Langevin Thompson Sampling with Logarithmic Communication: Bandits and Reinforcement Learning

Amin Karbasi [1 2 3] **Nikki Lijing Kuang** [* 4 5] **Yi-An Ma** [5] **Siddharth Mitra** [* 2]

Abstract

Thompson sampling (TS) is widely used in sequential decision making due to its ease of use and appealing empirical performance. However, many existing analytical and empirical results for TS rely on restrictive assumptions on reward distributions, such as belonging to conjugate families, which limits their applicability in realistic scenarios. Moreover, sequential decision making problems are often carried out in a batched manner, either due to the inherent nature of the problem or to serve the purpose of reducing communication and computation costs. In this work, we jointly study these problems in two popular settings, namely, stochastic multi-armed bandits (MABs) and infinite-horizon reinforcement learning (RL), where TS is used to learn the unknown reward distributions and transition dynamics, respectively. We propose batched *Langevin Thompson Sampling* algorithms that leverage MCMC methods to sample from approximate posteriors with only logarithmic communication costs in terms of batches. Our algorithms are computationally efficient and maintain the same order-optimal regret guarantees of $\mathcal{O}(\log T)$ for stochastic MABs, and $\mathcal{O}(\sqrt{T})$ for RL. We complement our theoretical findings with experimental results.

1. Introduction

Modern machine learning often needs to balance computation and communication budgets with statistical guarantees. Existing analyses of sequential decision making have been primarily focused on the statistical aspects of the problems (Tian et al., 2020), less is known about their computation and communication aspects. In particular, regret minimization in multi-armed bandits (MABs) and reinforcement learning (RL) (Jaksch et al., 2010; Wu et al., 2022; Jung et al., 2019) is often studied under the common assumptions that computation can be performed perfectly in time, and that communication always happens in real-time (Li et al., 2022; Jin et al., 2018; Haarnoja et al., 2018).

A question of particular importance is whether optimal decisions can still be made under reasonable computation and communication budgets. In this work, we study the exploration-exploitation problem with low computation and communication costs using Thompson Sampling (TS) (Thompson, 1933) (a.k.a. posterior sampling). To allow sampling from distributions that deviate from the standard restrictive assumptions, and to enable its deployment in settings where computing the exact posterior is challenging, we employ Markov Chain Monte Carlo (MCMC) methods.

TS operates by maintaining a posterior distribution over the unknown and is widely used owing to its strong empirical performance (Chapelle and Li, 2011). However, the theoretical understanding of TS in bandits typically relied on the restrictive conjugacy assumptions between priors and reward distributions. Recently, approximate TS methods for general posterior distributions start to be studied (Mazumdar et al., 2020; Xu et al., 2022), where MCMC algorithms are used in conjunction with TS to expand its applicability in fully-sequential settings. On the other hand, to the best of our knowledge, how to provably incorporate MCMC with posterior sampling in RL domains remains to be untackled.

Moreover, previous analyses of TS have been restricted to the fully-sequential settings, where feedback or reward is assumed to be immediately observable upon taking actions (i.e. before making the next decision). Nevertheless, it fails to account for the practical settings when delays take place in communication, or when feedback is only available in an aggregate or batched manner. Examples include clinical trials where feedback about the efficacy of medication is only available after a nontrivial amount of time, recommender systems where feedback from multiple users comes all at once and marketing campaigns (Schwartz

[*] Major contribution credited equally to Kuang and Mitra [1]Department of Electrical Engineering, Yale University, New Haven, USA [2]Department of Computer Science, Yale University, New Haven, USA [3]Department of Statistics and Data Science, Yale University, New Haven, USA [4]Department of Computer Science and Engineering, University of California San Diego, La Jolla, USA [5]Halıcıoğlu Data Science Institute, University of California San Diego, La Jolla, USA. Correspondence to: Nikki Lijing Kuang <l1kuang@ucsd.edu>, Siddharth Mitra <siddharth.mitra@yale.edu>, Amin Karbasi <amin.karbasi@yale.edu>, Yi-An Ma <yianma@ucsd.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

	TS-based Algorithms	Batching Scheme	MCMC Method	Regret	# of Batches
Stochastic MAB	(Karbasi et al., 2021)	Dynamic	-	$O(\log T)$	$O(\log T)$
	(Mazumdar et al., 2020)	-	SGLD, ULA	$O(\log T)$	$O(T)$
	This paper (Algorithm 2)	Dynamic	SGLD	$O(\log T)$	$O(\log T)$
TS for RL (PSRL)	(Osband et al., 2013)	-	-	$O(\sqrt{T})$	$O(T/H)$
	(Ouyang et al., 2017)	Dynamic	-	$O(\sqrt{T})$	$O(\sqrt{T})$
	(Theocharous et al., 2017b)	Static	-	$O(\sqrt{T})$	$O(\log T)$
	This paper (Algorithm 3)	Static	SGLD, MLD	$O(\sqrt{T})$	$O(\log T)$

Table 1: We compare our methods with existing TS-based algorithms in terms of batching schemes and MCMC methods adopted for approximation. Performance is measured by regret, while communication cost is quantified with the number of batches. Here, T is the time horizon, H is the fixed episode length in episodic MDP settings. Our methods achieve optimal performance, while reducing computation and communication costs due to batching, and are applicable in broader regimes.

et al., 2017). This issue has been studied in literature by considering static or dynamic batching schemes and by designing algorithms that acquire feedback in batches, where the learner typically receives reward information only at the end of the batch (Karbasi et al., 2021; Kalkanli and Ozgur, 2021; Vernade et al., 2020; Zhang et al., 2020). Nonetheless, the analysis of approximate TS in batched settings is unavailable for both bandits and RL.

In this paper, we tackle these challenges by incorporating TS with Langevin Monte Carlo (LMC) and batching schemes in stochastic MABs and infinite-horizon RL. Our algorithms are applicable to a broad class of distributions with only logarithmic rounds of communication between the learner and the environment, thus being robust to constraints on communication. We compare our results with other works in Table 1, and summarize our main contributions as follows:

- For stochastic MABs with time horizon T, we present *Langevin Thompson Sampling* (BLTS, Algorithm 2) along with Theorem 3, which achieves the optimal $O(\log T)$ regret with $O(\log T)$ batches[1]. The main technical contribution here is to show that when feedback is obtained in a batched manner where the posterior concentration is weaker (Theorem 1), the convergence guarantee of SGLD continues to hold.
- For large-scale infinite-horizon MDPs, we present *Langevin Posterior Sampling for RL* (LPSRL, Algorithm 3) along with Theorem 4 to show that SGLD with a static policy-switching[2] scheme achieves the optimal $O(\sqrt{T})$ Bayesian regret with $O(\log T)$ policy switches. For tabular MDPs, we show that LPSRL incorporated with the Mirrored Langevin Dynamics (MLD) achieves the optimal $O(\sqrt{T})$ Bayesian regret with $O(\log T)$ policy switches. The use of approximate sampling leads to an additive error where the true model and the sampled model are no longer identically distributed. This error can be properly handled with the convergence guarantees of LMC methods.
- Experiments are performed to demonstrate the effectiveness of our algorithms, which maintain the order-optimal regret with significantly lower communication costs compared to existing exact TS methods.

2. Problem Setting

In this section, we introduce the problem setting with relevant background information.

2.1. Stochastic Multi-armed Bandits

We consider the N-armed stochastic multi-armed bandit problem, where the set of arms is denoted by $\mathcal{A} = [N] = \{1, 2, \ldots, N\}$. Let T be the time horizon of the game. At $t = 1, 2, \ldots, T$, the learner chooses an arm $a_t \in \mathcal{A}$ and receives a real-valued reward r_{a_t} drawn from a fixed, unknown, parametric distribution corresponding to arm a_t. In the standard fully-sequential setup, the learner observes rewards immediately. Here, we consider the more general batched setting, where the learner observes the rewards for all timesteps within a batch at the end of it. We use B_k to denote the starting time of the k-th batch, $B(t)$ to represent the starting time of the batch that contains time t, and K as the total number of batches. The learner observes the set of rewards $\{r_{a_t}\}_{t=B_k}^{B_{k+1}-1}$ at the end of the k-th batch. Note that the batched setting reduces to the fully-sequential setting when the number of batches is T, each with size 1.

Suppose for each arm a, there exists a parametric reward distribution parameterized by $\theta_a \in \mathbb{R}^d$ such that the true reward distribution is given by $p_a(r) = p_a(r|\theta_a^*)$, where θ_a^* is an unknown parameter[3]. To ensure meaningful results,

[1] A T round game can be thought of as T many batches each of size 1.

[2] In MDP settings, the notion of a batch is more appropriately thought of as a policy-switch.

[3] Our results hold for the more general case of $\theta_a \in \mathbb{R}^{d_a}$, but for simplicity of exposition,

we impose the following assumptions on the reward distributions for all $a \in \mathcal{A}$:

- **Assumption 1:** $\log p_a(r|\theta_a)$ is L-smooth and m-strongly concave in θ_a.
- **Assumption 2:** $p_a(r|\theta_a^*)$ is ν strongly log-concave in r and $\nabla_\theta \log p_a(r|\theta_a^*)$ is L-Lipschitz in r.
- **Assumption 3:** The prior $\lambda_a(\theta_a)$ is concave with L-Lipschitz gradients for all θ_a.
- **Assumption 4:** Joint Lipschitz smoothness of (the bivariate) $\log p_a(r|\theta_a)$ in r and θ_a.

These properties include log-concavity and Lipschitz smoothness of the parametric families and prior distributions, which are standard assumptions in existing literature Mazumdar et al. (2020) and are satisfied by models like Gaussian bandits (Honda and Takemura, 2013). For the sake of brevity, We provide the mathematical statements of these assumptions in Appendix B.

Let μ_a denote the expected value of the true reward distribution for arm a. The goal of the learner is to minimize the expected regret, which is defined as follows:

$$R(T) := \mathbb{E}\left[\sum_{t=1}^{T} \mu^* - \mu_{a_t}\right] = \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}\left[k_a(T)\right], \quad (1)$$

where $\mu^* = \max_{a \in \mathcal{A}} \mu_a$, $\Delta_a = \mu^* - \mu_a$, and $k_a(t)$ represents the number of times arm a has been played up to time t. Without loss of generality, we will assume that arm 1 is the best arm. We discuss the MAB setting in Section 5.

2.2. Infinite-horizon Markov Decision Processes

We focus on average-reward MDPs with infinite horizon (Jaksch et al., 2010; Wei et al., 2021), which is underexplored compared to the episodic setting. It is a more realistic model for real-world tasks, such as robotics and financial-market decision making, where state reset is not possible. Specifically, we consider an undiscounted *weakly communicating* MDP $(\mathcal{S}, \mathcal{A}, p, \mathcal{R})$ with infinite time horizon[4] (Ouyang et al., 2017; Theocharous et al., 2017b), where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, p represents the parameterized transition dynamics, and $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function. We assume that $\theta \in \mathbb{R}^d$ parameterizes the transition dynamics and there exists a true unknown θ^* governing the next state of the learner. At each time t, the learner is in state s_t, takes action a_t, and transits into the next state s_{t+1}, which is drawn from $p(\cdot|s_t, a_t, \theta^*)$.

We consider two sub-settings based on the parameterization of the transition dynamics: the General Parameterization and the Simplex Parameterization. These sub-settings require different assumptions and setups, which we elaborate on in their respective sections (Section 6.2 and Section 6.3). In the MDP context, the notion of batch is more appropriately thought of as a policy switch. Therefore, B_k now represents the starting time of the k-th policy switch, and we additionally define T_k as the number of time steps between policy switch k and $k+1$. We consider stationary and deterministic policies, which are mappings from $\mathcal{S} \to \mathcal{A}$. Let π_k be the policy followed by the learner after the k-th policy switch. When the decision to update and obtain the k-th policy is made, the learner uses the observed data $\{s_t, a_t, \mathcal{R}(s_t, a_t), s_{t+1}\}_{t=B_k}^{B_{k+1}-1}$ collected after the $(k-1)$-th policy switch to sample from the updated posterior and compute π_k. The goal of the learner is to maximize the long-term average reward:

$$J^\pi(\theta) = \mathbb{E}\left[\limsup_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathcal{R}(s_t, a_t)\right].$$

Similar to other works (Ouyang et al., 2017; Osband et al., 2013; Theocharous et al., 2017b), we measure the performance using Bayesian regret[5] defined by:

$$R_B(T) := \mathbb{E}\left[\sum_{t=1}^{T} (J^{\pi^*}(\theta^*) - \mathcal{R}(s_t, a_t))\right], \quad (2)$$

where $J^{\pi^*}(\theta^*)$ denotes the average long-term reward after running the optimal policy under the true model.

It is known that weakly communicating MDPs satisfy the following Bellman optimality (Bertsekas, 2012; Ouyang et al., 2017; Wei et al., 2021) in infinite-horizon setting, and there exists some positive number H such that the span (Definition 2) satisfies $sp(h(\theta)) \leq H$ for all $\theta \in \mathbb{R}^d$.

Lemma 1 (Bellman Optimality)**.** *There exist optimal average reward $J \in \mathbb{R}$ and a bounded measurable function $h : \mathcal{S} \to \mathbb{R}$, such that for any $s, \in \mathcal{S}, \theta \in \mathbb{R}^d$, the Bellman optimality equation holds:*

$$J(\theta) + h(s, \theta) = \max_{a \in \mathcal{A}} \left\{\mathcal{R}(s, a) + \mathbb{E}_{s' \sim p(\cdot|s,a;\theta)}[h(s', \theta)]\right\}. \quad (3)$$

Here $J(\theta) = max_\pi J^\pi(\theta)$ under θ and is independent of initial state. Function $h^\pi(s, \theta) = \lim_{T \to \infty} \mathbb{E}[\sum_{t=1}^{T} (\mathcal{R}(s_t, \pi(s_t)) - J^\pi(s_t))|s_1 = s]$ quantifies the bias of policy π w.r.t the average-reward under θ, and $h(s, \theta) = h^{\pi^*}(s, \theta)$, where $\pi^* = \text{argmax}_\pi J^\pi(\theta)$.

Definition 2. *For any $\theta \in \mathbb{R}^d$, span of an MDP is defined as $sp(h(\theta)) := sup_{s,s' \in \mathcal{S}} |h(s, \theta) - h(s', \theta)| = max_{s \in \mathcal{S}} h(s, \theta) - min_{s \in \mathcal{S}} h(s, \theta)$.*

we consider the ambient dimension for the parameters of each arm to be the same.

[4] It is known that weakly communicating MDPs satisfy the Bellman Optimality.

[5] In Bayesian regret, expectation is taken $w.r.t$ the prior distribution of the true parameter θ^*, the randomness of algorithm and transition dynamics.

3. Related Work

Under the conjugacy assumptions on rewards, asymptotic convergence of TS was studied in stochastic MABs by Granmo (2010) and May et al. (2012). Later, finite-time analyses with $O(\log T)$ problem-dependent regret bound were provided (Agrawal and Goyal, 2012; Kaufmann et al., 2012; Agrawal and Goyal, 2013). However, in practice, exact posteriors are intractable for all but the simplest models (Riquelme et al., 2018), necessitating the use of approximate sampling methods with TS in complex problem domains. Recent progress has been made in understanding approximate TS in fully-sequential MABs (Lu and Van Roy, 2017; Mazumdar et al., 2020; Zhang, 2022; Xu et al., 2022). On the other hand, the question of learning with TS in the presence of batched data has evolved along a separate trajectory of works (Karbasi et al., 2021; Kalkanli and Ozgur, 2021; Vernade et al., 2020; Zhang et al., 2020). However, provably performing Langevin TS in batched settings remains unexplored, and in this paper, we aim at bridging these lines.

Moving to the more complex decision-making frameworks based on MDPs, TS is employed in model-based methods to learn transition models, which is known as *Posterior Sampling for Reinforcement Learning* (PSRL) (Strens, 2000). When exact posteriors are intractable, MCMC methods have been empirically studied for performing Bayesian inference in policy and reward spaces in RL (Brown et al., 2020; Imani et al., 2018; Bojun, 2020; Guez et al., 2014). MCMC is a family of approximate posterior inference methods that enables sampling without exact knowledge of posteriors (Ma et al., 2015; Welling and Teh, 2011). However, it is unclear how to provably incorporate MCMC methods in learning transition models for RL.

Furthermore, the analysis of undiscounted infinite-horizon MDPs (Abbasi-Yadkori and Szepesvári, 2015; Osband and Van Roy, 2016; Ouyang et al., 2017; Wei et al., 2020; 2021) poses greater challenges compared to the well-studied episodic MDPs with finite horizon and fixed episode length (Osband et al., 2013). Previous works on infinite-horizon settings include model-based methods that estimate environment dynamics and switch policies when the number of visits to state-action pairs doubles (Jaksch et al., 2010; Tossou et al., 2019; Agrawal and Jia, 2017; Bartlett and Tewari, 2012). Nevertheless, under such dynamic schemes, the number of policy switches can be as large as $O(\sqrt{T})$, making it computationally heavy and infeasible for continuous states and actions. To enable TS with logarithmic policy switches while maintaining optimal regret, we build upon an algorithmically-independent static scheme as in Theocharous et al. (2017b), and incorporate Langevin Monte Carlo (LMC) methods to sample from inexact posteriors.

4. SGLD for Langevin Thompson sampling

In the MAB and MDP settings, θ parameterizes the unknown reward or transition distributions respectively. TS maintains a distribution over the parameters and updates the distribution to (the new) posterior upon receiving new data. Given $p(X|\theta)$, prior $\lambda(\theta)$, and n data samples $\{X_i\}_{i=1}^n$[6], let ρ_n be the posterior distribution after receiving n data samples which satisfies: $\rho(\theta|\{X_i\}_{i=1}^n) \propto \exp(\sum_{i=1}^n \log p(X_i|\theta) + \log\lambda(\theta))$. In addition, consider the scaled posterior $\rho_n[\gamma]$ for some scaling parameter γ, which represents the density proportional to $\exp(\gamma(\sum_{i=1}^n \log p(X_i|\theta) + \log\lambda(\theta)))$.

The introduction of MCMC methods arises from the need for sampling from intractable posteriors in the absence of conjugacy assumptions. We resort to a gradient-based MCMC method that performs noisy updates based on Langevin dynamics: Stochastic Gradient Langevin Dynamics (SGLD). Algorithm 1 presents SGLD with bached data to generate samples from an approximation of the true posterior. For a detailed exposition, please refer to (Welling and Teh, 2011; Ma et al., 2015) and Appendix A. Algorithm 1 takes all available data $\{X_s\}_{s=1}^n$ at the start of a batch b as input, subsamples data, performs gradient updates by computing $\nabla\widehat{U}(\theta) = -\frac{n}{|D|}\sum_{X_s\in D}\nabla\log p(X_s|\theta) - \nabla\log\lambda(\theta)$, and outputs the posterior for batch b.

Algorithm 1 SGLD with Batched Data

Input: prior $\lambda(\theta)$, data $\{X_s\}_{s=1}^n$, sample from last batch θ^{b-1}, total iterations N, learning rate η, parameters L, scaling parameter γ.
Initialization: $\theta_0 \leftarrow \theta^{b-1}$
for $i = 1, \dots, N$ **do**
  Subsample $D \subseteq \{X_s\}_{s=1}^n$
  Compute $\nabla\widehat{U}(\theta_{i\eta})$ over D
  Sample $\theta_{(i+1)\eta} \sim \mathcal{N}(\theta_{i\eta} - \eta\nabla\widehat{U}(\theta_{i\eta}),\ 2\eta I)$
Output: $\theta^b \sim \mathcal{N}\big(\theta_{N\eta},\ \frac{1}{nL\gamma}I\big)$

In the batched setting, new data is received at the end of a batch, or when making the decision to perform a new policy switch. Due to the way that the learner receives new data and the fact that the batch data size may increase exponentially[7], the posterior concentrates slower. This differs from the fully-sequential problem where the distribution shift of successive true posteriors is small owing to data being received in an iterative manner. We show that in batched settings, with only constant computational complexity in terms of iterations, SGLD is able to provide strong convergence guarantee as in the fully-sequential setting (Mazumdar et al., 2020). Theorem 1 shows the convergence of

[6] Here data $\{X_i\}_{i=1}^n$ can be rewards for some arms or actual transitions of state-action pairs depending on the setting.
[7] Data received in batch k can be doubled compared to the previous batch.

SGLD in the Wasserstein-p distance can be achieved with a constant number of iterations and data.

Theorem 1 (SGLD convergence)**.** *Suppose that the parametric reward/transition families, priors, and true reward/transition distributions satisfy Assumptions 1-4. Let* $\kappa := \max\{L/m, L/\nu\}$, $|D| = O(\kappa^2)$, $\eta = O(1/n\kappa L)$, *and* $N = O(\kappa^2)$, *then for any* $\delta \in (0,1)$, *the following holds with probability* $\geq 1-\delta$*:*

$$W_p\left(\tilde{\rho}_n, \rho_n\right) \leq \sqrt{\frac{12}{nm}}(d + \log Q + (32 + 8d\kappa^2)p)^{1/2}$$

for all $p \geq 2$, *and where* $Q := \max_\theta \frac{\lambda(\theta)}{\lambda(\theta^*)}$ *measures the quality of prior distribution.*

ρ_n denotes the true posterior corresponding to n data samples and $\tilde{\rho}_n$ is the approximate posterior outputted by Algorithm 1. We also note that similar concentration bounds can be achieved by using the *Unadjusted Langevin Algorithm (ULA)* for batched data, which adopts full-batch gradient evaluations and therefore leads to a growing iteration complexity. The proofs of Theorem 1 are adapted to the batched setting, which differs from (Mazumdar et al., 2020).

5. Batched Langevin Thompson Sampling for Bandits

In this section, we introduce *Langevin Thompson Sampling* for batched stochastic MAB setting in Algorithm 2, namely, BLTS. It leverages SGLD and batching schemes to learn a wide class of unknown reward distributions while reducing communication and computation costs. We have previously discussed the results of SGLD in Section 4 for both MABs and MDPs. Here, we focus on the batching strategy in Algorithm 2 for bandits, and discuss the resulting regret guarantee.

5.1. Dynamic Doubling Batching Scheme

BLTS keeps track of the number of times each arm a has been played until time t with $k_a(t)$. Initially, all $\{k_a\}_{a\in\mathcal{A}}$ are set to 0. The size of each batch is determined by $\{k_a\}_{a\in\mathcal{A}}$ and the corresponding integers $\{l_a\}_{a\in\mathcal{A}}$. Once k_a reaches 2^{l_a} for some arm a, BLTS makes the decision to terminate the current batch, collects all rewards from the batch in a single request, and increases l_a by 1. BLTS thus starts a new batch whenever an arm is played twice as many times as in the previous batch, which results in growing batch sizes. As the decision to move onto the next batch depends on the sequence of arms that is played, it is considered as "dynamic". This batching scheme is similar to the one used in Karbasi et al. (2021). The total number of batches that BLTS carries out satisfies the following theorem, and its proof can be found in Appendix D.

Theorem 2. *BLTS ensures that the total number of batches is at most* $O(N \log T)$ *where* $N = |\mathcal{A}|$.

Gao et al. (2019) showed that $\Omega(\log T/\log\log T)$ batches are required to achieve the optimal logarithmic dependence in time horizon T for a batched MAB problem. This shows that the dependence on T in the number of batches BLTS requires is at most a factor of $\log\log T$ off the optimal. We now state and discuss the BLTS algorithm.

5.2. Regret of BLTS Algorithm

In Algorithm 2, denote by θ_a^k the output of Algorithm 1 for arm a at batch k. At the end of each batch, new data is acquired all at once and the posterior is being updated. It is important to note that upon receiving new data when we run Algorithm 1 for each arm, only that arm's data is fed into Algorithm 1. For each $a \in \mathcal{A}$, assume the existence of linear map α_a such that $\mathbb{E}_{X\sim p_a(X|\theta_a)}[X] = \alpha_a^\intercal\, \theta_a\ \forall \theta_a \in \mathbb{R}^d$, where $\|\alpha_a\|$ is bounded. Theorem 3 shows the regret guarantee of BLTS.

Theorem 3. *Assume that the parametric reward families, priors, and true reward distributions satisfy Assumptions 1 through 4 for each arm* $a \in \mathcal{A}$. *Then with the SGLD parameters specified as per Algorithm 1 and with* $\gamma = O(1/d\kappa^3)$ *(for* $\kappa := \max\{L/m, L/\nu\}$*), BLTS satisfies:*

$$R(T) \leq \sum_{a>1} \frac{C\sqrt{Q_1}}{m\Delta_a}\left(d + \log Q_1 + d\kappa^2 \log T + d^2\kappa^2\right) + \frac{C}{m\Delta_a}\left(d + \log Q_a + d^2\kappa^2 \log T\right) + 4\Delta_a,$$

where C *is a constant and* $Q_a := \max_\theta \frac{\lambda_a(\theta)}{\lambda_a(\theta^*)}$. *The total number of SGLD iterations used by BLTS is* $O(\kappa^2 N log T)$.

Discussion We show that BLTS achieves the optimal $O\left(\frac{\log T}{\triangle}\right)$ regret bound with exponentially fewer rounds of communication between the learner and the environment. Result of Theorem 3 relies on both the statistical guarantee provided by SGLD and the design of our batching scheme. In bached setting, one must carefully consider the trade-off between batch size and the number of batches. While it is desirable to reuse the existing posterior for sampling within a batch, the batching scheme must also ensure new data is collected in time to avoid significant distribution shifts. In addition, the use of SGLD allows BLTS to be applicable in a wide range of general settings with a low computation cost of $O(\kappa^2 N \log T)$.

In the regret bound of Theorem 3, Q_a measures the quality of prior for arm a. Specifically, if the prior is properly centered such that its mode is at θ_a^*, or if the prior is uninformative or flat everywhere, then $\log Q_a = 0$. In Section 7, we show that using either favorable priors or uninformative priors provides similar empirical performance as existing methods.

Algorithm 2 Batched Langevin Thompson Sampling (BLTS)

Input: priors $\lambda_a(\theta)\ \forall a \in \mathcal{A}$, scaling parameter γ, inputs for SGLD subroutine N, η, L.
Initialization: $k_a \leftarrow 0, l_a \leftarrow 0, n_a \leftarrow 0, \tilde{\rho}_{a,k} = \tilde{\rho}_{a,0} = \lambda_a\ \forall a \in \mathcal{A}$, batch index $k \leftarrow 0$.
for $t = 1, \ldots, T$ **do**
  Sample $\theta_{a,t} \sim \mathcal{N}\big(\theta_a^k,\ \frac{1}{n_a L\gamma} I\big)\ \forall a \in \mathcal{A}$
  Choose action $a_t = \operatorname{argmax}_{a\in\mathcal{A}} \alpha_a^\intercal \theta_{a,t}$
  Update $k_{a(t)} \leftarrow k_{a(t)} + 1$
  if $k_{a_t} = 2^{l_{a(t)}}$ **then**
    $l_{a(t)} \leftarrow l_{a(t)} + 1$
    Terminate batch k and observe rewards $\{r_{a_i}\}_{i=B_k}^t$
    for $a \in \mathcal{A}$ **do**
      Update n_a with the number of new samples
      Run Algorithm 1 to obtain $\tilde{\rho}_{a,k+1}$ and θ_a^{k+1}
      Update batch index $k \leftarrow k + 1$

6. Batched Langevin Posterior Sampling For RL

In RL frameworks, posterior sampling is commonly used in model-based methods to learn unknown transition dynamics and is known as PSRL[8]. In infinite-horizon settings, PSRL operates by sampling a model and solving for an optimal policy based on the sampled MDP at the beginning of each policy switch. The learner then follows the same policy until the next policy switch. In this context, the concept of a batch corresponds to a policy switch.

Previous analyses of PSRL have primarily focused on transition distributions that conform to well-behaved conjugate families. Handling transitions that deviate from these families and computing the corresponding posteriors has been heuristically left to MCMC methods. Here, we provably extend PSRL with LMC and introduce *Langevin Posterior Sampling for RL* (LPSRL, Algorithm 3) using a static doubling policy-switch scheme. Analyses of PSRL have crucially relied on the true transition dynamics θ^* and the sampled MDPs being identically distributed (Osband et al., 2013; Osband and Van Roy, 2016; Russo and Van Roy, 2014; Ouyang et al., 2017; Theocharous et al., 2017b). However, when the dynamics are sampled from an approximation of the true posterior, this fails to hold. To address the issue, we introduce the Langevin posterior sampling lemma (Lemma 3), which shows approximate sampling yields an additive error in the Wasserstein-1 distance.

Lemma 3. *(Langevin Posterior Sampling). Let t_k be the beginning time of policy-switch k, $\mathcal{H}_{t_k} := \{s_\tau, a_\tau\}_{\tau=1}^{t_k}$ be the history of observed states and actions till time t_k, and $\theta^k \sim \tilde{\rho}_{t_k}$ be the sampled model from the approximate posterior $\tilde{\rho}_{t_k}$ at time t_k. Then, for any $\sigma(\mathcal{H}_{t_k})$-measurable function f that is 1-Lipschitz, it holds that:*

$$\Big|\mathbb{E}[f(\theta^*)|\mathcal{H}_{t_k}] - \mathbb{E}[f(\theta^k)|\mathcal{H}_{t_k}]\Big| \le W_1(\tilde{\rho}_{t_k}, \rho_{t_k}). \quad (4)$$

By the tower rule, $\Big|\mathbb{E}[f(\theta^*)] - \mathbb{E}[f(\theta^k)]\Big| \le W_1(\tilde{\rho}_{t_k}, \rho_{t_k})$.

As demonstrated later, this error term can be effectively controlled and does not impact the overall regret (Theorems 4 and 6). It only requires the average reward function $J^\pi(\theta)$ to be 1-Lipschitz, as specified in **Assumption 5**[9]. Let us consider the parameterization of the transition dynamics p with $\theta \in \mathbb{R}^d$, where $\theta^* \in \mathbb{R}^d$ denotes the true (unknown) parameter governing the dynamics. We explore two distinct settings based on these parameterizations:

- **General Parameterization** (Section 6.2): In this setting, we consider modeling the full transition dynamics using $\theta^* \in \mathbb{R}^d$, where $d \ll |\mathcal{S}||\mathcal{A}|$. This parameterization can be particularly useful for tackling large-scale MDPs with large (or even continuous) state and action spaces. Towards this end, we consider $\mathcal{S} \cong \mathbb{R}$. Examples of General Parameterization include linear MDPs with feature mappings (Jin et al., 2020), RL with general function approximation (Yang et al., 2020), and the low-dimensional structures that govern the transition (Gopalan and Mannor, 2015; Yang and Wang, 2020). We provide a real-world example that adopts such parameterization in Appendix E.3.

  Despite Theocharous et al. (2017b) studies a similar setting, their work confines the parameter space to $\mathbb{R}$. To accommodate a broader class of MDPs, we generalize the parameter space to $\mathbb{R}^d$. As suggested by Theorem 4, our algorithm retains the optimal $O(\sqrt{T})$ regret with $O(\log T)$ policy switches, making it applicable to a wide range of general transition dynamics.

- **Simplex Parameterization** (Section 6.3): Here, we consider the classical tabular MDPs with finite states and actions. For each state-action pair, there exists a probability simplex $\Delta^{|\mathcal{S}|}$ that encodes the likelihood of transitioning into each state. Hence, in this case, $\theta^* \in \mathbb{R}^d$ with $d = |\mathcal{S}|^2|\mathcal{A}|$. This structure necessitates sampling transition dynamics from constrained distributions, which naturally leads us to instantiate LPSRL with the *Mirrored Langevin Dynamics* (Hsieh et al., 2018) (See Appendix A for more discussions). As proven in Theorem 6, LPSRL with MLD achieves the optimal $O(\sqrt{T})$ regret with $O(\log T)$ policy switches for general transition dynamics subject to the probability simplex constraints.

[8] We also depart from using TS for the RL setting and stick to the more popular posterior sampling terminology for RL.

[9] Mathematical statement is in Appendix B.

6.1. The LPSRL Algorithm

LPSRL (Algorithm 3) use `SamplingAlg` as a subroutine, where SGLD and MLD are invoked respectively depending on the parameterization. Unlike the BLTS algorithm in bandit settings, LPSRL adopts a static doubling batching scheme, in which the decision to move onto the next batch is independent of the dynamic statistics of the algorithm, and thus is algorithmically independent.

Let t_k be the starting time of policy-switch k and let $T_k := 2^{k-1}$ represent the total number of time steps between policy-switch k and $k+1$. At the beginning of each policy-switch k, we utilize `SamplingAlg` to obtain an approximate posterior distribution $\tilde{\rho}_{t_k}$ and sample dynamics θ^k from $\tilde{\rho}_{t_k}$. A policy π_k is then computed for θ^k with any planning algorithm[10]. The learner follows π_k to select actions and transit into new states during the remaining time steps before the next policy switch. New Data is collected all at once at the end of k. Once the total number of time steps is being doubled, i.e., t reaches $t_k + T_k - 1$, the posterior is updated using the latest data D, and the above process is repeated.

Algorithm 3 Langevin PSRL (LPSRL)

Input: MCMC scheme `SamplingAlg` initiated with prior $\lambda(\theta)$.
Initialization: time step $t \leftarrow 1$, $D \leftarrow \emptyset$
for *batch* $k = 1, \ldots, K_T$ **do**
  $T_k \leftarrow 2^{k-1}$
  $t_k \leftarrow 2^{k-1}$
  Run `SamplingAlg` and sample θ^k from posterior: $\theta^k \sim \tilde{\rho}_{t_k}(\theta|D)$
  Compute optimal policy π_k based on θ^k
  for $t = t_k, t_k+1, \cdots, t_k+T_k-1$ **do**
    Choose action $a_t \sim \pi_k$
    Generate immediate reward $\mathcal{R}(s_t, a_t)$, transit into new state s_{t+1}
  $D \leftarrow D \cup \{s_t, a_t, \mathcal{R}(s_t, a_t), s_{t+1}\}_{t=t_k}^{t_k+T_k-1}$

6.2. General Parametrization

In RL context, to study the performance of LPSRL instantiated with SGLD as `SamplingAlg`, Assumptions 1-4 are required to hold on the (unknown) transition dynamics, rather than the (unknown) rewards as in the bandit setting. Additionally, similar to Theocharous et al. (2017b), the General Parameterization requires $p(\cdot|\theta)$ to be Lipschitz in θ (**Assumption 6**). Mathematical statements of all assumptions are in Appendix B. We now state the main theorem for LPSRL under the General Parameterization.

Theorem 4. *Under Assumptions $1-6$, by instantiating `SamplingAlg` with SGLD and setting the hyperparameters as per Theorem 1, with $p = 2$, the regret of LPSRL (Algorithm 3) satisfies:*

$$R_B(T) \leq CH \log T \sqrt{\frac{T}{m}(d + \log Q + (32 + 8d\kappa^2)p)^{1/2}},$$

where C is some positive constant, H is the upper bound of the MDP span, and Q denotes the quality of the prior. The total number of iterations required for SGLD is $O(\kappa^2 \log T)$.

Discussion. LPSRL with SGLD maintains the same order-optimal regret as exact PSRL in (Theocharous et al., 2017b). Similar to Theorem 3, the regret bound has explicit dependence on the quality of prior imposed to transitions, where $\log Q = 0$ when prior is properly centered with its mode at θ^*, or when it is uninformative or flat. Let $\theta^{k,*}$ be the true posterior in policy-switch k. Our result relies on θ^* and $\theta^{k,*}$ being identically distributed, and the convergence of SGLD in $O(\log T)$ iterations to control the additive cumulative error in $\sum_{k=1}^{K_T} T_k W_1(\tilde{\rho}_{t_k}, \rho_{t_k})$ arising from approximate sampling.

6.3. Simplex Parametrization

We now consider the tabular setting where θ^* specifically models a collection of $|\mathcal{A}|$ transition matrices in $[0,1]^{|\mathcal{S}|\times|\mathcal{S}|}$. Each row of the transition matrices lies in a probability simplex $\Delta^{|\mathcal{S}|}$, specifying the transition probabilities for each corresponding state-action pair. In particular, if the learner is in state $s \in \mathcal{S}$ and takes action $a \in \mathcal{A}$, then it lands in state s' with $p(s') = p(s'|s, a, \theta^*)$. In order to run LPSRL on constrained space, we need to sample from probability simplexes and therefore appeal to the Mirrored Langevin Dynamics (MLD) (Hsieh et al., 2018) by using the entropic mirror map, which satisfies the requirements set forth by Theorem 2 in Hsieh et al. (2018). Under Assumptions 5 and 6, we have the following convergence guarantee for MLD and regret bound for LPSRL under the Simplex Parameterization.

Theorem 5. *At the beginning of each policy-switch k, for each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, sample transition probabilities over $\Delta^{|\mathcal{S}|}$ using MLD with the entropic mirror map. Let n_{t_k} be the number of data samples for any (s, a) at time t_k, then with step size chosen per Cheng and Bartlett (2018), running MLD with $O(n_{t_k})$ iterations guarantees that $W_2(\tilde{\rho}_{t_k}, \rho_{t_k}) = \tilde{O}\left(\sqrt{|\mathcal{S}|/n_{t_k}}\right)$.*

Theorem 6. *Suppose Assumptions 5 and 6 are satisfied, then by instantiating `SamplingAlg` with MLD (Algorithm 4), there exists some positive constant C such that the regret of LPSRL (Algorithm 3) in the Simplex Parameterization is bounded by*

$$R_B(T) \leq CH|\mathcal{S}|\sqrt{|\mathcal{A}|T\log(|\mathcal{S}||\mathcal{A}|T)},$$

[10] We assume the optimality of policies and focus on learning the transitions. When only sub-optimal policies are available in our setting, it can be shown that small approximation errors in policies only contribute additive non-leading terms to regret. See details in (Ouyang et al., 2017).

where C is some positive constant, H is the upper bound of the MDP span. The total number of iterations required for MLD is $O(|\mathcal{S}|^2|\mathcal{A}|^2T)$.

Discussion. In simplex parameterization, instantiating LPSRL with MLD achieves the same order-optimal regret, but the computational complexity in terms of iterations for MLD is linear in T as opposed to $\log T$ for SGLD in the General Parameterization. Nevertheless, given that the simplex parameterization implies simpler structures, we naturally have fewer assumptions for the theory to hold.

7. Experiments

In this section, we perform empirical studies in simulated environments for bandit and RL to corroborate our theoretical findings. By comparing the actual regret (average rewards) and the number of batches for interaction (maximum policy switches), we show *Langevin TS* algorithms empowered by LMC methods achieve appealing statistical accuracy with low communication cost. For additional experimental details, please refer to Appendix F.

7.1. Langevin TS in Bandits

We first study how Langevin TS behaves in learning the true reward distributions of log-concave bandits with different priors and batching schemes. Specifically, we construct two bandit environments[11] with Gaussian and Laplace reward distributions, respectively. While both environments are instances of log-concave families, Laplace bandits do not belong to conjugate families.

7.1.1. Gaussian Bandits

We simulate a Gaussian bandit environment with $N = 15$ arms. The existence of closed-form posteriors in Gaussian bandits allows us to benchmark against existing exact TS algorithms. More specifically, we instantiate Langevin TS with SGLD (SGLD-TS), and perform the following tasks:

- Compare SGLD-TS against both frequentist and Bayesian methods, including UCB1, Bayes-UCB, decaying ϵ-greedy, and exact TS.
- Apply informative priors and uninformative priors for Bayesian methods based on the availability of prior knowledge in reward distributions.
- Examine all methods under three batching schemes: fully-sequential mode, dynamic batch, static batch.

Results and Discussion. Figure 1(a) illustrates the cumulative regret for SGLD-TS and Exact-TS with favorable priors. Table 2 reports the regret upon convergence along with the total number of batches in interaction. Note that SGLD-TS equipped with dynamic batching scheme implements Algorithm 2 (BLTS). Empirical results demonstrate that SGLD-TS is comparable to Exact-TS under all batching schemes, and is empirically more appealing compared to UCB1 as well as Bayes-UCB. While static batch incurs slightly lower communication costs compared to dynamic batch, results show that all methods under dynamic batch scheme are more robust with smaller standard deviation. Our BLTS algorithm thus well balances the trade-off between statistical performance, communication, and computational efficiency by achieving the order-optimal regret with a small number of batches.

	SGLD-TS	**Exact-TS**	**UCB1**	**Bayes-UCB**	**Batches**
Fully sequential	99.66± 13.09	99.07± 12.23	154.13+ −4.10	160.55± 25.75	650.0± 0.0
Static batch	148.52± 39.28	145.94± 31.46	155.17± 5.06	231.80± 52.11	9.0 ± 0.0
Dynamic batch	99.80± 15.62	98.71± 12.10	153.31± 3.83	214.43± 0.5	22.93± 1.50

Table 2: Average regret with the standard deviation under different batching schemes. The last column quantifies communication cost $w.r.t$ the total number of batches for interaction. BLTS (SGLD-TS under dynamic batching scheme) achieves order-optimal regret with low communication cost.

7.1.2. Laplace Bandits

To demonstrate the applicability of Langevin TS in scenarios where posteriors are intractable, we construct a Laplace bandit environment with $N = 10$ arms. It is important to note that Laplace reward distributions do not have conjugate priors, rendering exact TS inapplicable in this setting. Therefore, we compare the performance of SGLD-TS with favorable priors against UCB1. Results presented in Figure 1(b) reveal that, similar to the Gaussian bandits, SGLD-TS with dynamic batching scheme achieves comparable performance as in the fully-sequential setting and significantly outperforms UCB1, highlighting its capability to handle diverse environments. In addition, the static batching scheme exhibits larger deviations compared to the dynamic batching, which aligns with results in Table 2.

7.2. Langevin PSRL in Average-reward MDPs

In MDP setting, we consider a variant of RiverSwim environment (Strehl and Littman, 2008), which is a common testbed for provable RL methods. Specifically, it models an agent swimming in the river with five states, two actions ($|\mathcal{S}| = 5, |\mathcal{A}| = 2$). In this tabular case, LPSRL (Algorithm 3) employs MLD (Algorithm 4 in Appendix A) as `SamplingAlg`, namely, MLD-PSRL. We benchmark the performance of MLD-PSRL against other mainstream

[11] Our theories apply to bandits with a more general family of reward distributions.

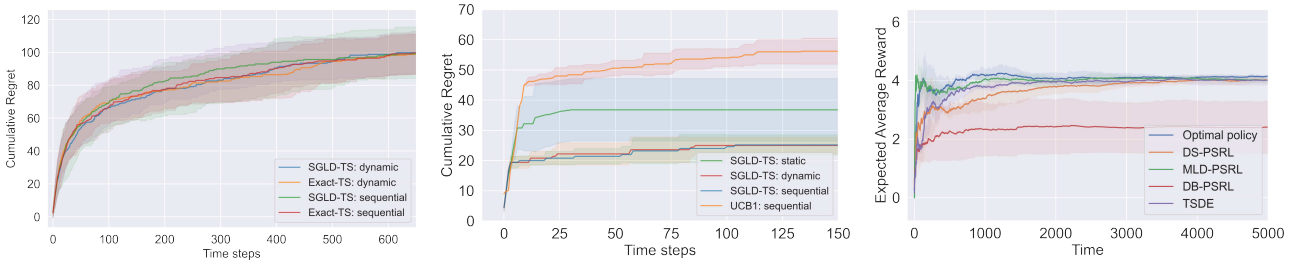

Figure 1: (a) Regret in Gaussian Bandits ($N = 15$): expected regret is reported over 10 experiments with informative priors. Results show SGLD-TS under dynamic batching scheme achieves optimal performance as in the sequential case without using approximate sampling. Results with uninformative priors yield the same conclusions (See Appendix F). (b) Regret in Laplace Bandits ($N = 10$): regret is reported over 10 experiments with informative priors. As in Gaussian Bandits, SGLD-TS with dynamic batching scheme achieves optimal regret and outperforms UCB1. (c) Average reward in RiverSwim: expected average reward is reported over 10 experiments. MLD-PSRL achieves optimal average reward upon convergence with a small number of policy switches.

model-based RL methods, including TSDE (Ouyang et al., 2017), DS-PSRL (Theocharous et al., 2017b) and DB-PSRL (exact-PSRL(Strens, 2000) with dynamic batch). Note that MLD-PSRL and DS-PSRL adopt the static doubling policy switch scheme discussed in section 6. Dynamic doubling policy switch scheme adopted by both DB-PSRL and TSDE is akin to the one we use in bandit setting, but based on the visiting counts of state-action pairs. We simulate 10 different runs of experiment, and report the average rewards obtained by each method in Figure 1(c). Mechanisms used by each method are summarized in Table 3, along with the average rewards achieved and maximum number of policy switches incurred.

	MLD-PSRL	**DS-PSRL**	**DB-PSRL**	**TSDE**	**Optimal policy**
Static ps	$\surd$	$\surd$			
Dynamic ps			$\surd$	$\surd$	
Linear growth				$\surd$	
Avg. reward	4.01 ± 0.11	4.02 ± 0.08	2.41 ± 0.91	4.01 ± 0.17	4.15 ± 0.04
Max. switches	12.0 ± 0.0	12.0 ± 0.0	15.33 ± 1.70	94.0 ± 3.56	-

Table 3: We report the average reward and the maximum number of policy switches all methods over 10 different runs. MLD-PSRL instantiates Alogrithm 3 in Section 6, which achieves order-optimal performance with small number of policy switches.

Results and Discussion. We demonstrate that MLD-PSRL achieves comparable performance compared to existing PSRL methods while significantly reducing communication costs through the use of static policy switches. In contrast, as illustrated in Figure 4 (Appendix F) and Table 3, TSDE achieves near-optimal performance but requires high communication costs. Additionally, our empirical results reveal that the static policy switch in the MDP setting outperforms the dynamic policy switch alone. This observation aligns with existing findings that frequent policy switches in MDPs can harm performance. Moreover, compared to DS-PSRL, MLD-PSRL is applicable to more general frameworks when closed-form posterior distributions are not available[12].

8. Conclusion

In this paper, we jointly address two challenges in the design and analysis of Thompson sampling (TS) methods. Firstly, when dealing with posteriors that do not belong to conjugate families, it is necessary to generate approximate samples within a reasonable computational budget. Secondly, when interacting with the environment in a batched manner, it is important to limit the amount of communication required. These challenges are critical in real-world deployments of TS, as closed-form posteriors and fully-sequential interactions are rare. In stochastic MABs, approximate TS and batched interactions are studied independently. We bridge the two lines of work by providing a Langevin TS algorithm that works for a wide class of reward distributions with only logarithmic communication. In the case of undiscounted infinite-horizon MDP settings, to the best of our knowledge, we are the first to provably incorporate approximate sampling with the TS paradigm. This enhances the applicability of TS for RL problems with low communication costs. Finally, we conclude with experiments to demonstrate the appealing empirical performance of the Langevin TS algorithms.

Acknowledgements

This work is supported in part by the National Science Foundation Grants NSF-SCALE MoDL(2134209) and NSF-CCF-2112665 (TILOS), the U.S. Department Of Energy, Office of Science, and the Facebook Research award. Amin Karbasi acknowledges funding in direct support of this work from NSF (IIS-1845032), ONR (N00014-19-1-2406), and the AI Institute for Learning-Enabled Optimization at Scale (TILOS).

[12] Different mirror maps can be applied to MLD method depending on parameterization.

References

Abbasi-Yadkori, Y. and Szepesvári, C. (2015). Bayesian optimal control of smoothly parameterized systems. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*. AUAI Press.

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

Agrawal, S. and Goyal, N. (2013). Further optimal regret bounds for thompson sampling. In *Artificial intelligence and statistics*, pages 99–107. PMLR.

Agrawal, S. and Jia, R. (2017). Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, 30.

Bartlett, P. L. and Tewari, A. (2012). Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. *arXiv preprint arXiv:1205.2661*.

Bertsekas, D. (2012). *Dynamic programming and optimal control: Volume I*, volume 1. Athena scientific.

Bojun, H. (2020). Steady state analysis of episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 33:9335–9345.

Brown, D., Coleman, R., Srinivasan, R., and Niekum, S. (2020). Safe imitation learning via fast bayesian reward inference from preferences. In *International Conference on Machine Learning*, pages 1165–1177. PMLR.

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

Cheng, X. and Bartlett, P. (2018). Convergence of langevin mcmc in kl-divergence. In *Algorithmic Learning Theory*, pages 186–211. PMLR.

Gao, Z., Han, Y., Ren, Z., and Zhou, Z. (2019). Batched multi-armed bandits problem.

Gopalan, A. and Mannor, S. (2015). Thompson Sampling for Learning Parameterized Markov Decision Processes. In *Proceedings of The 28th Conference on Learning Theory*. PMLR.

Granmo, O.-C. (2010). Solving two-armed bernoulli bandit problems using a bayesian learning automaton. *International Journal of Intelligent Computing and Cybernetics*.

Guez, A., Silver, D., and Dayan, P. (2014). Better optimism by bayes: Adaptive planning with rich models. *arXiv preprint arXiv:1402.1958*.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., et al. (2018). Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*.

Honda, J. and Takemura, A. (2013). Optimality of thompson sampling for gaussian bandits depends on priors.

Hsieh, Y.-P., Kavis, A., Rolland, P., and Cevher, V. (2018). Mirrored langevin dynamics. *Advances in Neural Information Processing Systems*, 31.

Imani, M., Ghoreishi, S. F., and Braga-Neto, U. M. (2018). Bayesian control of large mdps with unknown dynamics in data-poor environments. *Advances in neural information processing systems*, 31.

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is q-learning provably efficient? *Advances in neural information processing systems*, 31.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provablylearning to optimize via posterior sampling efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR.

Jung, Y. H., Abeille, M., and Tewari, A. (2019). Thompson sampling in non-episodic restless bandits. *arXiv preprint arXiv:1910.05654*.

Kalkanli, C. and Ozgur, A. (2021). Batched thompson sampling. *Advances in Neural Information Processing Systems*, 34:29984–29994.

Karbasi, A., Mirrokni, V., and Shadravan, M. (2021). Parallelizing thompson sampling. *Advances in Neural Information Processing Systems*, 34:10535–10548.

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

Li, T., Wu, F., and Lan, G. (2022). Stochastic first-order methods for average-reward markov decision processes. *arXiv preprint arXiv:2205.05800*.

Lu, X. and Van Roy, B. (2017). Ensemble sampling. *arXiv preprint arXiv:1705.07347*.

Ma, Y.-A., Chen, T., and Fox, E. (2015). A complete recipe for stochastic gradient mcmc. *Advances in neural information processing systems*, 28.

May, B. C., Korda, N., Lee, A., and Leslie, D. S. (2012). Optimistic bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research*, 13:2069–2106.

Mazumdar, E., Pacchiano, A., Ma, Y.-a., Bartlett, P. L., and Jordan, M. I. (2020). On approximate Thompson sampling with Langevin algorithms. In *ICML*, volume 119, pages 6797–6807.

Osband, I., Russo, D., and Van Roy, B. (2013). (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26.

Osband, I. and Van Roy, B. (2014). Model-based reinforcement learning and the eluder dimension. *Advances in Neural Information Processing Systems*, 27.

Osband, I. and Van Roy, B. (2016). Posterior sampling for reinforcement learning without episodes. *arXiv preprint arXiv:1608.02731*.

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. (2017). Learning unknown markov decision processes: A thompson sampling approach. *Advances in neural information processing systems*, 30.

Riquelme, C., Tucker, G., and Snoek, J. (2018). Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. *arXiv preprint arXiv:1802.09127*.

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

Schwartz, E. M., Bradlow, E. T., and Fader, P. S. (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522.

Strehl, A. L. and Littman, M. L. (2008). An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331.

Strens, M. (2000). A bayesian framework for reinforcement learning. In *ICML*, volume 2000, pages 943–950.

Theocharous, G., Vlassis, N., and Wen, Z. (2017a). An interactive points of interest guidance system. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces Companion*, IUI '17 Companion.

Theocharous, G., Wen, Z., Abbasi-Yadkori, Y., and Vlassis, N. (2017b). Posterior sampling for large scale reinforcement learning. *arXiv preprint arXiv:1711.07979*.

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

Tian, Y., Qian, J., and Sra, S. (2020). Towards minimax optimal reinforcement learning in factored markov decision processes. *Advances in Neural Information Processing Systems*, 33:19896–19907.

Tossou, A., Basu, D., and Dimitrakakis, C. (2019). Near-optimal optimistic reinforcement learning using empirical bernstein inequalities. *arXiv preprint arXiv:1905.12425*.

Vernade, C., Carpentier, A., Lattimore, T., Zappella, G., Ermis, B., and Brueckner, M. (2020). Linear bandits with stochastic delayed feedback. In *International Conference on Machine Learning*, pages 9712–9721. PMLR.

Wei, C.-Y., Jahromi, M. J., Luo, H., and Jain, R. (2021). Learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3007–3015. PMLR.

Wei, C.-Y., Jahromi, M. J., Luo, H., Sharma, H., and Jain, R. (2020). Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *International conference on machine learning*, pages 10170–10180. PMLR.

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

Wu, Y., Zhou, D., and Gu, Q. (2022). Nearly minimax optimal regret for learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3883–3913. PMLR.

Xu, P., Zheng, H., Mazumdar, E. V., Azizzadenesheli, K., and Anandkumar, A. (2022). Langevin monte carlo for contextual bandits. In *International Conference on Machine Learning*, pages 24830–24850. PMLR.

Yang, L. and Wang, M. (2020). Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR.

Yang, Z., Jin, C., Wang, Z., Wang, M., and Jordan, M. I. (2020). On function approximation in reinforcement

learning: Optimism in the face of large state spaces. *arXiv preprint arXiv:2011.04622*.

Zhang, T. (2022). Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857.

Zhang, W., Zhou, D., Li, L., and Gu, Q. (2020). Neural thompson sampling.

Appendices

A. MCMC Methods

A.1. Unconstrained Approximate Sampling

Suppose a target distribution ρ is parameterized by $\theta \in \mathbb{R}^d$, and observed data $\{X_i\}_{i=1}^n$ are independently identically distributed. A posterior distribution defined up to a normalization factor can be expressed via the Gibbs distribution form:

$$\rho(\theta|X_1,\ldots,X_n) \propto \lambda(\theta)\prod_{i=1}^n p(X_i;\theta) = \exp\left(-U(\theta)\right),$$

where $\lambda(\theta)$ is the prior distribution of θ, $p(X_i;\theta)$ is the likelihood function, and $U(\theta) := -\log\left(\lambda(\theta)\right) - \sum_{i=1}^n \log\left(p(X_i;\theta)\right)$ is the energy function.

Typical MCMC methods require computations over the whole dataset, which is inefficient in large-scale online learning. To overcome this issue, we adopt SGLD (Welling and Teh, 2011) as one of the approximate sampling methods, which is developed upon stochastic optimization over mini-batch data $D \subseteq \{X_i\}_{i=1}^n$. The update rule is based on the Euler-Murayama discretization of the Langevin *stochastic differential equation (SDE)*:

$$d\theta_t = \frac{1}{2}\left(\nabla\log\left(\lambda(\theta_0)\right) + \frac{n}{|D|}\sum_{i\in D}\nabla\log\left(p(x_i;\theta_t)\right)\right)dt + \sqrt{2}dB_t,$$

where B_t is a Brownian motion. To further improve computation, we reuse samples from previous batches to warm start the Markov chains (Algorithm 1). The resulting dependent structure in samples will complicate our analysis.

A.2. Constrained Approximate Sampling

While the convergence of SGLD methods is well-studied, it is only applicable to unconstrained settings. To enable sampling from constrained non log-concave distributions, such as probability simplex in transition dynamics of MDPs, reparameterization can be used in conjunction with SGLD. Alternatively, one can adopt MLD (Hsieh et al., 2018) which utilizes mirror maps for sampling from a dual unconstrained space (Algorithm 4). Let the probability measure of θ be $d\rho = e^{-U(\theta)}\mathrm{d}\theta$, where $\mathrm{dom}(U)$ is constrained. Suppose there exist a mirror map h that maps ρ to some unconstrained distribution $d\nu = e^{-W(\omega)}\mathrm{d}\omega$, denoted by $\nabla h\#\rho = \nu$. Then MLD has the following SDE:

$$\begin{cases} \mathrm{d}\omega_t = -(\nabla W \circ \nabla h)(\theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \\ \theta_t = \nabla h^*(\omega_t) \end{cases}, \tag{5}$$

where h^* is the dual of h, and $(\nabla h)^{-1} = \nabla h^*$.

Algorithm 4 Mirrored Langevin Dynamics (MLD)

Input: $\mathcal{S}, \mathcal{A}$, mirror map h, observed transitions $\{X_s\}_{s=1}^n$, total iterations N
for $i = 1,\ldots,N$ **do**
  Subsample $D \subseteq \{X_s\}_{s=1}^n$
  Sample $\omega_{i+1} \sim \nabla h\#e^{-U}$ from the unconstrained dual space
  Compute constrained sample $\theta_{i+1} = \nabla h^*(\omega_{i+1})$
Output: θ_N

In tabular settings of MDP, MLD needs to be run against each row of the $|\mathcal{A}| \times |\mathcal{S}|$ matrices to generate a sampled transition from simplex $\Delta_{|\mathcal{S}|}$ for each state-action pair. In this case, entropic mirror map will be adopted as h, which is given by

$$h(\theta) = \sum_{i=1}^{|\mathcal{S}|}\theta_i\log\theta_i + (1-\sum_{i=1}^{|\mathcal{S}|}\theta_i)\log(1-\sum_{i=1}^{|\mathcal{S}|}\theta_i), \quad \text{where } 0\log 0 := 0. \tag{6}$$

B. Assumptions

Here we explicitly mention all of the assumptions required in the paper. Assumptions 1-4 are required for SGLD to converge (Algorithm 1 and Theorem 1), Assumptions 5 and 6 are required in Section 6.

Assumption 1 (Assumption on the family $p(S|\theta)$ for approximate sampling). Assume that $\log p(s|\theta)$ is L-smooth and m-strongly concave over θ:

$$-\log p(s|\theta') - \nabla_\theta \log p(s|\theta')^\top(\theta - \theta') + \frac{m}{2}\|\theta - \theta'\|^2 \le -\log p(s|\theta)$$
$$\le -\log p(s|\theta') - \nabla_\theta \log p(s|\theta')^\top(\theta - \theta') + \frac{L}{2}\|\theta - \theta'\|^2 \quad \forall \theta, \theta' \in \mathbb{R}^d, s \in \mathcal{S}$$

Assumption 2 (Assumption on true reward/transition distribution $p(S|\theta^*)$). Assume that $p(S;\theta^*)$ is strongly log-concave in S with some parameter ν, and that $\nabla_\theta \log p(s|\theta^*)$ is L-Lipschitz in S:

$$-(\nabla_s \log p(s|\theta^*) - \nabla_s \log p(s'|\theta^*))^\top(s - s') \ge \nu \|s - s'\|^2, \quad \forall s, s' \in \mathbb{R}$$

$$\|\nabla_\theta \log p(s|\theta^*) - \nabla_\theta \log p(s'|\theta^*)\| \le L\|s - s'\|, \quad \forall s, s' \in \mathbb{R}$$

Assumption 3 (Assumption on the prior distribution). Assume that $\log \lambda(\theta)$ is concave with L-Lipschitz gradients for all $\theta \in \mathbb{R}^d$:

$$\|\nabla_\theta \lambda(\theta) - \nabla_\theta \lambda(\theta')\| \le L\|\theta - \theta'\|, \quad \forall \theta, \theta' \in \mathbb{R}^d$$

Assumption 4 (Joint Lipschitz smoothness of $\log p(S|\theta)$).

$$\|\nabla_\theta \log p(s|\theta) - \nabla_\theta \log p(s'|\theta)\| \le L\|\theta - \theta'\| + L\|s - s'\|, \quad \forall \theta, \theta' \in \mathbb{R}^d, s, s' \in \mathbb{R}$$

Assumption 5 (1- Lipschitzness of $J(\theta)$ in θ). The optimal average-reward function J satisfies

$$\|J(\theta) - J(\theta')\| \le \|\theta - \theta'\|, \quad \forall \theta, \theta' \in \mathbb{R}^d$$

where $J(\theta) = \max_\pi J^\pi(\theta)$.

Assumption 6 (Lipschitzness of transition in θ for RL). There exists a constant L_p such that the transition for each state-action pair is L_p-Lipschtiz in parameter space:

$$\|p(\cdot|s,a,\theta) - p(\cdot|s,a,\theta')\| \le L_p\|\theta - \theta'\|, \quad \forall \theta, \theta' \in \mathbb{R}^d, \quad s, a \in \mathcal{S} \times \mathcal{A}$$

C. Convergence of SGLD with Batched Data

In this section, we prove the convergence of SGLD in sequential decision making frameworks under the batch scheme, which is stated with precise hyperparameters as Theorem 7. We first state the supporting lemmas, followed by the proof of the convergence theorem.

Lemma C.4 (Lemma 5 in (Mazumdar et al., 2020)). *Denote $\widehat{U}$ as the stochastic estimator of U. Then for stochastic gradient estimate with k data points, we have,*

$$\mathbb{E}\left[\left\|\nabla\widehat{U}(\theta) - \nabla U(\theta)\right\|^p |\theta\right] \le 2\frac{n^{p/2}}{k^{p/2}}\left(\frac{\sqrt{dp}L_a^*}{\sqrt{\nu_a}}\right)^p.$$

Lemma C.5 (Lemma 6 from (Mazumdar et al., 2020)). *For a fixed arm a with n samples, suppose we run Algorithm 1 with step size $\eta \leq \frac{\hat{m}}{32\hat{L}^2}$ for N iterations to generate samples from posterior $\rho_n^* \propto \exp(-U)$, in which U is $\hat{m}-$strongly convex and $\hat{L}-$Lipschitz smooth. If at each step $i \in [N]$, the p-th moment between the true gradient and the stochastic gradient satisfies $\mathbb{E}\left[\|\nabla U(\theta_{i\eta}) - \nabla \hat{U}(\theta_{i\eta})\|^p \mid \theta_{i\eta}\right] \leq \Delta_p$, then:*

$$W_p^p(\tilde{\rho}_{i\eta,n}, \rho_n^*) \leq \left(1 - \frac{\hat{m}}{8}\eta\right)^{pi} W_p^p(\rho_0, \rho_n^*) + 2^{5p}\frac{\hat{L}^p}{\hat{m}^p}(dp)^{p/2}(\eta)^{p/2} + 2^{2p+3}\frac{\Delta_p}{\hat{m}^p}$$

where $\rho_0 = \tilde{\rho}_{0\eta,n}$.

Theorem 7 (SGLD convergence). *Fix an arm $a \in \mathcal{A}$ and suppose that Assumptions 1-4 are met for it. Let $\kappa := \max\{L/m, L/\nu\}$, n_k be the number of available rewards for arm a when running SGLD for the k-th time, ρ_{a,n_k} be the exact posterior of arm a after observing n_k samples, and $\tilde{\rho}_{a,n_k}$ be the corresponding approximate posterior obtained by SGLD. If $\mathbb{E}_{\theta\sim\rho_{a,n_k}}[\|\theta - \theta^*\|^p]^{1/p} \leq \frac{\tilde{D}}{\sqrt{n_k}}$ is satisfied by the posterior, then with mini-batch size $s = \frac{32L^2}{m\nu} = \mathcal{O}(\kappa^2)$, step size $\eta = \frac{mn_k}{32L^2(n_k+1)^2} = \mathcal{O}(\frac{1}{L\kappa n_k})$, and the number of steps $N = \frac{1280L^2(n_k+1)^2}{m^2 n_k^2} = \mathcal{O}(\kappa^2)$, SGLD in Algorithm 1 converges in Wasserstein-p distance:*

$$W_p\left(\tilde{\rho}_{a,n_k}, \rho_{a,n_k}\right) \leq \frac{2\tilde{D}}{\sqrt{n_k}}, \quad \forall \tilde{D} \geq \sqrt{\frac{32dp}{m}},\ p \geq 2.$$

Proof of Theorem 7 The proof follows similarly to that of Theorem 6 in (Mazumdar et al., 2020). Compared to the analysis in (Mazumdar et al., 2020), our proof is based on induction on the batches, as opposed to induction on the number of samples, as for us, SGLD is only executed at the end of the batch. Let B_k be the k-th batch. Now for the base case, i.e. when $k = 1$, we have that $n_k = 1$. And therefore the claim follows by the initialization of the algorithm (this is similar to the fully sequential case in (Mazumdar et al., 2020)).

Now, suppose that the claim holds for batch $k-1$. That is, suppose that all the necessary conditions are met and that $W_p\left(\tilde{\rho}_{a,n_{k-1}}, \rho_{a,n_{k-1}}\right) \leq \frac{2\tilde{D}}{\sqrt{n_{k-1}}}$.

Taking the initial condition $\rho_0 = \tilde{\rho}_{a,n_{k-1}}$ in Lemma C.5, we get that:

$$W_p^p(\tilde{\rho}_{i\eta,n_k}, \rho_{n_k}^*) \leq \left(1 - \frac{\hat{m}}{8}\eta\right)^{pi} W_p^p(\tilde{\rho}_{a,n_{k-1}}, \rho_{n_k}^*) + 2^{5p}\frac{\hat{L}^p}{\hat{m}^p}(dp)^{p/2}(\eta)^{p/2} + 2^{2p+3}\frac{\Delta_p}{\hat{m}^p}.$$

Now we know that:

$$\begin{aligned} W_p(\rho_{n_k}^*, \tilde{\rho}_{a,n_{k-1}}) &\leq W_p(\rho_{n_k}^*, \rho_{n_{k-1}}^*) + W_p(\rho_{n_{k-1}}^*, \tilde{\rho}_{a,n_{k-1}}) \\ &\leq \frac{\tilde{D}}{\sqrt{n_k}} + \frac{\tilde{D}}{\sqrt{n_{k-1}}} + \frac{2\tilde{D}}{\sqrt{n_{k-1}}} \\ &\leq \frac{8\tilde{D}}{\sqrt{n_k}} \end{aligned}$$

where the first inequality follows from triangle inequality, the second one follows from the assumption on the posterior and the induction hypothesis, and the last one just upper bounds the expression while also using the fact that $n_k \leq 2n_{k-1}$. This shows us that we can get the same upper bound as is seen in the fully sequential proof. The main point to note is that the proof has enough looseness in it, so that despite collecting at most double the data, the same bounds hold. With the choice of hyperparameters, taking $i = N$ and using Lemma C.4 leads us to the conclusion that $W_p\left(\tilde{\rho}_{a,n_k}, \rho_{a,n_k}\right) \leq \frac{2\tilde{D}}{\sqrt{n_k}}$.

■

We now state the concentration results provided by SGLD in Lemma C.6, which shows the probability that the sampled parameters (from the approximate posterior) are far away from the true dynamics is small. Lemma C.6 extends Lemma 11 in (Mazumdar et al., 2020) to the batched settings.

Lemma C.6 (Concentration of SGLD in bandits). *For a fixed arm $a \in \mathcal{A}$, say that it is pulled n_{k-1} times till batch $k-1$ and n_k times till batch k (where $n_k \leq 2n_{k-1}$). Suppose that Assumptions 1-4 are satisfied, then for $\delta \in (0,1)$, with parameters as specified in Theorem 7, the sampled parameter θ_a^k generated in the k-th batch satisfies,*

$$\mathbb{P}_{\theta_a^k \sim \tilde{\rho}_{a,n_k}[\gamma]}\left(\|\theta_a^k - \theta_a^*\|_2 > \sqrt{\frac{36e}{n_k m}\left(d + \log Q_a + 2\sigma \log 1/\delta + 2(\sigma + \frac{md}{18L\gamma}) \log 1/\delta\right)} \;\middle|\; Z_{k-1} \right) < \delta,$$

where $Z_{k-1} = \{ \left\|\theta_a^{k-1} - \theta_a^*\right\|_2 \leq C(n_k) \}$, $C(n_k) = \sqrt{\frac{18e}{n_k m}}(d + \log Q_a + 2\sigma \log 1/\delta)^{0.5}$, $\sigma = 16 + \frac{4dL^2}{\nu m}$.

Proof of Lemma C.6 The proof follows exactly as Lemma 11 from (Mazumdar et al., 2020) by replacing the notations in fully-sequential settings by those in batched settings, i.e., $\theta_{a,t}$ by θ_a^k, $\theta_{a,t-1}$ by θ_a^{k-1}.

■

D. Proofs of Langevin Thompson Sampling in Multi-armed Bandits

In this section, we provide the regret proofs of BLTS algorithm in the stochastic multi-armed bandit (MAB) setting, which are discussed in Section 5. In particular, we discuss the information exchange guarantees under dynamic batching scheme and its communication cost. We then utilize the convergence of SGLD in Appendix C and the above results to prove the problem-dependent regret bound in MAB setting.

D.1. Notations

We first introduce the notation being used in this section, which is summarized in Table 4.

Symbol	Meaning
$\mathcal{A}$	set of arms in bandit environment
N	number of arms in bandit environment, i.e., $\lvert\mathcal{A}\rvert$
T	time horizon
K	total number of batches
$B(t)$	starting time of the batch containing timestep t
B_k	starting time of the k-th batch
l_a	trigger of dynamic batches (a batch is formed when $k_a(t) = 2^{l_a}$), a monotonically-increasing integer for arm a
$k_a(t)$	the number of times that arm a has been pulled up to time t
$p_a(r\lvert\theta_a)$	reward distribution of arm a parameterized by $\theta_a \in \mathbb{R}^d$
θ_a	parameter of reward distribution for arm $a \in \mathcal{A}$
μ_a	expected reward of arm a, $\mu_a := \mathbb{E}[r_a\lvert\theta_a^*]$
$\hat{\mu}_a$	estimated expected reward of arm a, $\hat{\mu}_a := \mathbb{E}[\lVert\theta_a\rVert]$
Q_a	quality of prior for arm a, $Q_a := \max_\theta \frac{p_a(\theta)}{p_a(\theta_a^*)}$
κ	condition number of parameterized reward distribution, $\kappa := \max\{L/m, L/\nu\}$
$\lambda_a(\theta_a)$	prior distribution over $\theta_a \in \mathbb{R}^d$
U	energy function of posterior distribution ρ : $\rho \propto e^{-U}$
L	Lipschitz constant of the true reward distribution and likelihood families $p_a(r\lvert\theta^*)$ in r
m	strong log-concavity parameter of $p_a(r;\theta)$ in θ for all r
ν	strong log-concavity parameter of $p_a(r;\theta)$ in r

Table 4: Notations in multi-armed bandit setting.

D.2. Communication cost of Dynamic Doubling Batching Scheme

In batched setting, striking a balance between batch size and the number of batches is critical to achieving optimal performance. More specifically, it is crucial to balance the number of actions taken within each batch, with the frequency of starting new batches to collect new data and update the posteriors. According to Lemma D.1, dynamic doubling batching scheme guarantees an arm that has been pulled k times has at least $k/2$ observed rewards, indicating that communication between the learner and the environment is sufficient under this batching scheme.

Lemma D.1. *Let t be the current time step, $B(t)$ be the starting time of the current batch, $k_a(t)$ be the number of times that arm a has been pulled up to time t. For all $a \in \mathcal{A}$, the dynamic batch scheme ensures:*

$$\frac{1}{2}k_a(t) \leq k_a(B(t)) \leq k_a(t).$$

Proof of Lemma D.1 By the mechanism of our batch scheme, a new batch will begin when the number of times of any arm $a \in \mathcal{A}$ being pulled is doubled. It implies that the number of times that an arm is pulled within a batch is less than the number of times that it has been pulled at the beginning of this batch. At any time step $t \leq T$:

$$k_a(t) - k_a(B(t)) \leq k_a(B(t)),$$

which gives $\frac{1}{2}k_a(t) \leq k_a(B(t))$. On the other hand, $k_a(B(t)) \leq k_a(t)$ holds due to the fact that $B(t) \leq t$.

■

Next, we show that by employing the dynamic doubling batching scheme, BATS algorithm achieves optimal performance using only logarithmic rounds of communication (measured in terms of batches).

Theorem 2. *BLTS ensures that the total number of batches is at most $O(N \log T)$ where $N = |\mathcal{A}|$.*

Proof of Theorem 2 Denote by B_k the starting time of the k-th batch, and let $l_a(B_k)$ be the trigger integer for arm a at time B_k, K be the total number of rounds to interact with environment, namely, batches. Then for each arm $a \in \mathcal{A}$, $k_a(T) \leq T$, and

$$k_a(T) = \sum_{k=1}^{K-1} k_a(B_{k+1}) - k_a(B_k) \leq \sum_{k=1}^{K-1} k_a(B_k) = \sum_{k=1}^{K-1} 2^{l_a(B_k)-1} \leq \sum_{l=0}^{K-1} 2^l,$$

where the second and third step result from the dynamic batching scheme. Thus for each arm a, we have

$$K \leq \log(T+1).$$

The proof is then completed by multiplying the above result by N arms. ■

D.3. Regret Proofs in Multi-armed Bandit

With the convergence properties shown in Appendix C, we proceed to prove the regret guarantee of Langevin TS with SGLD. The general idea of our regret proof is to upper bound the total number of times that the sub-optimal arms are pulled over time horizon T. We remark that the dependence of approximate samples across batches complicates our analysis of TS compared to the existing analyses in bandit literature.

We first decompose the expected regret according to the events of concentration in approximate samples $\theta_{a,t}$ and the events of estimation accuracy in expected rewards of sub-optimal arms.

For approximate samples θ, define event $E_{\theta,a}(B_k) = \{\|\theta_{a,k} - \theta_a^*\| < C(n_k)\}$, which is guaranteed to happen with probability at least $(1-\delta_2)$ by Lemma C.6 for some $\delta_2 \in [0,1]$. Let $E_{\theta,a}(T) = \bigcap_{t=1}^{T} E_{\theta,a}(t)$, $E_{\theta,a}(K) = \bigcap_{k=1}^{K} E_{\theta,a}(B_k)$, where K is the total number of batches. Without loss of generality, we take $\|\alpha_a\| = 1$ for all arms in $\mathbb{E}_{X\sim p_a(X|\theta_a)}[X] = \alpha_a^\intercal\, \theta_a \leq \|\theta_a\|$ in the subsequent proofs

Let $\hat{\mu}_a(t)$ be the estimate of the expected reward for arm a at time step t, and denote the filtration up to time $B(t)$ as $\mathcal{F}_{B(t)} := \{a(\tau), r_{a(\tau),k_a(B(\tau))} \mid \tau \leq B(t)\}$. For any sub-optimal arm $a \neq 1$, define event $E_{\mu,a}(t) = \{\hat{\mu}_a(t) \geq \mu_1 - \epsilon\}$ with probability $p_{a,k_a(B(t))}(t) := \mathbb{P}(\hat{\mu}_a(t) \geq \mu_1 - \epsilon | \mathcal{F}_{B(t)})$ for some $\epsilon \in (0,1)$, which signifies the estimation of arm a is close to the true optimal expected reward.

Lemma D.2 (Regret Decomposition). *Let μ_a be the true expected reward of arm a, $\mu^* = \max_{a\in\mathcal{A}} \mu_a$, $\Delta_a := \mu^* - \mu_a$. The expected regret of Langevin TS with SGLD satisfies:*

$$R_T \leq \sum_{a\in\mathcal{A}} \Big(R_1 + R_2 + 2\Big)\Delta_a,$$

where $R_1 := \mathbb{E}\left[\sum_{t=1}^T \mathbb{I}(a(t)=a, E^c_{\mu,a}(t)) \mid E_{\theta,a}(K)\cap E_{\theta,1}(K)\right]$, $R_2 := \mathbb{E}\left[\sum_{t=1}^T \mathbb{I}(a(t)=a, E_{\mu,a}(t)) \mid E_{\theta,a}(K)\cap E_{\theta,1}(K)\right]$.

Proof of Lemma D.2. Recall that

$$R_T = \sum_{a\in\mathcal{A}} \Delta_a \cdot \mathbb{E}\left[k_a(T)\right], \quad \Delta_a = \mu^* - \mu_a.$$

For any sub-optimal arm $a \neq 1$, consider the event space $\mathcal{F}_\theta = \{\{E_{\theta,a}(T)\cap E_{\theta,1}(T)\}, \{E_{\theta,a}(T)\cap E_{\theta,1}(T)\}^C\}$, in which $E_{\theta,a}(T)\cap E_{\theta,1}(T)$ denotes the event that all approximate samples of arm a and optimal arm 1 are concentrated.

To bound the regret, we bound the largest number of times that each sub-optimal arm will be played:

$$\begin{aligned}
\mathbb{E}\left[k_a(T)\right] &= \mathbb{E}\left[\sum_{t=1}^T \mathbb{I}(a(t)=a, E_{\mu,a}(t)\cup E^c_{\mu,a}(t), E_{\theta,a}(T)\cap E_{\theta,1}(T))\right] + \mathbb{E}\left[\sum_{t=1}^T \mathbb{I}(a(t)=a, (E_{\theta,a}(T)\cap E_{\theta,1}(T))^c)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^T \mathbb{I}(a(t)=a, E_{\mu,a}(t)\cup E^c_{\mu,a}(t)) \mid E_{\theta,a}(T)\cap E_{\theta,1}(T)\right] + \mathbb{E}\left[\sum_{t=1}^T \mathbb{I}(a(t)=a, (E_{\theta,a}(T)\cap E_{\theta,1}(T))^c)\right].
\end{aligned}$$

where the second inequality results from $P(E_{\theta,a}(T)\cap E_{\theta,1}(T)) \leq 1$.

For any arm $a \in \mathcal{A}$ in each batch, approximate samples are independently generated from the identical approximate distribution $\tilde{\rho}_a(\theta_a|R_a)$. Thus, approximate samples for arm a are independent within the same batch, while being dependent across different batches, implying

$$\begin{cases}
\mathbb{P}(E_{\theta,a}(T)) & = \prod_{t=1}^T \mathbb{P}(E_{\theta,a}(t)|E_{\theta,a}(1),\ldots,E_{\theta,a}(t-1)) = \prod_{k=1}^{K-1} \mathbb{P}(E_{\theta,a}(B_{k+1})|E_{\theta,a}(B_k))^{T_{k+1}} \\
\mathbb{P}(E^c_{\theta,a}(T)) & = \mathbb{P}(\bigcup_{t=1}^T E^c_{\theta,a}(t)) = \sum_{t=1}^T \mathbb{P}(E^c_{\theta,a}(t)) = \sum_{k=1}^K T_k \mathbb{P}(E^c_{\theta,a}(B_k))
\end{cases},$$

where $T_k := B_{k+1} - B_k$ is the number of time steps in the k-th batch, namely, the length of the batch. By Lemma C.6, for each arm a in batch B_k, $\mathbb{P}(E^c_{\theta,a}(B_k)) \leq \delta_2, (1-\delta_2) \leq \mathbb{P}(E_{\theta,a}(B_k)) \leq 1$, which gives:

$$\begin{cases}
\mathbb{P}[E_{\theta,a}(T)\cap E_{\theta,1}(T)] \leq \mathbb{P}[E_{\theta,a}(K)\cap E_{\theta,1}(K)] \\
\mathbb{P}[E^c_{\theta,a}(T)\cup E^c_{\theta,1}(T)] \leq \mathbb{P}[E^c_{\theta,a}(T)] + \mathbb{P}[E^c_{\theta,1}(T)] \leq 2\delta_2 \sum_{k=1}^T T_k = 2\delta_2 T
\end{cases}.$$

Setting $\delta_2 = 1/T^2$ gives,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T \mathbb{I}(a(t)=a, (E_{\theta,a}(T)\cap E_{\theta,1}(T))^c)\right] &= \mathbb{E}\left[\sum_{t=1}^T \mathbb{I}(a(t)=a) \mid E_{\theta,a}(T)^c \cup E^c_{\theta,1}(T)\right] \mathbb{P}\left[E^c_{\theta,a}(T)\cup E^c_{\theta,1}(T)\right] \\
&\leq 2\delta_2 T \mathbb{E}\left[k_a(T) \mid E_{\theta,a}(T)^c \cup E_{\theta,1}(T)^c\right] \leq 2\delta_2 T^2 \leq 2.
\end{aligned}$$

Plugging in the results to the definition of regret yields,

$$\begin{aligned}
R(T) &\leq \sum_{a\in\mathcal{A}} \Delta_a \cdot \left(\mathbb{E}\left[\sum_{t=1}^T \mathbb{I}(a(t)=a, E_{\mu,a}(t)\cup E^c_{\mu,a}(t)) \,\middle|\, E_{\theta,a}(T)\cap E_{\theta,1}(T)\right] + 2\right) \\
&\leq \sum_{a\in\mathcal{A}} \Delta_a \cdot \left(\mathbb{E}\left[\sum_{t=1}^T \mathbb{I}(a(t)=a, E_{\mu,a}(t)\cup E^c_{\mu,a}(t)) \,\middle|\, E_{\theta,a}(K)\cap E_{\theta,1}(K)\right] + 2\right) \\
&\leq \sum_{a\in\mathcal{A}} \Delta_a \cdot \Bigg(\underbrace{\mathbb{E}\left[\sum_{t=1}^T \mathbb{I}(a(t)=a, E_{\mu,a}(t)) \,\middle|\, E_{\theta,a}(K)\cap E_{\theta,1}(K)\right]}_{R_1} + \underbrace{\mathbb{E}\left[\sum_{t=1}^T \mathbb{I}(a(t)=a, E^c_{\mu,a}(t)) \,\middle|\, E_{\theta,a}(K)\cap E_{\theta,1}(K)\right]}_{R_2} + 2\Bigg).
\end{aligned}$$

■

We then proceed to bound R_1 and R_2 respectively in Lemma D.3 and Lemma D.4, the key to maintaining optimal regret is to maximize the probability of pulling the optimal arm by ensuring the event $E_{\mu,a}(t)$ takes place with low probability for all sub-optimal arms.

Lemma D.3 (Bound term R_1)**.** *It can be shown that,*

$$R_1 \leq \mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{1}{p_{1,k_1(B(t))}(t)}-1\right)\,\middle|\, E_{\theta,1}(K)\right].$$

Proof of lemma D.3. Note that arm a is played at time t if and only if $\hat{\mu}_{a'} \leq \hat{\mu}_a,\ \forall a' \in \mathcal{A}$. Thus for a sub-optimal arm a, the following event relationship holds: $\{a(t)=a, E^c_{\mu,a}(t)\} = \{a(t)=a, E^c_{\mu,a}(t), \cap_{a'\neq a} E^c_{\mu,a'}(t)\} \subseteq \{\cap_{a'\in\mathcal{A}} E^c_{\mu,a'}(t)\}$, and $\{E_{\mu,1}(t), \cap_{a'\neq 1} E^c_{\mu,a'}(t)\} = \{a(t)=1, E_{\mu,1}(t), \cap_{a'\neq 1} E^c_{\mu,a'}(t)\} \subseteq \{a(t)=1\}$. We then have,

$$\begin{cases} \mathbb{P}\left[a(t)=a, E^c_{\mu,a}(t) \mid \mathcal{F}_{B(t)}\right] \leq \mathbb{P}\left[\bigcap_{a'\in\mathcal{A}} E^c_{\mu,a'}(t) \mid \mathcal{F}_{B(t)}\right] = \mathbb{P}\left[\bigcap_{a'\neq 1} E^c_{\mu,a'}(t) \mid \mathcal{F}_{B(t)}\right]\left(1-\mathbb{P}\left[E_{\mu,1}(t) \mid \mathcal{F}_{B(t)}\right]\right) \\ \mathbb{P}\left[a(t)=1 \mid \mathcal{F}_{B(t)}\right] \geq \mathbb{P}\left[E_{\mu,1}(t) \mid \mathcal{F}_{B(t)}\right]\mathbb{P}\left[\bigcap_{a'\neq 1} E^c_{\mu,a'}(t) \mid \mathcal{F}_{B(t)}\right] \end{cases}$$

Recall that $p_{1,k_1(B(t))}(t) := \mathbb{P}[E_{\mu,1}(t) \mid \mathcal{F}_{B(t)}]$. Combining the above two equations shows that the probability of pulling a sub-optimal arm a is bounded by the probability of pulling the optimal arm with an exponentially decaying coefficient:

$$\mathbb{P}\left[a(t)=a, E^c_{\mu,a}(t) \mid \mathcal{F}_{B(t)}\right] \leq \left(\frac{1}{p_{1,k_1(B(t))}(t)}-1\right)\mathbb{P}\left[a(t)=1 \mid \mathcal{F}_{B(t)}\right]. \tag{7}$$

Therefore, R_1 is upper bounded accordingly:

$$\begin{aligned} R_1 &= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{I}(a(t)=a, E^c_{\mu,a}(t) \mid \mathcal{F}_{B(t)}\right] \,\middle|\, E_{\theta,a}(K)\bigcap E_{\theta,1}(K)\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{P}\left[a(t)=a, E^c_{\mu,a}(t) \mid \mathcal{F}_{B(t)}\right] \,\middle|\, E_{\theta,a}(K)\bigcap E_{\theta,1}(K)\right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{1}{p_{1,k_1(B(t))}(t)}-1\right)\mathbb{P}\left[a(t)=1 \mid \mathcal{F}_{B(t)}\right] \,\middle|\, E_{\theta,a}(K)\bigcap E_{\theta,1}(K)\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{1}{p_{1,k_1(B(t))}(t)}-1\right)\mathbb{I}\left[a(t)=1\right] \,\middle|\, E_{\theta,1}(K)\right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{1}{p_{1,k_1(B(t))}(t)}-1\right) \,\middle|\, E_{\theta,1}(K)\right]. \end{aligned}$$

■

Lemma D.4 (Bound term R_2)**.** *It can be shown that,*

$$R_2 \leq 1 + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left(p_{a,k_a(B(t))}(t) > \frac{1}{T}\right) \,\middle|\, E_{\theta,a}(K)\right].$$

Proof of Lemma D.4. The proof closely follows (Agrawal and Goyal, 2012). Let $\mathcal{T} := \{t \mid p_{a,k_a(B(t))}(t) > \frac{1}{T}\}$. R_2 term can be rewritten as:

$$\begin{aligned} R_2 &= \mathbb{E}\left[\sum_{t\in\mathcal{T}}\mathbb{I}(a(t)=a, E_{\mu,a}(t)) \,\middle|\, E_{\theta,a}(K)\bigcap E_{\theta,1}(K)\right] + \mathbb{E}\left[\sum_{t\notin\mathcal{T}}\mathbb{I}(a(t)=a, E_{\mu,a}(t)) \,\middle|\, E_{\theta,a}(K)\bigcap E_{\theta,1}(K)\right] \\ &\leq \underbrace{\mathbb{E}\left[\sum_{t\in\mathcal{T}}\mathbb{I}(a(t)=a) \,\middle|\, E_{\theta,a}(K)\bigcap E_{\theta,1}(K)\right]}_{I} + \underbrace{\mathbb{E}\left[\sum_{t\notin\mathcal{T}}\mathbb{I}(E_{\mu,a}(t)) \,\middle|\, E_{\theta,a}(K)\bigcap E_{\theta,1}(K)\right]}_{II}. \end{aligned}$$

It follows that the first term satisfies,

$$I = \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}(a(t)=a, p_{a,k_a(B(t))}(t) > \frac{1}{T}) \,\middle|\, E_{\theta,a}(K)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}(p_{a,k_a(B(t))}(t) > \frac{1}{T}) \,\middle|\, E_{\theta,a}(K)\right],$$

and the second term satisfies,

$$II = \mathbb{E}\left[\sum_{t\notin\mathcal{T}}\mathbb{E}\Big[\mathbb{I}(E_{\mu,a}(t)) \mid \mathcal{F}_{B(t)}\Big] \,\middle|\, E_{\theta,a}(K)\bigcap E_{\theta,1}(K)\right] = \mathbb{E}\left[\sum_{t\notin\mathcal{T}} p_{a,k_a(B(t))}(t) \,\middle|\, E_{\theta,a}(T)\bigcap E_{\theta,1}(T)\right] \leq 1,$$

where the last inequality holds as $p_{a,k_a(B(t))}(t) \leq 1/T$ for $t \notin \mathcal{T}$. ■

Lemma D.5. *Assume that the prior and reward distributions satisfy Assumptions 1-4. Then at each time step $t \leq T$, if there are $k_1(B(t))$ observed rewards for arm 1, then Algorithm 1 ensures:*

$$\mathbb{E}\left[\frac{1}{p_{1,k_1(B(t))}(t)}\right] \leq 36\sqrt{Q_1},$$

where $Q_1 = \max_{\theta\in\mathbb{R}^d}\frac{p_1(\theta)}{p_1(\theta_1^*)}$ measures the quality of the prior distribution, $Q_1 \geq 1$.

Proof of Lemma D.5 For completeness, we provide the proof of this lemma, which closely follows the proof of Lemma 18 in (Mazumdar et al., 2020).

For each arm a, upon running SGLD with batched data in batch k, by Cauchy-Schwartz inequality, we have,

$$\mathbb{P}\left(\alpha_a^{\mathrm{T}}(\theta_a^k - \theta_{a,N\eta}) \geq \alpha_1^{\mathrm{T}}(\theta_a^* - \theta_{a,N\eta}) - \epsilon\right) \geq \mathbb{P}\left(Z \geq \|\theta_a^* - \theta_{a,N\eta}\|\right),$$

where $Z \sim \mathcal{N}(0, \frac{1}{nL\gamma}I)$. Let $\sigma^2 = \frac{1}{nL\gamma}I$, by anti-concentration of Gaussian random variables, for the optimal arm 1,

$$p_{1,k_1(B(t))}(t) \geq \sqrt{\frac{1}{2\pi}}\begin{cases}\frac{\sigma t}{t^2+\sigma^2}e^{-\frac{t^2}{2\sigma^2}}, & t > \sigma;\\ 0.34, & \text{otherwise.}\end{cases}$$

Taking expectations of both sides and by Cauchy-Schwartz inequality,

$$\mathbb{E}\left[\frac{1}{p_{1,k_1(B(t))}(t)}\right] \leq 3\sqrt{2\pi} + \sqrt{2\pi nL\gamma}\sqrt{\mathbb{E}\left[\|\theta_1^* - \theta_{1,Nh}\|^2\right]}\sqrt{\mathbb{E}\left[e^{nL\gamma\|\theta_1^*-\theta_{1,Nh}\|^2}\right]} + \sqrt{2\pi}\mathbb{E}\left[e^{\frac{nL\gamma}{2}\|\theta_1^*-\theta_{1,Nh}\|^2}\right].$$

By the convergence guarantee of SGLD in Theorem 7,

$$\mathbb{E}\left[\|\theta_1^* - \theta_{1,Nh}\|^2\right] \leq \frac{18}{mn}\left(d + \log Q + 32 + \frac{8dL^2}{\nu m}\right).$$

Note that $\|\theta_1^* - \theta_{1,Nh}\|^2$ is a sub-Gaussian random variable, when $\gamma \leq \frac{m}{32L\sigma}$,

$$\mathbb{E}[e^{nL\gamma\|\theta_1^*-\theta_{1,Nh}\|^2}] \leq 3/2\left(e^{\frac{4nL\gamma D}{m}} + 2.5\right).$$

Combining the above results together completes the proof. ■

With Lemma D.5, we now proceed to prove the terms in R_1 and R_2 that lead us to the final regret bound.

Lemma D.6. *Assume that Assumptions 1-4 are satisfied. Let $\sigma = 16 + \frac{4dL^2}{\nu m}$, $\gamma = \frac{m}{32L\sigma}$. running Algorithm 2 with samples generated from approximate posteriors using Algorithm 1, we have,*

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{1}{p_{1,k_1(B(t))}(t)} - 1\right) \,\middle|\, E_{\theta,1}(K)\right] \leq \frac{20736e}{m\Delta_a^2}\sqrt{Q_1}\big(d + \log Q_1 + 4\sigma\log T + 12d\sigma\log 2\big) + 1. \tag{8}$$

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left(p_{a,k_a(B(t))}(t) > \frac{1}{T}\right) \,\middle|\, E_{\theta,a}(K)\right] \leq \frac{576e}{m\Delta_a^2}\big(d + \log Q_a + 10d\sigma\log(T)\big). \tag{9}$$

Proof of lemma D.6. For ease of notation, let the number of observed rewards in batch k for arm $a \in \mathcal{A}$ be n_k. By definition,

$$p_{1,n_k} = \mathbb{P}\big(\hat{\mu}_1(t) \geq \mu_1 - \epsilon \,\big|\mathcal{F}_{B(t)}\big) \geq 1 - \mathbb{P}\big(\,\|\theta_1 - \theta_1^*\| > \epsilon \,\big|\mathcal{F}_{B(t)}\big)$$

With concentration property of approximate samples in Lemma C.6, it suggests the increasing number of observed rewards for the optimal arm leads to the increasing probability of being optimal. Thus by Lemma D.1, at any time step $t \leq T$,

$$p_{1,k_1(B(t))}(t) \geq p_{1,\frac{k_1(t)}{2}}(t).$$

Concentration is achieved only when sufficient number of rewards is observed, we thus require:

$$\mathbb{P}_{\theta_1 \sim \tilde{\rho}_{1,n_k}[\gamma]}\left(\|\theta_1 - \theta_1^*\| \geq \epsilon\right) \leq \exp\left(-\frac{1}{6d\sigma}\left(\frac{mn_k\epsilon^2}{36e} - \bar{D}_1\right)\right), \tag{10}$$

where $\bar{D}_1 = d + \log Q_1 + 4\sigma \log T, \sigma = 16 + \frac{4dL^2}{\nu m}$. Choose $\epsilon = (\mu_1 - \mu_a)/2 = \Delta_a/2$, and consider the time step t when arm 1 satisfies:

$$k_1(t) = 2^{\lceil \log_2 2l \rceil}, \qquad where \quad l = \frac{144e}{m\Delta_a^2}\big(\bar{D}_1 + 6d\sigma \log 2\big).$$

As $k_1(t) \geq 2l$, the number of observed rewards is guaranteed to be at least $\frac{36e}{m\epsilon^2}\bar{D}$, and $\mathbb{P}_{\theta_1 \sim \tilde{\rho}_{1,n_k}[\gamma]}\left(\|\theta_1 - \theta_1^*\| \geq \epsilon\right) \leq 1/2$. Thus, the individual term in R_1 follows:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{1}{p_{1,k_1(B(t))}(t)} - 1\right) \,\middle|\, E_{\theta,1}(K)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{1}{p_{1,\frac{k_1(t)}{2}}(t)} - 1\right) \,\middle|\, E_{\theta,1}(K)\right] \\
&\leq \mathbb{E}\left[\sum_{k_1(t)=0}^{T-1}\left(\frac{1}{p_{1,\frac{k_1(t)}{2}}(t)} - 1\right) \,\middle|\, E_{\theta,1}(K)\right] \\
&\leq \mathbb{E}\left[\sum_{k_1(t)=0}^{2^{\lceil \log_2 2l \rceil}}\left(\frac{1}{p_{1,\frac{k_1(t)}{2}}(t)} - 1\right) \,\middle|\, E_{\theta,1}(K)\right] + \mathbb{E}\left[\sum_{k_1(t)=2^{\lceil \log_2 2l \rceil}+1}^{T-1}\left(\frac{1}{p_{1,\frac{k_1(t)}{2}}(t)} - 1\right) \,\middle|\, E_{\theta,1}(K)\right].
\end{aligned} \tag{11}$$

In early stage when concentration has not been achieved, using results from Lemma D.5,

$$\mathbb{E}\left[\sum_{k_1(t)=0}^{2^{\lceil \log_2 2l \rceil}}\left(\frac{1}{p_{1,\frac{k_1(t)}{2}}(t)} - 1\right) \,\middle|\, E_{\theta,1}(K)\right] \leq 2^{\lceil \log_2 2l \rceil} 36\sqrt{B_1} \leq 2 \cdot 2l \cdot 36\sqrt{B_1}. \tag{12}$$

When sufficient rewards for the optimal arm has been accumulated,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k_1(t)=2^{\lceil \log_2 2l \rceil}+1}^{T-1}\left(\frac{1}{p_{1,\frac{k_1(t)}{2}}(t)} - 1\right) \,\middle|\, E_{\theta,1}(K)\right] \\
&\leq \mathbb{E}\left[\sum_{k_1(t)=0}^{T-1}\left(\frac{1}{p_{1,\frac{k_1(t)}{2}}(t)} - 1\right) \,\middle|\, E_{\theta,1}(K)\right] \\
&\leq \sum_{k_1(t)=0}^{T-1}\frac{1}{\exp\left(-\frac{1}{6d\sigma_1}\left(\frac{m_1\epsilon^2}{36e} \cdot \frac{k_1(t)}{2}\right)\right)} - 1 \\
&\leq \int_{z=0}^{\infty}\left(\frac{1}{\exp\left(-\frac{m\epsilon^2}{432ed\sigma_1}z\right)} - 1\right) dz
\end{aligned}$$

$$\leq 2 \cdot \frac{144e}{m\Delta_a^2} \cdot 6d\sigma \log 2 + 1. \tag{13}$$

Substituting equation (12) and (13) back to (11) yields,

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{1}{p_{1,k_1(B(t))}(t)} - 1\right) \,\middle|\, E_{\theta,1}(K)\right] \leq 4l \cdot 36\sqrt{Q_1} + 2 \cdot \frac{144e}{m\Delta_a^2} \cdot 6d\sigma \log 2 + 1$$
$$\leq 36\sqrt{Q_1}\frac{576e}{m\Delta_a^2}\big(\bar{D}_1 + 12d\sigma \log 2\big) + 1.$$

Similarly, for R_2 term with event $E_{\mu,a}(t) = \{\hat{\mu}_a(t) \geq \mu_1 - \epsilon\}$, let $\epsilon = (\mu_1 - \mu_a)/2 = \Delta_a/2$,

$$\begin{aligned} p_{a,k_a(B(t))}(t) &= \mathbb{P}(\hat{\mu}_a(t) - \mu_a \geq \mu_1 - \mu_a - \epsilon | \mathcal{F}_{B(t)}) \\ &= \mathbb{P}(\hat{\mu}_a(t) - \mu_a \geq \frac{\Delta_a}{2} | \mathcal{F}_{B(t)}) \\ &\leq \mathbb{P}(\hat{\mu}_a(t) - \mu_a \geq \frac{\Delta_a}{2} | \mathcal{F}_{\frac{k_a(t)}{2}}) \\ &= p_{a,\frac{k_a(t)}{2}}(t), \end{aligned}$$

which gives

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left(p_{a,k_a(B(t))}(t) > \frac{1}{T}\right) \,\middle|\, E_{\theta,a}(K)\right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left(\mathbb{P}(\hat{\mu}_a(t) - \mu_a \geq \frac{\Delta_a}{2} | \mathcal{F}_{\frac{k_a(t)}{2}}) > \frac{1}{T}\right) \,\middle|\, E_{\theta,a}(K)\right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left(\mathbb{P}(|\hat{\mu}_a(t) - \mu_a| \geq \frac{\Delta_a}{2} | \mathcal{F}_{\frac{k_a(t)}{2}}) > \frac{1}{T}\right) \,\middle|\, E_{\theta,a}(K)\right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left(\mathbb{P}_{\theta_a \sim \tilde{\rho}_{a,\frac{k_a(t)}{2}}}\big(\|\theta_a - \theta_a^*\| \geq \frac{\Delta_a}{2}\big) > \frac{1}{T}\right) \,\middle|\, E_{\theta,a}(K)\right]. \end{aligned}$$

With the same form of posterior as in equation 10, $\mathbb{P}_{\theta_a \sim \tilde{\rho}_{a,\frac{k_a(t)}{2}}}\big(\|\theta_a - \theta_a^*\| \geq \frac{\Delta_a}{2}\big) \leq \frac{1}{T}$ for arm $a \neq 1$ holds, when

$$k_a(t) > 2 \cdot 2 \cdot \frac{144e}{m\Delta_a^2}\big(\bar{D}_a + 6d\sigma \log(T)\big).$$

Here, the number of observed rewards is guaranteed to be at least $2^{\lceil \log_2 l \rceil}$, where $l = \frac{144e}{m\Delta_a^2}\big(\bar{D}_a + 6d\sigma \log(T)\big)$. Therefore, using the fact that $d > 1$, we have,

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left(p_{a,k_a(B(t))}(t) > \frac{1}{T}\right) \,\middle|\, E_{\theta,a}(K)\right] \leq \frac{576e}{m_a\Delta_a^2}\big(\bar{D}_a + 6d\sigma_a \log(T)\big).$$

■

We are ready to prove the final regret bound by combining results from the above Lemmas.

Theorem 3. *Assume that the parametric reward families, priors, and true reward distributions satisfy Assumptions 1 through 4 for each arm $a \in \mathcal{A}$. Then with the SGLD parameters specified as per Algorithm 1 and with $\gamma = O(1/d\kappa^3)$ (for $\kappa := \max\{L/m, L/\nu\}$), BLTS satisfies:*

$$R(T) \leq \sum_{a>1}\frac{C\sqrt{Q_1}}{m\Delta_a}\left(d + \log Q_1 + d\kappa^2 \log T + d^2\kappa^2\right)$$

$$+\frac{C}{m\Delta_a}\left(d+\log Q_a+d^2\kappa^2\log T\right)+4\Delta_a,$$

where C is a constant and $Q_a := \max_\theta \frac{\lambda_a(\theta)}{\lambda_a(\theta^*)}$. The total number of SGLD iterations used by BLTS is $O(\kappa^2 N log T)$.

Proof of Theorem 3. The proof is a direct result by combining Lemma D.2, D.3, D.4,D.6, which gives,

$$\begin{aligned}
R_T &\le \sum_{a\in\mathcal{A}}\Delta_a\cdot\left(R_1+R_2+2\right)\\
&\le \left(\sum_{a\in\mathcal{A}}4\cdot 36\sqrt{Q_1}\frac{144e}{m\Delta_a}\big(d+\log Q_1+4\left(16+\frac{4dL^2}{m\nu}\right)(\log T+3d\log 2)\big)+\Delta_a\right.\\
&\left.+\Delta_a+4\frac{144e}{m\Delta_a}\big(d+\log Q_a+10d\left(16+\frac{4dL^2}{m\nu}\right)\log(T)\big)+2\Delta_a\right)\\
&\le \sum_{a>1}\frac{C\sqrt{Q_1}}{m\Delta_a}\left(d+\log Q_1+d\kappa^2\log T+d^2\kappa^2\right)+\frac{C}{m\Delta_a}\left(d+\log Q_a+d^2\kappa^2\log T\right)+4\Delta_a.
\end{aligned}$$

■

E. Proofs of Langevin Posterior Sampling for Reinforcement Learning

In this section, we will present the regret proofs for Langevin Posterior Sampling algorithms in RL frameworks under different types of parameterization, and conclude with a real-world example where the General Parameterization from Section 6.2 is applicable.

E.1. Communication cost of Static Doubling Batching Scheme

We first show that under the static doubling batching scheme in RL setting, LPSRL algorithm achieves optimal performance using only logarithmic rounds of communication (measured in terms of batches, or equivalently policy switches).

Theorem 8. *Let T_k be the number of time steps between the $(k-1)$-th policy switch and the k-th policy switch, and K_T be the total number of policy switches for time horizon T. LPSRL ensures that*

$$K_T \le \log T + 1.$$

Proof of Theorem 8. By design of Algorithm 3,at the k-th policy switch, $T_k = 2^{k-1}$. Since the total number of time steps is determined by time horizon T, we can easily obtain $K_T = \lceil \log T \rceil$. ■

E.2. Regret Proofs in Average-reward MDPs

In this section, we proceed to prove the theorems in Section 6. To focus on the problem of model estimation, our results are developed under the optimality of policies[13].

While analyses of Bayes regret in existing works of PSRL crucially depend on the true transition dynamics θ^* being identically distributed as those of sampled MDP (Russo and Van Roy, 2014), we show that in Langevin PSRL, sampling from the approximate posterior instead of the true posterior will introduce a bias that can be upper bounded using Wasserstein-1 distance.

Lemma 3. *(Langevin Posterior Sampling). Let t_k be the beginning time of policy-switch k, $\mathcal{H}_{t_k} := \{s_\tau, a_\tau\}_{\tau=1}^{t_k}$ be the history of observed states and actions till time t_k, and $\theta^k \sim \tilde{\rho}_{t_k}$ be the sampled model from the approximate posterior $\tilde{\rho}_{t_k}$ at time t_k. Then, for any $\sigma(\mathcal{H}_{t_k})$-measurable function f that is 1-Lipschitz, it holds that:*

$$\left|\mathbb{E}[f(\theta^*)|\mathcal{H}_{t_k}] - \mathbb{E}[f(\theta^k)|\mathcal{H}_{t_k}]\right| \le W_1(\tilde{\rho}_{t_k}, \rho_{t_k}). \tag{4}$$

By the tower rule, $\left|\mathbb{E}[f(\theta^*)] - \mathbb{E}[f(\theta^k)]\right| \le W_1(\tilde{\rho}_{t_k}, \rho_{t_k})$.

[13] If only suboptimal policies are available in our setting, it can be shown that small approximation errors in policies only contribute additive non-leading terms to regret. See details in (Ouyang et al., 2017).

Proof of Lemma 3 Notice that both $\tilde{\rho}_{t_k}$ and ρ_{t_k} are measurable with respect to $\sigma(\mathcal{H}_{t_k})$. Therefore, condition on history $\mathcal{H}_{t_k}$, the only randomness under the expectation comes from the sampling procedure for approximate posterior, which gives,

$$\begin{aligned}\mathbb{E}[f(\theta^k)|\mathcal{H}_{t_k}] &= \int_{\mathbb{R}^d} f(\theta)\tilde{\rho}_{t_k}(d\theta) \\ &= \int_{\mathbb{R}^d} f(\theta)(\tilde{\rho}_{t_k} - \rho_{t_k} + \rho_{t_k} - \delta(\theta^*) + \delta(\theta^*))(d\theta) \\ &\leq \mathbb{E}[f(\theta^{k,*})|\mathcal{H}_{t_k}] - \mathbb{E}[f(\theta^*)|\mathcal{H}_{t_k}] + \mathbb{E}[f(\theta^*)|\mathcal{H}_{t_k}] + W_1(\tilde{\rho}_{t_k}, \rho_{t_k}) \\ &= \mathbb{E}[f(\theta^*)|\mathcal{H}_{t_k}] + W_1(\tilde{\rho}_{t_k}, \rho_{t_k}). \end{aligned} \tag{14}$$

The third inequality follows from the fact that given $\mathcal{H}_{t_k}$, ρ_{t_k} is the posterior of $\theta^{k,*}$ and the definition of dual representation for W_1 with respect to the 1-Lipschitz function f. The last equality follows from the standard posterior sampling lemma in the Bayesian setting (Osband et al., 2013; Osband and Van Roy, 2014), which suggests that at time t_k, given the sigma-algebra $\sigma(\mathcal{H}_{t_k})$, $\theta^{k,*}$ and θ^* are identically distributed:

$$\mathbb{E}[f(\theta^{k,*})|\mathcal{H}_{t_k}] = \mathbb{E}[f(\theta^*)|\mathcal{H}_{t_k}].$$

Following the same argument, condition on $\mathcal{H}_{t_k}$, we also have,

$$\mathbb{E}[f(\theta^*)|\mathcal{H}_{t_k}] = \mathbb{E}[f(\theta^{k,*})|\mathcal{H}_{t_k}] = \int_{\mathbb{R}^d} f(\theta)(\rho_{t_k} + \tilde{\rho}_{t_k} - \tilde{\rho}_{t_k})(d\theta) \leq \mathbb{E}[f(\theta^k)|\mathcal{H}_{t_k}] + W_1(\tilde{\rho}_{t_k}, \rho_{t_k}). \tag{15}$$

Combining Equation (14) and (15) yields Equation (4). Applying the tower rule concludes the proof. ■

Corollary 1 (Tabular Langevin Posterior Sampling)**.** *In tabular settings with finite states and actions, by running an approximate sampling method for each $(s,a) \in \mathcal{S} \times \mathcal{A}$ at time t_k, it holds that for each policy switch $k \in [K_T]$,*

$$\Big|\mathbb{E}[f(\theta^*)|\mathcal{H}_{t_k}] - \mathbb{E}[f(\theta^k)|\mathcal{H}_{t_k}]\Big| \leq \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} W_1(\tilde{\rho}_{t_k}(s,a), \rho_{t_k}(s,a)),$$

where $\tilde{\rho}_{t_k}(s,a)$ are the corresponding true posterior and approximate posterior for (s,a) at time t_k.

Proof of Corollary 1 Since we run the approximate sampling algorithm for each state-action pair at the beginning of each policy switch k, the total approximation error is equal to the sum of approximation error for each (s,a). ■

We first provide a general regret decomposition in Lemma E.2, which holds for any undiscounted weakly-communicating MDPs with infinite horizon, where approximate sampling is adopted and the transition is Lipschitiz.

Lemma E.2 (Regret decomposition.)**.** *For a weakly-communicating MDP with infinite time-horizon T, the Bayesian regret of Algorithm 3 instantiated with any approximate sampling method can be decomposed as follows:*

$$R_B(T) \leq \mathbb{E}\Big[\sum_{k=1}^{K_T} T_k W_1(\tilde{\rho}_{t_k}, \rho_{t_k})\Big] + H(\log T + 1) + HL_p\mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} \big\|\theta^* - \theta^k\big\|\Big], \tag{16}$$

where L_p is a Lipschitz constant, and H is the upper bound of span of MDP.

Proof of Lemma E.2 We adopt the greedy policy with respect to the sampled model, which gives $a_t = \text{argmax}_{a\in A} r(s_t, a)$ at each time step t. By Bellman Optimality equation in Lemma 1,

$$J^{\pi_k}(\theta^k) + h^{\pi_k}(s, \theta^k) = \mathcal{R}(s_t, a_t) + \int_{s'\in\mathcal{S}} p(s'|s_t, a_t; \theta^k) h^{\pi_k}(s', \theta^k) ds', \quad \forall t \in [t_k, t_{k+1} - 1]. \tag{17}$$

We then follow the standard analyses in RL literature (Osband et al., 2013; Osband and Van Roy, 2014) to decompose the regret into sum of Bellman errors. Plug in Equation (17) into the definition of Bayesian regret, we have,

$$R_B(T) = \mathbb{E}\left[\sum_{t=1}^{T} J^{\pi^*}(\theta^*) - \mathcal{R}(s_t, a_t)\right]$$

$$= \mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} J^{\pi^*}(\theta^*) - \mathcal{R}(s_t, \pi_k(s_t))\right]$$
$$= \underbrace{\mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} J^*(\theta^*) - J^{\pi_k}(\theta^k)\Big]}_{\text{(i)}} + \underbrace{\mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\left(\int_{s'\in\mathcal{S}} p(s'|s_t,\pi_k(s_t);\theta^k)h^{\pi_k}(s',\theta^k)ds' - h^{\pi_k}(s_t,\theta^k)\right)\right]}_{\text{(ii)}} \tag{18}$$

Term (i). By the property of approximate posterior sampling in Lemma 3 and the non-negativity of Wasserstein distance,

$$\text{(i)} \le |\text{(i)}| \le \mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\Big|J^*(\theta^*) - J^{\pi_k}(\theta^k)\Big|\Big] \le \mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} W_1(\rho_{t_k}, \tilde{\rho}_{t_k})\Big] = \mathbb{E}\Big[\sum_{k=1}^{K_T} T_k W_1(\tilde{\rho}_{t_k}, \rho_{t_k})\Big]. \tag{19}$$

We remark that this term differs from the exact PSRL where no approximate sampling method is used. To ensure the final regret is properly bounded, approximate sampling method being used must provide sufficient statistical guarantee of $\tilde{\rho}_{t_k}$ and ρ_{t_k} in terms of Wasserstein-1 distance.

Term (ii). We further decompose term (ii) into the model estimation errors.

$$\text{(ii)} = \mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\Big(\int_{s'\in\mathcal{S}} p(s'|s_t,\pi_k(s_t);\theta^k)h^{\pi_k}(s',\theta^k)ds' - h^{\pi_k}(s_t,\theta^k) + h^{\pi_k}(s_{t+1},\theta^k) - h^{\pi_k}(s_{t+1},\theta^k)\Big)\Big]$$
$$= \underbrace{\mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\Big(h^{\pi_k}(s_{t+1},\theta^k) - h^{\pi_k}(s_t,\theta^k)\Big)\Big]}_{\Delta_h}$$
$$+ \underbrace{\mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\int_{s'\in\mathcal{S}}\Big(p(s'|s_t,\pi_k(s_t),\theta^k) - p(s'|s_t,\pi_k(s_t),\theta^*)\Big)h^{\pi_k}(s',\theta^k)ds'\Big]}_{\Delta_{err}}.$$

To bound Δ_h, note that for each $k \in [1, K_T]$, $sp(h(\theta^k)) \le H$, and by Theorem 8,

$$\begin{aligned}\Delta_h &= \mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\Big(h^{\pi_k}(s_{t+1},\theta^k) - h^{\pi_k}(s_t,\theta^k)\Big)\Big]\\ &= \mathbb{E}\Big[\sum_{k=1}^{K_T}\Big(h^{\pi_k}(s_{t_{k+1}},\theta_k) - h^{\pi_k}(s_{t_k},\theta_k)\Big)\Big]\\ &\le \mathbb{E}\left[sp(h(\theta^k))K_T\right]\\ &\le H(\log T + 1).\end{aligned} \tag{20}$$

Thus, combining Equation (18),(19), (20), and by Lemma E.3, we conclude the proof.

■

Lemma E.3 (Bound estimation error). *Let* $\Delta_{err} = \mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\int_{s'\in\mathcal{S}}\Big(p(s'|s_t,\pi_k(s_t),\theta^k) - p(s'|s_t,\pi_k(s_t),\theta^*)\Big)h^{\pi_k}(s',\theta^k)ds'\Big]$. *Suppose Assumption 6 holds, then*

$$\Delta_{err} \le HL_p\mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\|\theta^* - \theta^k\|\Big]. \tag{21}$$

Proof of Lemma E.3 Recall that π_k is the optimal policy under θ^k, thus $h(\cdot, \theta^k) = h^{\pi_k}(\cdot, \theta^k)$, and span is properly bounded in weakly-communicating MDPs: $sp(h(\theta)) \leq H$ for any $\theta \in \mathbb{R}^d$. Then by Assumption 6 and Cauchy-Schwartz inequality,

$$\begin{aligned}
&\int_{s' \in \mathcal{S}} \Big(p(s'|s_t, \pi_k(s_t), \theta^k) - p(s'|s_t, \pi_k(s_t), \theta^*) \Big) h^{\pi_k}(s', \theta^k) ds' \\
&\leq \left\| p(\cdot|s_t, \pi_k(s_t), \theta^k) - p(\cdot|s_t, \pi_k(s_t), \theta^*) \right\| \left\| h(\cdot, \theta^k) \right\|_\infty \\
&= H L_p \left\| \theta^* - \theta^k \right\| .
\end{aligned}$$

Plugging the result into the definition of Δ_{err} concludes the proof. ■

The above regret decomposition in Lemma E.2 holds regardless of the approximate sampling methods being employed. To derive the final regret bounds, we discuss in the context of General Parmeteration and Simple parameterization respectively.

E.2.1. General Parametrization

The first term in Equation (16) corresponds to the accumulating approximation error over the time horizon T due to the use of approximate sampling method. Upper bounding this term relies on the statistical guarantee provided by the adopted approximate sampling method, which is the main novelty of LPSRL. In this section, we focus on the regret guarantee under the general parameterization.

To maintain the sub-linear regret guarantee, the convergence guarantee provided by SGLD is required to effectively upper bound the approximation error in the first term of Lemma E.2.

Lemma E.4. *Suppose Assumptions 1-4 are satisfied. Under the general parameterization of MDP, by instantiating LPSRL with SGLD, it holds that for any $p \geq 2$,*

$$\sum_{k=1}^{K_T} T_k W_1(\tilde{\rho}_{t_k}, \rho_{t_k}) \leq \sqrt{\frac{24T(\log T + 1)}{m}} (d + \log Q + (32 + 8d\kappa^2)p)^{1/2} . \tag{22}$$

Proof of Lemma E.4 By design of Algorithm 3 and the convergence guarantee of SGLD in Theorem 1, we have,

$$\begin{aligned}
\sum_{k=1}^{K_T} T_k W_1(\tilde{\rho}_{t_k}, \rho_{t_k}) &\leq \sum_{k=1}^{\log T + 1} 2^{k-1} W_p(\tilde{\rho}_{t_k}, \rho_{t_k}) \\
&\leq \sum_{k=1}^{\log T + 1} 2^{k-1} \sqrt{\frac{12}{2^{k-1} m}} (d + \log Q + (32 + 8d\kappa^2)p)^{1/2} \\
&= \sqrt{\frac{12}{m}} (d + \log Q + (32 + 8d\kappa^2)p)^{1/2} \sum_{k=1}^{\log T + 1} \sqrt{2^{k-1}} \\
&\leq \sqrt{\frac{24T(\log T + 1)}{m}} (d + \log Q + (32 + 8d\kappa^2)p)^{1/2}.
\end{aligned}$$

Here, the first inequality follows from the fact that $W_p \geq W_q$ for any $p \geq q$. The second equality directly follows from Theorem 1, and the last inequality follows from the Cauchy-Schwartz inequality.

■

To further upper bound Δ_{err} in Lemma E.3 under the General Parameterization, we establish the following concentration guarantee provided by SGLD under the static doubling batching scheme adopted by LPSRL.

Lemma E.5 (Concentration of SGLD)**.** *For any policy-switch $k \in [K_T]$, instantiating LPSRL with SGLD guarantees that*

$$\mathbb{E}\Big[T_k \|\theta^* - \theta^k\|^2 \Big] \leq \frac{960d}{m} \log T.$$

Proof of Lemma E.5 At time t_k, denote $\theta^{k,*}$ the parameter sampling from the true posterior ρ_{t_k}, and n_{t_k} the total number of available observations. By the triangle inequality,

$$\left\| \theta^* - \theta^k \right\|^2 \leq 3(\left\| \theta^* - \theta^{k,*} \right\|^2 + \left\| \theta^{k,*} - \theta^k \right\|^2).$$

Taking expectation and multiplying both sides by 2^{k-1} yields,

$$\mathbb{E}[T_k\|\theta^* - \theta^k\|^2] \leq 3\mathbb{E}[T_k\|\theta^* - \theta^{k,*}\|^2] + 3\mathbb{E}[T_k\|\theta^{k,*} - \theta^k\|^2]. \tag{23}$$

Under the General parameterization, we follow Assumption A2 in (Theocharous et al., 2017b) to focus on the MDPs that have proper concentration, which suggests the true parameter θ^* and the mode of the posterior $\theta^{k,*}$ satisfies

$$\mathbb{E}\left[\|\theta^* - \theta^{k,*}\|^2\right] \leq \frac{32d}{mn_{t_k}}\log T.$$

We provide an example in Appendix E.3 to show this assumption can be easily satisfied in practice. Let $\tilde{D}^2 := \frac{32d}{m}\log T$, then by Theorem 7 adapted to the MDP setting (i.e. with a change in notation), we have,

$$W_2^2(\widetilde{\rho}_{t_k}, \rho_{t_k}) \leq \frac{4\tilde{D}^2}{n_{t_k}}.$$

Note that $n_{t_k} = \sum_{k'=1}^{k-1} T_{k'}$ and by design of Algorithm 3, $n_{t_k} \leq T_k \leq 2n_{t_k}$. Combining the above results and Equation (23), we have

$$\mathbb{E}\left[T_k\|\theta^* - \theta^k\|^2\right] \leq 30\tilde{D}^2 \leq \frac{960d}{m}\log T, \quad \forall \tilde{D}^2 \geq \frac{32d}{m}\log T.$$

■

With all the above results, we are now ready to prove the main theorem for LPSRL with SGLD.

Theorem 4. *Under Assumptions $1-6$, by instantiating `SamplingAlg` with SGLD and setting the hyperparameters as per Theorem 1, with $p = 2$, the regret of LPSRL (Algorithm 3) satisfies:*

$$R_B(T) \leq CH\log T\sqrt{\frac{T}{m}}(d + \log Q + (32 + 8d\kappa^2)p)^{1/2},$$

where C is some positive constant, H is the upper bound of the MDP span, and Q denotes the quality of the prior. The total number of iterations required for SGLD is $O(\kappa^2\log T)$.

Proof of Theorem 4 First we further upper bound Δ_{err} using the concentration guarantee provided by SGLD. We first note that by Cauchy–Schwarz inequality,

$$\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\left\|\theta^* - \theta^k\right\| = \sum_{t=1}^{T}\left\|\theta^* - \theta^k\right\| \leq \sqrt{T\sum_{t=1}^{T}\|\theta^* - \theta^k\|^2} = \sqrt{T\sum_{k=1}^{K_T}T_k\|\theta^* - \theta^k\|^2}. \tag{24}$$

Combining Equation (21) in Lemma E.3 and (24), by Theorem 8, Lemma E.3 and E.5, we have,

$$\begin{aligned}
\Delta_{err} &\leq HL_p\sqrt{T\mathbb{E}\Big[\sum_{k=1}^{K_T}T_k\|\theta^* - \theta^k\|^2\Big]} \\
&\leq HL_p\sqrt{TK_T\max_k\mathbb{E}\Big[T_k\|\theta^* - \theta^k\|^2\Big]} \\
&\leq HL_p\sqrt{\frac{960d}{m}TlogT(\log T + 1)} \\
&\leq H(\log T + 1)\sqrt{\frac{960d}{m}T}.
\end{aligned} \tag{25}$$

Then combining Lemma E.2, E.4 and Equation 25, we have,

$$R_B(T) \leq H(\log T + 1) + H(\log T + 1)\sqrt{\frac{960d}{m}T} + \sqrt{\frac{24T(\log T + 1)}{m}}(d + \log Q + (32 + 8d\kappa^2)p)^{1/2}$$

$$\leq (1+\sqrt{960}+\sqrt{24})H(\log T+1)\sqrt{\frac{T}{m}}(d+\log Q+(32+8d\kappa^2)p)^{1/2}$$
$$\leq 38H(\log T+1)\sqrt{\frac{T}{m}}(d+\log Q+(32+8d\kappa^2)p)^{1/2}.$$

■

E.2.2. SIMPLEX PARAMETRIZATION

We now discuss the performance of LPSRL under simplex parametrization. Similar to the General Parameterization, the regret guarantee of LPSRL relies on the convergence guarantee of MLD, which is presented in the following theorem.

Theorem 5. *At the beginning of each policy-switch k, for each state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, sample transition probabilities over $\Delta^{|\mathcal{S}|}$ using MLD with the entropic mirror map. Let n_{t_k} be the number of data samples for any (s,a) at time t_k, then with step size chosen per Cheng and Bartlett (2018), running MLD with $O(n_{t_k})$ iterations guarantees that $W_2(\tilde{\rho}_{t_k},\rho_{t_k}) = \tilde{O}\left(\sqrt{|\mathcal{S}|/n_{t_k}}\right)$.*

Proof of Theorem 5 Theorem 5 follows from Theorems 2 and 3 from Hsieh et al. (2018) with step sizes given as per Theorem 3 from Cheng and Bartlett (2018). ■

Instantiating Algorithm 3 with MLD provides the following statistical guarantee to control the approximation error in terms of the Wasserstein-1 distance.

Lemma E.6. *Under the simplex parameterization of MDPs, we run MLD for each state-action pair $(s,a)\in\mathcal{S}\times\mathcal{A}$ at the beginning of each policy-switch $k\in[K_T]$ for $\tilde{O}(|\mathcal{S}||\mathcal{A}|n_{t_k})$ iterations. Suppose Assumption 5 and 6 are satisfied, then by instantiating LPSRL (Algorithm 3) with MLD (Algorithm 4) as* `SamplingAlg`*, we have,*

$$\sum_{k=1}^{K_T} T_k W_1(\tilde{\rho}_{t_k},\rho_{t_k}) \leq |\mathcal{S}|\sqrt{8|\mathcal{A}|T\log T}. \tag{26}$$

Proof of Lemma E.6 By Corollary 1, in tabular settings, the error term in the Wasserstein-1 distance can be further decomposed in terms of state-action pairs, suggesting

$$W_1(\tilde{\rho}_{t_k},\rho_{t_k}) = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} W_1(\tilde{\rho}_{t_k}(s,a),\rho_{t_k}(s,a)).$$

Then by design of Algorithm 3, we have,

$$\begin{aligned}\sum_{k=1}^{K_T} T_k W_1(\tilde{\rho}_{t_k},\rho_{t_k}) &= \sum_{k=1}^{K_T} T_k \sum_{s,a} W_1(\tilde{\rho}_{t_k}(s,a),\rho_{t_k}(s,a)) \\ &\leq \sum_{k=1}^{\log T+1} T_k \sum_{s,a} W_2(\tilde{\rho}_{t_k}(s,a),\rho_{t_k}(s,a)) \\ &\leq \sum_{k=1}^{\log T+1} T_k|\mathcal{S}||\mathcal{A}|\max_{s,a} W_2(\tilde{\rho}_{t_k}(s,a),\rho_{t_k}(s,a)),\end{aligned} \tag{27}$$

where the first inequality follows from the fact that $W_p \geq W_q$ for any $p\geq q$.

The convergence guarantee provided by Theorem 5 for MLD suggests, for each state-action pair $(s,a)\in\mathcal{S}\times\mathcal{A}$, upon running MLD for $\tilde{O}(|\mathcal{S}||\mathcal{A}|n_{t_k})$ iterations, where n_{t_k} is the number of data available for (s,a) at time t_k, we have

$$W_2(\tilde{\rho}_{t_k}(s,a),\rho_{t_k}(s,a)) \leq \sqrt{\frac{1}{|\mathcal{A}|n_{t_k}}}. \tag{28}$$

At time t_k, let n_{t_k} be the total number of available observations, which gives $n_{t_k} = \sum_{k'=1}^{k-1} T_{k'}$ and $t_k = n_{t_k} + 1$. By design of Algorithm 3, $n_{t_k} \leq T_k \leq 2n_{t_k}$. Then combining Equation (27) and (28) gives,

$$\begin{aligned}\sum_{k=1}^{K_T} T_k \sum_{s,a} W_1(\tilde{\rho}_{t_k}(s,a), \rho_{t_k}(s,a)) &\leq \sum_{k=1}^{\log T+1} |\mathcal{S}|\sqrt{2|\mathcal{A}|T_k} \\ &\leq \sum_{k=1}^{\log T+1} |\mathcal{S}|\sqrt{|\mathcal{A}|2^k} \\ &\leq |\mathcal{S}|\sqrt{|\mathcal{A}|(\log T+1)\sum_{k=1}^{\log T+1} 2^k} \\ &\leq |\mathcal{S}|\sqrt{8|\mathcal{A}|T\log T},\end{aligned}$$

where the third inequality follows from the Cauchy-Schwarz inequality.

■

Lemma E.6 suggests the approximation error in the first term of Lemma E.2 can be effectively bounded when instantiating `SamplingAlg` with MLD.

Lemma E.7 (Concentration of MLD)**.** *For any policy-switch $k \in [K_T]$, we run MLD (Algorithm 4) for each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ at time t_k for $\tilde{O}(|\mathcal{S}||\mathcal{A}|n_{t_k})$ iterations. Then instantiating LPSRL with MLD guarantees that*

$$\mathbb{E}[T_k\|\theta^{k,*} - \theta^k\|^2] \leq 2|\mathcal{S}|^2|\mathcal{A}|.$$

Proof of Lemma E.7 By tower's rule and the triangle inequality, we have

$$\begin{aligned}\mathbb{E}\left[T_k\|\theta^{k,*} - \theta^k\|^2\right] &= \mathbb{E}\left[\mathbb{E}\left[T_k\|\theta^{k,*} - \theta^k\|^2\right]\Big|\mathcal{H}_{t_k}\right] \\ &\leq \mathbb{E}\left[T_k W_2^2(\tilde{\rho}_{t_k}, \rho_{t_k})\Big|\mathcal{H}_{t_k}\right] \\ &\leq \mathbb{E}\left[T_k(|\mathcal{S}||\mathcal{A}|)^2 \max_{s,a} W_2^2(\tilde{\rho}_{t_k}(s,a), \rho_{t_k}(s,a))\Big|\mathcal{H}_{t_k}\right]. \qquad (29)\end{aligned}$$

where the last inequality follows from the fact that in tabular setting, $W_2^2(\tilde{\rho}_{t_k}, \rho_{t_k}) = \left(\sum_{s,a} W_2(\tilde{\rho}_{t_k}(s,a), \rho_{t_k}(s,a))\right)^2$.

By the convergence guarantee of MLD in Theorem 5, for each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, upon running MLD for $\tilde{O}(|\mathcal{S}||\mathcal{A}|n_{t_k})$ iterations, we have

$$W_2(\tilde{\rho}_{t_k}(s,a), \rho_{t_k}(s,a)) \leq \sqrt{\frac{1}{|\mathcal{A}|n_{t_k}}}. \qquad (30)$$

Combining Equation (29) and (30) and the fact that $n_{t_k} \leq T_k \leq 2n_{t_k}$ concludes the proof.

■

With the concentration guarantee between sample θ^k and $\theta^{k,*}$, as well as the concentration guarantee between $\theta^{k,*}$ and θ^* in exact PSRL, we are able to effectively upper bound the model estimation error Δ_{err} in tabular settings.

Lemma E.8 (Bound Δ_{err} in tabular settings)**.** *With the definition of model estimation error Δ_{err} in Lemma E.3, in tabular setting, the following upper bound holds for Δ_{err},*

$$\Delta_{err} \leq 66H|\mathcal{S}|\sqrt{|\mathcal{A}|T\log(2|\mathcal{S}||\mathcal{A}|T)}, \qquad (31)$$

where L_p is the Lipschitz constant for transition dynamics.

Proof of Lemma E.8 By the triangle inequality and Lemma E.3,

$$\Delta_{err} \leq HL_p\mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\big\|\theta^* - \theta^k\big\|\Big] \leq HL_p\left(\mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\big\|\theta^* - \theta^{k,*}\big\|\Big] + \mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\big\|\theta^{k,*} - \theta^k\big\|\Big]\right). \qquad (32)$$

Bound The first term. The first term can be upper bounded using standard concentration results of exact PSRL algorithms in Bayesian settings. Define the event

$$E_\theta = \left\{\theta : \forall (s,a) \in \mathcal{S}\times\mathcal{A}, \quad \left\|\theta(\cdot|s,a) - \widehat{\theta}^k(\cdot|s,a)\right\|_1 \le \beta_k(s,a)\right\}, \tag{33}$$

where $\widehat{\theta}^k$ is the empirical distribution at the beginning of policy switch k, $\beta_k(s,a) := \sqrt{\frac{14|\mathcal{S}|\log(2|\mathcal{S}||\mathcal{A}|t_kT)}{max(1,n_{t_k}(s,a))}}$ following (Jaksch et al., 2010; Ouyang et al., 2017; Osband et al., 2013) by setting $\delta = 1/T$. Then event E_θ happens with probability at least $1-\delta$. Note that for any vector x, $\|x\|_2 \le \|x\|_1$, and by the triangle inequality, we have

$$\left\|\theta^* - \theta^{k,*}\right\| \le \sum_{s'\in\mathcal{S}} \left|\theta^*(\cdot|s,a) - \theta^{k,*}(\cdot|s,a)\right| \le 2(\beta_k(s_t,a_t) + \mathbf{1}_{\{\theta^*\notin E_\theta\}}).$$

At any time $t \in [t_k, t_k + T_k - 1]$, $n_t \le 2n_{t_k}$ for any state-action pair (s_t,a_t), and by the fact that $t_k \le T$, we have

$$\mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\beta_k(s_t,a_t)\Big] \le \sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\sqrt{\frac{28|\mathcal{S}|\log(2|\mathcal{S}||\mathcal{A}|t_kT)}{max(1,n_t(s_t,a_t))}} \le \sum_{t=1}^{T}\sqrt{\frac{56|\mathcal{S}|\log(2|\mathcal{S}||\mathcal{A}|T)}{max(1,n_t(s_t,a_t))}}. \tag{34}$$

It then suffices to bound $\sum_{t=1}^T 1/\sqrt{max(1,n_t s_t,a_t)}$. Note that

$$\begin{aligned}
\sum_{t=1}^{T}\frac{1}{\sqrt{max(1,n_t(s_t,a_t))}} &= \sum_{(s,a)}\sum_{t=1}^{T}\frac{\mathbf{1}_{(s_t,a_t)=(s,a)}}{\sqrt{max(1,n_t(s,a))}} \\
&\le 4\sum_{(s,a)}\int_{z=0}^{n_{T+1}(s,a)} z^{-1/2}dz \\
&\le 4\sqrt{|\mathcal{S}||\mathcal{A}|\sum_{(s,a)} n_{T+1}(s,a)} \\
&\le 4\sqrt{|\mathcal{S}||\mathcal{A}|T}.
\end{aligned} \tag{35}$$

On the other hand, by definition of $\beta_k(s,a)$, $\mathbb{P}(\theta^* \notin E_\theta\}) \le 1/(Tt_k^6)$, which yields

$$\mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\mathbf{1}_{\{\theta^*\notin E_\theta\}}\right] \le \mathbb{E}\left[\sum_{k=1}^{K_T}T_k\mathbb{P}(\theta^*\notin E_\theta\})\right] \le \sum_{k=1}^{\infty}k^{-6} \le \sum_{k=1}^{\infty}k^{-2} \le 2. \tag{36}$$

Combining Equation (34), (35) and (36), we have,

$$HL_p\mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\left\|\theta^* - \theta^{k,*}\right\|\Big] \le 64HL_p|\mathcal{S}|\sqrt{|\mathcal{A}|T\log(2|\mathcal{S}||\mathcal{A}|T)} \tag{37}$$

Bound the second term. The second term arises from the use of approximate sampling. Note that by Cauchy–Schwarz inequality, this term in Equation (32) satisfies,

$$\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\left\|\theta^{k,*} - \theta^k\right\| = \sum_{t=1}^{T}\left\|\theta^{k,*} - \theta^k\right\| \le \sqrt{T\sum_{t=1}^{T}\left\|\theta^{k,*} - \theta^k\right\|^2} = \sqrt{T\sum_{k=1}^{K_T}T_k\left\|\theta^{k,*} - \theta^k\right\|^2}. \tag{38}$$

It then relies on the concentration guarantee provided by MLD for LPSRL under the static policy switch scheme. By Lemma E.7, we have,

$$HL_p\sqrt{T\mathbb{E}\Big[\sum_{k=1}^{K_T}T_k\left\|\theta^* - \theta^k\right\|^2\Big]} \le HL_p\sqrt{TK_T\max_k\mathbb{E}\left[T_k\left\|\theta^* - \theta^k\right\|^2\right]} \le HL_p|\mathcal{S}|\sqrt{4|\mathcal{A}|T\log T}. \tag{39}$$

Combining Equation (37) and (39) concludes the proof.

■

With all the above results, we now proceed to prove the regret bound for LPSRL with MLD.

Theorem 6. *Suppose Assumptions 5 and 6 are satisfied, then by instantiating* `SamplingAlg` *with MLD (Algorithm 4), there exists some positive constant C such that the regret of LPSRL (Algorithm 3) in the Simplex Parameterization is bounded by*

$$R_B(T) \leq CH|\mathcal{S}|\sqrt{|\mathcal{A}|T\log(|\mathcal{S}||\mathcal{A}|T)},$$

where C is some positive constant, H is the upper bound of the MDP span. The total number of iterations required for MLD is $O(|\mathcal{S}|^2|\mathcal{A}|^2T)$.

Proof of Theorem 6 By Lemma E.2, E.6 and E.8, we have

$$\begin{aligned} R_B(T) &\leq H(\log T + 1) + 66H|\mathcal{S}|\sqrt{|\mathcal{A}|T\log(2|\mathcal{S}||\mathcal{A}|T)} + |\mathcal{S}|\sqrt{8|\mathcal{A}|T\log T} \\ &\leq 2H\log T + 66H|\mathcal{S}|\sqrt{|\mathcal{A}|T\log(2|\mathcal{S}||\mathcal{A}|T)} + 4|\mathcal{S}|\sqrt{|\mathcal{A}|T\log T} \\ &\leq 72H|\mathcal{S}|\sqrt{|\mathcal{A}|T\log(2|\mathcal{S}||\mathcal{A}|T)}. \end{aligned}$$

By Lemma E.6, for each state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$ and policy-switch $k \in [K_T]$, the number of iterations required for MLD is $O(|\mathcal{S}||\mathcal{A}|2^{k-1})$. This suggests that for each state-action pair, the total number of iterations required for MLD is $O(|\mathcal{S}||\mathcal{A}|T)$ along the time horizon T. Summing over all possible state-action pairs, the computational cost of running MLD in terms of the total number of iterations is $O(|\mathcal{S}|^2|\mathcal{A}|^2T)$. ■

E.3. General Parameterization Example

Following (Theocharous et al., 2017a;b) we consider a points of interest (POI) recommender system where the system recommends a sequence of points that could be of interest to a particular tourist or individual. We will let the points of interest be denoted by points on $\mathbb{R}$. Following the perturbation model in (Theocharous et al., 2017a;b), the transition probabilities are $p(s|\theta) = p(s)^{1/\theta}$ if the chosen action is s and it is $p(s)/z(\theta)$ otherwise. Here s is a state or a POI and $z(\theta) = \frac{\sum_{x\neq s} p(x)}{1-p(s)^{1/\theta}}$. Furthermore, to fully specify p we consider $p(s|\theta) = \frac{1}{\sqrt{2\pi}}e^{-s^2/2\theta}$. One can see that Assumptions 1-4 are satisfied due to the Gaussian-like nature of the transition dynamics and the satisfiability of Assumption 5 follows from Lemma 5 in (Theocharous et al., 2017b).

F. Experimental Details

F.1. Additional Discussions of Langevin TS in Gaussian Bandits

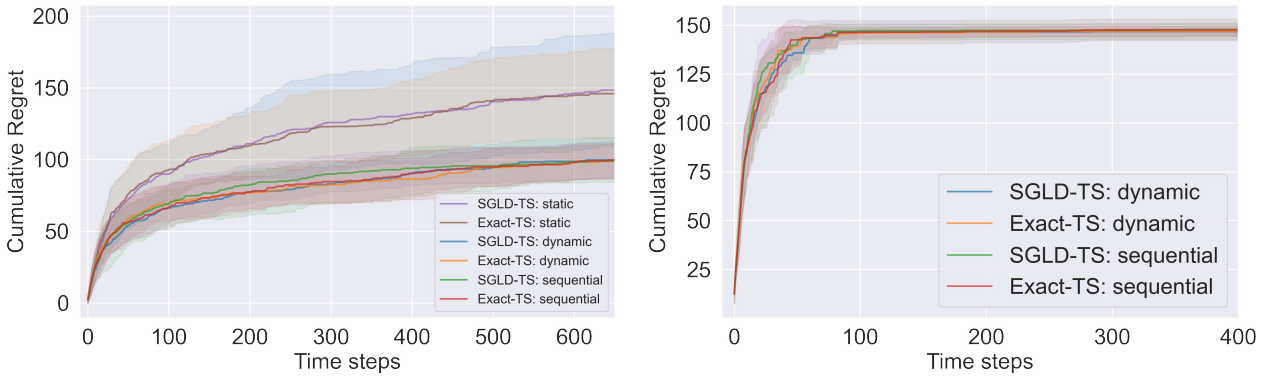

Figure 2: Left:(a) Expected regret for informative priors. Right:(b) Expected regret for uninformative priors. Results are reported over 10 experiments. In both scenarios, SGLD-TS under dynamic scheme achieves optimal performance as in sequential case without using approximate sampling.

In this section, we present additional empirical results for the Gaussian bandit experiments. In particular, we examine both informative priors and uninformative priors for Gaussian bandits with $N = 15$ arms, where each arm is associated with distinct expected rewards. We set the true expected rewards of all arms to be evenly spaced in the interval $[1, 20]$,

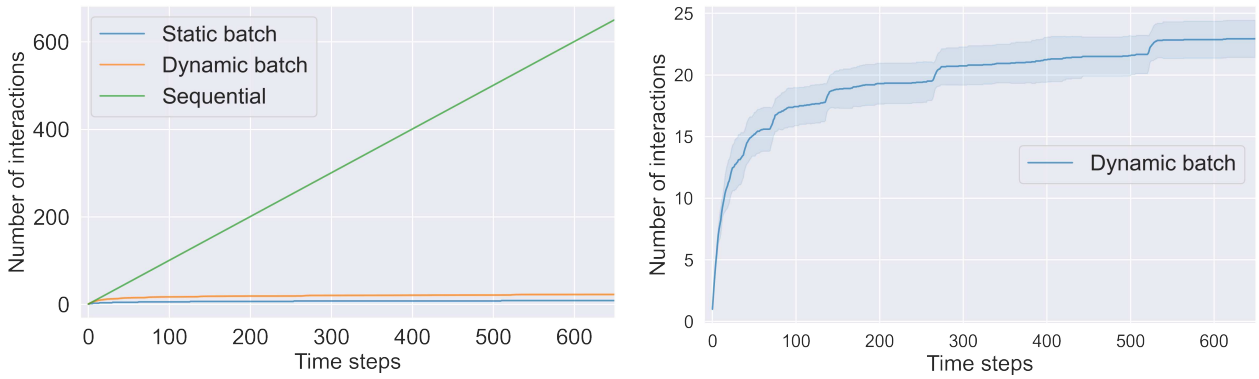

Figure 3: Left:(a) expected communication cost of three batching schemes: fully-sequential mode, dynamic batch, and static batch. Right:(b) expected communication cost under dynamic batching scheme for Gaussian bandits.

and the ordering of values is shuffled before assigning to arms. All arms share the same standard deviation of 0.5. We investigate the performance of SGLD-TS against UCB1, Bayes-UCB, and exact-TS under different interaction schemes: fully-sequential mode, dynamic batch scheme, and static batch scheme.

In the first setting, we assume prior knowledge of the ordering of expected rewards and apply informative priors to facilitate the learning process. Gaussian priors are adopted with means evenly spaced in $[14, 20]$, and inverted variance (i.e., precision) set to 0.375. The priors are assigned according to the ordering of the true reward distributions. Note that the exact knowledge of the true expected values is not required. In TS algorithms, the selection of arms at each time step is based on sampled values, therefore efficient learning is essential even with the knowledge of the correct ordering. The expected regret of all methods is reported over 10 experiments and results are illustrated in Figure 2(a). Results of both Figure 1(a) and Figure 2(a) demonstrate that SGLD-TS achieves optimal performance similar to exact-TS with conjugate families. Its appealing empirical performance in comparison to other popular methods (e.g., UCB1 and Bayes-UCB), along with its ability to handle complex posteriors using MCMC algorithms, make it a promising solution for challenging problem domains. Additionally, the introduction of the dynamic batch scheme ensures the computational efficiency of SGLD-TS. As depicted in Figure 3(a)(b) and Table 2 (column labeled "batches"), communication cost is significantly reduced from linear to logarithmic dependence on the time horizon, as suggested by Theorem 2. Furthermore, in bandit environments, our dynamic batch scheme exhibits greater robustness compared to the static batch scheme for both frequentist and Bayesian methods.

Furthermore, we explore the setting where prior information is absent, and uninformative priors are employed. In this case, we adopt the same Gaussian priors as $\mathcal{N}(14.0, 8.0)$ for all arms. Similar to the first setting, the same conclusion can be drawn for SGLD-TS from Figure 2(b).

F.2. Experimental Setup for Langevin TS in Laplace Bandits

In order to demonstrate the performance of Langevin TS in a broader class of general bandit environments where closed-form posteriors are not available and exact TS is not applicable, we construct a Laplace bandit environment consisting of $N = 10$ arms. Specifically, we set the expected rewards to be evenly spaced in the interval $[1, 10]$, and shuffle the ordering before assigning each arm a value. The reward distribution of each arm shares the same standard deviation of 0.8. We adopt favorable priors to incorporate the knowledge of the true ordering in Laplace bandits. It is important to note that our objective is to learn the expected rewards, and arm selection at each time step is based on the sampled values rather than the ordering. In particular, we adopt Gaussian priors with means evenly spaced in $[4, 10]$ (ordered according to prior knowledge). The inverted variance (i.e., precision) for all Gaussian priors is set to 0.875. We conduct the experiments 10 times and report the cumulative regrets in Figure 1(b).

By employing Langevin TS in the Laplace bandit environment, we aim to showcase the algorithm's effectiveness and versatility in scenarios where posteriors are intractable and exact TS cannot be directly applied.

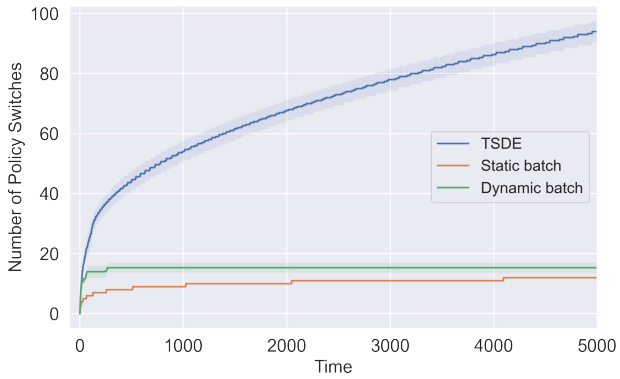

Figure 4: Number of interactions in RiverSwim over 10 experiments. Static policy switch scheme requires the least number of communication.

F.3. Experimental Setup for Langevin PSRL

In MDP setting, we consider a variant of RiverSwim environment being frequently used empirically (Strehl and Littman, 2008), in which the agent swimming in the river is modeled with five states, and two available actions: left and right. If the agent swims rightwards along the river current, the attempt to transit to the right is going to succeed with a large probability of $p = 0.8$. If the agent swims leftwards against the current, the transition probability to the left is small with $p = 0.2$. Rewards are zero unless the agent is in the leftmost state ($r = 2.0$) or the rightmost state ($r = 10.0$). The agent is assumed to start from the leftmost state. We implement MLD-PSRL and exact-PSRL under two policy switch schemes, one is the static doubling scheme discussed in section 6, and the other is the dynamic doubling scheme based on the visiting counts of state-action pairs. To ensure the performance of TSDE, we adopt its original policy switch criteria based on the linear growth restriction on episode length and dynamic doubling scheme. We run experiments 10 times, and report the average rewards of each method in Figure 1(c).

G. Societal Impacts

Our work focuses on sequential decision making in general reward/transition settings and with reduced communication costs. It is a theoretical work and there is no negative societal impact.