Isoperimetry is All We Need: Langevin Posterior Sampling for RL

Emilio Jorge
Chalmers University of Technology

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

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

Abstract

In Reinforcement Learning theory, we often assume restrictive assumptions, like linearity and RKHS structure on the model, or Gaussianity and log-concavity of the posteriors over models, to design an algorithm with provably sublinear regret. But RL in practice is known to work for a wider range of distributions and models. Thus, we study whether we can design efficient low-regret RL algorithms for any isoperimetric distribution, which includes and extends the standard setups in the literature to non-log-concave and perturbed distributions. Specifically, we show that the well-known PSRL (Posterior Sampling-based RL) algorithm yields sublinear regret if the sequence of posterior distributions satisfy the Log-Sobolev Inequality (LSI), which is a form of isoperimetry, with linearly growing constants. Further, for the cases where we cannot compute or sample from an exact posterior, we propose a Langevin sampling-based algorithm design scheme, namely LaPSRL. We show that LaPSRL also achieves order optimal regret if the posteriors satisfy LSI. Finally, we deploy a version of LaPSRL with a Langevin sampling algorithms, SARAH-LD. We numerically demonstrate their performances in different bandit and MDP environments. Experimental results validate the generality of LaPSRL across environments and its competitive performance with respect to the baselines.

1 Introduction

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

17th European Workshop on Reinforcement Learning (EWRL 2024).

with a bounded likelihood function and a log-Sobolev prior will be log-Sobolev, which would include complex setups such as some forms of Bayesian neural networks. In optimization and sampling 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). For our study, we focus on the popular PSRL algorithms (Osband et al., 2013; Russo et al., 2020), which are generalisation of Thompson sampling proposed for bandits (Thompson, 1933). PSRL is a Bayesian algorithm that begins with a prior distribution over the model parameters. As PSRL collects more data, it creates more informative posterior distributions, samples probable model parameters from the posteriors, and uses the sampled parameters for further planning. Since PSRL has been successful both theoretically and practically, we choose it as the base algorithm to study.

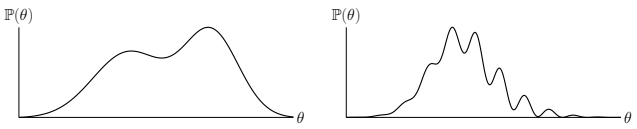


Figure 1: Examples of log-Sobolev distributions.

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

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

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

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

1. Prove that *PSRL can achieve sublinear regret for posteriors satisfying LSI* if we can compute and sample from the exact posteriors and the inequality constant scales linearly. This result broadens the scenarios where PSRL is proven to be efficient.

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

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

4. *Experimentally demonstrate that LaPSRL with SARAH-LD yields sublinear regret* for bandits with Gaussians and mixture of Gaussians as posteriors, and Linear Quadratic Regulators (LQRs) with approximate posteriors, and performs competitively with corresponding PSRL baselines.

Notation. We will use complexity notation O, Ω, Θ, with standard implications, and sometimes $\widetilde{O}, \widetilde{\Omega}, \widetilde{\Theta}$, which is the equivalent term but ignoring sub-logarithmic and poly-logarithmic terms.

2 Preliminaries: Reinforcement Learning, Sampling with Langevin Dynamics

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

Problem Formulation: Episodic Reinforcement Learning (RL). To perform RL, we consider the episodic finite-horizon MDPs (aka *Episodic RL*) (Azar et al., 2017; Osband et al., 2013). MDP in episodic RL is defined as $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, R, H \rangle$. M has states $s \in \mathcal{S}$ where $\mathcal{S} \in \mathbb{R}^d$, actions $a \in \mathcal{A}$. In episodic RL, the agent interacts with the environment in episodes of H steps. Any episode l starts with a state s_1^l. Then, for $t \in [H]$, the agent draws action a_t^l from a policy $\pi_t(s_t^l)$, observes the reward $R(s_t^l, a_t^l) \in \mathbb{R}$, and transits to a state $s_{t+1}^l \sim \mathcal{T}(. \mid s_t^l, a_t^l)$. This interaction is done for a total of τ (which is commonly unknown) episodes. The performance of a policy π is measured by the total expected reward V_1^π w.r.t. an initial state s. We define the value function and the Q-value function at $h \in [H]$

$$V_M^{\pi,h}(s) \triangleq \mathbb{E}\left[\sum_{t=h}^{H} R(s_t, a_t) \mid s_h = s\right], \quad \text{and} \quad Q_M^{\pi,h}(s,a) \triangleq \mathbb{E}\left[\sum_{t=h}^{H} R(s_t, a_t) \mid s_h = s, a_h = a\right].$$

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

Background: PSRL. A popular Bayesian approach, which has been very successful is to sample an MDP $M_l \sim P(M \mid D_l)$ and play the optimal policy for M_l for one episode before updating the posterior and resampling. This algorithm is known as PSRL (Osband et al., 2013). PSRL reduces to Thompson sampling (Thompson, 1933), when applied to MAB. In this paper, we will use some simplifying notation, z_i is shorthand for (s_{i+1}, s_i, a_i). The concept of regret is crucial to RL theory, it describes how much worse the policy is than the optimal policy. In the Bayesian regret, this is taken in expectation of value over the possible MDPs and evaluations and can be written $\sum_l^\tau \sum \mathbb{E}[V_{\pi_*,1}^{M_*}(s_{l,1}) - V_{\pi_l,1}^{M_*}(s_{l,1})]$. We also use notation $\Delta_{\max} = \max_\pi V_{\pi,1}^{M_*}(s_1) - \min_\pi V_{\pi,1}^{M_*}(s_1)$ to denote the maximal regret that could be obtained in one episode. In the paper, we use n to denote the amount of data samples we have observed. We denote to total interactions with the environment as $T = \tau H$.

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

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

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

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

$$\mathbb{E}_\nu[g^2 \log g^2] - \mathbb{E}_\nu[g^2] \log \mathbb{E}_\nu[g^2] \leq \frac{2}{\alpha} \mathbb{E}_\nu[||\nabla g||^2]. \tag{1}$$

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

In this paper, we will only cover a brief introduction to log-Sobolev distributions as needed, but there has been much work looking into the properties of log-Sobolev distributions, a summary of which can be found in (Chafaï & Lehec, 2023; Vempala & Wibisono, 2019). Also note that in some work an inverse definition is used where the constant is defined $\alpha' = \frac{1}{2\alpha}$, leading to some confusion.

Obtaining the LSI constant is not always trivial, but there are some tools. In some cases, the Bakry-Émery criterion can be used.

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

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

One example of an operation on log-Sobolev distributions that preserves the property is that of bounded perturbation, something that would generally break a log-concave property. The theorem is originally due to (Holley & Stroock, 1986) but presented here in the formulation of (Steiner, 2021).

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

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

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

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

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

Background: SARAH-LD (Kinoshita & Suzuki, 2022). There exists multiple algorithms for performing biased Langevin sampling on log-Sobolev distributions (Kinoshita & Suzuki, 2022; Vempala & Wibisono, 2019). In this paper, we focus on SARAH-LD (Algorithm 3), which is a variance-reduced version of Langevin dynamics which is the current state-of-the-art in terms of KL divergence concentration to the target distribution. SARAH-LD allows us to control the bias, and trade-off the computational complexity with the KL-divergence between sampled and target distributions, i.e. $KL(\rho \parallel \nu)$. The total amount of stochastic gradient evaluations that need to be done for any of the samples (also known as the gradient complexity) of SARAH-LD is $\widetilde{O}\left(\left(n + \frac{dn^{\frac{1}{2}}}{\epsilon}\right) \cdot \frac{\gamma^2 L^2}{\alpha^2}\right)$, complete result is deferred to Theorem 9.

3 Related work

In addition to the work discussed in the introduction, we present an overview of relevant related work.

Posterior sampling was introduced by Thompson (Thompson, 1933) in the context of clinical trials and was later used in the context of reinforcement learning by (Strens, 2000). It has since been found to yield good theoretical guarantees (Chowdhury & Gopalan, 2019; Chowdhury et al., 2021; Dai et al., 2022; Fan & Ming, 2021).

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

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

4 Exact posteriors

If the distributions fulfil the log-Sobolev inequality, we know that this implies sub-Gaussian concentration from Equation (2). We can then define confidence sets on $\bar{\mathcal{T}}_M$ and $\bar{R}_M$ and use simplifying notation $z = (s, a)$ with $Z = \mathcal{S} \times \mathcal{A}$.

$$C_{R,l} = \left\{ f : Z \to \mathbb{R} \;\middle|\; |f(z) - E_{P(\theta|D_l)}[\bar{R}(\theta)]| \leq \sqrt{\frac{L_{\bar{R}}^2 \log 1/\delta}{\alpha_{r,l}}} \right\} \tag{3}$$

$$C_{\mathcal{T},l} = \left\{ f : Z \to \mathbb{R}^d \;\middle|\; \|f(z) - E_{P(\theta|D_l)}[\bar{\mathcal{T}}(\theta)]\|_2 \leq \sqrt{\frac{dL_{\bar{\mathcal{T}}}^2 \log 1/\delta}{\alpha_{p,l}}} \right\} \tag{4}$$

to hold for a probability $0 \leq \delta \leq 1$. These confidence sets can then be used to create a PSRL regret analysis.

First, we must define a Lipschitz condition on the next step value functions. We define the one step future value function $U(\varphi)$ as the expected value of the optimal policy π_l in M_l where φ is the distribution of next state samples. This gives $U_h^{M_l}(\varphi) = \mathbb{E}_{s' \sim \varphi}\left[V^{\pi_l, h+1}(s')\right]$. We use this definition, which is also used in previous work, (Chowdhury & 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 2. *For any φ_1, φ_2 distributions over $\mathcal{S}$ with $1 \leq h \leq H$,*

$$|U_h^M(\varphi_1) - U_h^M(\varphi_2)| \leq L_M \|\overline{\varphi_1} - \overline{\varphi_2}\|_2 \tag{5}$$

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

We can then finally create the desired theorem on Bayes regret under log-Sobolev posteriors.

Theorem 3 (Bayes regret of PSRL under log-Sobolev posteriors)**.** *If the posteriors over the MDP M_l sampled at episode l, $\mathbb{P}(M_l \mid \mathcal{H}_l)$, fulfil LSI and with $M_l = (\mathcal{T}_{M_l}, R_{M_l})$ LSI constants are $\alpha_{p,l}$ and*

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

$\alpha_{r,l}$ and the mean reward for any MDP M $|\bar{R}_M(s)| \leq B_R \forall s$, the one step value function is Lipschitz in the state with parameter L_{M_*} as in Assumption 2 and that the mean reward and mean transitions are Lipschitz in θ with parameters $L_{\bar{R}}, L_{\bar{T}}$ respectively. We then obtain a PSRL Bayesian regret

$$\mathbb{E}[\textit{Regret}(T)] \leq 2H \left(L_{\bar{R}} \sqrt{\log 8T} \sum_{l=1}^{\tau} \frac{1}{\sqrt{\alpha_{r,l}}} + 2\,\mathbb{E}[L_{M_*}] L_{\bar{T}} \sqrt{d \log 8T} \sum_{l=1}^{\tau} \frac{1}{\sqrt{\alpha_{p,l}}} \right) + B_R \quad (6)$$

It is clear that this result leads to $\mathbb{E}[\text{Regret}(T)] = \widetilde{O}(\sqrt{T})$ if $\alpha_l = \Omega(T)$, in Section 6 we study how this holds for different families of distributions which is also summarized in Table 1.

The proof can be found in Appendix B and utilizes a result from (Chowdhury & Gopalan, 2019) which transforms the Bayesian regret into a function of the mean reward and transition functions. We can then use the confidence sets defined above and prove their concentration to obtain the final result.

5 Approximate posteriors

We know that constant approximation error for posteriors leads to linear regret in the context of Thompson sampling for multi-armed bandits (Phan et al., 2019). This also happens in other reinforcement learning setups, like contextual bandits and MDPs. Previous work has noted (Mazumdar et al., 2020) that proper decay of this error can allow for sublinear regret in multi armed bandits. The results of Mazumdar et al.; Zheng et al. were used to design an approximate algorithm for multi armed bandits and strongly log-concave posteriors. However their results do not include planning over episodes, which is essential for MDPs, and the regret analysis heavily depended on strong log-concavity assumptions. We aim to show that this philosophy of constructing approximate posteriors with proper concentration rates can be applied also to MDPs and with only an isoperimetric assumption (LSI in Definition 1) instead of log-concavity. To start, we derive Theorem 4 that shows how can we control the error rate of concentration of posteriors in RL.

Theorem 4. *Let a policy the start of episode l plan according to a posterior Q_l where $\min(KL((P_l \parallel Q_l), KL(Q_l \parallel P_l)) \leq \epsilon_{post,l}$ and where P_l is the true posterior at start of episode l and $|\bar{R}_M| \leq B_R$. Then the incurred regret from planning with an approximate posterior bounded by $\sqrt{2}\Delta_{max}\sqrt{\epsilon_{post,l}}$.*

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

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

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

5.1 LaPSRL

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

By combining Theorem 4 with log-Sobolev theory and SARAH-LD we obtain, for any log-Sobolev posterior, order optimal Bayesian regret while still limiting the computational gradient complexity of each episode to a quadratic polynomial.

Corollary 2. *For a posterior fulfiling the Assumption 1 and definition 1, a posterior sampling style algorithm can obtain an unchanged order of regret under SARAH-LD sampling under a gradient*

Algorithm 1 Langevin PSRL (LaPSRL)

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

Algorithm 2 LANGEVIN SAMPLE

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

complexity for episode l of

$$\textit{Gradient complexity} = \widetilde{O}\left(\frac{H^3l^3L^2}{\alpha_l^2} + \frac{dH^{2.5}l^{3.5}L^2}{\alpha_l^2 g(H,\mathcal{S},\mathcal{A})^2}\right) \tag{7}$$

In many cases, as seen in Section 6, $\alpha_l = \widetilde{\Omega}(Hl)$. This then becomes

$$\textit{Gradient complexity} \propto \widetilde{O}\left(HlL^2 + \frac{d\sqrt{H}l^{3/2}L^2}{g(H,\mathcal{S},\mathcal{A})^2}\right) \tag{8}$$

5.1.1 Chained samples

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

Theorem 5. *Let $\rho_*^l(\theta)$ be the final sample (i.e. after k steps) of the Langevin algorithm at episode l, approximating the true posterior $\mathbb{P}(\theta \mid D_l)$ with $KL(\rho_*^l(\theta) \parallel \mathbb{P}(\theta \mid D_l)) \leq \epsilon_{post,l}$. Additionally, if $\nabla_z \log P(z|\theta)$ is L_z-Lipschitz and α_z-Log Sobolev, we get $\mathbb{E}_{\mathbb{P}(z|D_l)} KL(\rho_*^l(\theta) \parallel \mathbb{P}(\theta \mid D_{l+1})) \leq \epsilon_{post,l} + \frac{L_z}{2\alpha_z} Var_{\rho_*^l(\theta)}(\theta)$, where $Var_{\rho_*^l(\theta)}(\theta)$ is the variance of the approximate posterior distribution $\rho_*^l(\theta)$.*

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

6 Applications of LaPSRL across Different Distributions

In this section, we study a variety of log-Sobolev distributions. We show their log-Sobolev constants and ultimately apply Theorem 3 to calculate the Bayesian regret of LaPSRL for such posteriors.

Table 1: Overview of log-Sobolev constants and Bayes regrets of LaPSRL for different families of distributions.

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

6.1 Univariate Gaussian

For illustrative purposes, we calculate the relevant constants for a Gaussian posterior with known variance σ^2. Here we also assume a Gaussian $(0, \sigma_0^2)$ prior over the mean μ. We have $P(\mu|D) \propto e^{-\left(\frac{\mu^2}{2{\sigma_0}^2} + \sum_{i=1}^n \frac{(\mu - x_i)^2}{2\sigma^2}\right)} = e^{-\left(n\frac{1}{n}\sum_{i=1}^n \left(\frac{\mu^2}{2n{\sigma_0}^2} + \frac{(\mu-x_i)^2}{2\sigma^2}\right)\right)}$

We can then see that we have $\gamma = n$, $f_i(\mu) = \left(\frac{\mu^2}{2n{\sigma_0}^2} + \frac{(\mu-x_i)^2}{2\sigma^2}\right)$. Since $\nabla_\mu^2 f_i(\mu) = \frac{1}{n\sigma_0^2} + \frac{1}{\sigma^2} \leq L$. Finally, we can use Theorem 1 to calculate α. Since $||\nabla_\mu^2 f_i(\mu)||$ is independent of i in this case, we can see that $\nabla_\mu^2 - \log P(\mu|D) = \nabla_\mu^2 \sum_{i=1}^n f_i(\mu) = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$

which gives $\alpha = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} = L\gamma$.

From Theorem 3 we then get the following.

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

6.2 Mixture Distributions

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

Theorem 6 (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* δ *between components, has* $\alpha_{\textit{Mixture}} \geq \frac{\delta \min p_i \min \alpha_i}{4k(1-\log(\min p_i))}$.

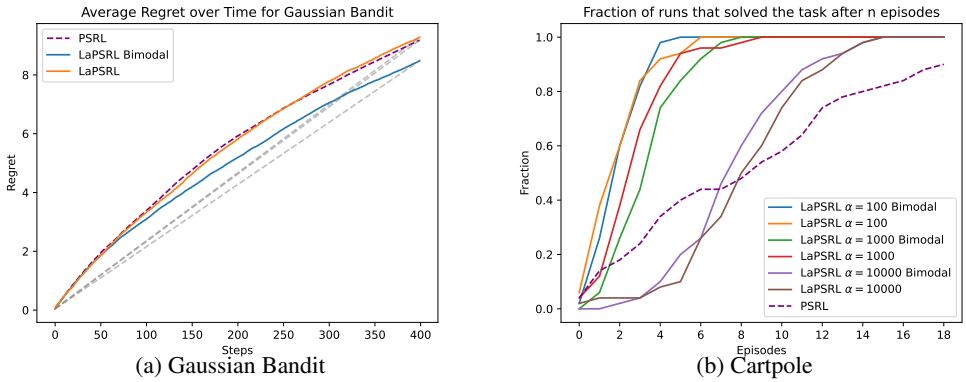
The overlap factor δ relates to integral over the minimum of the paired components, see (Koehler & Vuong, 2024) for more details. If the components are posteriors, this δ should go to 1 as the individual posteriors observe more data and converge.

6.3 Log-concave and Mixture of Log-concave Distributions

Theorem 7. *Any log-concave posterior will have* $\alpha_l = \Theta(n)$*. Similarly, for any posterior that is a mixture of log-concave distributions will have* $\alpha_{\textit{Mixture}} = \Omega\left(\frac{n \min p_i}{4k(1-\log(\min p_i))}\right)$

This result comes from the superadditivity of minimum eigenvalues of Hessians and therefore LSI constants for log-concave distributions. The mixture result follows from Theorem 6. A proof of the theorem can be found in Appendix D.

Combining Theorem 3 and Theorem 7 we obtain the following corollary



(a) Gaussian Bandit (b) Cartpole

Figure 2: We compare LaPSRL versus baseline PSRL. On the left we compare the expected regret for a Gaussian bandit algorithm, and on the right we compare how many episodes it takes to solve a Cartpole task. In both environments, we average over 50 independent runs. The plots are included full size in the appendix.

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

6.4 General log-Sobolev distributions

Theorem 8. *Any log-Sobolev prior with a likelihood ratio* $1/\Gamma \leq \frac{P(X|\theta)}{P(X|\theta')} \leq \Gamma$ *will have a log-Sobolev posterior.*

Proof. This result follows directly from the Holley-Stroock perturbation result found in Theorem 2. This also yields a vacuous bound $1/\Gamma^{2n}\alpha_{\text{prior}} \leq \alpha_l \leq \Gamma^{2n}\alpha_{\text{prior}}$. ■

Unfortunately, we have yet to prove that the log-Sobolev constant of a posterior, under some suitable conditions on the likelihood, will always scale as $\Omega(n)$. Although we conjecture that this is possible and this also matches the intuition from the asymptotic results of the Bernstein–von Mises theorem which gives a log-Sobolev constant of $\Theta(n)$ as $n \to \infty$.

7 Experimental Analysis

We run a set of experiments on two environments to verify that the LaPSRL is competitive. While these experiments are not exhaustive, they serve to show that the algorithm is sound. First, we deploy LaPSRL on a Gaussian multi-armed bandit task with two arms. Second, we perform experiments with a LQR (Kalman, 1960) setup on the Cartpole environment (Barto et al., 1983). We also perform experiments to visualize how SARAH-LD samples from posteriors.

7.1 Gaussian multi-armed bandits

We use LaPSRL on a Gaussian multi-armed bandit task with two arms. The arms generate rewards as $N(0, 0.25)$, $N(0.1, 0.25)$. As a baseline, we compare with the performance of PSRL from the true posterior. Both LaPSRL and Thompson sampling use a N(0,1) prior for the mean of each arm. Additionally, we compare with a LaPSRL algorithm that has a bimodal 1/2 N(0,1/4) + 1/2 N(1,1) prior over the arms. The results can be seen in Figure 2(a). There we see that LaPSRL performs almost identically to PSRL, which is to be expected. Additionally, the LaPSRL with a bimodal prior is converging faster to the correct arm, this could be due to the prior being better adapted to the true distribution but also could indicate a benefit of being able to have mixture distribution priors.

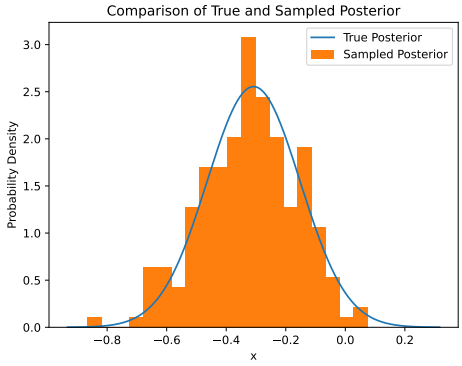


Figure 3: Samples vs. true distribution with $\epsilon = 0.1$ and a Gaussian posterior with 10 observations.

7.2 Continuous MDPs

To evaluate the performance on MDPs we evaluate on the Cartpole environment. We use a continuous control version of the task with states $s \in \mathbb{R}^4$ and a continuous action in [-1,1]. We use a Linear Quadratic Regulator model, where LaPSRL samples from a distribution over the A and B matrixes with a 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 $\alpha \in \left\{\frac{100}{n}, \frac{10000}{n}, \frac{10000}{n}\right\}$. To simplify the parameter search, we set the L parameter to αn. Instead of estimating $\log \frac{2KL(\rho_0 \parallel P(\theta|D_l))}{\epsilon_{\text{post},l}}$, we upper bound this with n. In each sampling step, we start with an initial sample from $\mathcal{N}(0,1)$. 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 Figure 2(b) where we plot what fraction of the 50 runs have solved the task (i.e. taking 200 steps without failing). Here we see that all versions successfully handle the task, even faster than the PSRL baseline. We can note that it takes longer for the experiments with larger α values to converge.

7.3 Evaluate posterior approximation

To illustrate the convergence of SARAH-LD to the true posterior, we also include experiments in Figure 2 which illustrates the correctness of the approximation. If anything, it seems the approximation has a somewhat lower variance than the true posterior.

8 Conclusions and future work

In this paper, we aim to understand whether we can design algorithms with sublinear regret for any isoperimetric distribution. We specifically study PSRL type algorithms for posteriors satisfying log-Sobolev inequalities. We show that if we can compute exact posteriors and sample from them, PSRL can achieve $\widetilde{O}(\sqrt{T})$ regret in an episodic MDP if log-Sobolev constant scales linearly, which we show is true in many cases. We further design a generic Langevin sampling based extension of PSRL, namely LaPSRL. We show that LaPSRL also achieves $\widetilde{O}(\sqrt{T})$ regret if the posterior for the Langevin sampling algorithm contracts at a linear rate with the number of episodes. We plug-in SARAH-LD as the Langevin sampling algorithm, and derive upper bounds on the required gradient complexity and chained sample complexity. We further specify LaPSRL's regret bound for gaussian, mixture of gaussians, log-concave and mixture of log-concave distributions showing LaPSRL can achieve sublinear regret in all these cases. Finally, we test LaPSRL in bandit and LQR environments with Gaussian and mixture priors. We show that the variants of LaPSRL perform competitevly with respect to classical PSRL in all these settings. In the future, it will be interesting to extend LaPSRL's analysis to neural tangent kernel's which can give a better understanding of deep RL methods.

Acknowledgments and Disclosure of Funding

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

References

Azar, M. G., Osband, I., & Munos, R. (2017). Minimax regret bounds for reinforcement learning. *International Conference on Machine Learning*, 263–272.

Bakry, D., Gentil, I., Ledoux, M., et al. (2014). *Analysis and geometry of markov diffusion operators* (Vol. 103). Springer.

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

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-13*(5), 834–846. https://doi.org/10.1109/TSMC.1983.6313077

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

Cattiaux, P., Guillin, A., & Wu, L.-M. (2010). A note on talagrand's transportation inequality and logarithmic sobolev inequality. *Probability theory and related fields*, *148*, 285–304.

Chafaï, D., & Lehec, J. (2023). *Logarithmic sobolev inequalities essentials*. https://djalil.chafai.net/docs/M2/chafai-lehec-m2-lsie-lecture-notes.pdf

Chen, H.-B., Chewi, S., & Niles-Weed, J. (2021a). Dimension-free log-sobolev inequalities for mixture distributions.

Chen, H.-B., Chewi, S., & Niles-Weed, J. (2021b). Dimension-free log-sobolev inequalities for mixture distributions. *Journal of Functional Analysis*, *281*(11), 109236. https://doi.org/https://doi.org/10.1016/j.jfa.2021.109236

Chowdhury, S. R., & Gopalan, A. (2019). Online learning in kernelized markov decision processes.

Chowdhury, S. R., Gopalan, A., & Maillard, O.-A. (2021). Reinforcement learning in parametric mdps with exponential families. In A. Banerjee & K. Fukumizu (Eds.), *Proceedings of the 24th international conference on artificial intelligence and statistics* (pp. 1855–1863). PMLR. https://proceedings.mlr.press/v130/chowdhury21b.html

Dai, Z., Shu, Y., Low, B. K. H., & Jaillet, P. (2022). Sample-then-optimize batch neural thompson sampling.

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

Fan, Y., & Ming, Y. (2021). Model-based reinforcement learning for continuous control with posterior sampling. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th international conference on machine learning* (pp. 3078–3087). PMLR. http://proceedings.mlr.press/v139/fan21b.html

Geramifard, A., Walsh, T. J., Tellex, S., Chowdhary, G., Roy, N., & How, J. P. (2013). A tutorial on linear function approximators for dynamic programming and reinforcement learning. *Foundations and Trends® in Machine Learning*, *6*(4), 375–451. https://doi.org/10.1561/2200000042

Holley, R., & Stroock, D. W. (1986). Logarithmic sobolev inequalities and stochastic ising models.

Huang, Z., Lam, H., Meisami, A., & Zhang, H. (2023). Optimal regret is achievable with bounded approximate inference error: An enhanced bayesian upper confidence bound framework. *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=vwr4bHHsRT

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

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

Karbasi, A., Kuang, N. L., Ma, Y.-A., & Mitra, S. (2023). Langevin thompson sampling with logarithmic communication: Bandits and reinforcement learning.

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

Kinoshita, Y., & Suzuki, T. (2022). Improved convergence rate of stochastic gradient langevin dynamics with variance reduction and its application to optimization. In A. H. Oh, A. Agarwal, D. Belgrave & K. Cho (Eds.), *Advances in neural information processing systems*. https://openreview.net/forum?id=Sj2z__i1wX-

Koehler, F., Heckett, A., & Risteski, A. (2023). Statistical efficiency of score matching: The view from isoperimetry. *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=TD7AnQjNzR6

Koehler, F., & Vuong, T.-D. (2024). Sampling multimodal distributions with the vanilla score: Benefits of data-based initialization. *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=oAMArMMQxb

Krishnamurthy, V., & Yin, G. (2021). Langevin dynamics for adaptive inverse reinforcement learning of stochastic gradient algorithms. *Journal of Machine Learning Research*, *22*(121), 1–49. http://jmlr.org/papers/v22/20-625.html

Kuang, N. L., Yin, M., Wang, M., Wang, Y.-X., & Ma, Y. (2023). Posterior sampling with delayed feedback for reinforcement learning with linear function approximation. *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=RiyH3z7oIF

Lattimore, T., & Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Ledoux, M. (2006). Concentration of measure and logarithmic sobolev inequalities. In *Seminaire de probabilites xxxiii* (pp. 120–216). Springer.

Mazumdar, E., Pacchiano, A., Ma, Y., Jordan, M., & Bartlett, P. (2020). On approximate thompson sampling with langevin algorithms. *International Conference on Machine Learning*, 6797–6807.

Minka, T. (2000). *Bayesian linear regression* (tech. rep.). Citeseer.

Nguyen-Tang, T., Yin, M., Uehara, M., Wang, Y.-X., Wang, M., & Arora, R. (2024). Posterior sampling via langevin monte carlo for offline reinforcement learning. https://openreview.net/forum?id=WwCirclMvl

Nitanda, A., Wu, D., & Suzuki, T. (2022). Convex analysis of the mean field langevin dynamics. In G. Camps-Valls, F. J. R. Ruiz & I. Valera (Eds.), *Proceedings of the 25th international conference on artificial intelligence and statistics* (pp. 9741–9757). PMLR. https://proceedings.mlr.press/v151/nitanda22a.html

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

Osband, I., & Van Roy, B. (2014). Model-based reinforcement learning and the eluder dimension. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence & K. Weinberger (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2014/file/1141938ba2c2b13f5505d7c424ebae5f-Paper.pdf

Osband, I., & Van Roy, B. (2017). Why is posterior sampling better than optimism for reinforcement learning? *International conference on machine learning*, 2701–2710.

Osband, I., Wen, Z., Asghari, S. M., Dwaracherla, V., Ibrahimi, M., Lu, X., & Van Roy, B. (2023). Approximate Thompson sampling via epistemic neural networks. In R. J. Evans & I. Shpitser (Eds.), *Proceedings of the thirty-ninth conference on uncertainty in artificial intelligence* (pp. 1586–1595). PMLR. https://proceedings.mlr.press/v216/osband23a.html

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

Phan, M., Abbasi Yadkori, Y., & Domke, J. (2019). Thompson sampling and approximate inference. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/f3507289cfdc8c9ae93f4098111a13f9-Paper.pdf

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

Sasso, R., Conserva, M., & Rauber, P. (2023). Posterior sampling for deep reinforcement learning. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato & J. Scarlett (Eds.), *Proceedings of the 40th international conference on machine learning* (pp. 30042–30061). PMLR. https://proceedings.mlr.press/v202/sasso23a.html

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

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

Strens, M. (2000). A Bayesian framework for reinforcement learning. *ICML 2000*, 943–950.

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

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

Vempala, S., & Wibisono, A. (2019). Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/65a99bb7a3115fdede20da98b08a370f-Paper.pdf

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

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

Xu, P., Zheng, H., Mazumdar, E. V., Azizzadenesheli, K., & Anandkumar, A. (2022). Langevin Monte Carlo for contextual bandits. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu & S. Sabato (Eds.), *Proceedings of the 39th international conference on machine learning* (pp. 24830–24850). PMLR. https://proceedings.mlr.press/v162/xu22p.html

Yamamoto, K., Oko, K., Yang, Z., & Suzuki, T. (2023). Mean field langevin actor-critic: Faster convergence and global optimality beyond lazy learning.

Zheng, H., Deng, W., Moya, C., & Lin, G. (2024). Accelerating approximate thompson sampling with underdamped langevin monte carlo.

A Algorithms

For completeness we include the SARAH-LD(Kinoshita & Suzuki, 2022) and PSRL(Osband et al., 2013) algorithms in Algorithm 3 and Algorithm 4 as well as a theorem on the gradient complexity of SARAH-LD in Theorem 9.

Algorithm 3 SARAH-LD

Input: step size $\eta > 0$, batch size B, epoch length m, inverse temperature $\gamma \geq 1$
Initialization: $X_0 = 0, X^{(0)} = X_0$
for $s = 0, 1, \ldots, (K/m)$ **do**
  $v_{sm} = \nabla F\left(X^{(s)}\right)$
  randomly draw $\epsilon sm \sim N\left(0, I_{d\times d}\right)$
  $X_{sm+1} = X_{sm} - \eta v_{sm} + \sqrt{2\eta/\gamma}\epsilon sm$
  for $l = 1, \ldots, m-1$ **do**
    $k = sm + l$
    randomly pick a subset I_k from $\{1, \ldots, n\}$ of size $|I_k| = B$
    randomly draw $\epsilon_{\text{post},l} \sim N\left(0, I_{d\times d}\right)$
    $v_k = \frac{1}{B}\sum_{i_k \in I_k}\left(\nabla f_{i_k}\left(X_k\right) - \nabla f_{i_k}\left(X_{k-1}\right)\right) + v_{k-1}$
    $X_{k+1} = X_k - \eta v_k + \sqrt{2\eta/\gamma}\epsilon_{\text{post},l}$
  end for
  $X^{(s+1)} = X_{(s+1)m}$
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 η such that $\eta \leq \frac{3\alpha\epsilon}{48\gamma\alpha^2}$, then a precision $KL(\rho_k \parallel \nu) \leq \epsilon$ is reached after $k \geq \frac{\gamma}{\alpha\eta}\log\frac{2KL(\rho_0 \parallel \nu)}{\epsilon}$ steps of SARAH-LD. Especially, if we take $B = m = \sqrt{n}$ and the largest permissible step size $\eta = \min(\frac{\alpha}{16\sqrt{2}L^2\sqrt{n}\gamma}, \frac{3\alpha\epsilon}{320dL^2\gamma})$, then the gradient complexity becomes $\widetilde{O}\left(\left(n + \frac{dn^{\frac{1}{2}}}{\epsilon}\right)\cdot\frac{\gamma^2L^2}{\alpha^2}\right)$.*

Algorithm 4 PSRL

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

B Proof on Bayes regret

Theorem 3 (Bayes regret of PSRL under log-Sobolev posteriors)**.** *If the posteriors over the MDP M_l sampled at episode l, $\mathbb{P}(M_l \mid \mathcal{H}_l)$, fulfil LSI and with $M_l = (\mathcal{T}_{M_l}, R_{M_l})$ LSI constants are $\alpha_{p,l}$ and $\alpha_{r,l}$ and the mean reward for any MDP M $|\bar{R}_M(s)| \leq B_R \forall s$, the one step value function is Lipschitz in the state with parameter L_{M_*} as in Assumption 2 and that the mean reward and mean transitions are Lipschitz in θ with parameters $L_{\bar{R}}, L_{\bar{\mathcal{T}}}$ respectively. We then obtain a PSRL Bayesian regret*

$$\mathbb{E}[\textit{Regret}(T)] \leq 2H\left(L_{\bar{R}}\sqrt{\log 8T}\sum_{l=1}^{\tau}\frac{1}{\sqrt{\alpha_{r,l}}} + 2\,\mathbb{E}[L_{M_*}]L_{\bar{\mathcal{T}}}\sqrt{d\log 8T}\sum_{l=1}^{\tau}\frac{1}{\sqrt{\alpha_{p,l}}}\right) + B_R \quad (6)$$

Proof. This proof follows the general flow from Chowdhury and Gopalan (2019) for Kernel PSRL but with major difference, to accommodate the different setting.

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 the under the observed history $\mathcal{H}_{l-1}$ we have

$\mathbb{E}[V^{M_l}_{\pi_l,1}(s_{l,1}) \mid \mathcal{H}_{l-1}] = \mathbb{E}[V^{M_*}_{\pi_*,1}(s_{l,1}) \mid \mathcal{H}_{l-1}]$, since they are both sampled from the same distribution. Marginalising we obtain:

$$\mathbb{E}[V^{M_*}_{\pi_*,1}(s_{l,1}) - V^{M_*}_{\pi_l,1}(s_{l,1})] = \mathbb{E}[V^{M_*}_{\pi_*,1}(s_{l,1}) - V^{M_l}_{\pi_l,1}(s_{l,1})] + \mathbb{E}[V^{M_l}_{\pi_l,1}(s_{l,1}) - V^{M_*}_{\pi_l,1}(s_{l,1})] \quad (9)$$

$$= \mathbb{E}[V^{M_l}_{\pi_l,1}(s_{l,1}) - V^{M_*}_{\pi_l,1}(s_{l,1})] \quad (10)$$

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

$$\mathbb{E}[\text{Regret}(T)] \triangleq \sum_{l=1}^{\tau} \mathbb{E}[V^{M_l}_{\pi_l,1}(s_{l,1}) - V^{M_*}_{\pi_l,1}(s_{l,1})] \quad (11)$$

$$\leq \mathbb{E}[\sum_{l=1}^{\tau}\sum_{h=1}^{H} \left[|\bar{R}_{M_l}(z_{l,h}) - \bar{R}_*(z_{l,h})| + L_{M_l}||\bar{\mathcal{T}}_{M_l}(z_{l,h}) - \bar{\mathcal{T}}_*(z_{l,h})||_2\right]] \quad (12)$$

where L_{M_l} is the global Lipschitz constant one step function of M_l under π_l. where $\bar{\mathcal{T}}_M$ and $\bar{R}_M$ are the mean of the transition and reward distributions for MDP M. Now we fix $0 \leq \delta \leq 1$ and for $1 \leq l \leq \tau$ define two confidence sets

$$C_{R,l} = \left\{ f : Z \to \mathbb{R} \;\middle|\; |f(z) - E_{P(\theta|D_l)}[\bar{R}(\theta)]| \leq \sqrt{\frac{L^2_{\bar{R}} \log 1/\delta}{\alpha_{r,l}}} \right\} \quad (13)$$

$$C_{\mathcal{T},l} = \left\{ f : Z \to \mathbb{R}^d \;\middle|\; \|f(z) - E_{P(\theta|D_l)}[\bar{\mathcal{T}}(\theta)]\|_2 \leq \sqrt{\frac{dL^2_{\bar{\mathcal{T}}} \log 1/\delta}{\alpha_{p,l}}} \right\} \quad (14)$$

Define events $E_* \triangleq \{\bar{R}_* \in C_{R,l}, \bar{\mathcal{T}}_* \in C_{\mathcal{T},l}, \forall 1 \leq l \leq \tau\}$ and $E_M \triangleq \{\bar{R}_{M_l} \in C_{R,l}, \bar{\mathcal{T}}_{M_l} \in C_{\mathcal{T},l}, \forall 1 \leq l \leq \tau\}$. 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 $\mathbb{P}(E^c) \leq \mathbb{P}(E^c_M) + \mathbb{P}(E^c_*) \leq 4\tau H\delta$. We also have that $\mathbb{E}[L_{M_l}] = \mathbb{E}[L_{M_*}]$ such that $\mathbb{E}[L_{M_l}|E] \leq \frac{\mathbb{E}[L_{M_*}]}{P(E)} \leq \frac{\mathbb{E}[L_{M_*}]}{1-4\tau H\delta}$.

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

$$\mathbb{E}[\sum_{l=1}^{\tau}\sum_{h=1}^{H} \left[|\bar{R}_{M_l}(z_{l,h}) - \bar{R}_*(z_{l,h})| \mid E\right] + \mathbb{E}\left[L_{M_l}||\bar{\mathcal{T}}_{M_l}(z_{l,h}) - \bar{\mathcal{T}}_*(z_{l,h})||_2 \middle| E\right] + 2B_R 4\tau H\delta \quad (15)$$

$$\leq 2H\left(L_{\bar{R}}\sqrt{\log 1/\delta}\sum_{l=1}^{\tau}\frac{1}{\sqrt{\alpha_{r,l}}} + \frac{\mathbb{E}[L_{M_*}]}{1-4\tau H\delta}L_{\bar{\mathcal{T}}}\sqrt{d\log 1/\delta}\sum_{l=1}^{\tau}\frac{1}{\sqrt{\alpha_{p,l}}}\right) + 8B_R\tau H\delta \quad (16)$$

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

$$\mathbb{E}[\text{Regret}(T)] \leq 2H\left(L_{\bar{R}}\sqrt{\log 8T}\sum_{l=1}^{\tau}\frac{1}{\sqrt{\alpha_{r,l}}} + 2\,\mathbb{E}[L_{M_*}]L_{\bar{\mathcal{T}}}\sqrt{d\log 8T}\sum_{l=1}^{\tau}\frac{1}{\sqrt{\alpha_{p,l}}}\right) + B_R \quad (17)$$

■

C Proofs on regret for approximate sampling and sample complexity.

Theorem 4. *Let a policy the start of episode l plan according to a posterior Q_l where $\min(KL((P_l \parallel Q_l), KL(Q_l \parallel P_l)) \leq \epsilon_{post,l}$ and where P_l is the true posterior at start of episode l and $|\bar{R}_M| \leq B_R$. Then the incurred regret from planning with an approximate posterior bounded by $\sqrt{2}\Delta_{max}\sqrt{\epsilon_{post,l}}$.*

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

$$E_{P_l,Q_l}[V_{\pi^*}^{\mu^*} - V_{\pi_l'}^{\mu^*}] = \int_{\mu^*}\int_{\mu_l'}(V_{\pi^*}^{\mu^*} - V_{\pi_l'(\mu_l')}^{\mu^*})P_l(\mu^*)Q_l(\mu_l') \tag{18}$$

$$= E_{P_l,Q_l}[V_{\pi^*}^{\mu^*} - V_{\pi_l'}^{\mu_l'} + V_{\pi_l'}^{\mu_l'} - V_{\pi_l'}^{\mu^*}] \tag{19}$$

$$= E_{P_l,Q_l}[V_{\pi^*}^{\mu^*} - V_{\pi_l}^{\mu_l} + V_{\pi_l}^{\mu_l} - V_{\pi_l'}^{\mu_l'} + V_{\pi_l'}^{\mu_l'} - V_{\pi_l'}^{\mu^*}] \tag{20}$$

$$= E_{P_l,Q_l}[[V_{\pi^*}^{\mu^*} - V_{\pi_l}^{\mu_l}] + [V_{\pi_l}^{\mu_l} - V_{\pi_l'}^{\mu_l'}] + [V_{\pi_l'}^{\mu_l'} - V_{\pi_l'}^{\mu^*}]] \tag{21}$$

$$\leq E_{P_l}[V_{\pi^*}^{\mu^*} - V_{\pi_l}^{\mu_l}] + \Delta_{\max}\sqrt{\frac{\epsilon_{\text{post},l}}{2}} + E_{P_l}[V_{\pi_l}^{\mu_l} - V_{\pi_l}^{\mu^*}] + \Delta_{\max}\sqrt{\frac{\epsilon_{\text{post},l}}{2}}] \tag{22}$$

$$= E_{P_l}[V_{\pi^*}^{\mu^*} - V_{\pi_l}^{\mu^*}] + \sqrt{2}\Delta_{\max}\sqrt{\epsilon_{\text{post},l}} \tag{23}$$

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

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

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

$$\widetilde{\mathcal{O}}(\mathbb{E}_P\, R(\pi_Q)) \leq \widetilde{\mathcal{O}}(\sqrt{\tau}g(H,\mathcal{S},\mathcal{A})) + \sqrt{2}\Delta_{\max}\sum_{k=1}^{\tau}\sqrt{\epsilon_{\text{post},l}} \tag{24}$$

$$\leq \widetilde{\mathcal{O}}(\sqrt{\tau}g(H,\mathcal{S},\mathcal{A})) + \sqrt{2}\Delta_{\max}\sum_{k=1}^{\tau}\sqrt{C}\frac{g(H,\mathcal{S},\mathcal{A})}{\sqrt{t}\Delta_{\max}} \tag{25}$$

$$= \widetilde{\mathcal{O}}(\sqrt{\tau}g(H,\mathcal{S},\mathcal{A})) + \sqrt{2}g(H,\mathcal{S},\mathcal{A})\sqrt{C}\sum_{k=1}^{\tau}\frac{1}{\sqrt{t}} \tag{26}$$

$$= \widetilde{\mathcal{O}}(\sqrt{\tau}g(H,\mathcal{S},\mathcal{A})) \tag{27}$$

■

Theorem 5. *Let $\rho_*^l(\theta)$ be the final sample (i.e. after k steps) of the Langevin algorithm at episode l, approximating the true posterior $\mathbb{P}(\theta \mid D_l)$ with $KL(\rho_*^l(\theta) \parallel \mathbb{P}(\theta \mid D_l)) \leq \epsilon_{post,l}$. Additionally, if $\nabla_z \log P(z|\theta)$ is L_z-Lipschitz and α_z-Log Sobolev, we get $\mathbb{E}_{\mathbb{P}(z|D_l)} KL(\rho_*^l(\theta) \parallel \mathbb{P}(\theta \mid D_{l+1})) \leq \epsilon_{post,l} + \frac{L_z}{2\alpha_z}Var_{\rho_*^l(\theta)}(\theta)$, where $Var_{\rho_*^l(\theta)}(\theta)$ is the variance of the approximate posterior distribution $\rho_*^l(\theta)$.*

Proof. For notation we write $P(\theta \mid D_{l+1}) = P(\theta \mid D_l, z_l)$ such that $P(\theta \mid D_{l+1}) = P(\theta \mid z_0, \ldots, z_{l-1})$. Note that $P(z_l \mid D_l, \theta) = P(z_l \mid \theta)$ and $\mathbb{E}_\theta\, P(z_l \mid D_l, \theta) = P(z_l \mid D_l)$.

$$KL(\rho_*^l \parallel \nu_{l+1} \mid z_l) \tag{28}$$

$$= \int_\theta \log\left(\frac{\rho_*^l(\theta)}{\mathbb{P}(\theta \mid D_l, z_l)}\right)\rho_*^l(\theta)d\theta \tag{29}$$

$$= \int_\theta \log\left(\frac{\rho_*^l(\theta)}{\frac{\mathbb{P}(\theta|D_l)P(z_l|\theta)}{P(z_l|D_l)}}\right)\rho_*^l(\theta)d\theta \tag{30}$$

$$= \int_\theta \left(\log\left(\frac{\rho_*^l(\theta)}{\mathbb{P}(\theta \mid D_l)}\right) + \log\left(\frac{P(z_l|D_l)}{P(z_l|\theta)}\right)\right)\rho_*^l(\theta)d\theta \tag{31}$$

$$= \int_\theta \log\left(\frac{\rho_*^l(\theta)}{\mathbb{P}(\theta \mid D_l)}\right)\rho_*^l(\theta)d\theta \tag{32}$$

$$+ \int_\theta \log\left(\frac{P(z_l|D_l)}{P(z_l|\theta)}\right)\rho_*^l(\theta)d\theta \tag{33}$$

$$= KL(\rho_*^l \parallel \nu_l) + \int_\theta \log\left(\frac{P(z_l|D_l)}{P(z_l|\theta)}\right)\rho_*^l(\theta)d\theta \tag{34}$$

$$\leq \epsilon_{\text{post},l} + \int_\theta \log\left(\frac{P(z_l|D_l)}{P(z_l|\theta)}\right)\rho_*^l(\theta)d\theta \tag{35}$$

$$= \epsilon_{\text{post},l} + \int_\theta \log\left(P(z_l|D_l)\right)\rho_*^l(\theta)d\theta - \int_\theta \log\left(P(z_l|\theta)\right)\rho_*^l(\theta)d\theta \tag{36}$$

$$\tag{37}$$

This then gives

$$\mathbb{E}_{z_l} KL(\rho_*^l \parallel \nu_{l+1} \mid z_l) \tag{38}$$

$$\leq \epsilon_{\text{post},l} + \int_{z_l}\int_\theta \log\left(\frac{P(z_l|D_l)}{P(z_l|\theta)}\right)\rho_*^l(\theta)d\theta P(z_l \mid D_l)dz_l \tag{39}$$

$$= \epsilon_{\text{post},l} + \int_\theta\int_{z_l} \log\left(\frac{P(z_l|D_l)}{P(z_l|\theta)}\right)\rho_*^l(\theta)P(z_l \mid D_l)dz_l d\theta \tag{40}$$

$$= \epsilon_{\text{post},l} + \int_\theta\int_{z_l} \log\left(\frac{P(z_l|D_l)}{P(z_l|\theta)}\right)\frac{P(z_l|D_l)}{P(z_l|\theta)}P(z_l|\theta)dz_l\rho_*^l(\theta)d\theta \tag{41}$$

$$\leq \epsilon_{\text{post},l} + \int_\theta 2/\alpha_z \int_{z_l} ||\nabla_z\sqrt{\frac{P(z_l|D_l)}{P(z_l|\theta)}}||^2 P(z_l|\theta)dz_l\rho_*^l(\theta)d\theta \tag{42}$$

$$= \epsilon_{\text{post},l} + \int_\theta 2/\alpha_z \int_{z_l} ||\frac{P(z_l|\theta)\nabla_z P(z_l|D_l) - P(z_l|D_l)\nabla_z P(z_l|\theta)}{2\sqrt{\frac{P(z_l|D_l)}{P(z_l|\theta)}}P(z_l|\theta)^2}||^2 P(z_l|\theta)dz_l\rho_*^l(\theta)d\theta \tag{43}$$

$$= \epsilon_{\text{post},l} + \int_\theta 2/\alpha_z \int_{z_l} ||\frac{P(z_l|\theta)\nabla_z P(z_l|D_l) - P(z_l|D_l)\nabla_z P(z_l|\theta)}{2\ P(z_l|D_l)P(z_l|\theta)} \times \sqrt{\frac{P(z_l|D_l)}{P(z_l|\theta)}}||^2 P(z_l|\theta)dz_l\rho_*^l(\theta)d\theta \tag{44}$$

$$= \epsilon_{\text{post},l} + \int_\theta 2/\alpha_z \int_{z_l} ||\frac{P(z_l|\theta)\nabla_z P(z_l|D_l) - P(z_l|D_l)\nabla_z P(z_l|\theta)}{2\ P(z_l|D_l)P(z_l|\theta)} \times \sqrt{\frac{P(z_l|D_l)}{P(z_l|\theta)}}||^2 P(z_l|\theta)dz_l\rho_*^l(\theta)d\theta \tag{45}$$

$$= \epsilon_{\text{post},l} + \int_\theta 2/\alpha_z \int_{z_l} ||\frac{P(z_l|\theta)\nabla_z P(z_l|D_l) - P(z_l|D_l)\nabla_z P(z_l|\theta)}{2\ P(z_l|D_l)P(z_l|\theta)}||^2 \frac{P(z_l|D_l)}{P(z_l|\theta)}P(z_l|\theta)dz_l\rho_*^l(\theta)d\theta \tag{46}$$

$$= \epsilon_{\text{post},l} + \int_\theta \frac{1}{2\alpha_z}\int_{z_l} ||\nabla_z \log P(z_l|D_l) - \nabla_z \log P(z_l|\theta)||^2 P(z_l|D_l)dz_l\rho_*^l(\theta)d\theta \tag{47}$$

$$\leq \epsilon_{\text{post},l} + \frac{L_z}{2\alpha_z}\int_\theta ||\theta_l - \theta||^2\rho_*^l(\theta)d\theta \tag{48}$$

$$= \epsilon_{\text{post},l} + \frac{L_z}{2\alpha_z}\text{Var}_{\rho_*^l(\theta)}(\theta) \tag{49}$$

■

The inequality in Equation (42) comes from the log-Sobolev inequality property of $P(z_l|\theta)$. The rest is algebra with the exception of the final inequality which comes from the Lipschitz property.

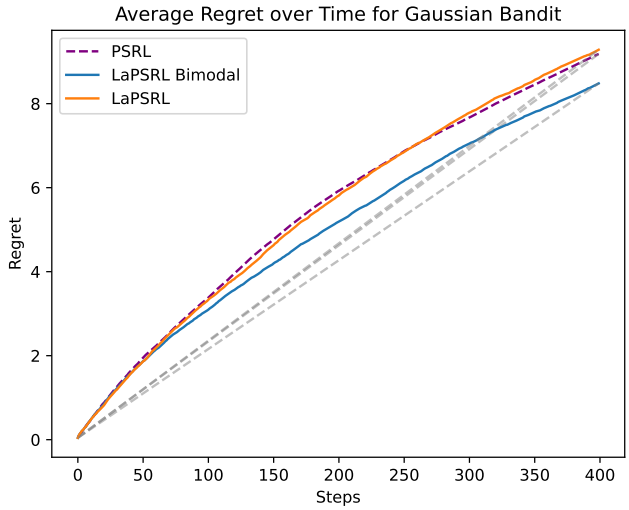
D Proofs for theorems on LSI constants.

Theorem 7. *Any log-concave posterior will have $\alpha_l = \Theta(n)$. Similarly, for any posterior that is a mixture of log-concave distributions will have $\alpha_{Mixture} = \Omega\left(\frac{n \min p_i}{4k(1-\log(\min p_i))}\right)$*

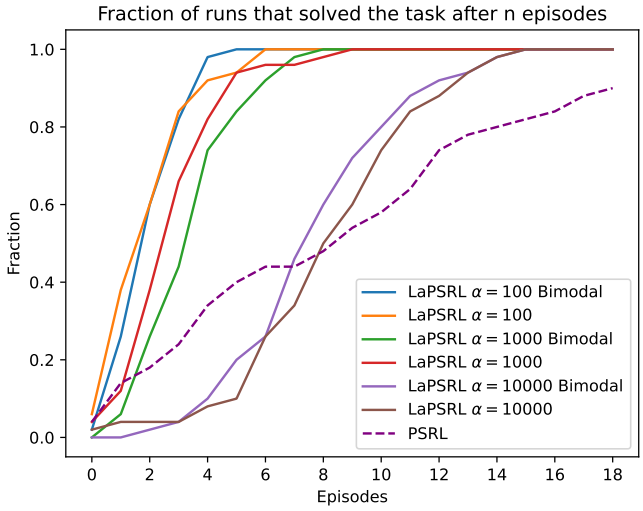
Proof. We can write the product of log-concave distributions $P(\theta \mid D_l) = P(\theta)\frac{\prod_{i=1}^{n} P_i(\theta)}{Z}$ where $P_i(\theta)$ is shorthand for $P(x_i \mid \theta)$ with x_i the datapoint at time i. Since products preserve log-concavity, we can use Theorem 1. From Weyl's inequality,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

Similarly, for mixtures of log-concave distributions we have from Theorem 6 that $\alpha_{\text{Mixture}} = \Omega\left(\frac{\min \alpha_i \min p_i}{4k(1-\log(\min p_i)}\right)$. Setting $\min_i \alpha_i = \Theta(n)$ completes the proof. ■

E Experimental results



(a) Gaussian Bandit



(b) Cartpole

Figure 4: We compare LaPSRL versus baseline PSRL. On the left we compare the expected regret for a Gaussian bandit algorithm, and on the right we compare how many episodes it takes to solve a Cartpole task. In both environments, we average over 50 independent runs.