

---

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

---

Emilio Jorge<sup>1,2</sup>  
 Christos Dimitrakakis<sup>3,4</sup>  
 Debabrota Basu<sup>5</sup>

## Abstract

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 theory, and others, such as many non-log-concave and perturbed distributions. First, we show that the Posterior Sampling-based RL (PSRL) algorithm yields sublinear regret if the data distributions satisfy LSI and 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 significant advances in Reinforcement Learning (RL), both in terms of theoretical understanding and practical applications. However, the theory does not always apply to real-world settings—exposing a theory-to-practice gap. For complex environments, RL algorithms often use a probabilistic approximation of the environment. In order to analyse them theoretically, we often assume linear (Geramifard et al., 2013), bilinear (Ouhamma et al., 2022), or reproducible kernel (Chowdhury & Gopalan, 2019) type parametric models, and Gaussian or log-concave posteriors for Bayesian RL algorithms (Chowdhury & Gopalan, 2019; Osband & Van Roy, 2017). In this paper, we aim to narrow this theory-to-practice gap by studying whether we can achieve the desired, i.e. sublinear, regret guarantees for the larger class of isoperimetric distributions.

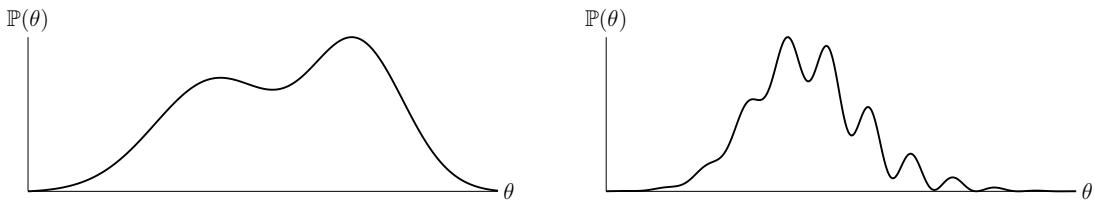


Figure 1: Examples of log-Sobolev distributions.

**Isoperimetry** relates to the ratio between the area of the boundary 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

<sup>1</sup>Chalmers University of Technology, Sweden <sup>2</sup>University of Gothenburg, Sweden <sup>3</sup>University of Neuchâtel, Switzerland <sup>4</sup>University of Oslo, Norway <sup>5</sup>Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 – CRIStAL. Correspondence to: Emilio Jorge <emilio@jorge.se>.

Table 1: Summary of assumptions and results of Exact and Approximate PSRL type algorithms with Bayesian regret.

Algorithm	$\text{BR}(T)$	Assumption	Total gradient complexity
PSRL (Osband et al., 2013b)	$\tilde{O}(HS\sqrt{AT})$	Tabular, exact posterior	-
(Moradipari et al., 2023)	$\tilde{O}(H^{1.5}\sqrt{SAT})$	Tabular, exact posterior	-
(Fan & Ming, 2021)	$\tilde{O}(dH^{1.5}\sqrt{T})$	Linear MDP, exact posterior	-
(Chowdhury & Gopalan, 2019)	$\tilde{O}(\sqrt{dHT})$	Kernel MDP, exact posterior	-
<b>PSRL (Lemma 1)</b>	$\tilde{O}(HT^{0.75})$	LSI $\mathcal{L}$ with constant $\alpha$ , exact posterior	-
<b>PSRL (Theorem 4)</b>	$\tilde{O}(\sqrt{dHT})$	LSI $\mathbb{P}(M)$ , linear growth on $\alpha$ , exact posterior	-
(Xu et al., 2022)	$\tilde{O}(d^{1.5}\sqrt{T})$	Lin. Bandits with cond. number $\kappa$ , subG rewards	$\tilde{O}(\kappa T^2)$
(Kuang et al., 2023)	$\tilde{O}(d^{1.5}H^{1.5}\sqrt{T})$	Linear MDP, episodic Delay	$\tilde{O}(T^2)$
(Haque et al., 2024)	$\tilde{O}(H^{1.5}d\sqrt{T})$	Linear MDP	$\tilde{O}(T^2/\sqrt{d})$
(Karbasi et al., 2023)	$\tilde{O}(d\varsigma\sqrt{T})$	Inf. horizon with span $\varsigma$ , $d \ll  \mathcal{S}  \mathcal{A} $ , strongly log-concave	$\tilde{O}(1)$ (due to log-conc.)
<b>LaPSRL (Lemma 2)</b>	$\tilde{O}(\sqrt{dHT})$	LSI $\mathbb{P}(M)$ , linear growth on $\alpha$	$\tilde{O}(T\tau + T^{1.5}\tau/d)$
<b>LaPSRL (Corollary 2)</b>	$\tilde{O}(\sqrt{Tg(\cdot)})$	LSI $\mathbb{P}(M)$ , policy with $\text{BR}(T) = \tilde{O}(\sqrt{T}g(\cdot))$ for exact post.	$\tilde{O}\left(\sum_{l=1}^{\tau} \frac{H^3 l^3}{\alpha_l^2} + \frac{dH^{4.5} l^{3.5}}{\alpha_l^2 g(\cdot)^2}\right)$

& Zegarlinski, 1992), and thus, motivates us to study isoperimetric distributions in RL. In addition, isoperimetric distributions include all the aforementioned setups studied in RL theory, and in addition, many non-log-concave and perturbed versions of log-concave distributions (Figure 1) 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 the optimization and sampling literature, isoperimetry is used as a minimal condition to conduct efficient and controlled sampling from target distribution(s) (Vempala & 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).** We focus on *Bayesian* RL, and in particular PSRL algorithms (Russo et al., 2020; Osband et al., 2013b). PSRL, also called Thompson sampling (Thompson, 1933), is a Bayesian algorithm requires a prior distribution to be defined over models and a posterior to be calculated as more data is collected. Periodically, a model is sampled from the posterior distribution, which is then used to create a policy. The policy is then used to act in the environment. PSRL has been successful both theoretically and experimentally, but its exact inference and sampling are only possible in very simple settings. On the other hand, naive approximations lead to linear regret. A recent line of research has examined how to obtain sublinear regret with approximate PSRL. However, this has so far been limited to either bandit problems or required strong assumptions. Our work is *more general*, both because it applies to general RL problems, and because the assumptions strictly generalise those of previous works.

**Langevin Sampling-based PSRLs.** A growing approach in this direction is to use Langevin-based approximate sampling methods, 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. Similarly, Xu et al. (2022) extend these ideas to linear contextual bandits and Ishfaq et al. (2024); Haque et al. (2024); Karbasi et al. (2023) bring Langevin-based PSRL to Markov Decision Processes (MDPs). A crucial limitation of all these works are that the regret guarantees are valid *only for log-concave or linear problems*, which we relax as shown in Table 1. Note that constant gradient complexity in (Karbasi et al., 2023) is possible due to strong log-concavity and exponential episode length, and is non-trivial to generalise.

However, the sampling literature has shown that Langevin methods are efficient for distribution fulfilling isoperimetric conditions, such as the ones satisfying LSI (Dubey et al., 2016; Vempala & Wibisono, 2019; Kinoshita & Suzuki, 2022). This motivates us to propose a generic algorithm that can work for any distribution satisfying LSI, both for bandits and MDPs, and also to study 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 result broadens the scenarios where PSRL is proven to be efficient to a new and wider family of posteriors. We show this for two cases: both when the data likelihood satisfies LSI and when the posterior distributions over MDPs satisfies LSI.
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  $\tilde{O}(\sqrt{T})$  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 (Kinoshita & Suzuki, 2022), which is an efficient Langevin sampling algorithm, we only need a polynomial number of samples with respect to the MDP parameters (both with and without chaining the samples). Conducting our regret analysis requires generalising the classical analysis of PSRL for LSI and studying the contraction of posteriors under Langevin dynamics.
3. Show that *LaPSRL deployed with SARAH-LD achieves sublinear regret across different environments*, including Gaussian, Mixtures of LSI distributions as well as any log-concave distribution or mixtures of them. We show that LaPSRL's regret is competitive w.r.t. the existing PSRL algorithms with approximate posteriors.
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. Numerical results validate that LaPSRL performs competitively with the baselines across environments and posterior distributions.

## 2. Problem Setup & 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.

**Notations.** Complexity notations  $\tilde{O}$ ,  $\tilde{\Omega}$ ,  $\tilde{\Theta}$  ignore sub- and poly-logarithmic terms. Table 2 summarises the notations.

**Problem Formulation: Episodic Reinforcement Learning (RL).** We consider episodic finite-horizon MDPs (aka *Episodic RL*) (Osband et al., 2013b; Azar et al., 2017). An MDP  $M$  is a tuple  $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, R, H \rangle$  with states  $s \in \mathcal{S} \subseteq \mathbb{R}^d$ , and actions  $a \in \mathcal{A}$ . In every episode, the agent interacts with the environment for  $H$  steps: The episode  $l$  starts at some 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  $s'_{l+1} \sim \mathcal{T}(\cdot | s_{l,t}, a_{l,t})$ . This interaction is done for a total of  $\tau$  episodes, which is commonly unknown a priori. We denote the total interactions with the environment as  $T \triangleq \tau H$ . When there is only one state, or the state does not depend on the action, this problem reduces to multi-armed bandits (Lattimore & Szepesvári, 2020).

The performance of a policy  $\pi$  is measured by the expected total reward  $V_1^\pi$  w.r.t. an initial state  $s$ . We define the value function at  $h \in [H]$  as  $V_M^{\pi,h}(s) \triangleq \mathbb{E}_{M,\pi} \left[ \sum_{t=h}^H R(s_t, a_t) | s_h = s \right]$ . The policy maximising these value functions is the optimal policy  $\pi^*$ .

**Background: Bayesian RL.** In the Bayesian setting, we first sample an MDP  $M$  from a prior distribution  $\mathbb{P}_0(M)$ . We then compute a policy for it and collect the data (aka history) by playing the policy, i.e.  $\mathcal{H}_l \triangleq \{s_{1,1}, a_{1,1}, \dots, s_{l-1,H}, a_{l-1,H}\}$ . We then construct a posterior distribution  $\mathbb{P}(M | \mathcal{H}_l)$  on  $M$  using  $\mathcal{H}_l$ .

Regret quantifies how much worse the learning policy is than an oracle policy. In Bayesian RL it is standard to consider the *Bayesian regret* (Osband et al., 2013b; Russo & Van Roy, 2014; O’Donoghue, 2021), i.e. the expected regret over the possible MDPs and trajectories. Given a prior  $\mathbb{P}_0(M)$  over MDPs, Bayesian regret of a policy  $\pi$  is

$$\text{BR}(T) \triangleq \mathbb{E}_{\mathbb{P}_0, \pi} \left[ \sum_{l=1}^{\tau} V_{\pi^*,1}^M(s_{l,1}) - V_{\pi_l,1}^M(s_{l,1}) \right]. \quad (1)$$

A sub-linear Bayesian regret means that the RL algorithm can solve all MDPs under prior, except for a set with measure zero (Osband et al., 2013b; Russo & Van Roy, 2014). To be specific, O’Donoghue (2021) states that sublinear Bayesian regret is preserved under a misspecified prior as long as the true prior is absolutely continuous w.r.t. the misspecified prior. Additionally, work has been done on setting these priors robustly as in (Buening et al., 2023). Our results do not just hold for a specifically chosen prior, but for *any prior* satisfying our assumptions.

---

**Algorithm 1** PSRL

---

```

1: Input: Likelihood  $\mathcal{L}(x|M)$ , Prior  $\mathbb{P}_0(M)$ 
2: for  $l = 1, 2, \dots$  do
3:   Sample  $M_l \sim \mathbb{P}(M | \mathcal{H}_l)$ 
4:   Play  $\pi^*(M_l)$  till horizon  $H$  to obtain  $\{x_i\}_{i=H(l-1)}^{Hl}$ 
5:    $\mathcal{H}_{l+1} \leftarrow \mathcal{H}_l \cup \{x_i\}_{i=H(l-1)+1}^{Hl}$ 
6: end for

```

---

A popular and successful Bayesian RL approach is to sample an MDP  $M_l \sim \mathbb{P}(M | \mathcal{H}_l)$  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., 2013b), or as Thompson sampling in bandits (Thompson, 1933). We illustrate a pseudocode of PSRL in Algorithm 1.

**Background: Sampling with Langevin dynamics.** Now, we discuss the other fundamental component of our work, i.e. Langevin sampling. This involves sampling from a target distribution  $d\nu \propto e^{-\gamma F(\theta)} d\theta$  over a set of parameters  $\theta \in \mathbb{R}^d$ , with  $F : \mathbb{R}^d \rightarrow \mathbb{R}$  defined as an  $n$ -sample average loss  $\frac{1}{n} \sum_{i=1}^n f_i(\theta)$  and  $\gamma = n$ , where  $f_i$  is the loss for sample  $x_i$ . In case of Bayesian posteriors, we define  $f_i(\theta) \triangleq -\frac{\log \mathbb{P}_0(\theta)}{n} - \log \mathcal{L}(x_i|\theta)$ , i.e. each  $f_i$  corresponds to the log-likelihood of data point  $x_i$  and its share of the log prior.

In continuous-time, Langevin methods can sample exactly from a posterior (Vempala & Wibisono, 2019). In practice, discretisation makes this impossible, but using a Langevin gradient descent algorithm allows for sampling from the target distribution with a controlled bias, under conditions of isoperimetry (Kinoshita & Suzuki, 2022). For this to work, we also need a property on smoothness:

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

We now start with an introduction to log-Sobolev distributions. There is a rich literature on log-Sobolev distributions—a summary of which can be found in (Chafaï & Lehec, 2023; Vempala & Wibisono, 2019).

**Definition 1** (log-Sobolev inequality). *A distribution  $\nu$  satisfies the log-Sobolev inequality (LSI) with a constant  $\alpha > 0$  if, for all smooth functions  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  with  $\mathbb{E}_\nu[g^2] \leq \infty$ ,*

$$\mathbb{E}_\nu[g^2 \log g^2] - \mathbb{E}_\nu[g^2] \log \mathbb{E}_\nu[g^2] \leq \frac{2}{\alpha} \mathbb{E}_\nu[\|\nabla g\|^2]. \quad (\text{LSI})$$

Obtaining the LSI constant<sup>1</sup>  $\alpha$  is not always trivial. We can determine whether a distribution is LSI through Lyapunov conditions, integral conditions, local inequalities and tools from optimal transport as well as mixture decomposition (Cattiaux et al., 2010; Wang, 2001; Barthe & Kolesnikov, 2008; Chen et al., 2021; Koehler et al., 2023). The most common tool is the Bakry-Émery criterion.

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

*Remark: Log-concavity vs. LSI.* Theorem 1 shows that log-concave distributions imply LSI. A distribution  $\nu(\theta)$  is log-concave if  $\log \nu(\theta)$  is concave in  $\theta$ . Log-concavity is a commonly used condition in sampling and RL (Mazumdar et al., 2020; Agrawal & Jia, 2017; Abeille & Lazaric, 2018) but this is significantly more restrictive than log-Sobolev. For example, log-concave distributions cannot be multimodal. Examples of LSI distributions given in Figure 1 show that multiple modes are generally not an issue for LSI.

An equivalent way of expressing (LSI) is  $\text{KL}(\rho \parallel \nu) \leq \frac{1}{2\alpha} J_\rho$ , where  $\rho \triangleq \frac{g^2 \nu}{\mathbb{E}_\nu[g^2]}$  and  $J_\rho \triangleq \mathbb{E}_\rho \left[ \left\| \nabla \log \frac{\rho}{\nu} \right\|^2 \right]$  is the relative Fisher information of  $\rho$  with respect to  $\nu$ .

Another operation on log-Sobolev distributions that preserves the LSI property while breaking log-concavity is a bounded perturbation. The result is due to (Holley & Stroock, 1987) but is presented here as per (Steiner, 2021).

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

---

<sup>1</sup>Note that some works use an inverse definition of LSI constant, i.e.  $\alpha' = \frac{1}{2\alpha}$ , leading to some confusion. We stick to Definition 1.

In some cases, unbounded perturbations to LSI distributions can still be LSI (Steiner, 2021). LSI is also preserved under a Lipschitz-transformation (Vempala & Wibisono, 2019), while log-concavity is not. Also, if the distribution is factorizable in log-Sobolev components, then the product is log-Sobolev with an LSI constant that is minimum among the factored components (Ledoux, 2006). Mixtures of log-Sobolev distributions are also log-Sobolev under conditions on the distance between the distributions (See Theorem 10).

*Sub-Gaussian Concentration from LSI.* An important consequence of a distribution  $\nu$  satisfying LSI with constant  $\alpha_\nu$  is obtaining the Gaussian concentration of any Lipschitz function around its mean (Bizeul, 2023). Specifically, for any function  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  with Lipschitz constant  $L_g$ ,

$$\mathbb{P}_\nu(|g - \mathbb{E}_\nu[g]| \geq t) \leq 2 \exp\left(-\frac{\alpha_\nu t^2}{L_g^2}\right). \quad (2)$$

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 & Suzuki, 2022). Multiple algorithms are developed to perform biased Langevin sampling on log-Sobolev distributions (Vempala & Wibisono, 2019; Kinoshita & Suzuki, 2022). In this paper, we focus on SARAH-LD (Algorithm 4), a variance-reduced version of Langevin dynamics that is state-of-the-art in terms of the KL divergence concentration between the sampled and the target distributions, i.e.  $\text{KL}(\hat{\nu} \parallel \nu)$ . SARAH-LD allows us to control the bias, and trade off the computational complexity w.r.t.  $\text{KL}(\hat{\nu} \parallel \nu)$ . Given  $n$  previous samples, SARAH-LD needs  $\tilde{O}\left(\left(n + \frac{d\sqrt{n}}{\text{KL}(\hat{\nu} \parallel \nu)}\right) \frac{1}{\alpha^2}\right)$  steps of stochastic gradient evaluations a new sample representative of the approximate posterior (also known as *gradient complexity*). The complete result is in Theorem 9.

### 3. PSRL for Exact posteriors

We first consider the convergence of posterior sampling (PSRL, Algorithm 1), when we have access to exact posterior distributions at each step. We ask: *can PSRL achieve sublinear regret if posterior or likelihood are isoperimetric?*

**Bounding Regret of PSRL.** First, we observe that following the series of works by (Osband & Van Roy, 2017; Chowdhury & Gopalan, 2019; Chowdhury et al., 2021), a generic three-step framework to bound the Bayesian regret ( $\text{BR}(T)$ ) of PSRL can be developed.

*Step 1:* At the first step of episode  $l$ , the total number of completed steps is  $n = (l-1)H$ . Osband et al. (2013b) observes that for any  $\sigma(\mathcal{H}_l)$  measurable function  $f$ , which includes the value function, we have  $\mathbb{E}[f(M_l)] = \mathbb{E}[f(M)]$ . Thus, we get from Equation (1) that  $\text{BR}(T) = \mathbb{E}\left[\sum_{l=1}^T V_{\pi_l,1}^{M_l}(s_{l,1}) - V_{\pi_l,1}^M(s_{l,1})\right]$ .

*Step 2:* Chowdhury & 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 between the next step value functions of the sampled and true MDPs. Specifically,  $\text{BR}(T) = \mathbb{E}\left[\sum_{l,h=1,1}^{\tau,H} \mathcal{T}_{M_l,h}^{\pi_l}(V_{\pi_l,h+1}^{M_l})(s_{l,h}) - \mathcal{T}_{M,h}^{\pi_l}(V_{\pi_l,h+1}^M)(s_{l,h})\right]$  and  $\mathcal{T}_{M,h}^{\pi_l}(V_{\pi_l,h+1}^M)(s_{l,h}) = R(s, \pi(s, h)) + \mathbb{E}_{s,\pi(s,h)}[V|M]$  is the Bellman operator at step  $h$  of the episode under policy  $\pi$  and MDP  $M$ .

*Step 3:* Finally, in the spirit of (Chowdhury et al., 2021), we use the transportation inequalities (Boucheron et al., 2003) to yield an upper bound of  $\text{BR}(T)$  as

$$\begin{aligned} \text{BR}(T) &\leq H\sigma_R \mathbb{E}\left[\sum_{l,h=1,1}^{\tau,H} \sqrt{2\text{KL}_{s_{l,h},a_{l,h}}(M||M_l)}\right] \\ &\leq B_R + 2H\sigma_R \sqrt{T\xi(T)}. \end{aligned}$$

Here,  $B_R$  and  $\sigma_R^2$  bound the mean and variance of rewards for any  $M$ , and  $\text{KL}_{s_{l,h},a_{l,h}}(M||M_l)$  denotes the KL divergence between  $\mathcal{T}_M(\cdot|s_{l,h}, a_{l,h})$  and  $\mathcal{T}_M(\cdot|s_{l,h}, a_{l,h})$ . *The last inequality holds if we can show that  $\text{KL}_{s_{l,h},a_{l,h}}(M, M_l)$  is upper bounded by a constant or monotonically increasing polylogarithmic function, say  $\xi(lh)$ , with probability at least  $1 - \frac{1}{lh}$  for any  $lh > 1$ .* Thus, our strategy for proving sublinear regret for PSRL is via such a concentration bound.

**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. In this setting we have  $n$  i.i.d. observations  $x^n = (x_1, \dots, x_n)$  sampled from a distribution with parameter  $\theta^*$ :  $x_i \sim \mathcal{L}(x|\theta^*)$ . We assume that the prior  $\mathbb{P}_0$  has a non-negligible mass around  $\theta^*$ :

**Assumption 2** (True-mass Prior). *There exists a measure  $\omega$  such that  $\mathbb{P}_0$  has non-zero mass in all closed, compact sets  $\Xi \subset \Theta$  around  $\theta^*$ :  $\int_{\theta \in \Xi} e^{-L_\theta \|\theta^* - \theta\|} \mathbb{P}_0(\theta) d\theta \geq \omega(\Xi)$ .*

This condition allows deriving the concentration result.

**Theorem 3.** *If the data likelihood  $\mathcal{L}(x | \theta^*)$  is  $\alpha_{\theta^*}$ -LSI, Assumption 2 holds, and the log-likelihood function  $\log \mathcal{L}(x | \theta)$  is  $L_x$  and  $L_\theta$  Lipschitz in  $x$  and  $\theta$ , respectively. Then, for any  $n > 0$  and  $\theta \sim \mathbb{P}(\theta | x^n)$ , we obtain with probability at least  $1 - 2\delta$  (for  $\delta \in (0, 1/2]$ )*

$$\text{KL}(\mathcal{L}(x|\theta^*) || \mathcal{L}(x|\theta)) \leq \sqrt{\frac{4L_x^2 \ln(\frac{2}{\delta})}{n\alpha_{\theta^*}}} + \frac{\ln(\frac{1}{\delta})}{n}. \quad (3)$$

The detailed proof is in Appendix C.1. This result shows a generic concentration of the posteriors for data likelihoods obeying LSI, and thus demands wider interest for any Bayesian learning framework. In RL, the data  $x$  is the state-action pairs  $(s_t, a_t)$  and  $\theta$  represents the MDP  $M$  while the data is generated by a policy  $\pi$  from  $M$ .

**Remark 1** (Prior Design). *Designing priors that have a small ball probability around the true parameter is common in Bayesian learning. For example, (Castillo et al., 2015) proposes such a prior for efficient Bayesian learning, and (Chakraborty et al., 2023) carefully designed prior to prove near-optimal learning in high-dimensional bandits.*

**Remark 2** (Lipschitz Log-likelihood). *To prove this result, we need an additional assumption of Lipschitzness of the log-likelihood function with respect to the data and the parameter. This holds true for most parametric distributions that does not consist of a Dirac distribution. For example, general exponential family distributions satisfy Lipschitzness 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.*

**Sublinear Regret of PSRL.** Now, we apply Theorem 3 to PSRL and bound its Bayesian regret.

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

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

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 be reduced to  $\sqrt{T}$  if the posterior  $\mathbb{P}(M | \mathcal{H}_t)$  satisfies LSI with linearly increasing constants. We formalise this observation now.

**Near-optimal Regret: Linear LSI Constant of Posteriors.** From (Chowdhury & Gopalan, 2019), we observe that if we assume the next step value functions  $\mathcal{T}_{M,h}^\pi(V_{\pi,h+1}^M)(s_{l,h})$  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  $\{\alpha_{\bar{T},l}\}$ , the mean reward for any MDP  $M$  is bounded:  $|\bar{R}_M(s)| \leq B_R \forall s$ , the one step value function is Lipschitz in the state with parameter  $L_M$  as given in Assumption 3, and the mean reward and mean transitions are  $L_{\bar{R}}$  and  $L_{\bar{T}}$  Lipschitz in  $M$ , Bayesian regret of PSRL is bounded by*

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

**Algorithm 2** Langevin PSRL (LaPSRL)

---

```

1: Input: Prior  $\mathbb{P}_0(M)$ , Horizon  $H$ , Regret order  $g(H, \mathcal{S}, \mathcal{A})$ , Likelihood function  $\mathcal{L}(x|M)$ .
2: for  $l = 1, 2, \dots$  do
3:    $\epsilon_{\text{post},l} = \frac{g(H, \mathcal{S}, \mathcal{A})}{l\Delta_{\max}^2}$ 
4:   if Chained sampling then
5:      $\rho_0 = M_{l-1}$   $\triangleright$  Reuse last sample from step  $l-1$ 
6:      $M_l \sim \text{LS}(\mathcal{L}(x|M), \hat{\mathbb{P}}(M|\mathcal{H}_{l-1}), \mathcal{H}_l, \epsilon_{\text{post},l}, \rho_0)$ 
7:   else
8:      $\rho_0 \sim \mathbb{P}_0(M)$   $\triangleright$  Resample from prior
9:      $M_l \sim \text{LS}(\mathcal{L}(x|M), \mathbb{P}_0(M), \mathcal{H}_l, \epsilon_{\text{post},l}, \rho_0)$ 
10:  end if
11:  Play  $\pi^*(M_l)$  until horizon  $H$  obtaining data  $\mathcal{H}_{l+1} = \mathcal{H}_l \cup \{(s_i, a_i)\}_{i=H(l-1)+1}^{Hl}$ .
12: end for

```

---

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

## 4. LaPSRL for Approximate Posteriors

In practice, the main bottleneck to deploy PSRL is constructing an exact posterior from high-dimensional data and data without any parametric assumption. In these cases, the common strategy is to approximate the posterior distribution computationally efficiently and use it to sample further. But we know that constant approximation error, in terms of  $\alpha$ -divergence, on the posterior leads to linear regret in the context of Thompson sampling for multi-armed bandits (Phan et al., 2019). This also happens in other RL setups, like contextual bandits and linear MDPs (Simchowitz et al., 2021). Previous work has noted (Mazumdar et al., 2020) that proper decay of this error can allow for sublinear regret in multi-armed bandits. Mazumdar et al.; Zheng et al.; Karbasi et al. designed approximate algorithms for multi-armed bandits and strongly log-concave posteriors. However, (Karbasi et al., 2023; Ishfaq et al., 2024; Haque et al., 2024) include planning over episodes, which is essential for MDPs, but the regret analysis heavily depends on strong log-concavity and linearity 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 isoperimetry (LSI in Definition 1) instead of log-concavity. To start, we derive Theorem 5 to control the error rate of concentration of posteriors in RL.

**Theorem 5.** *Let us sample an MDP from an approximate posterior  $\mathbb{Q}_l$  in episode  $l$  and use it for planning. If  $\mathbb{P}_l$  is the true posterior at  $l$  and  $\min(\text{KL}((\mathbb{P}_l \parallel \mathbb{Q}_l), \text{KL}(\mathbb{Q}_l \parallel \mathbb{P}_l)) \leq \epsilon_{\text{post},l}$ , then regret in an episode due to the approximate posterior is  $\mathcal{O}(HB_R\sqrt{\epsilon_{\text{post},l}})$ .*

This holds because KL-divergence of posterior controls the growth of Bayesian regret. Proof is in Appendix D.

**Corollary 1.** *If an algorithm incurs  $\tilde{\mathcal{O}}(\sqrt{T}g(H, \mathcal{S}, \mathcal{A}))$  regret for the true posteriors, it will incur the same order of regret for the approximate posteriors if  $\epsilon_{\text{post},l} \leq C \frac{g(H, \mathcal{S}, \mathcal{A})^2}{l\Delta_{\max}^2}$  for some  $C > 0$ . Here,  $\Delta_{\max} \triangleq \max_{\pi} V_{\pi,1}^M(s_1) - \min_{\pi} V_{\pi,1}^M(s_1) \leq 2HB_R$  is maximal regret in an episode.*

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  $\tilde{\mathcal{O}}(\sqrt{T})$  regret by running PSRL with such posteriors.

**LaPSRL.** With these results in mind, we design an algorithm, Langevin PSRL (LaPSRL). LaPSRL is detailed in Algorithm 2 and its sampling routine is in Algorithm 3. LaPSRL extends the PSRL template. In each episode  $l$ , a tolerable error  $\epsilon_{\text{post},l}$  is calculated. Then we use SARAH-LD to sample a  $M_l \in \mathbb{R}^D$ . Depending on the task at hand, SARAH-LD calculates the required step size and learning rate to reach the acceptable error in KL distance, and finally returns a desired sample. This sample is used to obtain an optimal policy, which is then played for that episode. We have two options for initializing the sampling in each episode, from the prior or reusing the previous sample (chained sample setting). Parametrisation of MDP  $M_l \in \mathbb{R}^D$  is not restricted but the LSI constant depends on it.

**Gradient Complexity.** By combining Theorem 5 with log-Sobolev theory and SARAH-LD, we obtain order-optimal

**Algorithm 3** LANGEVIN SAMPLE (LS)

- 
- 1: **Input:** Prior/posterior  $\mathbb{P}(M)$ , data  $\mathcal{H}_l$ , acceptable error  $\epsilon_{\text{post},l}$ , initial sample  $\rho_0$ , Likelihood function  $\mathcal{L}(x|M)$ .
  - 2: Learning rate  $\eta_l \leftarrow \min \left( \frac{\alpha_l}{16\sqrt{2}L^2|\mathcal{H}_l|^{3/2}}, \frac{3\alpha_l\epsilon_{\text{post},l}}{320dL^2|\mathcal{H}_l|} \right)$
  - 3: #Steps  $k_l \leftarrow \frac{|\mathcal{H}_l|}{\alpha_l\eta_l} \log \frac{2\text{KL}(\mathbb{P}(M) \parallel \mathbb{P}(M|\mathcal{H}_l))}{\epsilon_{\text{post},l}}$
  - 4: **Return**  $M_l \leftarrow \text{SARAH-LD}(\mathcal{L}(x|M), \mathcal{H}_l, \mathbb{P}(M), k_l, \eta_l)$
- 

Bayesian regret for any log-Sobolev posterior while limiting the computational gradient complexity of each episode to a low degree polynomial. Gradient complexity signifies the number of gradient steps  $\nabla_M \mathcal{L}(x|M)$  need to be performed.

**Corollary 2.** *For a posterior fulfilling the Assumption 1 and Definition 1, LaPSRL obtains the same order of regret as PSRL while SARAH-LD incurs a gradient complexity  $\tilde{O}\left(\frac{H^3l^3L^2}{\alpha_l^2} + \frac{dB_R^2H^{4.5}l^{3.5}L^2}{\alpha_l^2g(H,\mathcal{S},\mathcal{A})^2}\right)$  in episode  $l$ .*

**Lemma 2.** *For LSI posteriors (as in Theorem 4) with linearly growing LSI constants  $\alpha_l = \Omega(Hl)$ , the total gradient complexity of LaPSRL is  $\tilde{O}\left(\tau T + \tau T^{1.5}/\sqrt{d}\right)$  that yields regret  $\mathcal{O}(\sqrt{dHT})$ .*

Proofs are deferred to Appendix D. This implies that the computational complexity is sub-quadratic for the examples in Section 5. This could be improved by  $\tau$  using exponentially increasing episode lengths, requiring modifications to the regret proofs. This result can be compared with (Haque et al., 2024) that obtains a total sample complexity of  $\tilde{O}(T^2/\sqrt{d})$ , but the analysis is instantiated only for the linear MDPs with an additional  $H\sqrt{d}$  term in the regret compared to LaPSRL. Note that the regret obtained by Haque et al. is expected regret, but this can be transformed into Bayesian regret of the same order.

**Chained Samples.** The sample complexity to achieve an  $\epsilon$  approximation  $\hat{\nu}$  of  $\nu$  is controlled by  $\text{KL}(\hat{\nu} \parallel \nu)$ . The naïve approach is sampling only from a prior  $\mathbb{P}_0$ , such as an isotropic Gaussian, and the dependence is only logarithmic in  $\text{KL}(\mathbb{P}_0 \parallel \nu)$ . An alternative is to use the final sample from the previous time step as initialization for the next one. This can be seen as sampling  $\rho_0$  from the  $\epsilon_{\text{post},l-1}$ -approximate posterior  $\hat{\mathbb{P}}(M|\mathcal{H}_{l-1})$ . This allows for a more practical algorithm as it might be easier to shrink the divergence between two consecutive posteriors than between the prior and a posterior. We show that reusing samples bounds the KL distance to a function of the variance of  $M$ .

**Theorem 6.** *If  $\nabla_z \log \mathcal{L}(z|M)$  is  $L_z$ -Lipschitz and  $\alpha_z$ -Log Sobolev, with  $z$  being the data corresponding to an episode,*

$$\mathbb{E}_{\mathcal{L}(z|\mathcal{H}_l)} \text{KL}(\hat{\mathbb{P}}(M|\mathcal{H}_l) \parallel \mathbb{P}(M|\mathcal{H}_{l+1})) \leq \epsilon_{\text{post},l} + \frac{L_z^2}{\alpha_z} \epsilon_{\text{post},l}^2 \text{Var}(\mathbb{P}(M|\mathcal{H}_l)) + \frac{L_z^2}{2\alpha_z} \text{Var}(\hat{\mathbb{P}}(M|\mathcal{H}_l)), \quad (4)$$

where  $\text{Var}(P)$  is the variance of the distribution  $P$ . Note that LaPSRL ensures  $\epsilon_{\text{post},l} = \mathcal{O}(1/l)$ .

Chaining correlates the sampled parameters. Since Bayesian regret considers expectation, this does not affect the order of the regret. Rest is to bound the two variance terms. The variance of the true posterior  $\text{Var}(\mathbb{P}(M|\mathcal{H}_l))$  is bounded by  $(1-s)^l \text{Var}(\mathbb{P}_0(M))$ , where  $s \in (0, 1)$  is the posterior contraction rate of the posterior sampling (Ghosal, 1997; Wang & Ghosal, 2023; Mou et al., 2024). The remaining challenge is to control the variance of the approximate posterior  $\hat{\mathbb{P}}(M|\mathcal{H}_l)$ . But in practice, we know that the variance of the posterior distributions tends to decay as more data is observed, ensuring decay in  $\text{KL}(\hat{\mathbb{P}}(M|\mathcal{H}_l) \parallel \mathbb{P}(M|\mathcal{H}_{l+1}))$ . Specifically, for locally and globally log-concave distributions, it is known to decay with  $1/l$  for Langevin samplers (Mou et al., 2024; Zheng et al., 2024). But providing such a control for LSI demands analysing Langevin samplers independent of RL, which we defer to future work. In practice, we achieve control over posterior concentration and thus, regret of LaPSRL with/without chained samples.

## 5. Distributions with Linear LSI Constants

Now, we study LSI constants for families of distributions and apply Theorem 4 to calculate the Bayesian regret of PSRL for corresponding posteriors.

**Univariate Gaussian.** For illustration, we calculate the LSI constants for a Gaussian posterior with known variance  $\sigma^2$ . Here, we assume a Gaussian  $(0, \sigma_0^2)$  prior over the mean  $\mu$ . Then,  $\mathbb{P}(\mu|\mathcal{H}_n) \propto \exp\left\{-\left(\sum_{i=1}^n \left(\frac{\mu^2}{2n\sigma_0^2} + \frac{(\mu-x_i)^2}{2\sigma^2}\right)\right)\right\}$ . Then, we have  $\gamma = n$ ,  $f_i(\mu) = \left(\frac{\mu^2}{2n\sigma_0^2} + \frac{(\mu-x_i)^2}{2\sigma^2}\right)$ . Note that  $\nabla_\mu^2 f_i(\mu) = \frac{1}{n\sigma_0^2} + \frac{1}{\sigma^2} \leq L$ . Finally, we use Theorem 1 to calculate

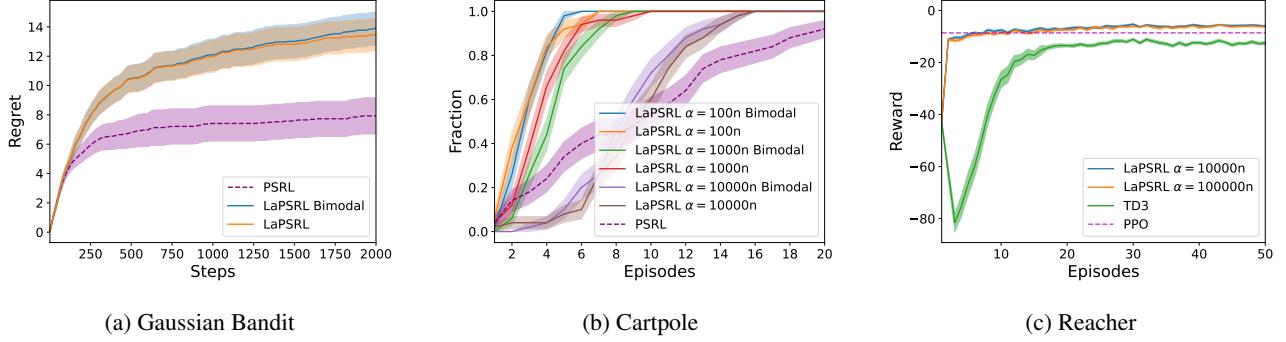


Figure 2: We compare LaPSRL against baselines. In the bandit and Cartpole experiments, we benchmark with PSRL, and in Reacher with TD3 and PPO. For the Gaussian bandits, we compare the expected regret and for Cartpole we evaluate how many episodes it takes to solve the task. Finally, in Reacher, we study the average regret per episode. In all environments, we average over 50 runs with the standard error highlighted around the average. Larger plots are in Appendix F.2.

$\alpha$ . Since  $\|\nabla_\mu^2 f_i(\mu)\|$  is independent of  $i$ , 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  $\alpha = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} = \Theta(n)$ .

**Corollary 3.** *PSRL and LaPSRL obtain  $BR(T) = \tilde{O}\left(\sqrt{T\sigma^2}\right)$  with univariate Gaussian posteriors.*

#### Log-concave and Mixture of Log-concave Distributions.

**Theorem 7.** *(a) Any log-concave posterior fulfills LSI with  $\alpha_n = \Theta(n)$ . (b) Any posterior that is a mixture of  $k$  log-concave distributions has  $\alpha_n^{Mixture} = \Omega\left(\frac{n \min p_i}{4k(1-\log(\min p_i))}\right)$ .*

This result comes from the superadditivity of minimum eigenvalues of Hessians, and thus, LSI constants for log-concave distributions. The result for mixtures follows from Theorem 10 (Appendix E). Combining Theorem 4 and Theorem 7, we obtain the sublinear regret bound.

**Corollary 4.** *If  $|\bar{R}_M(s)| \leq B_R \forall s$ , any log-concave posterior over MDPs  $M$  yields  $BR(T) = \tilde{O}\left(\sqrt{T}\right)$  for PSRL. Under the same conditions, PSRL obtains for any posterior that is a mixture of  $k$  log-concave posteriors with non-zero overlap  $BR(T) = \tilde{O}\left(\sqrt{\frac{4kT}{\min p_i}}\right)$ .*

#### The Final Goal: General Log-Sobolev Distributions.

**Theorem 8.** *A log-Sobolev distribution with bounded likelihood ratio:  $|\log \frac{\mathcal{L}(x|M)}{\mathcal{L}(x|M')}| \leq \Gamma$ , 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 could be possible. Similar assumptions of linear growth of constants are made in works involving other isoperimetric inequalities, namely Poincaré inequality, but in the offline setting (Haddouche et al., 2024). This also matches the intuition from the asymptotic results of the Bernstein–von Mises theorem which gives a log-Sobolev constant of  $\Theta(n)$  as  $n \rightarrow \infty$ . For now, in these settings, we can rely on the results in Lemma 1 exhibiting sublinear 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 setup on the Cartpole environment and finally with a neural network

on the Reacher environment. On Cartpole and Bandits, we additionally test a bimodal prior over the arms to demonstrate a non log-concave setting. Additional experimental details are found 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 follow the approximate shape of 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  $\alpha$  (See appendix). The Cartpole environment (Barto et al., 1983) is modified to have continuous action and we use parameterize the model as a Linear Quadratic Regulator model (Kalman, 1960). We use a PSRL algorithm which samples from Bayesian linear regression priors (Minka, 2000) as a baseline. The results from this experiment can be found in Figure 2(b) where we plot what fraction of the 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. It also takes longer for the experiments with larger  $\alpha$  values to converge. The slower convergence of PSRL can be due to the priors not being the same.

Reacher is a standard environment from the Gymnasium library (Towers et al., 2024). Here we use a neural network model, 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) and PPO (Schulman et al., 2017). Figure 2(c) shows that LaPSRL learns very quickly, while TD3 is unable to learn a policy that is equally good in a similar time frame. Due to no experience replay, PPO is significantly less data efficient. Thus, we present its average performance finally after 7000 episodes.

**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 perform well for varying domains and settings.

## 7. Extended Related Works

In addition to the works discussed previously, we present an overview of the related works with approximate posteriors.

Thompson sampling often requires approximations when exact posterior calculations or sampling become intractable (Wang et al., 2023; Sasso et al., 2023; Osband et al., 2023). While various algorithms for approximate sampling exist, they generally lack regret guarantees. Huang et al. (2023) does provide Bayesian regret bounds for approximate upper confidence bounds in bandits.

Recent works have extensively explored Langevin methods in bandits and RL (Kim, 2023; Dwaracherla & Van Roy, 2020; Yamamoto et al., 2024; Anonymous, 2025) but without regret guarantees. Further studies have applied these methods to specific settings: offline RL (Nguyen-Tang et al., 2024) and inverse RL (Krishnamurthy & Yin, 2021). Similarly, Hsu et al. (2024) establish regret bounds for multi-agent RL, but only under linear function approximations. (Kim et al., 2024) do the same for LQR systems under strongly log-concave assumptions. In contrast, our work provides a more general framework that overcomes these restrictive and eclectic assumptions of linearity, discreteness, or log-concavity. To conclude, there are works looking into using Langevin methods for RL but *none that comes with regret guarantees for a setting as general as LSI distributions*.

More work on Bayesian reinforcement learning includes (O'Donoghue et al., 2018; Osband et al., 2013a; 2023; Jorge et al., 2020; Dimitrakakis, 2011; O' Donoghue, 2021; Fellows et al., 2021; Luis et al., 2023; 2024b;a; Zintgraf et al., 2021; Chua et al., 2018; Fan & Ming, 2021; Eriksson et al., 2022; Ghavamzadeh et al., 2015; Cronrath et al., 2018; Dearden et al., 1999; Reisinger et al., 2008). A few of these works come with regret results, but they are limited to cases with easy posterior updates such as linear models or discrete state and actions spaces (Grover et al., 2020).

## 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  $\tilde{O}(\sqrt{T})$  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  $\tilde{O}(\sqrt{T})$  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 the 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 provided by the National Academic Infrastructure for Supercomputing in Sweden (NAIIS), 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

- Abeille, M. and Lazaric, A. Improved regret bounds for thompson sampling in linear quadratic control problems. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1–9. PMLR, 10–15 Jul 2018. URL <http://proceedings.mlr.press/v80/abeille18a.html>. 4
- Agrawal, S. and Jia, R. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper/2017/file/3621f1454cacf995530ea53652ddf8fb-Paper.pdf>. 4
- Anonymous. Langevin soft actor-critic: Efficient exploration through uncertainty-driven critic learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=FvQsk3la17>. 10
- Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017. 3
- Bakry, D., Gentil, I., Ledoux, M., et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014. 2, 5
- Barthe, F. and Kolesnikov, A. V. Mass transport and variants of the logarithmic sobolev inequality. *Journal of Geometric Analysis*, 18(4):921–979, 2008. 4
- Barto, A. G., Sutton, R. S., and Anderson, C. W. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):834–846, 1983. doi: 10.1109/TSMC.1983.6313077. 10, 26, 28
- Bizeul, P. On the log-sobolev constant of log-concave measures, 2023. 5
- Boucheron, S., Lugosi, G., and Bousquet, O. Concentration inequalities. In *Summer school on machine learning*, pp. 208–240. Springer, 2003. 5, 21
- 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. JAX: composable transformations of Python+NumPy programs, 2018. URL <http://github.com/jax-ml/jax>. 28
- Buening, T. K., Dimitrakakis, C., Eriksson, H., Grover, D., and Jorge, E. Minimax-bayes reinforcement learning. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and*

- Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7511–7527. PMLR, 25–27 Apr 2023. URL <https://proceedings.mlr.press/v206/buening23a.html>. 3
- Castillo, I., Schmidt-Hieber, J., and Van der Vaart, A. Bayesian linear regression with sparse priors. *The Annals of Statistics*, pp. 1986–2018, 2015. 6
- Cattiaux, P., Guillin, A., and Wu, L.-M. A note on talagrand’s transportation inequality and logarithmic sobolev inequality. *Probability theory and related fields*, 148:285–304, 2010. 4
- Chafaï, D. and Lehec, J. Logarithmic sobolev inequalities essentials, 2023. URL <https://djalil.chafai.net/docs/M2/chafai-lehec-m2-lsie-lecture-notes.pdf>. Accessed on 08/10/2024. 4
- Chakraborty, S., Roy, S., and Tewari, A. Thompson sampling for high-dimensional sparse linear contextual bandits. In *International Conference on Machine Learning*, pp. 3979–4008. PMLR, 2023. 6
- Chen, H.-B., Chewi, S., and Niles-Weed, J. Dimension-free log-sobolev inequalities for mixture distributions. *Journal of Functional Analysis*, 281(11):109236, 2021. ISSN 0022-1236. doi: <https://doi.org/10.1016/j.jfa.2021.109236>. 4, 25
- Chowdhury, S. R. and Gopalan, A. Online learning in kernelized markov decision processes, 2019. 1, 2, 5, 6, 21, 22
- Chowdhury, S. R., Gopalan, A., and Maillard, O.-A. Reinforcement learning in parametric mdps with exponential families. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1855–1863. PMLR, 13–15 Apr 2021. 5
- Chua, K., Calandra, R., McAllister, R., and Levine, S. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 10
- Cronrath, C., Jorge, E., Moberg, J., Jirstrand, M., and Lennartson, B. Bagger: A bayesian algorithm for safe and query-efficient imitation learning. 2018. 10
- Dearden, R., Friedman, N., and Andre, D. Model based Bayesian exploration. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pp. 150–159, 1999. 10
- Dimitrakakis, C. Robust Bayesian reinforcement learning through tight lower bounds. In *European Workshop on Reinforcement Learning (EWRL 2011)*, pp. 177–188, 2011. 10
- Dubey, K. A., J Reddi, S., Williamson, S. A., Poczos, B., Smola, A. J., and Xing, E. P. Variance reduction in stochastic gradient langevin dynamics. *Advances in neural information processing systems*, 29, 2016. 2
- Dwaracherla, V. and Van Roy, B. Langevin dqn. *arXiv preprint arXiv:2002.07282*, 2020. 10
- Efron, B. *Exponential families in theory and practice*. Cambridge University Press, 2022. 6
- Eriksson, H., Basu, D., Alibeigi, M., and Dimitrakakis, C. Sentinel: taming uncertainty with ensemble based distributional reinforcement learning. In Cussens, J. and Zhang, K. (eds.), *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 631–640. PMLR, 01–05 Aug 2022. URL <https://proceedings.mlr.press/v180/eriksson22a.html>. 10
- Fan, Y. and Ming, Y. Model-based reinforcement learning for continuous control with posterior sampling. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3078–3087. PMLR, 18–24 Jul 2021. URL <http://proceedings.mlr.press/v139/fan21b.html>. 2, 10
- Fellows, M., Hartikainen, K., and Whiteson, S. Bayesian bellman operators. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 13641–13656. Curran Associates, Inc., 2021. URL [https://proceedings.neurips.cc/paper\\_files/paper/2021/file/7180cffd6a8e829dacfc2a31b3f72ece-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2021/file/7180cffd6a8e829dacfc2a31b3f72ece-Paper.pdf). 10
- Ferguson, T. S. Location and scale parameters in exponential families of distributions. *The Annals of mathematical statistics*, 33(3):986–1001, 1962. 6

- Fujimoto, S., Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pp. 1587–1596. PMLR, 2018. [10](#), [27](#), [28](#)
- Geramifard, A., Walsh, T. J., Tellex, S., Chowdhary, G., Roy, N., and How, J. P. A tutorial on linear function approximators for dynamic programming and reinforcement learning. *Foundations and Trends® in Machine Learning*, 6(4):375–451, 2013. ISSN 1935-8237. doi: 10.1561/2200000042. [1](#)
- Ghavamzadeh, M., Mannor, S., Pineau, J., and Tamar, A. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(5–6):359–483, 2015. ISSN 1935-8245. doi: 10.1561/2200000049. URL <http://dx.doi.org/10.1561/2200000049>. [10](#)
- Ghosal, S. A review of consistency and convergence of posterior distribution. In *Varanashi Symposium in Bayesian Inference, Banaras Hindu University*. Citeseer, 1997. [8](#)
- Grover, D., Basu, D., and Dimitrakakis, C. Bayesian reinforcement learning via deep, sparse sampling. In *International Conference on Artificial Intelligence and Statistics*, pp. 3036–3045. PMLR, 2020. [10](#)
- Haddouche, M., Viallard, P., Simsekli, U., and Guedj, B. A pac-bayesian link between generalisation and flat minima. *arXiv preprint arXiv:2402.08508*, 2024. [9](#)
- Haque, I., Tan, Y., Yang, Y., Lan, Q., Lu, J., Mahmood, A. R., Precup, D., and Xu, P. More efficient randomized exploration for reinforcement learning via approximate sampling. *Reinforcement Learning Journal*, 3(1), 2024. [2](#), [7](#), [8](#)
- Holley, R. and Stroock, D. Logarithmic sobolev inequalities and stochastic ising models. *Journal of Statistical Physics*, 46 (5-6):1159–1194, 1987. [4](#)
- Hsu, H.-L., Wang, W., Pajic, M., and Xu, P. Randomized exploration in cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:2404.10728*, 2024. [10](#)
- Huang, S., Dossa, R. F. J., Ye, C., Braga, J., Chakraborty, D., Mehta, K., and Araújo, J. G. Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms. *Journal of Machine Learning Research*, 23(274):1–18, 2022. URL <http://jmlr.org/papers/v23/21-1342.html>. [28](#)
- Huang, Z., Lam, H., Meisami, A., and Zhang, H. 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*, 2023. URL <https://openreview.net/forum?id=vwr4bHHsRT>. [10](#)
- Huix, T., Zhang, M., and Durmus, A. Tight regret and complexity bounds for thompson sampling via langevin monte carlo. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 8749–8770. PMLR, 25–27 Apr 2023. URL <https://proceedings.mlr.press/v206/huix23a.html>. [2](#)
- Ishfaq, H., Lan, Q., Xu, P., Mahmood, A. R., Precup, D., Anandkumar, A., and Azizzadenesheli, K. Provable and practical: Efficient exploration in reinforcement learning via langevin monte carlo. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=nfIAEJFiBZ>. [2](#), [7](#)
- Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. A short note on concentration inequalities for random vectors with subgaussian norm, 2019. URL <https://arxiv.org/abs/1902.03736>. [19](#)
- Jorge, E., Eriksson, H., Dimitrakakis, C., Basu, D., and Grover, D. Inferential induction: A novel framework for Bayesian reinforcement learning. In *Proceedings on "I Can't Believe It's Not Better!" at NeurIPS Workshops*, volume 137 of *Proceedings of Machine Learning Research*, pp. 43–52. PMLR, 12 Dec 2020. URL <http://proceedings.mlr.press/v137/jorge20a.html>. [10](#)
- Kalman, R. E. A new approach to linear filtering and prediction problems. 1960. [10](#)
- Karbasi, A., Kuang, N. L., Ma, Y., and Mitra, S. Langevin thompson sampling with logarithmic communication: Bandits and reinforcement learning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15828–15860. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/karbasi23a.html>. [2](#), [7](#)

- Kim, G. Learning linear-quadratic regulators via thompson sampling with preconditioned langevin dynamics, 2023. 10
- Kim, Y., Kim, G., and Yang, I. Approximate thompson sampling for learning linear quadratic regulators with  $o(\sqrt{T})$  regret. *arXiv preprint arXiv:2405.19380*, 2024. 10
- Kinoshita, Y. and Suzuki, T. 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. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL [https://openreview.net/forum?id=Sj2z\\_i1wX-](https://openreview.net/forum?id=Sj2z_i1wX-). 2, 3, 4, 5, 18
- Koehler, F. and Vuong, T.-D. Sampling multimodal distributions with the vanilla score: Benefits of data-based initialization. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=oAMArMMQxb>. 25
- Koehler, F., Heckett, A., and Risteski, A. Statistical efficiency of score matching: The view from isoperimetry. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=TD7AnQjNzR6>. 4
- Krishnamurthy, V. and Yin, G. Langevin dynamics for adaptive inverse reinforcement learning of stochastic gradient algorithms. *Journal of Machine Learning Research*, 22(121):1–49, 2021. URL <http://jmlr.org/papers/v22/20-625.html>. 10
- Kuang, N. L., Yin, M., Wang, M., Wang, Y.-X., and Ma, Y. Posterior sampling with delayed feedback for reinforcement learning with linear function approximation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=RiyH3z7oIF>. 2
- Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020. 3
- Ledoux, M. Concentration of measure and logarithmic sobolev inequalities. In *Seminaire de probabilités XXXIII*, pp. 120–216. Springer, 2006. 2, 5
- Luis, C. E., Bottero, A. G., Vinogradska, J., Berkenkamp, F., and Peters, J. Model-based uncertainty in value functions. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 8029–8052. PMLR, 25–27 Apr 2023. URL <https://proceedings.mlr.press/v206/luis23a.html>. 10
- Luis, C. E., Bottero, A. G., Vinogradska, J., Berkenkamp, F., and Peters, J. Value-distributional model-based reinforcement learning. *Journal of Machine Learning Research*, 25(298):1–42, 2024a. URL <http://jmlr.org/papers/v25/23-0913.html>. 10
- Luis, C. E., Bottero, A. G., Vinogradska, J., Berkenkamp, F., and Peters, J. Model-based epistemic variance of values for risk-aware policy optimization, 2024b. URL <https://arxiv.org/abs/2312.04386>. 10
- Mazumdar, E., Pacchiano, A., Ma, Y., Jordan, M., and Bartlett, P. On approximate thompson sampling with langevin algorithms. In *International Conference on Machine Learning*, pp. 6797–6807. PMLR, 2020. 2, 4, 7
- Minka, T. Bayesian linear regression. Technical report, Citeseer, 2000. 10, 26
- Moradipari, A., Pedramfar, M., Zini, M. S., and Aggarwal, V. Improved bayesian regret bounds for thompson sampling in reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=2EVTB1idyR>. 2
- Mou, W., Ho, N., Wainwright, M., Bartlett, P., and Jordan, M. A diffusion process perspective on posterior contraction rates for parameters. *SIAM Journal on Mathematics of Data Science*, 6(2):553–577, 2024. 8
- Nguyen-Tang, T., Yin, M., Uehara, M., Wang, Y.-X., Wang, M., and Arora, R. Posterior sampling via langevin monte carlo for offline reinforcement learning, 2024. URL <https://openreview.net/forum?id=WwCirclMvl>. 10

- Nitanda, A., Wu, D., and Suzuki, T. Convex analysis of the mean field langevin dynamics. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 9741–9757. PMLR, 28–30 Mar 2022. URL <https://proceedings.mlr.press/v151/nitanda22a.html>. 2
- O’Donoghue, B. Variational bayesian reinforcement learning with regret bounds. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 28208–28221. Curran Associates, Inc., 2021. URL [https://proceedings.neurips.cc/paper\\_files/paper/2021/file/edb446b67d69adbfe9a21068982000c2-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2021/file/edb446b67d69adbfe9a21068982000c2-Paper.pdf). 3, 10
- O’Donoghue, B., Osband, I., Munos, R., and Mnih, V. The uncertainty bellman equation and exploration. In *International Conference on Machine Learning*, pp. 3836–3845, 2018. 10
- Osband, I. and Van Roy, B. Model-based reinforcement learning and the eluder dimension. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL [https://proceedings.neurips.cc/paper\\_files/paper/2014/file/1141938ba2c2b13f5505d7c424ebae5f-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2014/file/1141938ba2c2b13f5505d7c424ebae5f-Paper.pdf). 22
- Osband, I. and Van Roy, B. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, pp. 2701–2710. PMLR, 2017. 1, 5
- Osband, I., Russo, D., and Roy, B. V. (more) efficient reinforcement learning via posterior sampling. In *NIPS*, 2013a. 10
- Osband, I., Russo, D., and Van Roy, B. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013b. 2, 3, 4, 5, 20
- Osband, I., Wen, Z., Asghari, S. M., Dwaracherla, V., Ibrahim, M., Lu, X., and Van Roy, B. Approximate Thompson sampling via epistemic neural networks. In Evans, R. J. and Shpitser, I. (eds.), *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 1586–1595. PMLR, 31 Jul–04 Aug 2023. URL <https://proceedings.mlr.press/v216/osband23a.html>. 10
- Ouhamma, R., Basu, D., and Maillard, O.-A. Bilinear exponential family of mdps: Frequentist regret bound with tractable exploration and planning. *arXiv preprint arXiv:2210.02087*, 2022. 1, 6
- Phan, M., Abbasi Yadkori, Y., and Domke, J. Thompson sampling and approximate inference. In Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/file/f3507289cfdc8c9ae93f4098111a13f9-Paper.pdf>. 7
- Pinneri, C., Sawant, S., Blaes, S., Achterhold, J., Stueckler, J., Rolinek, M., and Martius, G. Sample-efficient cross-entropy method for real-time planning. In *Conference on Robot Learning 2020*, 2020. URL [https://corlconf.github.io/paper\\_217](https://corlconf.github.io/paper_217). 27
- Reisinger, J., Stone, P., and Miikkulainen, R. Online kernel selection for bayesian reinforcement learning. In *International Conference on Machine Learning*, pp. 816–823, 2008. 10
- Russo, D. and Van Roy, B. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4): 1221–1243, 2014. doi: 10.1287/moor.2014.0650. URL <https://doi.org/10.1287/moor.2014.0650>. 3
- Russo, D., Roy, B. V., Kazerouni, A., Osband, I., and Wen, Z. A tutorial on thompson sampling, 2020. 2
- Sasso, R., Conserva, M., and Rauber, P. Posterior sampling for deep reinforcement learning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 30042–30061. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/sasso23a.html>. 10
- Schlichting, A. Poincaré and log-sobolev inequalities for mixtures. *Entropy*, 21(1):89, 2019. 25
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017. URL <https://arxiv.org/abs/1707.06347>. 10, 27

- Simchowitz, M., Tosh, C., Krishnamurthy, A., Hsu, D. J., Lykouris, T., Dudik, M., and Schapire, R. E. Bayesian decision-making under misspecified priors with applications to meta-learning. *Advances in Neural Information Processing Systems*, 34:26382–26394, 2021. 7
- Steiner, C. A feynman-kac approach for logarithmic sobolev inequalities, 2021. 4, 5
- Stroock, D. W. and Zegarlinski, B. The equivalence of the logarithmic sobolev inequality and the dobrushin-shlosman mixing condition. *Communications in mathematical physics*, 144:303–323, 1992. 1
- Thompson, W. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of two Samples. *Biometrika*, 25(3-4):285–294, 1933. 2, 4
- Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109. 26, 28
- Towers, M., Kwiatkowski, A., Terry, J., Balis, J. U., De Cola, G., Deleu, T., Goulão, M., Kallinteris, A., Krimmel, M., KG, A., et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024. 10, 26, 28
- Vempala, S. and Wibisono, A. 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. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/file/65a99bb7a3115fdede20da98b08a370f-Paper.pdf>. 2, 4, 5
- Wang, C., Chen, Y., and Murphy, K. P. Model-based policy optimization under approximate bayesian inference. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023. 10
- Wang, F.-Y. Logarithmic sobolev inequalities: conditions and counterexamples. *Journal of Operator Theory*, pp. 183–197, 2001. 4
- Wang, K. and Ghosal, S. Posterior contraction and testing for multivariate isotonic regression. *Electronic Journal of Statistics*, 17(1):798–822, 2023. 8
- Xu, P., Zheng, H., Mazumdar, E. V., Azizzadenesheli, K., and Anandkumar, A. Langevin Monte Carlo for contextual bandits. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24830–24850. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/xu22p.html>. 2
- Yamamoto, K., Oko, K., Yang, Z., and Suzuki, T. Mean field langevin actor-critic: Faster convergence and global optimality beyond lazy learning. In *Forty-first International Conference on Machine Learning*, 2024. 10
- Zheng, H., Deng, W., Moya, C., and Lin, G. Accelerating approximate Thompson sampling with underdamped Langevin Monte Carlo. In Dasgupta, S., Mandt, S., and Li, Y. (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2611–2619. PMLR, 02–04 May 2024. URL <https://proceedings.mlr.press/v238/zheng24b.html>. 2, 7, 8
- Zintgraf, L., Schulze, S., Lu, C., Feng, L., Igl, M., Shiarlis, K., Gal, Y., Hofmann, K., and Whiteson, S. Varibad: variational bayes-adaptive deep rl via meta-learning. *The Journal of Machine Learning Research*, 22(1):13198–13236, 2021. 10

## A. Notation

Table 2: 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.
$\tau$	Total number of episodes.
$h$	Current step in episode.
$H$	Horizon, amount of steps in an episode.
$T$	Total amount of agent interactions, $T = \tau H$
$D$	Dimension of the parameter space of $M$
$d$	Dimension of the state space
$n$	Total amount of available data. Usually $n = lH$ at the start of episode $l + 1$ .
$\mathcal{H}_l$	The states, actions and transitions observed until start of episode $l$ .
$M$	An MDP and its parametrization in $\mathbb{R}^D$ .
$M_l$	Sampled MDP in episode $l$ .
$\pi$	Policy
$\pi^*$	Optimal policy for the true MDP
$\pi^*(M)$	Optimal policy for MDP $M$ .
$V_M^\pi$	Value of policy $\pi$ in MDP $M$ .
$\text{BR}(T)$	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.
$\alpha$	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 $l$ .
$\alpha_{\theta^*}$	Log-Sobolev constant for the data likelihood $\mathcal{L}(X \mid \theta^*)$ .
$\alpha_M$	Log-Sobolev constant for the likelihood of the true MDP $\mathcal{L}(X \mid M)$ .
$L,$	L-smooth constant of the log likelihood function.
$\gamma$	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 3
$L_{\bar{R}}, L_{\bar{\mathcal{T}}}$	Lipschitz parameters for mean transition and reward.
$\mathbb{P}_0(\cdot), \mathbb{P}(\cdot x)$	Prior, posterior and approximate posterior on the parameters.
$\hat{\mathbb{P}}(\cdot x)$	Approximate posterior on the parameters from the Langevin sampling.
$\mathcal{L}(x \mid \theta)$	Likelihood of data $x$ in model $\theta$ .
$\mathbb{E}$	Expectation
$\text{KL}(P \parallel Q)$	Kullback–Leibler divergence between $P$ and $Q$ .
$\omega(\Xi)$	Lower bound on prior mass from Assumption 2.
$g(H, \mathcal{S}, \mathcal{A})$	The non $T$ dependent part of the order of regret.
$\theta$	A set of parameters, in the case of MDPs this is written as $M$ .

## B. Algorithmic Details: SARAH-LD

For completeness, we describe the pseudocode of SARAH-LD (Kinoshita & Suzuki, 2022) algorithms in Algorithm 4 as well as a theorem on the gradient complexity of SARAH-LD in Theorem 9. We slightly modify SARAH-LD notation to Bayesian posterior setting for usage in LaPSRL.

### B.1. SARAH-LD

SARAH-LD, found in Algorithm 4, is a Langevin algorithm by Kinoshita & 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, 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

---

```

1: Input: step size  $\eta > 0$ , batch size  $B$ , epoch length  $m$ , data  $X$ , likelihood  $f$ , prior  $\mathbb{P}(\theta)$ , initial sample  $\rho_0$ .
2: Initialization:  $\theta_0 = \rho_0, \theta^{(0)} = \theta_0$ 
3: for  $s = 0, 1, \dots, (K/m)$  do
4:    $v_{sm} = \nabla \mathcal{L}(X | \theta^{(s)})$ 
5:   randomly draw  $\epsilon \sim N(0, I_{d \times d})$ 
6:    $\theta_{sm+1} = \theta_{sm} - \eta v_{sm} + \sqrt{2\eta/n}\epsilon$ 
7:   for  $l = 1, \dots, m-1$  do
8:      $k = sm + l$ 
9:     randomly pick a subset  $I_k$  from  $\{1, \dots, n\}$  of size  $|I_k| = B$ 
10:    randomly draw  $\epsilon \sim N(0, I_{d \times d})$ 
11:     $v_k = \frac{1}{B} \sum_{i_k \in I_k} (\nabla(\mathcal{L}(X_{i_k} | \theta_k) + 1/n\mathbb{P}(\theta_k)) - \nabla(\mathcal{L}(X_{i_k} | \theta_{k-1}) + 1/n\mathbb{P}(\theta_{k-1}))) + v_{k-1}$ 
12:     $\theta_{k+1} = \theta_k - \eta v_k + \sqrt{2\eta/n}\epsilon$ 
13:   end for
14:    $\theta^{(s+1)} = \theta_{(s+1)m}$ 
15: end for

```

---

**Theorem 9** (Corollary 2.1 of (Kinoshita & Suzuki, 2022)). *Under Assumption 1 and definition 1, for all  $\epsilon \geq 0$ , if we choose step size  $\eta$  such that  $\eta \leq \frac{3\alpha\epsilon}{48\gamma\alpha^2}$ , then a precision  $\text{KL}(\rho_k \| \nu) \leq \epsilon$  is reached after  $k \geq \frac{\gamma}{\alpha\eta} \log \frac{2\text{KL}(\rho_0 \| \nu)}{\epsilon}$  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  $\tilde{O}\left(\left(n + \frac{dn^{\frac{1}{2}}}{\epsilon}\right) \cdot \frac{\gamma^2 L^2}{\alpha^2}\right)$ .*

## C. Regret Bounds for PSRL with Exact Posteriors

### C.1. Confidence Intervals for Isoperimetric Data Distributions

**Theorem 3.** *If the data likelihood  $\mathcal{L}(x | \theta^*)$  is  $\alpha_{\theta^*}$ -LSI, Assumption 2 holds, and the log-likelihood function  $\log \mathcal{L}(x | \theta)$  is  $L_x$  and  $L_\theta$  Lipschitz in  $x$  and  $\theta$ , respectively. Then, for any  $n > 0$  and  $\theta \sim \mathbb{P}(\theta | x^n)$ , we obtain with probability at least  $1 - 2\delta$  (for  $\delta \in (0, 1/2]$ )*

$$\text{KL}(\mathcal{L}(x|\theta^*) || \mathcal{L}(x|\theta)) \leq \sqrt{\frac{4L_x^2 \ln(\frac{2}{\delta})}{n\alpha_{\theta^*}}} + \frac{\ln(\frac{1}{\delta})}{n}. \quad (3)$$

*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  $\mathcal{L}(x | \theta^*)$  is  $\alpha_{\theta^*}$  log-Sobolev, we have

from property of sub-Gaussian concentration of Lipschitz functions, as seen in Equation (2),

$$\mathbb{P}(|\log \mathcal{L}(x | \theta) - \mathbb{E}_{x \sim \mathcal{L}(x|\theta^*)} \log \mathcal{L}(x | \theta)| \geq t) \leq 2e^{-\frac{t^2 \alpha_{\theta^*}}{L_x^2}} \quad (5)$$

$$\implies |\log \mathcal{L}(x | \theta) - \mathbb{E}_{x \sim \mathcal{L}(x|\theta^*)} \log \mathcal{L}(x | \theta)| \leq \sqrt{\frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}, \quad (6)$$

with probability at least  $1 - \delta$ .

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

$$|\log \mathcal{L}(x^{(n)} | \theta) - \mathbb{E}_{x^{(n)} \sim \mathcal{L}(x^{(n)}|\theta^*)} \log \mathcal{L}(x^{(n)} | \theta)| \leq \sqrt{n \frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}, \quad (7)$$

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  $\mathbb{E}_{x \sim \mathcal{L}(x|\theta^*)} \log \mathcal{L}(x | \theta) = -\mathbb{H}[\mathcal{L}(x | \theta^*), \mathcal{L}(x | \theta)]$  where  $\mathbb{H}(\cdot, \cdot)$  is the cross entropy. If  $\theta = \theta^*$  this instead becomes the regular entropy  $\mathbb{H}[\mathcal{L}(x | \theta^*)]$ . This implies that

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

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

*Step 3: From bounding probabilities to concentration of posteriors.*

$$\mathbb{P}(\Theta \setminus \Xi | x^{(n)}, E_{\bar{x}}) = \frac{\int_{\Theta \setminus \Xi} \mathcal{L}(x^{(n)} | \theta, E_{\bar{x}}) d\mathbb{P}(\theta)}{\int_{\Theta} \mathcal{L}(x^{(n)} | \theta, E_{\bar{x}}) d\mathbb{P}(\theta)} \quad (10)$$

$$\leq \frac{\int_{\Theta \setminus \Xi} \mathcal{L}(x^{(n)} | \theta, E_{\bar{x}}) d\mathbb{P}(\theta)}{\int_{\Xi} \mathcal{L}(x^{(n)} | \theta, E_{\bar{x}}) d\mathbb{P}(\theta)} \quad (11)$$

$$\leq \frac{\int_{\Theta \setminus \Xi} e^{-\mathbb{H}(\mathcal{L}(x^{(n)} | \theta^*), \mathcal{L}(x^{(n)} | \theta)) + \sqrt{n \frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}} d\mathbb{P}(\theta)}{\int_{\Xi} e^{-\mathbb{H}(\mathcal{L}(x^{(n)} | \theta^*), \mathcal{L}(x^{(n)} | \theta)) - \sqrt{n \frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}} d\mathbb{P}(\theta)}. \quad (12)$$

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

Now, we proceed as follows

$$\mathbb{P}(\Theta \setminus \Xi | x^{(n)}, E_{\bar{x}}) \leq e^{2\sqrt{n \frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}} \frac{\int_{\Theta \setminus \Xi} e^{-\mathbb{H}(\mathcal{L}(x^{(n)} | \theta^*), \mathcal{L}(x^{(n)} | \theta))} d\mathbb{P}(\theta)}{\int_{\Xi} e^{-\mathbb{H}(\mathcal{L}(x^{(n)} | \theta^*), \mathcal{L}(x^{(n)} | \theta))} d\mathbb{P}(\theta)} \quad (13)$$

$$= e^{2\sqrt{n \frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}} \frac{\int_{\Theta \setminus \Xi} e^{-\text{KL}(\mathcal{L}(x^{(n)} | \theta^*) \| \mathcal{L}(x^{(n)} | \theta))} d\mathbb{P}(\theta)}{\int_{\Xi} e^{-\text{KL}(\mathcal{L}(x^{(n)} | \theta^*) \| \mathcal{L}(x^{(n)} | \theta))} d\mathbb{P}(\theta)} \quad (14)$$

$$\leq e^{2\sqrt{n \frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}} \frac{e^{-\inf_{\theta \notin \Xi} \text{KL}(\mathcal{L}(x^{(n)} | \theta^*) \| \mathcal{L}(x^{(n)} | \theta))} \int_{\Theta \setminus \Xi} d\mathbb{P}(\theta)}{\int_{\Xi} e^{-n L_\theta \| \theta^* - \theta \|} d\mathbb{P}(\theta)} \quad (15)$$

$$\leq \frac{1}{\mathbb{P}(\Xi) \omega(\Xi)^n} e^{-\inf_{\theta \notin \Xi} \text{KL}(\mathcal{L}(x^{(n)} | \theta^*) \| \mathcal{L}(x^{(n)} | \theta)) + 2\sqrt{n \frac{L_x^2}{\alpha_{\theta^*}} \ln\left(\frac{2}{\delta}\right)}} \quad (16)$$

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

from Assumption 2 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  $\Xi$  around  $\theta^*$ , 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 Assumption 2 is reasonable considering that the exponent is zero around  $\theta^*$  and the prior has a minimum mass around it.

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

$$\begin{aligned} & \inf_{\theta \notin \Xi} \text{KL}(\mathcal{L}(x^{(n)} | \theta^*) \| \mathcal{L}(x^{(n)} | \theta)) - 2\sqrt{n \frac{L_x^2}{\alpha_{\theta^*}} \ln \left( \frac{2}{\delta} \right)} \geq \ln \left( \frac{1}{\delta'(1 - \delta') \omega^n(\Xi)} \right) \\ \implies & \inf_{\theta \notin \Xi} \text{KL}(\mathcal{L}(x^{(n)} | \theta^*) \| \mathcal{L}(x^{(n)} | \theta)) \geq 2\sqrt{n \frac{L_x^2}{\alpha_{\theta^*}} \ln \left( \frac{2}{\delta} \right)} + \ln \left( \frac{1}{\delta'} \right) \\ \implies & \inf_{\theta \notin \Xi} \text{KL}(\mathcal{L}(x | \theta^*) \| \mathcal{L}(x | \theta)) \geq 2\sqrt{\frac{L_x^2}{n \alpha_{\theta^*}} \ln \left( \frac{2}{\delta} \right)} + \frac{1}{n} \ln \left( \frac{1}{\delta'} \right). \end{aligned}$$

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

If  $\Xi$  is defined as

$$\Xi = \left\{ \theta \mid \text{KL}(\mathcal{L}(x | \theta^*) \| \mathcal{L}(x | \theta)) \leq 2\sqrt{\frac{L_x^2}{n \alpha_{\theta^*}} \ln \left( \frac{2}{\delta} \right)} + \frac{1}{n} \ln \left( \frac{1}{\delta'} \right) \right\}. \quad (17)$$

Then, we get

$$\mathbb{P}(\theta \notin \Xi | x^{(n)}, E_{\bar{x}}) \leq \delta' \implies \mathbb{P}(\theta \notin \Xi | x^{(n)}) \leq \delta' + \delta. \quad (18)$$

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

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

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

*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  $M_l \sim \mathbb{P}(M \in \cdot | \mathcal{H}_n)$ , where the sampling is from the posterior distribution of  $M$ . (Osband et al., 2013b) observes that for any  $\sigma(\mathcal{H}_l)$  measurable function  $f$ , given the history of transitions  $\mathcal{H}_l \equiv \mathcal{H}_{(l-1)H} = \{(s_{t'}, h, a_{t'}, h, s_{t'}, h+1)_{t' < l, h \leq H}\}$ , we have  $\mathbb{E}[f(M_l)] = \mathbb{E}[f(M)]$ . This family of  $f$ 's includes the value function. Therefore, we have  $\mathbb{E}[V_{\pi_l, 1}^{M_l}(s_{l, 1})] = \mathbb{E}[V_{\pi^*, 1}^M(s_{l, 1})]$ . Hence, the Bayes regret (Equation (1)) of PSRL can be re-written as

$$\text{BR}(T) = \mathbb{E} \left[ \sum_{l=1}^{\tau} V_{\pi_l, 1}^{M_l}(s_{l, 1}) - V_{\pi_l, 1}^M(s_{l, 1}) \right].$$

**Step 2. Recursion with Bellman Equation.** (Chowdhury & 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,

$$\text{BR}(T) = \mathbb{E} \left[ \sum_{l=1}^{\tau} \sum_{h=1}^H \mathcal{T}_{M_l, h}^{\pi_l} (V_{\pi_l, h+1}^{M_l})(s_{l,h}) - \mathcal{T}_{M, h}^{\pi_l} (V_{\pi_l, h+1}^M)(s_{l,h}) + \sum_{l=1}^{\tau} \sum_{h=1}^H m_{l,h} \right].$$

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

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

$$\mathbb{E}_P[x] - \mathbb{E}_Q[x] \leq \sqrt{2\mathbb{V}_P(x)\text{KL}(Q \parallel P)}. \quad (19)$$

Then an application of the transportation inequality yields

$$\text{BR}(T) \leq H\mathbb{E} \left[ \sum_{l=1}^{\tau} \sum_{h=1}^H \sqrt{2\sigma_R^2 \text{KL}_{s_h^l, a_h^l}(M \parallel M_n)} \right],$$

Here,  $\sigma_R^2$  is the maximum variance of rewards at each step.

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

$$\text{BR}(T|\mathcal{E}^*) \leq H\sigma_R \sqrt{2T \sum_{l=1}^{\tau} \sum_{h=1}^H \text{KL}_{s_h^l, a_h^l}(M \parallel M_n)} \leq H\sigma_R \sqrt{2T \left( T \frac{L_x^2}{\alpha_M} \ln \left( \frac{2}{\delta} \right) \right)^{1/2} + 2T \ln \left( \frac{T}{\delta} \right)}.$$

Here,  $\mathcal{E}^*$  denotes the event of the distribution of data concentrating in the set  $\Xi_n$  around  $M$  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  $\text{BR}(T) \leq TB_R$ .

Thus, we have

$$\text{BR}(T) = \mathbb{E} [\mathcal{R}(T)\mathbb{I}_{\mathcal{E}^*} + \mathcal{R}(T)\mathbb{I}_{(\mathcal{E}^*)^c}] \leq \text{BR}(T|\mathcal{E}^*) + 2TB_R(1 - \mathbb{P}(\mathcal{E}^*))$$

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

$$\begin{aligned} \text{BR}(T) &\leq H\sigma_R \sqrt{2T \left( T \frac{L_x^2}{\alpha_M} \ln \left( \frac{2}{\delta} \right) \right)^{1/2} + 2T \ln \left( \frac{T}{\delta} \right) + 4TB_R\delta} \\ &\leq 2B_R + 2H\sigma_R \sqrt{T \ln(2T)} + \sqrt{2}H\sigma_R T^{3/4} \left( \frac{L_x^2}{\alpha_M} \right)^{1/4} \log^{1/4}(4T). \end{aligned} \quad (20)$$

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  $\pi_l$  in  $M_l$  where  $\varphi$  is the distribution of next state samples. This gives  $U_h^{M_l}(\varphi) = \mathbb{E}_{s' \sim \varphi}[V^{\pi_l, h+1}(s')]$ . We use this definition, which is also used in previous work, (Chowdhury & Gopalan, 2019; Osband & Van Roy, 2014), to make a Lipschitz assumption on the next step value function  $U$  with respect to the means of the distributions.

**Assumption 3** (One step value function is Lipschitz in the mean). *For any  $\varphi_1, \varphi_2$  distributions over  $\mathcal{S}$  with  $1 \leq h \leq H$ ,*

$$|U_h^M(\varphi_1) - U_h^M(\varphi_2)| \leq L_M \|\bar{\varphi}_1 - \bar{\varphi}_2\|_2 \quad (21)$$

where  $\bar{\varphi}_1$  and  $\bar{\varphi}_2$  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  $\{\alpha_{\bar{T},l}\}$ , the mean reward for any MDP  $M$  is bounded:  $|\bar{R}_M(s)| \leq B_R \forall s$ , the one step value function is Lipschitz in the state with parameter  $L_M$  as given in Assumption 3, and the mean reward and mean transitions are  $L_{\bar{R}}$  and  $L_{\bar{T}}$  Lipschitz in  $M$ , Bayesian regret of PSRL is bounded by*

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

*Proof.* This proof follows the general flow from Chowdhury & 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  $\mathbb{E}[V_{\pi_l,1}^{M_l}(s_{l,1}) | \mathcal{H}_{l-1}] = \mathbb{E}[V_{\pi_*,1}^M(s_{l,1}) | \mathcal{H}_{l-1}]$ , since they are both sampled from the same distribution. Marginalising we obtain:

$$\mathbb{E}[\text{Regret}(T)] \triangleq \sum_{l=1}^{\tau} \mathbb{E} [V_{\pi_*,1}^M(s_{l,1}) - V_{\pi_l,1}^M(s_{l,1})] \quad (22)$$

$$= \sum_{l=1}^{\tau} \mathbb{E} [V_{\pi_*,1}^M(s_{l,1}) - V_{\pi_l,1}^{M_l}(s_{l,1})] + \mathbb{E} [V_{\pi_l,1}^{M_l}(s_{l,1}) - V_{\pi_l,1}^M(s_{l,1})] \quad (23)$$

$$= \sum_{l=1}^{\tau} \mathbb{E} [V_{\pi_l,1}^{M_l}(s_{l,1}) - V_{\pi_l,1}^M(s_{l,1})] \quad (24)$$

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

$$\mathbb{E}[\text{Regret}(T)] \leq \mathbb{E} \left[ \sum_{l=1}^{\tau} \sum_{h=1}^H [| \bar{R}_{M_l}(s_{l,h}, a_{l,h}) - \bar{R}_*(s_{l,h}, a_{l,h}) | + L_{M_l} \| \bar{T}_{M_l}(s_{l,h}, a_{l,h}) - \bar{T}_*(s_{l,h}, a_{l,h}) \|_2] \right]. \quad (25)$$

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

We define two confidence sets

$$C_{R,l,h} = \left\{ |\bar{R}_M(s_{l,h}) - E_{\mathbb{P}(M|\mathcal{H}_l)}[\bar{R}_M(s_{l,h})]| \leq \sqrt{\frac{L_{\bar{R}}^2 \log 1/\delta}{\alpha_{\bar{R},l}}} \right\} \quad (26)$$

$$C_{\mathcal{T},l,h} = \left\{ \| \bar{T}_M(s_{l,h}, a_{l,h}) - E_{\mathbb{P}(M|\mathcal{H}_l)}[\bar{T}_M(s_{l,h}, a_{l,h})] \|_2 \leq \sqrt{\frac{d L_{\bar{T}}^2 \log 1/\delta}{\alpha_{\bar{T},l}}} \right\} \quad (27)$$

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

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

$$\mathbb{E}\left[\sum_{l=1}^{\tau} \sum_{h=1}^H \left[ |\bar{R}_{M_l}(s_{l,h}, a_{l,h}) - \bar{R}_*(s_{l,h}, a_{l,h})| \mid E \right] + \mathbb{E}\left[L_{M_l} \|\bar{T}_{M_l}(s_{l,h}, a_{l,h}) - \bar{T}_*(s_{l,h}, a_{l,h})\|_2 \mid E\right] + 2B_R 4\tau H \delta \right] \quad (28)$$

$$\leq 2H \left( L_{\bar{R}} \sqrt{\log 1/\delta} \sum_{l=1}^{\tau} \frac{1}{\sqrt{\alpha_{\bar{R},l}}} + \frac{\mathbb{E}[L_M]}{1-2\tau H \delta} L_{\bar{T}} \sqrt{d \log 1/\delta} \sum_{l=1}^{\tau} \frac{1}{\sqrt{\alpha_{\bar{T},l}}} \right) + 8B_R \tau H \delta \quad (29)$$

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

$$\mathbb{E}[\text{Regret}(T)] \leq 2H \left( L_{\bar{R}} \sqrt{\log 8T} \sum_{l=1}^{\tau} \frac{1}{\sqrt{\alpha_{\bar{R},l}}} + 2\mathbb{E}[L_M] L_{\bar{T}} \sqrt{d \log 8T} \sum_{l=1}^{\tau} \frac{1}{\sqrt{\alpha_{\bar{T},l}}} \right) + B_R \quad (30)$$

■

## D. Regret Bounds and Sample Complexity for LaPSRL with Approximate Posteriors

**Theorem 5.** Let us sample an MDP from an approximate posterior  $\mathbb{Q}_l$  in episode  $l$  and use it for planning. If  $\mathbb{P}_l$  is the true posterior at  $l$  and  $\min(\text{KL}((\mathbb{P}_l \parallel \mathbb{Q}_l), \text{KL}(\mathbb{Q}_l \parallel \mathbb{P}_l)) \leq \epsilon_{\text{post},l}$ , then regret in an episode due to the approximate posterior is  $\mathcal{O}(HB_R \sqrt{\epsilon_{\text{post},l}})$ .

*Proof.* Let  $M_l, M_l \sim \mathbb{P}(M)$ ,  $M'_l \sim \mathbb{Q}(M)$ . The policy  $\pi_l$  is the optimal policy corresponding to  $M_l$  and  $\pi'_l$  the policy corresponding to  $M'_l$ .

$$\mathbb{E}_{\mathbb{P}_l, \mathbb{Q}_l}[V_{\pi^*}^M - V_{\pi'_l}^M] = \mathbb{E}_{\mathbb{P}_l, \mathbb{Q}_l}[V_{\pi^*}^M - V_{\pi'_l}^{M_l} + V_{\pi'_l}^{M'_l} - V_{\pi'_l}^M] \quad (31)$$

$$= \mathbb{E}_{\mathbb{P}_l, \mathbb{Q}_l}[V_{\pi^*}^M - V_{\pi_l}^{M_l} + V_{\pi_l}^{M_l} - V_{\pi'_l}^{M'_l} + V_{\pi'_l}^{M'_l} - V_{\pi'_l}^M] \quad (32)$$

$$= \mathbb{E}_{\mathbb{P}_l, \mathbb{Q}_l}[[V_{\pi^*}^M - V_{\pi_l}^{M_l}] + [V_{\pi_l}^{M_l} - V_{\pi'_l}^{M'_l}] + [V_{\pi'_l}^{M'_l} - V_{\pi'_l}^M]] \quad (33)$$

$$\leq \mathbb{E}_{\mathbb{P}_l}[V_{\pi^*}^M - V_{\pi_l}^{M_l}] + \Delta_{\max} \sqrt{\frac{\epsilon_{\text{post},l}}{2}} + \Delta_{\max} \sqrt{\frac{\epsilon_{\text{post},l}}{2}} \quad (34)$$

$$= \mathbb{E}_{\mathbb{P}_l}[V_{\pi^*}^M - V_{\pi_l}^M] + \sqrt{2} \Delta_{\max} \sqrt{\epsilon_{\text{post},l}} \quad (35)$$

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

**Corollary 1.** If an algorithm incurs  $\tilde{\mathcal{O}}(\sqrt{T}g(H, \mathcal{S}, \mathcal{A}))$  regret for the true posteriors, it will incur the same order of regret for the approximate posteriors if  $\epsilon_{\text{post},l} \leq C \frac{g(H, \mathcal{S}, \mathcal{A})^2}{l \Delta_{\max}^2}$  for some  $C > 0$ . Here,  $\Delta_{\max} \triangleq \max_{\pi} V_{\pi,1}^M(s_1) - \min_{\pi} V_{\pi,1}^M(s_1) \leq 2H_B$  is maximal regret in an episode.

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

$$\tilde{\mathcal{O}}(\mathbb{E}_P R(\pi_Q)) \leq \tilde{\mathcal{O}}(\sqrt{\tau}g(H, \mathcal{S}, \mathcal{A})) + \sqrt{2} \Delta_{\max} \sum_{k=1}^{\tau} \sqrt{\epsilon_{\text{post},l}} \quad (36)$$

$$\leq \tilde{\mathcal{O}}(\sqrt{\tau}g(H, \mathcal{S}, \mathcal{A})) + \sqrt{2} \Delta_{\max} \sum_{k=1}^{\tau} \sqrt{C} \frac{g(H, \mathcal{S}, \mathcal{A})}{\sqrt{k} \Delta_{\max}} \quad (37)$$

$$= \tilde{\mathcal{O}}(\sqrt{\tau}g(H, \mathcal{S}, \mathcal{A})) + \sqrt{2} g(H, \mathcal{S}, \mathcal{A}) \sqrt{C} \sum_{k=1}^{\tau} \frac{1}{\sqrt{k}} \quad (38)$$

$$= \tilde{\mathcal{O}}(\sqrt{\tau}g(H, \mathcal{S}, \mathcal{A})) \quad (39)$$

■

**Corollary 2.** For a posterior fulfilling the Assumption 1 and Definition 1, LaPSRL obtains the same order of regret as PSRL while SARAH-LD incurs a gradient complexity  $\tilde{O}\left(\frac{H^3 l^3 L^2}{\alpha_l^2} + \frac{dB_R^2 H^{4.5} l^{3.5} L^2}{\alpha_l^2 g(H, \mathcal{S}, \mathcal{A})^2}\right)$  in episode  $l$ .

*Proof.* This result can be obtained by directly applying the  $\epsilon_{\text{post},l}$  obtained from Theorem 5 into Theorem 9 with  $\gamma = n$ . ■

**Lemma 2.** For LSI posteriors (as in Theorem 4) with linearly growing LSI constants  $\alpha_l = \Omega(Hl)$ , the total gradient complexity of LaPSRL is  $\tilde{O}\left(\tau T + \tau T^{1.5}/\sqrt{d}\right)$  that yields regret  $\mathcal{O}(\sqrt{dHT})$ .

*Proof.* Applying  $\alpha_l = \Omega(T)$  into Theorem 4 we have that  $\text{BR}(T) = \sqrt{dHT}$  with  $g(H, \mathcal{S}, \mathcal{A}) = \tilde{O}(\sqrt{dH})$ . Inserting  $g(H, \mathcal{S}, \mathcal{A})$  into Corollary 2 completes the proof. ■

**Theorem 6.** If  $\nabla_z \log \mathcal{L}(z|M)$  is  $L_z$ -Lipschitz and  $\alpha_z$ -Log Sobolev, with  $z$  being the data corresponding to an episode,

$$\mathbb{E}_{\mathcal{L}(z|\mathcal{H}_l)} \text{KL}(\hat{\mathbb{P}}(M|\mathcal{H}_l) \parallel \mathbb{P}(M|\mathcal{H}_{l+1})) \leq \epsilon_{\text{post},l} + \frac{L_z^2}{\alpha_z} \epsilon_{\text{post},l}^2 \text{Var}(\mathbb{P}(M|\mathcal{H}_l)) + \frac{L_z^2}{2\alpha_z} \text{Var}(\hat{\mathbb{P}}(M|\mathcal{H}_l)), \quad (4)$$

where  $\text{Var}(P)$  is the variance of the distribution  $P$ . Note that LaPSRL ensures  $\epsilon_{\text{post},l} = \mathcal{O}(1/l)$ .

*Proof.* For notation we write  $\mathbb{P}(M | \mathcal{H}_{l+1}) = \mathbb{P}(M | \mathcal{H}_l, z_l)$  such that  $\mathbb{P}(M | \mathcal{H}_{l+1}) = \mathbb{P}(M | z_0, \dots, z_l)$ . Here we have that  $z_l = \{(s_i, a_i)\}_{i=H(l-1)+1}^{Hl}$ , the data for episode  $l$ . Note that we can marginalize  $\mathcal{L}(z_l | \mathcal{H}_l, M) = \mathcal{L}(z_l | M)$  and  $\mathbb{E}_M \mathcal{L}(z_l | \mathcal{H}_l, M) = \mathcal{L}(z_l | \mathcal{H}_l)$ . As a reminder, we have  $\mathcal{L}(z | M)$  as the data likelihood for an episode, as such we have with Bayes rule  $\mathbb{P}(M | \mathcal{H}_l) = \frac{\mathbb{P}(M) \mathcal{L}(\mathcal{H}_l | M)}{\mathcal{L}(\mathcal{H}_l)}$ . Similarly,  $\mathcal{L}(x | M)$  is the likelihood of a single step.

$$\text{KL}(\hat{\mathbb{P}}(M|\mathcal{H}_l) \parallel \mathbb{P}(M|\mathcal{H}_{l+1}) | z_l) \quad (40)$$

$$= \int_M \log \left( \frac{\hat{\mathbb{P}}(M|\mathcal{H}_l)}{\mathbb{P}(M|\mathcal{H}_l, z_l)} \right) \hat{\mathbb{P}}(M|\mathcal{H}_l) dM \quad (41)$$

$$= \int_M \log \left( \frac{\hat{\mathbb{P}}(M|\mathcal{H}_l)}{\frac{\mathbb{P}(M|\mathcal{H}_l) \mathcal{L}(z_l|M)}{\mathcal{L}(z_l|\mathcal{H}_l)}} \right) \hat{\mathbb{P}}(M|\mathcal{H}_l) dM \quad (42)$$

$$= \int_M \left( \log \left( \frac{\hat{\mathbb{P}}(M|\mathcal{H}_l)}{\mathbb{P}(M|\mathcal{H}_l)} \right) + \log \left( \frac{\mathcal{L}(z_l|\mathcal{H}_l)}{\mathcal{L}(z_l|M)} \right) \right) d\hat{\mathbb{P}}(M|\mathcal{H}_l) \quad (43)$$

$$= \int_M \log \left( \frac{\hat{\mathbb{P}}(M|\mathcal{H}_l)}{\mathbb{P}(M|\mathcal{H}_l)} \right) d\hat{\mathbb{P}}(M|\mathcal{H}_l) \quad (44)$$

$$+ \int_M \log \left( \frac{\mathcal{L}(z_l|\mathcal{H}_l)}{\mathcal{L}(z_l|M)} \right) \hat{\mathbb{P}}(M|\mathcal{H}_l) dM \quad (45)$$

$$= \text{KL}(\hat{\mathbb{P}}(M|\mathcal{H}_l) \parallel \nu_l) + \int_M \log \left( \frac{\mathcal{L}(z_l|\mathcal{H}_l)}{\mathcal{L}(z_l|M)} \right) d\hat{\mathbb{P}}(M|\mathcal{H}_l) \quad (46)$$

$$\leq \epsilon_{\text{post},l} + \int_M \log \left( \frac{\mathcal{L}(z_l|\mathcal{H}_l)}{\mathcal{L}(z_l|M)} \right) d\hat{\mathbb{P}}(M|\mathcal{H}_l) \quad (47)$$

The second equality 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

$$\mathbb{E}_{z_l} \text{KL}(\hat{\mathbb{P}}(M|\mathcal{H}_l) \parallel \mathbb{P}(M|\mathcal{H}_{l+1}) | z_l) \quad (48)$$

$$\leq \epsilon_{\text{post},l} + \int_{z_l} \int_M \log \left( \frac{\mathcal{L}(z_l|\mathcal{H}_l)}{\mathcal{L}(z_l|M)} \right) d\hat{\mathbb{P}}(M|\mathcal{H}_l) \mathcal{L}(z_l | \mathcal{H}_l) dz_l \quad (49)$$

$$= \epsilon_{\text{post},l} + \int_M \int_{z_l} \log \left( \frac{\mathcal{L}(z_l | \mathcal{H}_l)}{\mathcal{L}(z_l | M)} \right) \mathcal{L}(z_l | \mathcal{H}_l) dz_l d\hat{\mathbb{P}}(M | \mathcal{H}_l) \quad (50)$$

$$= \epsilon_{\text{post},l} + \int_M \int_{z_l} \log \left( \frac{\mathcal{L}(z_l | \mathcal{H}_l)}{\mathcal{L}(z_l | M)} \right) \frac{\mathcal{L}(z_l | \mathcal{H}_l)}{\mathcal{L}(z_l | M)} \mathcal{L}(z_l | M) dz_l d\hat{\mathbb{P}}(M | \mathcal{H}_l) \quad (51)$$

$$\leq \epsilon_{\text{post},l} + \int_M 2/\alpha_z \int_{z_l} \|\nabla_z \sqrt{\frac{\mathcal{L}(z_l | \mathcal{H}_l)}{\mathcal{L}(z_l | M)}}\|^2 \mathcal{L}(z_l | M) dz_l d\hat{\mathbb{P}}(M | \mathcal{H}_l) \quad (52)$$

$$= \epsilon_{\text{post},l} + \int_M 2/\alpha_z \int_{z_l} \left\| \frac{\mathcal{L}(z_l | M) \nabla_z \mathcal{L}(z_l | \mathcal{H}_l) - \mathcal{L}(z_l | \mathcal{H}_l) \nabla_z \mathcal{L}(z_l | M)}{2 \sqrt{\frac{\mathcal{L}(z_l | \mathcal{H}_l)}{\mathcal{L}(z_l | M)}} \mathcal{L}(z_l | M)^2} \right\|^2 \mathcal{L}(z_l | M) dz_l d\hat{\mathbb{P}}(M | \mathcal{H}_l) \quad (53)$$

$$= \epsilon_{\text{post},l} + \int_M 2/\alpha_z \int_{z_l} \left\| \frac{\mathcal{L}(z_l | M) \nabla_z \mathcal{L}(z_l | \mathcal{H}_l) - \mathcal{L}(z_l | \mathcal{H}_l) \nabla_z \mathcal{L}(z_l | M)}{2 \mathcal{L}(z_l | \mathcal{H}_l) \mathcal{L}(z_l | M)} \times \sqrt{\frac{\mathcal{L}(z_l | \mathcal{H}_l)}{\mathcal{L}(z_l | M)}} \right\|^2 \mathcal{L}(z_l | M) dz_l d\hat{\mathbb{P}}(M | \mathcal{H}_l) \quad (54)$$

$$= \epsilon_{\text{post},l} + \int_M 2/\alpha_z \int_{z_l} \left\| \frac{\mathcal{L}(z_l | M) \nabla_z \mathcal{L}(z_l | \mathcal{H}_l) - \mathcal{L}(z_l | \mathcal{H}_l) \nabla_z \mathcal{L}(z_l | M)}{2 \mathcal{L}(z_l | \mathcal{H}_l) \mathcal{L}(z_l | M)} \right\|^2 \frac{\mathcal{L}(z_l | \mathcal{H}_l)}{\mathcal{L}(z_l | M)} \mathcal{L}(z_l | M) dz_l d\hat{\mathbb{P}}(M | \mathcal{H}_l) \quad (55)$$

$$= \epsilon_{\text{post},l} + \int_M \frac{1}{2\alpha_z} \int_{z_l} \|\nabla_z \log \mathcal{L}(z_l | \mathcal{H}_l) - \nabla_z \log \mathcal{L}(z_l | M)\|^2 \mathcal{L}(z_l | \mathcal{H}_l) dz_l d\hat{\mathbb{P}}(M | \mathcal{H}_l) \quad (56)$$

$$\leq \epsilon_{\text{post},l} + \frac{L_z^2}{2\alpha_z} \int_M \|\mathbb{E}_{\mathbb{P}(M | \mathcal{H}_l)}[M] - M\|^2 d\hat{\mathbb{P}}(M | \mathcal{H}_l) \quad (57)$$

$$\leq \epsilon_{\text{post},l} + \frac{L_z^2}{2\alpha_z} \int_M \|\mathbb{E}_{\mathbb{P}(M | \mathcal{H}_l)}[M] - \mathbb{E}_{\hat{\mathbb{P}}(M | \mathcal{H}_l)}[M]\|^2 d\hat{\mathbb{P}}(M | \mathcal{H}_l) + \frac{L_z^2}{2\alpha_z} \int_M \|\mathbb{E}_{\hat{\mathbb{P}}(M | \mathcal{H}_l)}[M] - M\|^2 d\hat{\mathbb{P}}(M | \mathcal{H}_l) \quad (58)$$

$$\leq \epsilon_{\text{post},l} + \frac{L_z^2}{\alpha_z} \epsilon_{\text{post},l}^2 \text{Var}(\mathbb{P}(M | \mathcal{H}_l)) + \frac{L_z^2}{2\alpha_z} \text{Var}(\hat{\mathbb{P}}(M | \mathcal{H}_l)) \quad (59)$$

The inequality in Equation (52) comes from the log-Sobolev inequality property of  $\mathcal{L}(z_l | M)$ . Equation (57) which comes from the Lipschitz property of the gradient, Equation (58) is an addition and subtraction with the norm triangle inequality and Equation (59) comes from the definition of variance together and the transportation result of Equation (19). The rest of the steps are algebraic manipulations. ■

## 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 3. The results for the Bayesian regret follow from combining the scaling of the log-Sobolev constant with Theorem 4.

**Mixture Distributions.** There has been multiple works on LSI constants for mixtures of log-Sobolev distributions (Koehler & Vuong, 2024; Chen et al., 2021; Schlichting, 2019). Generally, LSI of a mixture depends on LSI constants of the components and the distance between them.

**Theorem 10** (Informally from Theorem 2 (Koehler & 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  $\delta$  between components, has  $\alpha_{\text{Mixture}} \geq \frac{\delta \min p_i \min \alpha_i}{4k(1-\log(\min p_i))}$ .*

The overlap factor  $\delta$  relates to integral over the minimum of the paired components (Koehler & Vuong, 2024). If the components are posterior distributions,  $\delta \rightarrow 1$  as individual posteriors observe more data and converge.

**Theorem 7.** (a) Any log-concave posterior fulfills LSI with  $\alpha_n = \Theta(n)$ . (b) Any posterior that is a mixture of  $k$  log-concave distributions has  $\alpha_n^{\text{Mixture}} = \Omega\left(\frac{n \min p_i}{4k(1-\log(\min p_i))}\right)$ .

*Proof.* We can write the product of log-concave distributions  $\mathbb{P}(\theta | \mathcal{H}_l) = \mathbb{P}_0(\theta) \frac{\prod_{i=1}^n \mathcal{L}_i(\theta)}{Z}$  where  $\mathcal{L}_i(\theta)$  is shorthand for  $\mathcal{L}(x_i | \theta)$  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  $\alpha_l \geq \alpha_{\mathbb{P}_0(\theta)} + \sum_{i=1}^n \alpha_i \geq \alpha_{\mathbb{P}_0(\theta)} + n \min_i \alpha_i$ . Similarly, applying Weyl's inequality for the largest eigenvalue, we get that the largest eigenvalue of  $-\nabla^2 \log(\mathbb{P}(\theta | \mathcal{H}_l))$  is upper bounded by the sum of maximal eigenvalues which gives an upper bound of  $O(n)$  for  $\alpha_l$  since the smallest eigenvalue must be smaller than the largest.

Similarly, for mixtures of log-concave distributions we have from Theorem 10 that  $\alpha_{\text{Mixture}} = \Omega\left(\frac{\min_i \alpha_i \min p_i}{4k(1-\log(\min p_i))}\right)$ . Setting  $\min_i \alpha_i = \Theta(n)$  completes the proof. ■

**Theorem 8.** A log-Sobolev distribution with bounded likelihood ratio:  $|\log \frac{\mathcal{L}(x|M)}{\mathcal{L}(x|M')}| \leq \Gamma$ , has a log-Sobolev posterior.

*Proof.* This result follows directly from Theorem 2 where the log-likelihood ratio gives the difference between the maximum and minimum perturbations. ■

Table 3: Overview of log-Sobolev constants and Bayes regrets of LaPSRL for different families of distributions.

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

## 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.2 for increased visibility.

### F.1. Gaussian Multi-armed Bandits

We use LaPSRL on a Gaussian multi-armed bandit task with two arms. The arms generate rewards as  $\mathcal{N}(0, 0.25)$ ,  $\mathcal{N}(0.1, 0.25)$ . 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.2.

### 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  $\alpha/n$ . Instead of estimating  $\log \frac{2KL(\rho_0 \| \mathcal{L}(M|\mathcal{H}_t))}{\epsilon_{\text{post},t}}$ , 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.2 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  $\alpha$  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

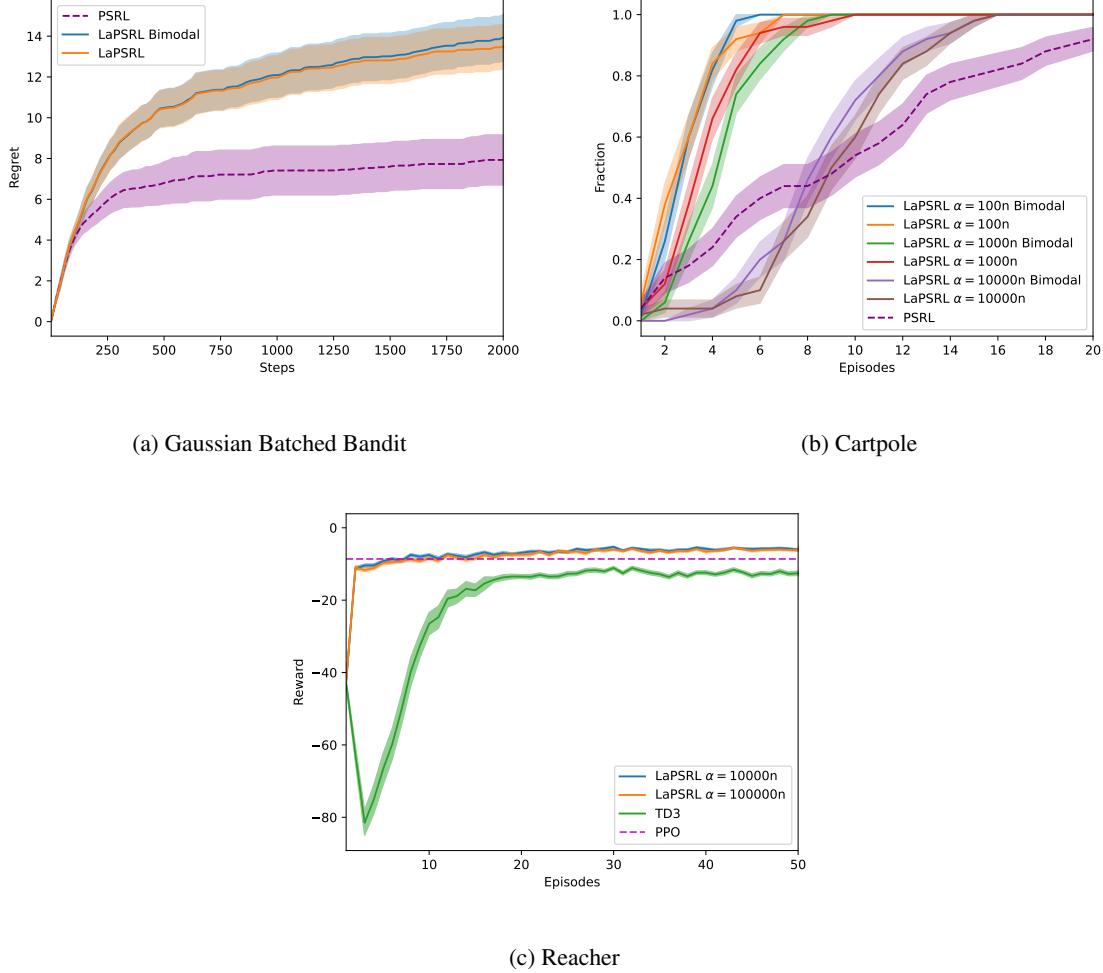


Figure 3: We compare LaPSRL versus baselines. In the bandit and Cartpole experiments we benchmark with PSRL, in Reacher with TD3. In a) we compare the expected regret for a Gaussian bandit algorithm. In b) we compare how many episodes it takes to solve a Cartpole task. In c) we study the average regret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

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 position 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  $\alpha/n$ . Here we benchmark with the TD3 (Fujimoto et al., 2018) and PPO (Schulman et al., 2017) algorithms. As PPO does not use a replay buffer and the CleanRL implementation has a large batch size, it is significantly less data efficient, and we present the average performance after 7000 episodes. We can see that Appendix F.2 that LaPSRL can learn the environment very quickly and reaches a performance that takes TD3 reaches eventually, but takes significantly longer.

### 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. For the experiments with TD3 (Fujimoto et al., 2018) we used the implementation from CleanRL (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. See the supplementary code for additional details.