Isoperimetry is All We Need: Langevin Posterior Sampling for RL

Anonymous Author(s)
Affiliation
Address
email

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. In this paper, 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. Specifically, we show that the well-known PSRL (Posterior Sampling-based RL) algorithm yields sublinear regret if the posterior distributions satisfy the Log-Sobolev Inequality (LSI), which is a form of isoperimetry. 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 sublinear regret if the posteriors only 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 competetive 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 impact 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, bilinear, or reproducible kernel [OBM22] type parametric models, and Gaussian or log-concave posteriors for Bayesian algorithms [CG19; OV17]. In this paper, we aim to narrow this gap further by studying whether we can achieve the desired regret guarantees for isoperimetric distributions. Isoperimetric distributions include all the aforementioned setups studied in RL theory, and also non-log-concave and perturbed versions of log-concave distributions. In optimization and sampling literature, isoperimetry is used to give efficient and controlled sampling from non-convex and perturbed 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 [VW19]. Among the different forms of isoperimetric inequalities, we consider the Log Sobolev Inequality (LSI).

Posterior Sampling-based RL (PSRL). For our study, we focus on the popular PSRL algorithms [Rus+20; ORV13], which are generalisation of Thompson sampling proposed for bandits [Tho33]. 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 posteri-

Submitted to 17th European Workshop on Reinforcement Learning (EWRL 2024). Do not distribute.

ors, and uses the sampled parameters for further planning. Since PSRL has been successful both theoretically and practically, we choose it as the base algorithm to study. Still, exact sampling and tracking of the posterior may be intractable for many distributions (e.g. in high dimensions). It is easy to show that approximation in the sampling can lead to linear regret unless sufficient care is taken. On the other hand, being limited to distributions 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.

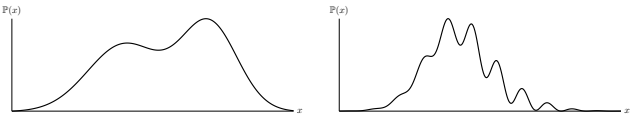


Figure 1: Examples of log-Sobolev distributions.

Langevin Sampling-based PSRLs. One of the growing approaches in this direction is to use Langevin-based approximate sampling methods [Maz+20; Zhe+24; Ish+23], which are known to be generic and efficient in optimisation, sampling, and deep learning literature. Mazumdar et al. [Maz+20] and Zheng et al. [Zhe+24] propose Langevin-based PSRL algorithms for multi-armed bandits that achieve order-optimal regret only for log-concave distributions. Similarly, Xu et al. [Xu+22] extends these ideas to linear contextual bandits but still with a linear dependence on the approximation error. Ishfaq et al. [Ish+23] 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, i.e. the ones that satisfy LSI or Poincaré inequalities. 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. This result broadens the scenarios where PSRL is proven to be efficient.

2. Propose a generic PSRL-algorithm, called **LaPSRL**, that u*ses 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, a well-studied 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.

Notations. 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*) [ORV13; AOM17]. MDP in episodic RL is defined

as $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, R, \tau, H \rangle$. 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 $s_{t+1}^l \sim \mathcal{T}(. \mid s_t^l, a_t^l)$. 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 $h \in [H]$

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

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) [LS20]. For bandits, the episode length $H = 1$.

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 [ORV13]. PSRL reduces to Thompson sampling [Tho33], 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 over the possible MDPs and evaluations and can be written $\sum_l^\tau \sum \mathbb{E}[V_{\pi_*,1}^{M_*}(s_{l,1}) - V_{\pi_l,1}^{M_*}(s_{l,1})]$ To calculate regret we use notation $\Delta_{\max} = \max_\pi V_\pi(s) - \min_\pi V_\pi(s)$. In the paper we use n to denote the amount of data samples we have observed. When in the context of scaling we use $T = \tau H$ instead.

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 [VW19]. 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 two 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.*

Assumption 2 (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 setting $\rho = \frac{g^2 \nu}{\mathbb{E}_\nu[g^2]}$, which gives $KL(\rho \parallel \nu) \leq \frac{1}{2\alpha} J_\rho$ 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. Generally, it is the smallest α that is known as the LSI constant, which is the one that will be indicated as α from now on.

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 [CL23; VW19]. Also note that in some work an inverse definition is used where the constant is defined $\alpha' = \frac{1}{2\alpha}$, leading to some confusion.

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

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

In other cases Lyaponov conditions [CGW10], integral conditions [Wan01] or decomposing into mixtures [CCN21b; KHR23] can be utilized. Additionally, since LSI implies the Poincare inequality (see Assumption 3) with the same constant [VW19], it can be easier to find the Poincare constant instead. Theorem 1 shows that log-concave distributions imply LSI, but log-Sobolev distributions are more general. Some examples of what log-Sobolev distributions could look like can be found in Figure 1. For example, a log-Sobolev distribution with added bounded perturbations (and sometimes even unbounded) would still fulfil LSI, but would generally break the log-concave property. [Ste21]. The LSI is also preserved under a Lipschitz-transformation,[VW19] 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 [Led06]. 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 [Biz23] 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 some curvature conditions, the reverse is true, Gaussian concentration implies that the distribution is log-Sobolev [BGL+14, Theorem 8.7.2].

Background: SARAH-LD [KS22]. There exists multiple algorithms for performing biased Langevin sampling on log-Sobolev distributions [VW19; KS22]. 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 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 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 7.

3 Related work

Posterior sampling was introduced by Thompson[Tho33] in the context of clinical trials and was later used in the context of reinforcement learning by Strens[Str00]. It has since been found to good theoretical guarantees[CG19; FM21; CGM21; Dai+22].

Sometimes approximations are required, either because calculating or sampling from the posterior is intractable[WCM23; SCR23; Osb+23]. While these papers have frameworks for approximate sampling, none of them comes with any regret guarantees.[1] Fan and Ming [FM21] also study the case of function approximation, but the theory does not hold there. The work of Huang et al. [Hua+23] 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 [Kim23; DV20; Yam+23], but this work comes without any theoretical guarantees. In Nguyen-Tang et al. [Ngu+24] they use Langevin for offline RL and in [Kua+23] it is for linear MDPs. The work of Karbasi et al. [Kar+23] 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 [Zhe+24], but with requirements on convexity. Finally, [Kua+23] uses these ideas for delayed feedback RL, but limited to Linear MDPs and Krishnamurthy and Yin [KY21] uses Langevin dynamics for inverse reinforcement learning.

4 Exact posteriors

We follow the general outline from Chowdhury and Gopalan [CG19] to create a generic regret proof for log-Sobolev posteriors.

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

Theorem 2 (Bayes regret of PSRL under log-Sobolev posteriors)**.** *If the posteriors $\mathbb{P}(M_l \mid \mathcal{H}_l)$ fulfill 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 $|\bar{R}_M(s)| \leq B_R \forall s$. We then obtain a PSRL Bayesian regret*

$$\mathbb{E}[\textit{Regret}(T)] \leq 2H \left(L_{\bar{R}} \sqrt{\log(B_R T H)} \sum_{l=1}^{\tau} \frac{1}{\sqrt{\alpha_{r,l}}} + L_{M_*} L_{\bar{\mathcal{T}}} \sqrt{d \log(B_R T H)} \sum_{l=1}^{\tau} \frac{1}{\sqrt{\alpha_{p,l}}} \right) + 8 \tag{3}$$

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

We prove this by Using Lemma 7 from [CG19] which transforms the Bayesian regret into a function of $\bar{\mathcal{T}}$ and $\bar{R}$ using a Lipschitz property. These functions are the mean of the transition and reward distributions for MDP M.

We then define confidence sets on $\bar{\mathcal{T}}_M$ and $\bar{R}_M$.

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

$$C_{\mathcal{T},l} = \left\{ f : Z \to \mathbb{R}^d \;\middle|\; \|f(z) - E_{P(\theta|D_l)}[\bar{\mathcal{T}}(\theta)]\|_2 \leq \sqrt{\frac{d L_{\bar{\mathcal{T}}}^2 \log 1/\delta}{\alpha_{p,l}}} \right\} \tag{5}$$

These can be then used together with the Gaussian concentration of log-Sobolev distributions from Equation (2) to hold for a probability $0 \leq \delta \leq 1$. The rest of the proof follows straightforwardly.

5 Approximate posteriors

It is a known result that even a small, but constant, approximation error will lead to linear regret in the context of Thompson sampling for multi-armed bandits [PAD19]. This is not unique to bandits and will also apply to reinforcement learning. Previous work has noted [Maz+20] that decay of this error can allow for unchanged regret complexity in bandits. An illustration of this is given in Theorem 3.

Theorem 3. *Let a policy the start of episode l plan according to a posterior Q_l where $\min(KL((P_l \parallel Q_l), KL(P_l \parallel Q_l)) \leq \epsilon_{post,l}$ 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 controls the total variation, the proof can be found in Appendix D.

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

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 3 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.

Algorithm 1 Langevin PSRL (LaPSRL)

Input: Likelihood $f(x|\theta)$, Prior $P(\theta)$, Horizon H, total episodes τ, Regret complexity $g(H, S, A, D)$
for $l = 1 : \tau$ **do**
  $\epsilon_{\text{post},l} = \frac{g(H,S,A,D)}{l\Delta_{\max}^2}$
  if Chained sampling **then**
    $\rho_0 = \theta_{l-1}$
  else
    $\rho_0 \sim P(\theta)$
  end if
  Sample $\theta_l = \text{Sample\_parameter}(f, P(\theta), x, \epsilon_{\text{post},l}, \rho_0)$
  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 Sample_parameter

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

Corollary 2. *For a posterior fulfilling the Assumptions 1 and 2, a posterior sampling style algorithm can obtain an unchanged regret complexity under SARAH-LD sampling under a gradient complexity for each episode of*

$$\textit{Gradient complexity} = \widetilde{O}\left(\frac{H^3l^3L^2}{\alpha_l^2} + \frac{dH^{2.5}l^{3.5}L^2}{\alpha_l^2 g(H,S,A,D)^2}\right) \tag{6}$$

If $\alpha_l = \Theta(Hl)$, this becomes

$$\textit{Gradient complexity} \propto \widetilde{O}\left(HlL^2 + \frac{d\sqrt{H}l^{3/2}L^2}{g(H,S,A,D)^2}\right) \tag{7}$$

We will see in Theorem 6 that in many cases, $\alpha_l = \Theta(Hl)$.

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 4. *Let $\rho_*^l(\theta)$ be the distribution final sample (i.e. after k steps) 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 $\rho_*^l(\theta)$.*

See Appendix D for the 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 will not affect the regret complexity. One problem is that the variance is taken under the approximative distribution $\rho_*^l(\theta)$, but in practice, we know that this is an $\epsilon_{\text{post},l}$ close approximation. We also know that

Table 1: Overview of log-Sobolev constant and PSRL Bayes regret for different families of distributions.

Posterior	log-Sobolev constant	PSRL BayesRegret
Gaussian	$\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$	$\widetilde{O}\left(\sqrt{H\tau\sigma^2}\right)$
Log-concave	$\Theta(n)$	$\widetilde{O}(\sqrt{\tau H})$
Mixture of Log-concave	$\Omega\left(\frac{\delta \min p_i \min \alpha_i}{4k(1-\log(\min p_i))}\right)$	$\widetilde{O}\left(\sqrt{\frac{4kH\tau}{\min p_i}}\right)$

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 2 to calculate the Bayesian regret of PSRL for such posteriors.

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 2 we then get the following.

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

6.2 Mixture distributions

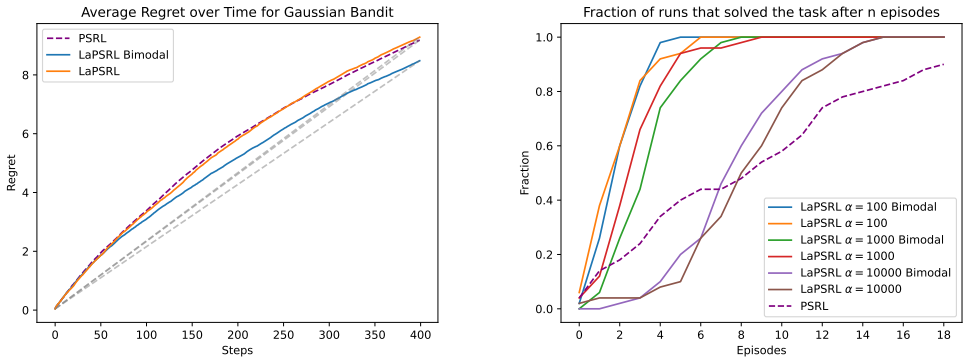
There has been multiple work looking into log-Sobolev constants for mixtures of log-Sobolev distributions[KV24; CCN21a; Sch19]. Generally, it depends on constants of the mixture components as well as a function of the distance between the components. Koehler and Vuong [KV24] show

Theorem 5 (Informally from Theorem 2 [KV24])**.** *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 [KV24] for more details. If the components are posteriors, this δ should go to 1 as the individual posteriors observe more data and converge.

Combining this result with Theorem 2 we obtain the following corollary.

Corollary 4. *Under the conditions of Theorem 5, a PSRL obtains* $\mathbb{E}[\textit{Regret}(T)] = \widetilde{O}\left(\sqrt{\frac{4kH\tau}{\min p_i}}\right)$ *with a mixture of log-concave posterior.*



(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.

6.3 Log-concave and Mixture of Log-concave Distributions

Theorem 6. *Any log-concave posterior will have* $\alpha_l = \Theta(n)$. *Similarly, for any posterior that is a mixture of log-concave distributions will have* $\alpha_{Mixture} = \Omega\left(\frac{n \min p_i}{4k(1-\log(\min p_i))}\right)$

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

Combining Theorem 2 and Theorem 6 we obtain the following corollary

Corollary 5. *Any log-concave posterior* $|\bar{R}_M(s)| \leq B_R \forall s$ *for all MDPs* M *will have* $\mathbb{E}[Regret(T)] = \tilde{O}\left(\sqrt{H\tau}\left(L_{\bar{R}} + L_{M_*} L_{\bar{T}}\right)\right)$. *for PSRL. Similarly, and under the same condition, any posterior that is a mixture of log-concave posteriors with* $\mathbb{E}[Regret(T)] = \widetilde{O}\left(\sqrt{\frac{4kH\tau}{\min p_i}}\left(L_{\bar{R}} + L_{M_*} L_{\bar{T}}\right)\right)$ *PSRL regret.*

7 Experimental Analysis

We run a set of experiments on two environments to verify that the LaPSRL is competetive. 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 [Kal60] setup on the Cartpole environment[BSA83]. 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 Section 7.1. 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.

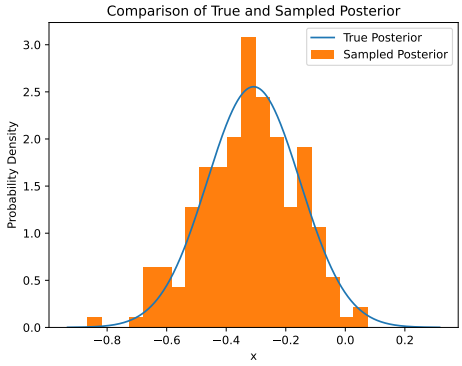


Figure 3: Samples vs true distribution with $\epsilon = 0.1$ and a Gaussian posterior with 10 observations.

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 $N(0, 1)$. As a baseline we, compare with an exact PSRL algorithm which samples from Bayesian linear regression priors [Min00]. Finally, we use a variant of LaPSRL with a multimodal prior over the A and B matrixes with a 1/2 N(0,1) + 1/2 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 Section 7.1 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 $\mathcal{O}(\sqrt{HT})$ regret in an episodic MDP. We further design a generic Langevin sampling based extension of PSRL, namely LaPSRL. We show that LaPSRL also achieves $\mathcal{O}(\sqrt{HT})$ regret if the posterior for the Langevin sampling algorithm contracts at a linear rate. 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 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.

References

[AOM17] Mohammad Gheshlaghi Azar, Ian Osband and Rémi Munos. 'Minimax regret bounds for reinforcement learning'. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 263–272.

[BGL+14] Dominique Bakry, Ivan Gentil, Michel Ledoux et al. *Analysis and geometry of Markov diffusion operators*. Vol. 103. Springer, 2014.

[Biz23] Pierre Bizeul. *On the log-Sobolev constant of log-concave measures*. 2023. arXiv: 2306.12997 [math.FA].

[BSA83] Andrew G. Barto, Richard S. Sutton and Charles W. Anderson. 'Neuronlike adaptive elements that can solve difficult learning control problems'. In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-13.5 (1983), pp. 834–846. DOI: 10.1109/TSMC.1983.6313077.

[CCN21a] Hong-Bin Chen, Sinho Chewi and Jonathan Niles-Weed. *Dimension-free log-Sobolev inequalities for mixture distributions*. 2021. arXiv: 2102.11476 [math.PR].

[CCN21b] Hong-Bin Chen, Sinho Chewi and Jonathan Niles-Weed. 'Dimension-free log-Sobolev inequalities for mixture distributions'. In: *Journal of Functional Analysis* 281.11 (2021), p. 109236. ISSN: 0022-1236. DOI: https://doi.org/10.1016/j.jfa.2021.109236.

[CG19] Sayak Ray Chowdhury and Aditya Gopalan. *Online Learning in Kernelized Markov Decision Processes*. 2019. arXiv: 1805.08052 [cs.LG].

[CGM21] Sayak Ray Chowdhury, Aditya Gopalan and Odalric-Ambrym Maillard. 'Reinforcement Learning in Parametric MDPs with Exponential Families'. In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. Ed. by Arindam Banerjee and Kenji Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, 13–15 Apr 2021, pp. 1855–1863. URL: https://proceedings.mlr.press/v130/chowdhury21b.html.

[CGW10] Patrick Cattiaux, Arnaud Guillin and Li-Ming Wu. 'A note on Talagrand's transportation inequality and logarithmic Sobolev inequality'. In: *Probability theory and related fields* 148 (2010), pp. 285–304.

[CL23] Djalil Chafaï and Joseph Lehec. 'Logarithmic Sobolev Inequalities Essentials'. 2023. URL: https://djalil.chafai.net/docs/M2/chafai-lehec-m2-lsie-lecture-notes.pdf.

[Dai+22] Zhongxiang Dai et al. *Sample-Then-Optimize Batch Neural Thompson Sampling*. 2022. arXiv: 2210.06850 [cs.LG].

[DV20] Vikranth Dwaracherla and Benjamin Van Roy. 'Langevin dqn'. In: *arXiv preprint arXiv:2002.07282* (2020).

[FM21] Ying Fan and Yifei Ming. 'Model-based Reinforcement Learning for Continuous Control with Posterior Sampling'. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 3078–3087. URL: http://proceedings.mlr.press/v139/fan21b.html.

[Hua+23] Ziyi Huang et al. '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.

[Ish+23] Haque Ishfaq et al. *Provable and Practical: Efficient Exploration in Reinforcement Learning via Langevin Monte Carlo*. 2023. arXiv: 2305.18246 [cs.LG].

[Kal60] Rudolph Emil Kalman. 'A new approach to linear filtering and prediction problems'. In: (1960).

[Kar+23] Amin Karbasi et al. *Langevin Thompson Sampling with Logarithmic Communication: Bandits and Reinforcement Learning*. 2023. arXiv: 2306.08803 [cs.LG].

[KHR23] Frederic Koehler, Alexander Heckett and Andrej Risteski. '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.

[Kim23] Gihun Kim. 'Learning Linear-Quadratic Regulators via Thompson Sampling with Preconditioned Langevin Dynamics'. In: (2023).

[KS22] Yuri Kinoshita and Taiji Suzuki. 'Improved Convergence Rate of Stochastic Gradient Langevin Dynamics with Variance Reduction and its Application to Optimization'. In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: https://openreview.net/forum?id=Sj2z__i1wX-.

[Kua+23] Nikki Lijing Kuang et al. '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.

[KV24] Frederic Koehler and Thuy-Duong Vuong. '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.

[KY21] Vikram Krishnamurthy and George Yin. 'Langevin Dynamics for Adaptive Inverse Reinforcement Learning of Stochastic Gradient Algorithms'. In: *Journal of Machine Learning Research* 22.121 (2021), pp. 1–49. URL: http://jmlr.org/papers/v22/20-625.html.

[Led06] Michel Ledoux. 'Concentration of measure and logarithmic Sobolev inequalities'. In: *Seminaire de probabilites XXXIII*. Springer, 2006, pp. 120–216.

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[Maz+20] Eric Mazumdar et al. 'On Approximate Thompson Sampling with Langevin Algorithms'. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 6797–6807.

[Min00] Thomas Minka. *Bayesian linear regression*. Tech. rep. Citeseer, 2000.

[Ngu+24] Thanh Nguyen-Tang et al. *Posterior Sampling via Langevin Monte Carlo for Offline Reinforcement Learning*. 2024. URL: https://openreview.net/forum?id=WwCirclMvl.

[OBM22] Reda Ouhamma, Debabrota Basu and Odalric-Ambrym Maillard. 'Bilinear exponential family of mdps: Frequentist regret bound with tractable exploration and planning'. In: *arXiv preprint arXiv:2210.02087* (2022).

[ORV13] Ian Osband, Daniel Russo and Benjamin Van Roy. '(More) efficient reinforcement learning via posterior sampling'. In: *Advances in Neural Information Processing Systems* 26 (2013).

[Osb+23] Ian Osband et al. 'Approximate Thompson Sampling via Epistemic Neural Networks'. In: *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. Ed. by Robin J. Evans and Ilya Shpitser. Vol. 216. Proceedings of Machine Learning Research. PMLR, 31 Jul–04 Aug 2023, pp. 1586–1595. URL: https://proceedings.mlr.press/v216/osband23a.html.

[OV17] Ian Osband and Benjamin Van Roy. 'Why is posterior sampling better than optimism for reinforcement learning?' In: *International conference on machine learning*. PMLR. 2017, pp. 2701–2710.

[PAD19] My Phan, Yasin Abbasi Yadkori and Justin Domke. 'Thompson Sampling and Approximate Inference'. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/f3507289cfdc8c9ae93f4098111a13f9-Paper.pdf.

[Rus+20] Daniel Russo et al. *A Tutorial on Thompson Sampling*. 2020. arXiv: 1707.02038 [cs.LG].

[Sch19] André Schlichting. 'Poincaré and log–sobolev inequalities for mixtures'. In: *Entropy* 21.1 (2019), p. 89.

[SCR23] Remo Sasso, Michelangelo Conserva and Paulo Rauber. 'Posterior Sampling for Deep Reinforcement Learning'. In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, 23–29 Jul 2023, pp. 30042–30061. URL: https://proceedings.mlr.press/v202/sasso23a.html.

[Ste21] Clément Steiner. *A Feynman-Kac approach for Logarithmic Sobolev Inequalities*. 2021. arXiv: 2002.01167 [math.FA].

[Str00] Malcolm Strens. 'A Bayesian framework for reinforcement learning'. In: *ICML 2000*. 2000, pp. 943–950.

[Tho33] W.R. Thompson. 'On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of two Samples'. In: *Biometrika* 25.3-4 (1933), pp. 285–294.

[VW19] Santosh Vempala and Andre Wibisono. 'Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices'. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/65a99bb7a3115fdede20da98b08a370f-Paper.pdf.

[Wan01] Feng-Yu Wang. 'Logarithmic Sobolev inequalities: conditions and counterexamples'. In: *Journal of Operator Theory* (2001), pp. 183–197.

[WCM23] Chaoqi Wang, Yuxin Chen and Kevin Patrick Murphy. 'Model-based Policy Optimization under Approximate Bayesian Inference'. In: *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*. 2023.

[Xu+22] Pan Xu et al. 'Langevin Monte Carlo for Contextual Bandits'. In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, 17–23 Jul 2022, pp. 24830–24850. URL: https://proceedings.mlr.press/v162/xu22p.html.

[Yam+23] Kakei Yamamoto et al. 'Mean Field Langevin Actor-Critic: Faster Convergence and Global Optimality beyond Lazy Learning'. In: (2023).

[Zhe+24] Haoyang Zheng et al. *Accelerating Approximate Thompson Sampling with Underdamped Langevin Monte Carlo*. 2024. arXiv: 2401.11665 [stat.ML].

A Algorithms

For completeness we include the SARAH-LD[KS22] and PSRL[ORV13] algorithms in Algorithm 3 and Algorithm 4 as well as a theorem on the gradient complexity of SARAH-LD in Theorem 7.

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 7 (Corollary 2.1 of [KS22])**.** *Under Assumptions 1 and 2, 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 Poincaré inequality

Assumption 3. *The probability distribution ν satisfies the Poincaré inequality with constant ϱ if for all smooth functions $g : \mathbb{R}^n \to \mathbb{R}$,*

$$Var_\nu(g) \leq \frac{1}{\varrho}\,\mathbb{E}_\nu\left[\|\nabla g\|^2\right] \tag{8}$$

C Proof on Bayes regret

Theorem 2 (Bayes regret of PSRL under log-Sobolev posteriors)**.** *If the posteriors $\mathbb{P}(M_l \mid \mathcal{H}_l)$ fulfill 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 $|\bar{R}_M(s)| \leq B_R \forall s$. We then obtain a PSRL Bayesian regret*

$$\mathbb{E}[\textit{Regret}(T)] \leq 2H\left(L_{\bar{R}}\sqrt{\log(B_RTH)}\sum_{l=1}^{\tau}\frac{1}{\sqrt{\alpha_{r,l}}} + L_{M_*}L_{\bar{\mathcal{T}}}\sqrt{d\log(B_RTH)}\sum_{l=1}^{\tau}\frac{1}{\sqrt{\alpha_{p,l}}}\right) + 8 \tag{3}$$

Proof. 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 $\mathbb{E}[V_{\pi_l,1}^{M_l}(s_{l,1}) \mid \mathcal{H}_{l-1}] = \mathbb{E}[V_{\pi_*,1}^{M_*}(s_{l,1}) \mid \mathcal{H}_{l-1}]$. Marginalising we obtain:

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

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

Next, we use Lemma 7 and observation after eq 50 from [CG19] and obtain

$$\sum_{l=1}^{\tau} \mathbb{E}[V_{\pi_l,1}^{M_l}(s_{l,1}) - V_{\pi_l,1}^{M_*}(s_{l,1})] \leq \mathbb{E}[\sum_{l=1}^{\tau}\sum_{h=1}^{H} \left[|\bar{R}_{M_l}(z_{l,h}) - \bar{R}_*(z_{l,h})| + L_{M_l}||\bar{\mathcal{T}}_{M_l}(z_{l,h}) - \bar{\mathcal{T}}_*(z_{l,h})||_2\right]] \quad (11)$$

where L_{M_l} is the Lipschitz constant for the mean of the transition kernel of M_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

$$C_{R,l} = \left\{ f : Z \to \mathbb{R} \;\middle|\; |f(z) - E_{P(\theta|D_l)}[\bar{R}(\theta)]| \leq \sqrt{\frac{L_{\bar{R}}^2 \log 1/\delta}{\alpha_{r,l}}} \right\} \quad (12)$$

$$C_{\mathcal{T},l} = \left\{ f : Z \to \mathbb{R}^d \;\middle|\; \|f(z) - E_{P(\theta|D_l)}[\bar{\mathcal{T}}(\theta)]\|_2 \leq \sqrt{\frac{dL_{\bar{\mathcal{T}}}^2 \log 1/\delta}{\alpha_{p,l}}} \right\} \quad (13)$$

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 $\mathbb{P}(E^c) \leq \mathbb{P}(E_M^c) + \mathbb{P}(E_*^c) \leq 4\tau H\delta$.

Combining the results we then get

$$\mathbb{E}[\sum_{l=1}^{\tau}\sum_{h=1}^{H} \left[|\bar{R}_{M_l}(z_{l,h}) - \bar{R}_*(z_{l,h})| \mid E\right] + \mathbb{E}\left[L_{M_l}||\bar{\mathcal{T}}_{M_l}(z_{l,h}) - \bar{\mathcal{T}}_*(z_{l,h})||_2 \middle| E\right] + 2B_R 4\tau H\delta \quad (14)$$

$$\leq 2H\left(L_{\bar{R}}\sqrt{\log 1/\delta}\sum_{l=1}^{\tau}\frac{1}{\sqrt{\alpha_{r,l}}} + L_{M_*}L_{\bar{\mathcal{T}}}\sqrt{d\log 1/\delta}\sum_{l=1}^{\tau}\frac{1}{\sqrt{\alpha_{p,l}}}\right) + 8B_R\tau H\delta \quad (15)$$

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

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

■

D Proofs on regret for approximate sampling and sample complexity.

Theorem 3. *Let a policy the start of episode l plan according to a posterior Q_l where $\min(KL((P_l \parallel Q_l), KL(P_l \parallel Q_l)) \leq \epsilon_{post,l}$ 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'.

$$E_{P_l,Q_l}[V_{\pi^*}^{\mu^*} - V_{\pi_l'}^{\mu^*}] = \int_{\mu^*}\int_{\mu_l'}(V_{\pi^*}^{\mu^*} - V_{\pi_l'(\mu_l')}^{\mu^*})P_l(\mu^*)Q_l(\mu_l') \tag{17}$$

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

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

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

$$\leq E_{P_l}[V_{\pi^*}^{\mu^*} - V_{\pi_l}^{\mu_l}] + \Delta_{\max}\sqrt{\frac{\epsilon_{\text{post},l}}{2}} + E_{P_l}[V_{\pi_l}^{\mu_l} - V_{\pi_l}^{\mu^*}] + \Delta_{\max}\sqrt{\frac{\epsilon_{\text{post},l}}{2}}] \tag{21}$$

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

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

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

$$\widetilde{\mathcal{O}}(\mathbb{E}_P\, R(\pi_Q)) \leq \widetilde{\mathcal{O}}(\sqrt{\tau}g(H,S,A,D)) + \sqrt{2}\Delta_{\max}\sum_{k=1}^{\tau}\sqrt{\epsilon_{\text{post},l}} \tag{23}$$

$$\leq \widetilde{\mathcal{O}}(\sqrt{\tau}g(H,S,A,D)) + \sqrt{2}\Delta_{\max}\sum_{k=1}^{\tau}\sqrt{C}\frac{g(H,S,A,D)}{\sqrt{t}\Delta_{\max}} \tag{24}$$

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

$$= \widetilde{\mathcal{O}}(\sqrt{\tau}g(H,S,A,D)) \tag{26}$$

■

Theorem 4. *Let $\rho_*^l(\theta)$ be the distribution final sample (i.e. after k steps) 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 $P(\theta \mid D_{l+1}) = P(\theta \mid D_l, z_l)$. Note that $P(z_l \mid D_l, \theta) = P(z_l \mid \theta)$ and $\mathbb{E}_\theta\, P(z_l \mid D_l, \theta) = P(z_l \mid D_l)$.

$$KL(\rho_*^l \parallel \nu_{l+1} \mid z_l) \tag{27}$$

$$= \int_{\rho_*^l} \rho_*^l(\theta)\log\left(\frac{\rho_*^l(\theta)}{\mathbb{P}(\theta \mid D_l, z_l)}\right)d\theta \tag{28}$$

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

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

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

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

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

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

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

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

$$\leq \epsilon_{\text{post},l} + \int_{z_l} \int_\theta \log \left(\frac{P(z_l|D_l)}{P(z_l|\theta)} \right) \rho_*^l(\theta) d\theta P(z_l \mid D_l) dz_l \tag{37}$$

$$= \epsilon_{\text{post},l} + \int_\theta \int_{z_l} \log \left(\frac{P(z_l|D_l)}{P(z_l|\theta)} \right) \rho_*^l(\theta) P(z_l \mid D_l) dz_l d\theta \tag{38}$$

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

$$\leq \epsilon_{\text{post},l} + \int_\theta 2/\alpha_z \int_{z_l} ||\nabla_z \sqrt{\frac{P(z_l|D_l)}{P(z_l|\theta)}}||^2 P(z_l|\theta) dz_l \rho_*^l(\theta) d\theta \tag{40}$$

$$= \epsilon_{\text{post},l} + \int_\theta 2/\alpha_z \int_{z_l} || \frac{P(z_l|\theta) \nabla_z P(z_l|D_l) - P(z_l|D_l) \nabla_z P(z_l|\theta)}{2\sqrt{\frac{P(z_l|D_l)}{P(z_l|\theta)}} P(z_l|\theta)^2} ||^2 P(z_l|\theta) dz_l \rho_*^l(\theta) d\theta \tag{41}$$

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

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

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

$$= \epsilon_{\text{post},l} + \int_\theta \frac{1}{2\alpha_z} \int_{z_l} ||\nabla_z \log P(z_l|D_l) - \nabla_z \log P(z_l|\theta)||^2 P(z_l|D_l) dz_l \rho_*^l(\theta) d\theta \tag{45}$$

$$\leq \epsilon_{\text{post},l} + \frac{L_z}{2\alpha_z} \int_\theta ||\theta_l - \theta||^2 \rho_*^l(\theta) d\theta \tag{46}$$

$$= \epsilon_{\text{post},l} + \frac{L_z}{2\alpha_z} \text{Var}_{\rho_*^l(\theta)}(\theta) \tag{47}$$

■

E Proofs for theorems on LSI constants.

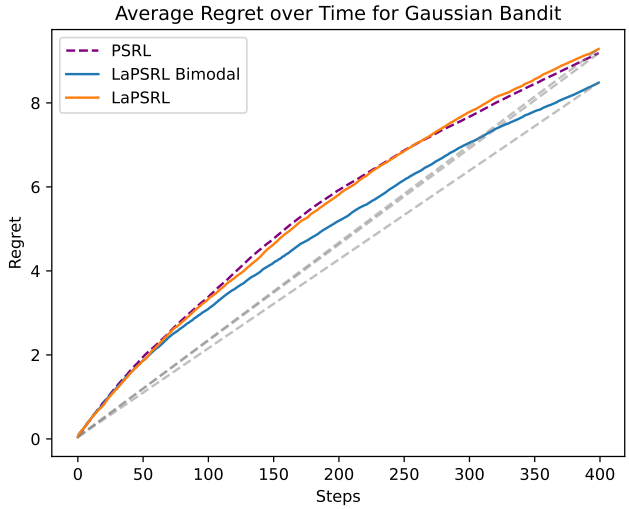
Theorem 6. *Any log-concave posterior will have $\alpha_l = \Theta(n)$. Similarly, for any posterior that is a mixture of log-concave distributions will have $\alpha_{\textit{Mixture}} = \Omega\left(\frac{n \min p_i}{4k(1-\log(\min p_i))}\right)$*

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 $P(x_i \mid \theta)$. Since products preserve log-concavity, we can use Theorem 1.

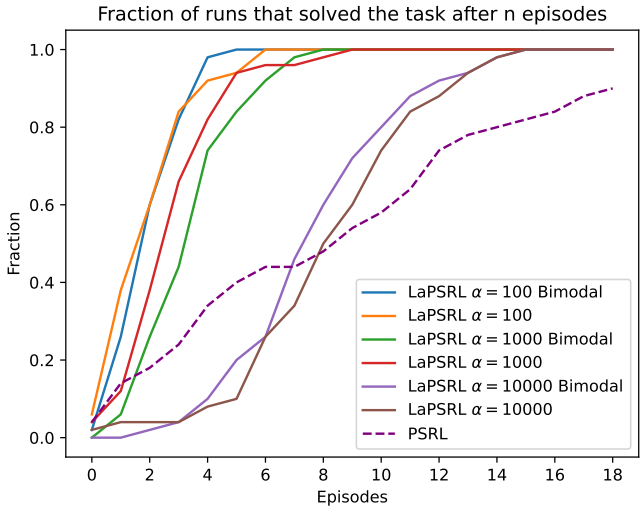
From Weyl's inequality, we have that the smallest eigenvalue a sum of two Hermitian is larger than the sum of the smallest eigenvalues of the two matrices. Since the Hessian is a Hermitian matrix, putting this into Theorem 1 this gives that $\alpha_l \geq \alpha_{P(\theta)} + \sum_{i=1}^{n} \alpha_i \geq \alpha_{P(\theta)} + n \min_i \alpha_i$. 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 5 that $\alpha_{\text{Mixture}} = \Omega\left(\frac{\min \alpha_i \min p_i}{4k(1-\log(\min p_i)}\right)$. Setting $\min_i \alpha_i = \Theta(n)$ completes the proof. ■

F Experimental 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.