMLR, 2021.

Kaufmann, E., Korda, N., and Munos, R. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pp. 199–213. Springer, 2012.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kveton, B., Zaheer, M., Szepesvari, C., Li, L., Ghavamzadeh, M., and Boutilier, C. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 2066–2076. PMLR, 2020.

Lale, S., Azizzadenesheli, K., Anandkumar, A., and Hassibi, B. Stochastic linear bandits with hidden low rank structure. *arXiv preprint arXiv:1901.09490*, 2019.

Langevin, P. On the theory of brownian motion. *CR Acad. Sci. Paris*, 146:530–533, 1908.

Langford, J. and Zhang, T. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20(1):96–1, 2007.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Li, L., Lu, Y., and Zhou, D. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pp. 2071–2080. PMLR, 2017.

Mazumdar, E., Pacchiano, A., Ma, Y., Jordan, M., and Bartlett, P. On approximate thompson sampling with Langevin algorithms. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 6797–6807. PMLR, 2020. URL https://proceedings.mlr.press/v119/mazumdar20a.html.

McCullagh, P. and Nelder, J. A. *Generalized linear models*. Routledge, 2019.

Ménard, P. and Garivier, A. A minimax and asymptotically optimal algorithm for stochastic bandits. In *International Conference on Algorithmic Learning Theory*, pp. 223–237. PMLR, 2017.

Phan, M., Abbasi Yadkori, Y., and Domke, J. Thompson sampling and approximate inference. *Advances in Neural Information Processing Systems*, 32:8804–8813, 2019.

Riquelme, C., Tucker, G., and Snoek, J. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SyYe6k-CW.

Roberts, G. O. and Tweedie, R. L. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pp. 341–363, 1996.

Russo, D. and Van Roy, B. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39 (4):1221–1243, 2014.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., and Wen, Z. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

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

Vempala, S. and Wibisono, A. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Wang, Z. and Zhou, M. Thompson sampling via local uncertainty. In *International Conference on Machine Learning*, pp. 10115–10125. PMLR, 2020.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688. Citeseer, 2011.

Wu, W., Yang, J., and Shen, C. Stochastic linear contextual bandits with diverse contexts. In *International Conference on Artificial Intelligence and Statistics*, pp. 2392–2401. PMLR, 2020.

Xu, P., Chen, J., Zou, D., and Gu, Q. Global convergence of langevin dynamics based algorithms for nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Xu, P., Wen, Z., Zhao, H., and Gu, Q. Neural contextual bandits with deep representation and shallow exploration. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xnYACQquaGV.

Zhang, W., Zhou, D., Li, L., and Gu, Q. Neural thompson sampling. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tkAtoZkcUnm.

Zhou, D., Li, L., and Gu, Q. Neural contextual bandits with ucb-based exploration. In *International Conference on Machine Learning*, pp. 11492–11502. PMLR, 2020.

Zou, D., Xu, P., and Gu, Q. Faster convergence of stochastic gradient langevin dynamics for non-log-concave sampling. In *Uncertainty in Artificial Intelligence*, pp. 1152–1162. PMLR, 2021.

A. Proof of the Regret Bound for Linear Contextual Bandits

In this section, we study the regret of Algorithm 1. Before we analyze the regret of the proposed algorithm, define $\mathbf{x}_t^* = \mathrm{argmax}_{\mathbf{x}\in\mathcal{X}_t} \mathbf{x}^\top\boldsymbol{\theta}^*$ to be the optimal arm at time step t. Define $\Delta_t(\mathbf{x}) = \mathbf{x}_t^*\boldsymbol{\theta}^* - \mathbf{x}^\top\boldsymbol{\theta}^*$ to be the gap between the expected reward of the best arm and an arbitrary arm $\mathbf{x}\in\mathcal{X}_t$ at time t.

We follow a similar proof in Agrawal & Goyal (2013) based on the notion of saturated and unsaturated arm sets, which are defined as follows respectively

$$\text{Saturated arms: } \mathcal{S}_t = \{\mathbf{x}\in\mathcal{X}_t : \Delta_t(\mathbf{x}) > g_t(\mathbf{x})\}, \tag{A.1}$$

$$\text{Unsaturated arms: } \mathcal{U}_t = \{\mathbf{x}\in\mathcal{X}_t : \Delta_t(\mathbf{x}) \le g_t(\mathbf{x})\}, \tag{A.2}$$

where $g_t(\mathbf{x})$ is a function of the time index t, arm $\mathbf{x}$, and historical data up to time t, which will be determined in later proofs. Intuitively, saturated arms are already well estimated and the suboptimal gap of these arms are large enough such that they can be easily distinguished from the best arm. Define the filtration $\mathcal{F}_t = \{\mathbf{x}_1, r_1, \ldots, \mathbf{x}_t\}$. The following proposition implies that the distribution of the iterate $\boldsymbol{\theta}_{t,K_t}$ of Algorithm 1 is a Gaussian distribution.

Proposition A.1. *Under filtration $\mathcal{F}_{t-1}$, the parameter $\boldsymbol{\theta}_{t,K_t}$ used in the decision at round t of Algorithm 1 follows a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_{t,K_t}, \boldsymbol{\Sigma}_{t,K_t})$, where the mean vector and the covariance matrix are defined as*

$$\boldsymbol{\mu}_{t,K_t} = \mathbf{A}_t^{K_t}\ldots\mathbf{A}_1^{k_1}\boldsymbol{\theta}_{1,0} + \sum_{i=1}^{t}\mathbf{A}_t^{K_t}\ldots\mathbf{A}_{i+1}^{k_{i+1}}\big(\mathbf{I}-\mathbf{A}_i^{K_t}\big)\widehat{\boldsymbol{\theta}}_i, \tag{A.3}$$

$$\boldsymbol{\Sigma}_{t,K_t} = \sum_{i=1}^{t}\frac{1}{\beta_i}\mathbf{A}_t^{K_t}\ldots\mathbf{A}_{i+1}^{k_{i+1}}\big(\mathbf{I}-\mathbf{A}_i^{2K_i}\big)\mathbf{V}_i(\mathbf{I}+\mathbf{A}_i)^{-1}\mathbf{A}_{i+1}^{k_{i+1}}\ldots\mathbf{A}_t^{K_t}, \tag{A.4}$$

where $\mathbf{A}_i = \mathbf{I} - 2\eta_i\mathbf{V}_i$ for $i = 1, 2, \ldots$.

It can be seen that the variance term depends more on the recent noise than those noises added in the history.

The following lemma upper bounds the distance between the solution $\widehat{\boldsymbol{\theta}}_t$ of the ridge regression in (3.3) and the true weight parameter $\boldsymbol{\theta}^*$ of the bandit model.

Lemma A.2 (Agrawal & Goyal (2013))**.** *Let $\delta\in(0,1)$. For any $\mathbf{x}\in\mathbb{R}^d$, it holds that*

$$\mathbb{P}\big(\big|\mathbf{x}^\top(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)\big| \le \big(R\sqrt{d\log(t^3/\delta)}+1\big)\|\mathbf{x}\|_{\mathbf{V}_t^{-1}}\big) \ge 1 - \frac{\delta}{t^2}. \tag{A.5}$$

For the simplicity of notations, we denote $g_R(t) = R\sqrt{d\log(t^3/\delta)}+1$ in the rest of the paper.

This can be bounded using the same bound of the least square solution in Agrawal & Goyal (2013).

Lemma A.3. *Let $R > 0$. For a R-subGaussian random variable ξ, it holds that $\mathbb{P}(|\xi| > x) \le \exp(1 - x^2/R^2)$. In other words, $\mathbb{P}(|\xi| > R\sqrt{1+2\log t}) \le 1/t^2$ for any $t > 1$. Define the following event*

$$E_{R,t} = \big\{|\mathbf{x}^\top\widehat{\boldsymbol{\theta}}_t - \mathbf{x}^\top\boldsymbol{\theta}^*| \le g_R(t)\|\mathbf{x}\|_{\mathbf{V}_t^{-1}}, \forall\mathbf{x}\in\mathcal{X}_t\big\} \cap \{|\xi_t| \le R\sqrt{1+2\log t}\}. \tag{A.6}$$

Then it holds that $\mathbb{P}(E_{R,t}) \ge 1 - 2/t^2$.

We define the event $E_{S,t}$ as follows, which provides an upper bound of the distance between the iterate $\boldsymbol{\theta}_{t,k}$ and the mode $\widehat{\boldsymbol{\theta}}_t$ of density $\pi_t \propto \exp(-\beta_t L_t(\boldsymbol{\theta}))$.

$$E_{S,t} = \left\{|\mathbf{x}^\top\boldsymbol{\theta}_{t,k} - \mathbf{x}^\top\widehat{\boldsymbol{\theta}}_t| \le \left(5\sqrt{\frac{2d\log t}{3\beta_T}} + 3/2\right)\|\mathbf{x}\|_{\mathbf{V}_t^{-1}}\right\}. \tag{A.7}$$

The following lemma shows that $E_{S,t}$ happens with high probability.

Lemma A.4. *Under event $E_{R,t}$, for any $\mathbf{x}\in\mathcal{X}_t$, with probability at least $1-1/t^2$ we have $\mathbb{P}(E_{S,t}) \ge 1 - 1/T^2$.*

Lemma A.5 (Optimistic estimation)**.** *Based on the parameter estimation at time step t, the best arm is over estimated. In particular, let $K_j \geq \kappa_j \log(3R\sqrt{2dT\log(T^3/\delta)})$ and $1/\beta_j = 4(R\sqrt{d\log(T^3/\delta)}+2)$ for all $j \in [T]$. Then conditional on event $E_{R,t}$, we have*

$$P\big(\mathbf{x}_t^{*\top}\boldsymbol{\theta}_{t,k} > \mathbf{x}_t^{*\top}\boldsymbol{\theta}^*\big) \geq \frac{1}{2\sqrt{2e\pi}}. \tag{A.8}$$

Lemma A.6. *The probability of playing unsaturated arms is at least a constant. That is,*

$$\mathbb{P}(\mathbf{x}_t \in \mathcal{U}_t | E_{R,t}) \geq \frac{1}{2\sqrt{2e\pi}} - \frac{1}{t^2}.$$

The following lemma is a combination of Lemma 10 and Lemma 11 in Abbasi-Yadkori et al. (2011), which is standard in the analysis of linear contextual bandits.

Lemma A.7. *Let $\{\mathbf{x}_t\}_{t=1}^\infty$ be a sequence in $\mathbb{R}^d$ and $\lambda > 0$. Suppose $\|\mathbf{x}_t\|_2 \leq 1$ and $\lambda \geq 1$. Define $\mathbf{V}_t = \lambda\mathbf{I} + \sum_{s=1}^t \mathbf{x}_t\mathbf{x}_t^\top$. Then we have*

$$\det(\mathbf{V}_t) \leq (\lambda + t/d)^d, \quad \textit{and} \sum_{t=1}^T \|\mathbf{x}_t\|^2_{\mathbf{V}_{t-1}^{-1}} \leq 2\log\frac{\det(\mathbf{V}_T)}{\det(\lambda\mathbf{I})} \leq 2d\log(1+T/(\lambda d)).$$

Recall the definition of in (A.1) and (A.2), where $g_t(\mathbf{x})$ was not immediately specified. Now based on the knowledge of Lemmas A.2 and A.4, let us choose $g_t(\mathbf{x})$ as follows:

$$g_t(\mathbf{x}) = \big(R\sqrt{d\log(t^3/\delta)} + 10Rd\sqrt{\log T \log(T^3/\delta)} + 5/2\big)\|\mathbf{x}_t\|_{\mathbf{V}_t^{-1}}. \tag{A.9}$$

Proof of Theorem 4.2. By definition, the regret of Algorithm 1 is $R(T) = \sum_{t=1}^T \Delta_t(\mathbf{x}_t)$. At each time step, define $\bar{\mathbf{x}}_t = \operatorname{argmin}_{\mathbf{x}\in\mathcal{U}_t} g_t(\mathbf{x})$ to be the arm in the unsaturated arm set that has the smallest value $\|\mathbf{x}\|_{\mathbf{V}_t^{-1}}$. Then conditional on events $E_{R,t}$ and $E_{S,t}$, we have

$$\begin{aligned}
\Delta_t(\mathbf{x}_t) &= \mathbf{x}_t^{*\top}\boldsymbol{\theta}^* - \mathbf{x}_t^\top\boldsymbol{\theta}^* \\
&= \mathbf{x}_t^{*\top}\boldsymbol{\theta}^* - \bar{\mathbf{x}}_t^\top\boldsymbol{\theta}^* + \bar{\mathbf{x}}_t^\top\boldsymbol{\theta}^* - \mathbf{x}_t^\top\boldsymbol{\theta}^* \\
&\leq g_t(\bar{\mathbf{x}}_t) + \big(\bar{\mathbf{x}}_t^\top\boldsymbol{\theta}_{t,K_t} + g_t(\bar{\mathbf{x}}_t)\big) - \big(\mathbf{x}_t^\top\boldsymbol{\theta}_{t,K_t} - g_t(\mathbf{x}_t)\big) \\
&\leq 2g_t(\bar{\mathbf{x}}_t) + g_t(\mathbf{x}_t),
\end{aligned} \tag{A.10}$$

where the first inequality is due to $\bar{\mathbf{x}}_t \in \mathcal{U}_t$, Lemma A.2 and Lemma A.4 respectively, and the last inequality is due to the choice of $\mathbf{x}_t$ in our algorithm. We denote $p = 1/(2\sqrt{2e\pi})$. Note that $g_t(\mathbf{x}) > 0$. We have

$$\begin{aligned}
\mathbb{E}[g_t(\mathbf{x}_t)|\mathcal{F}_t, E_{R,t}] &= \mathbb{E}[g_t(\mathbf{x}_t)|\mathcal{F}_t, E_{R,t}, \mathbf{x}_t \in \mathcal{U}_t]\mathbb{P}(\mathbf{x}_t \in \mathcal{U}_t) \\
&\quad + \mathbb{E}[g_t(\mathbf{x}_t)|\mathcal{F}_t, E_{R,t}, \mathbf{x}_t \in \mathcal{S}_t]\mathbb{P}(\mathbf{x}_t \in \mathcal{S}_t) \\
&\geq (p - 1/t^2)g_t(\bar{\mathbf{x}}_t),
\end{aligned} \tag{A.11}$$

where the inequality holds due to the definition of $\bar{\mathbf{x}}_t$ and Lemma A.6. Therefore, we have

$$\begin{aligned}
\mathbb{E}[\Delta_t(\mathbf{x}_t)|\mathcal{F}_t] &= \mathbb{E}[\Delta_t(\mathbf{x}_t)|\mathcal{F}_t, E_{R,t}]\mathbb{P}(E_{R,t}) + \mathbb{E}[\Delta_t(\mathbf{x}_t)|\mathcal{F}_t, E_R^c(t)]\mathbb{P}(E_R^c(t)) \\
&\leq \mathbb{E}[2g_t(\bar{\mathbf{x}}_t) + g_t(\mathbf{x}_t)|\mathcal{F}_t, E_{R,t}]\mathbb{P}(E_{R,t}) + \mathbb{P}(E_R^c(t)) \\
&\leq \bigg(\frac{2}{p - 1/t^2} + 1\bigg)\mathbb{E}[g_t(\mathbf{x}_t)|\mathcal{F}_t, E_{R,t}] + \frac{\delta}{t^2} \\
&\leq \frac{c}{p}\mathbb{E}[g_t(\mathbf{x}_t)|\mathcal{F}_t, E_{R,t}] + \frac{\delta}{t^2},
\end{aligned} \tag{A.12}$$

where we assume $\|\Delta_t(\mathbf{x})\| \leq 1$ for any $\mathbf{x} \in \mathcal{X}$ since both $\mathbf{x}$ and $\boldsymbol{\theta}^*$ are bounded, and $c > 0$ is a constant. Define $Y_t = \sum_{s=1}^t (\Delta_s(\mathbf{x}_s) - c/pg_s(\mathbf{x}_s) - \delta/s^2)$ and $Y_0 = 0$. Then we obtain $\mathbb{E}[Y_t - Y_{t-1}|\mathcal{F}_t] \leq 0$, which implies that $\{Y_t\}_{t=0,1,\ldots}$ is a super martingale, corresponding a filtration $\mathcal{F}_t$. Note that

$$|Y_t - Y_{t-1}| = |\Delta_t(\mathbf{x}_t) - c/pg_t(\mathbf{x}_t) - \delta/t^2| \leq 3c/pg_t(\mathbf{x}_t). \tag{A.13}$$

Let $\epsilon^2 = 2\log(1/\delta)\sum_{t=1}^{T} 9(c/p)^2 g_t(\mathbf{x}_t)^2$. By Azuma-Hoeffding inequality in Lemma C.3, we know with probability at least $1-\delta$ it holds that

$$Y_T = \sum_{t=1}^{T}(\Delta_t(\mathbf{x}_t) - c/pg_t(\mathbf{x}_t) - \delta/t^2) \le \sqrt{2\log(1/\delta)\sum_{t=1}^{T} 9(c/p)^2 g_t(\mathbf{x}_t)^2},$$

which immediately implies with probability at least $1-\delta$ that

$$R(T) \le (1 + 3\sqrt{2\log(1/\delta)})c/p\sum_{t=1}^{T} g_t(\mathbf{x}_t) + \sum_{t=1}^{T}\delta/t^2.$$

Recall the definition of $g_t(\mathbf{x}_t)$ in (A.9), we further have with probability at least $1-\delta$ it holds

$$\begin{aligned}
R(T) &\le (1 + 3\sqrt{2\log(1/\delta)})c/p\sum_{t=1}^{T}\big(R\sqrt{d\log(t^3/\delta)} + 10Rd\sqrt{\log T\log(T^3/\delta)} + 5/2\big)\|\mathbf{x}_t\|_{\mathbf{V}_t^{-1}} + \frac{\pi^2\delta}{6} \\
&\le C_0 Rd\sqrt{dT\log T\log(T^3/\delta)\log(1/\delta)\log(1+T/(\lambda d))} \\
&\le C_0 Rd\log(1/\delta)\sqrt{dT\log^3(1+T/(\lambda d))},
\end{aligned}$$

where in the first inequality we used the fact that $\sum_{t=1}^{\infty} 1/t^2 = \pi^2/6$, the second inequality is due to Lemmas A.5 and A.7 and Cauchy inequality, and C_0 is a constant independent of the problem. □

B. Proof of Technical Lemmas

B.1. Proof of Proposition A.1

Proof of Proposition A.1. By the updating rule of $\boldsymbol{\theta}_{t,K_t}$ in Algorithm 1, we have

$$\begin{aligned}
\boldsymbol{\theta}_{t,K_t} &= \boldsymbol{\theta}_{t,K_t-1} - 2\eta_t(\mathbf{V}_t\boldsymbol{\theta}_{t,K_t-1} - \mathbf{b}_t) + \sqrt{2\eta_t\beta_t^{-1}}\epsilon_{t,K_t} \\
&= (\mathbf{I} - 2\eta_t\mathbf{V}_t)\boldsymbol{\theta}_{t,K_t-1} + 2\eta_t\mathbf{b}_t + \sqrt{2\eta_t\beta_t^{-1}}\epsilon_{t,K_t} \\
&= (\mathbf{I} - 2\eta_t\mathbf{V}_t)^{K_t}\boldsymbol{\theta}_{t,0} + \sum_{l=0}^{K_t-1}(\mathbf{I} - 2\eta_t\mathbf{V}_t)^l\big(2\eta_t\mathbf{b}_t + \sqrt{2\eta_t\beta_t^{-1}}\epsilon_{t,k-l}\big) \\
&= (\mathbf{I} - 2\eta_t\mathbf{V}_t)^{K_t}\boldsymbol{\theta}_{t,0} + 2\eta_t\sum_{l=0}^{K_t-1}(\mathbf{I} - 2\eta_t\mathbf{V}_t)^l\mathbf{b}_t + \sqrt{2\eta_t\beta_t^{-1}}\sum_{l=0}^{K_t-1}(\mathbf{I} - 2\eta_t\mathbf{V}_t)^l\epsilon_{t,K_t-l}.
\end{aligned} \tag{B.1}$$

Recall that in the inner loop of Algorithm 1, we use a warm start from previous iteration, namely $\boldsymbol{\theta}_{t,0} = \boldsymbol{\theta}_{t-1,k_{t-1}}$. To simplify the notation, we denote $\mathbf{A}_i = \mathbf{I} - 2\eta_i\mathbf{V}_i$ for $i = 1, 2, \ldots$. Note that $\mathbf{A}_i$ is symmetric and satisfies $\mathbf{I} \succ \mathbf{I} - 2\eta_i\mathbf{V}_i \succ \mathbf{0}$ if the step size is chosen such that $0 < \eta_i < 1/(2\lambda_{\max}(\mathbf{V}_i))$. Therefore, we further have

$$\begin{aligned}
\boldsymbol{\theta}_{t,K_t} &= \mathbf{A}_t^{K_t}\boldsymbol{\theta}_{t-1,k_{t-1}} + \big(\mathbf{I} - \mathbf{A}_t^{K_t}\big)\widehat{\boldsymbol{\theta}}_t + \sqrt{2\eta_t\beta_t^{-1}}\sum_{l=0}^{K_t-1}\mathbf{A}_t^l\epsilon_{t,K_t-l} \\
&= \mathbf{A}_t^{K_t}\ldots\mathbf{A}_1^{k_1}\boldsymbol{\theta}_{1,0} + \sum_{i=1}^{t}\mathbf{A}_t^{K_t}\ldots\mathbf{A}_{i+1}^{k_{i+1}}\big(\mathbf{I} - \mathbf{A}_i^{K_t}\big)\widehat{\boldsymbol{\theta}}_i + \sum_{i=1}^{t}\sqrt{\frac{2\eta_i}{\beta_i}}\mathbf{A}_t^{K_t}\ldots\mathbf{A}_{i+1}^{k_{i+1}}\Bigg(\sum_{l=0}^{K_t-1}\mathbf{A}_i^l\boldsymbol{\epsilon}_{i,K_i-l}\Bigg),
\end{aligned}$$

where in the first equality we used $\mathbf{I} + \mathbf{A} + \ldots + \mathbf{A}^{n-1} = (\mathbf{I} - \mathbf{A}^n)(\mathbf{I} - \mathbf{A})^{-1}$ and $\mathbf{V}_t^{-1}\mathbf{b}_t = \widehat{\boldsymbol{\theta}}_t$. Conditional on $\mathcal{F}_{t-1}$ and the initialization $\boldsymbol{\theta}_{1,0}$, we know that $\boldsymbol{\theta}_{t,K_t}$ follows the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_{t,K_t}, \boldsymbol{\Sigma}_{t,K_t})$. Based on the property of multivariate Gaussian distribution, if $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d\times d})$, then we have $\mathbf{A}\boldsymbol{\epsilon} + \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{A}\mathbf{A}^\top)$ for any $\mathbf{A} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{\mu} \in \mathbb{R}^d$. Then the mean vector is defined as

$$\boldsymbol{\mu}_{t,K_t} = \mathbf{A}_t^{K_t}\ldots\mathbf{A}_1^{k_1}\boldsymbol{\theta}_{1,0} + \sum_{i=1}^{t}\mathbf{A}_t^{K_t}\ldots\mathbf{A}_{i+1}^{k_{i+1}}\big(\mathbf{I} - \mathbf{A}_i^{K_t}\big)\widehat{\boldsymbol{\theta}}_i. \tag{B.2}$$

Similarly, the covariance matrix is defined as

$$\begin{aligned}\boldsymbol{\Sigma}_{t,K_t} &= \sum_{i=1}^{t} \frac{2\eta_i}{\beta_i} \mathbf{A}_t^{K_t} \dots \mathbf{A}_{i+1}^{k_{i+1}} \sum_{l=0}^{K_i-1} \mathbf{A}_i^{2l} \mathbf{A}_{i+1}^{k_{i+1}} \dots \mathbf{A}_t^{K_t} \\ &= \sum_{i=1}^{t} \frac{2\eta_i}{\beta_i} \mathbf{A}_t^{K_t} \dots \mathbf{A}_{i+1}^{k_{i+1}} \big(\mathbf{I} - \mathbf{A}_i^{2K_i}\big)(\mathbf{I} - \mathbf{A}_i^2)^{-1} \mathbf{A}_{i+1}^{k_{i+1}} \dots \mathbf{A}_t^{K_t} \qquad \text{(B.3)} \\ &= \sum_{i=1}^{t} \frac{1}{\beta_i} \mathbf{A}_t^{K_t} \dots \mathbf{A}_{i+1}^{k_{i+1}} \big(\mathbf{I} - \mathbf{A}_i^{2K_i}\big) \mathbf{V}_i (\mathbf{I} + \mathbf{A}_i)^{-1} \mathbf{A}_{i+1}^{k_{i+1}} \dots \mathbf{A}_t^{K_t}, \qquad \text{(B.4)}\end{aligned}$$

where we used the fact that $\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \ldots + \mathbf{A}^{n-1} = (\mathbf{I} - \mathbf{A}^n)(\mathbf{I} - \mathbf{A})^{-1} = (\mathbf{I} - \mathbf{A})^{-1}(\mathbf{I} - \mathbf{A}^n)$ if $\mathbf{A}$ is symmetric. □

B.2. Proof of Lemma A.3

Proof of Lemma A.3. The first statement can be proved by standard concentration techniques (Vershynin, 2010). Note that

$$\begin{aligned}\mathbb{P}(E_{R,t}^c) &= \mathbb{P}\big(\{|\mathbf{x}^\top \widehat{\boldsymbol{\theta}}_t - \mathbf{x}^\top \boldsymbol{\theta}^*| \le g_R(t) \|\mathbf{x}\|_{\mathbf{V}_t^{-1}}, \forall \mathbf{x} \in \mathcal{X}_t\} \cup \{|\xi_t| \le R\sqrt{1 + 2\log t}\}\big) \\ &\le \mathbb{P}\big(\{|\mathbf{x}^\top \widehat{\boldsymbol{\theta}}_t - \mathbf{x}^\top \boldsymbol{\theta}^*| \le g_R(t) \|\mathbf{x}\|_{\mathbf{V}_t^{-1}}, \forall \mathbf{x} \in \mathcal{X}_t\}\big) + \mathbb{P}\big(\{|\xi_t| \le R\sqrt{1 + 2\log t}\}\big).\end{aligned}$$

Substituting the first statement and the result in Lemma A.2 into the above inequality, we know that $\mathbb{P}(E_{R,t}) \ge 1 - 2/t^2$. □

B.3. Proof of Lemma A.4

Proof of Lemma A.4. For any $\mathbf{x} \in \mathbb{R}^d$, we can decompose $\mathbf{x}^\top(\boldsymbol{\theta}_{t,k} - \widehat{\boldsymbol{\theta}}_t)$ as follows.

$$\mathbf{x}^\top(\boldsymbol{\theta}_{t,k} - \widehat{\boldsymbol{\theta}}_t) = \mathbf{x}^\top(\boldsymbol{\theta}_{t,k} - \boldsymbol{\mu}_{t,k}) + \mathbf{x}^\top(\boldsymbol{\mu}_{t,k} - \widehat{\boldsymbol{\theta}}_t). \qquad \text{(B.5)}$$

Bounding term $\mathbf{x}^\top(\boldsymbol{\theta}_{t,k} - \boldsymbol{\mu}_{t,k})$: we have

$$\big|\mathbf{x}^\top \boldsymbol{\theta}_{t,k} - \mathbf{x}^\top \boldsymbol{\mu}_{t,k}\big| \le \big\|\mathbf{x}^\top \boldsymbol{\Sigma}_{t,K_t}^{1/2}\big\|_2 \big\|\boldsymbol{\Sigma}_{t,K_t}^{-1/2}(\boldsymbol{\theta}_{t,k} - \boldsymbol{\mu}_{t,k})\big\|_2.$$

According to Proposition A.1, $\boldsymbol{\Sigma}_{t,K_t}^{-1/2}(\boldsymbol{\theta}_{t,k} - \boldsymbol{\mu}_{t,k})$ is a standard Gaussian random vector in $\mathbb{R}^d$. Therefore, we have

$$\mathbb{P}\Big(\big\|\boldsymbol{\Sigma}_{t,K_t}^{-1/2}(\boldsymbol{\theta}_{t,k} - \boldsymbol{\mu}_{t,k})\big\|_2 \ge \sqrt{4d\log t}\Big) \ge \frac{1}{t^2}. \qquad \text{(B.6)}$$

Note that when we choose $\eta_i \le 1/(4\lambda_{\max}(\mathbf{V}_i))$ for all i, we have

$$\begin{aligned}&\frac{1}{2}\mathbf{I} \prec \mathbf{A}_i = \mathbf{I} - 2\eta_i \mathbf{V}_i \prec (1 - 2\eta_i \lambda_{\min}(\mathbf{V}_i))\mathbf{I}, \\ &\frac{3}{2}\mathbf{I} \prec \mathbf{I} + \mathbf{A}_i = 2\mathbf{I} - 2\eta_i \mathbf{V}_i \prec 2\mathbf{I}.\end{aligned} \qquad \text{(B.7)}$$

By definition $\mathbf{A}_i = \mathbf{I} - 2\eta_i \mathbf{V}_i$ and $\mathbf{V}_i$ is symmetric. Therefore, $\mathbf{A}_i$ and $\mathbf{V}_i^{-1}$ commute, which implies

$$\begin{aligned}\mathbf{A}_i^{2K_i} \mathbf{V}_i^{-1} &= (\mathbf{I} - 2\eta_i \mathbf{V}_i) \dots (\mathbf{I} - 2\eta_i \mathbf{V}_i)(\mathbf{I} - 2\eta_i \mathbf{V}_i) \mathbf{V}_i^{-1} \\ &= (\mathbf{I} - 2\eta_i \mathbf{V}_i) \dots (\mathbf{I} - 2\eta_i \mathbf{V}_i) \mathbf{V}_i^{-1} (\mathbf{I} - 2\eta_i \mathbf{V}_i) \\ &= \mathbf{A}_i^{K_i} \mathbf{V}_i^{-1} \mathbf{A}_i^{K_i}. \qquad \text{(B.8)}\end{aligned}$$

Recall the definition of $\boldsymbol{\Sigma}_{t,K_t}$ in Proposition A.1, we have

$$\begin{aligned}
\mathbf{x}^\top \boldsymbol{\Sigma}_{t,K_t}\mathbf{x} &= \sum_{i=1}^{t} \frac{1}{\beta_i}\mathbf{x}^\top \mathbf{A}_t^{K_t}\dots\mathbf{A}_{i+1}^{k_{i+1}}\big(\mathbf{I}-\mathbf{A}_i^{2K_i}\big)\mathbf{V}_i^{-1}(\mathbf{I}+\mathbf{A}_i)^{-1}\mathbf{A}_{i+1}^{k_{i+1}}\dots\mathbf{A}_t^{K_t}\mathbf{x}\\
&\le \sum_{i=1}^{t} \frac{2}{3\beta_i}\mathbf{x}^\top \mathbf{A}_t^{K_t}\dots\mathbf{A}_{i+1}^{k_{i+1}}\big(\mathbf{V}_i^{-1}-\mathbf{A}_i^{K_i}\mathbf{V}_i^{-1}\mathbf{A}_i^{K_i}\big)\mathbf{A}_{i+1}^{k_{i+1}}\dots\mathbf{A}_t^{K_t}\mathbf{x}\\
&= \frac{2}{3\beta_T}\sum_{i=1}^{t-1}\mathbf{x}^\top \mathbf{A}_t^{K_t}\dots\mathbf{A}_{i+1}^{k_{i+1}}\big(\mathbf{V}_i^{-1}-\mathbf{V}_{i+1}^{-1}\big)\mathbf{A}_{i+1}^{k_{i+1}}\dots\mathbf{A}_t^{K_t}\mathbf{x}\\
&\quad - \frac{2}{3\beta_T}\mathbf{x}^\top \mathbf{A}_t^{K_t}\dots\mathbf{A}_1^{k_1}\mathbf{V}_1^{-1}\mathbf{A}_1^{k_1}\dots\mathbf{A}_t^{K_t}\mathbf{x} + \frac{2}{3\beta_T}\mathbf{x}^\top\mathbf{V}_t^{-1}\mathbf{x}, \qquad \text{(B.9)}
\end{aligned}$$

where the fist inequality is due to (B.7), and the last equality is due to the choice of $1/\beta_i = 1/\beta_T$ for all i. By the definition in (2.2) and Sherman-Morrison formula, we have

$$\mathbf{V}_i^{-1}-\mathbf{V}_{i+1}^{-1} = \mathbf{V}_i^{-1} - \big(\mathbf{V}_i + \mathbf{x}_i\mathbf{x}_i^\top\big)^{-1} = \frac{\mathbf{V}_i^{-1}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{V}_i^{-1}}{1+\|\mathbf{x}_i\|_{\mathbf{V}_i^{-1}}^2}, \qquad \text{(B.10)}$$

which immediately implies

$$\begin{aligned}
\mathbf{x}^\top \mathbf{A}_t^{K_t}\dots\mathbf{A}_{i+1}^{k_{i+1}}\big(\mathbf{V}_i^{-1}-\mathbf{V}_{i+1}^{-1}\big)\mathbf{A}_{i+1}^{k_{i+1}}\dots\mathbf{A}_t^{K_t}\mathbf{x} &= \mathbf{x}^\top \mathbf{A}_t^{K_t}\dots\mathbf{A}_{i+1}^{k_{i+1}}\frac{\mathbf{V}_i^{-1}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{V}_i^{-1}}{1+\|\mathbf{x}_i\|_{\mathbf{V}_i^{-1}}^2}\mathbf{A}_{i+1}^{k_{i+1}}\dots\mathbf{A}_t^{K_t}\mathbf{x}\\
&\le \big(\mathbf{x}^\top\mathbf{A}_t^{K_t}\dots\mathbf{A}_{i+1}^{k_{i+1}}\mathbf{V}_i^{-1}\mathbf{x}_i\big)^2\\
&\le \|\mathbf{A}_t^{K_t}\dots\mathbf{A}_{i+1}^{k_{i+1}}\mathbf{V}_i^{-1/2}\mathbf{x}\|_2^2\cdot\|\mathbf{V}_i^{-1/2}\mathbf{x}_i\|_2^2\\
&\le \prod_{j=i+1}^{t}(1-2\eta_j\lambda_{\min}(\mathbf{V}_j))^{2K_j}\|\mathbf{x}_i\|_{\mathbf{V}_i^{-1}}^2\|\mathbf{x}\|_{\mathbf{V}_i^{-1}}^2, \qquad \text{(B.11)}
\end{aligned}$$

where we used $0 < 1/i \le \|\mathbf{x}\|_{\mathbf{V}_i^{-1}} \le 1$ and the last inequality is due to (B.7). Therefore, we have

$$\mathbf{x}^\top\boldsymbol{\Sigma}_{t,k}\mathbf{x} \le \frac{2}{3\beta_T}\|\mathbf{x}\|_{\mathbf{V}_t^{-1}}^2 + \frac{2}{3\beta_T}\sum_{i=1}^{t-1}\prod_{j=i+1}^{t}(1-2\eta_j\lambda_{\min}(\mathbf{V}_j))^{2K_j}\|\mathbf{x}_i\|_{\mathbf{V}_i^{-1}}^2\|\mathbf{x}\|_{\mathbf{V}_i^{-1}}^2, \qquad \text{(B.12)}$$

which immediately implies

$$\|\mathbf{x}\|_{\boldsymbol{\Sigma}_{t,K_t}} \le \sqrt{\frac{2}{3\beta_T}}\Bigg(\|\mathbf{x}\|_{\mathbf{V}_t^{-1}} + \sum_{i=1}^{t-1}\prod_{j=i+1}^{t}(1-2\eta_j\lambda_{\min}(\mathbf{V}_j))^{K_j}\|\mathbf{x}_i\|_{\mathbf{V}_i^{-1}}\|\mathbf{x}\|_{\mathbf{V}_i^{-1}}\Bigg) := \hat{g}_t(\mathbf{x}).$$

Therefore, it holds that

$$\begin{aligned}
&\mathbb{P}\big(\big|\mathbf{x}^\top\boldsymbol{\theta}_{t,k} - \mathbf{x}^\top\boldsymbol{\mu}_{t,K_t}\big| \ge 2\hat{g}_t(\mathbf{x})\sqrt{d\log t}\big)\\
&\le \mathbb{P}\big(\big|\mathbf{x}^\top\boldsymbol{\theta}_{t,k} - \mathbf{x}^\top\boldsymbol{\mu}_{t,K_t}\big| \ge 2\sqrt{d\log t}\|\mathbf{x}\|_{\boldsymbol{\Sigma}_{t,K_t}}\big)\\
&\le \mathbb{P}\big(\big\|\mathbf{x}^\top\boldsymbol{\Sigma}_{t,K_t}^{1/2}\big\|_2\big\|\boldsymbol{\Sigma}_{t,K_t}^{-1/2}(\boldsymbol{\theta}_{t,k}-\boldsymbol{\mu}_{t,k})\big\|_2 \ge 2\sqrt{d\log t}\|\mathbf{x}\|_{\boldsymbol{\Sigma}_{t,K_t}}\big)\\
&\le \frac{1}{t^2}, \qquad \text{(B.13)}
\end{aligned}$$

where the last inequality follows from (B.6).

Bounding term $\mathbf{x}^\top(\boldsymbol{\mu}_{t,k} - \widehat{\boldsymbol{\theta}}_t)$: recall $\boldsymbol{\mu}_{t,k}$ defined in (B.2), we have

$$\begin{aligned}
\boldsymbol{\mu}_{t,K_t} &= \mathbf{A}_t^{K_t}\dots\mathbf{A}_1^{k_1}\boldsymbol{\theta}_{1,0} + \sum_{i=1}^{t}\mathbf{A}_t^{K_t}\dots\mathbf{A}_{i+1}^{k_{i+1}}\big(\mathbf{I}-\mathbf{A}_i^{K_t}\big)\widehat{\boldsymbol{\theta}}_i \\
&= \mathbf{A}_t^{K_t}\dots\mathbf{A}_1^{k_1}\boldsymbol{\theta}_{1,0} + \sum_{i=1}^{t-1}\mathbf{A}_t^{K_t}\dots\mathbf{A}_{i+1}^{k_{i+1}}\big(\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_{i+1}\big) - \mathbf{A}_t^{K_t}\dots\mathbf{A}_1^{k_1}\widehat{\boldsymbol{\theta}}_1 + \widehat{\boldsymbol{\theta}}_t \\
&= \mathbf{A}_t^{K_t}\dots\mathbf{A}_1^{k_1}(\boldsymbol{\theta}_{1,0} - \widehat{\boldsymbol{\theta}}_1) + \sum_{i=1}^{t-1}\mathbf{A}_t^{K_t}\dots\mathbf{A}_{i+1}^{k_{i+1}}\big(\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_{i+1}\big) + \widehat{\boldsymbol{\theta}}_t,
\end{aligned} \tag{B.14}$$

which immediately implies

$$\mathbf{x}^\top(\boldsymbol{\mu}_{t,K_t} - \widehat{\boldsymbol{\theta}}_t) = \underbrace{\mathbf{x}^\top\mathbf{A}_t^{K_t}\dots\mathbf{A}_1^{k_1}\big(\boldsymbol{\theta}_{1,0} - \widehat{\boldsymbol{\theta}}_1\big)}_{I_1} + \underbrace{\mathbf{x}^\top\sum_{i=1}^{t-1}\mathbf{A}_t^{K_t}\dots\mathbf{A}_{i+1}^{k_{i+1}}\big(\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_{i+1}\big)}_{I_2}. \tag{B.15}$$

For term I_1, recall the inequalities in (B.7), we have

$$\mathbf{x}^\top\mathbf{A}_t^{K_t}\dots\mathbf{A}_1^{k_1}\big(\boldsymbol{\theta}_{1,0} - \widehat{\boldsymbol{\theta}}_1\big) \le \prod_{i=1}^{t}(1-2\eta_i\lambda_{\min}(\mathbf{V}_i))^{K_i}\|\mathbf{x}\|_2\|\boldsymbol{\theta}_{1,0} - \widehat{\boldsymbol{\theta}}_1\|_2. \tag{B.16}$$

Recall that in Algorithm 1, we choose $\boldsymbol{\theta}_{1,0} = \mathbf{0}$ and $\widehat{\boldsymbol{\theta}}_1 = \mathbf{V}_1^{-1}\mathbf{b}_1 = \mathbf{0}$. We have $I_1 = 0$. Now let's look at term I_2. Note that by Sherman–Morrison formula we have

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_{t+1} - \widehat{\boldsymbol{\theta}}_t &= \mathbf{V}_{t+1}^{-1}\mathbf{b}_{t+1} - \mathbf{V}_t^{-1}\mathbf{b}_t \\
&= \frac{\mathbf{V}_t^{-1}\mathbf{x}_t\big(r_t - \mathbf{x}_t^\top\widehat{\boldsymbol{\theta}}_t\big)}{1+\mathbf{x}_t^\top\mathbf{V}_t^{-1}\mathbf{x}_t} \\
&= \frac{\mathbf{V}_t^{-1}\mathbf{x}_t\big(\mathbf{x}_t^\top\boldsymbol{\theta}^* - \mathbf{x}_t^\top\widehat{\boldsymbol{\theta}}_t\big)}{1+\mathbf{x}_t^\top\mathbf{V}_t^{-1}\mathbf{x}_t} + \frac{\mathbf{V}_t^{-1}\mathbf{x}_t\xi_t}{1+\mathbf{x}_t^\top\mathbf{V}_t^{-1}\mathbf{x}_t}.
\end{aligned}$$

Therefore, I_2 can be bounded as follows.

$$\begin{aligned}
I_2 &\le \left|\mathbf{x}^\top\sum_{i=1}^{t-1}\mathbf{A}_t^{K_t}\dots\mathbf{A}_{i+1}^{k_{i+1}}\mathbf{V}_i^{-1}\mathbf{x}_i\frac{\xi_i + \mathbf{x}_i^\top\big(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_t\big)}{1+\mathbf{x}_i^\top\mathbf{V}_i^{-1}\mathbf{x}_i}\right| \\
&\le \sum_{i=1}^{t-1}\big|\mathbf{x}^\top\mathbf{A}_t^{K_t}\dots\mathbf{A}_{i+1}^{k_{i+1}}\mathbf{V}_i^{-1}\mathbf{x}_i\big|\left|\frac{\xi_i + \mathbf{x}_i^\top\big(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_t\big)}{1+\mathbf{x}_i^\top\mathbf{V}_i^{-1}\mathbf{x}_i}\right| \\
&\le \sum_{i=1}^{t-1}\big|\mathbf{x}^\top\mathbf{A}_t^{K_t}\dots\mathbf{A}_{i+1}^{k_{i+1}}\mathbf{V}_i^{-1}\mathbf{x}_i\big|\big[|\xi_i| + \big|\mathbf{x}_i^\top\big(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_t\big)\big|\big] \\
&\le \sum_{i=1}^{t-1}\|\mathbf{A}_t^{K_t}\dots\mathbf{A}_{i+1}^{k_{i+1}}\mathbf{x}\|_{\mathbf{V}_i^{-1}}\|\mathbf{x}_i\|_{\mathbf{V}_i^{-1}}\big|\big[|\xi_i| + \big|\mathbf{x}_i^\top\big(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_t\big)\big|\big] \\
&\le \sum_{i=1}^{t-1}\prod_{j=i+1}^{t}(1-2\eta_j\lambda_{\min}(\mathbf{V}_j))^{K_j}\|\mathbf{x}\|_{\mathbf{V}_i^{-1}}\|\mathbf{x}_i\|_{\mathbf{V}_i^{-1}}\big|\big(|\xi_i| + \big|\mathbf{x}_i^\top\big(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_t\big)\big|\big).
\end{aligned} \tag{B.17}$$

Note that under event $E_{R,t}$, by Lemma A.3, we have $\big|\mathbf{x}_i^\top\big(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_t\big)\big| \le g_R(t)\|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}} \le g_R(t)$ and that $|\xi_i| \le R\sqrt{1+2\log t}$. Combining the above results, we have

$$|\mathbf{x}^\top(\boldsymbol{\mu}_{t,K_t} - \widehat{\boldsymbol{\theta}}_t)| \le \sum_{i=1}^{t-1}\prod_{j=i+1}^{t}(1-2\eta_j\lambda_{\min}(\mathbf{V}_j))^{K_j}\|\mathbf{x}_i\|_{\mathbf{V}_i^{-1}}\|\mathbf{x}\|_{\mathbf{V}_i^{-1}}\big(R\sqrt{1+2\log t} + g_R(i)\big). \tag{B.18}$$

Substituting (B.13) and (B.18) into (B.5), we have with probability at least $1-1/t^2$ that

$$\begin{aligned}
|\mathbf{x}^\top(\boldsymbol{\theta}_{t,K_t}-\widehat{\boldsymbol{\theta}}_t)| &\le \sum_{i=1}^{t-1}\prod_{j=i+1}^{t}(1-2\eta_j\lambda_{\min}(\mathbf{V}_j))^{K_j}\big(R\sqrt{1+2\log t}+g_R(i)\big)\|\mathbf{x}_i\|_{\mathbf{V}_i^{-1}}\|\mathbf{x}\|_{\mathbf{V}_i^{-1}}\\
&\quad+2\sqrt{\frac{2d\log t}{3\beta_T}}\bigg(\|\mathbf{x}\|_{\mathbf{V}_t^{-1}}+\sum_{i=1}^{t-1}\prod_{j=i+1}^{t}(1-2\eta_j\lambda_{\min}(\mathbf{V}_j))^{K_j}\|\mathbf{x}_i\|_{\mathbf{V}_i^{-1}}\|\mathbf{x}\|_{\mathbf{V}_i^{-1}}\bigg)\\
&:=Q, \qquad\qquad (\text{B.19})
\end{aligned}$$

where we denote the upper bound as Q. For any $j\in[t]$, recall that we require $\eta_j<1/(2\lambda_{\max}(\mathbf{V}_j))$. Now we choose $\eta_j=1/(4\lambda_{\max}(\mathbf{V}_j))$. Then it holds that

$$(1-2\eta_j\lambda_{\min}(\mathbf{V}_j))^{K_j}=(1-1/(2\kappa_j))^{K_j},$$

where $\kappa_j=\lambda_{\max}(\mathbf{V}_j)/\lambda_{\min}(\mathbf{V}_j)$. In order to ensure the above quantity be smaller than ϵ, we need

$$K_j\ge\frac{\log(1/\epsilon)}{\log\frac{1}{1-1/(2\kappa_j)}}. \qquad (\text{B.20})$$

Note that $e^{-x}>1-x$ for any $0<x<1$. Since $1/(2\kappa_j)\le 1/2$, we have $\log(1/(1-1/(2\kappa_j)))\ge 1/(2\kappa_j)$. Therefore, it suffices to set $K_j\ge 2\kappa_j\log(1/\epsilon)$ to ensure $(1-1/(2\kappa_j))^{K_j}\le\epsilon$. Recall the definition $g_R(t)=R\sqrt{d\log(t^3/\delta)}+1$. Then we have

$$\begin{aligned}
R\sqrt{1+2\log t}+g_R(i)&\le R\sqrt{2\log t}+R\sqrt{d\log(t^3/\delta)}+R+1\\
&\le 3R\sqrt{2d\log(t^3/\delta)}.
\end{aligned}$$

By setting $\epsilon=(3R\sqrt{2dt\log(t^3/\delta)})^{-1}$, we obtain

$$\begin{aligned}
Q&\le\sum_{i=1}^{t-1}\epsilon^{t-i}3R\sqrt{2d\log(t^3/\delta)}\|\mathbf{x}\|_2+2\sqrt{\frac{2d\log t}{3\beta_T}}\bigg(\|\mathbf{x}\|_{\mathbf{V}_t^{-1}}+\sum_{i=1}^{t-1}\epsilon^{t-i}\|\mathbf{x}\|_2\bigg)\\
&\le\sum_{i=1}^{t-1}\epsilon^{t-1-i}\|\mathbf{x}\|_{\mathbf{V}_t^{-1}}+2\sqrt{\frac{2d\log t}{3\beta_T}}\bigg(\|\mathbf{x}\|_{\mathbf{V}_t^{-1}}+\sum_{i=1}^{t-1}\epsilon^{t-1-i}\|\mathbf{x}\|_{\mathbf{V}_t^{-1}}\bigg)\\
&\le\bigg(5\sqrt{\frac{2d\log t}{3\beta_T}}+3/2\bigg)\|\mathbf{x}\|_{\mathbf{V}_t^{-1}}, \qquad\qquad (\text{B.21})
\end{aligned}$$

where the second inequality is due to $\|\mathbf{x}\|_{\mathbf{V}_t^{-1}}\ge 1/\sqrt{t}\|\mathbf{x}\|_2$, and the third inequality is due to $\sum_{i=1}^{t-1}\epsilon^{t-1-i}=\sum_{i=0}^{t-2}\epsilon^i<1/(1-\epsilon)\le 3/2$. Therefore, it holds that

$$\begin{aligned}
\mathbb{P}(E_{S,t})&\ge\mathbb{P}\big(\big|\mathbf{x}^\top(\boldsymbol{\theta}_{t,k}-\widehat{\boldsymbol{\theta}}_t)\big|\le Q\big)\\
&\ge 1-1/t^2,
\end{aligned}$$

where the last inequality is due to (B.19). □

B.4. Proof of Lemma A.5

Proof of Lemma A.5. Based on the mean and covariance matrix defined in (B.2) and (B.3), we have that $\mathbf{x}_t^{*\top}\boldsymbol{\theta}_{t,k}$ follows the distribution $\mathcal{N}(\mathbf{x}_t^{*\top}\boldsymbol{\mu}_{t,k},\mathbf{x}_t^{*\top}\boldsymbol{\Sigma}_{t,k}\mathbf{x}_t^*)$. By Lemma C.1, we have

$$\begin{aligned}
P\big(\mathbf{x}_t^{*\top}\boldsymbol{\theta}_{t,k}>\mathbf{x}_t^{*\top}\boldsymbol{\theta}^*\big)&=P\left(\frac{\mathbf{x}_t^{*\top}\boldsymbol{\theta}_{t,k}-\mathbf{x}_t^{*\top}\boldsymbol{\mu}_{t,k}}{\sqrt{\mathbf{x}_t^{*\top}\boldsymbol{\Sigma}_{t,k}\mathbf{x}_t^*}}>\frac{\mathbf{x}_t^{*\top}\boldsymbol{\theta}^*-\mathbf{x}_t^{*\top}\boldsymbol{\mu}_{t,k}}{\sqrt{\mathbf{x}_t^{*\top}\boldsymbol{\Sigma}_{t,k}\mathbf{x}_t^*}}\right)\\
&\ge\frac{1}{2\sqrt{2\pi}}e^{-Z_t^2/2}, \qquad\qquad (\text{B.22})
\end{aligned}$$

where the inequality holds when $|Z_t| < 1$ and we define $Z_t = \big(\mathbf{x}_t^{*\top}\boldsymbol{\theta}^* - \mathbf{x}_t^{*\top}\boldsymbol{\mu}_{t,k}\big)/\sqrt{\mathbf{x}_t^{*\top}\boldsymbol{\Sigma}_{t,k}\mathbf{x}_t^*}$. In the rest of the proof, we will show that $|Z_t| < 1$ under event $E_{R,t}$. First note that by triangle inequality we have

$$\begin{aligned} |\mathbf{x}_t^{*\top}\boldsymbol{\theta}^* - \mathbf{x}_t^{*\top}\boldsymbol{\mu}_{t,k}| &\le |\mathbf{x}_t^{*\top}\boldsymbol{\theta}^* - \mathbf{x}_t^{*\top}\widehat{\boldsymbol{\theta}}_t| + |\mathbf{x}_t^{*\top}\widehat{\boldsymbol{\theta}}_t - \mathbf{x}_t^{*\top}\boldsymbol{\mu}_{t,k}| \\ &\le \sum_{i=1}^{t-1}\prod_{j=i+1}^{t}\big(1-2\eta_j\lambda_{\min}(\mathbf{V}_j)\big)^{K_j}\|\mathbf{x}_i\|_{\mathbf{V}_i^{-1}}\|\mathbf{x}_t^*\|_{\mathbf{V}_i^{-1}}\big(R\sqrt{1+2\log t}+g_R(i)\big) \\ &\quad + g_R(t)\|\mathbf{x}_t^*\|_{\mathbf{V}_t^{-1}}, \end{aligned} \tag{B.23}$$

where in the second inequality we used the conclusion in Lemma A.3 since event $E_{R,t}$ holds and (B.18). When we choose $K_j \ge \kappa_j \log(3R\sqrt{2dt\log(t^3/\delta)})$, we further have

$$\begin{aligned} |\mathbf{x}_t^{*\top}\boldsymbol{\theta}^* - \mathbf{x}_t^{*\top}\boldsymbol{\mu}_{t,k}| &\le \sum_{i=1}^{t-1} t^{-(t-i)}\|\mathbf{x}_t^*\|_2\big(R\sqrt{1+2\log t}+g_R(t)\big) + g_R(t)\|\mathbf{x}_t^*\|_{\mathbf{V}_t^{-1}} \\ &\le \frac{1}{3R\sqrt{2dt\log(t^3/\delta)}}\|\mathbf{x}_t^*\|_2\big(R\sqrt{1+2\log t}+g_R(t)\big) + g_R(t)\|\mathbf{x}_t^*\|_{\mathbf{V}_t^{-1}} \\ &\le (R\sqrt{d\log(t^3/\delta)}+2)\|\mathbf{x}_t^*\|_{\mathbf{V}_t^{-1}}, \end{aligned} \tag{B.24}$$

where in the last inequality we used the fact that $g_R(t) = R\sqrt{d\log(t^3/\delta)}+1$.

On the other hand, recall the definition of $\boldsymbol{\Sigma}_{t,K_t}$ in Proposition A.1. Following similar proof as in the previous lemma, we have

$$\begin{aligned} \mathbf{x}_t^{*\top}\boldsymbol{\Sigma}_{t,k}\mathbf{x}_t^* &= \sum_{i=1}^{t}\frac{1}{\beta_i}\mathbf{x}_t^{*\top}\mathbf{A}_t^{K_t}\ldots\mathbf{A}_{i+1}^{k_{i+1}}\big(\mathbf{I}-\mathbf{A}_i^{2K_i}\big)\mathbf{V}_i^{-1}(\mathbf{I}+\mathbf{A}_i)^{-1}\mathbf{A}_{i+1}^{k_{i+1}}\ldots\mathbf{A}_t^{K_t}\mathbf{x}_t^* \\ &\ge \sum_{i=1}^{t}\frac{1}{2\beta_i}\mathbf{x}_t^{*\top}\mathbf{A}_t^{K_t}\ldots\mathbf{A}_{i+1}^{k_{i+1}}\big(\mathbf{I}-\mathbf{A}_i^{2K_i}\big)\mathbf{V}_i^{-1}\mathbf{A}_{i+1}^{k_{i+1}}\ldots\mathbf{A}_t^{K_t}\mathbf{x}_t^*. \end{aligned} \tag{B.25}$$

Note that by definition $\mathbf{A}_i = \mathbf{I} - 2\eta_i\mathbf{V}_i$ and $\mathbf{V}_i$ is symmetric. Therefore, $\mathbf{A}_i$ and $\mathbf{V}_i^{-1}$ commute, and it holds that

$$\begin{aligned} \mathbf{A}_i^{2K_i}\mathbf{V}_i^{-1} &= (\mathbf{I}-2\eta_i\mathbf{V}_i)\ldots(\mathbf{I}-2\eta_i\mathbf{V}_i)(\mathbf{I}-2\eta_i\mathbf{V}_i)\mathbf{V}_i^{-1} \\ &= (\mathbf{I}-2\eta_i\mathbf{V}_i)\ldots(\mathbf{I}-2\eta_i\mathbf{V}_i)\mathbf{V}_i^{-1}(\mathbf{I}-2\eta_i\mathbf{V}_i) \\ &= \mathbf{A}_i^{K_i}\mathbf{V}_i^{-1}\mathbf{A}_i^{K_i}. \end{aligned} \tag{B.26}$$

Hence we have

$$\begin{aligned} \mathbf{x}_t^{*\top}\boldsymbol{\Sigma}_{t,k}\mathbf{x}_t^* &\ge \sum_{i=1}^{t}\frac{1}{2\beta_i}\mathbf{x}_t^{*\top}\mathbf{A}_t^{K_t}\ldots\mathbf{A}_{i+1}^{k_{i+1}}\big(\mathbf{V}_i^{-1}-\mathbf{A}_i^{K_i}\mathbf{V}_i^{-1}\mathbf{A}_i^{K_i}\big)\mathbf{A}_{i+1}^{k_{i+1}}\ldots\mathbf{A}_t^{K_t}\mathbf{x}_t^* \\ &= \frac{1}{2\beta_T}\sum_{i=1}^{t-1}\mathbf{x}_t^{*\top}\mathbf{A}_t^{K_t}\ldots\mathbf{A}_{i+1}^{k_{i+1}}\big(\mathbf{V}_i^{-1}-\mathbf{V}_{i+1}^{-1}\big)\mathbf{A}_{i+1}^{k_{i+1}}\ldots\mathbf{A}_t^{K_t}\mathbf{x}_t^* \\ &\quad - \frac{1}{2\beta_T}\mathbf{x}_t^{*\top}\mathbf{A}_t^{K_t}\ldots\mathbf{A}_1^{k_1}\mathbf{V}_1^{-1}\mathbf{A}_1^{k_1}\ldots\mathbf{A}_t^{K_t}\mathbf{x}_t^* + \frac{1}{2\beta_T}\mathbf{x}_t^{*\top}\mathbf{V}_t^{-1}\mathbf{x}_t^*. \end{aligned} \tag{B.27}$$

where we used the choice of $1/\beta_i = 1/\beta_T$ for all i. By the definition in (2.2) and Sherman-Morrison formula, we have

$$\mathbf{V}_i^{-1} - \mathbf{V}_{i+1}^{-1} = \mathbf{V}_i^{-1} - \big(\mathbf{V}_i + \mathbf{x}_i\mathbf{x}_i^\top\big)^{-1} = \frac{\mathbf{V}_i^{-1}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{V}_i^{-1}}{1+\|\mathbf{x}_i\|_{\mathbf{V}_i^{-1}}^2}, \tag{B.28}$$

which immediately implies

$$
\begin{aligned}
\mathbf{x}_t^{*\top}\mathbf{A}_t^{K_t}\dots\mathbf{A}_{i+1}^{k_{i+1}}\big(\mathbf{V}_i^{-1}-\mathbf{V}_{i+1}^{-1}\big)\mathbf{A}_{i+1}^{k_{i+1}}\dots\mathbf{A}_t^{K_t}\mathbf{x}_t^* &= \mathbf{x}_t^{*\top}\mathbf{A}_t^{K_t}\dots\mathbf{A}_{i+1}^{k_{i+1}}\frac{\mathbf{V}_i^{-1}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{V}_i^{-1}}{1+\|\mathbf{x}_i\|_{\mathbf{V}_i^{-1}}^2}\mathbf{A}_{i+1}^{k_{i+1}}\dots\mathbf{A}_t^{K_t}\mathbf{x}_t^*\\
&\le \big(\mathbf{x}_t^{*\top}\mathbf{A}_t^{K_t}\dots\mathbf{A}_{i+1}^{k_{i+1}}\mathbf{V}_i^{-1}\mathbf{x}_i\big)^2\\
&\le \|\mathbf{A}_t^{K_t}\dots\mathbf{A}_{i+1}^{k_{i+1}}\mathbf{V}_i^{-1/2}\mathbf{x}_t^*\|_2^2\cdot\|\mathbf{V}_i^{-1/2}\mathbf{x}_i\|_2^2\\
&\le \prod_{j=i+1}^{t}(1-2\eta_j\lambda_{\min}(\mathbf{V}_j))^{2K_j}\|\mathbf{x}_i\|_{\mathbf{V}_i^{-1}}^2\|\mathbf{x}_t^*\|_{\mathbf{V}_i^{-1}}^2,
\end{aligned}
$$

where we used $0<1/i\le\|\mathbf{x}_t^*\|_{\mathbf{V}_i^{-1}}\le 1$. Therefore, we have

$$
\begin{aligned}
\mathbf{x}_t^{*\top}\mathbf{\Sigma}_{t,K_t}\mathbf{x}_t^* \ge \frac{1}{2\beta_T}\mathbf{x}_t^{*\top}\mathbf{V}_t^{-1}\mathbf{x}_t^* - \frac{1}{2\beta_T}\prod_{i=1}^{t}(1-2\eta_i\lambda_{\min}(\mathbf{V}_i))^{2K_i}\|\mathbf{x}_t^*\|_{\mathbf{V}_1^{-1}}^2\\
-\frac{1}{2\beta_T}\sum_{i=1}^{t-1}\prod_{j=i+1}^{t}(1-2\eta_j\lambda_{\min}(\mathbf{V}_j))^{2K_j}\|\mathbf{x}_i\|_{\mathbf{V}_i^{-1}}^2\|\mathbf{x}_t^*\|_{\mathbf{V}_i^{-1}}^2.
\end{aligned}
\tag{B.29}
$$

Similar to the proof of Lemma A.4, when we choose $K_j\ge\kappa_j\log(3\sqrt{t})$, we have

$$
\begin{aligned}
\|\mathbf{x}_t^*\|_{\mathbf{\Sigma}_{t,K_t}} &\ge \frac{1}{2\beta_T}\bigg(\|\mathbf{x}_t^*\|_{\mathbf{V}_t^{-1}}-\frac{\|\mathbf{x}_t^*\|_2}{(3\sqrt{t})^t}-\sum_{i=1}^{t-1}\frac{1}{(3\sqrt{t})^{t-i}}\|\mathbf{x}_t^*\|_2\bigg)\\
&\ge \frac{1}{2\beta_T}\bigg(\|\mathbf{x}_t^*\|_{\mathbf{V}_t^{-1}}-\frac{1}{3\sqrt{t}}\|\mathbf{x}_t^*\|_2-\frac{1}{6\sqrt{t}}\|\mathbf{x}_t^*\|_2\bigg)\\
&\ge \frac{1}{4\beta_T}\|\mathbf{x}_t^*\|_{\mathbf{V}_t^{-1}},
\end{aligned}
\tag{B.30}
$$

where we used the fact that $\lambda_{\min}(\mathbf{V}_t^{-1})\ge 1/t$.

Therefore, according to (B.24) and (B.30), it holds that

$$
|Z_t| = \left|\frac{\mathbf{x}_{a_t}^\top\boldsymbol{\theta}^*-\mathbf{x}_{a_t}^\top\boldsymbol{\mu}_{t,k}}{\sqrt{\mathbf{x}^\top\mathbf{\Sigma}_{t,k}\mathbf{x}}}\right| \le \frac{R\sqrt{d\log(t^3/\delta)}+2}{1/(4\beta_T)},
\tag{B.31}
$$

which implies $|Z_t|<1$ when $\beta_t^{-1}=4R\sqrt{d\log\frac{t^3}{\delta}}+8$. This completes our proof. □

B.5. Proof of Lemma A.6

Proof of Lemma A.6. Since the algorithm chooses the arm to pull based on the estimated reward $\mathbf{x}^\top\boldsymbol{\theta}_{t,k}$, as long as we can find an arm in the unsaturated set that beats all arms in the saturated set, we will have $\mathbf{x}_t\in\mathcal{U}_t$. Recall the definition in (A.2), we know that the best arm is in the unsaturated set, i.e., $\mathbf{x}_t^*\in\mathcal{U}_t$. Therefore, it holds that

$$
\{\mathbf{x}_t\in\mathcal{U}_t\}\supseteq\big\{\mathbf{x}_t^{*\top}\boldsymbol{\theta}_{t,k}>\mathbf{x}^\top\boldsymbol{\theta}_{t,k},\forall\mathbf{x}\in\mathcal{S}_t\big\}.
\tag{B.32}
$$

Conditional on event $E_{R,t}$, we have

$$
\begin{aligned}
\mathbb{P}\big(\mathbf{x}_t^{*\top}\boldsymbol{\theta}_{t,k}>\mathbf{x}_t^{*\top}\boldsymbol{\theta}^*\big) &= \mathbb{P}\big(\mathbf{x}_t^{*\top}\boldsymbol{\theta}_{t,k}>\mathbf{x}_t^{*\top}\boldsymbol{\theta}^*|E_{S,t}\big)\mathbb{P}(E_{S,t})+\mathbb{P}\big(\mathbf{x}_t^{*\top}\boldsymbol{\theta}_{t,k}>\mathbf{x}_t^{*\top}\boldsymbol{\theta}^*|E_{S,t}^c\big)\mathbb{P}(E_{S,t}^c)\\
&\le \mathbb{P}\big(\mathbf{x}_t^{*\top}\boldsymbol{\theta}_{t,k}>\mathbf{x}_t^{*\top}\boldsymbol{\theta}^*|E_{S,t}\big)+\mathbb{P}(E_{S,t}^c).
\end{aligned}
\tag{B.33}
$$

Recall the definition of the gap and the saturated set in (A.1), we have that $\mathbf{x}_t^{*\top}\boldsymbol{\theta}^*=\mathbf{x}^\top\boldsymbol{\theta}^*+\Delta(t)(\mathbf{x})\ge\mathbf{x}^\top\boldsymbol{\theta}^*+g_t(\mathbf{x})$ for any $\mathbf{x}\in\mathcal{S}_t$, where $g_t(\mathbf{x})$ is defined as in (A.9). Then it holds that

$$
\begin{aligned}
\mathbb{P}\big(\mathbf{x}_t^{*\top}\boldsymbol{\theta}_{t,k}>\mathbf{x}_t^{*\top}\boldsymbol{\theta}^*|E_{S,t}\big) &\le \mathbb{P}\big(\mathbf{x}_t^{*\top}\boldsymbol{\theta}_{t,k}>\mathbf{x}^\top\boldsymbol{\theta}^*+g_t(\mathbf{x}),\forall\mathbf{x}\in\mathcal{S}_t|E_{S,t}\big)\\
&\le \mathbb{P}\big(\mathbf{x}_t^{*\top}\boldsymbol{\theta}_{t,k}>\mathbf{x}^\top\boldsymbol{\theta}_{t,k},\forall\mathbf{x}\in\mathcal{S}_t|E_{S,t}\big),
\end{aligned}
\tag{B.34}
$$

where the second inequality is true since $|\mathbf{x}^\top(\boldsymbol{\theta}_{t,k} - \boldsymbol{\theta}^*)| \le g_t(\mathbf{x})$ based on the definition of events $E_{R,t}$ and $E_{S,t}$ in (A.6) and (A.7) respectively. Therefore, we have

$$\begin{aligned}\mathbb{P}(\mathbf{x}_t \in \mathcal{U}_t) &\ge \mathbb{P}\big(\mathbf{x}_t^{*\top}\boldsymbol{\theta}_{t,k} > \mathbf{x}_t^{*\top}\boldsymbol{\theta}^* | E_{S,t}\big) \\ &\ge \mathbb{P}\big(\mathbf{x}_t^{*\top}\boldsymbol{\theta}_{t,k} > \mathbf{x}_t^{*\top}\boldsymbol{\theta}^*\big) - \mathbb{P}(E_{S,t}^c(t)) \\ &\ge \frac{1}{2\sqrt{2e\pi}} - \frac{1}{t^2},\end{aligned}$$

where the last inequality holds due to Lemma A.4 and Lemma A.5. □

C. Auxiliary Lemmas

Lemma C.1. *(Abramowitz & Stegun, 1964) Suppose Z is a Gaussian random variable $Z \sim \mathcal{N}(\mu, \sigma^2)$, where $\sigma > 0$. For $0 \le z \le 1$, we have*

$$\mathbb{P}(Z > \mu + z\sigma) \ge \frac{1}{\sqrt{8\pi}} e^{-\frac{z^2}{2}}, \quad \mathbb{P}(Z < \mu - z\sigma) \ge \frac{1}{\sqrt{8\pi}} e^{-\frac{z^2}{2}}.$$

And for $z \ge 1$, we have

$$\frac{e^{-z^2/2}}{2z\sqrt{\pi}} \le \mathbb{P}(|Z - \mu| > z\sigma) \le \frac{e^{-\frac{z^2}{2}}}{z\sqrt{\pi}}.$$

Lemma C.2. *(Vershynin, 2010) Let $\boldsymbol{X}$ be a σ^2-subGaussian vector in $\mathbb{R}^d$. Then we have $\mathbb{E}[\|\boldsymbol{X}\|_2] \le 4\sigma\sqrt{d}$. For $\delta \in (0,1)$, with probability at least $1-\delta$ that $\|\boldsymbol{X}\|_2 \le 4\sigma\sqrt{d} + 2\sigma\sqrt{\log(1/\delta)}$.*

The following lemma introduces the Azuma-Hoeffding inequality for super-martingale.

Lemma C.3. *Suppose $\{X_k\}_{k=0,1,\ldots}$ is a super-martingale and satisfies $|X_{k+1} - X_k| < c_{k+1}$ for all $k \ge 0$. Then for any $\epsilon > 0$, we have*

$$\mathbb{P}(X_T - X_0 \ge \epsilon) \le \exp\bigg(-\frac{\epsilon^2}{2\sum_{t=1}^T c_t^2}\bigg).$$