Tight Regret and Complexity Bounds for Thompson Sampling via Langevin Monte Carlo

Tom Huix
CMAP, CNRS, Ecole Polytechnique,
Institut Polytechnique de Paris,
91120 Palaiseau, France.

Matthew Zhang
Department of Computer Science
at University of Toronto,
and Vector Institute.

Alain Durmus
CMAP, CNRS, Ecole Polytechnique,
Institut Polytechnique de Paris,
91120 Palaiseau, France.

Abstract

In this paper, we consider high dimensional contextual bandit problems. Within this setting, Thompson Sampling and its variants have been proposed and successfully applied to multiple machine learning problems. Existing theory on Thompson Sampling shows that it has suboptimal dimension dependency in contrast to upper confidence bound (UCB) algorithms. To circumvent this issue and obtain optimal regret bounds, (Zhang 2021) recently proposed to modify Thompson Sampling by enforcing more exploration and hence is able to attain optimal regret bounds. Nonetheless, this analysis does not permit tractable implementation in high dimensions. The main challenge therein is the simulation of the posterior samples at each step given the available observations. To overcome this, we propose and analyze the use of Markov Chain Monte Carlo methods. As a corollary, we show that for contextual linear bandits, using Langevin Monte Carlo (LMC) or Metropolis Adjusted Langevin Algorithm (MALA), our algorithm attains optimal regret bounds of $\tilde{\mathcal{O}}(d\sqrt{T})$. Furthermore, we show that this is obtained with $\tilde{\mathcal{O}}(dT^4)$, $\tilde{\mathcal{O}}(dT^2)$ data evaluations respectively for LMC and MALA. Finally, we validate our findings through numerical simulations and show that we outperform vanilla Thompson sampling in high dimensions.

1 Introduction

Bandit models have proven to be one of the most successful paradigms for decision making in random environments (Robbins 1952; Katehakis and Veinott 1987; Berry and Fristedt 1985; Auer, Cesa-Bianchi, and Fischer 2002; Lattimore and Szepesvári 2020). Formally, it models an agent which for some rounds has to choose between several potential actions. The agent selects each action according to its current policy and receives a reward once this action is made. In this paper, we are especially interested in the contextual bandit problem (Langford and Zhang 2007) which supposes that the set of actions at each round and the corresponding reward mean function depend on a context vector which is specified by the environment under consideration. This setting has been developed and studied intensively over the past decade (Langford and Zhang 2007; Filippi et al. 2010; Abbasi-Yadkori, Pál, and Szepesvári 2011; Chu et al. 2011; Agrawal and Goyal 2013; Li, Lu, and Zhou 2017; Lale et al. 2019; Kveton et al. 2020a) and has been successfully applied in various fields; see e.g. for applications in content recommendation, mobile health and finance (Li, Chu, et al. 2010; Agarwal, Bird, et al. 2016; Tewari and Murphy 2017; Bouneffouf, Rish, and Aggarwal 2020). To address this problem, bandits algorithms deal with the research and design of efficient algorithms that seek to optimize the cumulative reward. To this end, they recursively define a sequence of policies which is adjusted at each round given the previous historical state-action-reward tuples. The main challenge towards the adaptation and implementation of these policies is to find a compromise between (1) exploitation of the arms with good empirical expected rewards and (2) exploration of the worse arms with under-sampled rewards.

The approaches to maximizing cumulative reward (alternatively, minimizing cumulative regret) can be broadly divided into two categories. Maximum likelihood methods with optimistic adjustment (UCB) follow the principle of optimism in the face of uncertainty and were adopted in (Auer, Cesa-Bianchi, and Fischer 2002; Ménard and Garivier 2017; Chu et al. 2011; Abbasi-Yadkori, Pál, and Szepesvári 2011; Li, Lu, and Zhou 2017; Zhou, Li, and Gu 2020; Zenati et al. 2022; Foster and Rakhlin 2020). The second approach is based on the Bayesian paradigm, and involves the sampling of a sequence of posterior distributions associated with a statistical model for the reward function; see e.g. , (Thomp-

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

son 1933; Agrawal and Goyal 2012; Kaufmann, Cappé, and Garivier 2012; Russo and Van Roy 2016; Russo and Van Roy 2014; Jin et al. 2021). This is called Thompson sampling (TS). Both of these aim to inject uncertainty into the model in order to encourage "exploration"-type behaviour, and have demonstrated their efficiency and robustness in a wide range of applications. In addition, they come with important theoretical guarantees, complementing each other while providing comparable results empirically; see (Chapelle and Li 2011). However, existing regret bounds for TS are often sub-optimal when compared to analogous rates for UCB. In particular, the bounds established in (Agrawal and Goyal 2012) for Thompson Sampling (TS) applied to linear models are of order $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ where d is the dimension and T the time horizon. These bounds are worse by a factor d than the ones proved in (Dani, Hayes, and Kakade 2008; Abbasi-Yadkori, Pál, and Szepesvári 2011) for Linear UCB type algortihms. In fact (Zhang 2021) showed that this discrepancy between usual TS and UCB cannot be reduced, providing an instance where regret bounds for usual TS can be lower bounded by $\widetilde{\mathcal{O}}(T)$ whereas results on UCB from (Foster and Rakhlin 2020) achieve a cumulative regret of order $\widetilde{\mathcal{O}}(\sqrt{KT})$, where K is the number of possible actions. To circumvent this issue, (Zhang 2021) proposed to modify the likelihood function in TS by adding a penalty term to enforce more optimistic exploration. In addition, the author was able to show that this version of TS, coined Feel-Good Thompson sampling (FG-TS), comes with an upper bound for the cumulative regret which is of order $\widetilde{\mathcal{O}}(d\sqrt{T})$. This matches the minimax regret lower bound established in (Agarwal, Dudik, et al. 2012).

One defect in the methodology and the analysis of (Zhang 2021) is that they do not take into account that the sequence of posterior distributions associated with FG-TS is intractable to sample from in practice, even for linear contextual bandits. This is in contrast to the standard TS algorithm. The objective of the present paper is precisely to fill this gap. To address this problem, we propose the use of Markov Chain Monte Carlo methods at each round to obtain approximate samples from the target posterior distribution.

Organization of the paper and contributions. We summarize our contributions as follows:

• We first introduce sFG-TS, a version of FG-TS where in comparison to (Zhang 2021), the likelihood is a smooth function of the parameter. This results in posteriors which are also smooth under a smooth prior distribution, which is beneficial since targeting smooth distributions is generally easier for MCMC algorithms. This is especially true for MCMC methods which are based on gradient information (Durmus, Moulines, and Pereyra 2018).
• We propose MCMC-sFG-TS, in which the posteriors at each round are approximately sampled from a generic MCMC algorithm.
• We adapt and extend the analysis of (Zhang 2021) to the setting where only approximate samples from the posteriors are used in our TS algorithm.
• We apply our result to linear contextual bandit problems and show that our method achieves optimal regret bounds of order $\widetilde{\mathcal{O}}(d\sqrt{T})$ when using the Langevin Monte Carlo (LMC) or its metropolized version (MALA) at each round.
• In addition, we validate our results through some practical examples: firstly, with a toy Gaussian problem and secondly with the Yahoo! Front Page Today Module dataset Li, Chu, et al. 2010. Our algorithms consistently outperform the vanilla Thompson sampling benchmarks in both settings.

Notation For any two probability measures on a measurable space $(\mathsf{X}, \mathcal{X})$, we denote by $||\mu - \nu||_{\mathrm{TV}} = \sup|\int f \mathrm{d}\mu - \int f \mathrm{d}\nu|$ where the supremum is taken over the set of measurable and bounded (by one) functions from X to $\mathbb{R}$. For $n \geq 1$, we refer to the set of integers between 1 and n with the notation $[n]$. The d-multidimensional Gaussian probability distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ is denoted by $\mathrm{N}(\mu, \Sigma)$. The norm $\|\cdot\|$ will refer to the 2-norm for vectors, and the operator norm for matrices. By abuse of notation we will use the same symbol for both a measure and its density.

2 Contextual bandit and Thompson sampling methods

We describe the contextual Bandit framework below. Let X be a contextual set and $\mathcal{A} : \mathsf{X} \to 2^{\mathsf{A}}$ be a set-valued action map, where 2^{A} denotes the power set of the action space A. While we do not assume that A is finite, we suppose $\sup_{x\in\mathsf{X}} \mathrm{Card}(\mathcal{A}(x)) < \infty$. In the sequel, we consider policies $\pi : \mathsf{X} \to \mathsf{A}$ such that for any $x \in \mathsf{X}$, $\pi(x) \in \mathcal{A}(x)$, and π can be either deterministic or random. Given a horizon $T \in \mathbb{N}^*$ let the following procedure define the bandit framework:

Contextual bandit process. At each iteration $t \in [T]$ and given the past observations $\mathrm{D}_{t-1} = \{(x_s, a_s, r_s)\}_{s<t}$:

• The agent observes a contextual vector $x_t \in \mathsf{X}$;
• The agent chooses a policy π_t from some conditional distribution $\mathbb{Q}_t(\cdot|\mathrm{D}_{t-1})$ and sets its action to $a_t = \pi_t(x_t)$;
• The agent receives a reward r_t with conditional distribution $\mathrm{R}(\cdot|x_t, a_t)$ given D_{t-1} (where R is a Markov kernel on $(\mathsf{A} \times \mathsf{X}) \times \mathsf{R}$, where R is some subset of $\mathbb{R}$).

Given a sequence of conditionals $\mathbb{Q}_{1:T} = \{\mathbb{Q}_t\}_{t\leq T}$, this process defines a distribution on the sequence of policies $\pi_{1:T} = \{\pi_t\}_{t\leq T}$ still denoted by $\mathbb{Q}_{1:T}$ by abuse of notation. The bandit problem then consists in finding the conditional $\{\mathbb{Q}_t\}_{t\leq T}$ that minimizes the cumulative regret that we will define below. However, as the reward distribution R is unknown, the agent has to simultaneously learn this distribution and choose the best policy. This is a classical exploitation/exploration problem. First, define the expected reward under the optimal action and the expected reward under any particular action as the following respectively:

$$f_*(x) = \max_{a\in\mathcal{A}(x)} \int r\,\mathrm{R}(\mathrm{d}r|x,a)\;, \tag{2.1}$$
$$f(x,a) = \int r\,\mathrm{R}(\mathrm{d}r|x,a)\;.$$

We define then the regret at time s with respect to a policy π_s and a context x_s as

$$\mathrm{REG}_s^{\pi_s} = f_\star(x_s) - f(x_s,\pi_s(x_s))\;, \tag{2.2}$$

and finally, we seek to find $\mathbb{Q}_{1:T}$ such that the cumulative regret is minimized

$$\mathrm{CREG}(\mathbb{Q}_{1:T}) = \mathbb{E}_{\pi_{1:T}\sim\mathbb{Q}_{1:T}}[\textstyle\sum_{s\le T}\mathrm{REG}_s^{\pi_s}]\;. \tag{2.3}$$

Thompson sampling (TS) algorithm is a well known algorithm which achieves this goal, with strong performance in practice. First we present the standard Thompson sampling framework to highlight its limitations. Firstly, consider the Gaussian parametric model $\{\mathrm{R}_\theta^{(\mathrm{TS})} : \theta\in\mathbb{R}^d\}$ based on $g:\mathbb{R}^d\times\mathsf{X}\times\mathsf{A}\to\mathbb{R}$, where $\mathrm{R}_\theta^{(\mathrm{TS})}(\cdot|x,a)$ is the Gaussian distribution with mean = $g(\theta,x,a)$ and variance $1/(2\eta)$ for some $\eta>0$. For example, in linear contextual bandits (Chu et al. 2011; Abbasi-Yadkori, Pál, and Szepesvári 2011), $g(\theta,x,a)=\langle a,\theta\rangle$ assuming that $\mathcal{A}(x)\subset\mathbb{R}^d$ for any $x\in\mathsf{X}$. Under the same condition, generalized linear bandits (Filippi et al. 2010; Kveton et al. 2020a) consist of $g(\theta,x,a)=\sigma(\langle a,\theta\rangle)$ for some link function σ. Finally, in neural contextual bandits (Riquelme, Tucker, and Snoek 2018; Zhou, Li, and Gu 2020; Xu, Wen, et al. 2020), g is a neural network taking as input a pair (x,a) and θ stands for the weights of g. Then, the likelihood function associated with the observations D_t at step t is given by

$$\mathrm{L}_t^{(\mathrm{TS})}(\theta|\mathrm{D}_t) \propto \exp\left(-\sum_{s=1}^{t}\ell^{(\mathrm{TS})}(\theta|x_s,a_s,r_s)\right)\;, \tag{2.4}$$

where the negative log-likelihood $\ell^{(\mathrm{TS})}$ is given by

$$\ell^{(\mathrm{TS})}(\theta|x,a,r) = \eta(g(\theta,x,a)-r)^2\;.$$

Then, at each iteration $t\in[T]$, TS considers the policy π_t determined, for any x, by

$$\pi_t^{(\mathrm{TS})}(x) = a^{\theta_t}(x) \tag{2.5}$$

where $a^\theta(x) = \arg\max_a g(\theta,x,a)$. Here θ_t is a sample from the posterior distribution $\mu_t^{(\mathrm{TS})}(\theta|\mathrm{D}_{t-1}) \propto \mathrm{L}_t^{(\mathrm{TS})}(\theta|\mathrm{D}_{t-1})\mathrm{p}_0(\theta)$, where p_0 is the prior on θ. However, as mentioned in (Zhang 2021), the classic TS algorithm may yield to sub-optimal cumulative regret. They described a simple example where the cumulative regret defined in (2.3) is linear ($\mathcal{O}(T)$), which is sub-optimal compared to the regret bound of $\mathcal{O}(\sqrt{T\log T})$ achieved in (Foster and Rakhlin 2020) for UCB models. This behavior comes from the choice of Gaussians as the model, which leads to sub-exploration of the action space.

To overcome this difficulty, (Zhang 2021) proposes a new model where the classic negative log-likelihood is replaced by the Feel-Good negative log-likelihood, defined by

$$\ell^{(\mathrm{FG})}(\theta|x,a,r) = \eta(g(\theta,x,a)-r)^2 - \lambda\min(b,g_\star(\theta,x))\;,$$

where λ, η and b are hyperparameters in $\mathbb{R}_+$ and $g_\star(\theta,x) = \max_{a\in\mathcal{A}(x)} g(\theta,x,a)$. Then the Feel-Good Thompson sampling algorithm analysed in (Zhang 2021) considers the resulting sequence of likelihoods $\{\mathrm{L}_t^{(\mathrm{FG})}\}_{t\le T}$ and sequence of posteriors $\{\mu_t^{(\mathrm{FG})}\}_{t\le T}$ defined similarly to the classic TS method, and defines the sequence of policies $\{\pi_t^{(\mathrm{FG})}\}_{t\le T}$ as in (2.5) where this time θ_t is a sample from $\mu_t^{(\mathrm{FG})}(\cdot|\mathrm{D}_{t-1})$.

However, exact sampling from $\mu_t^{(\mathrm{FG})}(\cdot|\mathrm{D}_{t-1})$ is usually not tractable, and MCMC algorithms have to be used in their place. This difficulty is not tackled in (Zhang 2021). Consequently, the main objective and contribution of the present paper to extend the analysis by considering the additional complexity from using approximate samples of the posteriors. More precisely, we consider using gradient-based MCMC schemes to generate these approximate samples. The non-smoothness of the prior definition raises a challenge to this end. While gradient-based MCMC has been developed to sample from such non-smooth densities, they do not enjoy the same theoretical guarantees as smooth densities. For that reason, we propose to consider a smoothed posterior (sFG-TS) with the negative log-likelihood

$$\ell^{(\mathrm{sFG})}(\theta|x,a,r) = \eta(g(\theta,x,a)-r)^2 - \lambda[b-\phi_\varsigma(b-g_\star(\theta,x))]\;,$$

with $\phi_\varsigma(u) = \log(1+\exp(\varsigma u))/\varsigma$ for $u\in\mathbb{R}$ and $\varsigma>0$ is a hyperparameter which controls the regularity of $\ell^{(\mathrm{sFG})}$. Through an application of the Bayes theorem, assuming that the prior distribution p_0 is correctly specified, then the posterior distribution at time $t\le T$ can be defined as

$$\mu_t^{(\mathrm{sFG})}(\theta|\mathrm{D}_{t-1}) \propto \mathrm{e}^{-\sum_{s=1}^{t-1}\ell^{(\mathrm{sFG})}(\theta|x_s,a_s,r_s)}\mathrm{p}_0(\theta)\;. \tag{2.6}$$

For simplicity, we denote $\mu_{t-1}^{(\mathrm{sFG})}(\theta|\mathrm{D}_{t-1})$ by $\mu_{t-1}^{(\mathrm{sFG})}(\theta)$. With this notation, we present the MCMC-sFG-TS method in Algorithm 1. In this algorithm, the choice of the sequence of initial distributions $\{p_{t,0}\}_{t\ge T}$ and the sequence of Markov kernels $\{\mathrm{K}_t\}_{t\le T}$ is left arbitrary. Indeed, we first extend the analysis provided in (Zhang 2021) to this setting and derive general bounds depending on quantities related to the convergence of Markov chains with Markov kernels $\{\mathrm{K}_t\}_{t\le T}$ and initialized with $\{p_{t,0}\}_{t\ge T}$. We then illustrate our results by considering two examples of MCMC algorithms in particular, which we provide below.

(1) Langevin Monte Carlo: For a fixed step $t \in [T]$, given the target $\mu_t^{(\mathrm{sFG})}$ and an initial distribution $p_{t,0}$, Langevin Monte Carlo (LMC) follows the Markov chain $\{\theta_{t,k}^{\mathrm{L}}\}_{k=0}^{N_t}$ initialized $\theta_{t,0}^{\mathrm{L}} \sim p_{t,0}$, defined through the recursion:

$$\theta_{t,k+1}^{\mathrm{L}} = \theta_{t,k}^{\mathrm{L}} + \gamma_t \nabla \log \mu_t^{(\mathrm{sFG})}(\theta_{t,k}^{\mathrm{L}}) + \sqrt{2\gamma_t} Z_{t,k} \;, \quad (2.7)$$

where $N_t \in \mathbb{N}^*$ is a number of iterations, γ_t a step size, and $\{Z_{t,k}\}_{k \in [N_t]}$ are i.i.d. samples from the d-dimensional standard Gaussian. It amounts to choosing the Markov kernel $\mathrm{K}_t^{\mathrm{L}}$ with transition density given for $\theta_0, \theta_1 \in \mathbb{R}^d$ by

$$k_t^{\mathrm{L}}(\theta_0, \theta_1) \propto \exp\left[-\|\theta_1 - \theta_0 + \gamma_t \nabla \log \mu_t^{(\mathrm{sFG})}(\theta_0)\|^2 / (4\gamma_t)\right].$$

For better rates (Durmus, Majewski, and Miasojedow 2019), in our analysis we consider the final parameter to be the ergodic average after some burn-in time, i.e. $\theta_t^{\mathrm{L}} = 2/N_t \sum_{k=N_t/2}^{N_t} \theta_{t,k}$ for some even N_t.

LMC is the Euler discretization of the overdamped Langevin diffusion (Roberts and Tweedie 1996) and is a popular way to sample approximately from a smooth positive target density. The Langevin diffusion is a Markov process associated with solutions to the stochastic differential equation (SDE) $\mathrm{d}\boldsymbol{\theta}_{t,s} = \nabla \log \mu_t^{(\mathrm{sFG})}(\boldsymbol{\theta}_{t,s})\mathrm{d}s + \sqrt{2}\mathrm{d}\mathbf{B}_s$, where $(\mathbf{B}_s)_{s \geq 0}$ is a d-dimensional standard Brownian motion. However, while $\{\boldsymbol{\theta}_{t,s}\}_{s \geq 0}$ admits $\mu_t^{(\mathrm{sFG})}$ as its stationary distribution, this is not the case for the Markov kernel associated with (2.7). Therefore, LMC comes with a bias which is the same order as the stepsize γ_t under appropriate conditions (Talay and Tubaro 1990; Durmus and Eberle 2021).

(2) Metropolis Adjusted Langevin Algorithm: To correct the discretization bias of the Langevin SDE, a Metropolis filter can be applied at each iteration as suggested for example in (Roberts and Tweedie 1996). This corresponds to the Metropolis Adjusted Langevin Algorithm (MALA). For technical reasons, we study the $1/2$-lazy version of this algorithm, which defines the Markov chain $\{\theta_{t,k}^{\mathrm{M}}\}_{k=0}^{N_t}$ initialized with $\theta_{t,0}^{\mathrm{M}} \sim p_{t,0}$ following the recursion:

- generate a proposal $\tilde{\theta}_{t,k+1}^{\mathrm{M}} \sim \mathrm{K}_t^{\mathrm{L}}(\theta_{t,k}^{\mathrm{M}}, \cdot)$;
- with probability $1/\alpha_t^{\mathrm{M}}(\theta_{t,k}^{\mathrm{M}}, \tilde{\theta}_{t,k+1}^{\mathrm{M}})$ set $\theta_{t,k+1}^{\mathrm{M}} = \tilde{\theta}_{t,k+1}^{\mathrm{M}}$, otherwise set $\theta_{t,k+1}^{\mathrm{M}} = \theta_{t,k}^{\mathrm{M}}$, where

$$\alpha_t^{\mathrm{M}}(\theta_0, \theta_1) = 1 \wedge \frac{\mu_t^{(\mathrm{sFG})}(\theta_1) k_t^{\mathrm{L}}(\theta_1, \theta_0)}{\mu_t^{(\mathrm{sFG})}(\theta_0) k_t^{\mathrm{L}}(\theta_0, \theta_1)} \;.$$

For MALA, we take $\theta_t = \theta_{t,N_t}$ to be the last iterate.

We refer to the resulting methods as LMC-sFG-TS (resp. MALA-sFG-TS) in the sequel.

Related Works Appoximate sampling in TS algorithms is in general based on Laplace approximation (Chapelle and Li 2011), which fits the mean and the covariance matrix of a Gaussian distribution based on the target. This is then used to approximately sample from the posterior. However, high-dimensional Gaussian distribution with general covariance matrices may be expensive to compute. Further, in non-linear models such as generalized linear bandits and neural contextual bandits, the sequence of posteriors may be far from Gaussian distributions and Laplace approximation may fail in capturing their complex properties. Finally, Laplace approximation does not come with any theoretical guarantees on the quality of the resulting approximation.

The use of LMC or Stochastic Gradient Langevin Dynamics in Thompson Sampling for non-contextual bandits has been proposed in (Mazumdar et al. 2020). This idea has been recently been extended to contextual bandits in (Xu, Zheng, et al. 2022), which introduced LMC-TS. Algorithm 1 extends this method in two ways: (1) by considering the more complex likelihood (2.4), (2) taking as an input the MCMC algorithms which are used to sample in sFG-TS. Finally, (Xu, Zheng, et al. 2022) is only applicable for linear bandits, where the TS posteriors are Gaussian distributions. In contrast, we are able to establish very generic bounds for MCMC-sFG-TS by adapting and extending the FG-TS theory in (Zhang 2021). We specify these results in Section 3.3 to the particular instance of linear bandits, when the MCMC method used in MCMC-sFG-TS is LMC or MALA.

In Algorithm 1, the function F selects an appropriate sample θ_t from the MCMC algorithm. Usually, we take $F(\{\theta_{t,k}\}_{k \leq N_t}) = \theta_{t,N_t}$.

Algorithm 1 MCMC-sFG-TS

Initialize:
  $\mathrm{D}_0 = \emptyset$
for $t = 1, \ldots T$ **do**
  receive $x_t \in \mathsf{X}$
  initialize the Markov chain $\theta_{t,0} | \mathrm{D}_{t-1} \sim p_{t,0}$ where $p_{t,0}$ may depend on D_{t-1};
  for $k = 0, \ldots N_t - 1$ **do**
    $\theta_{t,k+1} | \mathrm{D}_{t-1} \sim \mathrm{K}_t(\theta_{t,k}, \cdot)$ where K_t is a Markov kernel which targets $\mu_t^{(\mathrm{sFG})}(\cdot | \mathrm{D}_{t-1})$, e.g., LMC or MALA
  end for
  choose $\theta_t = F(\{\theta_{t,k}\}_{k \leq N_t})$
  choose $a_t = \arg\max_{a \in \mathcal{A}(x_t)} g(\theta_t, x_t, a)$
  receive the reward $r_t \sim \mathrm{R}(\cdot | x_t, a_t)$
end for

3 Main results

3.1 Analysis of MCMC-sFG-TS

We make these assumptions on the reward distribution.

H 1. *(Sub-Gaussian Reward Distribution) There exists* $c >$

0 *such that for any* $x \in \mathsf{X}$, $a \in \mathcal{A}(x)$, $\rho > 0$,

$$\log \int \exp\{\rho(r - f(x,a))\}\mathrm{R}(\mathrm{d}r|x,a) \le c\rho^2 \ ,$$

where f *is defined in* (2.1). *Furthermore, assume* $\sup_{x\in\mathsf{X}\,,\,a\in\mathcal{A}(x)} |f(x,a)| \le b_f$.

Note that Assumption 1 is automatically satisfied if the rewards are bounded almost surely, i.e, for any x and a, $\mathrm{R}(\cdot|x,a)$ has a bounded support.

We state our main result regarding the cumulative regret for MCMC-FG-TS. First recall that we have assumed a finite action set A and therefore we can define $K = \max_{x\in\mathsf{X}} \mathrm{Card}(\mathcal{A}(x))$. Second, we denote by $\hat{\mu}_t^{(\mathrm{sFG})}$ the distribution of θ_t given D_{t-1}, as defined in Algorithm 1, and define for $t \in [T]$, $\delta_t = ||\hat{\mu}_t^{(\mathrm{sFG})} - \mu_t^{(\mathrm{sFG})}||_{\mathrm{TV}}$. Note that the sequence $(x_t, a_t, r_t, \theta_t)_{t=0}^T$, defined in Algorithm 1, is a Markov chain, possibly inhomogeneous, and we define by $\mathbb{E}_{\nu_0}^T$ and $\mathbb{P}_{\nu_0}^T$ the canonical expectation and probability respectively associated with this process and with initial distribution ν_0. Define the filtration $(\mathcal{F}_t)_{t\in[T]}$ by $\mathcal{F}_t = \sigma\{\{x_s, a_s, r_s\}_{s\in[t]}\}$. With this notation, the cumulative regret associated with the distribution $\mathbb{Q}_{1:T}^{(\mathrm{sFG})}$ defined by Algorithm 1 can be written as $\mathrm{CREG}(\mathbb{Q}_{1:T}^{(\mathrm{sFG})}) = \mathbb{E}_{\nu_0}^T[\sum_{s\le T} f_\star(x_s) - r_s]$.

Theorem 1. *Assume that H 1 holds and let* $\varsigma > 0$. *If* η *is chosen according to* (A.3) *with* $\epsilon \in]0,1[$, *then there exists* C_1, C_1, C_2 *and* C_3, *independent of* $\epsilon, \eta, \lambda, d, T, K$ *such that*

$$\mathrm{CREG}(\mathbb{Q}_{1:T}^{(\mathrm{sFG})}) \le \frac{\lambda}{\eta\epsilon} KT + \mathrm{C}_1 \lambda T - \frac{Z_T}{\lambda} + (\mathrm{C}_2 + \frac{\mathrm{C}_3}{\lambda}) \sum_{t=0}^{T} \mathbb{E}_{\nu_0}^T[\delta_t] \ ,$$

where

$$Z_T = \mathbb{E}_{\nu_0}^T \log \int \exp\Big(-\textstyle\sum_{s=1}^T \Delta\ell^{(\mathrm{sFG})}(\tilde{\theta}, x_s, a_s, r_s)\Big)\mathrm{dp}_0(\tilde{\theta}) \ ,$$

and

$$\Delta\ell^{(\mathrm{sFG})}(\theta,x,a) = \eta\big\{(g(\theta,x,a)-r)^2 - (f(x,a)-r)^2\big\} - \lambda\{b - \phi_\varsigma(b - g_\star(\theta,x)) - f_\star(x)\} \ . \quad (3.1)$$

Proof. We provide here the main steps leading to Theorem 1 based on Lemmas which are stated and proved in Section A.1 of the supplement.

(A) **Regret decomposition.** The first step of the proof is to decompose the expected regret at time s into two terms as follows

$$\mathbb{E}_{\nu_0}^T[\mathrm{REG}_s^{\pi_s}] = \mathbb{E}_{\nu}^T[\mathrm{B}_{x_s}(\theta_s, a^{\theta_s}(x_s))] - \mathbb{E}_{\nu}^T[\mathrm{FG}_{x_s}(\theta_s, a^{\theta_s}(x_s)] \quad (3.2)$$

where

$$\mathrm{B}_x : (\theta,a) \to g_b(\theta,x,a) - f(x,a) \ ,$$
$$g_b(\theta,x) = \max\{-b,\ \min(b,\ g_\star(\theta,x))\} \ ,$$
$$\mathrm{FG}_x \ : \ (\theta,a) \mapsto g_b(\theta,x,a) - f_*(x) \ .$$

On the right hand side, the first term is referred to as the Bellman error in the reinforcement learning literature (Bellman 1966), and the second one as the Feel-Good exploration term. The proof of the decomposition is provided in Lemma 6.

(B) **Bellman error.** By using Lemma 7 we can bound the Bellman error by

$$\mathbb{E}_{\nu_0}^T[\mathrm{B}_{x_s}(\theta_s, a^{\theta_s}(x_s))|x_s, \mathcal{F}_{s-1}] \le \inf_{\gamma>0}\left(\frac{K}{4\gamma} + \gamma\, \mathbb{E}_{\nu_0}^T[\psi(x_s, a^{\theta_s}(x_s))|x_s, \mathcal{F}_{s-1}]\right) ,$$

where $\psi(x_s, a) = \mathbb{E}_{\nu_0}^T[\mathrm{LS}_{x_s}^b(\theta_s, a)|x_s, \mathcal{F}_{s-1}]$, and

$$\mathrm{LS}_x^b \ : \ (\theta,a) \mapsto (g_b(\theta,x,a) - f(x,a))^2 \ . \quad (3.3)$$

This step allows us to decouple the contribution of the random parameter θ_s and its associated action $a^{\theta_s}(x_s)$ to the Bellman error. In the right hand side, we first take the expectation with respect to the parameter for a fixed action, and then with respect to the random action $a^{\theta_s}(x_s)$. This inequality holds for any $\gamma > 0$, in particular for $\gamma = 2C_\eta/(3\lambda)$, with

$$C_\eta = 1.5\eta(1-4c\eta)[1 \ - 0.75\eta(1-4c\eta)(b+b_f)^2] \ , \quad (3.4)$$

where c is the sub-Gaussian coefficient and b_f is the supremum of the true reward function, both defined in H 1. Lemma 12 shows that $2C_\eta/(3\lambda)$ is strictly positive. Hence, the Bellman error bound becomes

$$\mathbb{E}_{\nu_0}^T[\mathrm{B}_{x_s}(\theta_s, a^{\theta_s}(x_s))|x_s, \mathcal{F}_{s-1}] \le \frac{3K\lambda}{8C_\eta} + \frac{2C_\eta}{3\lambda}\, \mathbb{E}_{\nu_0}^T[\psi(x_s, a^{\theta_s}(x_s))|x_s, \mathcal{F}_{s-1}] \ . \quad (3.5)$$

In the next step of the proof, we focus on bounding the resulting error $\mathbb{E}_{\nu_0}^T[\psi(x_s, a_s^{\theta_s}(x_s))|x_s, \mathcal{F}_{s-1}]$. More precisely, given D_{s-1}, $x \in \mathsf{X}$, $a \in \mathcal{A}(x)$, Lemma 8 with $\tau = 3\eta(1-4c\eta)/2$ (which is positive according to Lemma 12) gives

$$C_\eta\, \mathbb{E}_{\theta\sim\hat{\mu}_s^{(\mathrm{sFG})}}[\mathrm{LS}_x^b(\theta,a)] \le -\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathrm{e}^{-3\eta(1-4c\eta)\mathrm{LS}_x(\theta,a)/2}] + C_\eta(b+b_f)^2\delta_s \ , \quad (3.6)$$

where LS_x is defined in (3.3), and

$$\mathrm{LS}_x \ : \ (\theta,a) \mapsto (g(\theta,x,a) - f(x,a))^2 \ .$$

Next, we will focus on the second term in the regret decomposition (3.2), the Feel-Good exploration term.

(C) **Feel Good exploration term.** Similarly, given D_{s-1}, for any $x \in \mathsf{X}$, Lemma 9 with $\tau = 3\lambda$ gives

$$\begin{aligned} &-\mathbb{E}_{\theta\sim\hat{\mu}_s^{(\mathrm{sFG})}}[\mathrm{FG}_x(\theta, a^\theta(x))] \\ &\quad\le -\frac{1}{3\lambda}\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathrm{e}^{3\lambda \mathrm{FG}_x(\theta,a^\theta(x))}] \\ &\qquad + \frac{3\lambda(b+b_f)^2}{2} + (b+b_f)\delta_s \;. \end{aligned} \tag{3.7}$$

Now, the Bellman error bound (3.5) and the Feel-Good bound (3.7) can be merged.

(D) **Combining the bounds.** The combination of (3.6) and (3.7) gives

$$\begin{aligned} &\mathbb{E}_{\theta\sim\hat{\mu}_s^{(\mathrm{sFG})}}\left[\frac{2C_\eta}{3\lambda}\mathrm{LS}_x^b(\theta,a) - \mathrm{FG}_x(\theta,a^\theta(x))\right] \\ &\le -\frac{2}{3\lambda}\log \mathbb{E}_{\theta\sim\hat{\mu}_s^{(\mathrm{sFG})}}[\mathrm{e}^{-3\eta(1-4c\eta)\mathrm{LS}_x(\theta,a)/2}] \\ &- \frac{1}{3\lambda}\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathrm{e}^{3\lambda\mathrm{FG}_x(\theta,a^\theta(x))}] \\ &+ \left[\frac{2C_\eta(b+b_f)^2}{3\lambda} + (b+b_f)\right]\delta_s + \frac{3\lambda}{2}(b+b_f)^2 \;. \end{aligned}$$

Moreover, given D_{s-1}, we can use Lemma 10 to get for any $x \in \mathsf{X}$ and $a \in \mathcal{A}(x)$,

$$\begin{aligned} &\mathbb{E}_{\theta\sim\hat{\mu}_s^{(\mathrm{sFG})}}\left[\frac{2C_\eta}{3\lambda}\mathrm{LS}_x^b(\theta,a) - \mathrm{FG}_x(\theta,a^\theta(x))\right] \\ &\quad\le -\frac{1}{\lambda}\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\Gamma(a,x)] \\ &\qquad + \left[\frac{2C_\eta(b+b_f)^2}{3\lambda} + (b+b_f)\right]\delta_s \\ &\qquad + \frac{3\lambda}{2}(b+b_f)^2 \;, \end{aligned} \tag{3.8}$$

setting $\Gamma(a,x) = \mathbb{E}_{r\sim\mathrm{R}(\cdot|x,a)}[\mathrm{e}^{-\Delta\ell^{(\mathrm{sFG})}(\theta,x,a,r)}]$. We now have all tools to bound the cumulative regret and conclude the proof.

(E) **Cumulative Regret Bound.** Using the regret decomposition (3.2) and the Bellman error bound (3.5), we have

$$\begin{aligned} &\mathbb{E}_{\nu_0}^T[\mathrm{REG}_s^{\pi_s}] \\ &\le \frac{3K\lambda}{8C_\eta} + \frac{2C_\eta}{3\lambda}\mathbb{E}_{\nu_0}^T\left[\mathbb{E}_{\theta\sim\hat{\mu}_s^{(\mathrm{sFG})}}[\mathrm{LS}_{x_s}^b(\theta,a_s)]\right] \\ &\quad - \mathbb{E}_{\nu_0}^T\left[\mathbb{E}_{\theta\sim\hat{\mu}_s^{(\mathrm{sFG})}}[\mathrm{FG}_{x_s}(\theta,a^\theta(x_s))]\right] . \end{aligned}$$

Then Eq. (3.8) gives

$$\begin{aligned} &\mathbb{E}_{\nu_0}^T[\mathrm{REG}_s^{\pi_s}] \\ &\le \frac{3\lambda K}{8C_\eta} - \frac{1}{\lambda}\mathbb{E}_{\nu_0}^T\left[\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\Gamma(a_s,x_s)]\right] \\ &\quad + \left[\frac{2C_\eta(b+b_f)^2}{3\lambda} + (b+b_f)\right]\mathbb{E}_{\nu_0}^T[\delta_s] \\ &\quad + \frac{3\lambda}{2}(b+b_f)^2 \;. \end{aligned}$$

Finally, we can use Lemma 11 to get,

$$Z_t - Z_{t-1} \le \mathbb{E}_{\nu_0}^T\left[\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\Gamma(a_s,x_s)]\right] .$$

We conclude the proof by summing over t to get,

$$\begin{aligned} &\mathrm{CREG}(\mathbb{Q}_{1:T}) = \sum_{s\le T}\mathbb{E}_{\nu_0}^T[\mathrm{REG}_s^{\pi_s}] \\ &\le \left[\frac{3\lambda K}{8C_\eta} + \frac{3\lambda}{2}(b+b_f)^2\right]T - \frac{Z_T}{\lambda} \\ &\quad + \left[\frac{2C_\eta(b+b_f)^2}{3\lambda} + (b+b_f)\right]\sum_{s\le T}\mathbb{E}_{\nu_0}^T[\delta_s] \\ &\le \frac{\lambda KT}{\epsilon\eta} + \mathrm{C}_1\lambda T + (\mathrm{C}_2 + \frac{\mathrm{C}_3}{\lambda})\sum_{s\le T}\mathbb{E}_{\eta_0}^T[\delta_s] - \frac{Z_T}{\lambda} \;, \end{aligned}$$

where $\mathrm{C}_1 = 3(b+b_f)^2/2$, $\mathrm{C}_2 = (b+b_f)$ and $\mathrm{C}_3 = (b+b_f)^2/4$, these constants do not depend neither on η nor in λ. The last inequality uses Lemmas 13-14.

3.2 Regret Bounds for Bandits

We now specify the bounds provided by Theorem 1 assuming the following condition on the prior distribution p_0 and the family of models $\{(x,a) \mapsto g(\theta,x,a) \,:\, \theta \in \mathbb{R}^d\}$.

H 2. *Assume that* $\log \mathrm{p}_0$ *is continuously differentiable,* L_0-*smooth and* m_0-*strongly concave for some* $L_0 \ge m_0 \ge 0$. *This implies that the following holds for all* $\theta_1, \theta_2 \in \mathbb{R}^d$*:*

$$\|\nabla \log \mathrm{p}_0(\theta_2) - \nabla \log \mathrm{p}_0(\theta_1)\| \le L_0\|\theta_1 - \theta_2\|$$

$$\langle \nabla \log \mathrm{p}_0(\theta_2) - \nabla \log \mathrm{p}_0(\theta_1), \theta_1 - \theta_2\rangle \ge \frac{m_0}{2}\|\theta_1 - \theta_2\|^2 \;.$$

In addition, we assume that the family of models $\{(x,a) \mapsto g(\theta,x,a) \,:\, \theta \in \mathbb{R}^d\}$ is regular enough and close to the true model, in the following senses.

H 3. *(Uniform Smoothness) Suppose that for all* $\theta_1, \theta_2 \in \mathbb{R}^d, x \in \mathsf{X}, a \in \mathcal{A}(x)$*, the following bound holds for some* $L_g \in \mathbb{R}_+$*:*

$$|g(\theta_1,x,a) - g(\theta_2,x,a)| \le L_g\|\theta_1 - \theta_2\| \;.$$

H 4. *(Well Specified Model) Suppose that there exist* $\theta_* \in \mathbb{R}^d$ *and* $\xi \in \mathbb{R}_+$ *such that for all* $x \in \mathsf{X}, a \in \mathcal{A}(x)$*:*

$$|g(\theta_*,x,a) - f(x,a)| \le \xi \;.$$

Corollary 2. *Let Assumptions H 1-4 hold. For* ω, η, λ *specified in* (A.4)*, and* T *large enough (specified in* (A.6)*), and for constants* $\mathrm{C}_4, \mathrm{C}_5, \mathrm{C}_6$ *not dependent on* $\omega, \epsilon, d, K, T$

$$\begin{aligned} &\mathrm{CREG}(\mathbb{Q}_{1:T}^{(\mathrm{sFG})}) \\ &\le \frac{\mathrm{C}_4}{\epsilon}\sqrt{\omega dKT\log(dT)} + \left(4\xi + \phi_\varsigma(\frac{L_g}{T} + \xi + b_f - b)\right)T \\ &\quad + \mathrm{C}_5\sqrt{\frac{\omega KT}{d\log(dT)}}\left(-\log \mathrm{p}_0(\theta_*) + L_g + \xi T + \xi^2 T\right) \\ &\quad + \mathrm{C}_6\left(1 + \sqrt{\frac{\omega KT}{d\log(dT)}}\right)\sum_{t=0}^{T}\mathbb{E}_{\nu_0}^T[\delta_t] + 4L_g. \end{aligned}$$

Here θ_* is the parameter in Assumption 4.

The proof of this result along with explicit bounds are given in Section A.2 of the supplement.

3.3 Linear Bandits

A concrete example where H 2-4 hold is the linear contextual bandits framework:

Example 3. (Linear Gaussian Function Class) Consider the function class with $f(x,a) = \langle \varphi(x,a), \theta_* \rangle$, with $\theta_* \in \mathbb{R}^d$ and $x \in \mathsf{X}, a \in \mathcal{A}(x)$, with $\varphi : \mathsf{X} \times \mathsf{A} \to \mathbb{R}^d$ being some feature map. Let the reward be absolutely bounded by some constant b_r almost surely, and let $\sup_{(x,a)\in \mathsf{X}\times\mathsf{A}(x)} \|\varphi(x,a)\| \leq \sqrt{M}$ with $0 < M < \infty$. Finally, let $|\mathcal{A}(x)| \leq d$ for all $x \in \mathsf{X}$.

Remark: The absolute bound on the reward is only needed to guarantee the almost sure complexity bounds on the gradient descent step.

We now define an appropriate notion of complexity, which is different from the typical definition seen in bandit literature.

Definition 4. *(Data Complexity) The agent has access to both the value $g(\theta,x,a)$ and the gradient $\nabla g(\theta,x,a)$ for any $\theta \in \mathbb{R}^d, x \in \mathsf{X}, a \in \mathcal{A}(x)$. Then, if g is evaluated $\mathfrak{a}_t$ times and ∇g is evaluated $\mathfrak{b}_t$ times at any timestep t, then we define $G_t = \mathfrak{a}_t + \mathfrak{b}_t$ as the **data complexity** at time t, and $\mathrm{CG} = \sum_{t\leq T} G_t$ be the **cumulative data complexity**.*

Theorem 5. *Consider Example 3 with the linear function class $g(\theta,x,a) = \langle \varphi(x,a), \theta \rangle$ and a Gaussian prior $\mathrm{N}(0, \mathtt{m}_0^{-1}\mathrm{I}_d)$, with $\mathtt{m}_0 > 0$. Assume H 1 holds, let $\varsigma, \omega_{\mathsf{LG}}, \lambda, \eta, b$ be as specified in (A.7), and let T be large enough (specified in (A.8)). Assume in addition let there exist $\kappa > 0$ such that almost surely, for any $t \in [T]$ the Hessian matrix of $-\log \mu_t^{(\mathrm{sFG})}(\theta)$ (2.6) satisfies for some $\mathtt{m}_t, \mathtt{L}_t > 0$:*

$$\mathtt{L}_t \mathrm{I}_d \succeq -\nabla^2 \log \mu_t^{(\mathrm{sFG})}(\theta) \succeq \mathtt{m}_t \mathrm{I}_d \quad , \quad \mathtt{L}_t/\mathtt{m}_t \leq \kappa \, .$$

(a) Then, starting from an initial point $\hat{\theta}_0^* = \theta_0$, we can find at each round recursively $\hat{\theta}_t^*$ satisfying $\|\hat{\theta}_t^* - \theta_t^*\| \leq \sqrt{d/(2\mathtt{L}_t)}$ using the gradient descent algorithm to maximize $\log \mu_t^{(\mathrm{sFG})}(\theta)$ and initialized with $\hat{\theta}_{t-1}^*$. Here θ_t^* is the maximizer of $\log \mu_t^{(\mathrm{sFG})}(\theta)$. The cumulative data complexity of this procedure is of order $\mathrm{C}_{\mathrm{GD}}\kappa T^2 \log(b_r \mathtt{L}_t \sqrt{M}T/\mathtt{m}_0)$, for some absolute constant C_{GD}, and the step size is $2/(\mathtt{L}_t + \mathtt{m}_t)$.

(b) In addition setting $p_{t,0} = \mathrm{N}(\hat{\theta}_t^*, (\mathtt{L}_t)^{-1}\mathrm{I}_d)$, for any of the following standard choices of Markov kernel, we attain the regret bound for some constant C_7 not dependent on $\omega_{\mathsf{LG}}, \epsilon, d, K, T, M$

$$\mathrm{CREG}(\mathbb{Q}_{1:T}^{(\mathrm{sFG})}) \leq \mathrm{C}_7 \sqrt{\omega_{\mathsf{LG}} T \log^3(dT)} \Big(d(\epsilon \wedge \mathtt{m}_0)^{-1} + \sqrt{M} + \mathtt{m}_0 \|\theta_*\|^2 \Big) \, ,$$

with the number of oracle calls stated below:

- *K^{L} (Langevin Monte Carlo): has $\mathrm{CG}^{\mathsf{LMC}} \leq \mathrm{C}_{\mathrm{L}}\mathrm{C}_\kappa d T^4 \log(4\sqrt{d\kappa}/\mathtt{m}_0)$ cumulative data complexity, with step-size $\gamma_t^{\mathrm{L}} = \mathrm{A}_{\mathrm{L}}/(\max(\kappa, \mathtt{L}_t) d T^2)$, $\mathrm{C}_\kappa = \max(\mathtt{L}_T/\mathtt{m}_0^2, \mathtt{L}_T)$.*
- *K^{M} (Metropolis Adjusted Langevin Monte Carlo): has $\mathrm{CG}^{\mathsf{MALA}} \leq \mathrm{C}_{\mathrm{M}}\kappa d T^2 (1 \vee \sqrt{\kappa d^{-1}}) \log(dT^2)$ cumulative data complexity, with step-size $\gamma_t^{\mathrm{M}} = \mathrm{A}_{\mathrm{M}}/(\mathtt{L}_t d \max(1, \sqrt{\kappa d^{-1}}))$*

Here C_{L}, C_{M}, $\mathrm{A}_{\mathrm{L}}, \mathrm{A}_{\mathrm{M}}$ are absolute constants depending on which MCMC algorithm was chosen.

The proof of this result along with explicit bounds are given in Section A.3 of the supplement.

Remarks: We note that the Gaussian prior can be replaced with an arbitrary prior satisfying Assumption 2, so long as a good bound on $\mathrm{p}_0(\theta_*)$ exists. The Lipschitz constant can be bounded by $\mathtt{L}_t \leq 2(\mathtt{m}_0 + t\lambda\varsigma\sqrt{M} + t\eta\sqrt{M})$.

We can compare Theorem 5 with (Xu, Zheng, et al. 2022, Theorem 4.2). (Xu, Zheng, et al. 2022, Theorem 4.2) has a bound on the cumulative data complexity for LMC-TS of order κT^2, which is used to obtain a cumulative regret of order $d^{3/2}T^{1/2}$. In contrast, for our results under MALA, we pay an extra factor of d in the cumulative data complexity in order to remove the suboptimal factor of $d^{1/2}$ in the resulting cumulative regret. We see this increased complexity as a necessary cost in order to obtain our tighter regret bounds. It may be possible to more finely balance this trade-off by e.g. annealing the Feel-Good parameter, but we defer this investigation to subsequent work.

4 Experiments

In this section, we illustrate the benefits of our methodology on several contextual bandit benchmarks associated with both synthetic and real data. In our comparisons, we first perform grid searches for the hyperparameters, and then fix the best ones. Additional details about experimental design are provided in Section B of the supplement.

4.1 Toy example

We first illustrate our approaches on a synthetic contextual bandit problem. At each round $t \in [T]$, the agent observes a contextual vector sampled from a 4 dimensional Gaussian distribution, i.e., $x_t \sim \mathrm{N}(0_4, \mathrm{I}_4)$. Then, the agent has to choose an action a_t between $K = 5$ arms, and finally, receives a reward $r_t = \varphi(x_t, a_t)^\top \theta^\star + \epsilon$ where $\epsilon \sim \mathrm{N}(0, \sigma^2)$, $\theta^\star \in \mathbb{R}^{20}$ is the true parameter of the model, σ is the noise level of the problem. Here φ allows us to transform the context vector and the arm index into a vector v such as, $\varphi(x, 0) = (x, 0, \dots, 0)$, $\varphi(x, 1) = (0, x, 0, \dots)$ and $\varphi(x, d-1) = (0, \dots, 0, x)$. We consider the corresponding model defined as $g(\theta, x, a) = \varphi(x, a)^\top \theta$. Under these settings, note that posterior distributions associated with TS are Gaussian distributions and are therefore tractable.

In Figure 1, we compare our methodology MCMC-sFG-TS using LMC and MALA with Linear TS, along with LMC-TS. For completeness, we also consider TS where at each iteration, we approximate the TS posterior (2.4) with MALA. This simply corresponds to MCMC-sFG-TS but choosing $\lambda = 0$. For these results, we only display the best combination of hyperparameters for each algorithm. More details for the experiment settings are provided in Section B. Note that for MALA-sFG-TS and MALA-TS, we initialize MALA with the output of a gradient descent scheme using full-batch gradient. Moreover, we also consider Linear UCB for which results can be found in Section B in the supplement. We observe that adding the Feel-Good framework allow us to converge to a better regret. Similarly, approximating the posterior using MALA seems to improve the algorithmic performance by converging faster to the target. Finally, by combining the Feel-Good adjustment with MALA, we obtain MALA-sFG-TS which provides the best cumulative regret.

Similar conclusions are drawn on different bandit settings, including logistic and quadratic bandits trained with benchmark algorithms; see Section B in the supplementary.

4.2 Real-World dataset

In this subsection, we compare the algorithms on the Yahoo! Front Page Today Module dataset, which is a standard benchmark for contextual bandits (Li, Chu, et al. 2010; Mellor and Shapiro 2013; Liu, Lee, and Shroff 2018). This seeks to model a user's interest in a specified news article using the contextual bandit framework. At each round, we consider a user and a pool of articles. Here, the context is composed by a user-features vector and user-article interaction information. In addition, the set of arms is the pool of articles. Then, given a current bandit model, we choose an article and check if it is clicked. If so, a reward of 1 is incurred; otherwise, the reward is 0. With this definition and our bandit formulation, we seek here to maximize the average expected cumulative reward $T^{-1}\mathbb{E}_{\Pi \sim \mathbb{Q}_{1:T}}[\sum_{t=1}^{T} f(x_t, \pi_s(x_t))]$, which is precisely the click-through rate (CTR) in (Li, Chu, et al. 2010). A more detailed description on the implementation can be found in (Li, Chu, et al. 2010). In our experiments, we consider just a subset of 500 thousand recommendations made the 3^{th} of May 2009, with the statistics reported over

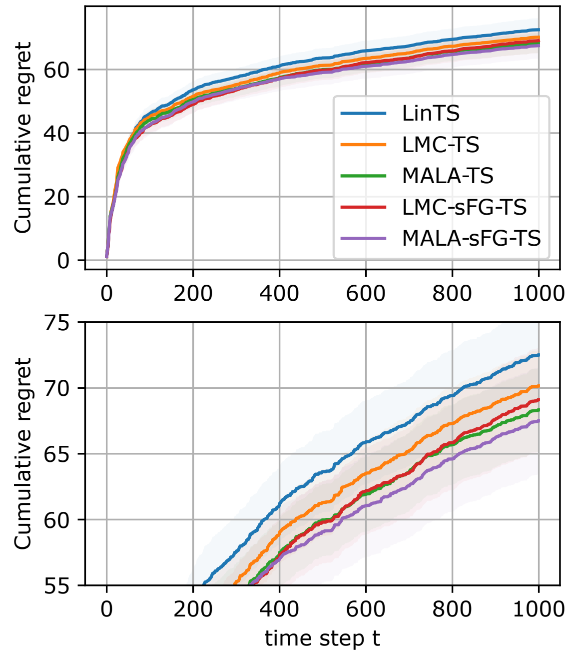

Figure 1: Cumulative regret for the toy example. Whole curve (Top) and its zoomed version (Bottom) are represented. Statistics are reported over 50 runs.

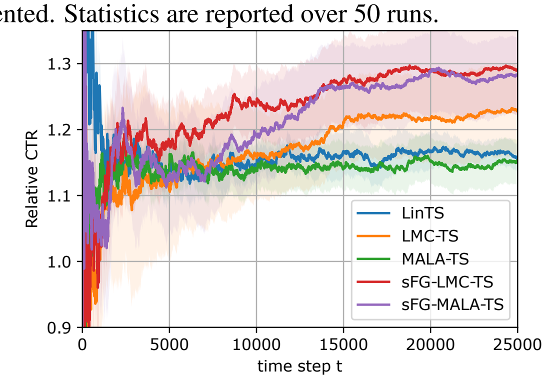

Figure 2: Relative CTR for the Yahoo recommendation task

10 trials. For each run the dataset is shuffled.

In Figure 2 we compare the different approaches using their relative CTR, which is the algorithm's CTR divided by that of a baseline random policy. It can be seen that LMC-sFG-TS and MALA-sFG-TS deliver the best recommendations amongst their competitors.

5 Conclusion

In this work we proposed and analyzed the MCMC-sFG-TS algorithm for contextual bandits, which is a tractable implementation of Thompson sampling with an optimistic Feel-Good adjustment term. We showed that this obtains the optimal regret bound of $\tilde{\mathcal{O}}(d\sqrt{T})$ in high dimensions, in contrast to the $\tilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ that was previously known for MCMC algorithms in the Thompson sampling setting. We also validated the superior performance of this algorithm in practice, relative to the standard Thompson sampling.

Further extensions to our approach include non-quadratic log-likelihoods, which would extend our results to classes such as logistic bandits and bandits with generalized linear models. Finally, applying our framework to some classes of reinforcement learning problems would be an important step towards a general understanding of Thompson sampling algorithms in that setting.

Acknowledgements

Part of this research has been carried out under the auspice of the Lagrange Center for Mathematics and Computing.

References

Abbasi-Yadkori, Y., D. Pál, and C. Szepesvári (2011). “Improved Algorithms for Linear Stochastic Bandits”. In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor et al. Vol. 24. Curran Associates, Inc.

Agarwal, A., S. Bird, et al. (2016). *Making Contextual Decisions with Low Technical Debt*.

Agarwal, A., M. Dudik, et al. (2012). “Contextual bandit learning with predictable rewards”. In: *Artificial Intelligence and Statistics*. PMLR, pp. 19–26.

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

– (17–19 Jun 2013). “Thompson Sampling for Contextual Bandits with Linear Payoffs”. In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by S. Dasgupta and D. McAllester. Vol. 28. Proceedings of Machine Learning Research 3. Atlanta, Georgia, USA: PMLR, pp. 127–135.

Auer, P., N. Cesa-Bianchi, and P. Fischer (2002). “Finite-time analysis of the multiarmed bandit problem”. In: *Machine learning* 47.2, pp. 235–256.

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

Bellman, R. (1966). “Dynamic programming”. In: *Science* 153.3731, pp. 34–37.

Berry, D. A. and B. Fristedt (1985). “Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability)”. In: *London: Chapman and Hall* 5.71-87, pp. 7–7.

Bouneffouf, D., I. Rish, and C. Aggarwal (2020). “Survey on applications of multi-armed and contextual bandits”. In: *2020 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, pp. 1–8.

Chapelle, O. and L. Li (2011). “An empirical evaluation of thompson sampling”. In: *Advances in neural information processing systems* 24.

Chen, Y. et al. (2020). “Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients.” In: *J. Mach. Learn. Res.* 21, pp. 92–1.

Chu, W. et al. (Nov. 2011). “Contextual Bandits with Linear Payoff Functions”. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Ed. by G. Gordon, D. Dunson, and M. Dudík. Vol. 15. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR, pp. 208–214.

Dani, V., T. P. Hayes, and S. M. Kakade (2008). “Stochastic Linear Optimization under Bandit Feedback”. In: *COLT*.

Durmus, A. and A. Eberle (2021). “Asymptotic bias of inexact Markov Chain Monte Carlo methods in high dimension”. In: *arXiv preprint arXiv:2108.00682*.

Durmus, A., S. Majewski, and B. Miasojedow (2019). “Analysis of Langevin Monte Carlo via convex optimization”. In: *The Journal of Machine Learning Research* 20.1, pp. 2666–2711.

Durmus, A., É. Moulines, and M. Pereyra (2018). “Efficient Bayesian Computation by Proximal Markov Chain Monte Carlo: When Langevin Meets Moreau”. In: *SIAM Journal on Imaging Sciences* 11.1, pp. 473–506.

Dwivedi, R. et al. (2018). “Log-concave sampling: Metropolis-Hastings algorithms are fast!” In: *Conference on learning theory*. PMLR, pp. 793–797.

Filippi, S. et al. (2010). “Parametric Bandits: The Generalized Linear Case”. In: *Advances in Neural Information Processing Systems*. Ed. by J. Lafferty et al. Vol. 23. Curran Associates, Inc.

Foster, D. and A. Rakhlin (2020). “Beyond ucb: Optimal and efficient contextual bandits with regression oracles”. In: *International Conference on Machine Learning*. PMLR, pp. 3199–3210.

Jin, T. et al. (18–24 Jul 2021). “MOTS: Minimax Optimal Thompson Sampling”. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 5074–5083.

Katehakis, M. N. and A. F. Veinott (1987). “The Multi-Armed Bandit Problem: Decomposition and Computation”. In: *Math. Oper. Res.* 12, pp. 262–268.

Kaufmann, E., O. Cappé, and A. Garivier (2012). “On Bayesian upper confidence bounds for bandit problems”. In: *Artificial intelligence and statistics*. PMLR, pp. 592–600.

Kveton, B. et al. (26–28 Aug 2020a). “Randomized Exploration in Generalized Linear Bandits”. In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by S. Chiappa and R. Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 2066–2076.

– (2020b). “Randomized exploration in generalized linear bandits”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2066–2076.

Lale, S. et al. (2019). "Stochastic Linear Bandits with Hidden Low Rank Structure". In: *CoRR* abs/1901.09490.

Langford, J. and T. Zhang (2007). "The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information". In: *Advances in Neural Information Processing Systems*. Ed. by J. Platt et al. Vol. 20. Curran Associates, Inc.

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

Li, L., W. Chu, et al. (2010). "A contextual-bandit approach to personalized news article recommendation". In: *Proceedings of the 19th international conference on World wide web*, pp. 661–670.

Li, L., Y. Lu, and D. Zhou (June 2017). "Provably Optimal Algorithms for Generalized Linear Contextual Bandits". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 2071–2080.

Liu, F., J. Lee, and N. Shroff (2018). "A change-detection based framework for piecewise-stationary multi-armed bandit problem". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1.

Mazumdar, E. et al. (13–18 Jul 2020). "On Approximate Thompson Sampling with Langevin Algorithms". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 6797–6807.

Mellor, J. and J. Shapiro (2013). "Thompson sampling in switching environments with Bayesian online change detection". In: *Artificial intelligence and statistics*. PMLR, pp. 442–450.

Ménard, P. and A. Garivier (15–17 Oct 2017). "A minimax and asymptotically optimal algorithm for stochastic bandits". In: *Proceedings of the 28th International Conference on Algorithmic Learning Theory*. Ed. by S. Hanneke and L. Reyzin. Vol. 76. Proceedings of Machine Learning Research. PMLR, pp. 223–237.

Nesterov, Y. et al. (2018). *Lectures on convex optimization*. Vol. 137. Springer.

Riquelme, C., G. Tucker, and J. Snoek (2018). "Deep Bayesian Bandits Showdown: An Empirical Comparison of Bayesian Deep Networks for Thompson Sampling". In: *International Conference on Learning Representations*.

Robbins, H. (1952). "Some aspects of the sequential design of experiments". In: *Bulletin of the American Mathematical Society* 58.5, pp. 527–535.

Roberts, G. O. and R. L. Tweedie (1996). "Exponential convergence of Langevin distributions and their discrete approximations". In: *Bernoulli* 2.4, pp. 341–363.

Russo, D. and B. Van Roy (2014). "Learning to optimize via posterior sampling". In: *Mathematics of Operations Research* 39.4, pp. 1221–1243.

– (2016). "An information-theoretic analysis of thompson sampling". In: *The Journal of Machine Learning Research* 17.1, pp. 2442–2471.

Talay, D. and L. Tubaro (1990). "Expansion of the global error for numerical schemes solving stochastic differential equations". In: *Stochastic analysis and applications* 8.4, pp. 483–509.

Tewari, A. and S. A. Murphy (2017). "From Ads to Interventions: Contextual Bandits in Mobile Health". In: *Mobile Health - Sensors, Analytic Methods, and Applications*.

Thompson, W. R. (1933). "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". In: *Biometrika* 25.3-4, pp. 285–294.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press.

Xu, P., Z. Wen, et al. (2020). "Neural contextual bandits with deep representation and shallow exploration". In: *arXiv preprint arXiv:2012.01780*.

Xu, P., H. Zheng, et al. (2022). "Langevin Monte Carlo for Contextual Bandits". In: *arXiv preprint arXiv:2206.11254*.

Zenati, H. et al. (2022). "Efficient Kernelized UCB for Contextual Bandits". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 5689–5720.

Zhang, T. (2021). "Feel-Good Thompson Sampling for Contextual Bandits and Reinforcement Learning". In: *arXiv preprint arXiv:2110.00871*.

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

A Postponed Proofs

A.1 Proof of Theorem 1

Lemma 6. *(Regret decomposition) The regret at time s can be decomposed into two terms as follows*

$$\begin{aligned}\mathbb{E}_{\nu_0}^T[\mathrm{REG}_s^{\pi_s}] = \mathbb{E}_{\nu_0}^T[g_b(\theta, x_s, a_s^\theta(x_s)) - f_\star(x_s, a_s^\theta(x_s))] \\ - \mathbb{E}_{\nu_0}^T[g_b(\theta, x_s, a_s^\theta(x_s)) - f_\star(x_s)] \ .\end{aligned}$$

Proof. Using the definition of $\mathrm{REG}_s^{\pi_s}$ in (2.2) and the definition of policy π_s in (2.5), we have

$$\begin{aligned}\mathbb{E}_{\nu_0}^T[\mathrm{REG}_s^{\pi_s}] &= \mathbb{E}_{\nu_0}^T[f_\star(x_s) - f(x_s, \pi_s(x_s))] \\ &= \mathbb{E}_{\nu_0}^T[f_\star(x_s) - f(x_s, a_s^\theta(x_s))] \\ &= \mathbb{E}_{\nu_0}^T[g_b(\theta, x_s, a_s^\theta(x_s)) - f(x_s, a_s^\theta(x_s))] - \mathbb{E}_{\nu_0}^T[g_b(\theta, x_s, a_s^\theta(x_s)) - f_\star(x_s)] \ .\end{aligned}$$

□

Lemma 7. *Let $b > 0$. Then, we have the following decoupling bound*

$$\begin{aligned}&\mathbb{E}_{\nu_0}^T[g_b(\theta_s, x_s, a^{\theta_s}(x_s)) - f(x_s, a^{\theta_s}(x_s))|x_s, \mathcal{F}_{s-1}] \\ &\leq \inf_{\gamma>0}\big(K/(4\gamma) + \gamma\,\mathbb{E}_{\nu_0}^T[\psi_{x_s}(a_s^{\theta_s}(x_s))|x_s, \mathcal{F}_{s-1}]\big) \ ,\end{aligned}$$

where $\psi_{x_s}(a) = \mathbb{E}_{\nu_0}^T[\mathrm{LS}_{x_s}^b(\theta_s, a)|x_s, \mathcal{F}_{s-1}]$.

Proof. Note first that

$$\begin{aligned}\mathbb{E}_{\nu_0}^T[g_b(\theta_s, x_s, a^{\theta_s}(x_s)) - f(x_s, a^{\theta_s}(x_s))|x_s, \mathcal{F}_{s-1}] &\leq \mathbb{E}_{\nu_0}^T[|g_b(\theta_s, x_s, a^{\theta_s}(x_s)) - f(x_s, a^{\theta_s}(x_s))||\mathcal{F}_{s-1}, x_s] \\ &= \sum_{a\in\mathcal{A}(x_s)} \mathbb{E}_{\nu_0}^T[\mathbb{1}\{a^{\theta_s}(x_s) = a\}|g_b(\theta_s, x_s, a) - f(x_s, a)||\mathcal{F}_{s-1}, x_s] \ .\end{aligned} \tag{A.1}$$

Consider for any $\tilde{a} \in \mathcal{A}(x_s)$, $\mathbf{q}(\tilde{a}|x_s) = \mathbb{E}_{\nu_0}^T[\mathbb{1}\{a^{\theta_s}(x_s) = \tilde{a}\}|\mathcal{F}_{s-1}, x_s]$. Then for any $\gamma > 0$, we have

$$\begin{aligned}&\mathbb{E}_{\nu_0}[\mathbb{1}\{a^{\theta_s}(x_s) = a\}|g_b(\theta_s, x_s, a) - f(x_s, a)||\mathcal{F}_{s-1}, x_s] \\ &\quad\leq \mathbb{E}_{\nu_0}^T[\frac{\mathbb{1}\{a^{\theta_s}(x_s) = a\}}{4\gamma\mathbf{q}(a|x_s)} + \gamma\mathbf{q}(a|x_s)(g_b(\theta_s, x_s, a) - f(x_s, a))^2|\mathcal{F}_{s-1}, x_s] \\ &\quad= 1/(4\gamma) + \gamma\mathbf{q}(a|x_s)\,\mathbb{E}_{\nu_0}^T[(g_b(\theta_s, x_s, a) - f(x_s, a))^2|\mathcal{F}_{s-1}, x_s]\end{aligned}$$

where the inequality comes from the algebraic inequality $z_1 \cdot z_2 \leq z_1^2/2 + z_2^2/2$ and the last equality from the definition of the distribution $\mathbf{q}$. Plugging the previous inequality in (A.1), and using that for any $x \in \mathsf{X}$, $\mathrm{Card}(\mathcal{A}(x)) \leq K$, then we have

$$\begin{aligned}&\mathbb{E}_{\nu_0}^T[g_b(\theta_s, x_s, a^{\theta_s}(x_s)) - f(x_s, a^{\theta_s}(x_s))|\mathcal{F}_{s-1}, x_s] \\ &\quad\leq K/(4\gamma) + \gamma\sum_{a\in\mathcal{A}(x_s)}\mathbf{q}(a|x_s)\,\mathbb{E}_{\nu_0}^T[(g_b(\theta_s, x_s, a) - f(x_s, a_s))^2|\mathcal{F}_{s-1}, x_s] \\ &\quad= K/(4\gamma) + \gamma\,\mathbb{E}_{\nu_0}^T[\psi(x_s, a_s^{\theta_s}(x_s))|x_s, \mathcal{F}_{s-1}] \ .\end{aligned}$$

□

Lemma 8. *Assume H 1. Given D_{s-1}, for any $x \in \mathsf{X}$, $a \in \mathcal{A}(x)$ and $\tau > 0$, it holds*

$$C_\tau\,\mathbb{E}_{\theta\sim\hat{\mu}_s^{(\mathrm{sFG})}}[\mathrm{LS}_x^b(\theta, a)] \leq -\log\mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\exp\{-\tau\mathrm{LS}_x(\theta, a)\}] + C_\tau(b + b_f)^2\delta_s \ ,$$

where

$$C_\tau = \tau[1 - \tau(b + b_f)^2/2] \ .$$

Proof. Since for any $z \leq 0$, we have $\exp z \leq z^2/2 + z + 1$, we obtain for any $\tau > 0$,

$$\begin{aligned}\mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\exp\{-\tau \mathrm{LS}_x(\theta,a)\}] &\leq \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\exp\{-\tau \mathrm{LS}_x^b(\theta,a)\}] \\ &\leq -\tau\,\mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathrm{LS}_x^b(\theta,a)] + \frac{\tau^2}{2}\mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathrm{LS}_x^b(\theta,a)^2] + 1 \\ &\leq -\tau[1 - \frac{\tau(b+b_f)^2}{2}]\,\mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathrm{LS}_x^b(\theta,a)] + 1 \\ &\leq -C_\tau\,\mathbb{E}_{\theta\sim\hat{\mu}_s^{(\mathrm{sFG})}}[\mathrm{LS}_x^b(\theta,a)] + 1 + C_\tau(b+b_f)^2\delta_s\ ,\end{aligned}$$

where the first inequality uses $\mathrm{LS}_x(\theta,a) \geq \mathrm{LS}_x^b(\theta,a)$, third inequality $\mathrm{LS}_x^b \leq (b+b_f)^2$ and the last inequality the definition of the total variation distance. Moreover, using $\log z \leq z - 1$ for $z \leq 1$, we have,

$$\begin{aligned}\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\exp\{-\tau \mathrm{LS}_x(\theta,a)\}] &\leq \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\exp\{-\tau \mathrm{LS}_x(\theta,a)\}] - 1 \\ &\leq -C_\tau\,\mathbb{E}_{\theta\sim\hat{\mu}_s^{(\mathrm{sFG})}}[\mathrm{LS}_x^b(\theta,a)] + C_\tau(b+b_f)^2\delta_s\ .\end{aligned}$$

□

Lemma 9. *Assume H 1. Given D_{s-1}, for any $x \in \mathsf{X}$, $a \in \mathcal{A}(x)$ and $\tau > 0$, the Feel-Good exploration term is bounded as follows*

$$-\mathbb{E}_{\theta\sim\hat{\mu}_s^{(\mathrm{sFG})}}[\mathrm{FG}_x(\theta,a^\theta(x))] \leq -\frac{1}{\tau}\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\exp(\tau \mathrm{FG}_x(\theta,a^\theta(x)))] + \frac{\tau}{2}(b+b_f)^2 + (b+b_f)\delta_s\ .$$

Proof. Using Hoeffding's lemma since $\mathrm{FG}_x(\theta,a^\theta(x)) \in [-(b+b_f),(b+b_f)]$, for any $\tau > 0$, we have

$$\begin{aligned}\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\exp\{\tau \mathrm{FG}_x(\theta,a^\theta(x))\}] &\leq \tau\,\mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathrm{FG}_x(\theta,a^\theta(x))] + \frac{\tau^2}{2}(b+b_f)^2 \\ &\leq \tau\,\mathbb{E}_{\theta\sim\hat{\mu}_s^{(\mathrm{sFG})}}[\mathrm{FG}_x(\theta,a^\theta(x))] + \frac{\tau^2}{2}(b+b_f)^2 + \tau(b+b_f)\delta_s\ ,\end{aligned}$$

where the second line uses the definition of the total variation distance. □

Lemma 10. *Assume H 1. Given D_{s-1}, for any $x \in \mathsf{X}$, and $a \in \mathcal{A}(x)$,*

$$\begin{aligned}&-\frac{2}{3}\log\Big(\mathbb{E}_{\theta\sim\mu_s(\mathrm{sFG})}[\mathrm{e}^{-3\eta(1-4c\eta)\mathrm{LS}_x(\theta,a)/2}]\Big) - \frac{1}{3}\log\Big(\mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathrm{e}^{3\lambda \mathrm{FG}_x(\theta,a^\theta(x))}]\Big) \\ &\qquad \leq -\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})},\ r\sim\mathrm{R}(\cdot|x,a)}[\mathrm{e}^{-\Delta\ell^{(\mathrm{sFG})}(\theta,x,a,r)}]\ ,\end{aligned}$$

where $\Delta\ell^{(\mathrm{sFG})}(\theta,x,a,r)$ is defined in (3.1).

Proof. Firstly, we can apply the Hölder's inequality with $p = 3/2$ and $q = 3$:

$$\begin{aligned}&\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathrm{e}^{-\eta(1-4c\eta)\mathrm{LS}_x(\theta,a)+\lambda \mathrm{FG}_x(\theta,a^\theta(x))}] \\ &\leq \frac{2}{3}\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathrm{e}^{-3\eta(1-4c\eta)\mathrm{LS}_x(\theta,a)/2}] + \frac{1}{3}\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathrm{e}^{3\lambda \mathrm{FG}_x(\theta,a^\theta(x))}]\ .\end{aligned} \tag{A.2}$$

Subsequently, by Assumption 1 with $\rho = 2\eta(f(x,a) - g(\theta,x,a))$, if we denote $\epsilon = r - f(x,a)$, we find that: $\exists\, c > 0$ such that

$$\begin{aligned}\int \exp\{-2\eta(f(x,a) - g(\theta,x,a))\epsilon\}\mathrm{R}(\mathrm{d}r|x,a) &\leq \exp\{4c\eta^2(f(x,a) - g(\theta,x,a))^2\} \\ &= \exp\{4c\eta^2 \mathrm{LS}_x(\theta,a)\}\ .\end{aligned}$$

Recall the definition of $\Delta\ell^{(\mathrm{sFG})}$ in (3.1). Then,

$$\begin{aligned}-\Delta\ell^{(\mathrm{sFG})}(\theta,x,a,r) &= -\eta(\epsilon + f(x,a) - g(\theta,x,a))^2 + \eta\epsilon^2 + \lambda(b - \phi_\varsigma(b,g_\star(\theta,x)) - f_\star(x)) \\ &= -2\eta\epsilon(f(x,a) - g(\theta,x,a)) - \eta(f(x,a) - g(\theta,x,a))^2 + \lambda(b - \phi_\varsigma(b,g_\star(\theta,x)) - f_\star(x)) \\ &\leq -2\eta\epsilon(f(x,a) - g(\theta,x,a)) - \eta(f(x,a_s) - g(\theta,x,a))^2 + \lambda(g_b(\theta,x) - f_\star(x)) \\ &= -2\eta\epsilon(f(x,a) - g(\theta,x,a)) - \eta \mathrm{LS}_x(\theta,a) + \lambda \mathrm{FG}_x(\theta,a^\theta(x))\ .\end{aligned}$$

Combining the sub-Gaussian equation with (A.2) and the bound of $-\Delta\ell^{(\mathrm{sFG})}$, we find

$$
\begin{aligned}
&\leq -\frac{2}{3}\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathrm{e}^{-3\eta(1-4c\eta)\mathrm{LS}_x(\theta,a_s)/2}] - \frac{1}{3}\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathrm{e}^{3\lambda \mathrm{FG}_x(\theta,a^\theta(x))}] \\
&\leq -\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})},\, r\sim \mathrm{R}(\cdot|x,a)}[\mathrm{e}^{-2\eta(f(x,a)-g(\theta,x,a))\epsilon-\eta\mathrm{LS}_x(\theta,a)+\lambda\mathrm{FG}_x(\theta,a^\theta(x))}] \\
&\leq -\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})},\, r\sim \mathrm{R}(\cdot|x,a)}[\mathrm{e}^{-\Delta\ell^{(\mathrm{sFG})}(\theta,x,a,r)}]\ .
\end{aligned}
$$

□

Lemma 11.

$$Z_t - Z_{t-1} \leq \mathbb{E}_{\nu_0}^T\left[\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathbb{E}_{r\sim R(\cdot|x_s,a_s)}[\mathrm{e}^{-\Delta\ell^{(\mathrm{sFG})}(\theta,x_s,a_s,r)}]]\right]$$

where

$$Z_t = \mathbb{E}_{\nu_0}^T \log \int \exp\left(-\sum_{s=1}^t \Delta\ell^{(\mathrm{sFG})}(\tilde\theta, x_s, a_s, r_s)\right)\mathrm{dp}_0(\tilde\theta)\ ,$$

Proof. The proof is provided in (Zhang 2021) but has been rewritten for completeness.

For ease of notation, let define $K_t(\theta|\mathrm{D}_t) = \exp\{-\sum_{s=1}^t \Delta\ell^{(\mathrm{sFG})}(\theta, x_s, a_s, r_s)\}$ such that $Z_t = \mathbb{E}_{\nu_0}^T[\log \mathbb{E}_{\theta\sim\mathrm{p}_0}[K_t(\theta|\mathrm{D}_t)]$. Then we have

$$
\begin{aligned}
Z_t - Z_{t-1} &= \mathbb{E}_{\nu_0}^T \log \frac{\mathbb{E}_{\theta\sim\mathrm{p}_0}[K_t(\theta|\mathrm{D}_t)]}{\mathbb{E}_{\theta\sim\mathrm{p}_0}[K_{t-1}(\theta|\mathrm{D}_{t-1})]} \\
&= \mathbb{E}_{\nu_0}^T \log \mathbb{E}_{\theta\sim\mathrm{p}_0}\left[\frac{K_t(\theta|\mathrm{D}_t)}{\mathbb{E}_{\tilde\theta\sim\mathrm{p}_0}[K_{t-1}(\tilde\theta|\mathrm{D}_{t-1})]}\right] \\
&= \mathbb{E}_{\nu_0}^T \log \mathbb{E}_{\theta\sim\mathrm{p}_0}\left[\frac{K_{t-1}(\theta|\mathrm{D}_{t-1})\mathrm{e}^{-\Delta\ell^{(\mathrm{sFG})}(\theta,x_t,a_t,r_t)}}{\mathbb{E}_{\tilde\theta\sim\mathrm{p}_0}[K_{t-1}(\tilde\theta|\mathrm{D}_{t-1})]}\right] \\
&= \mathbb{E}_{\nu_0}^T \log \mathbb{E}_{\theta\sim\mu_t^{(\mathrm{sFG})}}[\mathrm{e}^{-\Delta\ell^{(\mathrm{sFG})}(\theta,x_t,a_t,r_t)}] \\
&\leq \mathbb{E}_{\nu_0}^T\left[\log \mathbb{E}_{\theta\sim\mu_s^{(\mathrm{sFG})}}[\mathbb{E}_{r\sim R(\cdot|x_s,a_s)}[\mathrm{e}^{-\Delta\ell^{(\mathrm{sFG})}(\theta,x_s,a_s,r)}]]\right]\ ,
\end{aligned}
$$

where the last line uses Jensen's inequality.

□

A.1.1 Technical Lemmas

Lemma 12. *Let $c > 0$ be given in H 1. If η is chosen according to the following strategy, for any $\epsilon \in]0,1[$,*

$$0 < \eta \leq \begin{cases} 3/(16c) & \text{if } \frac{1}{16c^2} \leq \frac{1-\epsilon}{3c(b+b_f)^2} \\ \min\left(\frac{3}{16c},\ \frac{1}{8c} - \sqrt{\frac{1}{64c^2} - \frac{1-\epsilon}{3c(b+b_f)^2}}\right) & \text{otherwise.} \end{cases} \tag{A.3}$$

Then we have these useful properties

(i) $\eta > 0$,

(ii) $\eta \leq 3/(16c) < 1/(4c)$,

(iii) $1 - (3\eta(1-4c\eta)(b+b_f)^2)/4 \geq \epsilon$,

(iv) $C_\eta > 0$ *where* C_η *is defined in (3.4).*

Proof. The results for (i) and (ii) are obvious regarding the definition of η in (A.3).

Moreover, $\mathrm{P}(\eta) = \eta^2 - \eta/(4c) + (1-\epsilon)/(3c(b+b_f)^2 s)$ is a second order polynomial with determinant $\Delta_\mathrm{P} = 1/(16c^2) - 4(1-\epsilon)/(3c(b+b_f)^2)$.

If $\Delta_{\mathrm{P}} \leq 0 \Leftrightarrow \ (b+b_f)^2 \leq 64(1-\epsilon)c/3$, then P is always positive on its domain.

However, if $\Delta_{\mathrm{P}} > 0 \Leftrightarrow \ (b+b_f)^2 > 64(1-\epsilon)c/3$ then P admits two zeros

$$x_1 = \frac{1}{8c} - \sqrt{\frac{1}{64c^2} - \frac{1-\epsilon}{3c(b+b_f)^2}} \geq 0$$
$$x_2 = \frac{1}{8c} + \sqrt{\frac{1}{64c^2} - \frac{1-\epsilon}{3c(b+b_f)^2}} \geq 0$$

As x_1 is obviously positive, by taking $\eta \leq x_1$ we have $\mathrm{P}(\eta)$ positive and then (iii) is true.

Finally, given (i), (ii) and (iii), C_η is obvisouly strictly positive. □

Lemma 13. *If η is chosen according to A.3, then we have,*

$$\frac{3\lambda KT}{8C_\eta} \leq \frac{\lambda KT}{\epsilon\eta} \ .$$

Proof. By definition of C_η in (3.4) and using the property (iii) of Lemma 12, then we have

$$\begin{aligned} C_\eta &= 1.5\eta(1-4c\eta)[1-3\eta(1-4c\eta)(b+b_f)^2/4] \\ &\geq 1.5\eta(1-4c\eta)\epsilon \end{aligned}$$

Moreover $\eta \leq 3/(16c)$ we have $1-4c\eta \geq 1/4$. Hence,

$$C_\eta \geq \frac{3\epsilon}{8}\eta \ .$$

This last inequality concludes the proof. □

Lemma 14. *If η is chosen according to (A.3), then*

$$\frac{2C_\eta(b+b_f)^2}{3\lambda} \leq \frac{(b+b_f)^2}{4\lambda} \ ,$$

Proof. By definition of C_η in (3.4),

$$\begin{aligned} C_\eta &= 1.5\eta(1-4c\eta)[1-3\eta(1-4c\eta)(b+b_f)^2/4] \\ &\leq 1.5\eta \leq \frac{3}{8} \ , \end{aligned}$$

where the last inequality comes from (A.3). □

A.2 Proof of Corollary 2

Proof. Hereafter we specify the choice of

$$\omega = D_\eta^{-1} \vee L_g \vee 1, \qquad \eta = \frac{1}{\omega}, \qquad \lambda = \sqrt{\frac{d\log(dT)}{\omega KT}}, \tag{A.4}$$

where D_η is the RHS of equation (A.3).

Consider the compact set $\mathrm{B}_\gamma = \{\theta \in \mathbb{R}^d : \|\theta - \theta_*\| \leq \frac{1}{\gamma}\}$ for some $\gamma \geq 1$. By H 2, we know that for any $\theta \in \mathrm{B}_\gamma$, if $\tilde{\theta}_s = (1-s)\theta + s\theta_*$,

$$\begin{aligned} \log \mathrm{p}_0(\theta) - \log \mathrm{p}_0(\theta_*) &\geq -\int_0^1 \left\langle \nabla \log \mathrm{p}_0(\tilde{\theta}_s), \theta_* - \theta \right\rangle \mathrm{d}s \\ &\geq -\int_0^1 \langle \nabla \log \mathrm{p}_0(\theta_*), \theta_* - \theta \rangle \, \mathrm{d}s - L_0 \int_0^1 \|\theta - \theta_*\|^2 \, \mathrm{d}s \\ &\geq -\frac{\|\nabla \log \mathrm{p}_0(\theta_*)\|}{\gamma} - \frac{L_0}{2\gamma^2} \ . \end{aligned}$$

From H 3-4, we get for any $\theta \in \mathrm{B}_\gamma$,

$$\sup_{x \in \mathsf{X}, a \in \mathcal{A}(x)} |g(\theta, x, a) - f(x, a)| \leq \frac{L_g}{\gamma} + \xi \ . \tag{A.5}$$

Consequently for $\theta \in \mathrm{B}_\gamma$, if we let $a_*(x) = \arg\max_{a \in \mathcal{A}(x)} f(x, a)$,

$$\begin{aligned}
-\Delta\ell^{(\mathrm{sFG})}(\theta, x_s, a_s, r_s) &\geq -\eta(g(\theta, x_s, a_s) - f(x_s, a_s))^2 - 2\eta|g(\theta, x_s, a_s) - f(x_s, a_s)||r_s - f(x_s, a_s)| \\
&\quad - \lambda(f_*(x_s) - b + \phi_\varsigma(b - g_*(\theta, x_s))) \\
&\geq -\Big(\frac{\eta L_g}{\gamma} + \eta\xi + 2\eta|r_s - f(x_s, a_s)|\Big)(\xi + \frac{L_g}{\gamma}) - \lambda\big(f_*(x_s) - b + \phi_\varsigma(b - g(\theta, x_s, a^\theta(x_s)))\big)
\end{aligned}$$

In the last line, we used (A.5). Now, let's focus on the last term of the previous inequality

$$\begin{aligned}
f_\star(x_s) - b + \phi_\varsigma\big(b - g(\theta, x_s, a^\theta(x_s))\big) &= f_\star(x_s) - g(\theta, x_s, a^\theta(x_s)) + \phi_\varsigma\big(g(\theta, x_s, a^\theta(x_s)) - b\big) \\
&\leq \frac{L_g}{\gamma} + \xi + \phi_\varsigma(g(\theta, x_s, a^\theta(x_s)) - b) \ ,
\end{aligned}$$

In the first line, we used that $\phi_\varsigma(x) = x + \phi_\varsigma(-x)$. The second line comes from A.5 and that for any $a \in \mathcal{A}(x)$, $f_\star(x) - f_\star(x, a) \leq 0$. Moreover, as ϕ_ς is a growing function, we just have to found an upper bound of $g(\theta, x_s, a^\theta(x_s)) - b$ to bound the previous term.

$$\begin{aligned}
g(\theta, x_s, a^\theta(x_s)) - b &= g(\theta, x_s, a^\theta(x_s)) - f_\star(x_s, a^\theta(x_s)) + f_\star(x_s, a^\theta(x_s)) - b \\
&\leq \frac{L_g}{\gamma} + \xi + b_f - b \ .
\end{aligned}$$

Consequently,

$$-\Delta\ell^{(\mathrm{sFG})}(\theta, x_s, a_s, r_s) \geq -\Big(\frac{\eta L_g}{\gamma} + \eta\xi + \lambda + 2\eta|r_s - f(x_s, a_s)|\Big)(\xi + \frac{L_g}{\gamma}) - \lambda\phi_\varsigma(\frac{L_g}{\gamma} + \xi + b_f - b) \ .$$

Then, taking expectation and using H 1 to control $\mathbb{E}_{\nu_0}[|r_s - f(x_s, a_s)|] \leq \sqrt{2c}$ (see e.g. (Wainwright 2019), Theorem 2.6),

$$\begin{aligned}
\mathbb{E}[\inf_{\theta \in \mathrm{B}_\gamma} -\Delta\ell^{(\mathrm{sFG})}(\theta, x_s, a_s, r_s)] &\geq -\left(\eta\left(\frac{L_g}{\gamma} + \xi\right) + \lambda + 2\sqrt{2c}\eta\right)(\xi + \frac{L_g}{\gamma}) - \lambda\phi_\varsigma(\frac{L_g}{\gamma} + \xi + b_f - b) \\
&\geq -4(1 + \xi + \lambda)(\xi + \frac{L_g}{\gamma}) - \lambda\phi_\varsigma(\frac{L_g}{\gamma} + \xi + b_f - b).
\end{aligned}$$

The last line follows from our choice of η, and $\gamma \geq 1$. Finally, noting that the volume of a d-dimensional ball can be lower bounded by $\exp(-10d\log d)$, we can estimate the probability of B_γ under p_0 with the following

$$\begin{aligned}
\log \mathrm{p}_0(\mathrm{B}_\gamma) &\geq \inf_{\theta \in \mathrm{B}_\gamma} \log \mathrm{p}_0(\theta) - 10d\log\gamma d \\
&\geq \log \mathrm{p}_0(\theta_*) - \frac{L_0}{2\gamma^2} - 10d\log(\gamma d)
\end{aligned}$$

Then we can bound as follows:

$$\begin{aligned}
Z_T &= \mathbb{E}\left[\log \mathbb{E}_{\theta \sim \mathrm{p}_0}\left[\exp\left(-\sum_{s=1}^T \Delta\ell^{(\mathrm{sFG})}(\theta, x_s, a_s, r_s)\right)\right]\right] \\
&\geq \mathbb{E}\log\left(\mathrm{p}_0(\mathrm{B}_\gamma)\inf_{\theta \in \mathrm{B}_\gamma}\exp\left(-\sum_{s=1}^T \Delta\ell^{(\mathrm{sFG})}(\theta, x_s, a_s, r_s)\right)\right) \\
&\geq \log \mathrm{p}_0(\theta_*) - \frac{L_0}{2\gamma^2} - 10d\log(\gamma d) - \Big(4(1 + \xi + \lambda)(\xi + \frac{L_g}{\gamma}) + \lambda\phi_\varsigma(\frac{L_g}{\gamma} + \xi + b_f - b)\Big)T,
\end{aligned}$$

where in the last step we used our bound on $\mathrm{p}_0(\mathrm{B}_\gamma)$.

Finally, substituting $Z_T, \lambda, \eta, \gamma = T$ into Theorem 1, and expanding the product:

$$\begin{aligned}
\mathrm{CREG}(\mathbb{Q}^{(\mathrm{sFG})}_{1:T}) &\leq \frac{\lambda}{\eta\epsilon}KT + \mathrm{C}_1\lambda T - \frac{Z_T}{\lambda} + \left(\mathrm{C}_2 + \frac{\mathrm{C}_3}{\lambda}\right)\sum_{t=0}^{T}\mathbb{E}^T_{\nu_0}[\delta_t] \\
&\leq \frac{\sqrt{\omega dKT\log(dT)}}{\epsilon} + \mathrm{C}_1\sqrt{\frac{dT\log(dT)}{\omega K}} + \left(\mathrm{C}_2 + \mathrm{C}_3\sqrt{\frac{\omega KT}{d\log(dT)}}\right)\sum_{t=0}^{T}\mathbb{E}^T_{\nu_0}[\delta_t] \\
&\quad + \sqrt{\frac{\omega KT}{d\log(dT)}}\left(-\log\mathrm{p}_0(\theta_*) + \frac{L_0}{2T^2} + 10d\log(dT) + 4L_g\right) \\
&\quad + 4\xi\sqrt{\frac{\omega KT}{d\log(dT)}}(T + \xi T + L_g) + 4(\xi T + L_g) + \phi_\varsigma(\frac{L_g}{T} + \xi + b_f - b)T.
\end{aligned}$$

When T satisfies

$$T \geq \sqrt{\frac{L_0}{2d}} \vee L_g \vee e, \tag{A.6}$$

then the following inequalities hold:

$$\frac{L_0}{2T^2} \leq d, \qquad L_g \leq T, \qquad \log T \geq 1.$$

This is a mild assumption and does not impact the viability of the result; the second term is only needed to absorb ξL_g into ξT, and is not necessary when ξ is small.

Consequently, we can make some simplifications to find

$$\begin{aligned}
\mathrm{CREG}(\mathbb{Q}^{(\mathrm{sFG})}_{1:T}) \leq{}& \frac{\mathrm{C}_4\sqrt{\omega dKT\log(dT)}}{\epsilon} + \mathrm{C}_6\left(1 + \sqrt{\frac{\omega KT}{d\log(dT)}}\right)\sum_{t=0}^{T}\mathbb{E}^T_{\nu_0}[\delta_t] \\
&+ \mathrm{C}_5\sqrt{\frac{\omega KT}{d\log(dT)}}\big(-\log\mathrm{p}_0(\theta_*) + L_g + \xi T + \xi^2 T\big) + \big(4\xi + \phi_\varsigma(\frac{L_g}{T} + \xi + b_f - b)\big)T + 4L_g,
\end{aligned}$$

where here we define $\mathrm{C}_4 = 1 + 11\epsilon + \epsilon\mathrm{C}_1/(\omega K) \leq 14 + \mathrm{C}_1$, $\mathrm{C}_6 = \mathrm{C}_2 + \mathrm{C}_3$, $\mathrm{C}_5 = 8$, such that they can be loosely upper bounded by constants not depending on $\epsilon, \omega, d, K, T$. Note that this restriction on T is dimension-free and quite mild.

$\square$

A.3 Proof of Theorem 5

Let D_η again be the RHS of (A.3). Hereafter we specify the choice of

$$\omega_{\mathsf{LG}} = D_\eta^{-1} \vee \sqrt{M} \vee 1, \qquad \eta = \frac{1}{\omega_{\mathsf{LG}}}, \qquad \lambda = \sqrt{\frac{\log(dT)}{\omega_{\mathsf{LG}}T}}, \qquad \varsigma = \sqrt{T}, \qquad b \geq b_f. \tag{A.7}$$

Secondly, the condition on T is now

$$T \geq e \vee \sqrt{\frac{\mathtt{m}_0}{2d}}, \tag{A.8}$$

since as ξ is zero, the second condition in (A.6) is not necessary. Note that this assumption is not very restrictive on T, especially when the dimension is large.

Lemma 15. *If the MCMC method can output p_{t,N_t} such that $\delta_t \leq \frac{1}{T}$, then we obtain the bound for $\mathrm{C}_7 = (\mathrm{C}_4 + \mathrm{C}_5) \vee \mathrm{C}_6$ when the parameters satisfy (A.7), (A.8):*

$$\mathrm{CREG}(\mathbb{Q}_{1:T}^{(\mathrm{sFG})}) \leq \mathrm{C}_7 \sqrt{\omega_{\mathsf{LG}} T \log^3(dT)} \Big(d(\frac{1}{\epsilon} + \frac{1}{\mathtt{m}_0}) + \sqrt{M} + \mathtt{m}_0 \|\theta_*\|^2 \Big).$$

Proof.

The setting of Theorem 5 satisfies all the assumptions of Proposition 2 with $\xi = 0$, $L_g = \sqrt{M}$, $\omega = \omega_{\mathsf{LG}}$.

Let us first examine the term $\phi_\varsigma(\sqrt{M}/T + b_f - b)T$ for our choice of ς, b. In this case,

$$\begin{aligned}
\phi_\varsigma(\frac{L_g}{T} + b_f - b)T &= \frac{\log(1 + \exp(\sqrt{M}/\sqrt{T} + \sqrt{T}(b_f - b)))}{\sqrt{T}} \times T \\
&\leq \sqrt{T} \log(1 + \exp(\sqrt{\frac{M}{T}})) \\
&\leq \sqrt{M} + \sqrt{T}.
\end{aligned}$$

In the second line we use that $b \geq b_f$, and in the third line we use that $\log(1 + \exp(x)) \leq 1 + x$ for $x \geq 0$.

Subsequently, we get the following bound immediately, using that $K/d \leq 1$:

$$\begin{aligned}
\mathrm{CREG}(\mathbb{Q}_{1:T}^{(\mathrm{sFG})}) &\leq \sqrt{\omega_{\mathsf{LG}} T \log(dT)} \Big(2\mathrm{C}_4 \frac{d}{\epsilon} + 3\mathrm{C}_5 \sqrt{M} - \mathrm{C}_5 \log \mathrm{p}_0(\theta_*) + \mathrm{C}_6 \sum_{t=0}^{T} \mathbb{E}_{\nu_0}^{T}[\delta_t] \Big) \\
&\leq \sqrt{\omega_{\mathsf{LG}} T \log(dT)} \Big(2\mathrm{C}_4 \frac{d}{\epsilon} + 3\mathrm{C}_5 \sqrt{M} + \frac{\mathrm{C}_5 \mathtt{m}_0 \|\theta_*\|^2}{2} + \frac{\mathrm{C}_5 d \log 2\pi}{2\mathtt{m}_0} + \mathrm{C}_6 \sum_{t=0}^{T} \mathbb{E}_{\nu_0}^{T}[\delta_t] \Big),
\end{aligned}$$

where in the second line we substitute the density of the Gaussian prior. We absorb the $\sqrt{M}$, $\sqrt{T}$ tersm from ϕ_ς into $\mathrm{C}_4, \mathrm{C}_5$. Since $4 \leq 2\sqrt{wT \log(dT)} C_5$, the $4L_g$ term in Corollary 2 is absorbed into the $\mathrm{C}_5 \sqrt{M}$ seen above. If we substitute $\delta_t \leq 1/T$, this last part of the sum can be absorbed as a factor of $\log(T) \leq \log(dT)$, and then we choose $\mathrm{C}_7 = (2\mathrm{C}_4 + 3\mathrm{C}_5) \vee \mathrm{C}_6$ to complete the proof.

□

Remark: We can assume instead $K \leq \mathrm{C}_K d$ for some absolute constant C_K, with this constant subsequently appearing at multiple places in the proof. For ease of presentation, we do not do this.

Consequently, this allows us to use gradient descent to estimate the modes of the successive posteriors with negligible cost (with the previous mode for bootstrapping). We state a theorem for gradient descent which makes this rate rigorous:

Lemma 16 (Adapted from (Nesterov et al. 2018), Theorem 2.1.15)**.** *Given a μ-strongly convex, λ-smooth function g with condition number κ and an initial point θ_0, gradient descent with step-size $2/(\mu + \lambda)$ can find the mode $\theta^* = \arg\min_\theta g(\theta)$ with rate*

$$N \geq 2\kappa \log\Big(\frac{\|\theta_0 - \theta\|}{\epsilon} \Big) \implies \|\theta_N - \theta^*\| \leq \epsilon.$$

We will not discuss this result extensively as it is only necessary to furnish a modal estimate for MCMC methods. The use of gradient descent is standard and has been well-studied, e.g. in the aforementioned (Nesterov et al. 2018).

We show a polynomial in time bound on the norms of the iterates, which is crude but sufficient for our purposes.

Lemma 17. *Let θ_t^* be the mode of the posterior $\mu_t^{(\mathrm{sFG})}$. Then the following holds, where b_r is the a.s. bound on the reward:*

$$\|\theta_t^*\| \leq \frac{2t\sqrt{Mt}}{\mathtt{m}_0} (\frac{b_r}{\omega_{\mathsf{LG}}} + \lambda)$$

In particular, we immediately get the crude bound

$$\|\theta_t^* - \theta_{t-1}^*\| \leq \frac{4t\sqrt{Mt}}{\mathtt{m}_0} (\frac{b_r}{\omega_{\mathsf{LG}}} + \lambda). \tag{A.9}$$

Proof. First consider the minimizer of the posterior for Thompson sampling without feel-good adjustment ($\mu_t^{(\mathrm{TS})}$), and denote it by ζ_t^*. Then, since ζ_t^* is just the solution of a regularized least squares problem, we know the following bound on ζ_t^*:

$$\zeta_t^* = (\mathbf{\Phi}_t^\top \mathbf{\Phi}_t + \frac{\mathtt{m}_0}{\eta} I_d)^{-1} \mathbf{\Phi}_t^\top \mathbf{r}_t, \qquad \mathbf{\Phi}_t = \begin{bmatrix} \varphi(x_1, a_1) \\ \varphi(x_2, a_2) \\ \dots \\ \varphi(x_t, a_t) \end{bmatrix}, \qquad \mathbf{r}_t = \begin{bmatrix} r_1 \\ r_2 \\ \dots \\ r_t \end{bmatrix}.$$

Here $\mathbf{\Phi}_t$ is the data matrix which has $\varphi(x_i, a_i)$ in its i-th row. In particular, since the matrix $\mathbf{\Phi}_t^\top \mathbf{\Phi}_t + \frac{\mathtt{m}_0}{\eta} I_d \succeq \omega_{\mathsf{LG}} \mathtt{m}_0 I_d$, $\|\mathbf{\Phi}_t\| \le t\sqrt{M}$ and $\|\mathbf{r_t}\|_2 \le \sqrt{t} b_r$, we obtain

$$\|\zeta_t^*\| \le \frac{1}{\omega_{\mathsf{LG}} \mathtt{m}_0} \|\mathbf{\Phi_t}\| \|\mathbf{r_t}\| \le \frac{b_r t \sqrt{Mt}}{\omega_{\mathsf{LG}} \mathtt{m}_0}. \tag{A.10}$$

Secondly, writing the difference in negative log-likelihoods as:

$$-\log \mu_t^{(\mathrm{TS})}(\theta) = -\log \mu_t^{(\mathrm{sFG})}(\theta) + \underbrace{\lambda \sum_{s=1}^t \big[b - \phi_\varsigma(b - \langle \theta, \varphi(x_s, a^\theta(x_s)) \rangle)\big]}_{\mathsf{J}_t(\theta)}.$$

We now seek to estimate $\|\theta_t^* - \zeta_t^*\|$, using that ζ_t^*, θ_t^* minimize their respective posteriors:

$$\begin{aligned}
0 &= \Big\|\nabla \log \mu_t^{(\mathrm{sFG})}(\theta_t^*) - \nabla \log \mu_t^{(\mathrm{TS})}(\zeta_t^*)\Big\|^2 \\
&= \Big\|\nabla \log \mu_t^{(\mathrm{sFG})}(\theta_t^*) - \nabla \log \mu_t^{(\mathrm{sFG})}(\zeta_t^*) + \nabla \mathsf{J}_t(\zeta_t^*)\Big\|^2 \\
&= \Big\|\nabla \log \mu_t^{(\mathrm{sFG})}(\theta_t^*) - \nabla \log \mu_t^{(\mathrm{sFG})}(\zeta_t^*)\Big\|^2 + \|\nabla \mathsf{J}_t(\zeta_t^*)\|^2 + 2\Big\langle \nabla \log \mu_t^{(\mathrm{sFG})}(\theta_t^*) + \nabla \log \mu_t^{(\mathrm{sFG})}(\zeta_t^*), \nabla \mathsf{J}_t(\zeta_t^*) \Big\rangle \\
&\ge \Big\|\nabla \log \mu_t^{(\mathrm{sFG})}(\theta_t^*) - \nabla \log \mu_t^{(\mathrm{sFG})}(\zeta_t^*)\Big\|^2 + \|\nabla \mathsf{J}_t(\zeta_t^*)\|^2 - 2\Big|\Big\langle \nabla \log \mu_t^{(\mathrm{sFG})}(\theta_t^*) - \nabla \log \mu_t^{(\mathrm{sFG})}(\zeta_t^*), \nabla \mathsf{J}_t(\zeta_t^*) \Big\rangle\Big|.
\end{aligned}$$

Let us proceed to use Young's inequality $|\langle a, b \rangle| \le 1/4\|a\|^2 + \|b\|^2$, to find

$$\begin{aligned}
&\Big\|\nabla \log \mu_t^{(\mathrm{sFG})}(\theta_t^*) - \nabla \log \mu_t^{(\mathrm{sFG})}(\zeta_t^*)\Big\|^2 + \|\nabla \mathsf{J}_t(\zeta_t^*)\|^2 \\
&\quad \le 2\Big|\Big\langle \nabla \log \mu_t^{(\mathrm{sFG})}(\theta_t^*) - \nabla \log \mu_t^{(\mathrm{sFG})}(\zeta_t^*), \nabla \mathsf{J}_t(\zeta_t^*) \Big\rangle\Big| \\
&\quad \le \frac{1}{2}\Big\|\nabla \log \mu_t^{(\mathrm{sFG})}(\theta_t^*) - \nabla \log \mu_t^{(\mathrm{sFG})}(\zeta_t^*)\Big\|^2 + 2\|\nabla \mathsf{J}_t(\zeta_t^*)\|^2.
\end{aligned}$$

After some rearranging, we get

$$\Big\|\nabla \log \mu_t^{(\mathrm{sFG})}(\theta_t^*) - \nabla \log \mu_t^{(\mathrm{sFG})}(\zeta_t^*)\Big\|^2 \le 2\|\nabla \mathsf{J}_t(\zeta_t^*)\|^2.$$

We use triangle inequality and the boundedness of φ to get for all $\theta \in \mathbb{R}^d$

$$\begin{aligned}
\|\nabla \mathsf{J}_t(\theta)\| &= \left\|\lambda \sum_{s=1}^t \frac{\exp(\varsigma(b - \langle \theta, \varphi(x_s, a^\theta(x_s)) \rangle))}{\exp(\varsigma(b - \langle \theta, \varphi(x_s, a^\theta(x_s)) \rangle)) + 1} \varphi(x_s, a^\theta(x_s))\right\| \\
&\le \lambda t \sqrt{M}.
\end{aligned}$$

From the strong convexity of $-\log \mu_t^{(\mathrm{sFG})}$, we get $\Big\|\nabla \log \mu_t^{(\mathrm{sFG})}(\theta_t^*) - \nabla \log \mu_t^{(\mathrm{sFG})}(\zeta_t^*)\Big\|^2 \ge \mathtt{m}_t^2 \|\theta_t^* - \zeta_t^*\|^2 \ge \mathtt{m}_0^2 \|\theta_t^* - \zeta_t^*\|^2$. Finally, this implies

$$\|\theta_t^* - \zeta_t^*\| \le \frac{\sqrt{2M}\lambda t}{\mathtt{m}_0}.$$

Substituting (A.10) completes the proof. □

Remarks: Much better bounds are possible through more careful analysis, but since it is only necessary to provide very rough bounds (as gradient descent is a fast algorithm), this will suffice for our purposes.

First we formally state the warm-start condition:

Definition 18. *(Warm-Start Condition) Let μ, ν be two distributions on $\mathbb{R}^d$. We say that a distribution μ is a $c_W(\mu, \nu)$ warm-start for another distribution ν if*

$$\sup_{\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)} \frac{\mu(\mathsf{A})}{\nu(\mathsf{A})} \leq c_W(\mu, \nu) ,$$

where $\mathcal{B}(\mathbb{R}^d)$ is the Borel σ-field of $\mathbb{R}^d$.

Finally, we state the consequences of gradient descent for finding appropriate warm-starts for our MCMC methods.

Corollary 19. *Using gradient descent methods, at time t, we can find an approximate mode $\hat{\theta}_t^*$, so that when we construct the prior $p_{t,0} = \mathrm{N}(\hat{\theta}_t^*, (2\mathtt{L}_t^{(\mathrm{sFG})})^{-1}\mathrm{I}_d)$ (with I_d the d-dimensional identity matrix), then*

$$\log c_W(p_{t,0}, \mu_t^{(\mathrm{sFG})}) \leq d \log 2\kappa, \qquad \mathsf{KL}(p_{t,0} \parallel \mu_t^{(\mathrm{sFG})}) \leq d \log 2\kappa.$$

is satisfied with only $2\kappa \log^2(8b\mathtt{L}_t^{(\mathrm{sFG})}\sqrt{M}T/\mathtt{m}_0)$ iterations of gradient descent.

Proof. For each time t, we can first estimate $\hat{\theta}_t^*$ using gradient descent from $\hat{\theta}_{t-1}^*$. We choose the desired accuracy to be $\epsilon = \sqrt{d/(2\mathtt{L}_t^{(\mathrm{sFG})})}$ at each time t. Using Lemmas 16 and (A.9), this can be done with number of iterations $4\kappa \log(8b\mathtt{L}_t^{(\mathrm{sFG})}\sqrt{M}T/\mathtt{m}_0)$.

Then, Section 3.2.1 of (Dwivedi et al. 2018) shows that $p_{t,0}$ chosen here attains a warm-start with $c_W(p_{t,0}, \mu_t^{(\mathrm{sFG})}) \leq \exp(d\log(2\kappa))$. Finally, for the KL bound, we need only note that

$$\mathsf{KL}(p \parallel q) = \int \log \frac{p}{q} \, \mathrm{d}p \leq \log c_W(p \parallel q).$$

□

Remark: Summing the number of iterations over $t \in [T]$, and noting that each iteration of gradient descent is equal to a full pass through the data, this yields $4\kappa T^2 \log(8b\mathtt{L}_t^{(\mathrm{sFG})}\sqrt{M}T/\mathtt{m}_0)$ data complexity. This is dominated by the data complexity due to sampling in all cases.

A.3.1 Langevin Monte Carlo

For the result under LMC, we can give the following state-of-the-art rate, following the result of (Durmus, Majewski, and Miasojedow 2019).

Lemma 20 (Adapted from (Durmus, Majewski, and Miasojedow 2019), Corollary 11)**.** *For targets with condition number $\kappa = \mathtt{L}/\mathtt{m}$, ambient dimension d and error tolerance ϵ, if we take the ergodic distribution of the $N/2$ to N LMC iterates, for some N even, $2/N \sum_{k=N/2}^{N} \theta_k$ with the law of θ_k denoted p_k and the stationary distribution μ, we get*

$$N^{\mathrm{L}} = \frac{\mathrm{C_L}\tilde{\mathrm{C}}_\kappa d}{\delta^2} \log \frac{2\mathcal{W}_2(p_0 \parallel \mu)}{\delta^2} \implies ||\frac{2}{N}\sum_{k=N/2}^{N} p_k - \mu||_{\mathrm{TV}} \leq \delta,$$

for some absolute constant $\mathrm{C_L}$, with $\tilde{\mathrm{C}}_\kappa = \max(\mathtt{L}/\mathtt{m}^2, \mathtt{L})$ and $\mathcal{W}_2$ is the 2-Wasserstein distance between measures. Here the step size is chosen as

$$\gamma^{\mathrm{L}} = \mathrm{A_L}\frac{\delta_t^2}{(\kappa \vee \mathtt{L})d}.$$

where $\mathrm{A_L} > 0$ is an absolute constant.

Secondly, we state a lemma:

Lemma 21 (Talagrand's Inequality, (Bakry, Gentil, Ledoux, et al. 2014) Corollary 9.3.2)**.** *If p is strongly convex with constant α, then $\mathcal{W}_2^2(q \,\|\, p) \le 2/\alpha \mathsf{KL}(q \,\|\, p)$.*

Finally, we are ready to show the complexity for LMC.

Proof of Proposition 5, LMC: To show the MCMC complexity, it remains only to combine Lemma 21 with Corollary 19. This shows that the Wasserstein term can be bounded

$$\log(2\mathcal{W}_2(p_{t,0} \,\|\, \mu_t^{(\mathrm{sFG})})) \le \log\left(\frac{2}{\mathtt{m}_t}\sqrt{d\log 2\kappa}\right).$$

Consequently, we apply with the choice $\delta_t \le \frac{1}{T}$. which yields $N_t^{\mathrm{L}} \le \mathrm{C_L C_\kappa} dT^2 \log(4\sqrt{d\kappa}/\mathtt{m}_0)$, and $\gamma_t^{\mathrm{L}} = \mathrm{A_L}/((\kappa \vee \mathtt{L})dT^2)$. This implies that at time t, the data complexity is $G_t \le \mathrm{C_L C_\kappa} dT^3 \log(4\sqrt{d\kappa}/\mathtt{m}_0)$, and that cumulative data complexity is

$$\sum_{t=1}^{T} G_t \le \mathrm{C_L C_\kappa} dT^4 \log(4\sqrt{d\kappa}/\mathtt{m}_0)$$

A.3.2 Metropolis Algorithm

Let us state the conditions required for MALA to obtain a fast rate, seen e.g. in (Chen et al. 2020).

Proposition 22. *(One-Step Convergence of Bandit MALA (Chen et al. 2020), Theorem 5) Assume that the initial distribution p_0 satisfies Definition 18 with $\log c_W(p_0, \mu) \le d\log(2\kappa)$, where μ is the stationary distribution of the chain. Assume further that the potential has condition number κ. Then the MALA algorithm converges to the true posterior with the following rate:*

$$N \ge \mathrm{C_M}\kappa d \log\left(\frac{d}{\delta^2}\right)\left(1 \vee \sqrt{\frac{\kappa}{d}}\right) \implies ||p_N - \mu||_{\mathrm{TV}} \le \delta,$$

when we take the step size to be

$$\gamma^{\mathrm{M}} = \frac{\mathrm{A_M}}{\mathtt{L}d\max\left(1, \sqrt{\kappa/d}\right)},$$

with $\mathrm{A_M}$ again an absolute constant.

Immediately, we can see that the critically dependency on the error tolerance ϵ are significantly better when contrasted with the unadjusted Langevin algorithm.

Proof of Proposition 5, MALA: The warm-start condition for all $t \le T$ is immediately implied by Corollary 19. Consequently, recalling that we pick $\delta_t \le \frac{1}{T}$ at each iteration t, we only need to perform $N_t = \mathrm{C_M}\kappa d\log(dT^2)(1 \vee \kappa/d)$ MALA iterations at each time t. Since each MALA iteration contains t gradients, this has data complexity $G_t \le \mathrm{C_M}\kappa dT\log(dT^2)(1 \vee \kappa/d)$. Finally,

$$\sum_{t=1}^{T} G_t \le \mathrm{C_M}\kappa dT^2\log(dT^2)(1 \vee \kappa/d).$$

B Numerical experiments

B.1 Toy Example

In this section, we give additional details about the Toy example settings. As presented in the section 4.1, the reward distribution considered in this toy example is Gaussian and all parameters used to describe the problem are provided in Table 1.

For each algorithm, we studied a pool of hyperparameters, and Figure 1 represents the best combination of hyperparameter for each approach. Table 2 summarizes the pool of hyperparameters studied during the experiment. Notice that the step size, parameter λ, and the standard deviation of the prior depend on the parameter η. This choice is subjective but seems to be

Parameter dimension (d)	20
Context dimension (d_x)	4
Number of arms (K)	5
Noise level (σ)	1
Time horizon (T)	1000

Table 1: Environment hyperparameters

η	[1, 5, 10, 50, 100, 500, 1000]
Step Size	$[1/(t\eta), 0.5/(\eta t), 0.1/(\eta t), 0.05/(\eta t), 0.01/(\eta t)]$
λ	$[0.5\eta, 0.1\eta, 0.05\eta]$
Gaussian Prior Std	0.01η
Number of gradient updates	[25, 50, 100]
b	1000
Gradient descent steps for MALA / FG-MALA	20

Table 2: Algorithm hyperparameters

quite logical. The step size is also depending on the time step t. For MALA-TS and FG-MALA-TS, we initialize MALA with the output of a full-batch gradient descent during 20 steps.

The baseline algorithm LinUCB has been studied for different values of α. However, for clarity, figure 1 shows only the performance of LinTS, LMC-TS, MALA-TS, FG-LMC-TS and FG-MALA-TS. The study of LinUCb is provided in figure 3. Notice that the best α among the pool studied is 0.1 and with this setting LinUCB outperforms all algorithms except FG-MALA-TS.

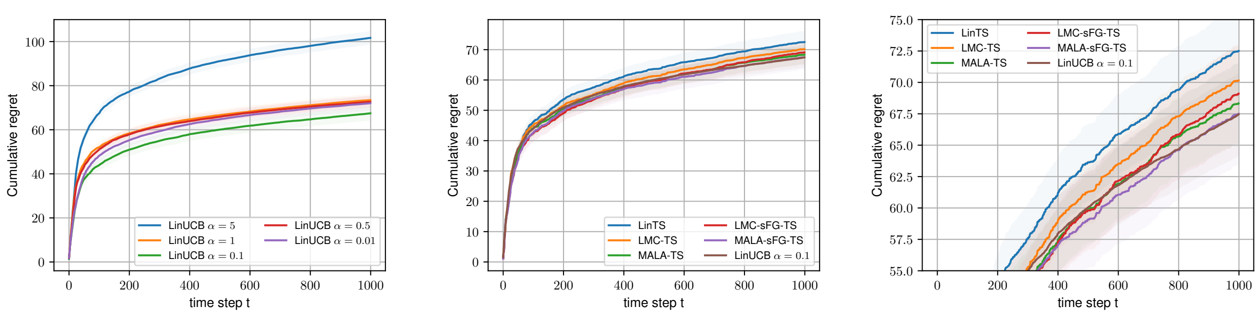

Figure 3: Linear UCB study

B.2 Real World Dataset

Table 3 summarizes the main parameters used for the Yahoo! Front Page Today Module Dataset. A more detailed description of the problem can be found in (Li, Chu, et al. 2010). Our implementation of this task is based on the git repository: https://github.com/antonismand/Personalized-News-Recommendation.

Parameter dimension (d)	12
Context dimension (d_x)	12
Number of arms (K)	22
Time horizon (T)	25000

Table 3: Environment hyperparameters

Similarly, Table 4 describes the pool of hyperparameters studied during this experiment. Therefore, Figure 2 shows only the best comparison among this pool.

η	[1, 3, 5, 10, 20, 30, 40, 50]
Step size	$0.1/(t\eta)$
λ	$[0.1\eta, 0.3\eta, 0.5\eta]$
Gaussian Prior std	0.01η
Number of gradient updates	100
b	1000
Gradient Descent steps for MALA/FG-MALA	20

Table 4: Algorithm hyperparameters

B.3 Logistic bandit

In this section we investigate the behavior of Feel-Good Thompson Sampling on a more complex setting: the logistic bandit. We follow the setting of Kveton et al. 2020b and Xu, Zheng, et al. 2022. We consider a contextual vector $x \in \mathbb{R}^{20}$ sampled from $\mathrm{N}(0_{20}, \mathrm{I}_{20})$ and scaled to unit norm. A fixed set of 50 arms. And a Bernoulli reward distribution such that $r \sim \mathrm{B}(\phi(\theta^{\star T} x))$ where $\theta^\star$ is the true parameter, sampled from $\mathrm{N}(0_{20}, \mathrm{I}_{20})$ and scaled to unit norm. The function $\phi(u) = 1/(1+e^{-u})$ is the logistic function.

Figure 4 shows the cumulative regret, ie, $\mathbb{E}_{\Pi \sim \mathbb{Q}_{1:T}}[\sum_{t=1}^{T} 1 - f(x_t, \pi_s(x_t))]$ for LMC-TS and FG-LMC-TS. For the later, we consider four different values of λ. We observe that for small λ (≤ 0.01) FG-LMC-TS outperforms LMC-TS. However, when λ is too high, FG-LMC-TS becomes unstable and linear. It means that in this setting, the parameter λ has to be carefully determined. The implementation is based on the repository git: `https://github.com/devzhk/LMCTS`. The hyperparameters used for this experiment are provided in Table 5

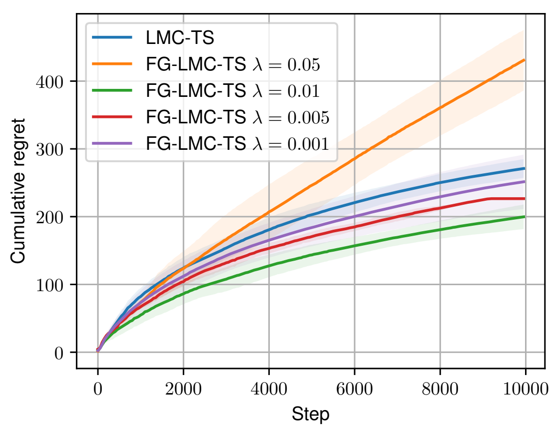

Figure 4: Comulative regret for logistic bandit over 10 runs

Time horizon (T)	10000
Number of LMC steps	500
Step size	0.001
Inverse temperature (β^{-1})	0.001

Table 5: Hyperparameters for logistic bandit