Isoperimetry is All We Need: Langevin Posterior Sampling for RL with Sublinear Regret

Emilio Jorge
Chalmers University of Technology
Gothenburg University

Christos Dimitrakakis
University of Neuchâtel
University of Oslo
Chalmers University of Technology

Debabrota Basu
Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 CRIStAL

Abstract

In Reinforcement Learning (RL) theory, we impose restrictive assumptions to design an algorithm with provably sublinear regret. Common assumptions, like linear or RKHS models, and Gaussian or log-concave posteriors over the models, do not explain practical success of RL across a wider range of distributions and models. Thus, we study how to design RL algorithms with sublinear regret for isoperimetric distributions, specifically the ones satisfying the Log-Sobolev Inequality (LSI). LSI distributions include the standard setups of RL and others, such as many non-log-concave and perturbed distributions. First, we show that the Posterior Sampling-based RL (PSRL) yields sublinear regret if the data distributions satisfy LSI under some mild additional assumptions. Also, when we cannot compute or sample from an exact posterior, we propose a Langevin sampling-based algorithm design: LaPSRL. We show that LaPSRL achieves order optimal regret and subquadratic complexity per episode. Finally, we deploy LaPSRL with a Langevin sampler– SARAH-LD, and test it for different bandit and MDP environments. Experimental results validate the generality of LaPSRL across environments and its competitive performance with respect to the baselines.

1 INTRODUCTION

The last decade has seen a significant advance in Reinforcement Learning (RL), both in terms of theoretical understanding and success in practical applications. However, still, the theoretical results do not always apply or explain RL in real-world settings. The central issue is that to operate on complex environments RL algorithms aim to learn a parametric functional approximation of the environment and to theoretically analyse them, we often assume linear (Geramifard et al., 2013), bilinear (Ouhamma et al., 2022), or reproducible kernel (Chowdhury and Gopalan, 2019) type parametric models, and Gaussian or log-concave posteriors for Bayesian algorithms (Chowdhury and Gopalan, 2019; Osband and Van Roy, 2017). In this paper, we aim to narrow this gap further by studying whether we can achieve the desired regret guarantees for isoperimetric distributions. Isoperimetry relates to the ratio between the area of the perimeter and the volume of a set. It is known that some isoperimetric condition is needed for rapid mixing of Markov chains to avoid the risk of getting stuck in bad regions (Stroock and Zegarlinski, 1992), motivating us to study isoperimetric distributions in RL. In addition, isoperimetric distributions include all the aforementioned setups studied in RL theory, and also non-log-concave and perturbed versions of log-concave distributions as well as mean field neural networks (Nitanda et al., 2022). In fact, we will see that any posterior with a bounded likelihood function and a log-Sobolev prior will be log-Sobolev, which would include complex setups such as some forms of Bayesian neural networks. In optimization and sampling

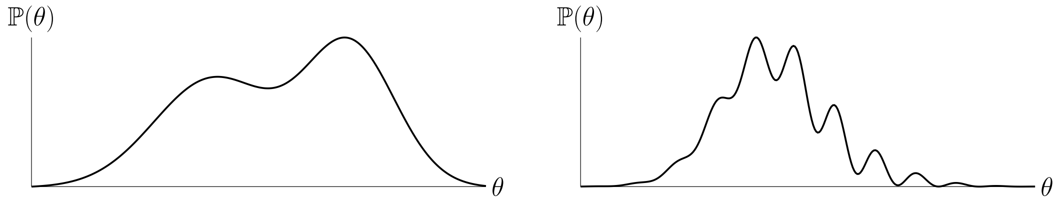


Figure 1: Examples of log-Sobolev distributions.

literature, isoperimetry is used as a minimal condition to conduct efficient and controlled sampling from target distribution(s) (Vempala and Wibisono, 2019) while ensuring proper concentration of empirical statistics (Ledoux, 2006). Among the different forms of isoperimetric inequalities (e.g. Poincaré, modified log Sobolev etc.), we consider the Log Sobolev Inequality (LSI) (Bakry et al., 2014) in this paper.

Posterior Sampling-based RL (PSRL). For our study, we focus on the popular PSRL algorithms (Russo et al., 2020; Osband et al., 2013), which are generalisation of Thompson sampling proposed for bandits (Thompson, 1933). PSRL is a Bayesian algorithm that begins with a prior distribution over the model parameters. As PSRL collects more data, it creates more informative posterior distributions, samples probable model parameters from the posteriors, and uses the sampled parameters for further planning. Since PSRL has been successful both theoretically and practically, we choose it as the base algorithm to study.

Still, exact sampling and tracking of the posterior may be intractable for many distributions (e.g. in high dimensions). It is easy to show that approximation in the sampling can lead to linear regret unless sufficient care is taken. On the other hand, being limited to distributions with exact sampling is insufficient for applications. Thus, there has been a series of works to relax PSRL with approximate posteriors and still to avoid linear regret.

Langevin Sampling-based PSRLs. One of the growing approaches in this direction is to use Langevin-based approximate sampling methods (Mazumdar et al., 2020; Zheng et al., 2024; Ishfaq et al., 2023; Karbasi et al., 2023), which are known to be generic and efficient in optimisation, sampling, and deep learning literature. Mazumdar et al. (2020); Zheng et al. (2024); Huix et al. (2023) propose Langevin-based PSRL algorithms for multi-armed bandits that achieve order-optimal regret only for log-concave distributions. Similarly, Xu et al. (2022) extends these ideas to linear contextual bandits but still with a linear dependence on the approximation error. Ishfaq et al. (2023) brings Langevin-based PSRL to Markov Decision Processes (MDPs) but the theoretical guarantees are available only for linear approximations. The work of Karbasi et al. (2023) also studies Langevin PSRL algorithms for bandits and reinforcement learning, however they also rely on strong concavity assumptions. However the sampling literature has shown that Langevin methods are efficient for isoperimetric distributions such as LSI. This motivates us to propose a generic algorithm that can work for any distribution satisfying LSI, and for bandits and MDPs, and also to study what are the minimum conditions required to achieve sublinear regret. Specifically, we ask:

1. *Is isoperimetry of posteriors enough to ensure efficient execution of PSRL-type algorithms?*
2. *Can we use Langevin sampling-based algorithms to approximate the isoperimetric posteriors and still obtain an efficient approximate PSRL algorithm?*

Our contributions address these questions affirmatively and more. Specifically, we

1. Prove that *PSRL can achieve sublinear regret for posteriors satisfying LSI* under some mild conditions if we can compute and sample from the exact posteriors. This new result broadens the scenarios where PSRL is proven to be efficient to a new and wider family of posteriors.

2. Propose a generic PSRL-algorithm, called **LaPSRL**, that *uses a Langevin-based sampling to compute approximate posterior distributions*. A generic regret analysis of LaPSRL shows it *can achieve $\mathcal{O}(\sqrt{T})$ regret if the approximate sampling algorithms allow the posterior to contract linearly*, where T is the number of interactions. Then, we show that if we deploy LaPSRL with SARAH-LD, which is an efficient Langevin sampling algorithm, we only need a polynomial number of samples w.r.t. the MDP parameters with and without chaining them. Conducting analysis

requires generalising the regret analysis with LSI and also studying the contraction of posterior over models under Langevin dynamics.

3. Show *LaPSRL with SARAH-LD achieves sublinear regret across different environments*, including Gaussian, Mixtures of LSI distributions as well as any log-concave distribution or mixture thereof.

4. *Experimentally demonstrate that LaPSRL with SARAH-LD yields sublinear regret* for bandits with Gaussians and mixture of Gaussians as posteriors, as well as continuous MDP experiments with Linear Quadratic Regulators (LQRs) and neural networks with Gaussian priors, and performs competitively with corresponding baselines.

2 PROBLEM FORMULATION & BACKGROUND

Before proceeding to the contributions, we first formally state the problem of episodic RL. Then we summarise PSRL for episodic RL and Langevin based sampling techniques, which are the main pillars of our work.

Notation. Throughout the paper we will use the following notation. Complexity notation O, Ω, Θ, with standard implications, and $\widetilde{O}, \widetilde{\Omega}, \widetilde{\Theta}$, which is the equivalent term but ignoring sub-logarithmic and poly-logarithmic terms. An overview of variable notation can be found in Table 1.

Problem Formulation: Episodic Reinforcement Learning (RL). To perform RL, we consider the episodic finite-horizon MDPs (aka *Episodic RL*) (Osband et al., 2013; Azar et al., 2017). MDP in episodic RL is defined as $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, R, H \rangle$. M has states $s \in \mathcal{S}$ where $\mathcal{S} \in \mathbb{R}^d$, actions $a \in \mathcal{A}$. We will also sometimes use $z = (s, a)$ to simplify notation. In episodic RL, the agent interacts with the environment in episodes of H steps. Any episode l starts with a state $s_{l,1}$. Then, for $t \in [H]$, the agent draws action $a_{l,t}$ from a policy $\pi_t(s_{l,t})$, observes the reward $R(s_{l,t}, a_{l,t}) \in \mathbb{R}$, and transits to a state $s^l_{t+1} \sim \mathcal{T}(. \mid s_{l,t}, a_{l,t})$. This interaction is done for a total of τ (which is commonly unknown) episodes. The performance of a policy π is measured by the total expected reward V_1^π w.r.t. an initial state s. We define the value function and the Q-value function at $h \in [H]$, $V_M^{\pi,h}(s) \triangleq \mathbb{E}\left[\sum_{t=h}^H R(s_t, a_t) \mid s_h = s\right]$, $Q_M^{\pi,h}(s,a) \triangleq \mathbb{E}\left[\sum_{t=h}^H R(s_t, a_t) \mid s_h = s, a_h = a\right]$.

The MDP is typically unknown. In the Bayesian approach, we construct a posterior distribution $\mathbb{P}(M \mid \mathcal{H}_l)$ over M given the data observed so far, i.e. $\mathcal{H}_l = \{s_{1,1}, a_{1,1}, \ldots, s_{l-1,H}, a_{l-1,H},\}$. When there is only one state, or the state does not depend on the action, this problem reduces to what is known as the multi-armed bandit problem (MAB) (Lattimore and Szepesvári, 2020). When $H = 1$ there is no sequential component and the problem becomes that of multi-armed bandits.

Background: PSRL. A popular Bayesian approach, which has been very successful is to sample an MDP $M_l \sim \mathbb{P}(M \mid \mathcal{H}_l)$ and play the optimal policy for M_l for one episode before updating the posterior and resampling. This algorithm is known as PSRL (Osband et al., 2013). PSRL reduces to Thompson sampling (Thompson, 1933), when applied to MAB. In this paper, we will use some simplifying notation, z_i is shorthand for (s_{i+1}, s_i, a_i). The concept of regret is crucial to RL theory, it describes how much worse the policy is than the optimal policy. In the Bayesian regret, this is taken in expectation of value over the possible MDPs and evaluations and can be written

$$BR(T) \triangleq \mathbb{E}\left[\sum_{l=1}^{\tau} V^{M_*}_{\pi_*,1}(s_{l,1}) - V^{M_*}_{\pi_l,1}(s_{l,1})\right] . \tag{1}$$

We also set $\Delta_{\max} \triangleq \max_\pi V^{M_*}_{\pi,1}(s_1) - \min_\pi V^{M_*}_{\pi,1}(s_1)$ to denote the maximal regret that could be obtained in one episode. In the paper, we use n to denote the amount of data samples we have observed. We denote to total interactions with the environment as $T = \tau H$.

Background: Sampling with Langevin dynamics. In the notation of Langevin sampling, we need to sample from a target distribution $d\nu \propto e^{-\gamma F}$, where $F : \mathbb{R}^d \to \mathbb{R}$. Specifically, we express $F(\theta) = 1/n \sum_{i=1}^n f_i(\theta)$, with each f_i representing the loss associated with a data point x_i, and F being the average loss. In the context of Bayesian posteriors, we can set $\gamma = n$ and define $f_i(\theta) = -1/n \log \mathbb{P}(\theta) - \log f(x_i|\theta)$, where each f_i corresponds to the log-likelihood for data point x_i and includes its share of the log prior.

In continuous time diffusion, Langevin methods can sample exactly from a posterior (Vempala and Wibisono, 2019). In practice, discretization makes this impossible, but using a Langevin gradient descent algorithm allows for sampling from the target distribution with a controlled bias, under conditions on isoperimetry. We need some assumptions on smoothness and isoperimetry.

Assumption 1 (L-smoothness). *If f_i is twice differentiable for all $i = 1 \dots, n$ and $\forall x, y \in \mathbb{R}^d$, $\left\|\nabla^2 f_i(x)\right\| \leq L$, then f_i is L-smooth. Additionally, this implies that F is also L-smooth.*

Definition 1 (log-Sobolev inequality). *A distribution ν satisfies the log-Sobolev inequality (LSI) with a constant α if, for all smooth functions $g : \mathbb{R}^n \to \mathbb{R}$ with $\mathbb{E}_\nu[g^2] \leq \infty$, the following holds:*

$$\mathbb{E}_\nu[g^2 \log g^2] - \mathbb{E}_\nu[g^2] \log \mathbb{E}_\nu[g^2] \leq \frac{2}{\alpha} \mathbb{E}_\nu[||\nabla g||^2]. \tag{2}$$

An equivalent way of writing the LSI, which is also commonly used and is found by defining the help function $\rho = \frac{g^2 \nu}{\mathbb{E}_\nu[g^2]}$, which reformulates the condition into $\mathrm{KL}(\rho \parallel \nu) \leq \frac{1}{2\alpha} J_\rho$ where $J_\rho := \mathbb{E}_\rho \left[\left\|\nabla \log \frac{\rho}{\nu}\right\|^2\right]$ is the relative Fisher information of ρ with respect to ν.

In this paper, we will only cover a brief introduction to log-Sobolev distributions as needed, but there has been much work looking into the properties of log-Sobolev distributions, a summary of which can be found in (Chafaï and Lehec, 2023; Vempala and Wibisono, 2019). Also note that in some work an inverse definition is used where the constant is defined $\alpha' = \frac{1}{2\alpha}$, leading to some confusion. Obtaining the LSI constant is not always trivial, but there are some tools. In some cases, the Bakry-Émery criterion can be used.

Theorem 1 (Bakry-Émery criterion). *If for distribution ν, $-\nabla_\theta^2 \log \nu \geq \alpha I_d$, where the inequality indicates the Loewner order and I_d the identity matrix of dimension d and θ the parametrization of ν, then ν fulfils LSI with constant α.*

There are plenty of other tools for analysing log-Sobolev constants such as Lyaponov conditions, integral conditions, local inequalities and tools from optimal transport as well as decomposing into mixtures (Cattiaux et al., 2010; Wang, 2001; Barthe and Kolesnikov, 2008; Chen et al., 2021a; Koehler et al., 2023). Log-concave distributions $P(\theta)$ are distributions where $\log P(\theta)$ is concave in θ, this is a commonly used condition, but is is much more restrictive than log-Sobolev. Theorem 1 shows that log-concave distributions imply LSI, but log-Sobolev distributions are more general. For example, log-concave distributions cannot be multimodal. Some examples of what log-Sobolev distributions could look like can be found in Figure 1.

One example of an operation on log-Sobolev distributions that preserves the property is that of bounded perturbation, something that would generally break a log-concave property. The theorem is originally due to Holley and Stroock (1987) but presented here as per Steiner (2021).

Theorem 2 (Steiner (2021)). *Assume $d\mu \propto e^{\Phi} d\nu$ where ν is a probability measure that satisfies LSI and Φ is continuous and bounded. Then μ satisfies a LSI with $\frac{1}{\alpha_\mu} \leq e^{2(\sup(\Phi) - \inf(\Phi))} \frac{1}{\alpha_\nu}$.*

In some cases, even unbounded perturbations could still fulfil LSI (Steiner, 2021). The LSI is also preserved under a Lipschitz-transformation (Vempala and Wibisono, 2019), and if the distribution is factorizable such that each part is log-Sobolev, then the product is log Sobolev with a constant that is equal to minimum constant among the factors (Ledoux, 2006). Mixtures of log-Sobolev distributions are also log-Sobolev under conditions on the distance between the distributions, more on that later.

The log-Sobolev inequality with constant α_ν implies Gaussian concentration of a function around its mean (Bizeul, 2023) such that for any locally Lipschitz function $g : \mathbb{R}^n \to \mathbb{R}$

$$\mathbb{P}_\nu(|g - E_\nu[g]| \geq t) \leq 2 \exp(-\frac{\alpha_\nu t^2}{L_g^2}) \tag{3}$$

where L_g is the Lipschitz constant of g. Under a curvature dimension condition, the reverse is also true, Gaussian concentration implies that the distribution is log-Sobolev, (Bakry et al., 2014, Theorem 8.7.2) .

Background: SARAH-LD (Kinoshita and Suzuki, 2022). There exists multiple algorithms for performing biased Langevin sampling on log-Sobolev distributions (Vempala and Wibisono, 2019;

Kinoshita and Suzuki, 2022). In this paper, we focus on SARAH-LD (Algorithm 4), which is a variance-reduced version of Langevin dynamics which is the current state-of-the-art in terms of KL divergence concentration to the target distribution. SARAH-LD allows us to control the bias, and trade-off the computational complexity with the KL-divergence between sampled and target distributions, i.e. $\mathrm{KL}(\rho \parallel \nu)$. The total amount of stochastic gradient evaluations that need to be done for any of the samples (also known as the gradient complexity) of SARAH-LD is $\widetilde{O}\left(\left(n + \frac{dn^{\frac{1}{2}}}{\epsilon}\right) \cdot \frac{\gamma^2 L^2}{\alpha^2}\right)$, complete result is deferred to Theorem 10.

3 PSRL FOR EXACT POSTERIORS

Now, we look into the convergence of Posterior Sampling-based RL algorithm (PSRL), when we have access to exact posterior distributions at each step. Specifically, we ask that *can PSRL achieve sublinear regret if it has access to isoperimetric data and prior distributions?*

First, we observe that following the series of works by Osband and Van Roy (2017); Chowdhury and Gopalan (2019); Chowdhury et al. (2021), a generic and minimal framework to bound the Bayesian regret ($BR(T)$) of PSRL is developed. It involves three steps. *Step 1:* If we consider the first step of an episode l, the total number of completed steps is $n = (l-1)H$. Osband et al. (2013) observes that for any $\sigma(\mathcal{H}_l)$ measurable function f, which includes the value function, we have $\mathbb{E}[f(\theta_n)] = \mathbb{E}[f(\theta^*)]$. Thus, we get from Equation (1)

$$BR(T) = \mathbb{E}\left[\sum_{l=1}^{\tau} V^{\theta_n}_{\pi_l,1}(s_{l,1}) - V^{\theta^\star}_{\pi_l,1}(s_{l,1})\right].$$

Step 2: Chowdhury and Gopalan (2019) further shows that by a recursive application of the Bellman equation, we can decompose this regret into the expectation of a martingale difference sequence, and the difference of the next step value functions in the sampled and true MDPs. Specifically, $BR(T) = \mathbb{E}\left[\sum_{l=1}^{\tau}\sum_{h=1}^{H} \mathcal{T}^{\pi_l}_{\theta_n,h}(V^{\theta_n}_{\pi_l,h+1})(s_{l,h}) - \mathcal{T}^{\pi_l}_{\theta^*,h}(V^{\theta^\star}_{\pi_l,h+1})(s_{l,h})\right]$. Here, $\mathcal{T}^{\pi}_{\theta,h}$ denotes the Bellman operator at step h of the episode due to a policy π and MDP θ, and is defined as $\mathcal{T}^{\pi}_{\theta,h}(V^{\theta}_{\pi,h+1})(s_{l,h}) = R(s,\pi(s,h)) + \mathbb{E}_{s,\pi(s,h)}[V|\theta]$.
Step 3: Finally, in the spirit of (Chowdhury et al., 2021), using the transportation inequalities (Boucheron et al., 2003) yields an upper bound of $BR(T)$ as

$$H\sigma_R \mathbb{E}\left[\sum_{l,h=1,1}^{\tau,H} \sqrt{2KL_{s_{l,h},a_{l,h}}(\theta^*,\theta_n)}\right] \leq B_R + 2H\sigma_R\sqrt{T\xi(T)}$$

Here, B_R and σ_R^2 are the bounds on mean and variance of rewards for any θ, and $KL_{s_{l,h},a_{l,h}}(\theta^*,\theta_n)$ denotes the KL divergence between $\mathcal{T}_{\theta^*}(\cdot|s_{l,h},a_{l,h})$ and $\mathcal{T}_{\theta}(\cdot|s_{l,h},a_{l,h})$. *The last inequality holds if we can show* that $KL_{s_{l,h},a_{l,h}}(\theta^*,\theta_n)$ is upper bounded by a constant or monotonically increasing polylogarithmic function, say $\xi(n)$, with probability at least $1-\frac{1}{n}$ for any $n > 1$. Thus, *we observe that the core to achieve a sublinear regret with PSRL is the ability to achieve such concentration bounds.*

A Generic Result: Concentration of Bayesian Posterior of LSI Distributions. Now, we prove an interesting and generic result that if the data distribution is isoperimetric and the prior is designed to have enough mass around the true MDP, we can achieve a polylogarithmic KL-divergence concentration rate under the posterior distribution.

Theorem 3. *Let us assume that the true data distribution $f(X \mid \theta^*)$ is α_{θ^*}-LSI, and the prior has non-zero mass in any closed and compact region $\Xi \subset \Theta$ around the true parameter $\theta^* \in \Theta$, and the log-likelihood function $\log f(X \mid \theta)$ is L_X and L_θ Lipschitz in X and θ, respectively. Then, for any $n > 0$ and $\theta \sim \mathbb{P}(\theta|X^n)$, we obtain with probability at least $1-2\delta$ (for $\delta \in (0,1/2]$)*

$$\begin{aligned} &KL(f(X|\theta^*) \parallel f(X|\theta)) \\ &\leq\ 2L_x\sqrt{\frac{1}{n\alpha_{\theta^*}}\ln\left(\frac{2}{\delta}\right)} + \frac{1}{n}\ln\left(\frac{1}{\delta}\right). \end{aligned} \tag{4}$$

The detailed proof is in Appendix C.1. This result shows a generic concentration of the posteriors for isoperimetric data distributions obeying LSI, and thus demands wider interest for any Bayesian learning framework.

Remark 1 (Lipschitz log-likelihood)**.** *To prove this result, we need additional assumption of Lipschitzness of the log-likelihood function with respect to the data and the parameter. This holds true for most of the parameteric distributions without a Dirac distribution in it. For example, general exponential family distributions satisfy this for their natural parameters and sufficient statistics over data (Efron, 2022). Bilinear exponential families (Ouhamma et al., 2022) and any location-scale family (Ferguson, 1962) of distributions also do the same for their natural parameters and sufficient statistics.*

Remark 2 (Prior Design)**.** *Designing priors that have a small ball probability around the true parameter is encountered often in Bayesian learning. For example, Castillo et al. (2015) proposes such a prior for efficient Bayesian learning, and Chakraborty et al. (2023) uses it carefully to ensure provably near-optimal learning in high-dimensional bandits.*

Sublinear Regret of PSRL. Now, we apply this result to PSRL and bound its Bayesian regret.

Lemma 1. *Under the conditions of Theorem 3 and the mean reward for the MDPs satisfying $|\bar{R}_M(s)| \leq B_R \forall s$, Bayesian regret of PSRL satisfies*

$$BR(T) = \tilde{\mathcal{O}}\left(B_R + H\sigma_R\sqrt{T} + H\sigma_R\left(\frac{L_x^2}{\alpha_{\theta^*}}\right)^{1/4} T^{3/4}\right).$$

The discussion earlier in this section are fleshed out in a detailed proof in Appendix C.1. The third term of $\mathcal{O}(T^{3/4})$ in the Bayesian regret would reduce to $\sqrt{T}$ if the n-fold data distribution $f(X^{(n)}|\theta^\star)$ satisfies LSI with increasing constants. We formalise this observation now.

Near-optimal Regret: Linear LSI Constant of Posteriors. From (Chowdhury and Gopalan, 2019), we observe that if we assume the next step value functions $\mathcal{T}^{\pi}_{\theta,h}(V^{\theta}_{\pi,h+1})(s_{l,h})$ are mean-Lipschitz with respect to the state distributions, we can obtain an alternative approach to prove sublinear regret of PSRL. This results holds if additionally the rewards are bounded and Lipschitz, and the transitions are Lipschitz. Under this condition, we get the following results.

Theorem 4. *If the posterior distributions for mean rewards and transitions separately satisfy LSI with constants $\{\alpha_{\bar{R},l}\}$ and $\{\alpha_{\bar{\mathcal{T}},l}\}$, the mean reward for any MDP M $|\bar{R}_M(s)| \leq B_R \forall s$, the one step value function is Lipschitz in the state with parameter L_{M_*} as Assumption 2, and the mean reward and mean transitions are $L_{\bar{R}}$ and $L_{\bar{\mathcal{T}}}$ Lipschitz in θ, Bayesian regret of PSRL is upper bounded by*

$$\tilde{\mathcal{O}}\left(H\left(\sum_{l=1}^{\tau}\frac{L_{\bar{R}}}{\sqrt{\alpha_{\bar{R},l}}} + \mathbb{E}[L_{M_*}]\sqrt{d}\sum_{l=1}^{\tau}\frac{L_{\bar{\mathcal{T}}}}{\sqrt{\alpha_{\bar{\mathcal{T}},l}}}\right)\right).$$

The proof is in the Appendix C.2 and follows from the sub-Gaussian concentration under LSI (Equation (3)). It implies that PSRL achieves $BR(T) = \widetilde{O}\left(\sqrt{T}\right)$ if $\alpha_l = \Omega(T)$. In Section 5, we show that this holds for different families of distributions studied in literature.

4 LAPSRL FOR APPROXIMATE POSTERIORS

We know that constant approximation error for posteriors leads to linear regret in the context of Thompson sampling for multi-armed bandits (Phan et al., 2019). This also happens in other reinforcement learning setups, like contextual bandits and MDPs. Previous work has noted (Mazumdar et al., 2020) that proper decay of this error can allow for sublinear regret in multi armed bandits. In the work of Mazumdar et al.; Zheng et al.; Karbasi et al., they designed an approximate algorithm for multi armed bandits and strongly log-concave posteriors. However, only Karbasi et al. (2023) include planning over episodes, which is essential for MDPs, and the regret analysis heavily depended on strong log-concavity assumptions. We aim to show that this philosophy of constructing approximate posteriors with proper concentration rates can be applied also to MDPs and with only an isoperimetric assumption (LSI in Definition 1) instead of log-concavity. To start, we derive Theorem 5 that shows how can we control the error rate of concentration of posteriors in RL.

Algorithm 1 Langevin PSRL (LaPSRL)

Input: Likelihood $f(x|\theta)$, Prior $\mathbb{P}(\theta)$, Horizon H, total episodes τ, Regret order $g(H, \mathcal{S}, \mathcal{A})$
for $l = 1 : \tau$ **do**
  $\epsilon_{\text{post},l} = \frac{g(H,\mathcal{S},\mathcal{A})}{l\Delta^2_{\text{max}}}$
  if Chained sampling **then**
    $\rho_0 = \theta_{l-1}$ # Reuse last sample from previous iteration
  else
    $\rho_0 \sim \mathbb{P}(\theta)$ # Resample from prior
  end if
  Sample $\theta_l = \text{LANGEVIN SAMPLE}(f(x \mid \theta), \mathbb{P}(\theta), \mathcal{H}_l, \epsilon_{\text{post},l}, \rho_0)$
  Play $\pi(\theta_l)$ until horizon H obtaining data $\mathcal{H}_{l+1} = \mathcal{H}_l \cup \{z_i\}_{i=H(l-1)}^{Hl}$.
end for

Algorithm 2 LANGEVIN SAMPLE

Input: Likelihood $f(x|\theta)$, Prior $\mathbb{P}(\theta)$, data $\mathcal{H}_l$, acceptable error $\epsilon_{\text{post},l}$, initial sample ρ_0.
Set $\eta_l = \min\left(\frac{\alpha_l}{16\sqrt{2}L^2(H(l-1))^{3/2}}, \frac{3\alpha_l\epsilon_{\text{post},l}}{320dL^2H(l-1)}\right)$ # Learning rate
Set $k_l = \frac{\gamma}{\alpha_l\eta}\log\frac{2\text{KL}(\rho_0 \parallel P(\theta|\mathcal{H}_l))}{\epsilon_{\text{post},l}}$ # Steps
return $\theta = \text{SARAH-LD}\left(f(x|\theta), \mathcal{H}_l, \mathbb{P}(\theta), k_l, \eta_l\right)$

Theorem 5. *Let the policy the start of episode l act by sampling a model from Q_l where $\min(KL((\mathbb{P}_l \parallel Q_l), KL(Q_l \parallel \mathbb{P}_l)) \leq \epsilon_{post,l}$ and where $\mathbb{P}_l$ is the exact posterior at start of episode l. Then the incurred regret from planning with an approximate posterior bounded by $\sqrt{2}\Delta_{max}\sqrt{\epsilon_{post,l}}$.*

The result comes from the fact that KL divergence of posterior controls the growth of Bayesian regret. The detailed proof is in Appendix D.

Corollary 1. *If a policy incurs $\widetilde{\mathcal{O}}(\sqrt{T}g(H, \mathcal{S}, \mathcal{A}))$ regret under distribution P, for some function g, it will incur the same order of regret under Q if $0 \leq \epsilon_{post,l} \leq C\frac{g(H,\mathcal{S},\mathcal{A})^2}{l\Delta^2_{max}}$ for some constant $C \geq 0$.*

Thus, Corollary 1 states that if the approximation error of the posterior distribution decays linearly with the number of episodes (l), then we can achieve $\widetilde{O}\left(\sqrt{T}\right)$ regret by running PSRL with such posteriors.

LaPSRL. With these results in mind, we design an algorithm, Langevin PSRL (LaPSRL). The algorithm can be seen in Algorithm 1 with its sampling routine in Algorithm 2. The algorithm works similarly to PSRL. In each episode l, a tolerable error $\epsilon_{\text{post},l}$ is calculated. Then we use SARAH-LD to sample a θ_l. Depending on the task at hand, SARAH-LD calculates the required step size and learning rate to reach the acceptable error in KL distance, returning the desired sample. This sample is used to obtain an optimal policy which is then played for the episode. We have two options for initializing the sampling in each episode, from some prior or taking the previous sample.

Gradient Complexity. By combining Theorem 5 with log-Sobolev theory and SARAH-LD we obtain, for any log-Sobolev posterior, order optimal Bayesian regret while still limiting the computational gradient complexity of each episode to a quadratic polynomial. Here gradient complexity signifies the amount of gradients $\nabla_\theta f(x_i)$ that need to be performed.

Corollary 2. *For a posterior fulfilling the Assumption 1 and definition 1, a posterior sampling style algorithm can obtain an unchanged order of regret under SARAH-LD sampling under a gradient complexity for episode l of gradient complexity $= \widetilde{O}\left(\frac{H^3l^3L^2}{\alpha_l^2} + \frac{dH^{2.5}l^{3.5}L^2}{\alpha_l^2 g(H,\mathcal{S},\mathcal{A})^2}\right)$ In many cases, as seen in Section 5, $\alpha_l = \widetilde{\Omega}(Hl)$. Similarly, as in Theorem 4, $g(H, \mathcal{S}, \mathcal{A}) \propto H$. This then becomes gradient complexity $\propto \widetilde{O}\left(HlL^2 + \frac{dl^{3/2}L^2}{H^{3/2}}\right)$.*

Chained samples. The sample complexity for a ϵ approximation of ν is controlled by initial distribution ρ_0 with $\text{KL}(\rho_0 \parallel \nu)$. The naive approach is having ρ_0 from a prior such as an isotropic Gaussian, the dependence is only logarithmic in $\text{KL}(\rho_0 \parallel \nu)$ which also does not grow very fast. An alternative is to use the final sample from the previous time step as initialization for the next one, this also allows for a more practical algorithm as it might be easier to estimate the divergence between the two sequential posteriors than between the prior and the posterior. We show that reusing samples bounds the KL distance to a function of the variance of θ.

Theorem 6. *Let $\rho_*^l(\theta)$ be the final sample of the Langevin algorithm at episode l, approximating the true posterior $\mathbb{P}(\theta \mid \mathcal{H}_l)$ with $KL(\rho_*^l(\theta) \parallel \mathbb{P}(\theta \mid \mathcal{H}_l)) \leq \epsilon_{post,l}$. Additionally, if $\nabla_z \log f(z|\theta)$ is L_z-Lipschitz and α_z-Log Sobolev, we get $\mathbb{E}_{f(z|\mathcal{H}_l)} KL(\rho_*^l(\theta) \parallel \mathbb{P}(\theta \mid \mathcal{H}_{l+1})) \leq \epsilon_{post,l} + \frac{L_z}{2\alpha_z} Var_{\rho_*^l(\theta)}(\theta)$, where $Var_{\rho_*^l(\theta)}(\theta)$ is the variance of the approximate posterior distribution $\rho_*^l(\theta)$.*

Appendix D contains the detailed proof. Chaining the samples will lead to correlations between the sampled parameters. While this could be problematic in some cases, since Bayesian regret is taken in expectation, this does not affect the order of the regret. One problem is that the variance is taken under the approximate distribution $\rho_*^l(\theta)$. But in practice, we know that this is an $\epsilon_{\text{post},l}$ close approximation. We also know that the variance of the posterior distributions tends to decay as more data is observed, meaning that this $\text{KL}(\rho_*^l(\theta) \parallel \mathbb{P}(\theta \mid \mathcal{H}_{l+1}))$ will decay. This is unlike the naive sampling from a prior, which will increase.

5 PSRL FOR FAMILIES OF DISTRIBUTIONS

In this section, we study a variety of log-Sobolev distributions. We show their log-Sobolev constants and ultimately apply Theorem 4 to calculate the Bayesian regret of PSRL for such posteriors.

Univariate Gaussian. For illustrative purposes, we calculate the relevant constants for a Gaussian posterior with known variance σ^2. Here we also assume a Gaussian $(0, \sigma_0^2)$ prior over the mean μ. We have $\mathbb{P}(\mu|\mathcal{H}_n) \propto \exp\{-\left(\sum_{i=1}^n \left(\frac{\mu^2}{2n\sigma_0{}^2} + \frac{(\mu-x_i)^2}{2\sigma^2}\right)\right)\}$. Then, we have $\gamma = n$, $f_i(\mu) = \left(\frac{\mu^2}{2n\sigma_0{}^2} + \frac{(\mu-x_i)^2}{2\sigma^2}\right)$, since $\nabla_\mu^2 f_i(\mu) = \frac{1}{n\sigma_0^2} + \frac{1}{\sigma^2} \leq L$. Finally, we can use Theorem 1 to calculate α. Since $||\nabla_\mu^2 f_i(\mu)||$ is independent of i in this case, we can see that $\nabla_\mu^2 - \log \mathbb{P}(\mu|\mathcal{H}_n) = \nabla_\mu^2 \sum_{i=1}^n f_i(\mu) = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$, which gives $\alpha = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} = L\gamma$.

Corollary 3. *PSRL obtains $\mathbb{E}[Regret(T)] = \widetilde{O}\left(\sqrt{T\sigma^2}\right)$ with univariate Gaussian posteriors.*

Mixture Distributions. There has been multiple work looking into log-Sobolev constants for mixtures of log-Sobolev distributions (Koehler and Vuong, 2024; Chen et al., 2021b; Schlichting, 2019). Generally, it depends the on constants of the mixture components as well as the distance between the components.

Theorem 7 (Informally from Theorem 2 (Koehler and Vuong, 2024))**.** *For k-mixture components $\mu = \sum_{i=1}^k p_i \mu_i, \sum_{i=1}^k p_i = 1$, where there is some overlap δ between components, has $\alpha_{Mixture} \geq \frac{\delta \min p_i \min \alpha_i}{4k(1-\log(\min p_i))}$.*

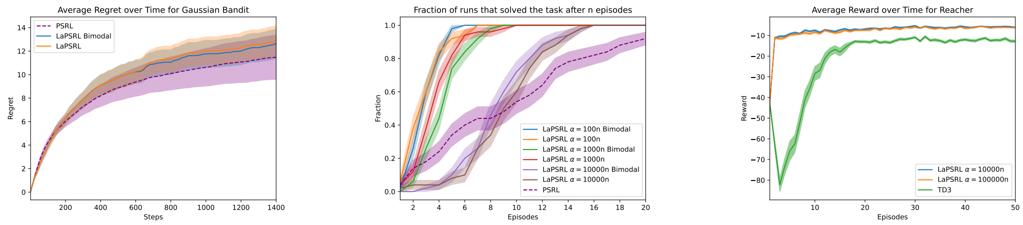
The overlap factor δ relates to integral over the minimum of the paired components, see (Koehler and Vuong, 2024) for more details. If the components are posteriors, this δ should go to 1 as the individual posteriors observe more data and converge.

Log-concave and Mixture of Log-concave Distributions.

Theorem 8. *Any log-concave posterior has $\alpha_l = \Theta(n)$. Any posterior that is a mixture of log-concave distributions has $\alpha_{Mixture} = \Omega\left(\frac{n \min p_i}{4k(1-\log(\min p_i))}\right)$.*

This result comes from the superadditivity of minimum eigenvalues of Hessians and therefore LSI constants for log-concave distributions. The mixture result follows from Theorem 7. A proof of the theorem is in Appendix E. Combining Theorem 4 and Theorem 8 we obtain the following corollary

Corollary 4. *Any log-concave posterior $|\bar{R}_M(s)| \leq B_R \forall s$ for all MDPs M will have $\mathbb{E}[Regret(T)] = \widetilde{O}\left(\sqrt{T}\right)$ for PSRL. Similarly, and under the same condition, any posterior that is a mixture of log-concave posteriors with sufficient overlap will obtain $\mathbb{E}[Regret(T)] = \widetilde{O}\left(\sqrt{\frac{4kT}{\min p_i}}\right)$ PSRL regret.*



(a) Gaussian Bandit (b) Cartpole (c) Reacher

Figure 2: We compare LaPSRL versus baselines. In the bandit and Cartpole experiments we benchmark with PSRL, in Reacher with PPO and TD3. On the left we compare the expected regret for a Gaussian bandit algorithm. In the middle we compare how many episodes it takes to solve a Cartpole task. On the right we study the average regret per episode in the Reacher environement. In all environments, we average over 50 independent runs with the standard error highlighted around the average. The plots are included full size in the appendix.

General log-Sobolev distributions.

Theorem 9. *Any log-Sobolev distribution with a likelihood ratio* $|\log \frac{f(X|\theta)}{f(X|\theta')}| \leq \Gamma$ *has a log-Sobolev posterior.*

This result is interesting because it means that a very wide family of settings have log-Sobolev posteriors. Unfortunately, we have been unable to prove that the log-Sobolev constant of a posterior, under some suitable conditions, will always scale as $\Omega(n)$. Although we conjecture that this is possible and this also matches the intuition from the asymptotic results of the Bernsteinvon Mises theorem which gives a log-Sobolev constant of $\Theta(n)$ as $n \to \infty$. Instead, in these settings, we can rely on the results in Lemma 1 to obtain the desired sub linear regret.

6 EXPERIMENTAL ANALYSIS

We run a set of experiments on two environments to verify that the LaPSRL is competitive. The goal of this section is to answer the following questions:

Can LaPSRL work for varying domains and settings?
Is LaPSRL efficient in terms of performance?

To demonstrate this we perform a set of experiments. First, we deploy LaPSRL on a Gaussian multi-armed bandit task with two arms. Second, we perform experiments with a LQR (Kalman, 1960) setup on the Cartpole environment (Barto et al., 1983) and finally with a neural network on the Reacher environment (Towers et al., 2024). For Cartpole and Bandits, we also use a LaPSRL algorithm that has a bimodal prior over the arms to demonstrate a non log-concave setting. Additional experimental details are in the appendix.

Gaussian batched bandits We use LaPSRL on a Gaussian multi-armed bandit task with two arms. To preserve computations, we use a batched approach, such that each action is taken 20 times each time it is sampled. As a baseline, we compare with the performance of PSRL from the true posterior. The results can be seen in Figure 2(a) where we plot the expected regret. We can see that both LaPSRL algorithms perform similarly, and only are only slightly worse than the theoretical priors. This is in line with the theoretical results of order optimality.

Continuous MDPs. We evaluate LaPSRL on two continuous environments, Cartpole and Reacher. For both experiments, we try different settings for α (details in appendix). The Cartpole environment (Barto et al., 1983) is modified to have continuous actions. Here we use a Linear Quadratic Regulator model. As a baseline we compare with an exact PSRL algorithm which samples from Bayesian linear regression priors (Minka, 2000). The results from this experiment can be found in Figure 2(b) where we plot what fraction of the 50 runs have solved the task (i.e. taking 200 steps without failing). Here we see that all versions successfully handle the task, even faster than the PSRL baseline. We can note that it takes longer for the experiments with larger α values to converge. The slower convergence of PSRL can be due to the priors not being the same.

The Reacher environment is a standard environment from the Gymnasium environments (Towers et al., 2024) Here we use a neural network model, this means that this is not necessarily log-Sobolev, but we wish to show that this approach still is useful. Here, we benchmark with TD3 (Fujimoto et al., 2018). We see in Figure 2(c) that LaPSRL learns very quickly, while TD3 is unable to learn a policy that is equally good.

Results and Discussions. To conclude, we find that in all three experiments, LaPSRL works well. These settings are quite different, and tell us that LaPSRL works efficiently in very different settings, supporting the claim that LaPSRL can work for varying domains and settings and with good performance.

7 RELATED WORKS

In addition to the work discussed before, we present an overview of the related works with approximate posteriors. Approximations are often required, either because calculating or sampling from the posterior is intractable (Wang et al., 2023; Sasso et al., 2023; Osband et al., 2023). While these papers have frameworks for approximate sampling, none of them comes with any regret guarantees.[1] Fan and Ming (2021) also study the case of function approximation, but the theory does not hold there. The work of Huang et al. (2023) has an approximate upper confidence bound algorithm which Bayesian regret bounds in the bandit setting.

In addition to the previously mentioned work, there has been a surge of recent work looking into the use of Langevin methods for bandits and reinforcement learning (Kim, 2023; Dwaracherla and Van Roy, 2020; Yamamoto et al., 2024), but this work comes without any theoretical guarantees. In Nguyen-Tang et al. (2024) they use Langevin for offline RL and in (Kuang et al., 2023) it is for linear MDPs. The work of Karbasi et al. (2023) also tries to tackle a similar problem as this paper, using Langevin dynamics for order optimal regret. An important difference is that they are limited to strongly log-concave distributions and to tabular MDPs, while we are much more general. Similarly, Hsu et al. (2024) obtain regret bounds for Multi-agent RL, but they assume linear function approximations. Also, concurrent work using Langevin methods are appearing, such as TS for bandits in (Zheng et al., 2024), but with requirements on convexity and (Kim et al., 2024) with Langevin for LQR but with strongly log-concave assumptions. Finally, (Kuang et al., 2023) uses these ideas for delayed feedback RL, but limited to Linear MDPs and Krishnamurthy and Yin (2021) uses Langevin dynamics for inverse reinforcement learning.

8 DISCUSSION & FUTURE WORKS

In this paper, we aim to understand whether we can design algorithms with sublinear regret for any isoperimetric distribution. We specifically study PSRL type algorithms for posteriors satisfying log-Sobolev inequalities. We show that if we can compute exact posteriors and sample from them, PSRL can achieve $\widetilde{O}(\sqrt{T})$ regret in an episodic MDP under log-Sobolev and some additional mild assumptions, this extends the setting where such results exist. We further design a generic Langevin sampling based extension of PSRL, namely LaPSRL. We show that LaPSRL also achieves $\widetilde{O}(\sqrt{T})$ in these settings. We plug-in SARAH-LD as the Langevin sampling algorithm, and derive upper bounds on the required gradient complexity and chained sample complexity. Finally, we test LaPSRL in bandit and continuous MDP environments. We show that the variants of LaPSRL perform competitively with respect to baselines in all these settings. In future, it will be interesting to extend LaPSRL's analysis to neural tangent kernel's yielding a better understanding of deep RL.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and the Norwegian Research Council Project "Algorithms and Models for Socially Beneficial AI". D. Basu acknowledges the Inria-Kyoto University Associate Team "RELIANT" for supporting the project, the CHIST-ERA CausalXRL project, the ANR JCJC for the REPUBLIC project (ANR-22-CE23-0003-01), and the PEPR project FOUNDRY (ANR23-PEIA-0003). The computations were enabled by resources

[1] It is worth noting that model sampling using subsamples does enjoy some theoretical properties.

provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725. The authors also wish to thank Hannes Eriksson for his assistance.

References

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR. (page 3)

Bakry, D., Gentil, I., Ledoux, M., et al. (2014). *Analysis and geometry of Markov diffusion operators*, volume 103. Springer. (page 2, 4)

Barthe, F. and Kolesnikov, A. V. (2008). Mass transport and variants of the logarithmic sobolev inequality. *Journal of Geometric Analysis*, 18(4):921–979. (page 4)

Barto, A. G., Sutton, R. S., and Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):834–846. (page 9, 24, 25)

Bizeul, P. (2023). On the log-sobolev constant of log-concave measures. (page 4)

Boucheron, S., Lugosi, G., and Bousquet, O. (2003). Concentration inequalities. In *Summer school on machine learning*, pages 208–240. Springer. (page 5, 19)

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. (2018). JAX: composable transformations of Python+NumPy programs. (page 25)

Castillo, I., Schmidt-Hieber, J., and Van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, pages 1986–2018. (page 6)

Cattiaux, P., Guillin, A., and Wu, L.-M. (2010). A note on talagrands transportation inequality and logarithmic sobolev inequality. *Probability theory and related fields*, 148:285–304. (page 4)

Chafaï, D. and Lehec, J. (2023). Logarithmic sobolev inequalities essentials. Accessed on 08/10/2024. (page 4)

Chakraborty, S., Roy, S., and Tewari, A. (2023). Thompson sampling for high-dimensional sparse linear contextual bandits. In *International Conference on Machine Learning*, pages 3979–4008. PMLR. (page 6)

Chen, H.-B., Chewi, S., and Niles-Weed, J. (2021a). Dimension-free log-sobolev inequalities for mixture distributions. *Journal of Functional Analysis*, 281(11):109236. (page 4)

Chen, H.-B., Chewi, S., and Niles-Weed, J. (2021b). Dimension-free log-sobolev inequalities for mixture distributions. (page 8)

Chowdhury, S. R. and Gopalan, A. (2019). Online learning in kernelized markov decision processes. (page 1, 5, 6, 19, 20)

Chowdhury, S. R., Gopalan, A., and Maillard, O.-A. (2021). Reinforcement learning in parametric mdps with exponential families. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1855–1863. PMLR. (page 5)

Dwaracherla, V. and Van Roy, B. (2020). Langevin dqn. *arXiv preprint arXiv:2002.07282*. (page 10)

Efron, B. (2022). *Exponential families in theory and practice*. Cambridge University Press. (page 6)

Fan, Y. and Ming, Y. (2021). Model-based reinforcement learning for continuous control with posterior sampling. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3078–3087. PMLR. (page 10)

Ferguson, T. S. (1962). Location and scale parameters in exponential families of distributions. *The Annals of mathematical statistics*, 33(3):986–1001. (page 6)

Fujimoto, S., Hoof, H., and Meger, D. (2018). Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR. (page 10, 25)

Geramifard, A., Walsh, T. J., Tellex, S., Chowdhary, G., Roy, N., and How, J. P. (2013). A tutorial on linear function approximators for dynamic programming and reinforcement learning. *Foundations and Trendső in Machine Learning*, 6(4):375–451. (page 1)

Holley, R. and Stroock, D. (1987). Logarithmic sobolev inequalities and stochastic ising models. *Journal of Statistical Physics*, 46(5-6):1159–1194. (page 4)

Hsu, H.-L., Wang, W., Pajic, M., and Xu, P. (2024). Randomized exploration in cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:2404.10728*. (page 10)

Huang, S., Dossa, R. F. J., Ye, C., Braga, J., Chakraborty, D., Mehta, K., and Araújo, J. G. (2022). Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms. *Journal of Machine Learning Research*, 23(274):1–18. (page 25)

Huang, Z., Lam, H., Meisami, A., and Zhang, H. (2023). Optimal regret is achievable with bounded approximate inference error: An enhanced bayesian upper confidence bound framework. In *Thirty-seventh Conference on Neural Information Processing Systems*. (page 10)

Huix, T., Zhang, M., and Durmus, A. (2023). Tight regret and complexity bounds for thompson sampling via langevin monte carlo. In Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 8749–8770. PMLR. (page 2)

Ishfaq, H., Lan, Q., Xu, P., Mahmood, A. R., Precup, D., Anandkumar, A., and Azizzadenesheli, K. (2023). Provable and practical: Efficient exploration in reinforcement learning via langevin monte carlo. (page 2)

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. (2019). A short note on concentration inequalities for random vectors with subgaussian norm. (page 17)

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. (page 9)

Karbasi, A., Kuang, N. L., Ma, Y., and Mitra, S. (2023). Langevin thompson sampling with logarithmic communication: Bandits and reinforcement learning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15828–15860. PMLR. (page 2, 6, 10)

Kim, G. (2023). Learning linear-quadratic regulators via thompson sampling with preconditioned langevin dynamics. (page 10)

Kim, Y., Kim, G., and Yang, I. (2024). Approximate thompson sampling for learning linear quadratic regulators with $o(\sqrt{\{T\}})$ regret. *arXiv preprint arXiv:2405.19380*. (page 10)

Kinoshita, Y. and Suzuki, T. (2022). Improved convergence rate of stochastic gradient langevin dynamics with variance reduction and its application to optimization. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*. (page 4, 5, 15, 16)

Koehler, F., Heckett, A., and Risteski, A. (2023). Statistical efficiency of score matching: The view from isoperimetry. In *The Eleventh International Conference on Learning Representations*. (page 4)

Koehler, F. and Vuong, T.-D. (2024). Sampling multimodal distributions with the vanilla score: Benefits of data-based initialization. In *The Twelfth International Conference on Learning Representations*. (page 8)

Krishnamurthy, V. and Yin, G. (2021). Langevin dynamics for adaptive inverse reinforcement learning of stochastic gradient algorithms. *Journal of Machine Learning Research*, 22(121):1–49. (page 10)

Kuang, N. L., Yin, M., Wang, M., Wang, Y.-X., and Ma, Y. (2023). Posterior sampling with delayed feedback for reinforcement learning with linear function approximation. In *Thirty-seventh Conference on Neural Information Processing Systems*. (page 10)

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press. (page 3)

Ledoux, M. (2006). Concentration of measure and logarithmic sobolev inequalities. In *Seminaire de probabilites XXXIII*, pages 120–216. Springer. (page 2, 4)

Mazumdar, E., Pacchiano, A., Ma, Y., Jordan, M., and Bartlett, P. (2020). On approximate thompson sampling with langevin algorithms. In *International Conference on Machine Learning*, pages 6797–6807. PMLR. (page 2, 6)

Minka, T. (2000). Bayesian linear regression. Technical report, Citeseer. (page 9, 24)

Nguyen-Tang, T., Yin, M., Uehara, M., Wang, Y.-X., Wang, M., and Arora, R. (2024). Posterior sampling via langevin monte carlo for offline reinforcement learning. (page 10)

Nitanda, A., Wu, D., and Suzuki, T. (2022). Convex analysis of the mean field langevin dynamics. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 9741–9757. PMLR. (page 1)

Osband, I., Russo, D., and Van Roy, B. (2013). (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26. (page 2, 3, 5, 15, 18)

Osband, I. and Van Roy, B. (2014). Model-based reinforcement learning and the eluder dimension. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc. (page 20)

Osband, I. and Van Roy, B. (2017). Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, pages 2701–2710. PMLR. (page 1, 5)

Osband, I., Wen, Z., Asghari, S. M., Dwaracherla, V., Ibrahimi, M., Lu, X., and Van Roy, B. (2023). Approximate Thompson sampling via epistemic neural networks. In Evans, R. J. and Shpitser, I., editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 1586–1595. PMLR. (page 10)

Ouhamma, R., Basu, D., and Maillard, O.-A. (2022). Bilinear exponential family of mdps: Frequentist regret bound with tractable exploration and planning. *arXiv preprint arXiv:2210.02087*. (page 1, 6)

Phan, M., Abbasi Yadkori, Y., and Domke, J. (2019). Thompson sampling and approximate inference. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. (page 6)

Pinneri, C., Sawant, S., Blaes, S., Achterhold, J., Stueckler, J., Rolinek, M., and Martius, G. (2020). Sample-efficient cross-entropy method for real-time planning. In *Conference on Robot Learning 2020*. (page 24)

Russo, D., Roy, B. V., Kazerouni, A., Osband, I., and Wen, Z. (2020). A tutorial on thompson sampling. (page 2)

Sasso, R., Conserva, M., and Rauber, P. (2023). Posterior sampling for deep reinforcement learning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 30042–30061. PMLR. (page 10)

Schlichting, A. (2019). Poincaré and log–sobolev inequalities for mixtures. *Entropy*, 21(1):89. (page 8)

Steiner, C. (2021). A feynman-kac approach for logarithmic sobolev inequalities. (page 4)

Stroock, D. W. and Zegarlinski, B. (1992). The equivalence of the logarithmic sobolev inequality and the dobrushin-shlosman mixing condition. *Communications in mathematical physics*, 144:303–323. (page 1)

Thompson, W. (1933). On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of two Samples. *Biometrika*, 25(3-4):285–294. (page 2, 3, 15)

Todorov, E., Erez, T., and Tassa, Y. (2012). Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE. (page 24, 25)

Towers, M., Kwiatkowski, A., Terry, J., Balis, J. U., De Cola, G., Deleu, T., Goulão, M., Kallinteris, A., Krimmel, M., KG, A., et al. (2024). Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*. (page 9, 10, 24, 25)

Vempala, S. and Wibisono, A. (2019). Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. (page 2, 4)

Wang, C., Chen, Y., and Murphy, K. P. (2023). Model-based policy optimization under approximate bayesian inference. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*. (page 10)

Wang, F.-Y. (2001). Logarithmic sobolev inequalities: conditions and counterexamples. *Journal of Operator Theory*, pages 183–197. (page 4)

Xu, P., Zheng, H., Mazumdar, E. V., Azizzadenesheli, K., and Anandkumar, A. (2022). Langevin Monte Carlo for contextual bandits. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 24830–24850. PMLR. (page 2)

Yamamoto, K., Oko, K., Yang, Z., and Suzuki, T. (2024). Mean field langevin actor-critic: Faster convergence and global optimality beyond lazy learning. In *Forty-first International Conference on Machine Learning*. (page 10)

Zheng, H., Deng, W., Moya, C., and Lin, G. (2024). Accelerating approximate Thompson sampling with underdamped Langevin Monte Carlo. In Dasgupta, S., Mandt, S., and Li, Y., editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 2611–2619. PMLR. (page 2, 6, 10)

A NOTATION

Table 1: Table of notations.

$a_{l,h}$	Action in timestep h of episode l
$s_{l,h}$	State in timestep h of episode l.
l	Episode index.
τ	Total number of episodes.
h	Current step in episode.
H	Horizon, amount of steps in an episode.
T	Total amount of agent interactions, $T = \tau H$
n	Total amount of agent interactions so far.
$z_{l,h}$	$(s_{l,h}, a_{l,h}, s_{l,h+1})$
$\mathcal{H}_l$	The states, actions and transitions observed until start of episode l.
M_l	Sampled MDP in episode l.
M_*	True MDP.
V_M^π	Value of policy π in MDP M.
$BR(T)$	Bayesian regret, Equation (1)
$\Delta_{\max}$	Maximal possible regret for an episode, $\max_\pi V_{\pi,1}^{M_*}(s_1) - \min_\pi V_{\pi,1}^{M_*}(s_1)$
B_R	Upper bound on absolute value of average reward in Theorem 4.
α	Log-Sobolev constant.
$\alpha_{\bar{\mathcal{T}},l}$	Log-Sobolev constant for the average transitions in episode l.
$\alpha_{\bar{R},l}$	Log-Sobolev constant for the average rewards in episode l.
α_{θ^*}	Log-Sobolev constant for the data distribution $f(X \mid \theta^*)$.
L,	L-smooth constant of the log likelihood function.
γ	Temperature parameter in Langevin dynamics, $\gamma = n$ in Bayesian posterior setting.
$\epsilon_{\text{post},l}$	Langevin sampling error
L_{M_*}	Mean-Lipschitz parameter from Assumption 2
$L_{\bar{R}}, L_{\bar{\mathcal{T}}}$	Lipschitz parameters for mean transition and reward.
$\mathbb{P}(\cdot), \mathbb{P}(\cdot\|x)$	prior, posterior on the parameters.
$f(X \mid \theta)$	Likelihood of data X in model θ.
$\mathbb{E}$	Expectation
KL$(P \parallel Q)$	KullbackLeibler divergence between P and Q.
$\nu(\Xi)$	Lower bound on prior mass from Equation (17).

B ALGORITHMIC DETAILS: PSRL AND SARH-LD

For completeness, we describe the pseudocodes of PSRL (Osband et al., 2013) and SARAH-LD (Kinoshita and Suzuki, 2022) algorithms in Algorithm 3 and Algorithm 4 as well as a theorem on the gradient complexity of SARAH-LD in Theorem 10. We slightly modify SARAH-LD to use in LaPSRL under Bayesian posterior setting.

B.1 PSRL

PSRL, found in Algorithm 3, comes from the idea of probability matching of Thompson (1933). In each episode, we sample an MDP M_l, parameterized by θ_l from the posterior, and play the optimal policy for that MDP$\pi^*(\theta_l)$ for an entire episode. At the end of the episode, we collect all the data, updating the posterior before starting again.

B.2 SARAH-LD

SARAH-LD, found in Algorithm 4, is a Langevin algorithm by Kinoshita and Suzuki (2022) that utilizes variance reduction techniques for reduced gradient complexity. In each epoch, it performs a gradient step on the full dataset, before taking smaller mini-batches for efficiency. In each parameter update, noise is added to the gradient in line with the Langevin dynamics. In the inner loop of mini batches, a difference of batches is done to reduce the variance of the estimate. At the end of the

Algorithm 3 PSRL

Input: Likelihood $f(x|\theta)$, Prior $\mathbb{P}(\theta)$
for $l = 1 : \tau$ **do**
  Sample $\theta_l \sim \mathbb{P}(\theta \mid \mathcal{H}_l)$
  Play $\pi^*(\theta_l)$ until horizon H obtaining data $\{x_i\}_{i=H(l-1)}^{Hl}$.
  $\mathcal{H}_{t+1} = \mathcal{H}_l \cup \{x_i\}_{i=H(l-1)}^{Hl}$
end for

algorithm, it returns a sample, which is bounded in distance from its target distribution. This can be seen more formally in the following theorem.

Algorithm 4 SARAH-LD

Input: step size $\eta > 0$, batch size B, epoch length m, data X, likelihood f, prior $\mathbb{P}(\theta)$, initial sample ρ_0.
Initialization: $\theta_0 = \rho_0, \theta^{(0)} = \theta_0$
for $s = 0, 1, \ldots, (K/m)$ **do**
  $v_{sm} = \nabla F\left(\theta^{(s)}\right)$
  randomly draw $\epsilon \sim N\left(0, I_{d\times d}\right)$
  $\theta_{sm+1} = \theta_{sm} - \eta v_{sm} + \sqrt{2\eta/n}\epsilon$
  for $l = 1, \ldots, m-1$ **do**
    $k = sm + l$
    randomly pick a subset I_k from $\{1, \ldots, n\}$ of size $|I_k| = B$
    randomly draw $\epsilon \sim N\left(0, I_{d\times d}\right)$
    $v_k = \frac{1}{B}\sum_{i_k \in I_k}\left(\nabla(f\left(X_{i_k} \mid \theta_k\right) + 1/n\mathbb{P}(\theta_k)) - \nabla(f\left(X_{i_k} \mid \theta_{k-1}\right) + 1/n\mathbb{P}(\theta_{k-1}))\right) + v_{k-1}$
    $\theta_{k+1} = \theta_k - \eta v_k + \sqrt{2\eta/n}\epsilon$
  end for
  $\theta^{(s+1)} = \theta_{(s+1)m}$
end for

Theorem 10 (Corollary 2.1 of (Kinoshita and Suzuki, 2022))**.** *Under Assumption 1 and definition 1, for all $\epsilon \geq 0$, if we choose step size η such that $\eta \leq \frac{3\alpha\epsilon}{48\gamma\alpha^2}$, then a precision KL$(\rho_k \parallel \nu) \leq \epsilon$ is reached after $k \geq \frac{\gamma}{\alpha\eta}\log\frac{2KL(\rho_0 \parallel \nu)}{\epsilon}$ steps of SARAH-LD. Especially, if we take $B = m = \sqrt{n}$ and the largest permissible step size $\eta = \min(\frac{\alpha}{16\sqrt{2}L^2\sqrt{n}\gamma}, \frac{3\alpha\epsilon}{320dL^2\gamma})$, then the gradient complexity becomes $\widetilde{O}\left(\left(n + \frac{dn^{\frac{1}{2}}}{\epsilon}\right)\cdot\frac{\gamma^2L^2}{\alpha^2}\right)$.*

C REGRET BOUNDS FOR PSRL WITH EXACT POSTERIORS

C.1 Confidence Intervals for Isoperimetric Data Distributions

Theorem 3. *Let us assume that the true data distribution $f(X \mid \theta^*)$ is α_{θ^*}-LSI, and the prior has non-zero mass in any closed and compact region $\Xi \subset \Theta$ around the true parameter $\theta^* \in \Theta$, and the log-likelihood function $\log f(X \mid \theta)$ is L_X and L_θ Lipschitz in X and θ, respectively. Then, for any $n > 0$ and $\theta \sim \mathbb{P}(\theta|X^n)$, we obtain with probability at least $1 - 2\delta$ (for $\delta \in (0, 1/2]$)*

$$KL(f(X|\theta^*) \parallel f(X|\theta)) \leq 2L_x\sqrt{\frac{1}{n\alpha_{\theta^*}}\ln\left(\frac{2}{\delta}\right)} + \frac{1}{n}\ln\left(\frac{1}{\delta}\right). \quad (4)$$

Proof. Here we will use notation $X^{(n)} = \{X_i\}_0^{n-1}$, to refer to the set of data of size n.

Step 1: From log-Sobolev to log-likelihood concentration. If the data distribution $f(X \mid \theta^*)$ is α_{θ^*} log-Sobolev, we have from property of sub-Gaussian concentration of Lipschitz functions, as seen in Equation (3),

$$\mathbb{P}(|\log f(X \mid \theta) - \mathbb{E}_{X\sim f(X|\theta^*)} \log f(X \mid \theta)| \geq t) \leq 2e^{-\frac{t^2\alpha_{\theta^*}}{L_x^2}} \tag{5}$$

$$\implies |\log f(X \mid \theta) - \mathbb{E}_{X\sim f(X|\theta^*)} \log f(X \mid \theta)| \leq \sqrt{\frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}, \tag{6}$$

with probability at least $1-\delta$.

Due to Hoeffding type bound on sum of conditionally independent sub-Gaussians (Jin et al., 2019), we get

$$|\log f(X^{(n)} \mid \theta) - \mathbb{E}_{X^{(n)}\sim f(X^{(n)}|\theta^*)} \log f(X^{(n)} \mid \theta)| \leq \sqrt{n\frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}, \tag{7}$$

with probability at least $1-\delta$. The final implication comes from the sum of n random variables that are $\frac{L_x^2}{\alpha_{\theta^*}}$ sub-Gaussians being $n\frac{L_x^2}{\alpha_{\theta^*}}$ sub-Gaussian. We denote the event from Equation (7) as $E_{\bar{X}}$.

Step 2: From log-likelihood concentration to bounding probabilities. Note that $\mathbb{E}_{X\sim f(X|\theta^*)} \log f(X \mid \theta) = -\mathbb{H}[f(X \mid \theta^*), f(X \mid \theta)]$ where $\mathbb{H}(\cdot,\cdot)$ is the cross entropy. If $\theta = \theta^*$ this instead becomes the regular entropy $\mathbb{H}[f(X \mid \theta^*)]$. This implies that

$$f(X^{(n)} \mid \theta) \geq e^{\mathbb{E}_{X^{(n)}\sim f(X^{(n)}|\theta^*)} \log f(X^{(n)}|\theta) - \sqrt{n\frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}} = e^{-\mathbb{H}(f(X^{(n)}|\theta^*), f(X^{(n)}|\theta)) - \sqrt{n\frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}} \tag{8}$$

$$f(X^{(n)} \mid \theta) \leq e^{\mathbb{E}_{X^{(n)}\sim f(X^{(n)}|\theta^*)} \log f(X^{(n)}|\theta) + \sqrt{n\frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}} = e^{-\mathbb{H}(f(X^{(n)}|\theta^*), f(X^{(n)}|\theta)) + \sqrt{n\frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}}. \tag{9}$$

Step 3: From bounding probabilities to concentration of posteriors.

$$\begin{aligned}
\mathbb{P}\left(\Theta \setminus \Xi \mid X^{(n)}, E_{\bar{X}}\right) =& \frac{\int_{\Theta\setminus\Xi} f\left(X^{(n)} \mid \theta, E_{\bar{X}}\right) d\mathbb{P}(\theta)}{\int_{\Theta} f\left(X^{(n)} \mid \theta, E_{\bar{X}}\right) d\mathbb{P}(\theta)} && (10)\\
\leq& \frac{\int_{\Theta\setminus\Xi} f\left(X^{(n)} \mid \theta, E_{\bar{X}}\right) d\mathbb{P}(\theta)}{\int_{\Xi} f\left(X^{(n)} \mid \theta, E_{\bar{X}}\right) d\mathbb{P}(\theta)} && (11)\\
\leq& \frac{\int_{\Theta\setminus\Xi} e^{-\mathbb{H}(f(X^{(n)}|\theta^*), f(X^{(n)}|\theta)) + \sqrt{n\frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}} d\mathbb{P}(\theta)}{\int_{\Xi} e^{-\mathbb{H}(f(X^{(n)}|\theta^*), f(X^{(n)}|\theta)) - \sqrt{n\frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}} d\mathbb{P}(\theta)}. && (12)
\end{aligned}$$

The last inequality holds due to Equation (8) and (9).

Now, we proceed as follows

$$\begin{aligned}
\mathbb{P}\left(\Theta \setminus \Xi \mid X^{(n)}, E_{\bar{X}}\right) \leq& e^{2\sqrt{n\frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}} \frac{\int_{\Theta\setminus\Xi} e^{-\mathbb{H}(f(X^{(n)}|\theta^*), f(X^{(n)}|\theta))} d\mathbb{P}(\theta)}{\int_{\Xi} e^{-\mathbb{H}(f(X^{(n)}|\theta^*), f(X^{(n)}|\theta))} d\mathbb{P}(\theta)} && (13)\\
=& e^{2\sqrt{n\frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}} \frac{\int_{\Theta\setminus\Xi} e^{-\mathrm{KL}(f(X^{(n)}|\theta^*) \,\|\, f(X^{(n)}|\theta))} d\mathbb{P}(\theta)}{\int_{\Xi} e^{-\mathrm{KL}(f(X^{(n)}|\theta^*) \,\|\, f(X^{(n)}|\theta))} d\mathbb{P}(\theta)} && (14)\\
\leq& e^{2\sqrt{n\frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}} \frac{e^{-\inf_{\theta\notin\Xi} \mathrm{KL}(f(X^{(n)}|\theta^*) \,\|\, f(X^{(n)}|\theta))} \int_{\Theta\setminus\Xi} d\mathbb{P}(\theta)}{\int_{\Xi} e^{-nL_\theta\|\theta^*-\theta\|} d\mathbb{P}(\theta)} && (15)\\
\leq& \frac{1}{\mathbb{P}(\Xi)\nu(\Xi)^n} e^{-\inf_{\theta\notin\Xi} \mathrm{KL}(f(X^{(n)}|\theta^*) \,\|\, f(X^{(n)}|\theta)) + 2\sqrt{n\frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}} && (16)
\end{aligned}$$

Equation (14) is from multiplying top and bottom with $e^{\mathbb{H}[f(X|\theta^*)]}$ and the definition of KL-divergence. Equation (15) comes from the Lipschitz property of the log-likelihood $f(X^{(n)} \mid \theta)$ and therefore also the KL-divergence. Equation (16) comes from the assumption that

$$\int_{\theta \in \Xi} e^{-L_\theta \|\theta^* - \theta\|} \mathbb{P}(\theta) d\theta \geq \nu(\Xi) \qquad \forall \Xi \ni \theta^*. \tag{17}$$

and Jensen's inequality using $\phi(y) = y^n$ which is convex for non-negative values of y. Specifically, by setting $y = e^{-L_\theta \|\theta^* - \theta\|}$ for $\theta \in \Xi$ and considering convexity and positivity of y^n for any closed Ξ around θ^*, we obtain through Jensen's inequality that

$$\int_{\Xi} \phi(\theta) d\mathbb{P}(\theta) = \mathbb{E}[\phi | \theta \in \Xi] \mathbb{P}(\Xi) = \mathbb{P}(\Xi) \int_{\Xi} \phi(\theta) d\mathbb{P}(\theta | \theta \in \Xi) \geq \mathbb{P}(\Xi) \phi \left(\int_{\Theta} f(\theta) d\mathbb{P}(\theta | \Xi) \right) .$$

Note that for conditional expectations $\mathbb{E}[f(x) | x \in S] = \int_{x \in S} f(x) P(x) / P(S)$. This completes the step together with bounding the probability by one.

The assumption in Equation (17) is reasonable considering that the exponent is zero around θ^* and the prior has a minimum mass around it.

Step 4. Constructing the confidence interval. Now, if we upper bound the probability $\mathbb{P}\left(\Theta \setminus \Xi \mid X^{(n)}, E_{\bar{X}}\right)$ with $\delta' \in (0,1)$, we get

$$\inf_{\theta \notin \Xi} \text{KL}(f(X^{(n)}|\theta^*) \parallel f(X^{(n)}|\theta)) - 2\sqrt{n \frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}) \geq \ln\left(\frac{1}{\delta'(1-\delta')\nu^n(\Xi)}\right)$$

$$\implies \inf_{\theta \notin \Xi} \text{KL}(f(X^{(n)}|\theta^*) \parallel f(X^{(n)}|\theta)) \geq 2\sqrt{n \frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}) + \ln\left(\frac{1}{\delta'}\right)$$

$$\implies \inf_{\theta \notin \Xi} \text{KL}(f(X|\theta^*) \parallel f(X|\theta)) \geq 2\sqrt{\frac{L_x^2}{n\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}) + \frac{1}{n} \ln\left(\frac{1}{\delta'}\right) .$$

The first implication is true due to the observations that $1 - \delta' < 1$ and $\nu(\Xi) < 1$. The second implication is true due to tensorisation property of KL-divergence.

If Ξ is defined as

$$\Xi = \left\{ \theta \;\middle|\; \text{KL}(f(X|\theta^*) \parallel f(X|\theta)) \leq 2\sqrt{\frac{L_x^2}{n\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)} + \frac{1}{n} \ln\left(\frac{1}{\delta'}\right) \right\}. \tag{18}$$

Then, we get

$$\mathbb{P}\left(\theta \notin \Xi \mid X^{(n)}, E_{\bar{X}}\right) \leq \delta' \quad \implies \quad \mathbb{P}\left(\theta \notin \Xi \mid X^{(n)}\right) \leq \delta' + \delta. \tag{19}$$

Setting $\delta' = \delta$ completes our proof. ∎

Lemma 1. *Under the conditions of Theorem 3 and the mean reward for the MDPs satisfying $|\bar{R}_M(s)| \leq B_R \forall s$, Bayesian regret of PSRL satisfies*

$$BR(T) = \tilde{\mathcal{O}}\left(B_R + H\sigma_R \sqrt{T} + H\sigma_R \left(\frac{L_x^2}{\alpha_{\theta^*}}\right)^{1/4} T^{3/4} \right).$$

Proof. **Step 1: Bayesian Regret in PSRL Scheme.** If we consider the first step of an episode l, the total number of completed steps is $n = (l-1)H$. In PSRL, we sample $\theta_n \equiv \theta_{(\ell-1)H} \sim \mu_n$, where $\mu_n \equiv \mu_{(\ell-1)H} = \mathbb{P}(\theta^\star \in \cdot \mid \mathcal{H}_n)$ is the posterior distribution of $\theta^\star$. Osband et al. (2013) observes that for any $\sigma(\mathcal{H}_l)$ measurable function f, given the history of transitions $\mathcal{H}_l \equiv \mathcal{H}_{(\ell-1)H} = \left\{(s_{t',h}, a_{l',h}, s_{l',h+1})_{l'<\ell, h\leq H}\right\}$, we have $\mathbb{E}[f(\theta_n)] = \mathbb{E}[f(\theta^*)]$. This family of f's includes the value function. Therefore, we have $\mathbb{E}\left[V^{\theta_n}_{\pi_\ell,1}(s_{\ell,1})\right] = \mathbb{E}\left[V^{\theta^*}_{\pi^*,1}(s_{\ell,1})\right]$. Hence, the Bayes regret (Equation (1)) of PSRL can be re-written as

$$BR(T) = \mathbb{E}\left[\sum_{l=1}^{\tau} V^{\theta_n}_{\pi_l,1}(s_{l,1}) - V^{\theta^\star}_{\pi_l,1}(s_{l,1})\right] .$$

Step 2. Recursion with Bellman Equation. Chowdhury and Gopalan (2019) further shows that by a recursive application of the Bellman equation, we can decompose this regret into the expectation of a martingale difference sequence, and the difference of the next step value functions in the sampled and true MDPs. Specifically,

$$BR(T) = \mathbb{E}\left[\sum_{l=1}^{\tau}\sum_{h=1}^{H}\mathcal{T}^{\pi_l}_{\theta_n,h}(V^{\theta_n}_{\pi_l,h+1})(s_{l,h}) - \mathcal{T}^{\pi_l}_{\theta^*,h}(V^{\theta^\star}_{\pi_l,h+1})(s_{l,h}) + \sum_{l=1}^{\tau}\sum_{h=1}^{H} m_{\ell,h}\right].$$

Here, $\mathcal{T}^{\pi}_{\theta,h}$ denotes the Bellman operator at step h of the episode due to a policy π and MDP θ, and is defined as $\mathcal{T}^{\pi}_{\theta,h}(V^{\theta}_{\pi,h+1})(s_{l,h}) = R(s,\pi(s,h)) + \mathbb{E}_{s,\pi(s,h)}[V|\theta]$. In addition, $m_{\ell,h} = \mathbb{E}^{\theta^\star}_{s_{\ell,h},a_{\ell,h}}\left[V^{\theta_n}_{\pi_\ell,h+1}(s_{\ell,h+1}) - V^{\theta^\star}_{\pi_\ell,h+1}(s_{\ell,h+1})\right] - \left(V^{\theta_n}_{\pi_\ell,h+1}(s_{\ell,h+1}) - V^{\theta^\star}_{\pi_\ell,h+1}(s_{\ell,h+1})\right)$ is a martingale difference sequence satisfying $\mathbb{E}\left[m_{\ell,h}\right] = 0$.

Step 3. From Value Function to KL. Now, from (Boucheron et al., 2003), we obtain the transportation inequalities stating that

$$\mathbb{E}_P[X] - \mathbb{E}_Q[X] \leq \sqrt{2\mathbb{V}_P(X)\mathrm{KL}(Q \parallel P)}\,. \tag{20}$$

Then an application of the transportation inequality yields

$$BR(T) \leq H\mathbb{E}\left[\sum_{\ell=1}^{\tau}\sum_{h=1}^{H}\sqrt{2\sigma_R^2\mathrm{KL}_{s^l_h,a^l_h}(\theta^*\|\theta_n)}\right],$$

Here, σ_R^2 is the maximum variance of rewards at each step.

Finally, using our concentration bounds on $\mathrm{KL}(\theta^* \parallel \theta)$ under posterior distributions and then Cauchy-Schawrtz inequality yields

$$BR(T|\mathcal{E}^\star) \leq H\sigma_R\sqrt{2T\sum_{l=1}^{\tau}\sum_{h=1}^{H}\mathrm{KL}_{s^l_h,a^l_h}(\theta^*\|\theta_n)} \leq H\sigma_R\sqrt{2T\left(T\frac{L_x^2}{\alpha_{\theta^*}}\ln\left(\frac{2}{\delta}\right)\right)^{1/2} + 2T\ln\left(\frac{T}{\delta}\right)}.$$

Here, $\mathcal{E}^\star$ denotes the event of the distribution of data concentrating in the set Ξ_n around θ^* under the n-th step posterior for any $n \geq 1$.

Step 4. Putting the Events Together. Due to bounded mean of the rewards, we can always bound $BR(T) \leq TB_{\mathcal{R}}$.

Thus, we have

$$BR(T) = \mathbb{E}\left[\mathcal{R}(T)\mathbb{I}_{\mathcal{E}^\star} + \mathcal{R}(T)\mathbb{I}_{(\mathcal{E}^\star)^c}\right] \leq BR(T|\mathcal{E}^\star) + TB_R\left(1 - \mathbb{P}\left(\mathcal{E}^\star\right)\right)$$

From Theorem 3, $\mathbb{P}\left(\mathcal{E}^\star\right) \geq 1 - 2\delta$. This implies for any $\delta \in (0,1)$ that the Bayes regret

$$\begin{aligned} BR(T) &\leq H\sigma_R\sqrt{2T\left(T\frac{L_x^2}{\alpha_{\theta^*}}\ln\left(\frac{2}{\delta}\right)\right)^{1/2} + 2T\ln\left(\frac{T}{\delta}\right)} + 2TB_R\delta \\ &\leq B_R + 2H\sigma_R\sqrt{T\ln(2T)} + \sqrt{2}H\sigma_R T^{3/4}\left(\frac{L_x^2}{\alpha_{\theta^*}}\right)^{1/4}\log^{1/4}(4T)\,. \end{aligned} \tag{21}$$

The proof is completed by setting $\delta = \frac{1}{2T}$. ■

C.2 Regret for Posteriors with Linear LSI Constants

To study the alternate approach we first need to define the one step future value function $U(\varphi)$ as the expected value of the optimal policy π_l in M_l where φ is the distribution of next state samples. This gives $U_h^{M_l}(\varphi) = \mathbb{E}_{s' \sim \varphi}\left[V^{\pi_l, h+1}(s')\right]$. We use this definition, which is also used in previous work, (Chowdhury and Gopalan, 2019; Osband and Van Roy, 2014), to make a Lipschitz assumption on the next step value function U with respect to the means of the distributions.

Assumption 2 (One step value function is Lipschitz in the mean)**.** *For any φ_1, φ_2 distributions over $\mathcal{S}$ with $1 \le h \le H$,*

$$|U_h^M(\varphi_1) - U_h^M(\varphi_2)| \le L_M \|\overline{\varphi_1} - \overline{\varphi_2}\|_2 \tag{22}$$

where $\overline{\varphi_1}$ and $\overline{\varphi_2}$ are the means of the respective distributions.

Theorem 4. *If the posterior distributions for mean rewards and transitions separately satisfy LSI with constants $\{\alpha_{\bar{R},l}\}$ and $\{\alpha_{\bar{\mathcal{T}},l}\}$, the mean reward for any MDP M $|\bar{R}_M(s)| \le B_R \forall s$, the one step value function is Lipschitz in the state with parameter L_{M_*} as Assumption 2, and the mean reward and mean transitions are $L_{\bar{R}}$ and $L_{\bar{\mathcal{T}}}$ Lipschitz in θ, Bayesian regret of PSRL is upper bounded by*

$$\tilde{\mathcal{O}}\left(H\left(\sum_{l=1}^{\tau} \frac{L_{\bar{R}}}{\sqrt{\alpha_{\bar{R},l}}} + \mathbb{E}[L_{M_*}]\sqrt{d}\sum_{l=1}^{\tau}\frac{L_{\bar{\mathcal{T}}}}{\sqrt{\alpha_{\bar{\mathcal{T}},l}}}\right)\right).$$

Proof. This proof follows the general flow from Chowdhury and Gopalan (2019) for Kernel PSRL but with totally different confidence bounds.

For PSRL, we have $\pi_l = \arg\max_\pi V_{\pi,1}^{M_l}$. We also denote the optimal policy for the true MDP M_* as $\pi_* = V_{\pi,1}^{M_*}$. With the observation that under any observed history $\mathcal{H}_{l-1}$ we have $\mathbb{E}[V_{\pi_l,1}^{M_l}(s_{l,1}) \mid \mathcal{H}_{l-1}] = \mathbb{E}[V_{\pi_*,1}^{M_*}(s_{l,1}) \mid \mathcal{H}_{l-1}]$, since they are both sampled from the same distribution. Marginalising we obtain:

$$\mathbb{E}[\text{Regret}(T)] \triangleq \sum_{l=1}^{\tau} \mathbb{E}\left[V_{\pi_*,1}^{M_*}(s_{l,1}) - V_{\pi_l,1}^{M_*}(s_{l,1})\right] \tag{23}$$

$$= \sum_{l=1}^{\tau} \mathbb{E}\left[V_{\pi_*,1}^{M_*}(s_{l,1}) - V_{\pi_l,1}^{M_l}(s_{l,1})\right] + \mathbb{E}\left[V_{\pi_l,1}^{M_l}(s_{l,1}) - V_{\pi_l,1}^{M_*}(s_{l,1})\right] \tag{24}$$

$$= \sum_{l=1}^{\tau} \mathbb{E}\left[V_{\pi_l,1}^{M_l}(s_{l,1}) - V_{\pi_l,1}^{M_*}(s_{l,1})\right] \tag{25}$$

Next, we use Lemma 7 and observation after eq 50 from (Chowdhury and Gopalan, 2019) and obtain

$$\mathbb{E}[\text{Regret}(T)] \le \mathbb{E}\left[\sum_{l=1}^{\tau}\sum_{h=1}^{H}\left[|\bar{R}_{M_l}(z_{l,h}) - \bar{R}_*(z_{l,h})| + L_{M_l}||\bar{\mathcal{T}}_{M_l}(z_{l,h}) - \bar{\mathcal{T}}_*(z_{l,h})||_2\right]\right]. \tag{26}$$

where $\bar{\mathcal{T}}_M$ and $\bar{R}_M$ are the mean of the transition and reward distributions for MDP M.

We define two confidence sets

$$C_{R,l,h} = \left\{|\bar{R}_\theta(s_{l,h}) - E_{\mathbb{P}(\theta|\mathcal{H}_l)}[\bar{R}_\theta(s_{l,h})]| \le \sqrt{\frac{L_{\bar{R}}^2 \log 1/\delta}{\alpha_{\bar{R},l}}}\right\} \tag{27}$$

$$C_{\mathcal{T},l,h} = \left\{\|\bar{\mathcal{T}}_\theta(s_{l,h}, a_{l,h}) - E_{\mathbb{P}(\theta|\mathcal{H}_l)}[\bar{\mathcal{T}}_\theta(s_{l,h}, a_{l,h})]\|_2 \le \sqrt{\frac{dL_{\bar{\mathcal{T}}}^2 \log 1/\delta}{\alpha_{\bar{\mathcal{T}},l}}}\right\} \tag{28}$$

Define events $E_* \triangleq \{\bar{R}_* \in C_{R,l,h}, \bar{\mathcal{T}}_* \in C_{\mathcal{T},l,h}, \forall 1 \le l \le \tau, 1 \le h \le H\}$ and $E_{M_l} \triangleq \{\bar{R}_{M_l} \in C_{R,l,h}, \bar{\mathcal{T}}_{M_l} \in C_{\mathcal{T},l,h}, \forall 1 \le l \le \tau, 1 \le h \le H\}$. Now We fix $0 \le \delta \le 1$ and from property on sub-Gaussian concentration for log-Sobolev posteriors in Equation (3), we get $\mathbb{P}(E_M) = \mathbb{P}(E_*) = 1 - 2H\tau\delta$. Taking the union of these events $E \triangleq E_M \cap E_*$ with $\mathbb{P}(E^c) \le \mathbb{P}(E_M^c) + \mathbb{P}(E_*^c) \le 4\tau H\delta$. We also have that $\mathbb{E}[L_{M_l}] = \mathbb{E}[L_{M_*}]$ such that $\mathbb{E}[L_{M_l}|E] \le \frac{\mathbb{E}[L_{M_*}]}{P(E)} \le \frac{\mathbb{E}[L_{M_*}]}{1-4\tau H\delta}$.

Combining the results we then get an upper bound on the Bayesian regret

$$\mathbb{E}[\sum_{l=1}^{\tau}\sum_{h=1}^{H}\left[|\bar{R}_{M_l}(z_{l,h}) - \bar{R}_*(z_{l,h})| \mid E\right] + \mathbb{E}\left[L_{M_l}||\bar{\mathcal{T}}_{M_l}(z_{l,h}) - \bar{\mathcal{T}}_*(z_{l,h})||_2 \big| E\right] + 2B_R 4\tau H\delta \tag{29}$$

$$\leq 2H\left(L_{\bar{R}}\sqrt{\log 1/\delta}\sum_{l=1}^{\tau}\frac{1}{\sqrt{\alpha_{\bar{R},l}}} + \frac{\mathbb{E}[L_{M_*}]}{1-2\tau H\delta}L_{\bar{\mathcal{T}}}\sqrt{d\log 1/\delta}\sum_{l=1}^{\tau}\frac{1}{\sqrt{\alpha_{\bar{\mathcal{T}},l}}}\right) + 8B_R\tau H\delta \tag{30}$$

Setting $\delta = \frac{1}{8\tau H}$ we obtain

$$\mathbb{E}[\text{Regret}(T)] \leq 2H\left(L_{\bar{R}}\sqrt{\log 8T}\sum_{l=1}^{\tau}\frac{1}{\sqrt{\alpha_{\bar{R},l}}} + 2\mathbb{E}[L_{M_*}]L_{\bar{\mathcal{T}}}\sqrt{d\log 8T}\sum_{l=1}^{\tau}\frac{1}{\sqrt{\alpha_{\bar{\mathcal{T}},l}}}\right) + B_R \tag{31}$$

■

D REGRET BOUNDS AND SAMPLE COMPLEXITY FOR LAPSRL WITH APPROXIMATE POSTERIORS

Theorem 5. *Let the policy the start of episode l act by sampling a model from Q_l where $\min(KL((\mathbb{P}_l \parallel Q_l), KL(Q_l \parallel \mathbb{P}_l)) \leq \epsilon_{post,l}$ and where $\mathbb{P}_l$ is the exact posterior at start of episode l. Then the incurred regret from planning with an approximate posterior bounded by $\sqrt{2}\Delta_{max}\sqrt{\epsilon_{post,l}}$.*

Proof. Let $\mu_l, \mu_l^* \sim \mathbb{P}(\mu_l)$, $\mu_l' \sim Q(\mu_l)$. The policy π_l is the optimal policy corresponding to μ_l and π_l' the policy corresponding to μ_l'.

$$\mathbb{E}_{P_l,Q_l}[V_{\pi^*}^{\mu^*} - V_{\pi_l'}^{\mu^*}] = \mathbb{E}_{P_l,Q_l}[V_{\pi^*}^{\mu^*} - V_{\pi_l'}^{\mu_l'} + V_{\pi_l'}^{\mu_l'} - V_{\pi_l'}^{\mu^*}] \tag{32}$$

$$= \mathbb{E}_{P_l,Q_l}[V_{\pi^*}^{\mu^*} - V_{\pi_l}^{\mu_l} + V_{\pi_l}^{\mu_l} - V_{\pi_l'}^{\mu_l'} + V_{\pi_l'}^{\mu_l'} - V_{\pi_l'}^{\mu^*}] \tag{33}$$

$$= \mathbb{E}_{P_l,Q_l}[[V_{\pi^*}^{\mu^*} - V_{\pi_l}^{\mu_l}] + [V_{\pi_l}^{\mu_l} - V_{\pi_l'}^{\mu_l'}] + [V_{\pi_l'}^{\mu_l'} - V_{\pi_l'}^{\mu^*}]] \tag{34}$$

$$\leq \mathbb{E}_{P_l}[V_{\pi^*}^{\mu^*} - V_{\pi_l}^{\mu_l}] + \Delta_{\max}\sqrt{\frac{\epsilon_{\text{post},l}}{2}} + \Delta_{\max}\sqrt{\frac{\epsilon_{\text{post},l}}{2}}] \tag{35}$$

$$= \mathbb{E}_{P_l}[V_{\pi^*}^{\mu^*} - V_{\pi_l}^{\mu^*}] + \sqrt{2}\Delta_{\max}\sqrt{\epsilon_{\text{post},l}} \tag{36}$$

The second term in the inequality comes from the total variation distance that can make MDPs with large values be more common in P than in Q. The third term is similar, we can sample the policy from P instead of Q, with the added worst case penalty for the terms that differ. ■

Corollary 1. *If a policy incurs $\widetilde{\mathcal{O}}(\sqrt{T}g(H,\mathcal{S},\mathcal{A}))$ regret under distribution P, for some function g, it will incur the same order of regret under Q if $0 \leq \epsilon_{post,l} \leq C\frac{g(H,\mathcal{S},\mathcal{A})^2}{l\Delta_{max}^2}$ for some constant $C \geq 0$.*

Proof. The regret for an algorithm following the approximate posterior Q is

$$\widetilde{\mathcal{O}}(\mathbb{E}_P R(\pi_Q)) \leq \widetilde{\mathcal{O}}(\sqrt{\tau}g(H,\mathcal{S},\mathcal{A})) + \sqrt{2}\Delta_{\max}\sum_{k=1}^{\tau}\sqrt{\epsilon_{\text{post},l}} \tag{37}$$

$$\leq \widetilde{\mathcal{O}}(\sqrt{\tau}g(H,\mathcal{S},\mathcal{A})) + \sqrt{2}\Delta_{\max}\sum_{k=1}^{\tau}\sqrt{C}\frac{g(H,\mathcal{S},\mathcal{A})}{\sqrt{t}\Delta_{\max}} \tag{38}$$

$$= \widetilde{\mathcal{O}}(\sqrt{\tau}g(H,\mathcal{S},\mathcal{A})) + \sqrt{2}g(H,\mathcal{S},\mathcal{A})\sqrt{C}\sum_{k=1}^{\tau}\frac{1}{\sqrt{t}} \tag{39}$$

$$= \widetilde{\mathcal{O}}(\sqrt{\tau}g(H,\mathcal{S},\mathcal{A})) \tag{40}$$

■

Corollary 2. *For a posterior fulfilling the Assumption 1 and definition 1, a posterior sampling style algorithm can obtain an unchanged order of regret under SARAH-LD sampling under a gradient complexity for episode l of gradient complexity* $= \widetilde{O}\left(\frac{H^3 l^3 L^2}{\alpha_l^2} + \frac{dH^{2.5} l^{3.5} L^2}{\alpha_l^2 g(H,\mathcal{S},\mathcal{A})^2}\right)$ *In many cases, as seen in Section 5,* $\alpha_l = \widetilde{\Omega}(Hl)$. *Similarly, as in Theorem 4,* $g(H,\mathcal{S},\mathcal{A}) \propto H$. *This then becomes gradient complexity* $\propto \widetilde{O}\left(HlL^2 + \frac{dl^{3/2}L^2}{H^{3/2}}\right)$.

Proof. This result can be obtained by directly applying the $\epsilon_{\text{post},l}$ obtained from Theorem 5 into Theorem 10 with $\gamma = n$. ■

Theorem 6. *Let $\rho_*^l(\theta)$ be the final sample of the Langevin algorithm at episode l, approximating the true posterior $\mathbb{P}(\theta \mid \mathcal{H}_l)$ with $KL(\rho_*^l(\theta) \parallel \mathbb{P}(\theta \mid \mathcal{H}_l)) \leq \epsilon_{post,l}$. Additionally, if $\nabla_z \log f(z|\theta)$ is L_z-Lipschitz and α_z-Log Sobolev, we get $\mathbb{E}_{f(z|\mathcal{H}_l)} KL(\rho_*^l(\theta) \parallel \mathbb{P}(\theta \mid \mathcal{H}_{l+1})) \leq \epsilon_{post,l} + \frac{L_z}{2\alpha_z} Var_{\rho_*^l(\theta)}(\theta)$, where $Var_{\rho_*^l(\theta)}(\theta)$ is the variance of the approximate posterior distribution $\rho_*^l(\theta)$.*

Proof. For notation we write $\mathbb{P}(\theta \mid \mathcal{H}_{l+1}) = \mathbb{P}(\theta \mid \mathcal{H}_l, z_l)$ such that $\mathbb{P}(\theta \mid \mathcal{H}_{l+1}) = \mathbb{P}(\theta \mid z_0, \ldots, z_l)$. Note that we can marginalize $f(z_l \mid \mathcal{H}_l, \theta) = f(z_l \mid \theta)$ and $\mathbb{E}_\theta f(z_l \mid \mathcal{H}_l, \theta) = f(z_l \mid \mathcal{H}_l)$. As a reminder, we have $f(z \mid \theta)$ as the data likelihood, as such we have with Bayes rule $\mathbb{P}(\theta \mid \mathcal{H}_l) = \frac{\mathbb{P}(\theta) f(\mathcal{H}_l|\theta)}{f(\mathcal{H}_l)}$.

$$
\begin{aligned}
&\mathrm{KL}(\rho_*^l \parallel \nu_{l+1} \mid z_l) && (41)\\
=&\int_\Theta \log\left(\frac{\rho_*^l(\theta)}{\mathbb{P}(\theta \mid \mathcal{H}_l, z_l)}\right)\rho_*^l(\theta)d\theta && (42)\\
=&\int_\Theta \log\left(\frac{\rho_*^l(\theta)}{\frac{\mathbb{P}(\theta|\mathcal{H}_l)f(z_l|\theta)}{f(z_l|\mathcal{H}_l)}}\right)\rho_*^l(\theta)d\theta && (43)\\
=&\int_\Theta \left(\log\left(\frac{\rho_*^l(\theta)}{\mathbb{P}(\theta \mid \mathcal{H}_l)}\right) + \log\left(\frac{f(z_l|\mathcal{H}_l)}{f(z_l|\theta)}\right)\right)d\rho_*^l(\theta) && (44)\\
=&\int_\Theta \log\left(\frac{\rho_*^l(\theta)}{\mathbb{P}(\theta \mid \mathcal{H}_l)}\right)d\rho_*^l(\theta) && (45)\\
&+\int_\Theta \log\left(\frac{f(z_l|\mathcal{H}_l)}{f(z_l|\theta)}\right)\rho_*^l(\theta)d\theta && (46)\\
=&\mathrm{KL}(\rho_*^l \parallel \nu_l) + \int_\Theta \log\left(\frac{f(z_l|\mathcal{H}_l)}{f(z_l|\theta)}\right)d\rho_*^l(\theta) && (47)\\
\leq&\epsilon_{\text{post},l} + \int_\Theta \log\left(\frac{f(z_l|\mathcal{H}_l)}{f(z_l|\theta)}\right)d\rho_*^l(\theta) && (48)\\
=&\epsilon_{\text{post},l} + \int_\Theta \log\left(f(z_l|\mathcal{H}_l)\right)d\rho_*^l(\theta) - \int_\Theta \log\left(f(z_l|\theta)\right)d\rho_*^l(\theta). && (49)
\end{aligned}
$$

The second inequality comes from Bayes rule together with the marginalizations from above, the rest is separating of logarithms and identifying the desired KL-divergence.

This then gives that in expectation

$$
\begin{aligned}
&\mathbb{E}_{z_l}\mathrm{KL}(\rho_*^l \parallel \nu_{l+1} \mid z_l) && (50)\\
\leq&\epsilon_{\text{post},l} + \int_{z_l}\int_\Theta \log\left(\frac{f(z_l|\mathcal{H}_l)}{f(z_l|\theta)}\right)d\rho_*^l(\theta)f(z_l \mid \mathcal{H}_l)dz_l && (51)\\
=&\epsilon_{\text{post},l} + \int_\Theta\int_{z_l} \log\left(\frac{f(z_l|\mathcal{H}_l)}{f(z_l|\theta)}\right)f(z_l \mid \mathcal{H}_l)dz_l d\rho_*^l(\theta) && (52)\\
=&\epsilon_{\text{post},l} + \int_\Theta\int_{z_l} \log\left(\frac{f(z_l|\mathcal{H}_l)}{f(z_l|\theta)}\right)\frac{f(z_l|\mathcal{H}_l)}{f(z_l|\theta)}f(z_l|\theta)dz_l d\rho_*^l(\theta) && (53)
\end{aligned}
$$

$$\leq \epsilon_{\text{post},l} + \int_{\Theta} 2/\alpha_z \int_{z_l} ||\nabla_z \sqrt{\frac{f(z_l|\mathcal{H}_l)}{f(z_l|\theta)}}||^2 f(z_l|\theta) dz_l d\rho_*^l(\theta) \tag{54}$$

$$= \epsilon_{\text{post},l} + \int_{\Theta} 2/\alpha_z \int_{z_l} ||\frac{f(z_l|\theta)\nabla_z f(z_l|\mathcal{H}_l) - f(z_l|\mathcal{H}_l)\nabla_z f(z_l|\theta)}{2\sqrt{\frac{f(z_l|\mathcal{H}_l)}{f(z_l|\theta)}} f(z_l|\theta)^2}||^2 f(z_l|\theta) dz_l d\rho_*^l(\theta) \tag{55}$$

$$= \epsilon_{\text{post},l} + \int_{\Theta} 2/\alpha_z \int_{z_l} ||\frac{f(z_l|\theta)\nabla_z f(z_l|\mathcal{H}_l) - f(z_l|\mathcal{H}_l)\nabla_z f(z_l|\theta)}{2\ f(z_l|\mathcal{H}_l) f(z_l|\theta)} \times \sqrt{\frac{f(z_l|\mathcal{H}_l)}{f(z_l|\theta)}}||^2 f(z_l|\theta) dz_l d\rho_*^l(\theta) \tag{56}$$

$$= \epsilon_{\text{post},l} + \int_{\Theta} 2/\alpha_z \int_{z_l} ||\frac{f(z_l|\theta)\nabla_z f(z_l|\mathcal{H}_l) - f(z_l|\mathcal{H}_l)\nabla_z f(z_l|\theta)}{2\ f(z_l|\mathcal{H}_l) f(z_l|\theta)}||^2 \frac{f(z_l|\mathcal{H}_l)}{f(z_l|\theta)} f(z_l|\theta) dz_l d\rho_*^l(\theta) \tag{57}$$

$$= \epsilon_{\text{post},l} + \int_{\Theta} \frac{1}{2\alpha_z} \int_{z_l} ||\nabla_z \log f(z_l|\mathcal{H}_l) - \nabla_z \log f(z_l|\theta)||^2 f(z_l|\mathcal{H}_l) dz_l d\rho_*^l(\theta) \tag{58}$$

$$\leq \epsilon_{\text{post},l} + \frac{L_z}{2\alpha_z} \int_{\Theta} ||\theta_l - \theta||^2 d\rho_*^l(\theta) \tag{59}$$

$$= \epsilon_{\text{post},l} + \frac{L_z}{2\alpha_z} \text{Var}_{\rho_*^l(\theta)}(\theta) \tag{60}$$

■

The inequality in Equation (54) comes from the log-Sobolev inequality property of $f(z_l|\theta)$. The rest is algebra with the exception of the final inequality which comes from the Lipschitz property.

E LAPSRL FOR DIFFERENT FAMILIES OF DISTRIBUTIONS

In this section we present the details on LaPSRL for different families of distributions which can be found in Table 2. The results for the Bayesian regret follow from combining the scaling of the log-Sobolev constant with Theorem 4.

Theorem 8. *Any log-concave posterior has $\alpha_l = \Theta(n)$. Any posterior that is a mixture of log-concave distributions has $\alpha_{\textit{Mixture}} = \Omega\left(\frac{n \min p_i}{4k(1-\log(\min p_i))}\right)$.*

Proof. We can write the product of log-concave distributions $\mathbb{P}(\theta \mid \mathcal{H}_l) = \mathbb{P}(\theta)\frac{\prod_{i=1}^n f_i(\theta)}{Z}$ where $f_i(\theta)$ is shorthand for $f(x_i \mid \theta)$ with x_i the datapoint at time i. Since products preserve log-concavity, we can use Theorem 1. From Weyl's inequality, we have that the smallest eigenvalue a sum of two Hermitian is larger than the sum of the smallest eigenvalues of the two matrices. Since the Hessian is a Hermitian matrix, putting this into Theorem 1 this gives that $\alpha_l \geq \alpha_{\mathbb{P}(\theta)} + \sum_{i=1}^n \alpha_i \geq \alpha_{\mathbb{P}(\theta)} + n \min_i \alpha_i$. Similarly, applying Weyl's inequality for the largest eigenvalue, we get that the largest eigenvalue of $-\nabla^2 \log(\mathbb{P}(\theta \mid \mathcal{H}_l))$ is upper bounded by the sum of maximal eigenvalues which gives an upper bound of $O(n)$ for α_l since the smallest eigenvalue must be smaller than the largest.

Similarly, for mixtures of log-concave distributions we have from Theorem 7 that $\alpha_{\text{Mixture}} = \Omega\left(\frac{\min \alpha_i \min p_i}{4k(1-\log(\min p_i)}\right)$. Setting $\min_i \alpha_i = \Theta(n)$ completes the proof. ■

F EXPERIMENTAL DETAILS

In this section we go into more neccesary details for the experiments. We also re-include the plots from the main paper in Appendix F.3 for increased visibility.

Table 2: Overview of log-Sobolev constants and Bayes regrets of LaPSRL for different families of distributions.

Posterior	log-Sobolev constant	LaPSRL Bayesian regret
Gaussian	$\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$	$\widetilde{O}\left((L_{\bar{R}} + \mathbb{E}[L_{M_*}]L_{\bar{\mathcal{T}}})\sqrt{T\sigma^2}\right)$
Log-concave	$\Theta(n)$	$\widetilde{O}\left((L_{\bar{R}} + \mathbb{E}[L_{M_*}]L_{\bar{\mathcal{T}}})\sqrt{T}\right)$
Mixture of Log-concave	$\Omega\left(\frac{\delta \min p_i \min \alpha_i}{4k(1-\log(\min p_i))}\right)$	$\widetilde{O}\left((L_{\bar{R}} + \mathbb{E}[L_{M_*}]L_{\bar{\mathcal{T}}})\sqrt{\frac{4kT}{\min p_i}}\right)$

F.1 Gaussian multi-armed bandits

We use LaPSRL on a Gaussian multi-armed bandit task with two arms. The arms generate rewards as $\mathcal{N}(0, 0.25)$, $N(0.1, 0.25)$. To preserve computations, we use a batched approach such that the same action is played for 20 steps. As a baseline, we compare with the performance of PSRL from the true posterior. Both LaPSRL and Thompson sampling use a $\mathcal{N}(0, 1)$ prior for the mean of each arm. Additionally, we compare with a LaPSRL algorithm that has a bimodal $1/2\mathcal{N}(0, 1/4) + 1/2\mathcal{N}(1, 1)$ prior over the arms. In this experiment we use chained sampling in the Langevin algorithm such that we initialize with the previous step. The results can be seen in Appendix F.3.

F.2 Continuous MDPs

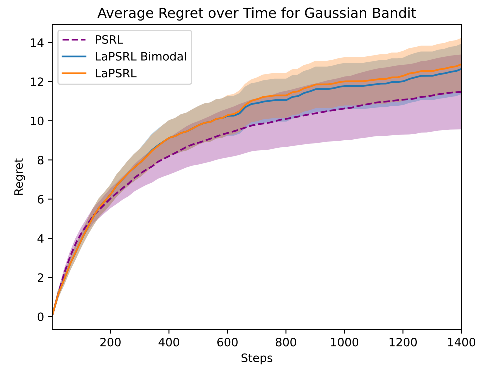
We evaluate LaPSRL on two continuous environments, Cartpole and Reacher. The Cartpole environment is a modified version of Cartpole environment (Barto et al., 1983) to have continuous actions, with states $s \in \mathbb{R}^4$ and a continuous action in [-1,1]. The goal is to control a cart such that the attached pole stays upright. We use a Linear Quadratic Regulator model, where LaPSRL samples from a distribution over the A and B matrixes with a $\mathcal{N}(0, 1)$ prior over the values. The policy can then be obtained through the Riccati equation. Instead of calculating the log-Sobolev constant for the posterior distribution, we just evaluate for a variety of $\alpha \in \{100n, 1000n, 10000n\}$. To simplify the parameter search, we set the L parameter to α/n. Instead of estimating $\log \frac{2\text{KL}(\rho_0 \parallel f(\theta|\mathcal{H}_l))}{\epsilon_{\text{post},l}}$, we upper bound this with n. In each sampling step, we start with an initial sample from $\mathcal{N}(0, 1)$. In the learning, we assume that the error is Gaussian standard deviation of 0.5. While Cartpole is not a linear MDP, but it is a good approximation and serves to show that LaPSRL can work even when the true model is not part of the posterior support. As a baseline we compare with an exact PSRL algorithm which samples from Bayesian linear regression priors (Minka, 2000). Finally, we use a variant of LaPSRL with a multimodal prior over the A and B matrixes with a $1/2\mathcal{N}(0, 1) + 1/2\mathcal{N}(1, 0.25)$ to demonstrate that it also works well for multimodal priors that are not log-concave. The results from this experiment can be found in Appendix F.3 where we plot what fraction of the 50 runs have solved the task (i.e. taking 200 steps without failing). Here we see that all versions successfully handle the task, even faster than the PSRL baseline. We can note that it takes longer for the experiments with larger α values to converge.

The Reacher environment is a standard environment from the Gymnasium environments (Towers et al., 2024) relying on MuJoCo (Todorov et al., 2012) physics simulations. In the Reacher environment the agent controls an arm and has to stay close to a randomly placed target. Here we use a neural network model, this means that this is not necessarily log-Sobolev, but we wish to show that this approach still is useful. The neural network parameters were sampled from $\mathcal{N}(0, 0.1)$. In the learning, we assume that the error is Gaussian standard deviation of 0.5. Here we let LaPSRL know the reward function, and the agent only models the movement of the robot arm. In the state of Reacher there are some states that are constant (the postion of the target) and some that are functions of other states (distance to the target), LaPSRL does not model these values. Policy rollouts are then done in the model using a cross-entropy method (iCEM) (Pinneri et al., 2020) to find a policy to use in the environment. For the iCEM algorithm we use 8 iterations, 48 trajectories in each and keep 5 elite samples. Since the environment is deterministic, we do not use any noise in the rollouts. We model the transitions using a neural network with two hidden layers with 128 neurons. In this Reacher experiment, we use chained sampling, but also fix the amount of steps in each sampling to 150000. We vary the LSI constant $\alpha \in \{10000n, 100000n\}$ and as in Cartpole experiment set the L

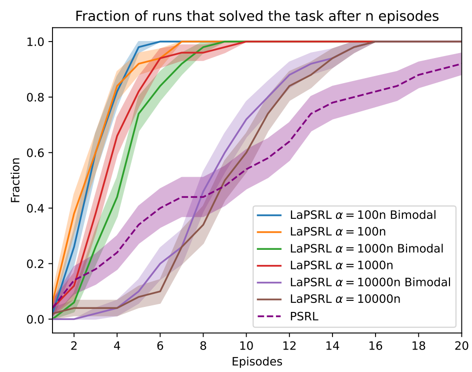
parameter to α/n. Here we benchmark with the TD3 (Fujimoto et al., 2018) algorithm and see in Appendix F.3 that LaPSRL can learn the environment very quickly.

F.3 Computational notes

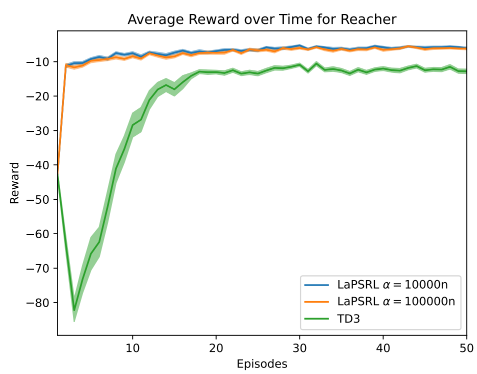
The experiments have been done primarily in Jax (Bradbury et al., 2018). The reacher environment is from the Gymnasium environments (Towers et al., 2024) relying on MuJoCo (Todorov et al., 2012) physics simulations. The Cartpole is a modified version of Cartpole environment (Barto et al., 1983) to have continuous actions. The for TD3 Fujimoto et al. (2018) we used the implementation from Huang et al. (2022). The experiments were run on an internal cluster to be able to run many experiments at once, but they will also run on a regular desktop. The continuous experiments use seeds 1-50, while the bandit experiments use seeds 100-150. These seeds being different is just a coincidence and has not been experimented with. See the supplementary code for additional details.



(a) Gaussian Batched Bandit



(b) Cartpole



(c) Reacher

Figure 3: We compare LaPSRL versus baselines. In the bandit and Cartpole experiments we benchmark with PSRL, in Reacher with PPO and TD3. On the left we compare the expected regret for a Gaussian bandit algorithm. In the middle we compare how many episodes it takes to solve a Cartpole task. On the right we study the average regret per episode in the Reacher environement. In all environments, we average over 50 independent runs with the standard error highlighted around the average.