Accelerating Approximate Thompson Sampling with Underdamped Langevin Monte Carlo

Haoyang Zheng
Purdue University

Wei Deng
Morgan Stanley

Christian Moya
Purdue University

Guang Lin
Purdue University

Abstract

Approximate Thompson sampling with Langevin Monte Carlo broadens its reach from Gaussian posterior sampling to encompass more general smooth posteriors. However, it still encounters scalability issues in high-dimensional problems when demanding high accuracy. To address this, we propose an approximate Thompson sampling strategy, utilizing underdamped Langevin Monte Carlo, where the latter is the go-to workhorse for simulations of high-dimensional posteriors. Based on the standard smoothness and log-concavity conditions, we study the accelerated posterior concentration and sampling using a specific potential function. This design improves the sample complexity for realizing logarithmic regrets from $\tilde{O}(d)$ to $\tilde{O}(\sqrt{d})$. The scalability and robustness of our algorithm are also empirically validated through synthetic experiments in high-dimensional bandit problems.

1 INTRODUCTION

Multi-armed bandit (MAB) problem is a classic decision-making dilemma to find the best option among multiple arms, which has been applied in various domains such as recommendation systems (May et al., 2012; Chapelle and Li, 2011), advertising (Graepel et al., 2010), etc. A fundamental challenge in the MAB problem is to balance exploration and exploitation, and Thompson sampling (TS) has proven to be an effective mechanism for addressing this, particularly appealing due to its simplicity and strong empirical performance. In TS, each arm is affiliated with a posterior distribution that approximates the model parameters. Arms with higher uncertainty in their approximate posterior may be selected to facilitate exploration. Post-selection, the observed rewards serve to iteratively update the corresponding posterior distribution, thereby improving the estimation and reducing variance (Russo and Van Roy, 2014). Thanks to recent extensive advancements in both empirical experimentation (Granmo, 2010) and theoretical formulations (Agrawal and Goyal, 2012; Russo and Van Roy, 2016), TS gained widespread recognition.

TS often models the posteriors with Laplace approximation for computational efficiency. However, such a simplification is computationally intensive and may inaccurately capture non-Gaussian posteriors (Huix et al., 2023). Given the limitations of the exclusive use of distributions with Laplace approximations, researchers are increasingly turning to Langevin Monte Carlo (LMC) as a robust method for exploring and exploiting the posterior distributions.

Overdamped Langevin Monte Carlo (LMC) in machine learning originated from stochastic gradient Langevin dynamics (SGLD) (Welling and Teh, 2011), which injects noise into stochastic gradient descent, evolving into a sampling algorithm as the learning rate decays. The theoretical guarantees on global optimization (Raginsky et al., 2017; Xu et al., 2018; Zhang et al., 2017) and uncertainty estimation (Teh et al., 2016) were further established, which provides specific guidance on the tuning of temperatures and learning rates. Despite the elegance, it is inevitable to suffer from scalability issues w.r.t. the dimension, accuracy target, and condition number. To tackle this issue, the go-to framework is the underdamped Langevin Monte Carlo (ULMC) (Chen et al., 2014; Cheng et al., 2018), which can be viewed as a variant of Hamiltonian Monte Carlo (HMC) (Neal, 2012; Hoffman and Gelman, 2014; Mangoubi and Vishnoi, 2018). The incorporation of momentum (or velocity) in the sampling process facilitates the exploration and results in more effective samples.

Our Contribution

In this work, we focus on enhancing the practicality and scalability of TS. The introduction of LMC to TS offers a potential route to avoid the Laplace approximation of the posterior distributions but falls short in high-dimensional settings when sample complexity is in high demand. Our primary contribution lies in the novel integration of ULMC

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

into the TS framework to address this issue.

We first analyze the posterior as the limiting distribution of stochastic differential equation (SDE) trajectories. Through a specific potential function, we delve into the analysis of SDE trajectories and provide a renewed perspective on posterior concentration rates. This analytical perspective further enables a more efficient posterior approximation and results in an enhancement of the sample complexity to realize logarithmic regrets from $\tilde{O}(d)$ to $\tilde{O}(\sqrt{d})$, where d represents the dimension of model parameters. Synthetic experiments across varied scenarios provide empirical results to further validate our theoretical statements and the robustness of our proposed algorithms.

2 RELATED WORKS

Thompson Sampling began with a primary focus on experimental studies (Russo et al., 2018) due to the theoretical challenges we introduced earlier. In recent MAB works, an in-depth analysis of TS from a theoretical perspective has become more prevalent across various reward and prior assumptions (Russo and Van Roy, 2016; Riou and Honda, 2020; Baudry et al., 2021). Recent works further explored TS with problem-independent bounds (Agrawal and Goyal, 2017), which was further improved to achieve minimax optimal performance (Jin et al., 2021, 2022, 2023). Its mechanism is straightforward when using closed-form posteriors, but in more general settings, the posterior sampling becomes challenging and requires further discussions. One challenge is the dependence of TS on prior quality, as improper priors may lead to insufficient explorations (Liu and Li, 2016). Another challenge is from approximate errors, specifically, small constant approximate errors may induce linear regrets. Drawing upon insights from (Phan et al., 2019), approximate TS with LMC achieves logarithmic regrets, with approximation errors diminishing over rounds (Mazumdar et al., 2020).

Beyond its well-known application in MABs, TS has broader implications in the domain of reinforcement learning. Specifically, it has been further validated in various contexts such as in combinatorial bandits (Wang and Chen, 2018; Perrault et al., 2020), contextual bandits (Agrawal and Goyal, 2013; Xu et al., 2022; Chakraborty et al., 2023), linear Markov Decision Processes (Ishfaq et al., 2024), and linear quadratic controls (Kargin et al., 2022), etc.

Langevin Monte Carlo becomes notable for convergence analysis in overdamped forms under strong log-concavity (Durmus and Éric Moulines, 2017), which achieves $O(d/\epsilon^2)$ complexity for ϵ error in 2-Wasserstein (W2) distance. This result was further extended to the stochastic gradient version (Dalalyan and Karagulyan, 2019). A similar convergence result, but based on the Total Variation metric, was achieved later (Dalalyan, 2017). Furthermore, LMC was shown to offer theoretical guarantees for handling multi-modal distributions in Raginsky et al. (2017).

To expedite convergence, Cheng et al. (2018) delved into the non-asymptotic convergence of the ULMC, revealing a computational complexity of $O(\sqrt{d}/\epsilon)$ to obtain ϵ error in W2. For the related Hamiltonian Monte Carlo, Mangoubi and Smith (2017) demonstrated that the convergence rate depends quadratically on the condition number. This result was significantly improved by Zongchen Chen (2019) using an elegant geometric approach, achieving a linear convergence rate. Further insights into the numerical analysis of the leapfrog scheme were provided by Chen et al. (2020) and Mangoubi and Vishnoi (2018). Some works have concentrated on improving sampling efficiency and accuracy by utilizing control variates (Zou et al., 2018a,b, 2019). For simulations of multi-modal distributions, the replica exchange Langevin Monte Carlo (Chen et al., 2019; Deng et al., 2020) ran multiple chains at different temperatures to balance exploration and exploitation; adaptive importance sampling algorithms (Deng et al., 2022b,a) were studied to simulate from an adaptively modified landscape to address the local trapping issue and further employed the importance weights to correct the bias.

3 PROBLEM SETTING

MABs are classical frameworks for modeling the trade-off between exploration and exploitation in sequential decision-making problems. Formally, MABs consist of a set of arms $\mathcal{A} = \{1, 2, \cdots, K\}$, each associated with an unknown reward distribution. At each round n ($n = 1, 2, \cdots, N$), the agent selects an arm $A_n \in \mathcal{A}$ and receives a random reward $\mathcal{R}_{a,n} \in \mathbb{R}$ sampled from the corresponding unknown reward distribution. Multiple rewards received from playing the same arm repeatedly are identically and independently distributed (i.i.d.) and are statistically independent of the rewards received from playing other arms.

Within N rounds, the objective of the TS algorithm is to identify the optimal arm that most likely incurs the lowest regrets $\Re(N)$:

$$\Re(N) = N\bar{\mathcal{R}}_1 - \sum_{n=1}^{N} \mathbb{E}[\mathcal{R}_{A_n,n}],$$

where $\bar{\mathcal{R}}_1$ denotes the expected reward of the first arm, and without loss of generality, we name the first arm as the optimal one. In the classical setting, the true posteriors of the arms are unknown and must be estimated through sampling. With each arm play, we sample a reward from its true posterior and then update the approximate posterior (denoted as μ_a) based on the given reward. The challenge lies in deciding which arm to play at each round, balancing the need to gather information about potentially suboptimal arms (exploration) against the desire to play the arm that currently seems best (exploitation). A general framework of TS for MABs is given in Algorithm 1:

Algorithm 1 A general framework of TS for MABs.

Input Bandit feature vectors α_a for $\forall a \in \mathcal{A}$.

for n=1 to N **do**
  Sample $(x_{a,n}, v_{a,n}) \sim \mu_a[\rho_a]$ for $\forall a \in \mathcal{A}$.
  Choose arm $A_n = \text{argmax}_{a\in\mathcal{A}} \langle \alpha_a, x_{a,n}\rangle$.
  Play arm A_n and receive reward $\mathcal{R}_n$.
  Update posterior distribution of arm A_n: $\mu_{A_n}[\rho_{A_n}]$.
  Calculate expected regret $\mathfrak{R}_n$.
end for

Output Expected total regret $\sum_{n=1}^{N} \mathfrak{R}_n$.

4 POSTERIOR ANALYSIS

The unique feature that distinguishes TS from frequentist methods (Sutton and Barto, 2018) is its use of posterior distributions to guide decision-making. This probabilistic sampling inherently incorporates both the mean and the uncertainty (including but not limited to variance) of our current beliefs about each arm. Therefore, understanding how these posteriors evolve over time becomes fundamental not only for explaining the algorithm's empirical efficiency but also for establishing its theoretical properties.

For the purposes of this study, we make the assumption that each arm's reward distribution is parameterized by a set of fixed parameters $x_{a,*} \in \mathbb{R}^d$, and the reward distribution $\mathrm{P}_a(\mathcal{R})$ is now parameterized: $\mathrm{P}_a(\mathcal{R}) = \mathrm{P}_a(\mathcal{R}, x_{a,*})$. To simplify our notation, we drop the arm-specific parameter a when discussing settings uniformly across arms in this section. To facilitate our analysis, we introduce a series of standard assumptions on both the reward distributions and distributions related to arms:

Assumption 1 (Local assumptions on the likelihood)**.** *For the function* $\log \mathrm{P}(\mathcal{R}|x)$*, its gradient w.r.t.* x *is L-Lipschitz and the function is m-strongly convex for constants* $L \geq 0$ *and* $m > 0$*. This means for all* $\mathcal{R} \in \mathbb{R}$ *and* $x, x_* \in \mathbb{R}^d$*, the following inequalities hold:*

$$\begin{aligned}&\langle \nabla_x \log \mathrm{P}(\mathcal{R}|x_*), x - x_*\rangle + \tfrac{m}{2}\|x - x_*\|_2^2 \\ &\leq \log \mathrm{P}(\mathcal{R}|x) - \log \mathrm{P}(\mathcal{R}|x_*) \leq \\ &\langle \nabla_x \log \mathrm{P}(\mathcal{R}|x_*), x - x_*\rangle + \tfrac{L}{2}\|x - x_*\|_2^2.\end{aligned}$$

Assumption 2 (Assumptions on the reward distribution)**.** *For* $\log \mathrm{P}(\mathcal{R}|x_*)$*, its gradient concerning* $\mathcal{R}$ *is L-Lipschitz and the function is* ν*-strongly convex, given constants* $L \geq 0$ *and* $\nu > 0$*. Consequently, For every* $\mathcal{R}, \mathcal{R}' \in \mathbb{R}$ *and* $x_* \in \mathbb{R}^d$*, we have the relationship:*

$$\begin{aligned}&\langle \nabla_{\mathcal{R}} \log \mathrm{P}(\mathcal{R}'|x_*), \mathcal{R} - \mathcal{R}'\rangle + \tfrac{\nu}{2}\|\mathcal{R} - \mathcal{R}'\|_2^2 \\ &\leq \log \mathrm{P}(\mathcal{R}|x_*) - \log \mathrm{P}(\mathcal{R}'|x_*) \leq \\ &\langle \nabla_{\mathcal{R}} \log \mathrm{P}(\mathcal{R}'|x_*), \mathcal{R} - \mathcal{R}'\rangle + \tfrac{L}{2}\|\mathcal{R} - \mathcal{R}'\|_2^2.\end{aligned}$$

Assumption 3 (Assumption on the prior)**.** *For the gradient of* $\pi(\mathcal{R}|x)$*, there exists a constant L such that for* $x, x' \in \mathbb{R}^d$*:*

$$\left\|\nabla\pi(x) - \nabla\pi(x')\right\|_2 \leq L\left\|x - x'\right\|_2.$$

Ignoring the difference across arms, this section presents theoretical investigations of the general convergence properties of the TS posterior distributions. We will focus on understanding the dynamics of the posterior distributions, and how ULMC samples from the posterior and approximates it. The main contribution regarding the sample complexity is also discussed subsequently.

4.1 Continuous-Time Diffusion Analysis

For the derivation of posterior concentration rates for parameters in terms of the accumulative number of likelihoods, we conduct an innovative analysis of the moments of a potential function along the trajectories of SDEs, where we consider the posterior as the limiting distribution given n rewards:

$$\begin{aligned}\mu[\rho] &= \lim_{t\to\infty} \mathrm{P}((x_t, v_t)|\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_n) \\ &\propto \exp\left(-\rho\left(nf_n(x) + n\|v\|_2^2/2u + \log\pi(x)\right)\right)\end{aligned}$$

almost surely. Here $x_t, v_t \in \mathbb{R}^d$ are the position and velocity terms over time $t > 0$, $f_n(x) = \frac{1}{n}\sum_{j=1}^{\mathcal{L}(n)} \log \mathrm{P}\left(\mathcal{R}_j|x\right)$ represents the average log-likelihood function and the limiting distribution is a scaled posterior distribution $\mu[\rho]$ with the scale parameter ρ. We denote here $\mathcal{L}(n)$ is the total number of rewards we have up to round n. By studying the evolution of the limiting joint distribution of x and v according to the following SDEs:

$$\begin{aligned}dv_t &= -\gamma v_t dt - u\nabla f_n(x_t)\, dt - \frac{u}{n}\nabla\log\pi(x_t)dt + 2\sqrt{\frac{\gamma u}{n\rho}}dB_t, \\ dx_t &= v_t dt\end{aligned} \tag{1}$$

as t goes to infinity, we are able to deduce the convergence rate of the scaled posterior distribution. Here γ is the friction coefficient, and u is the noise amplitude.

Posterior Concentration Analysis

Prior to presenting our findings on posterior concentration, we initially outline several fundamental points for achieving our results. Our derivation of the scaled posterior concentration $\mu[\rho]$ draws inspiration from Cheng et al. (2018), wherein a specific potential function is structured to infer the rates of posterior concentration. However, our focus is primarily on how the posterior concentrates around x_*, where x_* itself does not conform to SDEs. Therefore, we achieve the concentration rates by an extension of Grönwall's inequality in Dragomir (2003) rather than applying it directly. Moreover, the coefficients employed in constructing the potential function render $\gamma = 2$ and $u = \frac{1}{L}$ as the optimal selections. Subsequently, we clarify another additional essential property, indicating that, with high probability, $\|\nabla_x f_n(x_*)\|_2$ and the gradient of $\|v_*\|_2^2$ concentrating around zero, given the data $\mathcal{R}_1, \ldots, \mathcal{R}_n$. Specifically, our findings demonstrate that $\|\nabla_x f_n(x_*)\|_2$ exhibits sub-Gaussian tails, and we designate $\nabla_v\|v_*\|_2^2$ as zero to maintain simplicity.

Building upon the above analysis, we construct the following potential function given $\alpha > 0$:

$$V(x_t, v_t, t) = \frac{1}{2} e^{\alpha t} \|(x_t, x_t + v_t) - (x_*, x_* + v_*)\|_2^2,$$

which develops along the trajectories of SDEs (1). By delimiting the supremum of the given potential function, we obtain an upper bound for higher moments of $\|x - x_*\|$, where $x \sim \mu[\rho]$. These established moment bounds directly correspond to the posterior concentration rates of x around x_*, as illustrated in the following theorem:

Theorem 1. *Suppose that Assumptions 1 and 3 hold, and suppose $x \in \mathbb{R}^d$ follows SDEs (1), then for $x_* \in \mathbb{R}^d$ and $\delta \in \left(0, e^{-0.5}\right)$, the posterior satisfies:*

$$\mathbb{P}_{x\sim\mu[\rho]}\left(\|x - x_*\|_2 \geq \sqrt{\frac{2e}{mn}\left(D + 2\Omega \log\frac{1}{\delta}\right)}\right) \leq \delta,$$

with $D = \frac{8d}{\rho} + 2\log B$, $\Omega = 16\kappa^2 d + \frac{256}{\rho}$, $\kappa = \frac{L}{m}$ is the condition number, and $B = \max_x \frac{\pi(x)}{\pi(x_*)}$ represents prior quality.

We proceed to offer further insight from the theorem. The first term $\frac{8d}{\rho}$ originates from squaring the dB_t term in SDEs, which has a negligible influence on the results. The second term $2\log B$ is associated with the prior $\pi(x)$ and vanishes when $x_* = \text{argmax}[\pi(x)]$. Section 6 also provides extensive explorations into how the choice of prior affects TS performance. Term $16\kappa^2 d$ originates from the likelihood in SDEs, and the last term is associated with the noise present in SDEs, which substantially influences the contraction rate. Selecting a suitable value for ρ enables control of the posterior concentration scale. A detailed proof can be found in Appendix C.2, Theorem 6.

4.2 Discrete-Time Dynamics Analysis

We continue to delineate the methodology for employing ULMC to approximate posterior distributions by sampling positions and velocities. In essence, they provide a more efficient framework to draw samples sequentially based on current steps. Algorithm 2 outlines the execution of ULMC, detailing the sequential generation of samples for arm a. It initiates from the sample gathered at the last step of rounds $n-1$ and advances to the sample created for rounds n.

Within the ULMC framework, one can utilize either full gradients or stochastic gradients for the gradient estimation. The former is computed by summing all the gradients of likelihoods up to round n combined with the prior:

$$\nabla U(x_i) = -\sum_{j=1}^{\mathcal{L}(n)} \nabla \log \mathrm{P}\left(\mathcal{R}_j | x_i\right) - \nabla \log \pi(x_i),$$

which provide high-precision information for updates but might be computationally intensive. On the other hand,

Algorithm 2 (Stochastic Gradient) Underdamped Langevin Monte Carlo at round n.

Input Data $\{\mathcal{R}_1, \mathcal{R}_2, \cdots, \mathcal{R}_{\mathcal{L}(n-1)}\}$;
Input Sample $(x_{Ih^{(n-1)}}, v_{Ih^{(n-1)}})$ from last round;
Initialize $x_0 = x_{Ih^{(n-1)}}$ and $v_0 = v_{Ih^{(n-1)}}$.
for $i = 0, 1, \cdots, I-1$ **do**
  Subsample data set $\{\mathcal{R}_1, \cdots, \mathcal{R}_{|\mathcal{S}|}\} \subseteq \mathcal{S}$.
  Compute gradient estimate $\nabla \hat{U}(x_i)$.
  Sample (x_{i+1}, v_{i+1}) from (x_i, v_i).
end for
$x_{Ih^{(n)}} \sim \mathcal{N}\left(x_I, \frac{1}{nL\rho}\mathbf{I}_{d\times d}\right)$ and $v_{Ih^{(n)}} = v_I$
Output Sample $(x_{Ih^{(n)}}, v_{Ih^{(n)}})$ from current round.

stochastic gradients, approximated from a subset $\mathcal{S}$ of data:

$$\nabla \hat{U}(x_i) = -\frac{\mathcal{L}(n)}{|\mathcal{S}|} \sum_{\mathcal{R}_k \in \mathcal{S}} \nabla \log \mathrm{P}\left(\mathcal{R}_k | x_i\right) - \nabla \log \pi(x_i),$$

offer computational efficiency at the cost of precision. The choice between full or stochastic gradients depends on the specific requirements and constraints of the problem at hand, balancing between computational feasibility and approximation accuracy.

To transit consecutive samples from the index i to $i+1$, we followed by integrating the discrete underdamped Langevin dynamics up to h, which are given as follows:

$$\begin{bmatrix} x_{i+1} \\ v_{i+1} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbb{E}[x_{i+1}] \\ \mathbb{E}[v_{i+1}] \end{bmatrix}, \begin{bmatrix} \mathbb{V}_x & \mathbb{K}_{x,v} \\ \mathbb{K}_{v,x} & \mathbb{V}_v \end{bmatrix}\right). \tag{2}$$

Here the expectation of the position x_{i+1} and velocity v_{i+1} at step $i+1$ are given by the following equations:

$$\mathbb{E}[v_{i+1}] = v_i e^{-\gamma h} - \frac{u}{\gamma}(1 - e^{-\gamma h})\nabla U(x_i),$$
$$\mathbb{E}[x_{i+1}] = x_i + \frac{1}{\gamma}(1 - e^{-\gamma h})v_i - \frac{u}{\gamma}\left(h - \frac{1}{\gamma}\left(1 - e^{-\gamma h}\right)\right)\nabla U(x_i),$$

where $h < 1$ is the step size, and the choices of $\gamma = 2$ and $u = \frac{1}{L}$ maintain consistency. Moreover, the variance around the expected position and velocity at the step $i+1$, as well as the covariance between them, are encapsulated by the following equations:

$$\mathbb{V}_x = \frac{2u}{\gamma}\left[h - \frac{1}{2\gamma}e^{-2\gamma h} - \frac{3}{2\gamma} + \frac{2}{\gamma}e^{-\gamma h}\right] \cdot \mathbf{I}_{d\times d}$$
$$\mathbb{V}_v = u(1 - e^{-2\gamma h}) \cdot \mathbf{I}_{d\times d}$$
$$\mathbb{K}_{x,v} = \mathbb{K}_{v,x} = \frac{u}{\gamma}\left[1 + e^{-2\gamma h} - 2e^{-\gamma h}\right] \cdot \mathbf{I}_{d\times d}.$$

Following Ith iterations of the sampling process, we yield v_I directly. For x_I, we adopt a resampling strategy, adhering to a normal distribution with variance $\frac{1}{nL_a\rho_a}\mathbf{I}_{d\times d}$. A proper resampling step accelerates the mixing rate and facilitates our theoretical analysis.

Posterior Convergence Analysis

Sampling from an approximate posterior in TS demands careful control of approximation errors, as an improper approximation may lead to linear regrets (Phan et al., 2019). In pursuit of sublinear regrets, we employ the strategy from Mazumdar et al. (2020), which stipulates a $\tilde{O}\left(\frac{1}{n}\right)$ approximation error rate across rounds $n = 1, 2, \ldots, N$. To elaborate, W2 between the sample-generated measure $\hat{\mu}^{(n)}$ using ULMC and the posterior measure $\mu^{(n)}$ at each respective round should maintain $\tilde{O}\left(\frac{1}{n}\right)$. To attain such a rate, it is required to further assume that the log-likelihood is both globally strongly convex and Lipschitz smooth.

Assumption 4 (Global assumptions on the likelihood)**.** $\log \mathrm{P}(\mathcal{R}|x)$ *is characterized by its L-Lipschitz continuous gradient and its m-strong convexity, where* $L \geq 0$ *and* $m > 0$*. This means for all* $x, y \in \mathbb{R}^d$*, the following inequality holds:*

$$\begin{aligned}&\langle \nabla_x \log \mathrm{P}(\mathcal{R}|y), x - y\rangle + \tfrac{m}{2}\|x-y\|_2^2 \\ &\leq \log \mathrm{P}(\mathcal{R}|x) - \log \mathrm{P}(\mathcal{R}|y) \leq \\ &\langle \nabla_x \log \mathrm{P}(\mathcal{R}|y), x - y\rangle + \tfrac{L}{2}\|x-y\|_2^2.\end{aligned}$$

These prerequisites are imperative for the efficacious application of ULMC. In scenarios where the algorithm employs stochastic gradients, an additional prerequisite is the joint Lipschitz smoothness of the likelihood function concerning both rewards and parameters.

Assumption 5 (Further assumption on the likelihood)**.** *For the gradient of* $\log \mathrm{P}_a(\mathcal{R}|x)$*, there exists a constant* $L' \geq 0$[1] *such that for all* $x, y \in \mathbb{R}^d$*,* $\mathcal{R}, \mathcal{R}' \in \mathbb{R}$*, a stronger version of Lipschitz smooth condition holds:*

$$\left\|\nabla_x \log \mathrm{P}(\mathcal{R}|x) - \nabla_x \log \mathrm{P}(\mathcal{R}'|x)\right\|_2 \leq L\|x-y\|_2 + L'\left\|\mathcal{R} - \mathcal{R}'\right\|.$$

With this assumption and by wisely selecting the batch size for the number of stochastic gradient estimates, the algorithm can accomplish logarithmic regrets with a sample complexity of $\tilde{O}(\sqrt{d})$, indicative of the algorithm's efficacy in solving high-dimensional problems.

Theorem 2 (Convergence of Underdamped Langevin Monte Carlo)**.** *Assume that the likelihoods, rewards, and the prior satisfy Assumptions 2-5. If we take step size* $h = \tilde{O}\left(\frac{1}{\sqrt{d}}\right)$*, number of steps* $I = \tilde{O}\left(\sqrt{d}\right)$*, and batch size* $k = \tilde{O}\left(\kappa^2\right)$ *in Algorithm 2, we have convergence of ULMC in W2 to the posterior* $\mu^{(n)}$*:* $W_2\left(\hat{\mu}^{(n)}, \mu^{(n)}\right) \leq \frac{2}{\sqrt{n}}\hat{D}$*, where* $\hat{D} \geq 8\sqrt{\frac{d}{m}}$*.*

We focus on the convergence result for the stochastic gradient version, considering the full gradient approach as its more general counterpart. For deeper insights, readers can refer to Theorem 9 in Appendix E.2. Notably, the sample complexity $\tilde{O}(\sqrt{d})$ of our proposed algorithms outperforms the previous works with overdamped variations of $\tilde{O}(d)$, which is improved in terms of efficiency and reduced sample demands.

[1]For simplicity, we let the Lipschitz constants $L = L'$ to be consistent in our paper.

Posterior Concentration Analysis

The subsequent result provides guarantees for the convergence of samples generated from Algorithm 2. This theorem not only validates that the generated samples can approximate $\mu[\rho]$ well but also ensures the generated samples towards x_*.

Theorem 3. *Assume that the likelihood, rewards, and the prior satisfy Assumptions 2-5, and that arm a has been chosen* $\mathcal{L}_a(n)$ *times up to rounds n. We further state that the step size* $h^{(n)} = \tilde{O}\left(\frac{1}{\sqrt{d}}\right)$*, number of steps* $I = \tilde{O}\left(\sqrt{d}\right)$*, and batch size* $k = \tilde{O}\left(\kappa^2\right)$ *in Algorithm 2. Then for* $x \sim \hat{\mu}[\hat{\rho}]$ *follows (2),* $x_* \in \mathbb{R}^d$*, and* $\delta \in \left(0, e^{-0.5}\right)$*, the following concentration inequality holds:*

$$\mathbb{P}_{x\sim\hat{\mu}^{(n)}[\hat{\rho}]}\left(\|x - x_*\|_2 \geq \sqrt{\frac{36e}{mn}\left(D + 4\hat{\Omega}\log\frac{1}{\delta}\right)}\right) \leq \delta,$$

where $D = 8d + 2\log B$*,* $\hat{\Omega} = 16\kappa^2 d + 256 + \frac{d}{36\kappa\hat{\rho}}$*, and* $\hat{\mu}^{(n)}[\hat{\rho}]$ *serves as the scaled approximate posterior at round n, with the scale parameter* $\hat{\rho}$*, to approximate* $\mu[\rho]$*.*

In contrast to our results with the posterior concentration rates from Theorem 1, two notable distinctions are worthy of further discussion. Firstly, the magnitude increases from $\sqrt{\frac{2e}{mn}}$ to $\sqrt{\frac{36e}{mn}}$, which is an outcome of introducing approximation error into the posterior concentration analysis. Additionally, the term $\frac{\kappa d}{18\hat{\rho}}$ within $\hat{\Omega}$ serves to bound errors arising in the last resampling step of Algorithm 2. For an in-depth exploration, readers can be directed to E.3 in the Appendix.

5 REGRET ANALYSIS

We now turn our attention to analyzing the regrets incurred by TS when the posterior follows SDEs (1) or is approximated by ULMC (2). For the sake of clarity in analysis, we adopt the convention of treating the first arm as optimal, resulting in zero regrets upon its selection, while pulling other arms would lead to an increase in the expected regrets, without any loss of generality. A further discussion of the unique optimal arm setting can refer to Appendix A of Agrawal and Goyal (2012).

To facilitate our theoretical examination, we introduce several essential notations. Let $\mathcal{L}_a(n)$ represent the number of plays of the sub-optimal arm a after round n. We then define $\mathcal{F}_n = \sigma(A_1, \mathcal{R}_1, A_2, \mathcal{R}_2, A_3, \mathcal{R}_3, \cdots, A_n, \mathcal{R}_n)$ is a σ-algebra generated by Algorithm 1 after playing arms n times. We also denote an event $\mathcal{E}_a(n) = \left\{\mathcal{R}_{a,n} \geq \bar{\mathcal{R}}_1 - \epsilon\right\}$ to indicate the estimated reward of arm a at round n exceeds the expected

reward of optimal arm by at least a positive constant ϵ, and its probability is denoted as $\mathcal{G}_{a,n} = \mathrm{P}\left(\mathcal{E}_a(n)|\mathcal{F}_{n-1}\right)$. Lastly, the reward difference between the optimal arm and arm a is represented by $\Delta_a = \bar{\mathcal{R}}_1 - \bar{\mathcal{R}}_a$.

To understand the algorithm's behavior, we first establish an upper bound on the expected number of times the algorithm plays sub-optimal arm a ($a \in \mathcal{A}$, $a \neq 1$). It is clear from our setup that the regrets would increase only if we played suboptimal arms. As indicated by Agrawal and Goyal (2012); Lattimore and Szepesvári (2020), after playing sub-optimal arms extensively, their posteriors become precisely estimated, and they will no longer be played with high probability. Consequently, we can set upper bounds on their regrets. Based on the above analysis, the expected plays of suboptimal arm a are played after N rounds can be represented as $\mathbb{E}\left[\mathcal{L}_a(N)\right]$, which can be further separated into two parts:

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}_a(N)\right] &= \mathbb{E}\left[\sum_{n=1}^{N} \mathbb{I}\left(A_n = a, \mathcal{E}_a(n)\right) + \mathbb{I}\left(A_n = a, \mathcal{E}_a^c(n)\right)\right] \\ &= 1 + \mathbb{E}\left[\sum_{s=0}^{N-1} \mathbb{I}\left(\mathcal{G}_{a,s} > \frac{1}{N}\right)\right] + \mathbb{E}\left[\sum_{s=0}^{N-1}\left(\frac{1}{\mathcal{G}_{1,s}} - 1\right)\right].\end{aligned} \tag{3}$$

For a small positive ϵ, $\mathbb{E}\left[\sum \frac{1}{\mathcal{G}_{1,s}} - 1\right]$ reflects how closely the sample from the first arm at round n reach to its true posterior, which serves as a strong indicator favoring immediate exploitation of the first arm. However, the magnitude of this term diminishes with large variance, thereby suggesting a possible necessity for further exploration. The expectation of $\sum \mathbb{I}\left(\mathcal{G}_{a,s} > \frac{1}{N}\right)$, on the other hand, offers a contrasting perspective by evaluating the likelihood that the choice of arm a is nearly as optimal as the first arm, which functions as an effective metric for assessing the similarity between the selected arm and the optimal arm. A low value for this term, especially with low perturbation variance, indicates that the selected arm may not be worth further exploration. Therefore, achieving an optimal equilibrium between these two terms would enable more enlightened decision-making and keep the balance between exploration and exploitation.

We subsequently integrate the concentration findings from Section 4 with the exploration and exploitation terms in (3) to extract the relevant concentration guarantees.

5.1 Regrets for Exact Thompson Sampling

We first analyze the regrets of exact TS, where the posterior distribution is described by limiting distributions governed by SDEs (1). Within rounds $n = 1, 2, \ldots, N$, we establish a lower boundary for $\mathcal{G}_{1,n}$, which serves to establish a minimum threshold for the probability of the first arm being optimistic:

$$\mathbb{E}\left[\frac{1}{\mathcal{G}_{1,n}}\right] \leq C\sqrt{\kappa_1 B_1}.$$

This bound is dependent on the prior quality B_1 and the condition number κ_1 of the optimal arm. While it yields an almost sure concentration of the first arm's unscaled posterior to true posteriors when $\mathcal{L}_1(n)$ is sufficiently large, it may fail when $\mathcal{L}_1(n)$ is small. Intuitively, when $\mathcal{L}_1(n)$ is large, it implies that the obtained reward, $r_{1,n}$, tends to center around $\bar{r}_1$; however, a small $\mathcal{L}_a(n)$ can yield a reward falling below $\bar{r}_1 - \epsilon$. In our proof, setting ρ_1 as $\kappa_1^{-3}(8d)^{-1}$ requires the SDEs (1) to concentrate around a scaled posterior. This introduces a requisite variance in rewards to counterbalance any potential underestimation biases, thus yielding the highlighted anti-concentration result. The choice of ρ_1 is also constrained by the squaring term of the sub-Gaussian random variable $\|\nabla f_n(x_*)\|_2^2$. An overly large ρ_1 would also fail to derive the concentration bound.

Integrating the established anti-concentration bound with the conclusions in Theorem 1 enables more problem-dependent bounds of (3). Specifically, with the choice of $\rho_1 = \kappa_1^{-3}(8d)^{-1}$, we can further derive the following upper bounds in (3):

$$\mathbb{E}\left[\sum_{s=0}^{N-1}\left(\frac{1}{\mathcal{G}_{1,s}} - 1\right)\right] \leq \left\lceil \frac{C_1\sqrt{\kappa_1 B_1}}{m_1\Delta_a^2}\left(D_1 + \Omega_1\right)\right\rceil + 1,$$

$$\mathbb{E}\left[\sum_{s=0}^{N-1} \mathbb{I}\left(\mathcal{G}_{a,s} > \frac{1}{N}\right)\right] \leq \frac{C_2}{m_a\Delta_a^2}\left(D_a + \Omega_a\right),$$

where for all arms $a \in \mathcal{A}$, $D_a = \log B_a + d^2\kappa_a^3$, $\Omega_a = \kappa_a^2 d + \kappa_a^3 d$, and $C_1, C_2 > 0$ are constant which is not related to any parameter related to the algorithm.

Based on the above results to constrain the expected plays for the sub-optimal arms, we can now derive a bound for the total expected regrets of the proposed exact TS algorithm.

Theorem 4. *Suppose the likelihoods, rewards, and priors satisfy Assumptions 1-3. With the choice of $\rho_a = \kappa_a^{-3}(8d)^{-1}$, we have that the total expected regrets after $N > 0$ rounds of TS with exact sampling satisfies:*

$$\begin{aligned}\mathbb{E}[\Re(N)] \leq &\sum_{a>1}\frac{C_a}{\Delta_a}\left(\log B_a + d^2 + d\log N\right) \\ &+ \frac{C_1\sqrt{B_1}}{\Delta_a}\left(\log B_1 + d^2\right) + 2\Delta_a,\end{aligned}$$

where $C_1, C_a > 0$ are universal constants and are unaffected by parameters specific to the problem.

According to the results, the first term evaluates the probability of arm a being nearly optimal among all choices. Its logarithmic growth with increasing N highlights there is no need for further exploration for large N. The second term, $\sqrt{B_1}\left(\log B_1 + d^2\right)/\Delta_a$, estimates how closely the sample from the first arm in the nth round aligns with its true posterior, consistent across rounds. With a sufficient number of rounds, the proposed algorithm is inclined to select the

optimal arm and no further incur regrets. The proposed algorithms exhibit logarithmic regrets $\tilde{O}\left(\frac{\log N}{\Delta_a}\right)$, which is equivalent to the overdamped version of TS. Nonetheless, advancements in sample complexity denote an improvement over the previous works. Due to the considerable dependency of outcomes on the prior quality, further experimental analysis is conducted in Section 6 for the reliance of results on prior quality. For the in-depth derivation of the proof, one can refer to Appendices C.3 and C.4.

5.2 Regrets for Approximate Thompson Sampling

Building upon the regret analysis of exact TS and the concentration results for approximate posteriors produced by ULMC, we present the regret findings for approximate TS implemented with ULMC.

We first denote $\hat{\mathcal{G}}_{a,n} = \mathrm{P}\left(\mathcal{E}_a(n)|\mathcal{F}_{n-1}\right)$ as the distribution of samples from the approximate posterior $\hat{\mu}_a$ for arm a after n rounds. Consistent with the anti-concentration revelations of $\mathcal{G}_{a,n}$, we also provide concentration guarantees for $\hat{\mathcal{G}}_{a,n}$:

$$\mathbb{E}\left[\frac{1}{\hat{\mathcal{G}}_{1,n}}\right] \leq C\sqrt{B_1},$$

which uniquely do not depend on the condition number κ_1. This characteristic is due to the fact that in $\hat{\mathcal{G}}_{a,n}$, the generated sample $x_{a,Ih}$ conforms to a normal distribution, while in $\mathcal{G}_{a,n}$, a strongly-log concave characteristic is utilized for approximating normal distributions. As a result, the approximation introduces κ_a. Additionally, the choice of the scaled parameter, denoted as $\hat{\rho}_a = (8\kappa_a\Omega_a)^{-1}$, is strategically made to bound the moment generating function of the random variable $\left\|x_{a,Ih} - x_{a,*}\right\|_2^2$, where $\left\|x_{a,Ih} - x_{a,*}\right\|_2^2$ is denoted as the L2 norm between samples generated for TS and the point $x_{a,*}$. With the above information, we continue to derive a similar upper bound for the terms in (3):

$$\sum_{s=0}^{N-1}\mathbb{E}\left[\frac{1}{\hat{\mathcal{G}}_{1,s}} - 1\right] \leq \left\lceil \frac{C_1\sqrt{B_1}}{m_1\Delta_a^2}\left(\log B_1 + d\kappa_1^2\log N + d^2\kappa_1^2\right)\right\rceil + 1$$

$$\sum_{s=0}^{N-1}\mathbb{E}\left[\mathbb{I}\left(\hat{\mathcal{G}}_{a,s} > \frac{1}{N}\right)\right] \leq \frac{C_2}{m_a\Delta_a^2}\left(d + \log B_a + d^2\kappa_a^2\log N\right).$$

With these problem-dependent bounds, we reach the regret bounds for the proposed approximate TS algorithm:

Theorem 5. *Suppose the likelihoods, rewards, and priors satisfy Assumptions 2-5. Given $\hat{\rho}_a = (8\kappa_a\Omega_a)^{-1}$ and sampling schemes specified in Theorem 2, then the total expected regrets after N rounds of approximate TS conform to:*

$$\mathbb{E}[\Re(N)] \leq \sum_{a>1}\frac{\hat{C}_a}{\Delta_a}\left(d + \log B_a + d^2\kappa_a^2\log N\right) + \frac{\hat{C}_1\sqrt{B_1}}{\Delta_a}\left(\log B_1 + d\kappa_1^2\log N + d^2\kappa_1^2\right) + 4\Delta_a.$$

where $\hat{C}_1, \hat{C}_a > 0$ are universal constants that are independent of problem-dependent parameters.

Comparing the regrets provided in both exact and approximate TS, we observe that the first term still demonstrates a logarithmic rate at $\tilde{O}\left(\frac{\log N}{\Delta_a}\right)$. Interestingly, its dependence on the parameter dimensions increases from linear to quadratic. This change is attributed to the distance between the unscaled and scaled approximate posteriors when establishing the contraction results of the approximate posterior, which yields a dimension-related result. While the second term in exact TS is independent of N, the result in approximate TS behaves at a rate of $\tilde{O}\left(\frac{\sqrt{B_1}\log N}{\Delta_a}\right)$. This underscores the sensitivity of approximate TS to inferior prior quality.

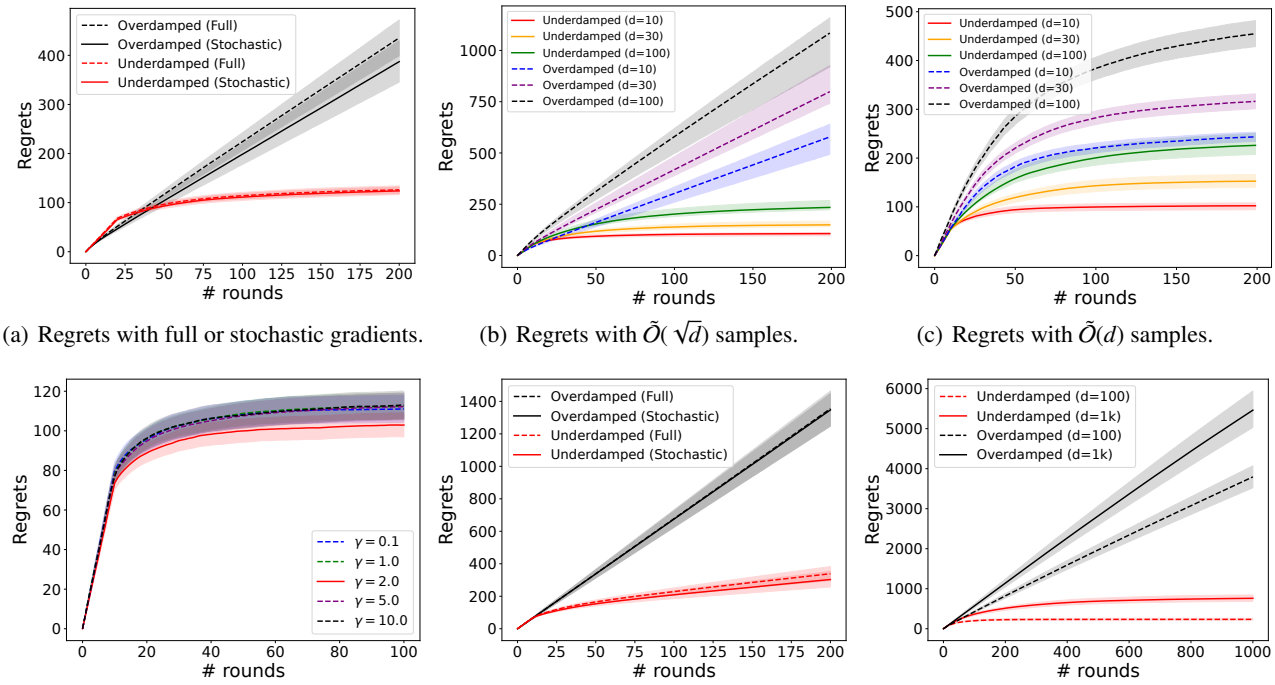
It should be noted that our proof achieves similar regrets as previous works (Mazumdar et al., 2020) but under less stringent sample complexity assumptions. With the same $\tilde{O}(\sqrt{d})$ samples, our method achieves logarithmic regrets, a guarantee absent in previous research. With $\tilde{O}(d)$ samples, our method accelerates posterior approximation and facilitates better convergence around the posterior mode, which indicates better regret bounds with the same $\tilde{O}(d)$ samples. Further discussion will be offered in the next section, and detailed theoretical derivations can be found in E.4.

6 EXPERIMENTS

In this section, we present numerical experiments designed to validate the benefits of the proposed methods on MAB from different perspectives[2]. We regard TS with LMC as a benchmark (both full gradient and stochastic gradient versions) across our tests. We first demonstrate general settings among all experiments without specific illustrations. For each round, the agent observes samples from ten-dimensional posterior distributions, and the agent has to choose action A_n between ten arms and receives rewards $\mathcal{R}_n$ to update the posterior of arm A_n. We generated 500 trajectories to get the expected mean of each result and then applied Bootstrapping to obtain their 95% confidence intervals. More details are provided in Appendix F.

We first compare the performance of the proposed algorithms and the benchmarks in Figure 1. Across all figures, the x-axis is the number of rounds to play arms, and the y-axis is the total expected regrets generated by different algorithms among various settings. Each line represents the average regrets among the generated trajectories, accompanied by a 95% confidence interval represented by a lighter shade. When we have the same step size and sample complexity $\tilde{O}(\sqrt{d})$, it is clear to see from Figure 1(a) that at a consistent sample complexity and step size, the proposed algorithm showcased logarithmic regrets while the overdamped algorithm incurred linear regrets. This disparity can be attributed to the overdamped version's inability to provide approximate error guarantees with large step sizes. Also, the figure indicates that a proper choice of batch size

[2]Code is available at the GitHub repository.



(a) Regrets with full or stochastic gradients. (b) Regrets with $\tilde{O}(\sqrt{d})$ samples. (c) Regrets with $\tilde{O}(d)$ samples.

(d) Regrets with different momentums γ. (e) Regrets with flat priors. (f) Regrets in high dimensional cases.

Figure 1: Regret comparisons of TS among different settings.

enables the proposed method to achieve logarithmic regrets relying only on stochastic gradients.

We continue to explore the derived regrets under different sample complexity and different dimension settings. Figure 1(b) demonstrates the regrets when we employ $\tilde{O}(\sqrt{d})$ samples and Figure 1(c) employs $\tilde{O}(d)$ samples. With $\tilde{O}(\sqrt{d})$ samples, it is clear that as the dimensions increase, there is a noticeable trend of slower regret convergence and higher regret magnitudes. Nevertheless, the algorithm persisted in delivering logarithmic regrets across all dimensional variations. On the contrary, the overdamped algorithm can only incur linear regrets. With $\tilde{O}(d)$ samples, the overdamped algorithm succeeds in achieving logarithmic regrets, but with the underdamped algorithm, the incurred regrets bounds are much tighter than the overdamped one.

We then further explore the stability of the proposed algorithm from different perspectives. We first explore the impact of various momentum (denoted by γ) values in Figure 1(d). Across different momentum values, the ULMC consistently displayed logarithmic regrets, which also indicates that our algorithms are stable among different momentum settings. It is worth noting that a gamma value of 2.0 incurs the smallest regrets, which aligns with our theoretical selection.

As we mentioned earlier in our analysis, the regrets may be heavily dependent on prior quality, and Figure 1(e) shows how the proposed method and benchmark behave when we consider flat priors. While the ULMC consistently outperformed the overdamped version, they are likely to incur linear regrets in smooth ascent. However, this milder linear regret growth hints at the algorithm's ability to avoid the worst decisions, even if it does not always identify the optimal ones.

Moreover, the advantages of our proposed algorithm over the conventional TS with LMC become significant in high-dimensional challenges, as depicted in Figure 1(f). Our approach exhibits much smaller regrets when dimensions increase to 100 and 1,000, which emphasizes the need for such advancements.

7 CONCLUSIONS AND DISCUSSIONS

To address the intricate task of sampling multiple approximate posteriors and guiding sequential decisions toward the optimal posteriors, we introduced a novel strategy using TS with ULMC to improve the approximation accuracy. The main contribution is the novel posterior analysis in the use of a specific potential function, which offers new insights into posterior concentration rates in TS. Based on this, the proposed algorithm offers a more favorable sample complexity $\tilde{O}(\sqrt{d})$ relative to the overdamped counterpart with $\tilde{O}(d)$ – a claim validated through both theoretical analysis and experimental validation.

Our algorithm demonstrates a consistent performance by

incurring logarithmic regrets with the sample complexity of both $\tilde{O}(\sqrt{d})$ and $\tilde{O}(d)$ comparable to that of the overdamped variant, which, by contrast, exhibits worse regret performance. Furthermore, the integration of variance reduction techniques, particularly control variates (Zou et al., 2018a, 2019), into the approximate TS algorithm with stochastic gradients is worthy of additional discussion here. These techniques have proven to significantly diminish noise variance in stochastic gradients, which contributes to accelerating sampling convergence. Consistent with the prior theoretical framework outlined in Mazumdar et al. (2020), our results also imply that, with variance reduction techniques, the sample complexity of approximate TS can be improved with a better constant, but not necessarily with a better order.

The robustness of the proposed algorithm was further supported by its consistent performance across various experiments: Our algorithms consistently perform well across a wide range of momentum values, which is consistent with our theory. In the face of challenges from the curse of dimensionality, it tends towards near-optimal arms and maintains logarithmic regret convergence. Its consistent and excellent performance becomes significant when comparing it and TS with LMC in high-dimensional settings. We also provide experiments to compare algorithms considering non-informative priors, but for our theoretical analysis, we highlight that we assume sufficiently good priors to target logarithmic regrets. It should be noted that current literature on prior sensitivity focuses more on simple cases and does not align with our settings. The impact of considering non-informative priors is an important aspect we aim to explore in future work.

This study has tentatively shown that our algorithms can align with theoretical regret bounds under weaker sample complexity assumptions and empirical evidence suggests that our algorithms could potentially derive tighter regret bounds. However, it is critical to recognize that we fall short of offering a theoretical guarantee for achieving consistently tighter regret bounds with equivalent sample complexities. Additionally, our method does not reach the minimax optimal regret benchmarks established in simpler TS settings (Jin et al., 2021). Our subsequent research will aim to develop a tight, optimal regret bound specifically to the approximate TS algorithm for MAB problems.

Another interesting topic for approximate TS is the extension to non-convex scenarios, which is an important but non-trivial future direction. With the dissipative or log-Sobolev assumptions, we aim to achieve exponential convergence in continuous time. Achieving this would not only validate the approach theoretically but also enhance the practical deployment of approximate TS across more diverse settings, which marks a substantial leap forward in the domain of TS and MCMC.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. G. Lin acknowledges the support of the National Science Foundation (DMS-2053746, DMS-2134209, ECCS-2328241, and OAC-2311848), and U.S. Department of Energy (DOE) Office of Science Advanced Scientific Computing Research program DE-SC0023161, and DOE–Fusion Energy Science, under grant number: DE-SC0024583.

References

Agrawal, S. and Goyal, N. (2012). Analysis of Thompson Sampling for the Multi-Armed Bandit Problem. In *Proc. of Conference on Learning Theory (COLT)*, pages 39–1. JMLR Workshop and Conference Proceedings.

Agrawal, S. and Goyal, N. (2013). Thompson Sampling for Contextual Bandits with Linear Payoffs. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 127–135. PMLR.

Agrawal, S. and Goyal, N. (2017). Near-Optimal Regret Bounds for Thompson Sampling. *Journal of the ACM*, 64(5):1–24.

Basu, S. and DasGupta, A. (1997). The Mean, Median, and Mode of Unimodal Distributions: a Characterization. *Theory of Probability & Its Applications*, 41(2):210–223.

Baudry, D., Gautron, R., Kaufmann, E., and Maillard, O. (2021). Optimal Thompson Sampling Strategies for Support-Aware CVaR Bandits. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 716–726. PMLR.

Boas, M. L. (2006). *Mathematical Methods in the Physical Sciences*. John Wiley & Sons.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford university press.

Brezis, H. and Brézis, H. (2011). *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, volume 2. Springer.

Chakraborty, S., Roy, S., and Tewari, A. (2023). Thompson Sampling for High-Dimensional Sparse Linear Contextual Bandits. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 3979–4008. PMLR.

Chapelle, O. and Li, L. (2011). An Empirical Evaluation of Thompson Sampling. *Advances in Neural Information Processing Systems (NeurIPS)*, 24.

Chen, T., Fox, E. B., and Guestrin, C. (2014). Stochastic Gradient Hamiltonian Monte Carlo. In *Proc. of the International Conference on Machine Learning (ICML)*.

Chen, Y., Chen, J., Dong, J., Peng, J., and Wang, Z. (2019). Accelerating Nonconvex Learning via Replica Exchange Langevin Diffusion. In *Proc. of the International Conference on Learning Representation (ICLR)*.

Chen, Y., Dwivedi, R., Wainwright, M. J., and Yu, B. (2020). Fast Mixing of Metropolized Hamiltonian Monte Carlo: Benefits of Multi-step Gradients. *Journal of Machine Learning Research (JMLR)*, 21:92–1.

Cheng, X., Chatterji, N. S., Bartlett, P. L., and Jordan, M. I. (2018). Underdamped Langevin MCMC: A Non-Asymptotic Analysis. In *Proc. of Conference on Learning Theory (COLT)*, pages 300–323. PMLR.

Dalalyan, A. S. (2017). Theoretical Guarantees for Approximate Sampling from Smooth and Log-concave Densities . *Journal of the Royal Statistical Society: Series B*, 79(3):651–676.

Dalalyan, A. S. and Karagulyan, A. (2019). User-Friendly Guarantees for the Langevin Monte Carlo with Inaccurate Gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311.

Deng, W., Feng, Q., Gao, L., Liang, F., and Lin, G. (2020). Non-Convex Learning via Replica Exchange Stochastic Gradient MCMC. In *Proc. of the International Conference on Machine Learning (ICML)*.

Deng, W., Liang, S., Hao, B., Lin, G., and Liang, F. (2022a). Interacting Contour Stochastic Gradient Langevin Dynamics. In *Proc. of the International Conference on Learning Representation (ICLR)*.

Deng, W., Lin, G., and Liang, F. (2022b). An Adaptively Weighted Stochastic Gradient MCMC Algorithm for Monte Carlo Simulation and Global Optimization. *Statistics and Computing*, pages 32–58.

Dragomir, S. S. (2003). Some Gronwall Type Inequalities and Applications. *Science Direct Working Paper*, .(S1574-0358):04.

Durmus, A. and Éric Moulines (2017). Non-asymptotic Convergence Analysis for the Unadjusted Langevin Algorithm. *Annals of Applied Probability*, 27:1551–1587.

Durmus, A. and Moulines, É. (2016). Sampling from a Strongly Log-Concave Distribution with the Unadjusted Langevin Algorithm. Preliminary version.

Graepel, T., Candela, J. Q. n., Borchert, T., and Herbrich, R. (2010). Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine. In *Proc. of the International Conference on Machine Learning (ICML)*, page 13–20. Omnipress.

Granmo, O.-C. (2010). Solving Two-Armed Bernoulli Bandit Problems Using a Bayesian Learning Automaton. *International Journal of Intelligent Computing and Cybernetics*, 3(2):207–234.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research (JMLR)*, pages 1593–1623.

Huix, T., Zhang, M., and Durmus, A. (2023). Tight Regret and Complexity Bounds for Thompson Sampling via Langevin Monte Carlo. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 8749–8770. PMLR.

Ishfaq, H., Lan, Q., Xu, P., Mahmood, A. R., Precup, D., Anandkumar, A., and Azizzadenesheli, K. (2024). Provable and Practical: Efficient Exploration in Reinforcement Learning via Langevin Monte Carlo. In *Proc. of the International Conference on Learning Representation (ICLR)*.

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. (2019). A Short Note on Concentration Inequalities for Random Vectors with Subgaussian Norm. *arXiv preprint arXiv:1902.03736*.

Jin, T., Xu, P., Shi, J., Xiao, X., and Gu, Q. (2021). MOTS: Minimax Optimal Thompson Sampling. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 5074–5083. PMLR.

Jin, T., Xu, P., Xiao, X., and Anandkumar, A. (2022). Finite-Time Regret of Thompson Sampling Algorithms for Exponential Family Multi-armed Bandits. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:38475–38487.

Jin, T., Yang, X., Xiao, X., and Xu, P. (2023). Thompson Sampling with Less Exploration Is Fast and Optimal. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 15239–15261. PMLR.

Kargin, T., Lale, S., Azizzadenesheli, K., Anandkumar, A., and Hassibi, B. (2022). Thompson Sampling Achieves $\tilde{O}(\sqrt{T})$ Regret in Linear Quadratic Control. In *Proc. of Conference on Learning Theory (COLT)*, pages 3235–3284. PMLR.

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

Ledoux, M. (2001). *The Concentration of Measure Phenomenon*. American Mathematical Soc.

Ledoux, M. (2006). Concentration of Measure and Logarithmic Sobolev Inequalities. In *Seminaire de probabilites XXXIII*, pages 120–216. Springer.

Liu, C.-Y. and Li, L. (2016). On the Prior Sensitivity of Thompson Sampling. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 321–336. Springer.

Mangoubi, O. and Smith, A. (2017). Rapid Mixing of Hamiltonian Monte Carlo on Strongly Log-Concave Distributions. *arXiv preprint arXiv:1708.07114*.

Mangoubi, O. and Vishnoi, N. K. (2018). Dimensionally Tight Running Time Bounds for Second-Order Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems (NeurIPS)*.

May, B. C., Korda, N., Lee, A., and Leslie, D. S. (2012). Optimistic Bayesian Sampling in Contextual-Bandit Problems. *Journal of Machine Learning Research (JMLR)*, 13:2069–2106.

Mazumdar, E., Pacchiano, A., Ma, Y., Jordan, M., and Bartlett, P. (2020). On Approximate Thompson Sampling with Langevin Algorithms. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 6797–6807. PMLR.

Mou, W., Ho, N., Wainwright, M. J., Bartlett, P., and Jordan, M. I. (2019). A Diffusion Process Perspective on Posterior Contraction Rates for Parameters. *arXiv preprint arXiv:1909.00966*.

Neal, R. M. (2012). MCMC Using Hamiltonian Dynamics. In *Handbook of Markov Chain Monte Carlo*, volume 54, pages 113–162. .

Pavliotis, G. A. (2014). *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*, volume 60. Springer.

Perrault, P., Boursier, E., Valko, M., and Perchet, V. (2020). Statistical Efficiency of Thompson Sampling for Combinatorial Semi-Bandits. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:5429–5440.

Phan, M., Abbasi Yadkori, Y., and Domke, J. (2019). Thompson Sampling and Approximate Inference. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Raginsky, M., Rakhlin, A., and Telgarsky, M. (2017). Nonconvex Learning via Stochastic Gradient Langevin Dynamics: a Nonasymptotic Analysis. In *Proc. of Conference on Learning Theory (COLT)*.

Ren, Y.-F. (2008). On the Burkholder-Davis-Gundy Inequalities for Continuous Martingales. *Statistics & probability letters*, 78(17):3034–3039.

Riou, C. and Honda, J. (2020). Bandit Algorithms Based on Thompson Sampling for Bounded Reward Distributions. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 777–826. PMLR.

Russo, D. and Van Roy, B. (2014). Learning to Optimize Via Posterior Sampling. *Mathematics of Operations Research*, 39(4):1221–1243.

Russo, D. and Van Roy, B. (2016). An Information-Theoretic Analysis of Thompson Sampling. *Journal of Machine Learning Research (JMLR)*, 17(1):2442–2471.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z., et al. (2018). A Tutorial on Thompson Sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96.

Saumard, A. and Wellner, J. A. (2014). Log-Concavity and Strong Log-Concavity: a Review. *Statistics surveys*, 8:45.

Silvester, J. R. (2000). Determinants of Block Matrices. *The Mathematical Gazette*, 84(501):460–467.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT press.

Teh, Y. W., Thiery, A., and Vollmer, S. (2016). Consistency and Fluctuations for Stochastic Gradient Langevin Dynamics. *Journal of Machine Learning Research (JMLR)*, 17:1–33.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press.

Wang, S. and Chen, W. (2018). Thompson Sampling for Combinatorial Semi-Bandits. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 5114–5122. PMLR.

Welling, M. and Teh, Y. W. (2011). Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 681–688.

Xu, P., Chen, J., Zou, D., and Gu, Q. (2018). Global Convergence of Langevin Dynamics Based Algorithms for Nonconvex Optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xu, P., Zheng, H., Mazumdar, E. V., Azizzadenesheli, K., and Anandkumar, A. (2022). Langevin Monte Carlo for Contextual Bandits. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 24830–24850. PMLR.

Yosida, K. and Taylor, J. (1967). Functional Analysis. *Physics Today*, 20(1):127–129.

Zhang, Y., Liang, P., and Charikar, M. (2017). A Hitting Time Analysis of Stochastic Gradient Langevin Dynamics. In *Proc. of Conference on Learning Theory (COLT)*, pages 1980–2022.

Zongchen Chen, S. S. V. (2019). Optimal Convergence Rate of Hamiltonian Monte Carlo for Strongly Logconcave Distributions. In *Approximation, Randomization, and Combinatorial Optimization*.

Zou, D., Xu, P., and Gu, Q. (2018a). Stochastic Variance-Reduced Hamilton Monte Carlo Methods. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 6028–6037. PMLR.

Zou, D., Xu, P., and Gu, Q. (2018b). Subsampled Stochastic Variance-Reduced Gradient Langevin Dynamics. In *Proc. of the Conference on Uncertainty in Artificial Intelligence (UAI)*.

Zou, D., Xu, P., and Gu, Q. (2019). Stochastic Gradient Hamiltonian Monte Carlo Methods with Recursive Variance Reduction. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/Not Applicable]
   Notes: sections 3, 4, 4.2
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable]
   Notes: section 4
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]
   Notes: this is theoretical work with just simple empirical validation

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [Yes/No/Not Applicable]
   (c) Clear explanations of any assumptions. [Yes/No/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable]
   Notes: theoretical work with simple validation.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/Not Applicable]
   (b) The license information of the assets, if applicable. [Yes/No/Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable]
   (d) Information about consent from data providers/curators. [Yes/No/Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you included:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/Not Applicable]

Supplementary Materials

A INTRODUCTION TO THOMPSON SAMPLING

Before we move on to the detailed proof, we here remind the general framework for the Thompson sampling algorithm. Thompson sampling is a probabilistic approach used for solving the multi-armed bandit problem, emphasizing the trade-off between the exploration of less-known arms and the exploitation of the best-known arms. As outlined in Algorithm 3, for each round, the algorithm samples from the scaled posterior distributions of each arm and then chooses the arm with the highest probability of getting optimal rewards. Once an arm is played and the reward is observed, the associated posterior distribution is updated. Through continuous updates and sampling, it enables the algorithm to quantify its performance in terms of total expected regrets.

Algorithm 3 A general framework of Thompson sampling for multi-armed bandits.

Input Bandit feature vectors α_a for $\forall a \in \mathcal{A}$.
Input Scaled posterior distribution $\mu_a[\rho_a]$ for $\forall a \in \mathcal{A}$.
for n=1 to N **do**
  Sample $(x_{a,n}, v_{a,n}) \sim \mu_a[\rho_a]$ for $\forall a \in \mathcal{A}$.
  Choose arm $A_n = \text{argmax}_{a\in\mathcal{A}} \langle \alpha_a, x_{a,n} \rangle$.
  Play arm A_n and receive reward $\mathcal{R}_n$.
  Update posterior distribution of arm A_n: $\mu_{A_n}[\rho_{A_n}]$.
  Calculate expected regrets $\mathfrak{R}_n$.
end for
Output Total expected regrets $\sum_{n=1}^{N} \mathfrak{R}_n$.

B PROOF OUTLINE

To make the reader easier to follow the proof, we outline the general idea of our proof as follows:

1. Decompose regrets into exploration and exploitation terms (Section C.3).
2. Derive the posterior concentration rates (Section C.2).
3. Apply posterior concentration rates to bound regret terms in the first step (Section C.4).
4. Determine sample size to attain $\tilde{O}(1/n)$ approximate error rates (Section E.2).
5. With approximate error guarantee, derive the ULMC-generated sample concentration rates (Section E.3).
6. Apply approximate posterior concentration rates to bound regret terms in the first step (Section E.4).

C ANALYSIS OF EXACT THOMPSON SAMPLING

C.1 Notation

Table 1 presents the key symbols used throughout our study on the posterior concentration analysis and the regret analysis of exact Thompson sampling, which will repeatedly appear in the following theoretical analysis. The choice of these notations helps clarity and consistency in our analyses and discussions. Readers are encouraged to refer to the table for a comprehensive understanding of the symbols employed.

Table 1: Notation summary in the analysis of exact Thompson sampling.

Symbols	Explanations
$[\![A, B]\!]$	set of all integers from A to B ($A, B \in \mathbb{Z}$)
d	model parameter dimensions
t	time horizon $t \in [0, T)$
n	number of rounds to play arm ($n \in [\![1, N]\!]$)
a	arm index in the Thompson sampling ($a \in \mathcal{A}$)
A_n	action (the selected arm) in round n
$\mathcal{R}_{a,n}$	reward in round n by pulling arm a
$\bar{\mathcal{R}}_a$	expected reward for arm a
$\Re(n)$	total expected regret in round n
x_* (v_*)	fixed position (velocity)
x_a (v_a)	inherent position (velocity) parameter for arm a
$x_{a,n}$ ($v_{a,n}$)	sampled position (velocity) parameter in round n for arm a
L_a	Lipschitz smooth constant for arm a
m_a	strongly convex constant on the likelihood for arm a
ν_a	strongly convex constant on the reward for arm a
ρ_a	parameter to scale posterior for arm a
κ_a	condition number for the likelihood of arm a: $\kappa_a = \frac{L_a}{m_a}$
B_a	prior quality of arm a: $B_a = \frac{\max_x \pi_a(x)}{\pi_a(x_*)}$
γ	friction coefficient
u	noise amplitude
α_a	bandit inherent parameters for arm a
ω_a	norm of α for arm a
Δ_a	distance between expected rewards of optimal arm and arm a
$\mu_a^{(n)}$	probability measure of posterior distribution of arm a after n rounds
$\mu_a^{(n)}[\rho_a]$	probability measure of scaled posterior distribution of arm a after n rounds

C.2 Posterior Concentration Analysis

Suppose the posterior distribution $\mu_a^{(n)}[\rho_a] \propto \exp\left(-\rho_a\left(nf_n(x_a) + \log\pi(x_a) + n\|v_a\|_2^2/2u\right)\right)$, where $x_a, v_a \in \mathbb{R}^d$ represent the vector for position and velocity. Then the posterior distribution satisfies the following theorem.

Theorem 6. *Suppose that the likelihood and the prior follow Assumptions 1 and 3 hold. Given rewards $\mathcal{R}_{a,1}, \mathcal{R}_{a,2}, \ldots, \mathcal{R}_{a,n}$ then for $x_a, v_a \in \mathbb{R}^d$ and $\delta_1 \in \left(0, e^{-0.5}\right)$, the posterior distribution satisfies:*

$$\mathbb{P}_{x_a \sim \mu_a\left[\rho_a^{(n)}\right]}\left(\|x_a - x_*\|_2 \geq \sqrt{\frac{2e}{m_a n}\left(D_a + 2\Omega_a \log 1/\delta_1\right)}\right) \leq \delta_1 \tag{4}$$

with $D_a = \frac{8d}{\rho_a} + 2\log B_a$, $\Omega_a = \frac{16L_a^2 d}{m\nu_a} + \frac{256}{\rho_a}$. Here $x_* \in \mathbb{R}^d$ represents the fixed pair toward which the posterior distribution tends to concentrate.

Proof. The proof incorporates the strategies utilized in establishing Theorem 1 from Mou et al. (2019) and Theorem 5 from Cheng et al. (2018). Throughout this posterior concentration proof, we generalize the conclusion across all arms. For simplicity and clarity, we omitted the arm-specific subscript in this proof. We first define the following stochastic differential equations (SDEs):

$$\begin{aligned} dv_t &= -\gamma v_t dt - u\nabla f(x_t)\,dt - \frac{u}{n}\nabla \log \pi(x_t)dt + 2\sqrt{\frac{\gamma u}{n\rho}}dB_t, \\ dx_t &= v_t dt, \end{aligned} \tag{5}$$

where $x_t, v_t \in \mathbb{R}^d$ are position and velocity at time t, $f_n(x) \coloneqq \frac{1}{n}\sum_{j=1}^{n} \log \mathrm{P}(\mathcal{R}_j|x)$ is the log-likelihood function (we omit the subscript n in the following parts for simplicity), $\log \pi(x)$ the prior, γ is the friction coefficient, u is the noise amplitude, and B_t is a standard Wiener process (Brownian motion). Using a chosen value of $\alpha > 0$, we define a potential function $V : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^+$:

$$V(x, v, t) = \frac{1}{2}e^{\alpha t}\left\|(x, x+v) - (x_*, x_*+v_*)\right\|_2^2. \tag{6}$$

Importantly, as t approaches infinity, the p-th moments of $V(x_t, v_t, t)$ can translate to the corresponding moments of $V(x, v, t)$, due to the convergence of (x_t, v_t) to $(x, v) \sim \mu[\rho]$. Applying Itô's Lemma (Pavliotis, 2014) to $V(x_t, v_t, t)$, we can decompose the potential function as follows:

$$\begin{aligned} V(x_s, v_s, t) &= \frac{\alpha}{2}\int_0^t e^{\alpha s}\left\|x_s - x_*\right\|_2^2 + e^{\alpha s}\left\|(x_s+v_s) - (x_*+v_*)\right\|_2^2 ds \\ &\quad + \int_0^t e^{\alpha s}\left[\left\langle (1-\gamma)v_s - u\nabla f(x_s) - \frac{u}{n}\nabla \log \pi(x_t), (x_s+v_s) - (x_*+v_*)\right\rangle + \langle v_s, x_s - x_*\rangle\right] ds \\ &\quad + \frac{2\gamma u d}{n\rho}\int_0^t e^{\alpha s} ds + 2\sqrt{\frac{\gamma u}{n\rho}}\int_0^t e^{\alpha s}\langle (x_s+v_s) - (x_*+v_*), dB_s\rangle \\ &\overset{(i)}{=} \frac{\alpha}{2}\int_0^t e^{\alpha s}\left\|x_s - x_*\right\|_2^2 + e^{\alpha s}\left\|(x_s+v_s) - (x_*+v_*)\right\|_2^2 ds \\ &\quad + \int_0^t e^{\alpha s}\underbrace{\left[\langle (1-\gamma)(v_s - v_*) - u\left(\nabla f(x_s) - \nabla f(x_*)\right), (x_s+v_s) - (x_*+v_*)\rangle + \langle v_s - v_*, x_s - x_*\rangle\right]}_{T1} ds \\ &\quad + \underbrace{\frac{2\gamma u d}{n\rho}\int_0^t e^{\alpha s} ds}_{T2} + \underbrace{2\sqrt{\frac{\gamma u}{n\rho}}\int_0^t e^{\alpha s}\langle (x_s+v_s) - (x_*+v_*), dB_s\rangle}_{T3} \\ &\quad + \underbrace{\int_0^t e^{\alpha s}\left\langle (1-\gamma)v_* - u\nabla f(x_*), (x_s+v_s) - (x_*+v_*)\right\rangle ds}_{T4} \\ &\quad + \frac{u}{n}\underbrace{\int_0^t e^{\alpha s}\left\langle -\nabla \log \pi(x_t), (x_s+v_s) - (x_*+v_*)\right\rangle ds}_{T5}, \end{aligned} \tag{7}$$

where (i) holds by adding and subtracting terms related to $\nabla f(x_*)$ and v_*. Next, we proceed to bound the terms in (7) step-by-step. We start by bounding $T1$ as follows:

$$\begin{aligned} T1 &= \langle (1-\gamma)(v_s - v_*) - u\left(\nabla f(x_s) - \nabla f(x_*)\right), (x_s+v_s) - (x_*+v_*)\rangle + \langle v_s - v_*, x_s - x_*\rangle \\ &\overset{(i)}{\leq} -\frac{m}{2L}\left(\left\|x_s - x_*\right\|_2^2 + \left\|(x_s+v_s) - (x_*+v_*)\right\|_2^2\right), \end{aligned} \tag{8}$$

where the elaborate procedure to derive (i) can be found in Lemma 5. $T2$ can be easily obtained by integration rules:

$$T2 = \frac{2\gamma u d}{n\rho}\int_0^t e^{\alpha s} ds = \frac{2\gamma u d}{n\rho\alpha}\left(e^{\alpha t} - 1\right) \leq \frac{2\gamma u d}{n\rho\alpha}e^{\alpha t}. \tag{9}$$

We proceed to bound $T3$. Suppose $M_t = \int_0^t e^{\alpha s} \langle (x_s + v_s) - (x_* + v_*), dB_s \rangle$, and $T3$ can be rewritten as $2\sqrt{\frac{\gamma u}{n\rho}} M_t$. Then we can derive a bound for M_t as follows:

$$
\begin{aligned}
\mathbb{E}\left[\sup_{0\le t\le T} |M_t|^p\right] &\overset{(i)}{\le} (8p)^{\frac{p}{2}}\, \mathbb{E}\left[\langle M_t, M_t\rangle_T^{\frac{p}{2}}\right] \\
&\overset{(ii)}{\le} (8p)^{\frac{p}{2}}\, \mathbb{E}\left[\left(\int_0^T e^{2\alpha s} \|(x_s + v_s) - (x_* + v_*)\|_2^2\, ds\right)^{\frac{p}{2}}\right] \\
&= (8p)^{\frac{p}{2}}\, \mathbb{E}\left[\left(\int_0^T e^{2\alpha s} \|(x_s + v_s) - (x_* + v_*)\|_2^2\, ds\right)^{\frac{p}{2}}\right] \\
&\overset{(iii)}{\le} (8p)^{\frac{p}{2}}\, \mathbb{E}\left[\left(\sup_{0\le t\le T} \frac{1}{2} e^{\alpha t} \|(x_t + v_t) - (x_* + v_*)\|_2^2 \int_0^T 2e^{\alpha s} ds\right)^{\frac{p}{2}}\right] \\
&= (8p)^{\frac{p}{2}}\, \mathbb{E}\left[\left(\sup_{0\le t\le T} \frac{1}{2} e^{\alpha t} \|(x_t + v_t) - (x_* + v_*)\|_2^2 \frac{2e^{\alpha T} - 2}{\alpha}\right)^{\frac{p}{2}}\right] \\
&\overset{(iv)}{\le} \left(\frac{16pe^{\alpha T}}{\alpha}\right)^{\frac{p}{2}} \mathbb{E}\left[\left(\sup_{0\le t\le T} \frac{1}{2} e^{\alpha t} \|(x_t + v_t) - (x_* + v_*)\|_2^2\right)^{\frac{p}{2}}\right],
\end{aligned}
\tag{10}
$$

where (i) is bounded by Burkholder-Davis-Gundy inequalities (Theorem 2 in Ren (2008)); (ii) follows quadratic variation of Itô's Lemma; (iii) proceeds by factoring the supremum out of the integral; (iv) holds because of the linearity of the expectation and $e^{\alpha T} - 1 < e^{\alpha T}$. For $T4$, we have:

$$
\begin{aligned}
T4 &= \int_0^t e^{\alpha s} \langle (1-\gamma) v_* - u\nabla f(x_*), (x_s + v_s) - (x_* + v_*)\rangle\, ds \\
&\overset{(i)}{\le} \int_0^t e^{\alpha s} \|(\gamma - 1) v_* + u\nabla f(x_*)\|_2 \|(x_s + v_s) - (x_* + v_*)\|_2\, ds \\
&\overset{(ii)}{\le} \int_0^t e^{\alpha s} \|u\nabla f(x_*)\|_2 \|(x_s + v_s) - (x_* + v_*)\|_2\, ds \\
&\overset{(iii)}{\le} \frac{u^2 L}{m} \int_0^t e^{\alpha s} \|\nabla f(x_*)\|_2^2 + \frac{m}{4L} \int_0^t e^{\alpha s} \|(x_s + v_s) - (x_* + v_*)\|_2^2 ds \\
&\le \frac{u^2 L}{m\alpha} e^{\alpha t} \|\nabla f(x_*)\|_2^2\, ds + \frac{m}{4L} \int_0^t e^{\alpha s} \|(x_s + v_s) - (x_* + v_*)\|_2^2 ds
\end{aligned}
\tag{11}
$$

where (i) is from Cauchy-Schwartz inequality and the selection of γ ($\gamma \ge 1$) demonstrated in Lemma 5, (ii) proceeds by the choice of $\|v_*\|_2 = 0$, and (iii) is obtained by Young's inequality for products. We finally bound $T5$ as follows:

$$
\begin{aligned}
T5 &= \int_0^t e^{\alpha s} \left\langle -\nabla \log \pi(x_s), (x_s + v_s) - (x_* + v_*)\right\rangle ds \\
&\overset{(i)}{\le} \frac{\log B}{\alpha} \left(e^{\alpha t} - 1\right) \\
&\le \frac{\log B}{\alpha} e^{\alpha t}
\end{aligned}
\tag{12}
$$

where (i) is from Proposition 1. Incorporating the upper bound of $T1$-$T5$ into (7), we can now express the upper bound for $V(x_t, v_t, t)$ as:

$$
\begin{aligned}
V(x_t, v_t, t) &= \frac{1}{2} e^{\alpha t} \|(x_t, x_t + v_t) - (x_*, x_* + v_*)\|_2^2 \\
&\leq \frac{\alpha}{2} \int_0^t e^{\alpha s} \left(\|x_s - x_*\|_2^2 + \|(x_s + v_s) - (x_* + v_*)\|_2^2 \right) ds \\
&\quad - \frac{m}{2L} \int_0^t e^{\alpha s} \left(\|x_s - x_*\|_2^2 + \|(x_s + v_s) - (x_* + v_*)\|_2^2 \right) ds \\
&\quad + \frac{2\gamma u d}{n\rho\alpha} e^{\alpha t} + 2\sqrt{\frac{\gamma u}{n\rho}} M_t + \frac{u^2 L}{m\alpha} e^{\alpha t} \|\nabla f(x_*)\|_2^2 \\
&\quad + \frac{m}{4L} \int_0^t e^{\alpha s} \|(x_s + v_s) - (x_* + v_*)\|_2^2 ds + \frac{u \log B}{n\alpha} e^{\alpha t} \\
&\overset{(i)}{\leq} \frac{4\gamma u d L}{n\rho m} e^{\alpha t} + 2\sqrt{\frac{\gamma u}{n\rho}} M_t + \frac{2u^2 L^2}{m^2} e^{\alpha t} \|\nabla f(x_*)\|_2^2 + \frac{2uL \log B}{nm} e^{\alpha t} \\
&= \left(\frac{4\gamma u d L}{n\rho m} + \frac{2u^2 L^2}{m^2} \|\nabla f(x_*)\|_2^2 + \frac{2uL}{nm} \log B \right) e^{\alpha t} + 2\sqrt{\frac{\gamma u}{n\rho}} M_t,
\end{aligned} \tag{13}
$$

where (i) holds by the selection of $\alpha = \frac{m}{2L}$ and $\|x_t - x_*\|_2^2 \geq 0$. We now turn our attention to a detailed exploration of the moment of $V(x_t, v_t, t)$:

$$
\begin{aligned}
\mathbb{E}\left[\left(\sup_{0\leq t\leq T} V(x_t, v_t, t) \right)^p \right]^{\frac{1}{p}} &\leq \mathbb{E}\left[\left(\sup_{0\leq t\leq T} \frac{1}{2} e^{\alpha t} \left(\|x_t - x_*\|_2^2 + \|(x_t + v_t) - (x_* + v_*)\|_2^2 \right) \right)^p \right]^{\frac{1}{p}} \\
&\leq \mathbb{E}\left[\left(\sup_{0\leq t\leq T} \left(\frac{4\gamma u d L}{n\rho m} + \frac{2u^2 L^2}{m^2} \|\nabla f(x_*)\|_2^2 + \frac{2uL}{nm} \log B \right) e^{\alpha t} + 2\sqrt{\frac{\gamma u}{n\rho}} |M_t| \right)^p \right]^{\frac{1}{p}} \\
&\overset{(i)}{\leq} \underbrace{\mathbb{E}\left[\sup_{0\leq t\leq T} \left[\left(\frac{4\gamma u d L}{n\rho m} + \frac{2uL}{nm} \log B \right) e^{\alpha t} \right]^p \right]^{\frac{1}{p}}}_{T1} + \underbrace{\mathbb{E}\left[\sup_{0\leq t\leq T} \left[\frac{2u^2 L^2}{m^2} \|\nabla f(x_*)\|_2^2 e^{\alpha t} \right]^p \right]^{\frac{1}{p}}}_{T2} \\
&\quad + \underbrace{\mathbb{E}\left[\left(\sup_{0\leq t\leq T} 2\sqrt{\frac{\gamma u}{n\rho}} |M_t| \right)^p \right]^{\frac{1}{p}}}_{T3},
\end{aligned} \tag{14}
$$

where (i) is from Minkowski inequality (Yosida and Taylor, 1967). We first bound $T1$:

$$
\begin{aligned}
T1 &= \mathbb{E}\left[\sup_{0\leq t\leq T} \left[\left(\frac{4\gamma u d L}{n\rho m} + \frac{2uL}{nm} \log B \right) e^{\alpha t} \right]^p \right]^{\frac{1}{p}} \\
&\leq \left(\frac{4\gamma u d L}{n\rho m} + \frac{2uL}{nm} \log B \right) e^{\alpha T}.
\end{aligned} \tag{15}
$$

We proceed to bound $T2$:

$$
\begin{aligned}
T2 &= \mathbb{E}\left[\sup_{0\leq t\leq T} \left[\left(\frac{2u^2 L^2}{m^2} \|\nabla f(x_*)\|_2^2 \right) e^{\alpha t} \right]^p \right]^{\frac{1}{p}} \\
&\leq \frac{2u^2 L^2}{m^2} \mathbb{E}\left[\|\nabla f(x_*)\|_2^{2p} \right]^{\frac{1}{p}} e^{\alpha T}. \\
&\overset{(i)}{\leq} \frac{16 u^2 L^4 d p}{m^2 n \nu} e^{\alpha T},
\end{aligned} \tag{16}
$$

where (i) proceeds because from Proposition 2, we know that $\|\nabla f(x_*)\|_2$ is a $L\sqrt{\frac{d}{n\nu}}$-sub-Gaussian vector, and its corresponding expectation follows

$$\mathbb{E}\left[\|\nabla f(x_*)\|_2^{2p}\right]^{\frac{1}{p}} \leq \left(2L\sqrt{\frac{2dp}{n\nu}}\right)^2.$$

We then bound $T3$:

$$\begin{aligned}
T3 &= \mathbb{E}\left[\left(\sup_{0\leq t\leq T} 2\sqrt{\frac{\gamma u}{n\rho}}\,|M_t|\right)^p\right]^{\frac{1}{p}} \\
&\overset{(i)}{\leq} \mathbb{E}\left[\left(\frac{64\gamma upL}{n\rho m}e^{\alpha T}\right)^{\frac{p}{2}}\left(\sup_{0\leq t\leq T}\frac{1}{2}e^{\alpha t}\|(x_t+v_t)-(x_*+v_*)\|_2^2\right)^{\frac{p}{2}}\right]^{\frac{1}{p}} \\
&\overset{(ii)}{\leq} \mathbb{E}\left[2^{p-2}\left(\frac{64\gamma upL}{n\rho m}e^{\alpha T}\right)^{p}+\frac{1}{2^p}\left(\sup_{0\leq t\leq T}\frac{1}{2}e^{\alpha t}\|(x_t+v_t)-(x_*+v_*)\|_2^2\right)^{p}\right]^{\frac{1}{p}} \\
&\overset{(iii)}{\leq} 128\mathbb{E}\left[\left(\frac{\gamma upL}{n\rho m}e^{\alpha T}\right)^{p}\right]^{\frac{1}{p}}+\frac{1}{2}\mathbb{E}\left[\left(\sup_{0\leq t\leq T}\frac{1}{2}e^{\alpha t}\|(x_t+v_t)-(x_*+v_*)\|_2^2\right)^{p}\right]^{\frac{1}{p}} \\
&= \frac{128\gamma upL}{n\rho m}e^{\alpha T}+\frac{1}{2}\mathbb{E}\left[\left(\sup_{0\leq t\leq T}\frac{1}{2}e^{\alpha t}\|(x_t+v_t)-(x_*+v_*)\|_2^2\right)^{p}\right]^{\frac{1}{p}},
\end{aligned} \tag{17}$$

where (i) is from (10); inequality (ii) follows Young's inequality; (iii) proceeds by Minkowski inequality, the linearity of the expectation, and $2^{\frac{p-2}{p}} \leq 2$. By incorporating (15)-(17) into (14), we can get

$$\begin{aligned}
\mathbb{E}\left[\left(\sup_{0\leq t\leq T} V(x_t,v_t,t)\right)^p\right]^{\frac{1}{p}} &\leq T1+T2+T3 \\
&\overset{(i)}{\leq} \left(\frac{4\gamma udL}{n\rho m}+\frac{2uL}{nm}\log B+\frac{16u^2L^4dp}{m^2n\nu}+\frac{128\gamma upL}{n\rho m}\right)e^{\alpha T} \\
&\quad+\frac{1}{2}\mathbb{E}\left[\left(\sup_{0\leq t\leq T}\frac{1}{2}e^{\alpha t}\|(x_t+v_t)-(x_*+v_*)\|_2^2\right)^{p}\right]^{\frac{1}{p}} \\
&\overset{(ii)}{\leq} \left(\frac{4\gamma udL}{n\rho m}+\frac{2uL}{nm}\log B+\frac{16u^2L^4dp}{m^2n\nu}+\frac{128\gamma upL}{n\rho m}\right)e^{\alpha T} \\
&\quad+\frac{1}{2}\mathbb{E}\left[\left(\sup_{0\leq t\leq T}\frac{1}{2}e^{\alpha t}\|(x_t,x_t+v_t)-(x_*,x_*+v_*)\|_2^2\right)^{p}\right]^{\frac{1}{p}},
\end{aligned} \tag{18}$$

where (i) follows (15)-(17), (ii) holds because $e^{\alpha t}\|x_t-x_*\|_2^2 \geq 0$. From (14) we know that the last term on the right-hand side of (18) holds the equality

$$\frac{1}{2}\mathbb{E}\left[\left(\sup_{0\leq t\leq T}\frac{1}{2}e^{\alpha t}\left(\|x_t-x_*\|_2^2+\|(x_t+v_t)-(x_*+v_*)\|_2^2\right)\right)^p\right]^{\frac{1}{p}} = \frac{1}{2}\mathbb{E}\left[\left(\sup_{0\leq t\leq T}V(x_t,v_t,t)\right)^p\right]^{\frac{1}{p}},$$

and thus we derive the following bounds:

$$\begin{aligned}
\mathbb{E}\left[\left(\sup_{0\leq t\leq T} V(x_t,v_t,t)\right)^p\right]^{\frac{1}{p}} &\leq \left(\frac{8\gamma udL}{n\rho m}+\frac{4uL}{nm}\log B+\frac{32u^2L^4dp}{m^2n\nu}+\frac{256\gamma upL}{n\rho m}\right)e^{\alpha T} \\
&\overset{(i)}{=} \left(\frac{16d}{n\rho m}+\frac{4}{nm}\log B+\frac{32L^2dp}{m^2n\nu}+\frac{512p}{n\rho m}\right)e^{\alpha T} \\
&\overset{(ii)}{=} \frac{2}{mn}(D+\Omega p)\,e^{\alpha T},
\end{aligned} \tag{19}$$

where (i) is by the choice of $\gamma = 2$, $u = \frac{1}{L}$ from Lemma 5, and (ii) is by defining $D = \frac{8d}{\rho} + 2\log B$ and $\Omega = \frac{16L^2 d}{m\nu} + \frac{256}{\rho}$. With the defined bound on the moments of $V(x_t, v_t, t)$'s supremum and considering the known expression for $V(x, v, t)$, we proceed to determine the p-th moments of $\|(x_T, x_t + v_t) - (x_*, x_* + v_*)\|$:

$$\begin{aligned}
\mathbb{E}\left[\|(x_T, x_T + v_T) - (x_*, x_* + v_*)\|^p\right]^{\frac{1}{p}} &= \mathbb{E}\left[e^{-\frac{p\alpha T}{2}} V(x_T, v_T, T)^{\frac{p}{2}}\right]^{\frac{1}{p}} \\
&\overset{(i)}{\leq} \mathbb{E}\left[e^{-\frac{p\alpha T}{2}} \left(\sup_{0\leq t\leq T} V(x_t, v_t, t)\right)^{\frac{p}{2}}\right]^{\frac{1}{p}} \\
&= e^{-\frac{\alpha T}{2}} \left(\mathbb{E}\left[\left(\sup_{0\leq t\leq T} V(x_t, v_t, t)\right)^{\frac{p}{2}}\right]^{\frac{2}{p}}\right)^{\frac{1}{2}} \\
&\overset{(ii)}{\leq} e^{-\frac{\alpha T}{2}} \left[\frac{2}{mn}(D + \Omega p) e^{\alpha T}\right]^{\frac{1}{2}} \\
&= \sqrt{\frac{2}{mn}(D + \Omega p)},
\end{aligned} \tag{20}$$

where (i) holds by taking supremum of $V(\cdot, \cdot, \cdot)$, and (ii) is from (19). Upon taking the limit for $T \to \infty$ and by referring to Fatou's Lemma (Brezis and Brézis, 2011), we can upper bound the moments of $\mathbb{E}\left[\|x - x_*\|^p\right]^{\frac{1}{p}}$ with a confidence of $1 - \delta_1$:

$$\begin{aligned}
\mathbb{E}\left[\|x_T - x_*\|^p\right]^{\frac{1}{p}} &\leq \mathbb{E}\left[\|(x_T, x_T + v_T) - (x_*, x_* + v_*)\|^p\right]^{\frac{1}{p}} \\
&\leq \sqrt{\frac{2}{mn}(D + \Omega p)}.
\end{aligned} \tag{21}$$

Taking the limit as $T \to \infty$ and using Fatou's Lemma (Brezis and Brézis, 2011), we therefore have that the moments of $\mathbb{E}\left[\|x - x_*\|^p\right]^{\frac{1}{p}}$, with probability at least $1 - \delta_1$:

$$\begin{aligned}
\mathbb{E}\left[\|x - x_*\|^p\right]^{\frac{1}{p}} &\leq \liminf_{T\to\infty} \mathbb{E}\left[\|x_T - x_*\|^p\right]^{\frac{1}{p}} \\
&= \sqrt{\frac{2}{mn}(D + \Omega p)}.
\end{aligned} \tag{22}$$

By employing Markov's inequality (Boucheron et al., 2013), we further adapt our bound as follows:

$$\begin{aligned}
\mathbb{P}_{x\sim\mu_a^{(n)}}(\|x - x_*\| \geq \epsilon) &\leq \frac{\mathbb{E}\left[\|x - x_*\|^p\right]}{\epsilon^p} \\
&\leq \left(\frac{1}{\epsilon}\sqrt{\frac{2}{mn}(D + \Omega p)}\right)^p.
\end{aligned} \tag{23}$$

By setting $p = 2\log 1/\delta_1$ and defining $\epsilon = \sqrt{\frac{2e}{mn}(D + \Omega p)}$, we derive the required bound:

$$\mathbb{P}_{x\sim\mu_a^{(n)}[\rho_a]}\left(\|x - x_*\|_2 \geq \sqrt{\frac{2e}{mn}(D + 2\Omega\log 1/\delta_1)}\right) \leq \delta_1, \tag{24}$$

for $\delta_1 \leq e^{-0.5}$.

□

C.3 Regret Analysis of Thompson Sampling

After establishing the contraction results for the posterior distributions, we next delve into the analysis of regret, building upon the foundations laid by the concentration results. Specifically, under the assumptions we have made, we demonstrate that using Thompson sampling with samples drawn from the posteriors can achieve optimal regret guarantees within a finite time frame. To elucidate this, we employ a typical approach used in regret proofs associated with Thompson sampling

(Lattimore and Szepesvári, 2020). Our primary goal is to quantify the number of times, denoted as $\mathcal{L}_a(N)$, the sub-optimal arm is selected up to time N.

For clarity and simplicity in our discourse, we assume, without the loss of generality, that the first arm is optimal. The filtration corresponding to an execution of the algorithm is denoted as $\mathcal{F}_n = \sigma(A_1, \mathcal{R}_1, A_2, \mathcal{R}_2, A_3, \mathcal{R}_3, \cdots, A_n, \mathcal{R}_n)$, where it is a σ-algebra generated by Algorithm 3 after playing n times. We also denote an event $\mathcal{E}_a(n) = \left\{\mathcal{R}_{a,n} \geq \bar{\mathcal{R}}_1 - \epsilon\right\}$ to indicate the estimated reward of arm a at round n exceeds the expected reward of optimal arm by at least a positive constant ϵ. We also define its probability as $\mathrm{P}(\mathcal{E}_a(n)|\mathcal{F}_{n-1}) = \mathcal{G}_{a,n}$. Next, we analyze the expected number of times that suboptimal arms are played by breaking it down into two parts:

$$\begin{aligned}\mathbb{E}[\mathcal{L}_a(N)] &= \mathbb{E}\left[\sum_{n=1}^{N} \mathbb{I}(A_n = a)\right] \\ &= \mathbb{E}\left[\sum_{n=1}^{N} \mathbb{I}(A_n = a, \mathcal{E}_a(n))\right] + \mathbb{E}\left[\sum_{n=1}^{N} \mathbb{I}(A_n = a, \mathcal{E}_a^c(n))\right].\end{aligned} \tag{25}$$

We now turn our attention to the proof of the upper bounds for the terms appearing in (25).

Lemma 1. *For a sub-optimal arm $a \in \mathcal{A}$, we have an upper bound as outlined below:*

$$\mathbb{E}\left[\sum_{n=1}^{N} \mathbb{I}(A_n = a, \mathcal{E}_a^c(n))\right] \leq \mathbb{E}\left[\sum_{s=0}^{N-1}\left(\frac{1}{\mathcal{G}_{1,s}} - 1\right)\right], \tag{26}$$

where $\mathcal{G}_{1,n} \coloneqq \mathrm{P}\left(\mathcal{R}_{1,n} > \bar{\mathcal{R}}_1 - \epsilon|\mathcal{F}_{n-1}\right)$, for some $\epsilon > 0$.

Proof. It should be noted that $A_n = \operatorname{argmax}_{a\in\mathcal{A}} \langle \alpha_a, x_{a,n}\rangle$ is the arm that has the highest sample reward at round n. Additionally, we set arm $A'_n = \operatorname{argmax}_{a\in\mathcal{A}, a\neq 1} \langle \alpha_a, x_{a,n}\rangle$ to be the one that attains the maximum sample reward except for the optimal arm A_n. It is worth noting that the following inequality holds:

$$\begin{aligned}\mathbb{P}(A_n = a, \mathcal{E}_a^c(n)|\mathcal{F}_{n-1}) &\overset{(i)}{\leq} \mathbb{P}\left(A'_n = a, \mathcal{E}_a^c(n), \mathcal{R}_{1,n} < \bar{\mathcal{R}}_1 - \epsilon|\mathcal{F}_{n-1}\right) \\ &\overset{(ii)}{=} \mathbb{P}(A'_n = a, \mathcal{E}_a^c(n)|\mathcal{F}_{n-1})\,\mathbb{P}\left(\mathcal{R}_{1,n} < \bar{\mathcal{R}}_1 - \epsilon|\mathcal{F}_{n-1}\right) \\ &= \mathbb{P}(A'_n = a, \mathcal{E}_a^c(n)|\mathcal{F}_{n-1})\,(1 - \mathbb{P}(\mathcal{E}_1(n)|\mathcal{F}_{n-1})) \\ &\overset{(iii)}{\leq} \frac{\mathbb{P}(A_n = 1, \mathcal{E}_a^c(n)|\mathcal{F}_{n-1})}{\mathbb{P}(\mathcal{E}_1(n)|\mathcal{F}_{n-1})}\,(1 - \mathbb{P}(\mathcal{E}_1(n)|\mathcal{F}_{n-1})) \\ &\overset{(iv)}{\leq} \mathbb{P}(A_n = 1|\mathcal{F}_{n-1})\left(\frac{1}{\mathbb{P}(\mathcal{E}_1(n)|\mathcal{F}_{n-1})} - 1\right) \\ &= \mathbb{P}(A_n = 1|\mathcal{F}_{n-1})\left(\frac{1}{\mathcal{G}_{1,n}} - 1\right),\end{aligned} \tag{27}$$

where (i) holds because event $\{A_n = a, \mathcal{E}_a^c(n)\} \subseteq \left\{A'_n = a, \mathcal{E}_a^c(n), \mathcal{R}_{1,n} < \bar{\mathcal{R}}_1 - \epsilon\right\}$ given $\mathcal{F}_{n-1}$, (ii) is valid because events $\{A'_n = a, \mathcal{E}_a^c(n)\}$ and $\left\{\mathcal{R}_{1,n} < \bar{\mathcal{R}}_1 - \epsilon\right\}$ are independent, (iii) is from $\mathbb{P}(A'_n = a, \mathcal{E}_a^c(n)|\mathcal{F}_{n-1})\,\mathbb{P}(\mathcal{E}_1(n)|\mathcal{F}_{n-1}) \leq \mathbb{P}(A_n = 1, \mathcal{E}_a^c(n)|\mathcal{F}_{n-1})$ since $\left\{A'_n = a, \mathcal{E}_a^c(n), \mathcal{E}_1^c(n)\right\} \subseteq \left\{A_n = 1, \mathcal{E}_a^c(n), \mathcal{E}_1^c(n)\right\}$. The fact that $\{A_n = 1, \mathcal{E}_a^c(n)\} \subseteq \{A_n = 1\}$ leads to the inequality (iv). Now by summing the result in (27) from $n = 1$ to N we have:

$$\begin{aligned}\mathbb{E}\left[\sum_{n=1}^{N}\mathbb{I}\left(A_n=a,\mathcal{E}_a^c(n)\right)\right] &\overset{(i)}{\leq} \mathbb{E}\left[\sum_{n=1}^{N}\mathbb{P}\left(A_n=1|\mathcal{F}_{n-1}\right)\left(\frac{1}{\mathcal{G}_{1,\mathcal{L}_1(n-1)}}-1\right)\right]\\ &\overset{(ii)}{=} \mathbb{E}\left[\sum_{n=1}^{N}\mathbb{I}\left(A_n=1\right)\left(\frac{1}{\mathcal{G}_{1,\mathcal{L}_1(n-1)}}-1\right)\right]\\ &\overset{(iii)}{\leq} \mathbb{E}\left[\sum_{s=0}^{N-1}\left(\frac{1}{\mathcal{G}_{1,s}}-1\right)\right],\end{aligned} \tag{28}$$

where (i) holds because of the law of total expectation and (27), (ii) is also derived from the law of total expectation, and (iii) is valid due to the constraint that, for a given $s \in [\![1, N]\!]$, both $\mathbb{I}(A_n = 1)$ and $\mathbb{I}(\mathcal{L}_1(n) = s)$ can simultaneously hold true at most once.

□

Lemma 2. *Considering a sub-optimal arm $a \in \mathcal{A}$, we derive an upper bound expressed as*

$$\mathbb{E}\left[\sum_{n=1}^{N}\mathbb{I}\left(A_n=a,\mathcal{E}_a(n)\right)\right] \leq 1+\mathbb{E}\left[\sum_{s=0}^{N-1}\mathbb{I}\left(\mathcal{G}_{a,s}>\frac{1}{N}\right)\right], \tag{29}$$

where $\mathcal{G}_{a,n} \coloneqq \mathtt{P}\left(\mathcal{R}_{a,n} > \bar{\mathcal{R}}_1 - \epsilon|\mathcal{F}_{n-1}\right)$, for a certain $\epsilon > 0$.

Proof. The derivation of the upper bound for Lemma 2 is identical to that presented in Agrawal and Goyal (2012); Lattimore and Szepesvári (2020), and we revisit this proof herein for comprehensive exposition. With a defined set $\mathcal{N} = \left\{n : \mathcal{G}_{a,\mathcal{L}_a(n)} > \frac{1}{N}\right\}$, we decompose the expression in (29) into two constituent terms:

$$\mathbb{E}\left[\sum_{n=1}^{N}\mathbb{I}\left(A_n=a,\mathcal{E}_a(n)\right)\right] \leq \underbrace{\mathbb{E}\left[\sum_{n\in\mathcal{N}}\mathbb{I}\left(A_n=a\right)\right]}_{T1} + \underbrace{\mathbb{E}\left[\sum_{n\notin\mathcal{N}}\mathbb{I}\left(\mathcal{E}_a(n)\right)\right]}_{T2} \tag{30}$$

For term $T1$, we have:

$$\begin{aligned}T1 &= \sum_{n=1}^{N}\mathbb{I}\left(A_n=a\right)\\ &\overset{(i)}{\leq} \sum_{n=1}^{N}\sum_{s=1}^{N}\mathbb{I}\left(\mathcal{L}_a(n)=s,\mathcal{L}_a(n-1)=s-1,\mathcal{G}_{a,\mathcal{L}_a(n-1)}>\frac{1}{N}\right)\\ &= \sum_{s=1}^{N}\mathbb{I}\left(\mathcal{G}_{a,s-1}>\frac{1}{N}\right)\sum_{n=1}^{N}\mathbb{I}\left(\mathcal{L}_a(n)=s,\mathcal{L}_a(n-1)=s-1\right)\\ &\overset{(ii)}{=} \sum_{s=1}^{N}\mathbb{I}\left(\mathcal{G}_{a,s-1}>\frac{1}{N}\right),\end{aligned}$$

where (i) uses the fact that when $A_n = a, \mathcal{L}_a(n) = s$ and $\mathcal{L}_a(n-1) = s-1$ for some $s \in [\![1, N]\!]$ and $n \in [\![1, N]\!]$ implies that $G_{a,s-1} > 1/N$, then (ii) holds because for any $s \in [\![1, N]\!]$, there is at most one time point $n \in [\![1, N]\!]$ such that both $\mathcal{L}_a(n) = s$

and $\mathcal{L}_a(n-1) = s-1$ hold true. For the next inequality, note that

$$\begin{aligned}
\mathbb{E}\left[\sum_{n\notin\mathcal{N}} \mathbb{I}\left(\mathcal{E}_a(n)\right)\right] &= \sum_{n\notin\mathcal{N}} \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\left(\mathcal{E}_a(n), \mathcal{G}_{a,\mathcal{L}_a(n-1)} \leq \frac{1}{N}\right)|\mathcal{F}_{n-1}\right]\right] \\
&\overset{(i)}{=} \sum_{n\notin\mathcal{N}} \mathbb{E}\left[\mathcal{G}_{a,\mathcal{L}_a(n-1)}\mathbb{I}\left(\mathcal{G}_{a,\mathcal{L}_a(n-1)} \leq \frac{1}{N}\right)\right] \\
&\leq \sum_{n\notin\mathcal{N}} \mathbb{E}\left[\frac{1}{N}\mathbb{I}\left(\mathcal{G}_{a,\mathcal{L}_a(n-1)} \leq \frac{1}{N}\right)\right] \\
&\leq \mathbb{E}\left[\sum_{n\notin\mathcal{N}} \frac{1}{N}\right] \\
&\leq 1,
\end{aligned}$$

where (i) stems from the fact that $\mathbb{I}\left(\mathcal{G}_{a,\mathcal{L}_a(n-1)} \leq \frac{1}{N}\right)$ is $\mathcal{F}_{n-1}$-measurable and the following equality holds:

$$\mathbb{E}\left[\mathbb{I}\left(\mathcal{E}_a(n)\right)|\mathcal{F}_{n-1}\right] = 1 - \mathbb{P}\left(\mathcal{R}_{a,n} < \bar{\mathcal{R}}_1 - \epsilon|\mathcal{F}_{n-1}\right) = \mathcal{G}_{a,\mathcal{L}_a(n-1)}.$$

By employing the upper bounds established for terms $T1$ and $T2$ as delineated in (30), we achieve the targeted result in (29). □

C.4 Regret Analysis of Exact Thompson Sampling

Drawing from insights in Lemmas 1 and 2, we delve deeper to derive the upper bound in the context of exact Thompson sampling. To derive the bound, we introduce two fundamental lemmas, which are crucial for the definitive proof of total expected regrets. Notably, our first lemma prescribes a probability threshold for an arm being thoroughly explored, as a function of prior quality.

Lemma 3. *Suppose the likelihood satisfies Assumption 1, and the prior satisfies Assumption 3. Then for every $n = 1, \cdots, N$ and with the selection of $\rho_1 = \frac{\nu_1 m_1^2}{8dL_1^3}$, we will have:*

$$\mathbb{E}\left[\frac{1}{\mathcal{G}_{1,n}}\right] \leq 16\sqrt{\kappa_1 B_1}.$$

Proof. For notational simplicity within this proof, we have chosen to exclude the arm-specific subscript, except when explicitly required. Our primary focus is directed towards $\|(x_*, v_*) - (x_u, v_u)\|_2^2$. In this context, (x_u, v_u) serves as the mode of the posterior for the first arm once it has received n samples in alignment with the following condition:

$$(\gamma - 1)\, v_u + u\nabla f\left(x_u\right) + \frac{u}{n}\nabla \log \pi(x_u) = 0$$

Using the provided definition and taking $\bar{x} = x_u - x_*$ and $\bar{v} = v_u - v_*$, we have that:

$$\begin{aligned}
(\gamma - 1)\,\bar{v} + u\left(\nabla f(x_u) - \nabla f(x_*)\right) &= -(\gamma - 1)\, v_* - u\nabla f(x_*) - \frac{u}{n}\nabla \log \pi(x_u) \\
-\langle \bar{x}, \bar{v}\rangle - \langle \bar{x} + \bar{v}, (\gamma - 1)\,\bar{v} + u\left(\nabla f(x_u) - \nabla f(x_*)\right)\rangle &= -\langle \bar{x}, \bar{v}\rangle - \left\langle \bar{x} + \bar{v}, (\gamma - 1)\, v_* + u\nabla f(x_*) + \frac{u}{n}\nabla \log \pi(x_u)\right\rangle.
\end{aligned} \tag{31}$$

From the above equation, we try to upper bound the distance between fixed points x_* and the posterior mode x_u:

$$\begin{aligned}
\|x_* - x_u\|_2^2 &\leq \|\bar{x}\|_2^2 + \|\bar{x} + \bar{v}\|_2^2 \\
&\overset{(i)}{\leq} -\frac{2L}{m}\left(\langle \bar{x}, \bar{v}\rangle + \langle \bar{x} + \bar{v}, (\gamma - 1)\bar{v} + u\left(\nabla f(x_u) - \nabla f(x_*)\right)\rangle\right) \\
&\overset{(ii)}{=} -\frac{2L}{m}\left(\langle \bar{x}, \bar{v}\rangle + \left\langle \bar{x} + \bar{v}, (\gamma - 1)\, v_* + u\nabla f(x_*) + \frac{u}{n}\nabla \log \pi(x_u)\right\rangle\right) \\
&\overset{(iii)}{\leq} -\frac{2L}{m}\langle \bar{x}, \bar{v}\rangle + \|\bar{x} + \bar{v}\|_2^2 + \frac{u^2L^2}{m^2}\|\nabla f(x_*)\|_2^2 + \frac{2uL}{mn}\log B \\
&\overset{(iv)}{\leq} \frac{2L}{m}\varepsilon + \|\bar{x} + \bar{v}\|_2^2 + \frac{u^2L^2}{m^2}\|\nabla f(x_*)\|_2^2 + \frac{2uL}{mn}\log B,
\end{aligned}$$

where we derive (i) from Lemma 5. we refer to (31) to derive (ii), and (iii) is from the application of Young's inequality and the insights of (11). In (iv), we adopt the constraint that $0 \leq |\langle \bar{x}, \bar{v} \rangle| \leq \varepsilon$. Building upon the definition $\kappa = L/m$, we then proceed to derive the subsequent inequality:

$$\begin{aligned}\|x_* - x_u\|_2^2 &\leq \frac{2L}{m}\varepsilon + \frac{2uL}{mn}\log B + \frac{u^2L^2}{m^2}\|\nabla f(x_*)\|_2^2 \\ &= 2\kappa\varepsilon + \frac{2u\kappa}{n}\log B + u^2\kappa^2\|\nabla f(x_*)\|_2^2.\end{aligned}$$

Noting that $|\langle \alpha, \bar{x} \rangle| \leq \sqrt{\omega^2\|x_* - x_u\|_2^2}$ (ω is the norm of α) we find that:

$$\begin{aligned}\mathcal{G}_{1,s} &= \mathrm{P}\left(\langle \alpha_1, x - x_u \rangle \geq \langle \alpha_1, x_* - x_u \rangle - \epsilon\right) \\ &\geq \mathrm{P}\left(\langle \alpha_1, x - x_u \rangle \geq \underbrace{\sqrt{2\omega_1^2\kappa_1\varepsilon + \frac{2\omega_1^2 u\kappa}{n}\log B_1 + \omega_1^2u^2\kappa_1^2\|\nabla f(x_*)\|_2^2}}_{=t}\right).\end{aligned}$$

Based on the information that the posterior over (x, v) aligns proportionally to the marginal distribution described by $\exp\left(-\rho_1\left(nf(x) + \pi(x)\right)\right)$, and given its attributes as being $\rho_1 L_1(n+1)$ Lipschitz smooth and $\rho_1 m_1 n$ strongly log-concave with mode x_u, we can infer from Theorem 3.8 in Saumard and Wellner (2014) that the marginal density of $\langle \alpha_1, x \rangle$ exhibits the same smoothness and log-concavity. Then we can derive the following relationship:

$$\mathrm{P}\left(\langle \alpha_1, x - x_u \rangle \geq t\right) \geq \sqrt{\frac{nm_1}{(n+1)L_1}}\mathrm{P}(\mathcal{X} \geq t)$$

where $\mathcal{X} \sim \mathcal{N}\left(0, \frac{\omega_1^2}{\rho_1(n+1)L_1}\right)$. Upon applying a lower bound derived from the cumulative density function of a Gaussian random variable, we ascertain that, for $\sigma^2 = \frac{\omega_1^2}{\rho_1(n+1)L_1}$:

$$\mathcal{G}_{1,s} \geq \sqrt{\frac{nm_1}{2\pi(n+1)L_1}}\begin{cases}\frac{\sigma t}{t^2+\sigma^2}e^{-\frac{t^2}{2\sigma^2}} & : t > \frac{\omega_1}{\sqrt{\rho_1(n+1)L_1}} \\ 0.34 & : t \leq \frac{\omega_1}{\sqrt{\rho_1(n+1)L_1}}\end{cases}$$

Thus we have that:

$$\begin{aligned}\frac{1}{\mathcal{G}_{1,s}} &\leq \sqrt{\frac{2\pi(n+1)L_1}{nm_1}}\begin{cases}\frac{t^2+\sigma^2}{\sigma t}e^{\frac{t^2}{2\sigma^2}} & : t > \frac{\omega_1}{\sqrt{\rho_1(n+1)L_1}} \\ \frac{1}{0.34} & : t \leq \frac{\omega_1}{\sqrt{\rho_1(n+1)L_1}}\end{cases} \\ &\leq \sqrt{\frac{2\pi(n+1)L_1}{nm_1}}\begin{cases}\left(\frac{t}{\sigma}+1\right)e^{\frac{t^2}{2\sigma^2}} & : t > \frac{\omega_1}{\sqrt{\rho_1(n+1)L_1}} \\ 3 & : t \leq \frac{\omega_1}{\sqrt{\rho_1(n+1)L_1}}\end{cases}\end{aligned}$$

After determining the expectation on both sides considering the samples $\mathcal{R}_1, \ldots, \mathcal{R}_n$, we deduce that:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\mathcal{G}_{1,s}}\right] &\leq 3\sqrt{\frac{2\pi(n+1)L_1}{nm_1}} + \sqrt{\frac{2\pi(n+1)L_1}{nm_1}}\mathbb{E}\left[\left(\frac{t}{\sigma}+1\right)e^{\frac{t^2}{2\sigma^2}}\right] \\ &\overset{(i)}{\leq} 6\sqrt{2\pi\kappa_1} + 2\sqrt{2\pi\kappa_1}\mathbb{E}\left[\left(\frac{t}{\sigma}+1\right)e^{\frac{t^2}{2\sigma^2}}\right],\end{aligned}$$

where (i) proceeds by $0 < \frac{(n+1)L_1}{nm_1} \leq \frac{2L_1}{m_1} = 2\kappa_1$. Our attention now turns to firstly bounding the term $\mathbb{E}\left[\left(\frac{t}{\sigma}+1\right)e^{\frac{t^2}{2\sigma^2}}\right]$, and

then we proceed to derive an upper bound for $\mathbb{E}\left[\frac{1}{\mathcal{G}_{1,s}}\right]$ (we drop arm-specific parameters here to simplify the notation).

$$
\begin{aligned}
\mathbb{E}\left[\left(\frac{t}{\sigma}+1\right)e^{\frac{t^2}{2\sigma^2}}\right] &= \mathbb{E}\left[\left(\sqrt{\rho(n+1)L\left(2\kappa\varepsilon+\frac{2u\kappa}{n}\log B+u^2\kappa^2\|\nabla f(x_*)\|_2^2\right)}+1\right)e^{\frac{\rho(n+1)L}{2}\left(2\kappa\varepsilon+\frac{2u\kappa}{n}\log B+u^2\kappa^2\|\nabla f(x_*)\|_2^2\right)}\right] \\
&\overset{(i)}{=} \mathbb{E}\left[\left(\sqrt{\rho(n+1)L\left(\frac{2L}{m}\varepsilon+\frac{2\log B}{mn}+\frac{\|\nabla f(x_*)\|_2^2}{m^2}\right)}+1\right)e^{\frac{\rho(n+1)L}{2}\left(\frac{2L}{m}\varepsilon+\frac{2}{mn}\log B+\frac{1}{m^2}\|\nabla f(x_*)\|_2^2\right)}\right] \\
&\overset{(ii)}{\leq} e^{\frac{\rho(n+1)L}{mn}\left(Ln\varepsilon+\log B\right)}\left(\sqrt{\frac{2\rho(n+1)L\left(Ln\varepsilon+\log B\right)}{mn}}+1\right)\mathbb{E}\left[e^{\frac{\rho(n+1)L}{2m^2}\|\nabla f(x_*)\|_2^2}\right] \\
&\quad + e^{\frac{\rho(n+1)L}{mn}\left(Ln\varepsilon+\log B\right)}\sqrt{\frac{\rho(n+1)L}{m^2}}\sqrt{\mathbb{E}\left[\|\nabla f(x_*)\|_2^2\right]\mathbb{E}\left[e^{\frac{\rho(n+1)L}{2m^2}\|\nabla f(x_*)\|_2^2}\right]} \\
&\overset{(iii)}{\leq} e^{\frac{\rho(n+1)L}{mn}\left(Ln\varepsilon+\log B\right)+\frac{1}{2}}\left(\sqrt{\frac{2\rho(n+1)L}{mn}\left(Ln\varepsilon+\log B\right)}+1+\sqrt{\frac{2\rho L^3 d}{m^2\nu}\frac{n+1}{n}}\right), \\
&\overset{(iv)}{\leq} \sqrt{e}\,(B)^{1/4}\left(\sqrt{\log B}+1+\frac{\sqrt{2}}{2}\right),
\end{aligned}
\tag{32}
$$

where from the selections of $u=\frac{1}{L_1}$ and $\kappa=\frac{L_1}{m_1}$, we obtain (i). (ii) is further extracted from an intrinsic inequality:

$$
\sqrt{\frac{2L_1}{m_1}\varepsilon+\frac{2\log B_1}{m_1 n}+\frac{\|\nabla f(x_*)\|_2^2}{m_1^2}} \leq \sqrt{\frac{2L_1}{m_1}\varepsilon+\frac{2\log B_1}{m_1 n}}+\frac{\|\nabla f(x_*)\|_2}{m_1}.
$$

(iii) holds because $\|\nabla f(x_*)\|_2$ is $L_1\sqrt{\frac{d}{n\nu_1}}$ sub-Gaussian and its squared $\|\nabla f(x_*)\|_2^2$ form as sub-exponential, then by selecting $\lambda<\frac{n\nu_1}{4dL_1^2}$ and $\rho_1=\frac{m_1^2\nu_1}{8L_1^3 d}$, we can bound the following random variables as:

$$
\begin{aligned}
\mathbb{E}\left[e^{\lambda\|\nabla f(x_*)\|_2^2}\right] &\leq e, \\
\mathbb{E}\left[\|\nabla f(x_*)\|_2^2\right] &\leq 2\frac{L_1^2 d}{\nu_1 n}.
\end{aligned}
$$

We finally obtain (iv) with the choice of $\varepsilon\leq\frac{\log B_1}{L_1 n}$, $\frac{m_1}{L_1}\leq 1$, $\frac{1}{d}\leq 1$, $\frac{n+1}{n}\leq 2$, and $\frac{\nu_1}{L_1}\leq 1$ without loss of generality. Then we turn our attention to bound the expected value of $\frac{1}{\mathcal{G}_{1,s}}$:

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{\mathcal{G}_{1,s}}\right] &\leq 3\sqrt{2\pi\kappa_1}+\sqrt{2\pi\kappa_1}\mathbb{E}\left[\left(\frac{t}{\sigma}+1\right)e^{\frac{t^2}{2\sigma^2}}\right] \\
&\overset{(i)}{\leq} 3\sqrt{2\pi\kappa_1}+\sqrt{2\pi\kappa_1 e}\,(B_1)^{1/4}\left(\sqrt{\log B_1}+1+\frac{\sqrt{2}}{2}\right) \\
&\leq \sqrt{2\pi\kappa_1 e}\,(B_1)^{1/4}\left(\sqrt{\log B_1}+1+\frac{\sqrt{2}}{2}+\frac{3}{\sqrt{e}}\right) \\
&\overset{(ii)}{<} 16\sqrt{\kappa_1 B_1},
\end{aligned}
$$

where (i) is from (32), and (ii) uses the fact that

$$
\sqrt{2\pi e}x^{1/4}\left(\sqrt{\log x}+1+\frac{\sqrt{2}}{2}+\frac{3}{\sqrt{e}}\right)<16\sqrt{x}
$$

when $x\geq 1$. We then derive the final upper bound.

□

We introduce the next technical lemma below, which provides problem-specific upper-upper bounds for the terms described in Lemmas 1 and 2.

Lemma 4. *Suppose the likelihood, rewards, and priors satisfy Assumptions 1-3, then for $a \in \mathcal{A}$ and $\rho_a = \frac{\nu_a m_a^2}{8dL_a^2}$:*

$$\sum_{s=0}^{N-1} \mathbb{E}\left[\frac{1}{\mathcal{G}_{1,s}} - 1\right] \leq 16\sqrt{\frac{L_1}{m_1} B_1}\left\lceil \frac{8e\omega_1^2}{m_1\Delta_a^2}(D_1 + 2\Omega_1)\right\rceil + 1, \tag{33}$$

$$\sum_{s=0}^{N-1} \mathbb{E}\left[\mathbb{I}\left(\mathcal{G}_{a,s} > \frac{1}{N}\right)\right] \leq \frac{8e\omega_a^2}{m_a\Delta_a^2}\left(D_a + 2\Omega_a \log N\right), \tag{34}$$

where from the preceding derivation, it follows that for all arms $a \in \mathcal{A}$, D_a and Ω_a are specified as $D_a = 2\log B_a + \frac{8d}{\rho_a}$ and $\Omega_a = \frac{16L_a^2 d}{m_a \nu_a} + \frac{256}{\rho_a}$ respectively.

Proof. We begin by showing that for $a \in \mathcal{A}$, the lower bound for $\mathcal{G}_{a,s}$:

$$\begin{aligned}
\mathcal{G}_{a,s} &= \mathbb{P}(\mathcal{R}_{a,s} > \bar{\mathcal{R}}_a - \epsilon | \mathcal{F}_{n-1}) \\
&= 1 - \mathbb{P}(\mathcal{R}_{a,s} - \bar{\mathcal{R}}_a \leq -\epsilon | \mathcal{F}_{n-1}) \\
&\geq 1 - \mathbb{P}(|\mathcal{R}_{a,s} - \bar{\mathcal{R}}_a| \geq \epsilon | \mathcal{F}_{n-1}) \\
&\overset{(i)}{\geq} 1 - \mathbb{P}_{x\sim\mu_a^{(s)}}\left(\|x - x_*\| \geq \frac{\epsilon}{\omega_a}\right),
\end{aligned} \tag{35}$$

where (i) lies in the ω_a-Lipschitz continuity of both $\mathcal{R}_{a,s}$ and $\bar{\mathcal{R}}_a$ w.r.t. $x \sim \mu_a^{(s)}$ and x_* respectively. We then use the above result to upper bound $\sum_s \mathbb{E}\left[\frac{1}{\mathcal{G}_{1,s}} - 1\right]$ and $\sum_s \mathbb{E}\left[\mathbb{I}\left(\mathcal{G}_{a,s} > \frac{1}{N}\right)\right]$. For (33), we have:

$$\begin{aligned}
\sum_{s=0}^{N-1} \mathbb{E}\left[\frac{1}{\mathcal{G}_{1,s}} - 1\right] &= \sum_{s=0}^{\ell-1} \mathbb{E}\left[\frac{1}{\mathcal{G}_{1,s}} - 1\right] + \sum_{s=\ell}^{N-1} \mathbb{E}\left[\frac{1}{\mathcal{G}_{1,s}} - 1\right] \\
&\overset{(i)}{\leq} \sum_{s=0}^{\ell-1} \mathbb{E}\left[\frac{1}{\mathcal{G}_{1,s}} - 1\right] + \sum_{s=\ell}^{N-1}\left[\frac{1}{1 - \mathbb{P}_{x\sim\mu_1^{(s)}}(\|x - x_*\| \geq \epsilon/\omega_1)} - 1\right] \\
&\leq \sum_{s=0}^{\ell-1} \mathbb{E}\left[\frac{1}{\mathcal{G}_{1,s}} - 1\right] + \int_{s=\ell}^{\infty}\left[\frac{1}{1 - \mathbb{P}_{x\sim\mu_1^{(s)}}(\|x - x_*\| \geq \epsilon/\omega_1)} - 1\right] ds \\
&\overset{(ii)}{\leq} \sum_{s=0}^{\ell-1} \mathbb{E}\left[\frac{1}{\mathcal{G}_{1,s}} - 1\right] + \int_{s=1}^{\infty}\left[\frac{1}{1 - \frac{1}{2}\exp\left(-\frac{m_1\epsilon^2}{4e\Omega_1\omega_1^2}s\right)} - 1\right] ds \\
&\overset{(iii)}{\leq} \sum_{s=0}^{\ell-1} \mathbb{E}\left[\frac{1}{\mathcal{G}_{1,s}} - 1\right] + \frac{16e\Omega_1\omega_1^2}{m_1\Delta_a^2}\log 2 + 1 \\
&\overset{(iv)}{\leq} 16\sqrt{\frac{L_1}{m_1}B_1}\left\lceil\frac{8e\omega_1^2}{m_1\Delta_a^2}(D_1 + 2\Omega_1)\right\rceil + 1
\end{aligned} \tag{36}$$

where (i) is by inducing (35). (ii) is because from Theorem 6 we have the following posterior concentration inequality:

$$\mathbb{P}_{x\sim\mu_1^{(s)}}\left(\|x - x_*\| > \frac{\epsilon}{A_1}\right) \leq \exp\left(-\frac{1}{2\Omega_1}\left(\frac{m_1 n\epsilon^2}{2e\omega_1^2} - D_1\right)\right), \tag{37}$$

where the given bound holds substantive value only if $n > \frac{2e\omega_1^2}{\epsilon^2 m_1} D_1$, and thus we specify that the lower bound of the definite integral in (ii) is greater than n ($\ell \geq n > \frac{2e\omega_1^2}{\epsilon^2 m_1} D_1$). Then (iii) holds by the selection of $\epsilon = (\bar{\mathcal{R}}_1 - \bar{\mathcal{R}}_a)/2 = \Delta_a/2$ and ℓ:

$$\ell = \left\lceil \frac{8e\omega_1^2}{m_1\Delta_a^2}(D_1 + 2\Omega_1 \log 2)\right\rceil, \tag{38}$$

and by performing elementary manipulations on the definite integral:

$$\int_{s=1}^{\infty} \frac{1}{1 - \frac{1}{2}g(s)} - 1 ds = \frac{\log 2 - \log(2e^c - 1)}{c} + 1 \leq \frac{\log 2}{c} + 1 = \frac{16e\Omega_1\omega_1^2}{m_1\Delta_a^2}\log 2 + 1.$$

Here function $g(\cdot)$ is defined as $g(s) = \exp\left(-\frac{m_1\epsilon^2}{4e\Omega_1\omega_1^2}s\right)$ for $s \geq \left\lceil \frac{8e\omega_1^2}{m_1\Delta_a^2}\left(D_1 + 2\Omega_1\log 2\right)\right\rceil$ and $c = \frac{m_1\Delta_a^2}{16e\Omega_1\omega_1^2}$. Then the final step (iv) proceeds by the choice of ℓ mentioned in (38), Lemma 3, and some simple mathematical operations.

The verification of (34) is achieved through deriving a similar from as (37):

$$\begin{aligned}
\sum_{s=0}^{N-1}\mathbb{E}\left[\mathbb{I}\left(\mathcal{G}_{a,s} > \frac{1}{N}\right)\right] &= \sum_{s=0}^{N-1}\mathbb{E}\left[\mathbb{I}\left(\mathbb{P}\left(\mathcal{R}_{a,s} - \bar{\mathcal{R}}_a > \Delta_a - \epsilon\middle|\mathcal{F}_{n-1}\right) > \frac{1}{N}\right)\right] \\
&\overset{(i)}{=} \sum_{s=0}^{N-1}\mathbb{E}\left[\mathbb{I}\left(\mathbb{P}\left(\mathcal{R}_{a,s} - \bar{\mathcal{R}}_a > \frac{\Delta_a}{2}\middle|\mathcal{F}_{n-1}\right) > \frac{1}{N}\right)\right] \\
&\leq \sum_{s=0}^{N-1}\mathbb{E}\left[\mathbb{I}\left(\mathbb{P}\left(|\mathcal{R}_{a,s} - \bar{\mathcal{R}}_a| > \frac{\Delta_a}{2}\middle|\mathcal{F}_{n-1}\right) > \frac{1}{N}\right)\right] \\
&\leq \sum_{s=0}^{N-1}\mathbb{E}\left[\mathbb{I}\left(\mathbb{P}_{x\sim\mu_a^{(s)}[\rho_a]}\left(\|x - x_*\| > \frac{\Delta_a}{2\omega_a}\right) > \frac{1}{N}\right)\right] \\
&\overset{(ii)}{\leq} \frac{8e\omega_a^2}{m_a\Delta_a^2}\left(D_a + 2\Omega_a\log N\right),
\end{aligned}$$

where (i) is from our previous defined $\epsilon = \Delta_a/2$, and (ii) proceeds by inducing (37) with the choice of

$$n \geq \left\lceil \frac{8e\omega_1^2}{m_1\Delta_a^2}\left(D_1 + 2\Omega_1\log N\right)\right\rceil,$$

which completes the proof.

□

With the support of Lemma 3 and Lemma 4, the proof of Theorem 7 becomes straightforward and is detailed below.

Theorem 7 (Regret of exact Thompson sampling)**.** *For likelihood, reward distributions, and priors that meet Assumptions 1-3, and given that $\rho_a = \frac{\nu_a m_a^2}{8dL_a^3}$ holds for every $a \in \mathcal{A}$, the Thompson sampling with exact sampling yields the following total expected regrets after $N > 0$ rounds:*

$$\mathbb{E}[\mathfrak{R}(N)] \leq \sum_{a>1}\left[\frac{C_a}{\Delta_a}\left(\log B_a + d^2 + d\log N\right) + \frac{C_1}{\Delta_a}\sqrt{B_1}\left(\log B_1 + d^2\right) + 2\Delta_a\right],$$

where C_1 and C_a are universal constants and are unaffected by parameters specific to the problem.

Proof. From the above derivation, the total expected regrets by performing the exact Thompson sampling algorithm can be upper bounded by:

$$\begin{aligned}
\mathbb{E}[\mathfrak{R}(N)] &= \sum_{a>1}\mathbb{E}\left[\mathcal{L}_a(N)\right] \\
&\overset{(i)}{\leq} \sum_{a>1}\left\{\sum_{s=0}^{N-1}\mathbb{E}\left[\frac{1}{\mathcal{G}_{1,s}} - 1\right] + \Delta_a + \sum_{s=0}^{N-1}\mathbb{E}\left[\mathbb{I}\left(1 - \mathcal{G}_{a,s} > \frac{1}{N}\right)\right]\right\} \\
&\overset{(ii)}{\leq} \sum_{a>1}\left[\frac{8e\omega_a^2}{m_a\Delta_a}\left(D_a + 2\Omega_a\log N\right) + \sqrt{\kappa_1 B_1}\frac{128e\omega_1^2}{m_1\Delta_a}\left(D_1 + 2\Omega_1\right) + 2\Delta_a\right] \\
&= \sum_{a>1}\frac{16e\omega_a^2}{m_a\Delta_a}\left(\log B_a + \frac{4d}{\rho_a} + 16\left(\frac{L_a^2 d}{m_a\nu_a} + \frac{32}{\rho_a}\right)\log N\right) \\
&\quad + \sqrt{\kappa_1 B_1}\frac{256e\omega_1^2}{m_1\Delta_a}\left(\log B_1 + \frac{4d}{\rho_1} + 16\left(\frac{L_1^2 d}{m_1\nu_1} + \frac{32}{\rho_1}\right)\right) + 2\Delta_a \\
&\leq \sum_{a>1}\left[\frac{C_a}{\Delta_a}\left(\log B_a + d^2 + d\log N\right) + \frac{C_1}{\Delta_a}\sqrt{B_1}\left(\log B_1 + d^2\right) + 2\Delta_a\right]
\end{aligned}$$

where (i) holds by Lemma 1 and Lemma 2, the derivation of (ii) draws directly from the result established in Lemma 4. Through the appropriate selection of constants C_1 and C_a, the desired regret bound is consequently derived. □

C.5 Supporting Proofs for Exact Thompson Sampling

We start by presenting some propositions pertaining to the prior $\log \pi(x)$ and log-likelihood function $\log \mathrm{P}(\mathcal{R}|x)$, which serve as foundational elements for the previous proof of Theorem 6.

Proposition 1 (Proposition 2 in Mazumdar et al. (2020))**.** *If the prior distribution over $x \in \mathbb{R}^d$ satisfies Assumption 3, then the following statement holds for all x:*

$$\sup_x \langle \nabla \log \pi(x), x - x_* \rangle \leq \max_x [\log \pi(x)] - \log \pi(x_*).$$

Proof. From Assumption 3 we have:

$$\langle \nabla \log \pi(x), x - x_* \rangle \leq \log \pi(x) - \log \pi(x_*).$$

Taking the supremum of both sides completes the proof. □

Remark 1. *We here define $\log B := \max_x [\log \pi(x)] - \log \pi(x_*)$. When the prior is ideally centered around the point x_*, the implication is that $\log B = 0$. Consequently, B becomes a key parameter affecting our posterior concentration rates.*

We proceed to demonstrate that the empirical likelihood function evaluated at x_* can be characterized as a sub-Gaussian random variable.

Proposition 2 (Proposition 3 in Mazumdar et al. (2020))**.** *Suppose $f_n = \frac{1}{n}\sum_{j=1}^n \log \mathrm{P}(\mathcal{R}_j|x)$, where $\log \mathrm{P}(\mathcal{R}_j|x)$ follows Assumption 2. Then $\|\nabla f_n(x_*)\|_2$ is $L\sqrt{\frac{d}{n\nu}}$-sub-Gaussian random variable.*

Proof. We start by showing that $\nabla \log \mathrm{P}(\mathcal{R}|x_*)$ is $\frac{L}{\sqrt{\nu}}$ sub-Gaussian random variable. Let $\mathrm{u} \in \mathbb{S}_d$ be an arbitrary point in the d-dimensional sphere. Then we have:

$$\begin{aligned}
\left|\langle \nabla \log \mathrm{P}(\mathcal{R}_1|x_*), \mathrm{u}\rangle - \langle \nabla \log \mathrm{P}(\mathcal{R}_2|x_*), \mathrm{u}\rangle\right| &= \left|\langle \nabla \log \mathrm{P}(\mathcal{R}_1|x_*) - \nabla \log \mathrm{P}(\mathcal{R}_2|x_*), \mathrm{u}\rangle\right| \\
&\overset{(i)}{\leq} \left\|\nabla \log \mathrm{P}(\mathcal{R}_1|x_*) - \nabla \log \mathrm{P}(\mathcal{R}_2|x_*)\right\|_2 \|\mathrm{u}\|_2 \\
&= \left\|\nabla \log \mathrm{P}(\mathcal{R}_1|x_*) - \nabla \log \mathrm{P}(\mathcal{R}_2|x_*)\right\|_2 \\
&\overset{(ii)}{\leq} L\|\mathcal{R}_1 - \mathcal{R}_2\|
\end{aligned}$$

where (i) follows by Cauchy-Schwartz inequality, and (ii) by Assumption 2. Following an elementary invocation of Proposition 2.18 in Ledoux (2001), we arrive at the conclusion that $\langle \nabla \log \mathrm{P}(\mathcal{R}|x_*), \mathrm{u}\rangle$ is sub-Gaussian with parameter $\frac{L}{\sqrt{\nu}}$.

Given that the projection of $\nabla \log \mathrm{P}(\mathcal{R}|x_*)$ onto an arbitrary unit vector u adheres to sub-Gaussian properties with a $\frac{L}{\sqrt{\nu}}$-independent parameter, we deduce that the random vector $\nabla \log \mathrm{P}(\mathcal{R}|x_*)$ is also sub-Gaussian, characterized by the same parameter $\frac{L}{\sqrt{\nu}}$. It follows that $\nabla f_{a,n}(x_*)$, which is the mean of n independent and identically distributed (i.i.d.) sub-Gaussian vectors, remains sub-Gaussian with parameter $\frac{L}{\sqrt{n\nu}}$. Therefore, as a direct application of Lemma 1 from Jin et al. (2019), the vector $\nabla f_{a,n}(x_*)$ can be further identified as norm sub-Gaussian with parameter $L\sqrt{\frac{d}{n\nu}}$, which completes the proof. □

Lemma 5. *Let pairs (x_t, v_t) and (x_*, v_*) are both in $\mathbb{R}^{2d}$ and suppose $f(x) : \mathbb{R}^d \to \mathbb{R}$ fulfills Assumption 1, then the following inequality holds:*

$$\langle (\gamma - 1)\bar{v} + u(\nabla f(x_t) - \nabla f(x_*)), \bar{x} + \bar{v}\rangle - \langle \bar{v}, \bar{x}\rangle \geq \frac{m}{2L}\left(\|\bar{x}\|_2^2 + \|\bar{x} + \bar{v}\|_2^2\right), \tag{39}$$

where $\bar{x} = x_t - x_*$ and $\bar{v} = v_t - v_*$.

Proof. According to the Mean Value Theorem for Integrals, we derive the relationship between the gradient of $f(\cdot)$ and the Hessian matrix of $f(\cdot)$:

$$\begin{aligned}\nabla f(x_t) - \nabla f(x_*) &= \int_0^1 \frac{d}{dh} \nabla f(hx_t + (1-h)x_*) \\ &= \underbrace{\left[\int_0^1 \nabla^2 f(x_t + h(x_* - x_t))\, dh\right]}_{:=\psi_t} (x_t - x_*).\end{aligned}$$

Under the given Assumption 1, the Hessian matrix of $f(\cdot)$ is positive definite over the domain, with the inequality $m\mathbf{I}_{d\times d} \preceq \nabla^2 f(\cdot) \preceq L\mathbf{I}_{d\times d}$ holds. Next we follow the equations given in (5) to derive:

$$\begin{aligned}&\langle (\gamma-1)(v_t - v_*) + u(\nabla f(x_t) - \nabla f(x_*)), (x_t + v_t) - (x_* + v_*)\rangle - \langle x_t - x_*, v_t - v_*\rangle \\ &\overset{(i)}{=} [(x_t - x_*)^\intercal, (x_t - x_*)^\intercal + (v_t - v_*)^\intercal] \underbrace{\begin{bmatrix} \mathbf{I}_{d\times d} & -\mathbf{I}_{d\times d} \\ (1-\gamma)\mathbf{I}_{d\times d} + u\psi_t & (\gamma-1)\mathbf{I}_{d\times d}\end{bmatrix}}_{:=\mathrm{P}_t} \begin{bmatrix} x_t - x_* \\ (x_t - x_*) + (v_t - v_*)\end{bmatrix} \\ &\overset{(ii)}{=} [(x_t - x_*)^\intercal, (x_t - x_*)^\intercal + (v_t - v_*)^\intercal] \left(\frac{\mathrm{P}_t + \mathrm{P}_t^\intercal}{2}\right) \begin{bmatrix} x_t - x_* \\ (x_t - x_*) + (v_t - v_*)\end{bmatrix},\end{aligned} \tag{40}$$

where (i) can be obtained through simple matrix operations, and (ii) is because for any vector $(x, v) \in \mathbb{R}^{2d}$ and matrix $\mathrm{P} \in \mathbb{R}^{2d\times 2d}$, the quadratic form $(x, v)^\intercal \mathrm{P}(x, v)$ is equal to the from with a symmetric matrix $(x, v)^\intercal(\mathrm{P} + \mathrm{P}^\intercal)(x, v)/2$. We now focus on the analysis of the eigenvalues of the matrix $(\mathrm{P}_t + \mathrm{P}_t^\intercal)/2$. Taking the determinant of $(\mathrm{P}_t + \mathrm{P}_t^\intercal)/2 - \Lambda\mathbf{I}_{2d\times 2d}$ to be 0, we have:

$$\det\begin{bmatrix} (1-\Lambda)\mathbf{I}_{d\times d} & \frac{u\psi_t - \gamma\mathbf{I}_{d\times d}}{2} \\ \frac{u\psi_t - \gamma\mathbf{I}_{d\times d}}{2} & (\gamma - 1 - \Lambda)\mathbf{I}_{d\times d}\end{bmatrix} = 0.$$

From Lemma 6, we can rewrite the above equation as:

$$\det\left((1-\Lambda)(\gamma - 1 - \Lambda) - \tfrac{1}{4}\left(\gamma - u\tilde{\Lambda}_j\right)^2\right) = 0.$$

where $\tilde{\Lambda}_j$ $(j = 1, 2, \cdots, d)$ are the eigenvalues of the matrix ψ_t. According to Assumption 1, we know that for any j the inequality $0 < m \le \tilde{\Lambda}_j \le L$ holds. By the selection of $\gamma = 2$ and $u = \frac{1}{L}$ we have the solution to each eigenvalue Λ_j:

$$\Lambda_j = 1 \pm \left(1 - \frac{\tilde{\Lambda}_j}{2L}\right).$$

As the value of $\tilde{\Lambda}_j$ is between m and L, it can be concluded that the minimum eigenvalue among Λ_j is guaranteed to be greater than or equal to $\frac{m}{2L}$. Therefore, we have

$$\begin{aligned}&[(x_t - x_*)^\intercal, (x_t - x_*)^\intercal + (v_t - x_*)^\intercal] \begin{bmatrix} \mathbf{I}_{d\times d} & -\mathbf{I}_{d\times d} \\ (1-\gamma)\mathbf{I}_{d\times d} + u\psi_t & (\gamma-1)\mathbf{I}_{d\times d}\end{bmatrix} \begin{bmatrix} x_t - x_* \\ (x_t - x_*) + (v_t - x_*)\end{bmatrix} \\ =&[(x_t - x_*)^\intercal, (x_t - x_*)^\intercal + (v_t - x_*)^\intercal] \left(\frac{\mathrm{P}_t + \mathrm{P}_t^\intercal}{2}\right) \begin{bmatrix} x_t - x_* \\ (x_t - x_*) + (v_t - x_*)\end{bmatrix} \\ \ge&\frac{m}{2L}[(x_t - x_*)^\intercal, (x_t - x_*)^\intercal + (v_t - x_*)^\intercal] \begin{bmatrix} x_t - x_* \\ (x_t - x_*) + (v_t - x_*)\end{bmatrix} \\ =&\frac{m}{2L}\left(\|x_t - x_*\|_2^2 + \|(x_t + v_t) - (x_* + v_*)\|_2^2\right).\end{aligned}$$

By incorporating the findings from (40), we deduce the conclusive inequality. □

Lemma 6 (Theorem 3 in Silvester (2000))**.** *Considering square matrices* $\mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3$ *and* P_4 *with dimension d, and given the commutativity of* P_3 *and* P_4*, we can the following results:*

$$\det\left(\begin{bmatrix} \mathrm{P}_1 & \mathrm{P}_2 \\ \mathrm{P}_3 & \mathrm{P}_4\end{bmatrix}\right) = \det(\mathrm{P}_1\mathrm{P}_4 - \mathrm{P}_2\mathrm{P}_3)$$

Lemma 7 (Sandwich Inequality)**.** *Suppose $x, x_* \in \mathbb{R}^d$ and $v, v_* \in \mathbb{R}^d$ are position and velocity terms correspondingly. Then with the Euclidean norm, the following relationship holds:*

$$\|(x, v) - (x_*, v_*)\|_2 \leq 2\,\|(x, x + v) - (x_*, x_* + v_*)\|_2 \leq 4\,\|(x, v) - (x_*, v_*)\|_2\,. \tag{41}$$

Proof. We begin our analysis by the derivation of the first inequality:

$$\begin{aligned}
\|(x, v) - (x_*, v_*)\|_2^2 &= \|x - x_*\|_2^2 + \|v - v_*\|_2^2 \\
&= \|x - x_*\|_2^2 + \|(x + v - x) - (x_* + v_* - x_*)\|_2^2 \\
&\overset{(i)}{\leq} \|x - x_*\|_2^2 + 2\,\|(x + v) - (x_* + v_*)\|_2^2 + 2\,\|x - x_*\|_2^2 \\
&\leq 4\,\|(x, x + v) - (x_*, x_* + v_*)\|_2^2\,,
\end{aligned}$$

where (i) proceeds by Young's inequality. Then we move our attention to the second inequality:

$$\begin{aligned}
\|(x, x + v) - (x_*, x_* + v_*)\|_2^2 &= \|x - x_*\|_2^2 + \|(x + v) - (x_* + v_*)\|_2^2 \\
&\overset{(i)}{\leq} \|x - x_*\|_2^2 + 2\,\|x - x_*\|_2^2 + 2\,\|v - v_*\|_2^2 \\
&\leq 4\,\|(x, v) - (x_*, v_*)\|_2^2
\end{aligned}$$

where (i) is another application of Young's inequality. Taking square roots from the derived inequalities gives us the required results. □

D INTRODUCTION TO UNDERDAMPED LANGEVIN MONTE CARLO

We first construct a continuous-time Langevin dynamics to target the posterior concentration of $\mu_a^{(n)}$. The continuous-time underdamped Langevin dynamics is derived from the subsequent stochastic differential equations:

$$\begin{aligned} dv_t &= -\gamma v_t dt - u\nabla U(x_t)\, dt + \sqrt{2\gamma u} dB_t, \\ dx_t &= v_t dt. \end{aligned} \tag{42}$$

Suppose t is the time between rounds n and $n+1$, and we have a set of rewards up to round n: $\{\mathcal{R}_1, \mathcal{R}_1, \ldots, \mathcal{R}_n\}$, then we can express $U(x)$ using the following equation:

$$U(x_t) = \sum_{j=1}^{n} \mathbb{I}\Big(A_j = a\Big) \log \mathrm{P}_a(\mathcal{R}_j | x_t) + \log \pi_a(x_t), \tag{43}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Underdamped Langevin dynamics is characterized by its invariant distribution, notably proportional to $e^{-U(x_t)}$, which enables the sampling of the unscaled posterior distribution $\mu_a^{(n)}$. By designating the potential function as (43), we obtain a continuous-time dynamic that guarantees trajectories converging rapidly toward the posterior distribution $\mu_a^{(n)}$. To transition from theoretical continuous-time dynamics to a practical algorithm, we integrate the discrete underdamped Langevin dynamics to derive (42). Namely, for time t between round n and round $n+1$, we uniformly discretize the time as I segments, where $i \in [\![1, I]\!]$ denote the step-index between rounds n and $n+1$, which leads to the following underdamped Langevin Monte Carlo:

$$\begin{bmatrix} x_{i+1} \\ v_{i+1} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbb{E}[x_{i+1}] \\ \mathbb{E}[v_{i+1}] \end{bmatrix}, \begin{bmatrix} \mathbb{V}(x_{i+1}) & \mathbb{K}(x_{i+1}, v_{i+1}) \\ \mathbb{K}(v_{i+1}, x_{i+1}) & \mathbb{V}(v_{i+1}) \end{bmatrix} \right), \tag{44}$$

where x_i, v_i are positions and velocities at step i, $\mathbb{E}[x_{i+1}]$, $\mathbb{E}[v_{i+1}]$, $\mathbb{V}(x_{i+1})$, $\mathbb{V}(v_{i+1})$, and $\mathbb{K}(v_{i+1}, x_{i+1})$ are obtained from the following computations:

$$\begin{aligned} \mathbb{E}[v_{i+1}] &= v_i e^{-\gamma h} - \frac{u}{\gamma}(1 - e^{-\gamma h})\nabla U(x_i) \\ \mathbb{E}[x_{i+1}] &= x_i + \frac{1}{\gamma}(1 - e^{-\gamma h}) v_i - \frac{u}{\gamma}\left(h - \frac{1}{\gamma}\left(1 - e^{-\gamma h}\right)\right)\nabla U(x_i) \\ \mathbb{V}(x_{i+1}) &= \mathbb{E}\Big[(x_{i+1} - \mathbb{E}[x_{i+1}])(x_{i+1} - \mathbb{E}[x_{i+1}])^\top\Big] = \frac{2u}{\gamma}\left[h - \frac{1}{2\gamma}e^{-2\gamma h} - \frac{3}{2\gamma} + \frac{2}{\gamma}e^{-\gamma h}\right] \cdot \mathbf{I}_{d\times d} \\ \mathbb{V}(v_{i+1}) &= \mathbb{E}\Big[(v_{i+1} - \mathbb{E}[v_{i+1}])(v_{i+1} - \mathbb{E}[v_{i+1}])^\top\Big] = u(1 - e^{-2\gamma h}) \cdot \mathbf{I}_{d\times d} \\ \mathbb{K}(x_{i+1}, v_{i+1}) &= \mathbb{E}\Big[(x_{i+1} - \mathbb{E}[x_{i+1}])(v_{i+1} - \mathbb{E}[v_{i+1}])^\top\Big] = \frac{u}{\gamma}\left[1 + e^{-2\gamma h} - 2e^{-\gamma h}\right] \cdot \mathbf{I}_{d\times d}, \end{aligned}$$

where h is the step size of the algorithm. A detailed proof can be found in Appendix A of Cheng et al. (2018). Within this update rule, the gradient of the potential function, $\nabla U(x_i)$, is proportional to the dataset size n:

$$\nabla U(x_i) = \sum_{j=1}^{\mathcal{L}_a(n)} \nabla \log \mathrm{P}_a\left(\mathcal{R}_{a,j} | x_i\right) + \nabla \log \pi_a(x_i). \tag{45}$$

To address the increasing terms in $\nabla U(x_i)$, we adopt stochastic gradient methods. Specifically, we define the calculation of the stochastic gradient, $\hat{U}(x_i)$, as:

$$\nabla \hat{U}(x_i) = \frac{\mathcal{L}_a(n)}{|\mathcal{S}|} \sum_{\mathcal{R}_{a,j} \in \mathcal{S}} \nabla \log \mathrm{P}_a\left(\mathcal{R}_{a,j} | x_i\right) + \nabla \log \pi_a(x_i), \tag{46}$$

with $\mathcal{S}$ representing a subset of the dataset. Typically, $\mathcal{S}$ is obtained through subsampling from $\{\mathcal{R}_{a,1}, \cdots, \mathcal{R}_{a,j} \cdots, \mathcal{R}_{a,\mathcal{L}_a(n)}\}$. We further clarify that the cardinality of $\mathcal{S}$, denoted as $|\mathcal{S}|$ or k, is the batch size for the stochastic gradient estimate. It is noteworthy that, early in the Thompson sampling algorithm, the round n being played might be less than the designated batch size k. Consequently, we redefine the batch size to be $|\mathcal{S}| = \min\{\mathcal{L}_a(n), k\}$. Incorporating the stochastic gradient $\nabla \hat{U}$ as a replacement for the full gradient ∇U within our update procedures yields the formulation of the Stochastic Gradient underdamped Langevin Monte Carlo, as is shown in Algorithm 4:

Algorithm 4 (Stochastic Gradient) underdamped Langevin Monte Carlo for arm a at round n.

Input Data $\{\mathcal{R}_{a,1}, \mathcal{R}_{a,2}, \cdots, \mathcal{R}_{a,\mathcal{L}_a(n-1)}\}$;
Input Sample $(x_{a,Ih^{(n-1)}}, v_{a,Ih^{(n-1)}})$ from last round;
Initialize $x_0 = x_{a,Ih^{(n-1)}}$ and $v_0 = v_{a,Ih^{(n-1)}}$.
for $i = 0, 1, \cdots, I-1$ **do**
    Uniformly subsample data set $\mathcal{S} \subseteq \{\mathcal{R}_{a,1}, \mathcal{R}_{a,2}, \cdots, \mathcal{R}_{a,\mathcal{L}_a(n-1)}\}$.
    Compute $\nabla U(x_i)$ with (45) (or $\nabla \hat{U}(x_i)$ with (46)).
    Sample (x_{i+1}, v_{i+1}) with (44) based on $\nabla U(x_i)$ ($\nabla \hat{U}(x_i)$) and (x_i, v_i).
end for
$x_{a,Ih^{(n)}} \sim \mathcal{N}\left(x_I, \frac{1}{nL_a\rho_a}\mathbf{I}_{d\times d}\right)$ and $v_{a,Ih^{(n)}} = v_I$
Output Sample $(x_{a,Ih^{(n)}}, v_{a,Ih^{(n)}})$ from current round;

E ANALYSIS OF APPROXIMATE THOMPSON SAMPLING

This section initially investigates the potential for underdamped Langevin Monte Carlo to accelerate the convergence rate in Thompson sampling, incorporating a quantitative examination of the required sample complexity for effective posterior approximation. Next, we evaluate the concentration properties of approximate samples yielded by underdamped Langevin Monte Carlo, covering both full and stochastic gradient cases. At the end of this section, we provide the regret analysis for the proposed Thompson sampling with underdamped Langevin Monte Carlo, under the condition of effective posterior sample approximation to reach sub-linear regrets.

Similar to our earlier analysis on exact Thompson sampling, we begin by providing a summary table outlining the notation that will be invoked in the following sections.

E.1 Notation

To further our investigations into approximate Thompson sampling, we have to introduce additional notations, as shown in Table 2. This table elaborates the symbols in the analysis of underdamped Langevin Monte Carlo and the regret analysis of approximate Thompson sampling. As delineated in Section D, underdamped Langevin Monte Carlo encompasses two algorithmic versions based on gradient estimation: the full gradient and the stochastic gradient versions. The symbols in full gradient version are represented using the "˜" notation (as in $\tilde{x}$, $\tilde{v}$, $\tilde{\mu}$), and the stochastic gradient version adopts the "ˆ" notation. It should be noted that, for regret analysis of approximate Thompson sampling, we ensure a congruent posterior concentration rate across both versions, facilitated by an informed choice of step size and number of samples.

Table 2: Additional notation in the analysis of approximate Thompson sampling.

Symbols	Explanations
i	number of steps for the current round n ($i \in [\![1, I]\!]$)
k	batch size for the stochastic gradient estimate
ζ	optimal couplings between two measures
$\bar{\rho}_a$	parameter to scale approximate posterior for arm a
$h^{(n)}$	step size of the approximate sampling algorithm for round n from one of the arm
$\delta(\cdot)$	Dirac delta distribution
$\tilde{x}_a$ ($\tilde{v}_a$)	sampled position (velocity) of arm a follows underdamped Langevin Monte Carlo with full gradient
$\hat{x}_a$ ($\hat{v}_a$)	sampled position (velocity) of arm a follows underdamped Langevin Monte Carlo with stochastic gradient
$\tilde{\mu}_a^{(n)}$ $\left(\hat{\mu}_a^{(n)}\right)$	probability measure of posterior distribution after n rounds approximated by underdamped Langevin Monte Carlo
$\tilde{\mu}_{ih^{(n)}}$ $(\hat{\mu}_{ih^{(n)}})$	probability measure of posterior distribution at ith step of round n approximated by underdamped Langevin Monte Carlo
$\tilde{\mu}_a^{(n)}[\bar{\rho}_a]$ $\left(\hat{\mu}_a^{(n)}[\bar{\rho}_a]\right)$	probability measure of scaled posterior distribution after n rounds approximated by underdamped Langevin Monte Carlo

E.2 Posterior Convergence Analysis

When the likelihood meets Assumption 4, complemented by the prior conforming to Assumption 3, we can employ underdamped Langevin Monte Carlo for the sampling process, ensuring convergence within the 2-Wasserstein distance. Algorithm 4 indicates that for n-th round, the sample starts from the last step of the $(n-1)$-th round and ends on the last step (step I) of the n-th round. Based on the above analysis, we provide convergence guarantees among rounds from 1 to N.

Theorem 8 (Posterior Convergence of underdamped Langevin Monte Carlo with full gradient)**.** *Suppose the log-likelihood function follows Assumption 4, the prior follows Assumption 3, and the posterior distribution fulfills the concentration inequality* $\mathbb{E}_{(x,v)\sim\mu_a^{(n)}}\left[\|(x,v)-(x_*,v_*)\|_2^2\right]^{\frac{1}{2}} \leq \frac{1}{\sqrt{n}}\tilde{D}_a$.

By selecting the step size $h^{(n)} = \frac{\tilde{D}_a}{80\kappa_a}\sqrt{\frac{m_a}{nd}} = \tilde{O}\left(\frac{1}{\sqrt{d}}\right)$ *and the number of steps* $I \geq \frac{2\kappa_a}{h^{(n)}}\log 24 = \tilde{O}\left(\sqrt{d}\right)$*, we are able to bound*

the convergence of Algorithm 4 in 2-Wasserstein distance w.r.t. the posterior distribution $\mu_a^{(n)}$*:* $W_2\left(\tilde{\mu}_a^{(n)}, \mu_a^{(n)}\right) \leq \frac{2}{\sqrt{n}}\tilde{D}_a$*, where* $\tilde{D}_a \geq 8\sqrt{\frac{d}{m_a}}$.

Proof. To prove Theorem 8, we employ an inductive approach, with the starting point being $n = 1$. When $n = 1$, we invoke Theorem 14 with initial condition $\mu_0 = \delta(x_0, v_0)$ to have the convergence of Algorithm 4 in the i-th iteration after the first pull to arm a:

$$\begin{aligned}
W_2\left(\tilde{\mu}_a^{(1)}, \mu_a^{(1)}\right) &= W_2\left(\tilde{\mu}_{ih^{(1)}}, \mu_a^{(1)}\right) \\
&\overset{(i)}{\leq} 4e^{-ih^{(1)}/2\kappa_a} W_2\left(\delta(x_0, v_0), \mu_a^{(1)}\right) + \frac{20\left(h^{(1)}\right)^2}{1 - e^{-h^{(1)}/2\kappa_a}}\sqrt{\frac{d}{m_a}} \\
&\overset{(ii)}{\leq} 4e^{-ih^{(1)}/2\kappa_a}\left[W_2\left(\delta(x_*, v_*), \mu_a^{(1)}\right) + W_2\left(\delta(x_0, v_0), \delta(x_*, v_*)\right)\right] + \frac{20\left(h^{(1)}\right)^2}{1 - e^{-h^{(1)}/2\kappa_a}}\sqrt{\frac{d}{m_a}} \\
&\overset{(iii)}{\leq} 4e^{-ih^{(1)}/2\kappa_a}\left(6\sqrt{\frac{d}{m_a}} + \sqrt{\frac{6d}{m_a}}\right) + \frac{20\left(h^{(1)}\right)^2}{1 - e^{-h^{(1)}/2\kappa_a}}\sqrt{\frac{d}{m_a}} \\
&\overset{(iv)}{\leq} 24\sqrt{2}e^{-ih^{(1)}/2\kappa_a}\sqrt{\frac{d}{m_a}} + 80h^{(1)}\kappa_a\sqrt{\frac{d}{m_a}}
\end{aligned} \tag{47}$$

where (i) holds by the results of Theorem 14, (ii) is from triangle inequality, (iii) is from Lemma 14 and Lemma 16, (iv) is because $6 + \sqrt{6} < 6\sqrt{2}$ and $1/(1 - e^{-h^{(1)}/2\kappa_a}) \leq 4\kappa_a/h^{(1)}$ holds when $h^{(1)}/\kappa_a < 1$. With the selection of $h^{(1)} = \frac{\tilde{D}_a}{80\kappa_a}\sqrt{\frac{m_a}{d}}$ and $I \geq \frac{2\kappa_a}{h^{(1)}}\log\left(\frac{24}{\tilde{D}_a}\sqrt{\frac{2d}{m_a}}\right)$, we can guarantee that Algorithm 4 converge to $2\tilde{D}_a$.

After pulling arm a for the $(n-1)$-th round and before the n-th round, following the result given by (47), we know that $W_2\left(\tilde{\mu}_a^{(n-1)}, \mu_a^{(n-1)}\right) \leq \frac{2}{\sqrt{n-1}}\tilde{D}_a$ can be guaranteed. We now continue to prove that, post the n-th pull, the constraint further refines to $W_2\left(\tilde{\mu}_a^{(n)}, \mu_a^{(n)}\right) \leq \frac{2}{\sqrt{n}}\tilde{D}_a$. From the above analysis, we have:

$$\begin{aligned}
W_2\left(\tilde{\mu}_a^{(n)}, \mu_a^{(n)}\right) &= W_2\left(\tilde{\mu}_{ih^{(n)}}, \mu_a^{(n)}\right) \\
&\leq 4e^{-ih^{(n)}/2\kappa_a} W_2\left(\tilde{\mu}_a^{(n-1)}, \mu_a^{(n)}\right) + \frac{20\left(h^{(n)}\right)^2}{1 - e^{-h^{(n)}/2\kappa_a}}\sqrt{\frac{d}{m_a}} \\
&\leq 4e^{-ih^{(n)}/2\kappa_a}\left[W_2\left(\tilde{\mu}_a^{(n-1)}, \mu_a^{(n-1)}\right) + W_2\left(\mu_a^{(n-1)}, \mu_a^{(n)}\right)\right] + \frac{20\left(h^{(n)}\right)^2}{1 - e^{-h^{(n)}/2\kappa_a}}\sqrt{\frac{d}{m_a}} \\
&\overset{(i)}{\leq} 4e^{-ih^{(n)}/2\kappa_a}\left[\frac{3}{\sqrt{n}}\tilde{D}_a + W_2\left(\mu_a^{(n-1)}, \delta(x_*, v_*)\right) + W_2\left(\delta(x_*, v_*), \mu_a^{(n)}\right)\right] + \frac{20\left(h^{(n)}\right)^2}{1 - e^{-h^{(n)}/2\kappa_a}}\sqrt{\frac{d}{m_a}} \\
&\overset{(ii)}{\leq} 24e^{-ih^{(n)}/2\kappa_a}\frac{\tilde{D}_a}{\sqrt{n}} + \frac{20\left(h^{(n)}\right)^2}{1 - e^{-h^{(n)}/2\kappa_a}}\sqrt{\frac{d}{m_a}} \\
&\leq 24e^{-ih^{(n)}/2\kappa_a}\frac{\tilde{D}_a}{\sqrt{n}} + 80h^{(n)}\kappa_a\sqrt{\frac{d}{m_a}},
\end{aligned} \tag{48}$$

where (i) is because $W_2\left(\tilde{\mu}_a^{(n-1)}, \mu_a^{(n-1)}\right) \leq \frac{2}{\sqrt{n-1}}\tilde{D}_a \leq \frac{3}{\sqrt{n}}\tilde{D}_a$, (ii) proceeds by our posterior assumption:

$$W_2\left(\mu_a^{(n-1)}, \delta(x_*)\right) + W_2\left(\delta(x_*), \mu_a^{(n)}\right) \leq \left(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{n}}\right)\tilde{D}_a \leq \frac{3}{\sqrt{n}}\tilde{D}_a.$$

With the selection of $h^{(n)} = \frac{\tilde{D}_a}{80\kappa_a}\sqrt{\frac{m_a}{nd}} = \tilde{O}\left(\frac{1}{\sqrt{d}}\right)$ and $I \geq \frac{2\kappa_a}{h^{(n)}}\log 24 = \tilde{O}\left(\sqrt{d}\right)$, we can guarantee that the approximate error after $(n-1)$-th round and before n-th round can be upper bounded by $2\tilde{D}_a/\sqrt{n}$. □

Given the log-likelihood function also satisfies Assumption 5, we observe similar convergence guarantees for the stochastic gradient version of the underdamped Langevin Monte Carlo.

Theorem 9 (Posterior Convergence of underdamped Langevin Monte Carlo with stochastic gradient)**.** *Given that the log-likelihood aligns with Assumptions 4 and 5, and the prior adheres to Assumption 3, we define the parameters for Algorithm 4 as follows: stochastic gradient samples are taken as* $k = \tilde{O}\left(\kappa_a^2\right)$*, step size as* $h^{(n)} = \tilde{O}\left(\frac{1}{\sqrt{d}}\right)$*, and the total steps count as* $I = \tilde{O}\left(\sqrt{d}\right)$*. Given the posterior distribution follows the prescribed concentration inequality* $\mathbb{E}_{(x,v)\sim\mu_a^{(n)}}\left[\|(x,v)-(x_*,v_*)\|_2^2\right]^{\frac{1}{2}} \leq \frac{1}{\sqrt{n}}\hat{D}_a$*, we have convergence of the underdamped Langevin Monte Carlo in 2-Wasserstein distance to the posterior* $\mu_a^{(n)}$*:* $W_2\left(\hat{\mu}_a^{(n)}, \mu_a^{(n)}\right) \leq \frac{2}{\sqrt{n}}\hat{D}_a$*, where* $\hat{D}_a \geq 8\sqrt{\frac{d}{m_a}}$.

Proof. To demonstrate this theorem, we employ an inductive methodology as outlined in Theorem 8. Specifically, for $n = 1$, we examine the convergence after i-th iterations to the first pull of arm a:

$$
\begin{aligned}
W_2\left(\hat{\mu}_a^{(1)}, \mu_a^{(1)}\right) &= W_2\left(\hat{\mu}_{ih^{(1)}}, \mu_a^{(1)}\right) \\
&\overset{(i)}{\leq} 4e^{-ih^{(1)}/2\kappa_a} W_2\left(\delta(x_0,v_0), \mu_a^{(1)}\right) + \frac{4}{1-e^{-h^{(1)}/2\kappa_a}}\left(5\left(h^{(1)}\right)^2\sqrt{\frac{d}{m_a}} + 4uh^{(1)}L_a\sqrt{\frac{nd}{k_a\nu_a}}\right) \\
&\overset{(ii)}{\leq} 4e^{-ih^{(1)}/2\kappa_a}\left[W_2\left(\delta(x_*,v_*), \mu_a^{(1)}\right) + W_2\left(\delta(x_0,v_0), \delta(x_*,v_*)\right)\right] \\
&\quad + \frac{4}{1-e^{-h^{(1)}/2\kappa_a}}\left(5\left(h^{(1)}\right)^2\sqrt{\frac{d}{m_a}} + 4uh^{(1)}L_a\sqrt{\frac{nd}{k_a\nu_a}}\right) \\
&\overset{(iii)}{\leq} 4e^{-ih^{(1)}/2\kappa_a}\left(6\sqrt{\frac{d}{m_a}} + \sqrt{\frac{6d}{m_a}}\right) + \frac{4}{1-e^{-h^{(1)}/2\kappa_a}}\left(5\left(h^{(1)}\right)^2\sqrt{\frac{d}{m_a}} + 4uh^{(1)}L_a\sqrt{\frac{nd}{k_a\nu_a}}\right) \\
&\overset{(iv)}{\leq} 24\sqrt{2}e^{-ih^{(1)}/2\kappa_a}\sqrt{\frac{d}{m_a}} + 80\kappa_a h^{(n)}\sqrt{\frac{d}{m_a}} + 64\kappa_a u L_a\sqrt{5\frac{nd}{k\nu_a}},
\end{aligned}
\tag{49}
$$

where (i) holds by the results of Theorem 15, (ii) is from triangle inequality, (iii) follows Lemma 14 and Lemma 16, (iv) is because $6 + \sqrt{6} < 6\sqrt{2}$ and $1/(1-e^{-h^{(1)}/2\kappa_a}) \leq 4\kappa_a/h$ hold when $h^{(1)}/\kappa_a < 1$. With the selection of $h^{(1)} = \frac{\tilde{D}_a}{120\kappa_a}\sqrt{\frac{m_a}{nd}}$ and $I \geq \frac{2\kappa_a}{h^{(1)}}\log\left(\frac{36}{\tilde{D}_a}\sqrt{\frac{2d}{m_a}}\right)$, we can guarantee that the first two terms converges to $\frac{2\hat{D}_a}{3}$. If we select $k = 720\frac{L_a^2}{m_a\nu_a}$, then we can upper bound this term by $\frac{2\hat{D}_a}{3}$:

$$
64\kappa_a u L_a\sqrt{\frac{5d}{k\nu_a}} = \frac{16}{3}\sqrt{\frac{d}{m_a}} \leq \frac{2\hat{D}_a}{3},
$$

which means Algorithm 4 converge to $2\hat{D}_a$. With this upper bound, we continue to bound the distance between $\hat{\mu}_a^{(n)}$ and $\mu_a^{(n)}$

given $W_2\left(\hat{\mu}_a^{(n-1)}, \mu_a^{(n-1)}\right) \leq \frac{2}{\sqrt{n-1}}\hat{D}_a$:

$$
\begin{aligned}
W_2\left(\hat{\mu}_a^{(n)}, \mu_a^{(n)}\right) &= W_2\left(\hat{\mu}_{ih^{(n)}}, \mu_a^{(n)}\right) \\
&\overset{(i)}{\leq} 4e^{-ih^{(n)}/2\kappa_a} W_2\left(\hat{\mu}_a^{(n-1)}, \mu_a^{(n)}\right) + \frac{4}{1-e^{-h^{(n)}/2\kappa_a}}\left(5\left(h^{(n)}\right)^2\sqrt{\frac{d}{m_a}} + 4uh^{(n)}L_a\sqrt{\frac{nd}{k_a\nu_a}}\right) \\
&\leq 4e^{-ih^{(n)}/2\kappa_a}\left[W_2\left(\hat{\mu}_a^{(n-1)}, \mu_a^{(n-1)}\right) + W_2\left(\mu_a^{(n-1)}, \mu_a^{(n)}\right)\right] + \frac{4}{1-e^{-h^{(n)}/2\kappa_a}}\left(5\left(h^{(n)}\right)^2\sqrt{\frac{d}{m_a}} + 4uh^{(n)}L_a\sqrt{\frac{nd}{k_a\nu_a}}\right) \\
&\overset{(ii)}{\leq} 4e^{-ih^{(n)}/2\kappa_a}\left[\frac{3}{\sqrt{n}}\hat{D}_a + W_2\left(\mu_a^{(n-1)}, \delta(x_*, v_*)\right) + W_2\left(\delta(x_*, v_*), \mu_a^{(n)}\right)\right] \\
&\quad + \frac{4}{1-e^{-h^{(n)}/2\kappa_a}}\left(5\left(h^{(n)}\right)^2\sqrt{\frac{d}{m_a}} + 4uh^{(n)}L_a\sqrt{\frac{nd}{k_a\nu_a}}\right) \\
&\overset{(iii)}{\leq} 24e^{-ih^{(n)}/2\kappa_a}\frac{\hat{D}_a}{\sqrt{n}} + \frac{4}{1-e^{-h^{(n)}/2\kappa_a}}\left(5\left(h^{(n)}\right)^2\sqrt{\frac{d}{m_a}} + 4uh^{(n)}L_a\sqrt{\frac{nd}{k_a\nu_a}}\right) \\
&\leq 24e^{-ih^{(n)}/2\kappa_a}\frac{\hat{D}_a}{\sqrt{n}} + \underbrace{80\kappa_a h^{(n)}\sqrt{\frac{d}{m_a}}}_{T1} + \underbrace{64\kappa_a uL_a\sqrt{5\frac{nd}{k\nu_a}}}_{T2},
\end{aligned}
$$

where (i) is from Theorem 15, (ii) and (iii) follow the same idea as (i) and (ii) in (48).

For terms $T1$ and $T2$, if we take step size $h^{(n)} = \frac{\hat{D}_a}{120\kappa_a}\sqrt{\frac{m_a}{nd}} = \tilde{O}\left(\frac{1}{\sqrt{d}}\right)$ and $k = 720\frac{L_a^2}{m_a\nu_a}$, we will have $T1 \leq \frac{2\hat{D}_a}{3\sqrt{n}}$ and $T2 \leq \frac{2\hat{D}_a}{3\sqrt{n}}$. If the number of steps taken in the SGLD algorithm from $(n-1)$-th pull till n-th pull is selected as $I \geq \frac{2\kappa_a}{h^{(n)}}\log 16 = \tilde{O}\left(\sqrt{d}\right)$, we then have

$$
\begin{aligned}
W_2\left(\hat{\mu}_a^{(n)}, \mu_a^{(n)}\right) &\leq 24e^{Ih^{(n)}/2\kappa_a}\frac{\hat{D}_a}{\sqrt{n}} + 8\kappa_a uL_a\sqrt{5\frac{nd}{k\nu_a}} + 16\kappa_a h^{(n)}\sqrt{\frac{104}{5}\frac{d}{m}} \\
&\leq \frac{2}{\sqrt{n}}\hat{D}_a.
\end{aligned}
$$

As we move from the $(n-1)$-th pull to the n-th for arm a and given that the first round has been proved to converge to $2\hat{D}_a$, the number of steps for the n-th round can be determined as $\tilde{O}\left(\sqrt{d}\right)$. □

E.3 Posterior Concentration Analysis

Based on the convergence analysis of the 2-Wasserstein distance metrics in underdamped Langevin Monte Carlo, we continue to deduce the following posterior concentration bounds of the algorithm.

Theorem 10. *Under conditions where the log-likelihood and the prior are consistent with Assumptions 3 and 4, and considering that arm a has been chosen $\mathcal{L}_a(n)$ times till the nth round of the Thompson sampling procedure. By setting the step size $h^{(n)} = \tilde{O}\left(\frac{1}{\sqrt{d}}\right)$ and number of steps $I = \tilde{O}\left(\sqrt{d}\right)$ in Algorithm 4, then the following concentration inequality holds:*

$$
\mathbb{P}_{x_{a,n}\sim\tilde{\mu}_a^{(n)}[\bar{\rho}_a]}\left(\|x_{a,n} - x_*\|_2 > 6\sqrt{\frac{e}{m_a n}\left(D_a + 2\Omega_a\log 1/\delta_1 + 2\tilde{\Omega}_a\log 1/\delta_2\right)}\middle|\mathcal{Z}_{n-1}\right) < \delta_2, \tag{50}
$$

where $D_a = 2\log B_a + 8d$, $\Omega_a = 256 + \frac{16dL_a^2}{m_a\nu_a}$, $\tilde{\Omega}_a = 256 + \frac{16dL_a^2}{m_a\nu_a} + \frac{m_a d}{18L_a\bar{\rho}_a}$, and we define $\mathcal{Z}_{n-1}$ as the following set:

$$
\mathcal{Z}_{n-1} = \left\{\left\|x_{a,n-1} - x_*\right\|_2 \leq \beta(n)\right\},
$$

where $x_{a,n-1}$ is from the output of Algorithm 4 from the round $n-1$, and $\beta : \mathbb{N} \to (0,\infty)$ is a constructed confidence interval related to round n:

$$
\beta(n) \coloneqq 3\sqrt{\frac{2e}{m_a n}\left(D_a + 2\Omega_a\log 1/\delta_1\right)}.
$$

Remark 2. *Using extensions from Theorems 6 and 8, we can establish the upper bound for $\beta(n)$: for arm a in round n, Theorem 6 provides an upper bound $\sqrt{\frac{2e}{m_a n}\left(D_a + 2\Omega_a\log 1/\delta_1\right)}$ for the posterior concentration, while Theorem 8 caps the distance between the posterior and its approximation as $\sqrt{\frac{8e}{m_a n}\left(D_a + 2\Omega_a\log 1/\delta_1\right)}$.*

Proof. We start by analyzing the upper bound of the 2-Wasserstein distance between $\tilde{\mu}_a^{(n)}[\bar{\rho}_a]$ and $\delta(x_*, v_*)$, and then conclude the posterior concentration rates by taking the marginal posterior distribution. By triangle inequality, we can separate the distance into three terms and we will upper bound these terms one by one:

$$W_2\left(\tilde{\mu}_a^{(n)}[\bar{\rho}_a], \delta(x_*, v_*)\right) \leq \underbrace{W_2\left(\tilde{\mu}_a^{(n)}[\bar{\rho}_a], \tilde{\mu}_{Ih^{(n)}}\right)}_{T1} + \underbrace{W_2\left(\tilde{\mu}_{Ih^{(n)}}, \mu_a^{(n)}\right)}_{T2} + \underbrace{W_2\left(\mu_a^{(n)}, \delta(x_*, v_*)\right)}_{T3}$$

For term $T1$, since the sample returned by the Langevin Monte Carlo is given by: $\tilde{x}_a = \tilde{x}_{Ih^{(n)}} + \mathcal{X}_a$, $\tilde{v}_a = \tilde{v}_{Ih^{(n)}}$, where $\mathcal{X}_a \sim \mathcal{N}\left(0, \frac{1}{nL_a\bar{\rho}_a}\mathbf{I}_{d\times d}\right)$, it remains to bound the distance between the approximate posterior $\tilde{\mu}_a^{(n)}$ of x_a and the distribution of $x_{Ih^{(n)}}$. Since $\tilde{x}_a - \tilde{x}_{Ih^{(n)}}$ follows a normal distribution, we can upper bound $T1$ by taking expectation of $\mathcal{X}_a$:

$$\begin{aligned}
T1 = W_2\left(\tilde{\mu}_a^{(n)}[\bar{\rho}_a], \mu_{Ih^{(n)}}\right) &= \left(\inf_{\bar{\rho}\in\Gamma\left(\tilde{\mu}_a^{(n)}[\bar{\rho}_a], \mu_{Ih^{(n)}}\right)} \int \|\tilde{x}_a - \tilde{x}_{Ih^{(n)}}\|_2^2 \, d\tilde{x}_a d\tilde{x}_{Ih^{(n)}}\right)^{1/2} \\
&\leq \sqrt{\mathbb{E}\left[\|\tilde{x}_a - \tilde{x}_{Ih^{(n)}}\|_2^2\right]} \\
&\overset{(i)}{\leq} \mathbb{E}\left[\|\tilde{x}_a - \tilde{x}_{Ih^{(n)}}\|^p\right]^{1/p} \\
&\overset{(ii)}{\leq} \sqrt{\frac{d}{nL_a\bar{\rho}_a}}\left(\frac{2^{p/2}\mathbf{\Gamma}\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}\right)^{1/p} \\
&\overset{(iii)}{\leq} \sqrt{\frac{dp}{nL_a\bar{\rho}_a}},
\end{aligned} \tag{51}$$

where (i) follows a simple application of the Hölder's inequality for any even integer $p \geq 2$, (ii) proceeds by Stirling's approximation for the Gamma function (Boas, 2006), and (iii) holds because $\mathbf{\Gamma}\left(\frac{p+1}{2}\right) \leq \sqrt{\pi}\left(\frac{p}{2}\right)^{p/2}$. We next bound $T2$ following the idea from Theorem 8:

$$\begin{aligned}
T2 = W_2\left(\tilde{\mu}_{Ih^{(n)}}, \mu_a^{(n)}\right) &\overset{(i)}{\leq} 4e^{-Ih^{(n)}/2\kappa_a} W_2\left(\delta(x_{a,n-1}, v_{a,n-1}), \mu_a^{(n)}\right) + \frac{20\left(h^{(n)}\right)^2}{1-e^{-h^{(n)}/2\kappa_a}}\sqrt{\frac{d}{m_a}} \\
&\overset{(ii)}{\leq} 4e^{-Ih^{(n)}/2\kappa_a}\left[W_2\left(\delta(x_{a,n-1}, v_{a,n-1}), \delta(x_*, v_*)\right) + W_2\left(\delta(x_*, v_*), \mu_a^{(n)}\right)\right] + \frac{20\left(h^{(n)}\right)^2}{1-e^{-h^{(n)}/2\kappa_a}}\sqrt{\frac{d}{m_a}} \\
&\overset{(iii)}{\leq} 4e^{-Ih^{(n)}/2\kappa_a}\left[W_2\left(\delta(x_{a,n-1}, v_{a,n-1}), \mu_a^{(n-1)}\right) + W_2\left(\mu_a^{(n-1)}, \delta(x_*, v_*)\right) + \frac{\tilde{D}'_a}{\sqrt{n}}\right] + \frac{20\left(h^{(n)}\right)^2}{1-e^{-h^{(n)}/2\kappa_a}}\sqrt{\frac{d}{m_a}} \\
&\overset{(iv)}{\leq} 4e^{-Ih^{(n)}/2\kappa_a}\left[\frac{3}{\sqrt{n-1}}\tilde{D}_a + \frac{\tilde{D}'_a}{\sqrt{n}}\right] + \frac{20\left(h^{(n)}\right)^2}{1-e^{-h^{(n)}/2\kappa_a}}\sqrt{\frac{d}{m_a}} \\
&\overset{(v)}{\leq} 4e^{-Ih^{(n)}/2\kappa_a}\frac{7}{\sqrt{n}}\bar{D}_a + \frac{20\left(h^{(n)}\right)^2}{1-e^{-h^{(n)}/2\kappa_a}}\sqrt{\frac{d}{m_a}} \\
&\overset{(vi)}{\leq} \frac{2\bar{D}_a}{\sqrt{n}},
\end{aligned} \tag{52}$$

where (i) is derived from Theorem 14, (ii) proceeds by triangle inequality, (iii) arise from the concentration rate for arm a at round n as delineated in Theorem 6:

$$W_2\left(\delta(x_*, v_*), \mu_a^{(n)}\right) \leq \frac{\tilde{D}'_a}{\sqrt{n}} = \sqrt{\frac{2}{m_a n}\left(D_a + \tilde{\Omega}_a p\right)}.$$

Subsequently, (iv) proceeds by the facts in Theorem 8 that $W_2\left(\delta(x_{a,n-1}, v_{a,n-1}), \mu_a^{(n-1)}\right) \leq \frac{2}{\sqrt{n-1}}\tilde{D}_a$ (posterior concentration

results in Theorem 6) and $W_2\left(\mu_a^{(n-1)}, \delta(x_*, v_*)\right) \leq \frac{1}{\sqrt{n-1}}\tilde{D}_a$. For (v), we define

$$\bar{D}_a = \sqrt{\frac{2}{m_a}\left(D_a + 2\Omega_a \log 1/\delta_1 + \tilde{\Omega}_a p\right)},$$

which can be easily validated that $\bar{D}_a \geq \max\left\{\sqrt{\frac{2}{m_a}\left(D_a + 2\Omega_a \log 1/\delta_1\right)}, \sqrt{\frac{2}{m_a}\left(D_a + \tilde{\Omega}_a p\right)}\right\}$. Then we follow a similar procedure of proving Theorem 8 to arrive at the fact in (vi).

For $T3$, we can easily bound it with the result of posterior concentration inequality in Theorem 8:

$$T3 = W_2\left(\mu_a^{(n)}, \delta(x_*, v_*)\right) \leq \frac{\tilde{D}_a}{\sqrt{n}}. \tag{53}$$

We now move our attention back to upper bound $W_2\left(\tilde{\mu}_a^{(n)}[\bar{\rho}_a], \delta\left(x_*, v_*\right)\right)$ and derive the following results:

$$\begin{aligned}
W_2\left(\tilde{\mu}_a^{(n)}[\bar{\rho}_a], \delta\left(x_*, v_*\right)\right) &\leq W_2\left(\tilde{\mu}_a^{(n)}[\bar{\rho}_a], \tilde{\mu}_{Ih^{(n)}}\right) + W_2\left(\tilde{\mu}_{ih^{(n)}}, \mu_a^{(n)}\right) + W_2\left(\mu_a^{(n)}, \delta\left(x_*, v_*\right)\right) \\
&\overset{(i)}{\leq} \sqrt{\frac{dp}{nL_a\bar{\rho}_a}} + \frac{2\bar{D}_a}{\sqrt{n}} + \frac{\tilde{D}_a}{\sqrt{n}} \\
&\overset{(ii)}{\leq} \sqrt{\frac{dp}{nL_a\bar{\rho}_a}} + \frac{3\bar{D}_a}{\sqrt{n}} \\
&\leq \sqrt{\frac{36}{m_a n}\left(d + \log B_a + 2\Omega_a \log 1/\delta_1 + \left(\Omega_a + \frac{d}{18L_a\bar{\rho}_a}\right)p\right)}
\end{aligned}$$

where (i) proceeds by the results of (51)-(53), (ii) holds because $\tilde{D}_a \leq \bar{D}_a$. Following the same idea in (22)-(24) and with the selection of $p = 2\log 1/\delta_2$ and $\tilde{\Omega}_a = \Omega_a + \frac{d}{18L_a\bar{\rho}_a}$, the derived upper bound finally indicates a marginal posterior concentration inequality:

$$\mathbb{P}_{x_{a,n}\sim\tilde{\mu}_a^{(n)}[\bar{\rho}_a]}\left(\|x_{a,n} - x_*\|_2 > 6\sqrt{\frac{e}{m_a n}\left(D_a + 2\Omega_a \log 1/\delta_1 + 2\tilde{\Omega}_a \log 1/\delta_2\right)}\middle|\mathcal{Z}_{n-1}\right) < \delta_2.$$

□

Using a similar reasoning process, we can deduce that the next Theorem stays true.

Theorem 11. *Under conditions where the log-likelihood is consistent with Assumptions 4, the prior aligns with Assumption 3, and considering that arm a has been chosen $\mathcal{L}_a(n)$ times till the nth round of the Thompson sampling. By setting the batch size for stochastic gradient estimates as $k = \tilde{O}\left(\kappa_a^2\right)$, the step size and number of steps in Algorithm 3 as $h^{(n)} = \hat{O}\left(\frac{1}{\sqrt{d}}\right)$ and $I = \hat{O}\left(\sqrt{d}\right)$ accordingly, we can derive the following posterior concentration bound:*

$$\mathbb{P}_{x_{a,n}\sim\hat{\mu}_a^{(n)}[\bar{\rho}_a]}\left(\|x_{a,n} - x_*\|_2 > \sqrt{\frac{36e}{m_a n}\left(D_a + 2\Omega_a \log 1/\delta_1 + 2\hat{\Omega}_a \log 1/\delta_2\right)}\middle|\mathcal{Z}_{n-1}\right) < \delta_2,$$

where $D_a = 2\log B_a + 8d$, $\Omega_a = 256 + \frac{16dL_a^2}{m_a v_a}$, $\hat{\Omega}_a = 256 + \frac{16dL_a^2}{m_a v_a} + \frac{m_a d}{18L_a\rho_a}$, and the definition of $\mathcal{Z}_n$ can refer to Theorem 10 and Remark 2:

$$\mathcal{Z}_n = \left\{\left\|x_{a,n} - x_*\right\|_2 \leq 3\sqrt{\frac{2e}{m_a n}\left(D_a + 2\Omega_a \log 1/\delta_1\right)}\right\}.$$

Proof. Similar to the proof in Theorem 10, we first upper bound the distance between $\hat{\mu}_a^{(n)}[\bar{\rho}_a]$ and $\delta\left(x_*\right)$:

$$\begin{aligned}
W_2\left(\tilde{\mu}_a^{(n)}[\bar{\rho}_a], \delta\left(x_*, v_*\right)\right) &\leq W_2\left(\hat{\mu}_a^{(n)}[\bar{\rho}_a], \hat{\mu}_{Ih^{(n)}}\right) + W_2\left(\hat{\mu}_{Ih^{(n)}}, \mu_a^{(n)}\right) + W_2\left(\mu_a^{(n)}, \delta\left(x_*, v_*\right)\right) \\
&\overset{(i)}{\leq} \sqrt{\frac{dp}{nL_a\bar{\rho}_a}} + W_2\left(\hat{\mu}_{Ih^{(n)}}, \mu_a^{(n)}\right) + W_2\left(\mu_a^{(n)}, \delta\left(x_*, v_*\right)\right) \\
&\overset{(ii)}{\leq} \sqrt{\frac{dp}{nL_a\bar{\rho}_a}} + \underbrace{W_2\left(\hat{\mu}_{Ih^{(n)}}, \mu_a^{(n)}\right)}_{T1} + \frac{\hat{D}_a}{\sqrt{n}},
\end{aligned} \tag{54}$$

where in (i), the distance between $\hat{\mu}_a^{(n)}[\bar{\rho}_a]$ and $\hat{\mu}_{Ih^{(n)}}$ parallels the re-sampling mechanism at the last step of Langevin Monte Carlo, characterized by $\hat{x}_a - \hat{x}_{Ih^{(n)}} \sim \mathcal{N}\left(0, \frac{1}{nL_a\bar{\rho}_a}\mathbf{I}_{d\times d}\right)$. Similar to the upper bound $W_2\left(\tilde{\mu}_a^{(n)}[\bar{\rho}_a], \tilde{\mu}_{Ih^{(n)}}\right) \leq \sqrt{\frac{dp}{nL_a\bar{\rho}_a}}$ by applying Stirling's approximation for the Gamma function in Theorem 9, we establish the same upper bound for $W_2\left(\hat{\mu}_a^{(n)}[\bar{\rho}_a], \hat{\mu}_{Ih^{(n)}}\right)$. Additionally, (ii) derives its result by incorporating the insights from Theorem 9 and an extension of (53). Following the idea in (52), we can subsequently bound term $T1$ as follows:

$$
\begin{aligned}
T1 &= W_2\left(\hat{\mu}_{Ih^{(n)}}, \mu_a^{(n)}\right) \\
&\leq 4e^{Ih^{(n)}/2\kappa_a} W_2\left(\delta(x_{a,n-1}, v_{a,n-1}), \mu_a^{(n)}\right) + \frac{4}{1-e^{-h^{(n)}/2\kappa_a}}\left(5\left(h^{(n)}\right)^2\sqrt{\frac{d}{m_a}} + 4uh^{(n)}L_a\sqrt{\frac{nd}{k_a\nu_a}}\right) \\
&\leq 4e^{Ih^{(n)}/2\kappa_a}\left[W_2\left(\delta(x_{a,n-1}, v_{a,n-1}), \delta(x_*, v_*)\right) + W_2\left(\delta(x_*, v_*), \mu_a^{(n)}\right)\right] \\
&\quad + \frac{4}{1-e^{-h^{(n)}/2\kappa_a}}\left(5\left(h^{(n)}\right)^2\sqrt{\frac{d}{m_a}} + 4uh^{(n)}L_a\sqrt{\frac{nd}{k_a\nu_a}}\right) \\
&\leq 4e^{Ih^{(n)}/2\kappa_a}\left[W_2\left(\delta(x_{a,n-1}, v_{a,n-1}), \mu_a^{(n-1)}\right) + W_2\left(\mu_a^{(n-1)}, \delta(x_*, v_*)\right) + W_2\left(\delta(x_*, v_*), \mu_a^{(n)}\right)\right] \\
&\quad + \frac{4}{1-e^{-h^{(n)}/2\kappa_a}}\left(5\left(h^{(n)}\right)^2\sqrt{\frac{d}{m_a}} + 4uh^{(n)}L_a\sqrt{\frac{nd}{k_a\nu_a}}\right) \\
&\overset{(i)}{\leq} 4e^{Ih^{(n)}/2\kappa_a}\left[\frac{3}{\sqrt{n-1}}\hat{D}_a + \frac{\hat{D}'_a}{\sqrt{n}}\right] + \frac{4}{1-e^{-h^{(n)}/2\kappa_a}}\left(5\left(h^{(n)}\right)^2\sqrt{\frac{d}{m_a}} + 4uh^{(n)}L_a\sqrt{\frac{nd}{k_a\nu_a}}\right) \\
&\leq 4e^{Ih^{(n)}/2\kappa_a}\left[\frac{3}{\sqrt{n-1}}\hat{D}_a + \frac{\hat{D}'_a}{\sqrt{n}}\right] + 80\kappa_a h^{(n)}\sqrt{\frac{d}{m_a}} + 64\kappa_a uL_a\sqrt{5\frac{nd}{k\nu_a}} \\
&\leq \frac{2\bar{D}_a}{\sqrt{n}}.
\end{aligned}
\tag{55}
$$

For (i) we define the following concentration rate for arm a at round n as:

$$
\begin{aligned}
W_2\left(\delta(x_*, v_*), \mu_a^{(n)}\right) \leq \frac{\hat{D}'_a}{\sqrt{n}} &= \sqrt{\frac{2}{m_a n}\left(D_a + \hat{\Omega}_a p\right)} \\
W_2\left(\mu_a^{(n-1)}, \delta(x_*, v_*)\right) \leq \frac{1}{\sqrt{n-1}}\hat{D}_a &\leq \frac{2}{\sqrt{n}}\hat{D}_a.
\end{aligned}
$$

We further define $\bar{D}_a = \sqrt{\frac{2}{m_a}\left(D_a + 2\Omega_a \log 1/\delta_1 + \hat{\Omega}_a p\right)}$ and derive the upper bound for $T1$. Applying this finding to (54), we derive the inequality:

$$
W_2\left(\hat{\mu}_a^{(n)}[\bar{\rho}_a], \delta\left(x_*, v_*\right)\right) \leq \sqrt{\frac{36}{m_a n}\left(d + \log B_a + 2\Omega_a \log 1/\delta_1 + \hat{\Omega}_a p\right)}.
$$

Following the idea in (22)-(24) and with the selection of $p = 2\log 1/\delta_2$, we derive the approximate marginal posterior concentration inequality:

$$
\mathbb{P}_{x_{a,n}\sim\hat{\mu}_a^{(n)}[\bar{\rho}_a]}\left(\|x_{a,n} - x_*\|_2 > 6\sqrt{\frac{e}{m_a n}\left(D_a + 2\Omega_a \log 1/\delta_1 + 2\hat{\Omega}_a \log 1/\delta_2\right)}\middle|\mathcal{Z}_{n-1}\right) < \delta_2.
$$

□

E.4 Regret Analysis of Approximate Thompson Sampling

Employing either the full gradient or stochastic gradient estimates, we can obtain a consistent posterior concentration rate as highlighted in Theorems 10-11. Building on this, we extend our study to the regrets considering the approximate Thompson sampling algorithm. While the proof strategy for Theorem 12 is similar to that of Theorem 7, the regret analysis becomes more intricate due to our choice of sampling strategy. This complexity emerges because the generated samples lose their conditional independence when the filtration starts with the previous sample. To address this, it requires an additional related

lemma and we start with the definition of the following event:

$$\mathcal{Z}_a(N) = \bigcap_{n=1}^{N-1} \mathcal{Z}_{a,n},$$

where the definition of $\mathcal{Z}_{a,n}$ utilized in Theorems 10-11 adheres to the formulation initially given in Lemma 10:

$$\mathcal{Z}_{a,n} = \left\{ \left\| x_{a,n} - x_* \right\|_2 < 6 \sqrt{\frac{e}{m_a n} \left(D_a + 2\Omega_a \log 1/\delta_1 \right)} \right\},$$

where $D_a = 2 \log B_a + 8d$ and $\Omega_a = \frac{16 d L_a^2}{m_a \gamma_a} + 256$.

Lemma 8. *Given that the likelihood, rewards, and priors adhere to Assumptions 3-5, and taking into account the settings from Theorems 8-9 for step size, number of steps, and stochastic gradient estimates, the regret from Thompson sampling using approximate posterior sampling can be expressed as:*

$$\mathbb{E}[\Re(N)] \leq \sum_{a>1} \Delta_a \mathbb{E}\left[\mathcal{L}_a(N) \middle| \mathcal{Z}_a(N) \cap \mathcal{Z}_1(N) \right] + 2\Delta_a$$

Proof. The derivation follows the same idea presented in Mazumdar et al. (2020), and we revisit this proof herein for comprehensive exposition. The general idea is because of Lemma 3, the probability of each event in $\mathcal{Z}_a(N)^c$ and $\mathcal{Z}_1(N)^c$ is less than Δ_a. Then we can construct "$\mathbb{P}(\mathcal{Z}_a(N)^c \cup \mathcal{Z}_1(N)^c)$" and thus we can upper bound this term through a function of Δ_a.

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}_a(N)\right] &= \mathbb{E}\left[\mathcal{L}_a(N) \middle| \mathcal{Z}_a(N) \cap \mathcal{Z}_1(N) \right] \mathrm{P}\left(\mathcal{Z}_a(N) \cap \mathcal{Z}_1(N)\right) \\
&\quad + \mathbb{E}\left[\mathcal{L}_a(N) \middle| \mathcal{Z}_a(N)^c \cup \mathcal{Z}_1(N)^c \right] \mathrm{P}\left(\mathcal{Z}_a(N)^c \cup \mathcal{Z}_1(N)^c\right) \\
&\overset{(i)}{\leq} \mathbb{E}\left[\mathcal{L}_a(N) \middle| \mathcal{Z}_a(N) \cap \mathcal{Z}_1(N) \right] + \mathbb{E}\left[\mathcal{L}_a(N) \middle| (\mathcal{Z}_a(N)^c \cup \mathcal{Z}_1(N)^c) \right] \mathrm{P}\left(\mathcal{Z}_a(N)^c \cup \mathcal{Z}_1(N)^c\right) \\
&\overset{(ii)}{\leq} \mathbb{E}\left[\mathcal{L}_a(N) \middle| \mathcal{Z}_a(N) \cap \mathcal{Z}_1(N) \right] + 2N\delta_2 \mathbb{E}\left[\mathcal{L}_a(N) \middle| (\mathcal{Z}_a(N)^c \cup \mathcal{Z}_1(N)^c) \right] \\
&\overset{(iii)}{\leq} \mathbb{E}\left[\mathcal{L}_a(N) \middle| \mathcal{Z}_a(N) \cap \mathcal{Z}_1(N) \right] + 2,
\end{aligned}$$

where (i) use the fact that $1 - \mathrm{P}\left(\mathcal{Z}_a(N) \cap \mathcal{Z}_1(N)\right) \leq 1$, and (ii) holds because from Lemma 10 we know that for $a \in \mathcal{A}$ we have:

$$\mathbb{P}\left(\mathcal{Z}_a(N)^c \cup \mathcal{Z}_1(N)^c\right) \leq \mathbb{P}(\mathcal{Z}_1(N)^c) + \mathbb{P}\left(\mathcal{Z}_a(N)^c\right) = 2N\delta_2.$$

For part (iii), we invoke a straightforward observation $\mathcal{L}_a(N) \leq N$ and the choice of $\delta_2 = \frac{1}{N^2}$. By summing $\mathbb{E}\left[\Delta_a \mathcal{L}_a(N)\right]$ over the range of the second arm ($a = 2$) to the last arm ($a = |\mathcal{A}|$), the proof for Lemma 8 is thereby concluded. □

Employing a similar method delineated in Lemma 3, we can extend anti-concentration guarantees to the approximate posterior distributions.

Lemma 9. *Suppose the likelihood, rewards, and priors satisfy Assumptions 2-5. We follow the sampling schemes given in Theorems 8-9, where the step size is $h = \tilde{O}\left(\frac{1}{\sqrt{d}}\right)$, number of steps $I = \tilde{O}\left(\sqrt{d}\right)$, and batch size $k = \tilde{O}\left(\kappa_a^2\right)$. With the choice of letting $\bar{\rho}_1 = \frac{m_1}{8 L_1 \Omega_1}$, it follows that the underdamped Langevin Monte Carlo yields approximate distributions that are capped by the following upper bound guarantee for each round $n \in [\![1, N]\!]$:*

$$\mathbb{E}\left[\frac{1}{\hat{\mathcal{G}}_{1,n}} \right] \leq 33 \sqrt{B_1}.$$

Remark 3. *Following the definition of $\mathcal{G}$ given in Section C.3, here we denote $\hat{\mathcal{G}} = \mathrm{P}\left(\hat{\mathcal{E}}_a(n) | \mathcal{F}_{n-1}\right)$, where the event $\hat{\mathcal{E}}_a(n) = \left\{ \hat{\mathcal{R}}_{a,n} \geq \bar{\mathcal{R}}_1 - \epsilon \right\}$ indicates the estimated reward of arm a at round n (the reward is estimated by employing approximate Thompson sampling with underdamped Langevin Monte Carlo) exceeds the expected reward of optimal arm by at least a positive constant ϵ.*

Proof. For any rounds n encompassed within $[\![1, N]\!]$, our initial analysis focuses on how samples generated from Algorithm 3 exhibit the requisite anti-concentration characteristics. Specifically, with $x_{1,n}$ following a normal distribution given by $x_{1,n} \sim \mathcal{N}(x_{1,Ih}, \frac{1}{nL_1\bar{\rho}_1}\mathbf{I}_{d\times d})$, we can infer a corresponding lower bound for $\hat{\mathcal{G}}_{1,n}$:

$$\begin{aligned}\hat{\mathcal{G}}_{1,n} &= \mathsf{P}\left(\langle \alpha_1, x_{1,n} - x_{1,Ih}\rangle \geq \langle \alpha_1, x_* - x_{1,Ih}\rangle - \epsilon\right) \\ &\geq \mathsf{P}\left(Z \geq \underbrace{\omega_1 \|x_{1,Ih} - x_*\|_2}_{:=t}\right),\end{aligned}$$

where $Z \sim \mathcal{N}(0, \frac{\omega_1^2}{nL_1\bar{\rho}_1}\mathbf{I}_{d\times d})$ by construction. Building upon the analysis presented in Lemma 3, and considering $\sigma^2 = \frac{\omega_1^2}{nL_1\bar{\rho}_1}$, we arrive that the cumulative density function $\hat{\mathcal{G}}_{1,n}$ is bounded by:

$$\hat{\mathcal{G}}_{1,n} \geq \sqrt{\frac{1}{2\pi}} \begin{cases} \dfrac{\sigma t}{t^2+\sigma^2} e^{-\frac{t^2}{2\sigma^2}}, & t > \dfrac{\omega_1}{\sqrt{nL_1\bar{\rho}_1}}, \\ 0.34, & t \leq \dfrac{\omega_1}{\sqrt{nL_1\bar{\rho}_1}}. \end{cases}$$

We subsequently derive an upper bound for $\frac{1}{\hat{\mathcal{G}}_{1,n}}$:

$$\frac{1}{\hat{\mathcal{G}}_{1,n}} \leq \sqrt{2\pi} \begin{cases} \left(\dfrac{t}{\sigma} + 1\right) e^{\frac{t^2}{2\sigma^2}}, & t > \dfrac{\omega_1}{\sqrt{nL_1\bar{\rho}_1}}, \\ 3, & t \leq \dfrac{\omega_1}{\sqrt{nL_1\bar{\rho}_1}}. \end{cases}$$

By setting the parameters in $\omega_1\|x_{1,Ih} - x_*\|_2$ as $\sigma = \frac{\omega_1}{\sqrt{nL_1\bar{\rho}_1}}$, and subsequently taking the expectation relative to rewards $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_n$, we can infer that:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\hat{\mathcal{G}}_{1,n}}\right] &\leq 3\sqrt{2\pi} + \sqrt{2\pi}\mathbb{E}\left[\left(\sqrt{nL_1\rho_1}\|x_{1,Ih} - x_*\|_2 + 1\right) e^{nL_1\rho_1\|x_{1,Ih}-x_*\|_2^2}\right] \\ &= 3\sqrt{2\pi} + \sqrt{2\pi}\mathbb{E}\left[e^{nL_1\rho_1\|x_{1,Ih}-x_*\|_2^2}\right] \\ &\quad + \sqrt{2\pi nL_1\rho_1}\sqrt{\mathbb{E}\left[\|x_{1,Ih} - x_*\|_2^2\right]}\mathbb{E}\left[e^{nL_1\rho_1\|x_{1,Ih}-x_*\|_2^2}\right] \\ &\overset{(i)}{\leq} 3\sqrt{2\pi} + \sqrt{2\pi}\left(9\sqrt{\frac{L_1\rho_1}{2m_1}(D_1 + 2\Omega_1)} + \frac{3}{2}\right)\left(e^{\frac{4L_1D_1}{m_1}\rho_1} + 1\right) \\ &\overset{(ii)}{=} 3\sqrt{2\pi} + \sqrt{2\pi}\left(9\sqrt{\frac{L_1\rho_1}{m_1}\left(4d + \log B_1 + \frac{16L_1^2 d}{m_1\nu_1} + 256\right)} + \frac{3}{2}\right)\left(e^{\frac{4L_1}{m_1}(8d_1 + 2\log B_1)\rho_1} + 1\right) \\ &\overset{(iii)}{\leq} 3\sqrt{2\pi} + \frac{3}{2}\sqrt{2\pi}\left(\frac{3}{2}\sqrt{\frac{1}{136}\log B_1 + \frac{75}{34}} + 1\right)\left(e^{\frac{1}{68} + \frac{1}{272}\log B_1} + 1\right) \\ &\overset{(iv)}{\leq} 3\sqrt{2\pi} + 25\sqrt{B_1} \\ &\overset{(v)}{\leq} 33\sqrt{B_1}\end{aligned}$$

where (i) is from Lemma 19, (ii) is by having $D_1 = 8d + 2\log B_1$ and $\Omega_1 = \frac{16L_1^2 d}{m_1\nu_1} + 256$. (iii) proceeds by the same selection of $\bar{\rho}_1 \leq \frac{m_1}{8L_1\Omega_1}$ in Lemma 19 and because without loss of generality we know that $\frac{m_1}{L_1} \leq 1$, $\frac{m_1^2\nu_1}{L_1^3} \leq 1$ and $\frac{1}{d} \leq 1$ hold, we thus have $\bar{\rho}_1 \leq \frac{m_1}{8L_1\Omega_1} \leq \frac{m_1}{2176L_1 d} \leq \frac{m_1}{2176L_1}$. (iv) holds by $\frac{3\sqrt{2\pi}}{2}\left(\frac{3}{2}\sqrt{\frac{1}{136}\log x + \frac{75}{34}} + 1\right)\left(e^{\frac{1}{68}}x^{\frac{1}{272}} + 1\right) < 25\sqrt{x}$ for $x \geq 1$, and finally (v) is because of $3\sqrt{2\pi} + 25\sqrt{x} < 33\sqrt{x}$ for $x \geq 1$. □

With the insights gained from Lemma 4, we can extend it to finalize the proof of Lemma 10 and then Theorem 12.

Lemma 10. *Suppose the likelihood, rewards, and priors satisfy Assumptions 2-5 and given that the samples are drawn according to the sampling methods delineated in Theorems 8 and 9* $\left(h = \tilde{O}\left(\frac{1}{\sqrt{d}}\right),\ I = \tilde{O}\left(\sqrt{d}\right),\ \text{and } k = \tilde{O}\left(\kappa_a^2\right)\right)$*. With the choice of* $\bar{\rho}_a = \frac{m_a}{8L_a\Omega_a}$*, the following inequalities for approximate probability concentrations holds true:*

$$\sum_{s=0}^{N-1}\mathbb{E}\left[\frac{1}{\hat{\mathcal{G}}_{1,s}}-1\middle|\mathcal{Z}_1(N)\right]\leq 33\sqrt{B_1}\left|\frac{144e\omega_1^2}{m_1\Delta_a^2}(8d+2\log B_1+4\Omega_1\log N+3d\Omega_1)\right|+1 \tag{56}$$

$$\sum_{s=0}^{N-1}\mathbb{E}\left[\mathbb{I}\left(\hat{\mathcal{G}}_{a,s}>\frac{1}{N}\right)\middle|\mathcal{Z}_a(N)\right]\leq\frac{144e\omega_a^2}{m_a\Delta_a^2}\left(8d+2\log B_a+7d\Omega_a\log N\right), \tag{57}$$

where the approximate posterior $\hat{\mu}_a$ *yields a distribution,* $\hat{\mathcal{G}}_{a,s}$*, after gathering* s *samples. Additionally, for arm* $a\in\mathcal{A}$*, and the relation* $\Omega_a=\frac{16dL_a^2}{m_a\gamma_a}+256$ *holds.*

Proof. According to the definition of $\mathcal{G}_{a,s}$ and the derivation in (35), we can derive a similar lower bound for $\hat{\mathcal{G}}_{a,s}$ as follows:

$$\begin{aligned}\hat{\mathcal{G}}_{1,s}&=\mathbb{P}(\mathcal{R}_{1,s}>\bar{\mathcal{R}}_1-\epsilon|\mathcal{F}_{n-1})\\&=1-\mathbb{P}(\mathcal{R}_{1,s}-\bar{\mathcal{R}}_1<-\epsilon|\mathcal{F}_{n-1})\\&\geq 1-\mathbb{P}(|\mathcal{R}_{1,s}-\bar{\mathcal{R}}_1|>\epsilon|\mathcal{F}_{n-1})\\&\geq 1-\mathbb{P}_{x\sim\hat{\mu}_1^{(s)}}\left(\|x-x_*\|>\frac{\epsilon}{\omega_1}\right).\end{aligned} \tag{58}$$

Then the result given in (56) can be derived following the steps given below:

$$\begin{aligned}\sum_{s=0}^{N-1}\mathbb{E}\left[\frac{1}{\hat{\mathcal{G}}_{1,s}}-1\middle|\mathcal{Z}_1(N)\right]&=\sum_{s=0}^{\ell-1}\mathbb{E}\left[\frac{1}{\hat{\mathcal{G}}_{1,s}}-1\middle|\mathcal{Z}_1(N)\right]+\sum_{s=\ell}^{N-1}\mathbb{E}\left[\frac{1}{\hat{\mathcal{G}}_{1,s}}-1\middle|\mathcal{Z}_1(N)\right]\\&\overset{(i)}{\leq}\sum_{s=0}^{\ell-1}\mathbb{E}\left[\frac{1}{\hat{\mathcal{G}}_{1,s}}-1\middle|\mathcal{Z}_1(N)\right]+\sum_{s=\ell}^{N-1}\left[\frac{1}{1-\mathbb{P}_{x\sim\hat{\mu}_1^{(s)}}(\|x-x_*\|\geq\epsilon/\omega_1)}-1\right]\\&\leq\sum_{s=0}^{\ell-1}\mathbb{E}\left[\frac{1}{\hat{\mathcal{G}}_{1,s}}-1\middle|\mathcal{Z}_1(N)\right]+\int_{s=\ell}^{\infty}\left[\frac{1}{1-\mathbb{P}_{x\sim\hat{\mu}_1^{(s)}}(\|x-x_*\|\geq\epsilon/\omega_1)}-1\right]ds\\&\overset{(ii)}{\leq}\sum_{s=0}^{\ell-1}\mathbb{E}\left[\frac{1}{\hat{\mathcal{G}}_{1,s}}-1\middle|\mathcal{Z}_1(N)\right]+\frac{144e\omega_1^2}{m_1\Delta_a^2}3d\Omega_1\log 2+1\\&\overset{(iii)}{\leq}33\sqrt{B_1}\left|\frac{144e\omega_1^2}{m_1\Delta_a^2}(8d+2\log B_1+4\Omega_1\log N+3d\Omega_1)\right|+1,\end{aligned}$$

where (i) is from (58), and (ii) is because from Lemmas 10 and 11, with the selection of $\delta_1=\frac{1}{Te^2}$ and $\bar{\rho}_a=\frac{m_a}{8L_a\Omega_a}$ we have:

$$\mathbb{P}_{x\sim\tilde{\mu}_a^{(n)}[\bar{\rho}_a]}\left(\|x-x_*\|_2>6\sqrt{\frac{e}{m_1n}\left(D_1+4\Omega_1\log N+3d\Omega_1\log 1/\delta_2\right)}\middle|\mathcal{Z}_{n-1}\right)<\delta_2. \tag{59}$$

Equivalently, we can have:

$$\mathbb{P}_{x\sim\tilde{\mu}_1^{(s)}[\rho_1]}\left(\|x-x_*\|>\frac{\epsilon}{\omega_1}\right)\leq\exp\left(-\frac{1}{3d\Omega_1}\left(\frac{m_1n\epsilon^2}{36e\omega_1^2}-8d-2\log B_1-4\Omega_1\log N\right)\right), \tag{60}$$

which holds true when $n\geq\frac{36e\omega_1^2}{\epsilon^2m_1}\left(8d+2\log B_1+4\Omega_1\log N\right)$. If we select the following ℓ:

$$\ell=\left|\frac{144e\omega_1^2}{m_1\Delta_a^2}(8d+2\log B_1+4\Omega_1\log N+3d\Omega_1\log 2)\right|$$

and let $\epsilon=\Delta_a/2$, following the same idea in (36) we can get (ii). Then (iii) proceeds by the selection of ℓ and Lemma 9.

To show that (57) holds, we have:

$$\begin{aligned}\sum_{n=1}^{N}\mathbb{E}\left[\mathbb{I}\left(\hat{\mathcal{G}}_{a,s}>\frac{1}{N}\right)\middle|\mathcal{Z}_1(N)\right] &= \sum_{n=1}^{N}\mathbb{E}\left[\mathbb{I}\left(\mathbb{P}\left(\mathcal{R}_{a,s}-\bar{\mathcal{R}}_a>\Delta_a-\epsilon|\mathcal{F}_{n-1}\right)>\frac{1}{N}\right)\middle|\mathcal{Z}_a(N)\right]\\ &\overset{(i)}{\leq}\sum_{n=1}^{N}\mathbb{E}\left[\mathbb{I}\left(\mathbb{P}_{x\sim\mu_a^{(s)}[\rho_a]}\left(\|x-x_*\|>\frac{\Delta_a}{2\omega_a}\right)>\frac{1}{N}\right)\middle|\mathcal{Z}_a(N)\right]\\ &\overset{(ii)}{\leq}\frac{144e\omega_a^2}{m_a\Delta_a^2}\left(8d+2\log B_a+10d\Omega_a\log N\right),\end{aligned}$$

where (i) is from our previous defined $\epsilon=\Delta_a/2$, and (ii) proceeds by inducing (60) with the choice of

$$n\geq\left\lceil\frac{8e\omega_1^2}{m\Delta_a^2}(8d+2\log B_a+4\Omega_a\log N+3d\Omega_a\log N)\right\rceil,$$

which completes the proof. □

Combining the conclusions drawn from Lemma 9 and Lemma 10, we present our concluding theorem:

Theorem 12 (Regret of approximate Thompson sampling with (stochastic gradient) underdamped Langevin Monte Carlo)**.** *When the likelihood, rewards, and priors satisfy Assumptions 2-5, scaling parameter $\bar{\rho}_a=\frac{m_a}{8L_a\Omega_a}$, and the sampling schemes of Thompson sampling follow the setting in Theorems 8-9, namely, the step size is $h=\tilde{O}\left(\frac{1}{\sqrt{d}}\right)$, number of steps $I=\tilde{O}\left(\sqrt{d}\right)$, and batch size $k=\tilde{O}\left(\kappa_a^2\right)$. We then have that the total expected regrets after $N>0$ rounds of Thompson sampling with the (stochastic gradient) underdamped Langevin Monte Carlo method satisfy:*

$$\begin{aligned}\mathbb{E}[\Re(N)]\leq\sum_{a>1}\frac{\hat{C}_a}{\Delta_a}\left(\log B_a+d+d^2\kappa_a^2\log N\right)\\ +\frac{\hat{C}_1\sqrt{B_1}}{\Delta_a}\left(\log B_1+d^2\kappa_1^2+d\kappa_1\log N\right)+4\Delta_a.\end{aligned}$$

where $\hat{C}_a,\hat{C}_1>0$ are universal constants that are independent of problem-dependent parameters and $\kappa_a=L_a/m_a$.

Proof. Our approach to proving Theorem 12 begins by taking the expected number of arm a pulls over N rounds, and multiplying it with the distance in expected rewards from the optimal and the chosen arm a. It is important to highlight that the expected number of arm a selections over N rounds is conditional upon $\mathcal{Z}_1(N)\cap\mathcal{Z}_a(N)$. This condition implies the concentration of the log-likelihood and the inclination of the approximate samples to be predominantly located in the high-probability areas of the posteriors. This upper bound is further detailed using the findings from Lemma 10. The detailed proof is given below:

$$\begin{aligned}\mathbb{E}[\Re(N)]&\overset{(i)}{\leq}\sum_{a>1}\Delta_a\mathbb{E}\left[\mathcal{L}_a(N)\middle|\mathcal{Z}_a(N)\cap\mathcal{Z}_1(N)\right]+2\Delta_a\\ &\overset{(ii)}{\leq}\sum_{a>1}\left\{\Delta_a\sum_{s=0}^{N-1}\mathbb{E}\left[\frac{1}{\mathcal{G}_{1,s}}-1\middle|\mathcal{Z}_1(N)\right]+\Delta_a\sum_{s=0}^{N-1}\mathbb{E}\left[\mathbb{I}\left(1-\mathcal{G}_{a,s}>\frac{1}{N}\right)\middle|\mathcal{Z}_a(N)\right]+3\Delta_a\right\}\\ &\overset{(iii)}{\leq}\sum_{a>1}33\sqrt{B_1}\left\lceil\frac{144e\omega_1^2}{m_1\Delta_a}\left(8d+2\log B_1+4\Omega_1\log N+3d\Omega_1\right)\right\rceil\\ &\quad+\sum_{a>1}\frac{144e\omega_a^2}{m_a\Delta_a}\left(8d+2\log B_a+7d\Omega_a\log N\right)+4\sum_{a>1}\Delta_a\\ &\leq\sum_{a>1}\left[\frac{\hat{C}_a}{\Delta_a}\left(\log B_a+d+d^2\kappa_a^2\log N\right)+\frac{\hat{C}_1}{\Delta_a}\sqrt{B_1}\left(\log B_1+d^2\kappa_1^2+d\kappa_1\log N\right)+4\Delta_a\right].\end{aligned}$$

where (i) is drawn from Lemma 8, (ii) is attributed to Lemmas 1 and 2, and (iii) can derived by applying Lemma 10. By selecting constants $\hat{C}_1$ and $\hat{C}_a$ to upper bound some problem-independent constants, we arrive at the final regret bound.

□

E.5 Supporting Proofs for Approximate Thompson Sampling

The following section provides additional theoretical support for the analysis of Langevin Monte Carlo and approximate Thompson sampling.

Continuous Time Analysis

Theorem 13 (Convergence for the Continuous-Time Process)**.** *Suppose $(x_t, v_t) \sim \mu_t$ follows Langevin dynamics described by (42), μ_* the posterior distribution described by (x_*, v_*). Then the following inequality holds:*

$$W_2(\mu_t, \mu_*) \leq 2e^{-t/2\kappa} W_2(\mu_0, \mu_*). \tag{61}$$

Proof. We start by defining η_t to be the distribution of $(x_t, x_t + v_t)$, η_* the posterior distribution of $(x_t', x_t' + v_t')$, and let $\zeta_t \in \Gamma(\eta_t, \eta_*)$ be the optimal coupling between η_t and η_* such that $\mathbb{E}_{\zeta_0}\left[\left\|x_0 - x_0'\right\|_2^2 + \|x_0 - x_0' + v_0 - v_0'\|_2^2\right] = W_2^2(\eta_0, \eta_*)$. Then we have:

$$\begin{aligned}
W_2^2(\mu_t, \mu_*) &\overset{(i)}{\leq} W_2^2(\eta_t, \eta_*) \\
&\overset{(ii)}{\leq} \mathbb{E}_{\zeta_0}\left[\mathbb{E}_{\zeta_t}\left[\left\|x_t - x_t'\right\|_2^2 + \left\|x_t - x_t' + v_t - v_t'\right\|_2^2 | x_0, x_0', v_0, v_0'\right]\right] \\
&\overset{(iii)}{\leq} \mathbb{E}_{\zeta_0}\left[e^{-t/\kappa}\left(\left\|x_0 - x_0'\right\|_2^2 + \left\|x_0 - x_0' + v_0 - v_0'\right\|_2^2\right)\right] \\
&\overset{(iv)}{=} e^{-t/\kappa} W_2^2(\eta_0, \eta_*) \\
&\overset{(v)}{=} 4e^{-t/\kappa} W_2^2(\mu_0, \mu_*),
\end{aligned}$$

where (i) and (v) proceed by Lemma 7. (ii) is a direct consequence of the definition of 2-Wasserstein distance, which is based on optimal coupling, and the tower property of expectation. Subsequently, (iii) is derived from Lemma 11, while (iv) is from the specific choice of ζ_0 as the optimal coupling. Taking square roots from both sides completes the proof. □

Lemma 11. *Let diffusion $(x_t, x_t + v_t) \sim \eta_t$ and posterior distribution $(x_t', x_t' + v_t') \sim \eta_*$ are both in $\mathbb{R}^{2d}$ and follow the dynamics dictated by SDEs (5). Suppose $f(\cdot)$ fulfills Assumption 4, then the following inequality holds:*

$$\mathbb{E}\left[\left\|x_t - x_t'\right\|_2^2 + \left\|x_t - x_t' + v_t - v_t'\right\|_2^2\right] \leq e^{-t/\kappa}\mathbb{E}\left[\left(\left\|x_0 - x_0'\right\|_2^2 + \left\|x_0 - x_0' + v_0 - v_0'\right\|_2^2\right)\right] \tag{62}$$

Proof. This proof substantially relies on the results of Lemma 5, and we consequently exclude some repetitive parts. The proof starts with the derivation of an upper bound for the time-dependent derivative of $\left\|x_t - x_t'\right\|_2^2 + \left\|(x_t + v_t) - (x_t' + v_t')\right\|_2^2$ w.r.t. time t:

$$\begin{aligned}
&\frac{d}{dt}\left(\left\|x_t - x_t'\right\|_2^2 + \left\|(x_t + v_t) - (x_t' + v_t')\right\|_2^2\right) \\
&\overset{(i)}{=} -2\left\langle x_t - x_t', v_t - v_t'\right\rangle + 2\left\langle (x_t - x_t') + (v_t - v_t'), (\gamma - 1)(v_t - v_t') + u\left(\nabla f(x_t) - \nabla f(x_t')\right)\right\rangle \\
&\overset{(ii)}{\leq} -\frac{m}{L}\left(\left\|x_t - x_t'\right\|_2^2 + \left\|(x_t + v_t) - (x_t' + v_t')\right\|_2^2\right)
\end{aligned}$$

where (i) proceed by taking derivatives of x_t, x_t', v_t, and v_t' w.r.t. time, (ii) applies the conclusion draw from Lemma 5. The convergence rate in Lemma 11 can be derived as a direct consequence of Grönwall's inequality, as detailed in Corollary 3 from Dragomir (2003). □

Discretization Analysis

We turn our attention to a single step within the underdamped Langevin diffusion in a discrete version, characterized by the following stochastic differential equations:

$$\begin{aligned}
d\tilde{v}_t &= -\gamma\tilde{v}_t dt - u\nabla f(\tilde{x}_0)\, dt + \left(\sqrt{2\gamma u}\right) dB_t \\
d\tilde{x}_t &= \tilde{v}_t dt,
\end{aligned} \tag{63}$$

where $t \in [0, h)$ for some small h, and h represents a single step in the underdamped Langevin Monte Carlo. The solution $(\tilde{x}_t, \tilde{v}_t)$ of (63) can be expressed as follows

$$\begin{aligned} \tilde{v}_t &= \tilde{v}_0 e^{-\gamma t} - u\left(\int_0^t e^{-\gamma(t-s)}\tilde{\nabla} f(\tilde{x}_0) ds\right) + \sqrt{2\gamma u}\int_0^t e^{-\gamma(t-s)} dB_s \\ \tilde{x}_t &= \tilde{x}_0 + \int_0^t v_s ds. \end{aligned} \tag{64}$$

To get the first term in (64), one may multiply the velocity term in (63) by $e^{-\gamma t}$ on both sides and then compute the definite integral from 0 to t. Then the second term in (64) can be derived by taking the definite integral of the position term. We skip the detailed derivation here.

Our focus in the following lemma is to quantify the discretization error between the continuous and discrete processes, as represented by (5) and (63) respectively, both initiating from identical distributions. Specifically, our bound pertains to $W_2(\mu_t, \tilde{\mu}_t)$ over the interval $t \in [ih, (i+1)h)$. While μ_t is a continuous-time diffusion detailed in Theorem 13, $\tilde{\mu}_t$ is a measure corresponding to (63).Our findings here illuminate the convergence rate defined in Lemma 12.

Throughout this analysis, we work under the assumption that the kinetic energy, representing the second moment of velocity in a continuous-time process, adheres to the boundary defined as follows: $\mathbb{E}_{\mu_t}\|v_t\|_2^2 \le 26(d/m + \mathcal{D}^2)$. A comprehensive explanation of this bound can be found in Lemma 15. Following this, we provide a detailed exploration of our discretization analysis findings.

Lemma 12 (Discretization Analysis of underdamped Langevin Monte Carlo with full gradient)**.** *Consider (x_t, v_t) distributed as μ_t and $(\tilde{x}_t, \tilde{v}_t)$ as $\tilde{\mu}_t$, representing the measures from SDEs (5) and (63), respectively. Given an initial distribution μ_0 and a step size $h < 1$, we select $u = 1/L$ and $\gamma = 2$. Consequently, the difference between continuous-time and discrete-time processes within:*

$$W_2(\mu_t, \tilde{\mu}_t) \le 2h^2\sqrt{\frac{26}{5}\frac{d}{m}}. \tag{65}$$

Proof. Considering the following synchronous coupling $\zeta_t \in \Gamma(\mu_t, \tilde{\mu}_t)$, where μ_t and $\tilde{\mu}_t$ are both anchored to the initial distribution μ_0 and are governed by the same Brownian motion, B_t. Our focus initially settles on bounding the velocity deviation, and referencing the descriptions of v_t and $\tilde{v}_t$ in (63), we deduce that:

$$\begin{aligned} \mathbb{E}_{\zeta_t}\left[\|v_t - \tilde{v}_t\|_2^2\right] &\overset{(i)}{=} \mathbb{E}\left[\left\|u\int_0^s e^{-2(t-s)}\left(\nabla f(x_s) - \nabla f(x_0)\right) ds\right\|_2^2\right] \\ &= u^2\mathbb{E}\left[\left\|\int_0^s e^{-2(t-s)}\left(\nabla f(x_s) - \nabla f(x_0) ds\right)\right\|_2^2\right] \\ &\overset{(ii)}{\le} tu^2\int_0^s \mathbb{E}\left[\left\|e^{-2(t-s)}\left(\nabla f(x_s) - \nabla f(x_0)\right)\right\|_2^2\right] ds \\ &\overset{(iii)}{\le} tu^2\int_0^s \mathbb{E}\left[\|(\nabla f(x_s) - \nabla f(x_0))\|_2^2\right] ds \\ &\overset{(iv)}{\le} tu^2L^2\int_0^s \mathbb{E}\left[\|x_s - x_0\|_2^2\right] ds \\ &\overset{(v)}{=} tu^2L^2\int_0^s \mathbb{E}\left[\left\|\int_0^r v_w dw\right\|_2^2\right] ds \\ &\overset{(vi)}{\le} tu^2L^2\int_0^s r\left(\int_0^r \mathbb{E}\left[\|v_w\|_2^2\right] dw\right) ds \\ &\overset{(vii)}{\le} 26tu^2L^2\left(\frac{d}{m} + \mathcal{D}^2\right)\int_0^s r\left(\int_0^r dw\right) ds \\ &= \frac{26t^4u^2L^2}{3}\left(\frac{d}{m} + \mathcal{D}^2\right), \end{aligned} \tag{66}$$

where (i) follows from (63) and $v_0 = \tilde{v}_0$, (ii) proceeds by an application of Jensen's inequality, (iii) follows as $\left\|e^{-4(t-s)}\right\| \le 1$, (iv) is by Lipschitz smooth property of $f(x)$, (v) arises from the explicit definition of x_r. The inferences in (vi) and (vii) stem from Jensen's inequality and the uniform kinetic energy bound, where an in-depth exploration of this bound is offered in Lemma 15.

Having delineated the bounds for the velocity aspect, our subsequent task is to determine the discretization error's bound within the position variable.

$$\begin{aligned}\mathbb{E}_{\zeta_t}\left[\|x_t-\hat{x}_t\|_2^2\right] &\overset{(i)}{=} \mathbb{E}\left[\left\|\int_0^t (v_s-\hat{v}_s)\,ds\right\|_2^2\right]\\ &\overset{(ii)}{\leq} t\int_0^t \mathbb{E}\left[\|v_s-\tilde{v}_s\|_2^2\right]ds\\ &\overset{(iii)}{\leq} t\int_0^t \left(\frac{d}{m}+\mathcal{D}^2\right)\frac{26u^2L^2s^4}{3}ds\\ &= \left(\frac{d}{m}+\mathcal{D}^2\right)\frac{26u^2L^2}{15}t^6,\end{aligned}$$

where (i) relies on the coupling associated with the initial distribution μ_0, (ii) is by the following Jensen's inequality:

$$\left\|\int_0^t v_s ds\right\|_2^2 = \left\|\frac{1}{N}\int_0^t t\cdot v_s ds\right\|_2^2 \leq t\int_0^t \|v_s\|_2^2\,ds,$$

(iii) uses the derived bound in (66). With the selection of $t=h$ and $u=\frac{1}{L}$, we identify an upper bound for the 2-Wasserstein distance between μ_h and $\tilde{\mu}_h$:

$$\begin{aligned}W_2(\mu_h,\tilde{\mu}_h) &\leq \sqrt{26\left(\frac{h^4}{3}+\frac{h^6}{15}\right)\left(\frac{d}{m}+\mathcal{D}^2\right)}\\ &\overset{(i)}{\leq} h^2\sqrt{\frac{52}{5}\left(\frac{d}{m}+\mathcal{D}^2\right)}\\ &\overset{(ii)}{\leq} 2h^2\sqrt{\frac{26}{5}\frac{d}{m}},\end{aligned}$$

where (i) holds because without loss of generality h is assumed to be less than 1, and (ii) proceeds by $\mathcal{D}^2 = d/m$ from Lemma 17, which completes the proof.

□

With the convergence rate of the continuous-time SDE captured in (61) and a discretization error bound delineated in (63), our next step is to formulate the following conclusion relative to underdamped Langevin Monte Carlo with full gradient.

Theorem 14 (Convergence of the discrete underdamped Langevin Monte Carlo with full gradient)**.** *Suppose $(x_t, v_t) \sim \mu_t$ and $(x_t, x_t + v_t) \sim \eta_t$ are corresponding to the measures of distribution described by SDEs (5), $(\tilde{x}_t, \tilde{v}_t) \sim \tilde{\mu}_t$ and $(\tilde{x}_t, \tilde{x}_t + \tilde{v}_t) \sim \tilde{\eta}_t$ by SDEs (63) respectively. Suppose μ_* and η_* the posterior distribution described by (x_*, v_*) and $(x_*, x_* + v_*)$. By defining $u = \frac{1}{L}$ and setting $\gamma = 2$, we can establish an upper bound on the 2-Wasserstein distance differentiating the continuous-time from the discrete-time process as follows:*

$$W_2(\mu_{ih},\mu_*) \leq 4e^{-ih/2\kappa}W_2(\mu_0,\mu_*) + \frac{20h^2}{1-e^{-h/2\kappa}}\sqrt{\frac{d}{m}}, \tag{67}$$

where we denote $t \in [ih, (i+1)h)$, $i \in [\![1, I]\!]$.

Proof. From Theorem 13 we have that for any $i \in [\![1, I]\!]$:

$$W_2(\eta_{ih},\eta_*) \leq e^{-h/2\kappa}W_2(\eta_{(i-1)h},\eta_*),$$

Through the application of the discretization error bound provided in Lemma 12 and Lemma 7, we can obtain:

$$W_2(\eta_{ih},\tilde{\eta}_{ih}) \leq 2W_2(\mu_{ih},\tilde{\mu}_{ih}) \leq 4h^2\sqrt{\frac{26}{5}\frac{d}{m}}. \tag{68}$$

Invoking the triangle inequality yields the following results:

$$\begin{aligned} W_2(\tilde{\eta}_{ih}, \eta_*) &\leq W_2(\eta_{ih}, \eta_*) + W_2(\eta_{ih}, \tilde{\eta}_{ih}) \\ &\leq e^{-h/2\kappa} W_2(\eta_{(i-1)h}, \eta_*) + 4h^2 \sqrt{\frac{26}{5}\frac{d}{m}}. \end{aligned} \tag{69}$$

With the above results, we derive the upper bound for the distance between $\tilde{\mu}_{ih}$ and μ_* shown in (67):

$$\begin{aligned} W_2(\tilde{\mu}_{ih}, \mu_*) &\overset{(i)}{\leq} 2W_2(\tilde{\eta}_{ih}, \eta_*) \\ &\overset{(ii)}{\leq} 2e^{-ih/2\kappa} W_2(\eta_0, \eta_*) + \left(1 + e^{-h/2\kappa} + \cdots + e^{-(i-1)h/2\kappa}\right) 8h^2 \sqrt{\frac{26}{5}\frac{d}{m}} \\ &\overset{(iii)}{\leq} 4e^{-ih/2\kappa} W_2(\mu_0, \mu_*) + \frac{8h^2}{1 - e^{-h/2\kappa}} \sqrt{\frac{26}{5}\frac{d}{m}} \\ &\leq 4e^{-ih/2\kappa} W_2(\mu_0, \mu_*) + \frac{20h^2}{1 - e^{-h/2\kappa}} \sqrt{\frac{d}{m}}. \end{aligned}$$

In this derivation, step (i) is from Lemma 7, (ii) holds by applying (69) iteratively i times, (iii) is concluded by summing up the geometric series and by another application of Lemma 7. Hence, we deduce the desired final bound. □

Stochasticity Analysis

Taking into account stochastic gradients, we commence by outlining the underdamped Langevin Monte Carlo (Algorithm 4 considering stochastic noise in gradient estimate).

Suppose $(\hat{x}_t, \hat{v}_t)$ is the pair of position and velocity at time t considering the stochastic gradient underdamped Langevin diffusion:

$$\begin{aligned} d\hat{v}_t &= -\gamma \hat{v}_t dt - u\nabla \hat{f}(\hat{x}_0)\, dt + \left(\sqrt{2\gamma u}\right) dB_t \\ d\hat{x}_t &= \hat{v}_t dt, \end{aligned} \tag{70}$$

Then the solution $(\hat{x}_t, \hat{v}_t)$ of (70) is

$$\begin{aligned} \hat{v}_t &= \hat{v}_0 e^{-\gamma t} - u\left(\int_0^t e^{-\gamma(t-s)} \nabla \hat{f}(\hat{x}_0) ds\right) + \sqrt{2\gamma u} \int_0^t e^{-\gamma(t-s)} dB_s \\ \hat{x}_t &= \hat{x}_0 + \int_0^t v_s ds. \end{aligned} \tag{71}$$

It is noteworthy that by setting $t = 0$ and adopting the same initial conditions as in (63), the resultant differential equations are almost identical to (63), except the stochastic gradient estimation. Therefore, we opt to omit detailed proof in this context.

Drawing on the notational framework and Assumptions outlined in 4, our next step is to delineate the discretization discrepancies. These discrepancies arise when comparing discrete process (63) with full gradient and another discrete process in (70) with stochastic gradient, where both are from the identical initial distribution.

Lemma 13 (Convergence of underdamped Langevin Monte Carlo with stochastic gradient)**.** *Consider* $(\tilde{x}_t, \tilde{v}_t) \sim \tilde{\mu}_t$ *and* $(\hat{x}_t, \hat{v}_t) \sim \hat{\mu}_t$ *be measures corresponding to the distributions as defined in* (63) *and* (70)*. Then for any h satisfying* $0 < h < 1$*, the subsequent relationship is valid:*

$$W_2(\tilde{\mu}_h, \hat{\mu}_h) \leq uhL \sqrt{\frac{15nd}{k\nu}},$$

where k is the number of data points for stochastic gradient estimates, n the total number of data points we have for gradient estimate $(n \geq k)$*, and L is a Lipschitz constant for the stochastic estimate of gradients, which follows Assumption 5.*

Proof. Drawing from the dynamics outlined in (63) and (70), and taking into account the definition of $\nabla \hat{f}(x)$, we can deduce the following relationship:

$$\begin{aligned} v_h &= \hat{v}_h + u\left(\int_0^h e^{-\gamma(s-h)}ds\right)\xi \\ x_h &= \hat{x}_h + u\left(\int_0^h \left(\int_0^r e^{-\gamma(s-r)}ds\right)dr\right)\xi \end{aligned} \tag{72}$$

where ξ is a zero-mean random variance to represent the difference between full gradient and stochastic gradient estimate, and it is independent of the Brownian motion. Let ζ be the optimal coupling between $\tilde{\mu}_h$ and $\hat{\mu}_h$. Then we can upper bound the distance between $\hat{\mu}_h$ and $\tilde{\mu}_h$ as follows:

$$\begin{aligned} W_2^2(\tilde{\mu}_h, \hat{\mu}_h) &= \mathbb{E}_\zeta\left[\|(\hat{x}, \hat{v}) - (\tilde{x}, \tilde{v})\|_2^2\right] \\ &= \mathbb{E}_\zeta\left[\left\| \begin{matrix} u\left(\int_0^h \left(\int_0^r e^{-\gamma(s-r)}ds\right)dr\right)\xi \\ u\left(\int_0^h \left(\int_0^r e^{-\gamma(s-r)}ds\right)dr + \int_0^h e^{-\gamma(s-h)}ds\right)\xi \end{matrix} \right\|_2^2\right] \\ &\overset{(i)}{\leq} 3u^2\left[\left(\int_0^h \left(\int_0^r e^{-\gamma(s-r)}ds\right)dr\right)^2 + \left(\int_0^h e^{-\gamma(s-h)}ds\right)^2\right]\mathbb{E}\|\xi\|_2^2 \\ &\overset{(ii)}{\leq} 3u^2\left(\frac{h^4}{4} + h^2\right)\mathbb{E}\|\xi\|_2^2 \\ &\overset{(iii)}{<} \frac{15}{4}u^2h^2\mathbb{E}\|\xi\|_2^2 \\ &\overset{(iv)}{<} \frac{15nd}{k\nu}u^2h^2L^2, \end{aligned}$$

where (i) is by independence and unbiasedness of ξ and the inequality $a^2 + (a+b)^2 \leq 3(a^2+b^2)$, (ii) uses the upper bound $e^{-\gamma(s-r)} \leq 1$ and $e^{-\gamma(s-t)} \leq 1$, (iii) proceeds by the assumption, made without loss of generality, that $h < 1$, and finally (iv) is derived from a result in Lemma 18. Taking square roots from both sides completes the proof.

□

By integrating Theorem 13, Lemma 12, and Lemma 13, we establish a convergence bound between $\hat{\mu}_{nh}$ and μ_* as follows:

Theorem 15 (Convergence of the underdamped Langevin Monte Carlo with stochastic gradient)**.** *Let $(x_t, v_t) \sim \mu_t$ and $(x_t, x_t + v_t) \sim \eta_t$ are corresponding to the measures of distribution described by SDEs (5), $(\tilde{x}_t, \tilde{v}_t) \sim \hat{\mu}_t$ and $(\tilde{x}_t, \tilde{x}_t + \tilde{v}_t) \sim \tilde{\eta}_t$ by SDEs (70) respectively. Let μ_* and η_* the posterior distribution described by (x_*, v_*) and $(x_*, x_* + v_*)$. Following the previous choice of $u = 1/L$ and $\gamma = 2$, and let $t \in [ih, (i+1)h)$, an upper bound for the 2-Wasserstein distance between the continuous-time and discrete-time processes, with stochastic gradient, is given by:*

$$W_2(\hat{\mu}_{ih}, \mu_*) \leq 4e^{-ih/2\kappa}W_2(\mu_0, \mu_*) + \frac{4}{1 - e^{-h/2\kappa}}\left(5h^2\sqrt{\frac{d}{m}} + 4uhL\sqrt{\frac{nd}{k\nu}}\right). \tag{73}$$

Proof. We follow the proof framework in Theorem 14 to derive this bound. Suppose $i \in [\![1, I]\!]$, then by the triangle inequality we will have:

$$\begin{aligned} W_2(\hat{\eta}_{ih}, \eta_*) &\leq W_2(\eta_{ih}, \eta_*) + W_2(\eta_{ih}, \tilde{\eta}_{ih}) + W_2\left(\tilde{\eta}_{ih}, \hat{\eta}_{ih}\right) \\ &\overset{(i)}{\leq} W_2(\eta_{ih}, \eta_*) + 2W_2(\mu_{ih}, \tilde{\mu}_{ih}) + 2W_2\left(\tilde{\mu}_{ih}, \hat{\mu}_{ih}\right) \\ &\overset{(ii)}{\leq} e^{-h/2\kappa}W_2(\eta_{(n-1)h}, \eta_*) + 4h^2\sqrt{\frac{26}{5}\frac{d}{m}} + 2uhL\sqrt{\frac{15nd}{k\nu}}, \end{aligned} \tag{74}$$

where (i) follows the same idea as (68), (ii) proceeds by Lemma 13 and (69). Finally, a repeated application of (74) i times

yields the following results:

$$W_2(\hat{\eta}_{ih}, \eta_*) \leq e^{-ih/2\kappa} W_2(\eta_0, \eta_*) + \left(1 + e^{-h/2\kappa} + \cdots + e^{-(i-1)h/2\kappa}\right)\left(4h^2\sqrt{\frac{26}{5}\frac{d}{m}} + 2uhL\sqrt{\frac{15nd}{k\nu}}\right)$$
$$< 2e^{-ih/2\kappa} W_2(\mu_0, \mu_*) + \frac{2}{1 - e^{-h/2\kappa}}\left(5h^2\sqrt{\frac{d}{m}} + 4uhL\sqrt{\frac{nd}{k\nu}}\right).$$

By leveraging Lemma 7 again we derive the results as depicted in (73).

□

Concentration on the Posterior Distribution

Lemma 14 (Lemma 10 in Mazumdar et al. (2020))**.** *When the marginal posterior μ'_* ($x \sim \mu'_*$) is characterized by m-strong log-concavity, the following result holds for $x_U = argmax_x \mu'_*(x)$:*

$$W_2\left(\mu'_*, \delta\left(x_U\right)\right) \leq 6\sqrt{\frac{d}{m}}.$$

Proof. We start by segmenting $W_2\left(\mu'_*, \delta\left(x_U\right)\right)$ into two distinct terms and subsequently establish an upper bound for both.

$$\begin{aligned} W_2\left(\mu'_*, \delta\left(x_U\right)\right) &\leq W_2\left(\mu'_*, \delta\left(\mathbb{E}_{x\sim\mu'_*}[x]\right)\right) + \left\|x_U - \mathbb{E}_{x\sim\mu'_*}[x]\right\|_2 \\ &\overset{(i)}{\leq} W_2\left(\mu'_*, \delta\left(\mathbb{E}_{x\sim\mu'_*}[x]\right)\right) + \sqrt{\frac{3}{m}} \\ &\overset{(ii)}{\leq} \left[\int \left\|x - \mathbb{E}_{x\sim\mu'_*}[x]\right\|_2^2 \, \mathrm{d}\mu'_*(x)\right]^{1/2} + \sqrt{\frac{3}{m}} \\ &\overset{(iii)}{\leq} e^{1/e}\sqrt{\frac{8d}{m}} + \sqrt{\frac{3}{m}} \\ &\overset{(iv)}{<} 6\sqrt{\frac{d}{m}}, \end{aligned}$$

where (i) builds upon an application of Theorem 7 in Basu and DasGupta (1997), whereby the selection of $\alpha = 1$, the relationship between the expectation and mode in unimodal distributions can be easily derived. In (ii), the coupling relationship between μ'_* and $\delta\left(\mathbb{E}_{x\sim\mu'_*}[x]\right)$ is from the adoption of the product measure. In (iii), we use the Herbst argument as detailed in (Ledoux, 2006), yielding bounds on the second moment. Notably, an m-strongly log-concave distribution reveals a log Sobolev constant equivalent to m. Thus, by application of the Herbst argument, $x \sim \mu'_*$ can be described as a sub-Gaussian random vector characterized by the parameter $\sigma^2 = \frac{1}{2m}$ with $\|\mathrm{u}\|_2 = 1$:

$$\int e^{\lambda\langle \mathrm{u}, x - \mathrm{E}_{x\sim\mu'_*}[x]\rangle} d\mu'_* \leq e^{\lambda^2/4m}.$$

Therefore, x is $2\sqrt{\frac{d}{m}}$ norm-sub-Gaussian and the following inequality holds:

$$\mathbb{E}_{x\sim\mu'_*}\left[\left\|x - \mathbb{E}_{x\sim\mu'_*}[x]\right\|_2^2\right] \leq \frac{8d}{m}e^{2/e}.$$

Given that $d \geq 1$ and $e^{1/e}\sqrt{8} + \sqrt{3} < 6$, the conclusion specified in (iv) naturally follows and thus finalize the proof.

□

Remark 4. *It is important to highlight that Lemma 14 provides an upper bound for the 2-Wasserstein distance between the marginal distribution μ'_* w.r.t. x and $\delta(x_U)$. After this, based on the relationship $\mu_* \propto \exp\left(-f(x) - \|v\|_2^2/2u\right)$, we can readily establish that the 2-Wasserstein distance between the marginal posterior distribution relative to v and $\delta(v_U)$ has an upper bound of $6\sqrt{du}$. By adopting $u = \frac{1}{L}$ and, without loss of generality, knowing that $m \leq L$, it follows that the 2-Wasserstein distance between the joint posterior distribution of (x, v) and $\delta(x_U, v_U)$ is upper-bounded by $6\sqrt{\frac{d}{m}}$.*

Controlling the Kinetic Energy

This subsection presents a definitive bound for the kinetic energy, which helps in managing the discretization error at every step.

Lemma 15 (Kinetic Energy Bound)**.** *For $(x_t, v_t) \sim \mu_t$ governed by the dynamics in (42), and given the delta measure $\delta(x_0, 0)$ when $t = 0$. With the established distance criterion $\|x_0 - x_*\|_2^2 \le \mathcal{D}^2$ and defining $u = 1/L$, then for $i = 1, \dots I$ and $t \in [ih, (i+1)h)$, we have the bound:*

$$\mathbb{E}_{v_t \sim \mu_t}\left[\|v\|_2^2\right] \le 26(d/m + \mathcal{D}^2).$$

Proof. We first denote μ_* the posterior described by (x, v). We further define $(x, x+v) \sim \eta_*$, $(x_t, x_t + v_t) \sim \eta_t$, and define $\zeta_t \in \Gamma(\eta_t, \eta_*)$ be the optimal coupling between η_t and η_*. Then we can upper bound the expected kinetic energy as:

$$\begin{aligned}
\mathbb{E}_{\mu_t}\|v_t\|_2^2 &= \mathbb{E}_{\zeta_t}\left[\|v_t - v + v\|_2^2\right] \\
&\overset{(i)}{\le} 2\mathbb{E}_{\mu_*}\|v\|_2^2 + 2\mathbb{E}_{\zeta_t}\left[\|v_t - v\|_2^2\right] \\
&\overset{(ii)}{\le} 2\mathbb{E}_{\mu_*}\|v\|_2^2 + 4\mathbb{E}_{\zeta_t}\left[\|(x_t + v_t) - (x+v)\|_2^2 + \|x_t - x\|_2^2\right] \\
&\overset{(iii)}{=} 2\underbrace{\mathbb{E}_{\mu_*}\|v\|_2^2}_{T1} + 4\underbrace{W_2^2(\eta_t, \eta_*)}_{T2},
\end{aligned} \tag{75}$$

where (i) and (ii) are applications of Young's inequality, and (iii) holds by optimality of ζ_t. Given that $\mu_* \propto \exp\left(-f(x) - \frac{I}{2}\|v\|_2^2\right)$, it follows that $T1$ conforms to the subsequent result:

$$T1 = \mathbb{E}_{\mu_*}\|v\|_2^2 = \frac{d}{L}.$$

Next, we proceed to bound term $T2$:

$$\begin{aligned}
T2 &= W_2^2(\eta_t, \eta_*) \\
&\overset{(i)}{\le} W_2^2(\eta_0, \eta_*) \\
&\overset{(ii)}{\le} 2\mathbb{E}_{\eta_*}\left[\|v\|_2^2\right] + 2\mathbb{E}_{x_0 \sim \mu_0, x \sim \eta_*}\left[\|x_0 - x\|_2^2\right] \\
&\overset{(iii)}{\le} \frac{2d}{L} + 2\mathbb{E}_{x_0 \sim \mu_0, x \sim \eta_*}\left[\|x_0 - x\|_2^2\right] \\
&= \frac{2d}{L} + 2\mathbb{E}_{x \sim \eta_*}\left[\|x - x_0\|_2^2\right] \\
&\overset{(iv)}{\le} \frac{2d}{L} + 4\left(\mathbb{E}_{x \sim \eta_*}\left[\|x - x_*\|_2^2\right] + \|x_0 - x_*\|_2^2\right) \\
&\overset{(v)}{\le} \frac{2d}{L} + 4\left(\frac{d}{m} + \mathcal{D}^2\right),
\end{aligned}$$

where the derivation of (i) is based on the repeated application of Theorem 13. For a more detailed explanation, readers are referred to Step 4 in Lemma 12 (Cheng et al., 2018). Both (ii) and (iv) derive from Young's inequality. The combination of $T1$ and Theorem 13 informs (iii). Finally, (v) aligns with our predefined assumption $\|x_0 - x_*\|_2^2 \le \mathcal{D}^2$ and is supported by Lemma 17.

By incorporating the established upper bounds for $T1$ and $T2$ into (75), we successfully derive the upper bound as mentioned in Lemma 15:

$$\begin{aligned}
\mathbb{E}_{v \sim \mu_t}\|v\|_2^2 &\le 2\mathbb{E}_{\mu_*}\|v\|_2^2 + 4W_2^2(\eta_t, \eta_*) \\
&\le \frac{2d}{L} + \frac{8d}{L} + \frac{16d}{m} + 16\mathcal{D}^2 \\
&\le 26\left(\frac{d}{m} + \mathcal{D}^2\right).
\end{aligned} \tag{76}$$

□

Subsequently, we establish a bound for the distance between μ_0 and μ_*.

Lemma 16 (Initial Kinetic Energy Bound)**.** *Assume $(x_t, v_t) \sim \mu_t$ which adheres to the dynamic defined in (42), the posterior distribution $(x, v) \sim \mu_*$ that satisfies $\exp\left(-f(x) - \frac{L}{2}\|v\|_2^2\right)$, a delta distribution $\delta(x_0, 0)$ at $t = 0$, and the starting distance from the optimal being $\|x_0 - x_*\|_2^2 \leq \mathcal{D}^2$ with the choice of $u = 1/L$. Then with $t = 0$, the following result holds:*

$$W_2^2(\mu_0, \mu_*) \leq 3\left(\frac{d}{m} + \mathcal{D}^2\right). \tag{77}$$

Proof. As μ_t at time $t = 0$ can be represented by a delta distribution $\delta(x_0, v_0)$, the distance between measures μ_0 and μ_* can be upper bounded by:

$$\begin{aligned}
W_2^2(\mu_0, \mu_*) &= W_2^2(\delta(x_0, v_0), \mu_*) \\
&= \mathbb{E}_{(x,v)\sim\mu_*}\left[\|x - x_0\|_2^2 + \|v\|_2^2\right] \\
&= \mathbb{E}_{(x,v)\sim\mu_*}\left[\|x - x_* + x_* - x_0\|_2^2 + \|v\|_2^2\right] \\
&\overset{(i)}{\leq} 2\mathbb{E}_{x\sim\mu_*}\left[\|x - x_*\|_2^2\right] + 2\mathcal{D}^2 + \mathbb{E}_{v\sim\mu_*}\left[\|v\|_2^2\right] \\
&\overset{(ii)}{\leq} 2\mathbb{E}_{x\sim\mu_*}\left[\|x - x_*\|_2^2\right] + \frac{d}{L} + 2\mathcal{D}^2 \\
&\overset{(iii)}{\leq} 3\left(\frac{d}{m} + \mathcal{D}^2\right),
\end{aligned} \tag{78}$$

where we note that (i) is from Young's inequality and the definition of $\mathcal{D}^2$. (ii) is built on the observation that the marginal distribution of μ_* w.r.t. v correlates with $\exp\left(-\frac{L}{2}\|v\|_2^2\right)$, thereby deducing $\mathbb{E}_{v\sim\mu_*}\|v\|_2^2 = d/L$, and the conclusion in (iii) is grounded in Lemma 17, which completes the proof.

□

Lemma 17 (Bounding the Position, Theorem 1 in Durmus and Moulines (2016))**.** *For all $t > 0$ and $x \sim \mu_t$ follows (42) we have:*

$$\mathbb{E}_{\mu_t}\left[\|x - x_*\|_2^2\right] \leq \mathcal{D}^2 = \frac{d}{m}.$$

Bounding the Stochastic Gradient Estimates

Lemma 18 (Lemma 5 in Mazumdar et al. (2020))**.** *Consider $f : \mathbb{R}^d \to \mathbb{R}$ as the function representing the mean of $\log \mathbb{P}(\mathcal{R}_i|x)$ for $i = 1, 2, \ldots, n$, and we also consider $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ as the stochastic estimator of f, which is derived by randomly sampling $\log \mathbb{P}(\mathcal{R}_j|x)$ k times without replacement from $[\![1, \mathcal{L}(n)]\!]$. We further denote $\mathcal{S}$ as the set of k samples $\log \mathbb{P}(\mathcal{R}_j|x)$ we have, and it follows trivially that $k = |\mathcal{S}|$. Provided that Assumptions 2 and 5 hold for $\log \mathbb{P}(\mathcal{R}_i|x)$ for all i, an upper bound can be derived for the expected difference between the full gradient and its stochastic estimate computed over k samples:*

$$\mathbb{E}\left[\left\|\nabla\hat{f}(x) - \nabla f(x)\right\|_2^2 \middle| x\right] \leq \frac{4ndL^2}{k\nu}.$$

Proof. The proof outline is as follows: our first task involves the analysis of the Lipschitz smoothness of the gradient estimates corresponding to the likelihood function. Given this Lipschitz smoothness and the convexity assumption, a sub-Gaussian bound is achieved via the application of logarithmic Sobolev inequalities. We finally derive a more restrictive sub-Gaussian bound on the sum of the gradient of the likelihood, which serves to upper-bound the expected value of the gradient estimate. We start by formulating the expression as the gradient of $\log \mathrm{P}(\cdot|\cdot)$, denoted as:

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla\hat{f}(x) - \nabla f(x)\right\|_2^2\right] &= n^2\mathbb{E}\left[\left\|\frac{1}{k}\sum_{j=1}^{k}\nabla\log \mathrm{P}\left(\mathcal{R}_j|x\right) - \frac{1}{n}\sum_{i=1}^{n}\nabla\log \mathrm{P}\left(\mathcal{R}_i|x\right)\right\|_2^2\right] \\
&= \frac{n^2}{k^2}\mathbb{E}\left[\left\|\sum_{j=1}^{k}\left(\nabla\log \mathrm{P}\left(\mathcal{R}_j|x\right) - \frac{1}{n}\sum_{i=1}^{n}\nabla\log \mathrm{P}\left(\mathcal{R}_i|x\right)\right)\right\|_2^2\right].
\end{aligned} \tag{79}$$

It should be noted that one term in $\nabla \log \mathrm{P}(\mathcal{R}_j|x) - \frac{1}{n}\sum_{i=1}^n \nabla \log \mathrm{P}(\mathcal{R}_i|x)$ has been canceled out:

$$\nabla \log \mathrm{P}(\mathcal{R}_j|x) - \frac{1}{n}\sum_{i=1}^{n} \nabla \log \mathrm{P}(\mathcal{R}_i|x) = \nabla \log \mathrm{P}(\mathcal{R}_j|x) - \frac{1}{n}\sum_{i\neq j} \nabla \log \mathrm{P}(\mathcal{R}_i|x),$$

which indicates that the sum of $\nabla \log \mathrm{P}(\mathcal{R}_i|x)$ should have $n-1$ terms remaining. We now turn our attention to analyzing the term $\nabla \log \mathrm{P}(\mathcal{R}_j|x) - \nabla \log \mathrm{P}(\mathcal{R}_i|x)$. Given the joint Lipschitz smoothness and the strong convexity outlined in Assumption 5, the application of Theorem 3.16 from Wainwright (2019) is employed to establish that $\nabla \log \mathrm{P}(\mathcal{R}_j|x) - \nabla \log \mathrm{P}(\mathcal{R}_i|x)$ is $\frac{2L}{\sqrt{\nu}}$-sub-Gaussian, where L is the Lipschitz smooth constant and ν strongly convex constant.

As $\nabla \log \mathrm{P}(\mathcal{R}_j|x) - \nabla \log \mathrm{P}(\mathcal{R}_i|x)$ is an i.i.d random variable, we can use the Azuma-Hoeffding inequality for martingale difference sequences (Wainwright, 2019) to characterize the summation of such variables across $n-1$ as $\frac{2L}{n}\sqrt{\frac{n-1}{\nu}}$-sub-Gaussian. Building upon this, another application of the Azuma-Hoeffding inequality allows us to have that $\sum_{j=1}^k \left(\nabla \log \mathrm{P}\left(\mathcal{R}_j|x\right) - \frac{1}{n}\sum_{i=1}^n \nabla \log \mathrm{P}\left(\mathcal{R}_i|x\right)\right)$ is $\frac{2L}{n}\sqrt{\frac{k(n-1)}{\nu}}$-sub-Gaussian. Therefore, we can derive the following upper bound through the sub-Gaussian property:

$$\mathbb{E}\left[\left\|\sum_{j=1}^{k}\left(\nabla \log \mathrm{P}\left(\mathcal{R}_j|x\right) - \frac{1}{n}\sum_{i=1}^{n} \nabla \log \mathrm{P}\left(\mathcal{R}_i|x\right)\right)\right\|_2^2\right] \leq 4\frac{dk(n-1)L^2}{n^2\nu} \leq 4\frac{dkL^2}{n\nu}.$$

The proof is finalized by integrating the above results into (79):

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla f(x) - \nabla \hat{f}(x)\right\|_2^2\right] &= \frac{n^2}{k^2}\mathbb{E}\left[\left\|\sum_{j=1}^{k}\frac{1}{n}\sum_{i=1}^{n} \nabla \log \mathrm{P}(\mathcal{R}_i|x) - \nabla \log \mathrm{P}(\mathcal{R}_j|x)\right\|_2^2\right] \\ &\leq 4\frac{ndL^2}{k\nu}.\end{aligned}$$

□

Controlling the Moment Generating Function

Lemma 19. *Suppose for samples $x \in \mathbb{R}^d$ given by Algorithm 4 and fixed point $x_* \in \mathbb{R}^d$ satisfy the following bound:*

$$\mathbb{E}\left[\|x - x_*\|_2^2\right] \leq \frac{18}{mn}(D + \Omega p), \tag{80}$$

where $D = 8d + 2\log B$, $\Omega = \frac{16L^2 d}{m\nu} + 256$. Then for $\bar{\rho} \leq \frac{m}{8L\Omega}$ we will have the following bound for its moment generating function:

$$\mathbb{E}\left[e^{nL\bar{\rho}\|x-x_*\|_2^2}\right] \leq \frac{3}{2}\left(e^{\frac{4LD}{m}\bar{\rho}} + 1\right).$$

Proof. From Theorem 6, Lemma 10, and Lemma 11 we can get the following upper bound from (80):

$$\mathbb{E}\left[\|x - x_*\|_2^{2p}\right] \leq 3\left[\frac{2}{mn}(D + \Omega p)\right]^p.$$

Given the term $\|x - x_*\|_2^2$, we identify it as a sub-exponential random variable. This characterization is supported when its

moment generating function is unfolded via Taylor expansion $\mathbb{E}[e^{nL\bar{\rho}\|x-x_*\|_2^2}] = 1 + \sum_{p=1}^{\infty} \mathbb{E}\left[\frac{(nL\bar{\rho})^p \|x-x_*\|_2^{2p}}{p!}\right]$, which yields:

$$\begin{aligned}
\mathbb{E}\left[e^{nL\bar{\rho}\|x-x_*\|_2^2}\right] &= \sum_{p=0}^{\infty} \mathbb{E}\left[\frac{(nL\bar{\rho})^p}{p!} \|x - x_*\|_2^{2p}\right] \\
&\leq 1 + 3\sum_{p=1}^{\infty} \frac{(nL\bar{\rho})^p}{p!} \left(\frac{2}{mn}\right)^p (D + \Omega p)^p \\
&\overset{(i)}{\leq} 1 + \frac{3}{2}\sum_{p=1}^{\infty} \frac{1}{p!} \left(\frac{2LD}{m}\bar{\rho}\right)^p + \frac{3}{2}\sum_{p=1}^{\infty} \frac{1}{p!} \left(\frac{2L\Omega}{m}\bar{\rho}p\right)^p \\
&\overset{(ii)}{\leq} \frac{3}{2} e^{\frac{4LD}{m}\bar{\rho}} + \frac{3}{2}\sum_{p=1}^{\infty} \left(\frac{4L\Omega}{m}\bar{\rho}\right)^p \\
&\overset{(iii)}{\leq} \frac{3}{2}\left(e^{\frac{4LD}{m}\bar{\rho}} + 1\right)
\end{aligned}$$

where for (i), we use a specific case of Young's inequality for products: $(x+y)^p \leq 2^{p-1}(x^p + y^p)$, which holds for all $p \geq 1$ and $x, y \geq 0$, (ii) proceeds by Taylor expansion and $\frac{1}{p!} \leq \left(\frac{2}{p}\right)^p$, and (iii) holds by the selection of $\bar{\rho} \leq \frac{m}{8L\Omega}$.

□

F EXPERIMENTAL DETAILS

In our experimental setup designed to evaluate Thompson sampling's efficacy, we adopted a systematic hyper-parameter configuration as delineated in Table 3. The table encapsulates the consideration for hyper-parameters among simulations of Langevin Monte Carlo and general Thompson sampling algorithms. A crucial distinction between our approach and the baseline is the inclusion of the friction coefficient and noise amplitude, specifically tailored for the underdamped Langevin Monte Carlo. It is pertinent to highlight that when $\mathcal{L}_a(n) \leq k$, we denote the batch size directly as $\mathcal{L}_a(n)$.

Table 3: Hyper-parameters used in Thompson sampling algorithms.

Names	Ranges
Parameter dimension (d)	[10, 30, 100, 300, 1000]
Number of arms ($\lvert\mathcal{A}\rvert$)	10
Time horizon (N)	1000
Step size (h)	$[10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}]$
Number of steps (I)	$[10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}]$
Batch size (k)	[2, 5, 10, 20]
Friction coefficient (γ)	[0.1, 1.0, 2.0, 5.0, 10.0]
Noise amplitude (u)	1.0

F.1 Additional experiments

Here we further provide a practical exploration task aimed at investigating the highest-rated restaurants, with the use of real Google Maps ratings as our foundational dataset. The average ratings of the selected restaurants span from 3.8 to 4.8, which serves as a real-world benchmark for our statistical investigation. To simulate a realistic decision-making process, we consider an agent to visit restaurants sequentially. At each visit, the agent assigns a rating on an integer scale from 1 to 5, where these ratings are sampled from true customer ratings. This methodology allows us to dynamically evaluate each restaurant and identify the restaurant with the highest average ratings.

We apply approximate Thompson Sampling with ULMC and LMC for the restaurant recommendations. We record the expected regrets associated with these restaurants and the regrets are illustrated in Figure 2, where the red solid line represents regrets obtained from the underdamped algorithm, and the black dashed line is from the overdamped algorithm. The comparative analysis of these regrets further indicates the superiority of our proposed algorithm for real-world applications.

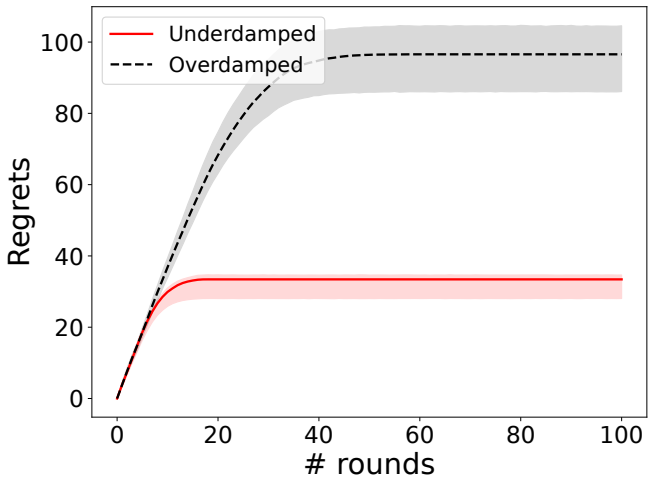


Figure 2: Restaurant example