# On Approximate Thompson Sampling with Langevin Algorithms

**Eric Mazumdar** [*1] **Aldo Pacchiano** [*1] **Yi-An Ma** [*2 3] **Peter L. Bartlett** [1 4] **Michael I. Jordan** [1 4]

## Abstract

Thompson sampling for multi-armed bandit problems is known to enjoy favorable performance in both theory and practice. However, its wider deployment is restricted due to a significant computational limitation: the need for samples from posterior distributions at every iteration. In practice, this limitation is alleviated by making use of approximate sampling methods, yet provably incorporating approximate samples into Thompson Sampling algorithms remains an open problem. In this work we address this by proposing two efficient Langevin MCMC algorithms tailored to Thompson sampling. The resulting approximate Thompson Sampling algorithms are efficiently implementable and provably achieve optimal instance-dependent regret for the Multi-Armed Bandit (MAB) problem. To prove these results we derive novel posterior concentration bounds and MCMC convergence rates for log-concave distributions which may be of independent interest.

## 1. Introduction

Sequential decision making under uncertainty has become one of the fastest developing fields of machine learning. A central theme in such problems is addressing exploration-exploitation tradeoffs (Auer et al., 2002; Lattimore and Szepesvári, 2020), wherein an algorithm must balance between exploiting its current knowledge and exploring previously unexplored options.

[*]Equal contribution [1]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA [2]Google Research [3]Halıcıoğlu Data Science Institute, University of California, San Diego, USA [4]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA. Correspondence to: Eric Mazumdar <mazumdar@berkeley.edu>, Aldo Pacchiano <pacchiano@berkeley.edu>, Yi-An Ma <yianma@google.com>, Peter Bartlett <peter@berkeley.edu>, Michael I. Jordan <jordan@cs.berkeley.edu>.

*Proceedings of the $37^{th}$ International Conference on Machine Learning*, Online, PMLR 119, 2020. 

The classic stochastic multi-armed bandit problem has provided a theoretical laboratory for the study of exploration/exploitation tradeoffs (Lai and Robbins, 1985). A vast literature has emerged that provides algorithms, insights, and matching upper and lower bounds in many cases. The dominant paradigm in this literature has been that of *frequentist analysis*; cf. in particular the analyses devoted to the celebrated upper confidence bound (UCB) algorithm (Auer et al., 2002). Interestingly, however, Thompson sampling, a Bayesian approach first introduced almost a century ago (Thompson, 1933) has been shown to be competitive and sometimes outperform UCB algorithms in practice (Scott, 2010; Chapelle and Li, 2011). Further, the fact that Thompson sampling, being a Bayesian method, explicitly makes use of prior information, has made it particularly popular in industrial applications (see, e.g., Russo et al., 2017, and the references therein).

Although most theory in the bandit literature is focused on non-Bayesian methods, there is a smaller, but nontrivial, theory associated with Thompson sampling. In particular, Thompson sampling has been shown to achieve optimal risk bounds in multi-armed bandit settings with Bernoulli rewards and beta priors (Kaufmann et al., 2012; Agrawal and Goyal, 2013a), Gaussian rewards with Gaussian priors (Agrawal and Goyal, 2013a), one-dimensional exponential family models with uninformative priors (Korda et al., 2013), and finitely-supported priors and observations (Gopalan et al., 2014). Thompson sampling has further been shown to asymptotically achieve optimal instance-independent performance (Russo and Van Roy, 2016).

Despite these appealing foundational results, the deployment of Thompson sampling in complex problems is often constrained by its use of samples from posterior distributions, which are often difficult to generate in regimes where the posteriors do not have closed forms. A common solution to this has been to use *approximate* sampling techniques to generate samples from *approximations* of the posteriors (Russo et al., 2017; Chapelle and Li, 2011; Gómez-Uribe, 2016; Lu and Van Roy, 2017). Such approaches have been demonstrated to work effectively in practice (Riquelme et al., 2018; Urteaga and Wiggins, 2018), but it is unclear how to maintain performance over arbitrary time horizons while using approximate sampling. Indeed, to the best of our knowledge the strongest regret guarantees for Thompson

sampling with approximate samples, due to Lu and Van Roy (2017), require a model whose complexity grows with the time horizon to guarantee optimal performance. Further, it was recently shown theoretically by Phan et al. (2019) that a naïve usage of approximate sampling algorithms with Thompson sampling can yield a drastic drop in performance.

**Contributions** In this work we analyze Thompson sampling with approximate sampling methods in a class of multi-armed bandit algorithms where the rewards are unbounded, but their distributions are log-concave. In Section 3 we derive posterior contraction rates for posteriors when the rewards are generated from such distributions and under general assumptions on the priors. Using these rates, we show that Thompson sampling with samples from the true posterior achieves finite-time optimal frequentist regret. Further, the regret guarantee we derive has explicit constants and explicit dependencies on the dimension of the parameter spaces, variance of the reward distributions, and the quality of the prior distributions.

In Section 4 we present a simple counterexample demonstrating the relationship between the approximation error to the posterior and the resulting regret of the algorithm. Building on the insight provided by this example, we propose two approximate sampling schemes based on Langevin dynamics to generate samples from approximate posteriors and analyze their impact on the regret of Thompson sampling. We first analyze samples generated from the unadjusted Langevin algorithm (ULA) and specify the runtime, hyperparameters, and initialization required to achieve an approximation error which provably maintains the optimal regret guarantee of exact Thompson sampling over finite-time horizons. Crucially, we initialize the ULA algorithm from the approximate sample generated in the previous round to make use of the posterior concentration property and ensure that only a *constant* number of iterations are required to achieve the optimal regret guarantee. Under slightly stronger assumptions, we then demonstrate that a stochastic gradient variant called *stochastic gradient Langevin dynamics* (SGLD) requires only a *constant* batch size in addition to the constant number of iterations to achieve logarithmic regret. Since the computational complexity of this sampling algorithm does not scale with the time horizon, the proposed method is a true "anytime" algorithm. Finally, we conclude in Section F by validating these theoretical results in numerical simulations where we find that Thompson sampling with our approximate sampling schemes maintain the desirable performance of exact Thompson sampling.

Our results suggest that the tailoring of approximate sampling algorithms to work with Thompson sampling can overcome the phenomenon studied in Phan et al. (2019), where approximation error in the samples can yield linear regret. Indeed, our results suggest that it is possible for Thompson sampling to achieve order-optimal regret guarantees with an efficiently implementable approximate sampling algorithm.

## 2. Preliminaries

In this work we analyze Thompson sampling strategies for the $K$-armed stochastic multi-armed bandit (MAB) problem. In such problems, there is a set of $K$ options, or "arms," $\mathcal{A} = \{1, ..., K\}$, from which a player must choose at each round $t = 1, 2, ....$ After choosing an arm $A_t \in \mathcal{A}$ in round $t$, the player receives a real-valued reward $X_{A_t}$ drawn from a fixed yet unknown distribution associated with the arm, $p_{A_t}$. The random rewards obtained from playing an arm repeatedly are i.i.d. and independent of the rewards obtained from choosing other arms.

Throughout this paper, we assume that the reward distribution for each arm is a member of a parametric family, parametrized by $\theta_a \in \mathbb{R}^{d_a}$, such that the true reward distribution is $p_a(X) = p_a(X; \theta_a^*)$, where $\theta_a^*$ is unknown. Moreover, we assume throughout this paper that the parametric families are log-concave and Lipschitz smooth in $\theta_a$:

**Assumption 1-Local** (Assumption on the family $p_a(X|\theta_a)$ around $\theta_a^*$)**.** *Assume that* $\log p_a(x|\theta_a)$ *is* $L_a$*-smooth and* $m_a$*-strongly concave around* $\theta_a^*$*. For all* $x \in \mathbb{R}$ *and* $\theta_a \in \mathbb{R}^{d_a}$*:*

$$
\begin{aligned}
&-\nabla_\theta \log p_a(x|\theta_a^*)^\top (\theta_a - \theta_a^*) + \frac{m_a}{2}\|\theta_a - \theta_a^*\|^2 \\
&\qquad \le -\big(\log p_a(x|\theta_a) - \log p_a(x|\theta_a^*)\big) \le \\
&\qquad -\nabla_\theta \log p_a(x|\theta_a^*)^\top (\theta_a - \theta_a^*) + \frac{L_a}{2}\|\theta_a - \theta_a^*\|^2.
\end{aligned}
$$

Additionally we make assumptions on the true distribution of the rewards:

**Assumption 2** (Assumption on true reward distribution $p_a(X|\theta_a^*)$). For every $a \in \mathcal{A}$ assume that $p_a(X; \theta_a^*)$ is strongly log-concave in $X$ with some parameter $\nu_a$, and that $\nabla_\theta \log p_a(x|\theta_a^*)$ is $L_a$-Lipschitz in $X$. $\forall x, x' \in \mathbb{R}$:

$$
\begin{aligned}
&-\left(\nabla_x \log p_a(x|\theta_a^*) - \nabla_x \log p_a(x'|\theta_a^*)\right)^{\mathrm{T}}(x - x') \\
&\qquad\qquad \ge \nu_a \|x - x'\|_2^2. \\
&\|\nabla_\theta \log p_a(x|\theta_a^*) - \nabla_\theta \log p_a(x'|\theta_a^*)\| \le L_a \|x - x'\|_2.
\end{aligned}
$$

Parameters $\nu_a$ and $L_a$ provide lower and upper bounds to the sub- and super-Gaussianity of the true reward distributions. We further define $\kappa_a = \max\{L_a/m_a, L_a/\nu_a\}$ to be the condition number of the model class. Finally, we assume that for each arm $a \in \mathcal{A}$ there is a linear map such that for all $\theta_a \in \mathbb{R}^{d_a}$, $\mathbb{E}_{x \sim p_a(x|\theta_a)}[X] = \alpha_a^T \theta_a$, with $\|\alpha_a\| = A_a$.

We now review Thompson sampling, the pseudo-code for which is presented in Algorithm 1. A key advantage of Thompson sampling over frequentist algorithms for multi-armed bandit problems is its flexibility in incorporating prior information. In this paper, we assume that the prior

distributions over the parameters of the arms have smooth log-concave densities:

**Assumption 3** (Assumptions on the prior distribution). For every $a \in \mathcal{A}$ assume that $\log \pi_a(\theta_a)$ is concave with $L_a$-Lipschitz gradients for all $\theta_a \in \mathbb{R}^{d_a}$:[1]

$$\|\nabla_\theta \pi_a(\theta_a) - \nabla_\theta \pi_a(\theta'_a)\| \leq L_a \|\theta_a - \theta'_a\|, \ \forall \theta_a, \theta'_a \in \mathbb{R}^{d_a}.$$

Thompson sampling proceeds by maintaining a posterior distribution over the parameters of each arm $a$ at each round $t$. Given the likelihood family, $p(X|\theta_a)$, the prior, $\pi(\theta_a)$, and the $n$ data samples from an arm $a$, $X_{a,1}, \cdots, X_{a,n}$, let $F_{n,a} : \mathbb{R}^{d_a} \to \mathbb{R}$ be $F_{n,a}(\theta_a) = \frac{1}{n}\sum_{i=1}^{n} \log p_a(X_{a,i}|\theta_a)$, be the average log-likelihood of the data. Then the posterior distribution over the parameter $\theta_a$ at round $t$, denoted $\mu_a^{(n)}$, satisfies:

$$\begin{aligned} p_a(\theta_a|X_{a,1}, \cdots, X_{a,n}) &\propto \pi_a(\theta_a) \prod_{i=1}^{t} \left(p_a(X_t|\theta_a)\right)^{\mathbb{I}\{A_t = a\}} \\ &= \exp\left(nF_{n,a}(\theta_a) + \log \pi(\theta_a)\right). \end{aligned}$$

For any $\gamma_a > 0$ we denote the scaled posterior[2] as $\mu_a^{(n)}[\gamma_a]$, whose density is proportional to:

$$\exp\left(\gamma_a(nF_{n,a}(\theta_a) + \log \pi(\theta_a))\right). \tag{1}$$

Letting $T_a(t)$ be the number of samples received from arm $a$ after $t$ rounds, a Thompson sampling algorithm, at each round $t$, first samples the parameters of each arm $a$ from their (scaled) posterior distributions, $\theta_{a,t} \sim \mu_a^{(T_a(t))}[\gamma_a]$, and then chooses the arm for which the sample has the highest value:

$$A_t = \operatorname*{argmax}_{a \in \mathcal{A}} \alpha_a^T \theta_{a,t}.$$

A player's objective in MAB problems is to maximize her cumulative reward over any fixed time horizon $T$. The measure of performance most commonly used in the MAB literature is known as the *expected regret*, $R(T)$, which corresponds to the expected difference between the accrued reward and the reward that would have been accrued had the learner selected the action with the highest mean reward during all steps $t = 1, \cdots, T$.[3] Recalling that $\bar{r}_a$ is the mean reward for arm $a \in \mathcal{A}$, the regret is given by:

$$R(T) := \mathbb{E}\left[\sum_{t=1}^{T} \bar{r}_{a^*} - \bar{r}_{A_t}\right],$$

**Algorithm 1** Thompson sampling

**Input :** Priors $\pi_a$ for $a \in \mathcal{A}$; posterior scaling: $\gamma_a$

Set $\mu_{a,t} = \pi_a$ for $a \in \mathcal{A}$.

**for** $t = 0, 1, \cdots$ **do**

  Sample $\theta_{a,t} \sim \mu_a^{(T_a(t))}[\gamma_a]$.

  Choose action $A_t = \operatorname{argmax}_{a \in \mathcal{A}} \alpha_a^T \theta_{a,t}$.

  Receive reward $X_{A_t}$.

  Update posterior distribution for arm $A_t$: $\mu_a^{(T_a(t+1))}$.

where $\bar{r}_{a^*} = \max_{a \in \mathcal{A}} \bar{r}_a$. Without loss of generality, we assume throughout this paper that the optimal arm, $a^* = \operatorname{argmax}_{a \in \mathcal{A}} \bar{r}_a$, is arm 1. Further, we assume that the optimal arm is unique[4]: $\bar{r}_1 > \bar{r}_a$ for $a > 1$.

Traditional treatment of Thompson sampling algorithms often overlooks one of its most critical aspects: ensuring compatibility between the mechanism that produces samples from the posterior distributions and the algorithm's regret guarantees. This issue is usually addressed by assuming that the prior distributions and the reward distributions are conjugate pairs. Although this approach is simple and prevalent in the literature (see, e.g., Russo et al., 2017), it fails to capture more complex distributional families for which this assumption may not hold. Indeed, it was recently shown in Phan et al. (2019) that if the samples come from distributions that approximate the posteriors with a constant error, the regret may grow at a linear rate. A more nuanced understanding of the relationship between the quality of the samples and the regret of the algorithms is, however, still lacking.

In the following sections we analyze Thompson sampling in two settings. In the first, the algorithm uses samples corresponding to the *true* scaled posterior distributions, $\{\mu_a^{(T_a(t))}[\gamma_a]\}_{a \in \mathcal{A}}$, at each round. In the second, Thompson sampling makes use of samples coming from two approximate sampling schemes that we propose, such that the samples can be seen as corresponding to *approximations* of the scaled posteriors, $\{\bar{\mu}_a^{(T_a(t))}[\gamma_a]\}_{a \in \mathcal{A}}$. We refer to the former as *exact* Thompson sampling, and the latter as *approximate* Thompson sampling.

For the analysis of *exact* Thompson sampling in Section 3 we derive posterior concentration theorems which characterize the rate at which the posterior distributions for the arms, $\mu_a^{(n)}$, converge to delta functions centered at $\theta_a^*$ as a function of the number of $n$, the number of samples received from the arm. We then use these rates to show that Thompson sampling in this family of multi-armed bandit problems achieves the optimal finite-time regret. Further, our results demonstrate an explicit dependence on the quality of the pri-

[1] We remark that the Lipschitz constants are all assumed to be the same to simplify notation.

[2] In Section 3 we explain the use of scaled posteriors is required to obtain optimal regret guarantees for our bandit algorithms.

[3] We remark that the analysis of Thompson sampling has often been focused on a different quantity known as the Bayes regret, which is simply the expectation of $R(T)$ over the priors: $\mathbb{E}_\pi[R(T)]$. However, in an effort to demonstrate that Thompson sampling is an effective alternative to frequentist methods like UCB, we analyze the frequentist regret $R(T)$.

[4] We introduce this assumption merely for the purpose of simplifying our analysis.

ors and other problem-dependent constants, which improve upon prior works.

In Section 4, we propose two efficiently implementable Langevin-MCMC-based sampling schemes for which the regret of approximate Thompson sampling still achieves the optimal logarithmic regret. To do so, we derive new results for the convergence of Langevin-MCMC-based sampling schemes in the Wasserstein-$p$ distance which we then use to prove optimal regret bounds.

## 3. Exact Thompson Sampling

In this section we first derive posterior concentration rates on the parameters of the reward distributions when the data, the priors, and the likelihoods satisfy our assumptions. We then make use of these concentration results to give finite-time regret guarantees for exact Thompson sampling in log-concave bandits.

### 3.1. Posterior Concentration Results

Core to the analysis of Thompson sampling is understanding the behavior of the posterior distributions over the parameters of the arms' distributions as the algorithm progresses and samples from the arms are collected.

The literature on understanding how posteriors evolve as data is collected goes back to Doob (1949) and his proof of the asymptotic normality of posteriors. More recently, there has been a line of work (see, e.g., van der Vaart and van Zanten, 2008; Ghosal and van der Vaart, 2007) that derives rates of convergence of posteriors in various regimes, mostly following the framework first developed in Ghosal et al. (2000) for finite- and infinite-dimensional models. Such results, though quite general, do not have explicit constants or forms which make them amenable for use in analyzing bandit algorithms. Indeed, finite-time rates remain an active area of research but have been developed using information-theoretic arguments (Shen and Wasserman, 2001), and more recently through the analysis of stochastic differential equations (Mou et al., 2019), though in both cases the assumptions, burn-in times, and lack of precise constants make them difficult to integrate with the analysis of Thompson sampling. Due to this, Thompson sampling has, for the most part, been only well understood for conjugate prior/likelihood families like beta/Bernoulli and Gaussian/Gaussian (Agrawal and Goyal, 2013a), or in more generality in well-behaved families such as one-dimensional exponential families with uninformative priors (Korda et al., 2013) or finitely supported prior/likelihood pairs (Gopalan et al., 2014).

To derive posterior concentration rates for parameters in $d$-dimensions and for a large class of priors and likelihoods we analyze the moments of a potential function along trajectories of a stochastic differential equation for which the posterior is the limiting distribution. Our results expand upon the recent derivation of novel contraction rates for posterior distributions presented in Mou et al. (2019) to hold for a finite number of samples and may be of independent interest. We make use of these concentration results to show that Thompson sampling with such priors and likelihoods results in order-optimal regret guarantees.

To begin, we note that classic results (Øksendal, 2003) guarantee that, as $t \to \infty$ the distribution $P_t$ of $\theta_t$ which evolves according to:

$$d\theta_t = \frac{1}{2}\nabla_\theta F_{n,a}(\theta_t)dt + \frac{1}{2n}\nabla_\theta \log \pi_a(\theta_t))dt + \frac{1}{\sqrt{n\gamma_a}}dB_t, \tag{2}$$

is given by:

$$\lim_{t\to\infty} P_t(\theta|X_1,...,X_n) \propto \exp(-\gamma_a\,(nF_{n,a}(\theta) + \log\ \pi_a(\theta))),$$

almost surely. Comparing with Eq. (1), this limiting distribution is the scaled posterior distribution $\mu_a^{(n)}[\gamma_a]$. Thus, by analyzing the limiting properties of $\theta_t$ as it evolves according to the stochastic differential equation, we can derive properties of the scaled posterior distribution.

To do so, we first show that with high probability the gradient of $F_{n,a}(\theta^*)$ concentrates around zero (given the data $X_1,...,X_n$). More precisely in Appendix B we show, using well-known results on the concentration of Lipschitz functions of strongly log-concave random variables, that $\nabla_\theta F_{a,n}(\theta_a^*)$ has sub-Gaussian tails:

**Proposition 1.** *The random variable $\|\nabla_\theta F_{a,n}(\theta_a^*)\|$ is $L_a\sqrt{\frac{d_a}{n\nu_a}}$-sub-Gaussian.*

Given this proposition, we then analyze how the potential function,

$$V(\theta_t) = \frac{1}{2}e^{ct}\|\theta_t - \theta^*\|_2^2,$$

evolves along trajectories of the stochastic differential equation (2), where $c > 0$. By bounding the supremum of $V(\theta_t)$, we construct bounds on the higher moments of the random variable $\|\theta_a - \theta_a^*\|$ where $\theta_a \sim \mu_a^{(n)}[\gamma_a]$. These moment bounds translate directly into the posterior concentration bound of $\theta_a \sim \mu_a^{(n)}[\gamma_a]$ around $\theta^*$ presented in the following theorem (the proof of which is deferred to Appendix B).

**Theorem 1.** *Suppose that Assumptions 1-3 hold, then for $\delta \in (0, e^{-1/2})$:*

$$\mathbb{P}_{\theta\sim\mu_a^{(n)}[\gamma_a]}\left(\|\theta_a - \theta_a^*\|_2 > \Gamma_a\right) < \delta,$$

*where:*

$$\Gamma_a(\delta) = \sqrt{\frac{2e}{m_a n}\left(\frac{d_a}{\gamma_a} + \log B_a + \left(\frac{32}{\gamma_a} + 8d_a\kappa_a^2\right)\log\frac{1}{\delta}\right)}$$

*and $B_a = \max_{\theta\in\mathbb{R}^d}\frac{\pi_a(\theta)}{\pi_a(\theta_a^*)}$.*

Theorem 1 guarantees that the scaled posterior distribution over the parameters of the arms concentrate at rate $\frac{1}{\sqrt{n}}$, where $n$ is the number of times the arm has been pulled.

We remark that this posterior concentration result has a number of desirable properties. Through the presence of $B_a$, it reflects an explicit dependence on the quality of the prior. In particular, $\log B_a = 0$ if the prior is properly centered such that its mode is at $\theta^*$ or if the prior is uninformative or nearly flat everywhere. We further remark that the concentration result also scales with the variance of $\theta_a$ which is on the order of $d_a/(\gamma_a m_a n)$. Lastly, we remark that this concentration result holds for any $n > 0$ and the constants are explicitly defined in terms of the smoothness and structural assumptions on the priors, likelihoods, and reward distributions. This makes it more amenable for use in constructing regret guarantees, since we do not have to wait for a burn-in period for the result to hold, as in Shen and Wasserman (2001) and Mou et al. (2019). Moreover, the dependence on the dimension of the parameter space and constants are explicit.

### 3.2. Exact Regret for Thompson Sampling

We now show that, under our assumptions, Thompson sampling with samples from the scaled posterior enjoys optimal finite-time regret guarantees. To provide these results we proceed as is common in regret proofs for multi-armed bandits by upper bounding $T_a(T)$, the number of times a sub-optimal arm $a \in \mathcal{A}$ is pulled up to time $T$. Without loss of generality we assume throughout this section that arm 1 is the optimal arm, and define the filtration associated with a run of the algorithm as $\mathcal{F}_t = \{A_1, X_1, A_2, X_2, ..., A_t, X_t\}$.

To upper bound the expected number of times a sub-optimal arm is pulled up to time $T$, we first define the low-probability event that the mean calculated from the value of $\theta_{a,t}$ sampled from the posterior at time $t \leq T$, $r_{a,t}(T_a(t))$, is greater than $\bar{r}_1 - \epsilon$ (recall that $\bar{r}_1$ is the optimal arm's mean): $E_a(t) = \{r_{a,t}(T_a(t)) \geq \bar{r}_1 - \epsilon\}$ for some $\epsilon > 0$. Given these events, we proceed to decompose the expected number of pulls of a sub-optimal arm $a \in \mathcal{A}$ as:

$$\mathbb{E}[T_a(T)] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}(A_t = a)\right] = \tag{3}$$
$$\underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}(A_t = a, E_a^c(t))\right]}_{\text{I}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}(A_t = a, E_a(t))\right]}_{\text{II}}.$$

These two terms satisfy the following standard bounds (see, e.g., Lattimore and Szepesvári (2020)):

**Lemma 1** (Bounding I and II). *For a sub-optimal arm $a \in \mathcal{A}$, we have that:*

$$\text{I} \leq \mathbb{E}\left[\sum_{s=1}^{T-1} \frac{1}{p_{1,s}} - 1\right]; \tag{4}$$

$$\text{II} \leq 1 + \mathbb{E}\left[\sum_{s=1}^{T} \mathbb{I}\left(p_{a,s} > \frac{1}{T}\right)\right], \tag{5}$$

*where $p_{a,s} = \mathbb{P}(r_{a,t}(s) > \bar{r}_1 - \epsilon | \mathcal{F}_{t-1})$, for some $\epsilon > 0$.*

The proof of these results are standard for the regret of Thompson sampling and can be found in Appendix E, Lemmas 13 and 14, for completeness.

Given Lemma 1, we see that to bound the regret of Thompson Sampling it is sufficient to bound the two terms I and II.

To bound term I, we first show that for all times $t = 1, ..., T$, and number of samples collected from arm 1, the probability $p_{1,n} = \mathbb{P}(r_{1,t}(n) > \bar{r}_1 - \epsilon | \mathcal{F}_{t-1})$ is lower bounded by a constant depending only on the quality of the prior for arm 1. This guarantees the posterior for the optimal arm is approximately optimistic with at least a constant probability, and requires a proper choice of $\gamma_1$. We note the unscaled posterior provides the correct concentration with respect to the number of data samples $T_a(t)$, when $T_a(t)$ is large. This is sufficient to upper bound the trailing terms of I, that is, summands in Equation 4 for large $s$. Unfortunately concentration is not enough to bound term I, since the early summands of Equation 4 corresponding to small values of $s$ could be extremely large. Intuitively, the random variable $r_{1,t}(s)$ can be thought of as centered around the posterior mean of arm 1. Though this is close to the true value of $\bar{r}_1$ with high probability, when $T_1(t)$ is small, concentration alone does not preclude the possibility that the posterior mean underestimates $\bar{r}_1$ by a value of at least $\epsilon$. In order to ensure $p_{1,s}$ is large enough in these cases, we require $r_{1,t}(s)$ to have sufficient variance to overcome this potential underestimation bias. We show that a scaled posterior $\mu_a^{(T_a(t))}[\gamma_a]$ with $\gamma_a = \left(8 d_a \kappa_a^3\right)^{-1}$ in Algorithm 1 ensures $r_{1,t}(s)$ has enough variance.

**Lemma 2.** *Suppose the likelihood and reward distributions satisfy Assumptions 1-3, then for all $n = 1, ..., T$ and $\gamma_1 = \frac{1}{8 d_1 \kappa_1^3}$:*

$$\mathbb{E}\left[\frac{1}{p_{1,n}}\right] \leq C\sqrt{B_1 \kappa_1},$$

*where C is a universal constant independent of the problem dependent parameters.*

**Remark 1.** *We find that a proper choice of $\gamma_1$ is required to ensure that that the posterior on the optimal arm has a large enough variance to guarantee a degree of optimism despite the randomness in its mean. Scaling up the posterior was also noted to be necessary in linear bandits (see, e.g.,*

*Agrawal and Goyal (2013b); Abeille and Lazaric (2017)) to ensure optimal regret. In practice, since we do not* a priori *know which is the optimal arm, we must scale the posterior of each arm by a parameter* $\gamma_a$.

The quantity $B_1 = \frac{\max_\theta \pi_1(\theta)}{\pi_1(\theta_1^*)}$ captures a worst case dependence on the quality of the prior for the optimal arm, and can be seen as the expected number of samples from the prior until an optimistic sample is observed.

By using this upper bound in conjunction with the posterior concentration result derived in Theorem 1, we can further bound I and II. We note that in contrast with simple sub-Gaussian concentration bounds, our posterior concentration rates have a bias term decreasing at a rate of $1/\sqrt{\text{number of samples}}$. In our analysis we carefully track and control the effects of this bias term ensuring it does not compromise our $\log$-regret guarantees. Indeed, using the posterior concentration in the bounds from Lemma 1 we show that, for $\gamma_a = \frac{1}{8 d_a \kappa_a^3}$, there are two universal constants $C_1, C_2 > 0$ independent of the problem-dependent parameters such that:

$$\mathrm{I} \leq C_1 \sqrt{\kappa_1 B_1} \left\lceil \frac{A_1^2}{m_1 \Delta_a^2} (D_1 + \sigma_1) \right\rceil + 1;$$

$$\mathrm{II} \leq \frac{C_2 A_a^2}{m_a \Delta_a^2} (D_a + \sigma_a \log(T)),$$

where for $a \in \mathcal{A}$, $D_a$ and $\sigma_a$ are given by:

$$D_a = \log B_a + d_a^2 \kappa_a^3, \qquad \sigma_a = d_a \kappa_a^3 + d_a \kappa_a^2.$$

Finally, combining all these observations we obtain the following regret guarantee:

**Theorem 2** (Regret of Exact Thompson Sampling)**.** *When the likelihood and true reward distributions satisfy Assumptions 1-3 and* $\gamma_a = \frac{1}{8 d_a \kappa_a^3}$ *we have that the expected regret after* $T > 0$ *rounds of Thompson sampling with exact sampling satisfies:*

$$\mathbb{E}[R(T)] \leq \sum_{a>1} \frac{C A_a^2}{m_a \Delta_a} \left( \log B_a + d_a^2 \kappa_a^3 + d_a \kappa_a^3 \log(T) \right) + \sqrt{\kappa_1 B_1} \frac{C A_1^2}{m_1 \Delta_a} \left( 1 + \log B_1 + d_1^2 \kappa_1^3 \right) + \Delta_a,$$

*where* $C$ *is a universal constant independent of problem-dependent parameters.*

The proof of the theorem is deferred to Appendix E, where we also provide the exact value of the universal constant $C$. We remark that this regret bound gives an $O\left(\frac{\log(T)}{\Delta}\right)$ asymptotic regret guarantee, but holds for any $T > 0$. This further highlights that Thompson sampling is a competitive alternative to UCB algorithms since it achieves the optimal problem-dependent rate for multi-armed bandit algorithms first presented in Lai and Robbins (1985).

Our bound also has explicit dependencies on the dimension of the parameter space of the likelihood distributions for each arm, as well as on the quality of the priors through the presence of $B_a$ and $B_1$. We note that the dependence on the priors does not distinguish between "good" and "bad" priors. Indeed, the parameter $B_a \geq 1$ is worst case, and does not capture the potential advantages of good priors in Thompson sampling, that we observe in our numerical experiments in Section F. Further, we remark that our bound exhibits a worse dependence on the prior for the optimal arm ($O(\sqrt{B_1}\log(B_1))$) than for sub-optimal arms ($O(\log(B_a))$). This is also a worst case dependence which captures the expected number of samples from the prior until an approximately optimistic sample is observed, which we believe to be unavoidable.

Finally, we note that our regret bound scales with the variances of the reward and likelihood families since $\frac{1}{m_a}$ and $\frac{1}{\nu_a}$ reflect the variance of the likelihoods in $\theta$ and the rewards $X_a$ respectively.

Thus, through the use of the posterior contraction rates we are able to get finite-time regret bounds for Thompson sampling with multi-dimensional log-concave families and arbitrary log-concave priors. This generalizes the result of Korda et al. (2013) to a more general class or priors and higher-dimensional parametric families.

## 4. Approximate Thompson Sampling

In this section we present two approximate sampling schemes for generating samples from approximations of the (scaled) posteriors at each round. For both, we give the values of the hyperparameters and computation time needed to guarantee an approximation error which does *not* result in a drastic change in the regret of the Thompson sampling algorithm.

Before doing so, however, we first present a simple counterexample to illustrate that in the worst case, Thompson sampling with approximate samples incurs an irreducible regret dependent on the error between the posterior and the approximation to the posterior. In particular, by allowing the approximation error to decrease over time, we extract a relationship between the order of the regret and the level of approximation.

**Example 1.** *Consider a Gaussian bandit instance of two arms* $\mathcal{A} = \{1, 2\}$ *having mean rewards* $\bar{r}_1$ *and* $\bar{r}_2$ *and known unit variances. Further assume that the unknown parameters are the means of the distributions such that* $\theta_a^* = \bar{r}_a$*, and consider the case where the learner makes use of a zero-mean, unit-variance Gaussian prior over* $\theta_a$ *for* $a = 1, 2$*. Under these assumptions, after obtaining samples* $X_{a,1}, \cdots, X_{a,n}$*, the posterior updates satisfy the following well-known formula:*

**Algorithm 2** (Stochastic Gradient) Langevin Algorithm for Arm $a$

**Input** : Data $\{x_{a,1}, \cdots, x_{a,n}\}$;
MCMC sample $\theta_{a,Nh^{(n-1)}}$ from last round

Set $\theta_0 = \theta_{a,t-1}$ for $a \in \mathcal{A}$

**for** $i = 0, 1, \cdots N$ **do**

  Uniformly subsample $\mathcal{S} \subseteq \{x_{a,1}, \cdots, x_{a,n}\}$.

  Compute $\nabla \widehat{U}(\theta_{ih^{(n)}}) = -\frac{n}{|\mathcal{S}|} \sum_{x_k \in \mathcal{S}} \nabla \log p_a(x_k|\theta_{ih^{(n)}}) - \nabla \log \pi_a(\theta_{ih^{(n)}})$.

  Sample $\theta_{(i+1)h^{(n)}} \sim \mathcal{N}\left(\theta_{ih^{(n)}} - h^{(n)} \nabla \widehat{U}(\theta_{ih^{(n)}}), 2h^{(n)} \mathrm{I}\right)$.

**Output** : $\theta_{a,Nh^{(n)}} = \theta_{Nh^{(n)}}$; $\theta_{a,t} \sim \mathcal{N}\left(\theta_{Nh^{(n)}}, \frac{1}{nL_a\gamma_a} I\right)$

$$P_{a,n}(\theta_a) \propto \mathcal{N}\left(\frac{n}{n+1}, \frac{1}{n+1}\right).$$

*Let $\bar{r}_1 = 1$ and $\bar{r}_2 = 0$ such that arm 1 is optimal. We now show there exists an approximate posterior $\tilde{P}_{a,t}$ of arm 2, satisfying $\mathrm{TV}(\tilde{P}_{2,t}, P_{2,t}) \leq n^{-\alpha}$ and such that if samples from $P_{1,t}$ and $\tilde{P}_{2,t}$ were to be used by a Thompson sampling algorithm, its regret would satisfy $R(T) = \Omega(T^{1-\alpha})$.*

*We substantiate this claim by a simple construction. Let $\tilde{P}_{a,t}$ be $(1 - n^{-\alpha})P_{a,t} + n^{-\alpha}\delta_2$, where $\delta_2$ denotes a delta mass centered at 2. $\tilde{P}_{a,t}$ is a mixture distribution between the true posterior and a point mass.*

*Clearly, for all $t \geq C$ for some universal constant $C$, with probability at least $n^{-\alpha}$ the posterior sample from arm 2 will be larger than the sample from arm 1. Since $t > n$, $t^{-\alpha} < n^{-\alpha}$ for $\alpha > 0$ and since the suboptimality gap equals 1, we conclude $R(T) = \Omega(\sum_{t=1}^{T} t^{-\alpha})$. Thus, to incur logarithmic regret, one needs $TV(\tilde{P}_{2,t}, P_{2,t}) = \Omega(\frac{1}{n})$.*

Example 1 builds on the insights in Phan et al. (2019), who showed that constant approximation error can incur linear regret, which highlights the fact that to achieve logarithmic regret the total variation distance between the approximation of the posterior $\bar{\mu}_a^{(n)}[\gamma_a]$ and the true posterior $\mu_a^{(n)}$ must decrease as samples are collected. In particular it illustrates that the rate at which the approximation error decreases is directly linked to the resulting regret bound.

Given this result, we first propose an unadjusted Langevin algorithm (ULA) (Durmus and Moulines, 2017), which generates samples from an approximate posterior which monotonically approaches the true posterior as data is collected and provably maintains the regret guarantee of exact Thompson sampling. Important to this effort, we demonstrate that the number of steps inside the ULA procedure does not scale with the time horizon, though the number of gradient evaluations scale with the number of times an arm has been pulled. To over this issue arising from full gradient evaluation, we propose a stochastic gradient Langevin dynamics (SGLD) (Welling and Teh, 2011) variant of ULA which has appealing computational benefits: under slightly stronger assumptions, SGLD takes a constant number of iterations as well as a constant number of data samples in the stochastic gradient estimate while maintaining the order-optimal regret of the exact Thompson sampling algorithm.

### 4.1. Convergence of (Stochastic Gradient) Langevin Algorithms

As described in Algorithm 2, in each round $t$ we run the (stochastic gradient) Langevin algorithm for $N$ steps to generate a sample of desirable quality for each arm. In particular, we first run a Langevin MCMC algorithm to generate a sample from an approximation to the unscaled posterior. To achieve the scaling with $\gamma_a$ that we require for the analysis of the regret, we add zero-mean Gaussian noise with variance $\frac{1}{\gamma_a L_a n}$ to this sample. The distribution of the resulting sample has the same characteristics as those from the scaled posterior analyzed in Sec. 3.

Given Assumptions 1-Uniform and 3, we prove (in Theorem 5 in the Appendix) that running ULA with exact gradients provides appealing convergence properties. In particular, for a number of iterations independent of the number of rounds $t$ or the number of samples from an arm, $n = T_a(t)$, ULA converges to an accuracy in Wasserstein-$p$ distance which maintains the logarithmic regret of the exact algorithm (for more information on such metrics see Villani (2009)). We note parenthetically that working with the Wasserstein-$p$ distance provides us with a tighter MCMC convergence analysis (than with the total variation distance used in Example 1) that helps in conjunction with the regret bounds. The proofs of the ULA and SGLD convergence require a uniform strong log-concavity and Lipschitz smoothness condition of the family $p_a(X|\theta_a)$ over the parameter $\theta_a$, a strengthening of Assumption 1-Local.

**Assumption 1-Uniform** (Assumption on the family $p_a(X|\theta_a)$: strengthened for approximate sampling)**.** *Assume that $\log p_a(x|\theta_a)$ is $L_a$-smooth and $m_a$-strongly concave over the parameter $\theta_a$. For all $\theta_a, \theta'_a \in \mathbb{R}^{d_a}, x \in \mathbb{R}$:*

$$-\nabla_\theta \log p_a(x|\theta'_a)^\top (\theta_a - \theta'_a) + \frac{m_a}{2}\|\theta_a - \theta'_a\|^2 \leq -\big(\log p_a(x|\theta_a) - \log p_a(x|\theta'_a)\big) \leq -\nabla_\theta \log p_a(x|\theta'_a)^\top (\theta_a - \theta'_a) + \frac{L_a}{2}\|\theta_a - \theta'_a\|^2.$$

Although the number of iterations required for ULA to converge is constant with respect to the time horizon $t$, the number of gradient computations over the likelihood function within each iteration is $T_a(t)$. To tackle this issue, we sub-sample the data at each iteration and use a stochastic gradient MCMC method (Ma et al., 2015). To get conver-

gence guarantees despite the larger variance this method incurs, we make a slightly stronger Lipschitz smoothness assumption on the parametric family of likelihoods.

**Assumption 4** (Joint Lipschitz smoothness of the family $\log p_a(X|\theta_a)$: for SGLD). Assume a joint Lipschitz smoothness condition, which strengthens Assumptions 1-Uniform and 2 to impose the Lipschitz smoothness on the entire bivariate function $\log p_a(x;\theta)$. For all $\theta_a, \theta'_a \in \mathbb{R}^{d_a}, x, x' \in \mathbb{R}$ :[5]

$$\|\nabla_\theta \log p_a(x|\theta_a) - \nabla_\theta \log p_a(x'|\theta_a)\| \leq L_a \|\theta_a - \theta'_a\| + L_a^* \|x - x'\|.$$

Under this stronger assumption, we prove the fast convergence of the SGLD method in the following Theorem 3. Specifically, we demonstrate that for a suitable choice of stepsize $h^{(n)}$, number of iterations $N$, and size of the minibatch $k = |\mathcal{S}|$, samples generated by Algorithm 2 are distributed sufficiently close to the true posterior to ensure the optimal regret guarantee. By examining the number of iterations, $N$, and size of the minibatch, $k$, we confirm that the algorithmic and sample complexity of our method do not grow with the number of rounds $t$, as advertised.

**Theorem 3** (SGLD Convergence)**.** *Assume that the family $\log p_a(x;\theta)$, prior distributions, and that the true reward distributions satisfy Assumptions 1-Uniform through 4. If we take the batch size $k = \mathcal{O}\left(\kappa_a^2\right)$, step size $h^{(n)} = \mathcal{O}\left(\frac{1}{n}\frac{1}{\kappa_a L_a}\right)$ and number of steps $N = \mathcal{O}\left(\kappa_a^2\right)$ in the SGLD algorithm, then for $\delta_1 \in (0,1)$, with probability at least $1-\delta_1$ with respect to $X_{a,1}, ...X_{a,n}$, we have convergence of the SGLD algorithm in the Wasserstein-$p$ distance. In particular, between the $n$-th and the $(n+1)$-th pull of arm $a$, samples $\theta_{a,t}$ approximately follow the posterior $\mu_a^{(n)}$:*

$$W_p\left(\widehat{\mu}_a^{(n)}, \mu_a^{(n)}\right) \leq \sqrt{\frac{8}{nm_a}}\left(d_a + \log B_a + \left(32 + 8d_a\kappa_a^2\right)p\right)^{\frac{1}{2}},$$

*where $\widehat{\mu}_a^{(n)}$ is the probability measure associated with any of the sample(s) $\theta_{a,Nh_a^{(n)}}$ between the $n$-th and the $(n+1)$-th pull of arm $a$.*

We are able to keep the number of iterations, $N$, for both algorithms constant by initializing the current round of the approximate sampling algorithm using the output of the last round of the Langevin MCMC algorithm. If we initialized the algorithm independently from the prior, we would need $O(\log T_a(t))$ iterations to achieve this result, which would in turn yield a Thompson sampling algorithm for which the computational complexity grows with the time horizon.

[5]For simplicity of notation, we let Lipschitz constants $L_a^* = L_a$ in the main paper.

This warm-starting complicates the regret proof for the approximate Thompson sampling algorithms since the samples used by Thompson sampling are no longer independent.

By scrutinizing the stepsize $h^{(n)}$ and the accuracy level of the sample distribution $W_p\left(\widehat{\mu}_a^{(n)}, \mu_a^{(n)}\right)$, we note that we are taking smaller steps to get increasingly accurate MCMC samples as more data are being collected. This is due to the need of decreasing the error incurred by discretizing the continuous Langevin dynamics and stochastically estimating the gradient of the log posterior. However, the number of iterations and subsampled gradients are not increasing since the concentration of the posterior provides us with stronger contraction of the continuous Langevin dynamics and requires less work because $\mu_a^{(n)}$ and $\mu_a^{(n+1)}$ are closer.

We restate Theorem 3 and give explicit values of the hyperparameters in Theorem 6 in the appendix, but remark that the proof of this theorem is novel in the MCMC literature. It builds upon and strengthens Durmus and Moulines (2016) by taking into account the discretization and stochastic gradient error to achieve strong convergence guarantees in the Wasserstein-$p$ distance up to any finite order $p$. Other related works on the convergence of ULA can provide upper bounds in the Wassertein distances up to the second order (i.e., for $p \leq 2$) (see, e.g., Dalalyan and Karagulyan, 2019; Cheng and Bartlett, 2018; Ma et al., 2019; Vempala and Wibisono, 2019). This bound in the Wasserstein-$p$ distance for arbitrarily large $p$ is necessary in guaranteeing the following Lemma 3, a similar concentration result as in Theorem 1 for the approximate samples $\theta_{a,t} \sim \bar{\mu}_a^{(n)}[\gamma_a]$.

**Lemma 3.** *If Assumptions 1-Uniform through 4 hold, then for $\delta \in (0, e^{-1/2})$, the sample $\theta_{a,t}$ resulting from running the (stochastic gradient) ULA with $N$ steps, a step size of $h^{(n)}$, and a batch-size $k$ as defined in Theorem 3 satisfies:*

$$\mathbb{P}_{\theta_{a,t} \sim \bar{\mu}_a^{(n)}[\gamma_a]}\left(\|\theta_{a,t} - \theta_a^*\|_2 > \Gamma_a(\delta)\right) < \delta,$$

*where:*

$$\Gamma_a(\delta) = \sqrt{\frac{36e}{m_a n}\left(d_a + \log B_a + 2\left(\sigma_a + \frac{d_a}{18\kappa_a\gamma_a}\right)\log\frac{1}{\delta}\right)}$$

*and $\sigma_a = 16 + 4d_a\kappa_a^2$.*

### 4.2. Thompson Sampling Regret with (Stochastic Gradient) Langevin Algorithms

Given that the concentration results of the samples from ULA and SGLD have the same form as that of exact Thompson sampling, we now show that approximate Thompson sampling achieves the same *finite*-time optimal regret guarantees (up to constant factors) as the exact Thompson sampling algorithm. To show this, we require a result analogous to Lemma 2 on the anti-concentration properties of the approximations to the scaled posteriors:

**Lemma 4.** *Suppose the likelihood and true reward distributions satisfy Assumptions 1-Uniform–4: then if* $\gamma_1 = O\left(\frac{1}{d_1\kappa_1^3}\right)$*, for all* $n = 1, ..., T$ *all samples from the (stochastic gradient) ULA method with the hyperparameters and runtime as described in Theorem 3 satisfy:*

$$\mathbb{E}\left[\frac{1}{p_{1,n}}\right] \leq C\sqrt{B_1},$$

*where* $C$ *is a universal constant independent of problem-dependent parameters.*

The proof of Lemma 4 is similar to that of 2, but we are able to save a factor of $\sqrt{\kappa_1}$ due to the fact that the last step of the approximate sampling scheme samples $\theta_{a,t}$ from a Gaussian distribution as opposed to a strongly-log concave distribution which we must approximate with a Gaussian.

Given this lemma and our concentration results presented in the previous section, the proof of logarithmic regret is the same as that of the regret for exact Thompson sampling. However, more care has to be taken to deal with the fact that the samples from the approximate posteriors are no longer independent because we warm-start our proposed sampling algorithms using previous samples. We cope with this issue by constructing concentration rates (of a similar form as in Lemma 3) on the distributions of the samples given the initial sample is sufficiently well behaved (see Lemmas 11 and 12). We then show that this happens with sufficiently high probability to maintain similar upper bounds on terms $I$ and $II$ from Lemma 1 in Lemma 17, which in turn allows us to prove the following Theorem in Appendix E.2.

**Theorem 4** (Regret of Thompson sampling with a (stochastic gradient) Langevin algorithm)**.** *When the likelihood and true reward distributions satisfy Assumptions 1-Uniform–4: the expected regret of* $T > 0$ *rounds of Thompson sampling with the (stochastic gradient) ULA method with hyperparameters and runtime as in Theorem 3 satisfies:*

$$\mathbb{E}[R(T)] \leq \sum_{a>1} \frac{CA_a^2}{m_a\Delta_a}\left(\log B_a + d_a + d_a^2\kappa_a^2 \log T\right)$$
$$+ \frac{\sqrt{B_1}CA_1^2}{m_1\Delta_a}\left(1 + \log B_1 + d_1^2\kappa_1^2 + d_1\kappa_1^2 \log T\right) + 3\Delta_a,$$

*where* $C$ *is a universal constant and the scale parameter* $\gamma_a = O\left(\frac{1}{d_a\kappa_a^3}\right)$.

Theorem 3 allows for SGLD to be implemented with a constant number of steps per iteration and a constant batch size with only the step size decreasing linearly with the number of samples. Combining this with our regret guarantee shows that an anytime algorithm for Thompson sampling with approximate samples can achieve logarithmic regret.

Further, we remark that this bound exhibits a *worse* dependence on the quality of the prior on the optimal arm than in the exact sampling regime. In particular, we pay $d_1^2\sqrt{B_1}\log T$ in this bound as opposed to $d_1^2\sqrt{B_1}$. Our regret bound in the approximate sampling regime exhibits a slightly better dependence on the condition number of the family. This, we believe, is an artifact of our analysis and is due to the fact that a lower bound on the exact posterior was needed to invoke Gaussian anti-concentration results which were not needed in the approximate sampling regime due to the design of the proposed sampling algorithm.

## 5. Conclusions

Although Thompson sampling has been long been used successfully in real-world problems there remains a lack of understanding of how approximate sampling affects its regret guarantees.

In this work we derived new posterior contraction rates for log-concave likelihood families with arbitrary log-concave priors which capture key dependencies between the posterior distributions and various problem-dependent parameters such as the prior quality and the parameter dimension. We used these rates to show that exact Thompson sampling in MAB problems where the reward distributions are log-concave achieves the optimal finite-time regret guarantee for MAB bandit problems. As a direction for future work, we note that although our regret bound demonstrates a dependence on the quality of the prior, it still is unable to capture the potential advantages of good priors.

We then demonstrated that Thompson sampling using samples generated from ULA, and under slightly stronger assumptions, SGLD, could still achieve the optimal regret guarantee with constant algorithmic as well as sample complexity in the stochastic gradient estimate. Thus, by designing approximate sampling algorithms specifically for use with Thompson sampling, we were able to construct a computationally tractable, anytime approximate Thompson sampling algorithm with end-to-end guarantees of logarithmic regret.

## References

M. Abeille and A. Lazaric. Linear Thompson sampling revisited. *Electron. J. Statist.*, 11(2):5165–5197, 2017.

S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, pages 39.1–39.26, 2012.

S. Agrawal and N. Goyal. Further optimal regret bounds for Thompson sampling. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 99–107, 2013a.

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 127–135. 2013b.

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

S. Basu and A. DasGupta. The mean, median, and mode of unimodal distributions: a characterization. *Teor. Veroyatnost. i Primenen.*, 41:336–352, 1996.

O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems 24 (NeurIPS)*, pages 2249–2257, 2011.

X. Cheng and P. L. Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of the 29th International Conference on Algorithmic Learning Theory (ALT)*, pages 186–211, 2018.

A. S. Dalalyan and A. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stoch. Process. and their Appl.*, 129(12):5278–5311, 2019.

J. L. Doob. Application of the theory of martingales. *Le Calcul des Probabilites et ses Applications*, pages 23–27, 1949.

A. Durmus and E. Moulines. Sampling from strongly log-concave distributions with the Unadjusted Langevin Algorithm. arXiv preprint, 2016.

A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 06 2017.

S. Ghosal and A. W. van der Vaart. Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35(1):192–223, 2007.

S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2): 500–531, 04 2000.

C. A. Gómez-Uribe. Online algorithms for parameter mean and variance estimation in dynamic regression. arXiv preprint, 2016.

A. Gopalan, S. Mannor, and Y. Mansour. Thompson sampling for complex online problems. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, pages 100–108, 2014.

C. Jin, R. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan. A short note on concentration inequalities for random vectors with subGaussian norm. arXiv preprint, 2019.

E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Proceedings of the 29th International Conference on Algorithmic Learning Theory (ALT)*, 2012.

N. Korda, E. Kaufmann, and R. Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems 26 (NeurRIPS)*, pages 1448–1456, 2013.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 6:4–22, 1985.

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

M. Ledoux. Concentration of measure and logarithmic Sobolev inequalities. In *Seminaire de probabilites XXXIII*, pages 120–216. Springer, 1999.

M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical surveys and monographs. American Mathematical Society, 2001.

X. Lu and B. Van Roy. Ensemble sampling. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 3260–3268, 2017.

Y.-A Ma, T. Chen, and E. Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems 28 (NeurIPS)*, pages 2917–2925, 2015.

Y.-A. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I. Jordan. Sampling can be faster than optimization. *Proc. Natl. Acad. Sci. U.S.A.*, 116(42):20881–20885, 2019.

W. Mou, N. Ho, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan. A diffusion process perspective on posterior contraction rates for parameters. arXiv preprint, 2019.

B. Øksendal. *Stochastic Differential Equations*. Springer, Berlin, 6th edition, 2003.

M. Phan, Y. A. Yadkori, and J. Domke. Thompson sampling and approximate inference. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 8804–8813, 2019.

Y. Ren. On the Burkholder-Davis-Gundy inequalities for continuous martingales. *Stat. Probabil. Lett.*, 78(17): 3034–3039, 2008.

C. Riquelme, Tucker G., and J. Snoek. Deep Bayesian bandits showdown: An empirical comparison of Bayesian deep networks for Thompson sampling. arXiv preprint, 2018.

D. Russo and B. Van Roy. An information-theoretic analysis of Thompson sampling. *J. Mach. Learn. Res.*, 17:1–30, 2016.

D. Russo, B. V. Roy, A. Kazerouni, and I. Osband. A tutorial on Thompson sampling. arXiv preprint, 2017.

A. Saumard and J. A. Wellner. Log-concavity and strong log-concavity: A review. *Statist. Surv.*, 8:45–114, 2014.

S. L. Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *Ann. Statist.*, 29(3):687–714, 06 2001.

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

I. Urteaga and C. Wiggins. Variational inference for the multi-armed contextual bandit. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 698–706, 2018.

A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, 36(3):1435–1463, 06 2008.

S. Vempala and A. Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 8094–8106, 2019.

C. Villani. *Optimal Transport: Old and New*. Wissenschaften. Springer, Berlin, 2009.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

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