Thompson Sampling with Approximate Inference

My Phan
College of Information and Computer Science
University of Massachusetts
Amherst, MA
myphan@cs.umass.edu

Yasin Abbasi-Yadkori
VinAI
Hanoi, Vietnam
yasin.abbasi@gmail.com

Justin Domke
College of Information and Computer Science
University of Massachusetts
Amherst, MA
domke@cs.umass.edu

Abstract

We study the effects of approximate inference on the performance of Thompson sampling in the k-armed bandit problems. Thompson sampling is a successful algorithm for online decision-making but requires posterior inference, which often must be approximated in practice. We show that even small constant inference error (in α-divergence) can lead to poor performance (linear regret) due to under-exploration (for $\alpha < 1$) or over-exploration (for $\alpha > 0$) by the approximation. While for $\alpha > 0$ this is unavoidable, for $\alpha \leq 0$ the regret can be improved by adding a small amount of forced exploration even when the inference error is a large constant.

1 Introduction

The stochastic k-armed bandit problem is a sequential decision making problem where at each time-step t, a learning agent chooses an action (arm) among k possible actions and observes a random reward. Thompson sampling (Russo et al., 2018) is a popular approach in bandit problems based on sampling from a posterior in each round. It has been shown to have good performance both in term of frequentist regret and Bayesian regret for the k-armed bandit problem under certain conditions.

This paper investigates Thompson sampling when only an *approximate* posterior is available. This is motivated by the fact that in complex models, approximate inference methods such as Markov Chain Monte Carlo or Variational Inference must be used. Along this line, Lu & Van Roy (2017) propose a novel inference method – Ensemble sampling – and analyze its regret for linear contextual bandits. To the best of our knowledge this is the most closely related theoretical analysis of Thompson sampling with approximate inference.

This paper analyzes the regret of Thompson sampling with approximate inference. Rather than considering a particular inference algorithm, we parameterize the error using the α-divergence, a typical measure of inference accuracy. Our contributions are as follows:

- **Even small inference errors can lead to linear regret with naive Thompson sampling.** Given any error threshold $\epsilon > 0$ and any α we show that approximate posteriors with error at most ϵ in α-divergence at all times can result in linear regret (both frequentist and Bayesian). For $\alpha > 0$ and for any reasonable prior, we show linear regret due to over-exploration by the approximation (Theorem 1, Corrolary 1). For $\alpha < 1$ and for priors satisfying certain

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

conditions, we show linear regret due to under-exploration by the approximation, which prevents the posterior from concentrating (Theorem 2, Corrolary 2).

- **Forced exploration can restore sub-linear regret.** For $\alpha \leq 0$ we show that adding forced exploration to Thompson sampling can make the posterior concentrate and restore sub-linear regret (Theorem 3) even when the error threshold is a very large constant. We illustrate this effect by showing that the performances of Ensemble sampling (Lu & Van Roy, 2017) and mean-field Variation Inference (Blei et al., 2017) can be improved in this way either theoretically (Section 5.1) or in simulations (Section 6).

2 Background and Notations.

2.1 The k-armed Bandit Problem.

We consider the k-armed bandit problem parameterized by the mean reward vector $m = (m_1, ..., m_k) \in \mathcal{R}^k$, where m_i^* denotes the mean reward of arm (action) i. At each round t, the learner chooses an action A_t and observes the outcome Y_t which, conditioned on A_t, is independent of the history up to and not including time t, $H_{t-1} = (A_1, Y_1, ..., A_{t-1}, Y_{t-1})$. For a time horizon T, the goal of the algorithm π is to maximize the expected cumulative reward up to time T.

Let $\Omega \subseteq \mathcal{R}^k$ be the domain of the mean and $\Omega_i \subseteq \Omega$ denote the region where the ith arm has the largest mean. Let the function $A^* : \Omega \to \{a_1, ..., a_k\}$ denoting the best action be defined as: $A^*(m) = i$ if $m \in \Omega_i$.

In the frequentist setting we assume that there exists a true mean m^* which is fixed and unknown to the learner. Therefore, a policy π^* that always chooses $A^*(m^*)$ will get the highest reward. The performance of policy π is measured by its expected regret compared to an optimal policy π^*, which is defined as:

$$\text{Regret}(T, \pi, m^*) = T m^*_{A^*(m^*)} - \mathbb{E} \sum_{t=1}^{T} m^*_{A_t} \ . \tag{1}$$

On the other hand, in the Bayesian setting, an agent expresses her beliefs about the mean vector in terms of a prior Π_0, and therefore, the mean is treated as a random variable $M = (M_1, ..., M_k)$ distributed according to the prior Π_0. The Bayesian regret is the expectation of the regret under the prior of parameter M:

$$\text{BayesRegret}(T, \pi) = \mathbb{E}_{\Pi_0} \text{Regret}(T, \pi, M) \ . \tag{2}$$

2.2 Thompson Sampling with Approximate Inference

In the frequentist setting, in order to perform Thompson sampling we define a prior which is only used in the algorithm. On the other hand, in the Bayesian setting the prior is given.

Let Π_t be the posterior distribution of $M|H_{t-1}$ with density function $\pi_t(m)$. Thompson sampling obtains a sample $\widehat{m}$ from Π_t and then selects arm A_t as follow: $A_t = i$ if $\widehat{m} \in \Omega_i$. In each round, we assume an approximate sampling method is available that generates sample from an approximate distribution Q_t. We use q_t to denote the density function of Q_t.

Popular approximate sampling methods include Markov Chain Monte Carlo (MCMC) (Andrieu et al., 2003), Sequential Monte Carlo (Doucet & Johansen, 2009) and Variational Inference (VI) (Blei et al., 2017). There are packages that conveniently implement VI and MCMC methods, such as Stan (Carpenter et al., 2017), Edward (Tran et al., 2016), PyMC (Salvatier et al., 2016) and infer.NET (Minka et al., 2018).

To provide a general analysis of approximate sampling methods, we will use the α-divergence (Section 2.3) to quantify the distance between the posterior Π_t and the approximation Q_t.

2.3 The Alpha Divergence

The α-divergence between two distributions P and Q with density functions $p(x)$ and $q(x)$ is defined as:

$$D_\alpha(P,Q) = \frac{1 - \int p(x)^\alpha q(x)^{1-\alpha} dx}{\alpha(1-\alpha)}. \tag{3}$$

α-divergence generalizes many divergences, including $KL(Q,P)$ ($\alpha \to 0$), $KL(P,Q)$ ($\alpha \to 1$), Hellinger distance ($\alpha = 0.5$) and χ^2 divergence ($\alpha = 2$) and is a common way to measure errors in inference methods. MCMC errors are measured by the Total Variation distance, which can be upper bounded by the KL divergence using Pinsker's inequality ($\alpha = 0$ or $\alpha = 1$). Variational Inference tries to minimize the reverse KL divergence (information projection) between the target distribution and the approximation ($\alpha = 0$). Ensemble sampling (Lu & Van Roy, 2017) provides error guarantees using reverse KL divergence ($\alpha = 0$). Expectation Propagation tries to minimize the KL divergence ($\alpha = 1$) and χ^2 Variational Inference tries to minimize the χ^2 divergence ($\alpha = 2$).

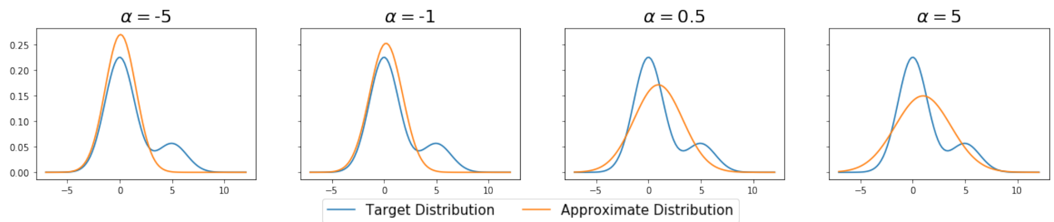


Figure 1: The Gaussian Q which minimizes $D_\alpha(P,Q)$ for different values of α where the target distribution P is a mixture of two Gaussians. Based on Figure 1 from (Minka, 2005)

When α is small, the approximation fits the posterior's dominant mode. When α is large, the approximation covers the posterior's entire support (Minka, 2005) as illustrated in Figure 1. Therefore changing α will affect the exploration-exploitation trade-off in bandit problems.

2.4 Problem Statement.

Problem Statement. For the k-armed bandit problem, given α and $\epsilon > 0$, if at all time-steps t we sample from an approximate distribution Q_t such that $D_\alpha(\Pi_t, Q_t) < \epsilon$, will the regret be sub-linear in t?

3 Motivating Example

In this section we present a simple example to show the effects of inference errors on the frequentist regret.

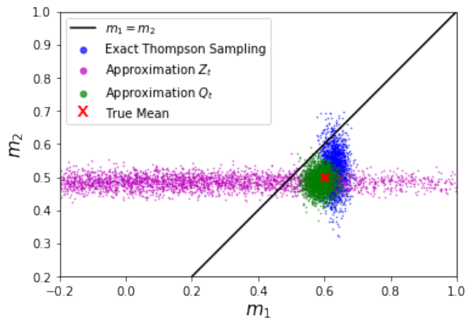
Example. Consider a 2-armed bandit problem where the reward distributions are $\mathrm{Norm}(0.6, 0.2^2)$ and $\mathrm{Norm}(0.5, 0.2^2)$ for arm 1 and 2 respectively. The prior Π_0 is $\mathrm{Norm}\left(\mu_0^T, 0.5^2 I\right)$ where $\mu_0 = [0.1, 0.9]$ is the vector of prior means of arm 1 and 2 respectively, and I denotes the identity matrix.

Let $\Pi_t = \mathrm{Norm}(\mu_t, \Sigma_t)$ be the posterior at time t. Approximations Q_t and Z_t are calculated such that $\mathrm{KL}(\Pi_t, Q_t) = 2$ and $\mathrm{KL}(Z_t, \Pi_t) = 1.5$ by multiplying the covariance Σ_t by a constant: $Q_t = \mathrm{Norm}(\mu_t, 4.5^2\Sigma_t)$ and $Z_t = \mathrm{Norm}(\mu_t, 0.3^2\Sigma_t)$. The KL divergence between two Gaussian distributions is provided in Appendix F.

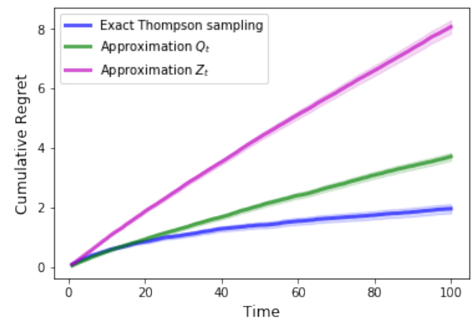
We perform the following simulations 1000 times and plot the mean cumulative regret up to time $T = 100$ in Figure 2b using three different policies:

1. (**Exact Thompson Sampling**) At each time-step t, sample from the true posterior Π_t.
2. (**Approximation** Q_t) At each time-step t, compute Q_t from Π_t and sample from Q_t.
3. (**Approximation** Z_t) At each time-step t, compute Z_t from Π_t and sample from Z_t.

The regrets of sampling from the approximations Q_t and Z_t are in both cases larger than that of exact Thompson sampling. Intuitively, the regret of Q_t is larger because Q_t explores more than the true



(a) Over-dispersed (approximation Q_t) and under-dispersed sampling (approximation Z_t) yield different posteriors after $T = 100$ time-steps. m_1 and m_2 are the means of arms 1 and 2. Q_t picks arm 2 more often than exact Thompson sampling and Z_t mostly picks arm 2. The posteriors of exact Thompson sampling and Q_t concentrate mostly in the region where $m_1 > m_2$ while Z_t's spans both regions.



(b) The regret of sampling from the approximations Q_t and Z_t are both larger than that of exact Thompson sampling from the true posterior Π_t. Shaded regions show 95% confidence intervals.

Figure 2: Approximation Q_t (with high variance) and approximation Z_t (with small variance) are defined in Section 3 where $D_1(\Pi_t, Q_t) = 2$ and $D_0(\Pi_t, Z_t) = 1.5$. Arm 1 is the true best arm.

posterior (Figure 2a). In Section 4 we show that when $\alpha > 0$ the approximation can incur this type of error, leading to linear regret. On the other hand, the regret of Z_t is larger because Z_t explores less than the exact Thompson sampling algorithm and therefore commits to the sub-optimal arm (Figure 2a). In Section 5 we show that when $\alpha < 1$ the approximation can change the posterior concentration rate, leading to linear regret. We also show that adding a uniform sampling step can help the posterior to concentrate when $\alpha \leq 0$, and make the regret sub-linear.

4 Regret Analysis When $\alpha > 0$

In this section we analyze the regret when $\alpha > 0$. Our result shows that the approximate method might pick the sub-optimal arm with constant probability in every time-step, leading to linear regret.

Theorem 1 (Frequentist Regret)**.** *Let $\alpha > 0$, the number of arms be $k = 2$ and $m_1^* > m_2^*$. Let Π_0 be a prior where $\mathbb{P}_{\Pi_0}(M_2 > M_1) > 0$. For any error threshold $\epsilon > 0$, there is a deterministic mapping $f(\Pi)$ such that for all $t \geq 0$:*

1. *Sampling from $Q_t = f(\Pi_t)$ chooses arm 2 with a constant probability.*
2. *$D_\alpha(\Pi_t, Q_t) < \epsilon$.*

Therefore sampling from Q_t for $T/10$ time-steps and using any policy for the remaining time-steps will cause linear frequentist regret.

Typically, approximate inference methods minimize divergences. Broadly speaking, this theorem shows that making a divergence a small constant, alone, is not enough to guarantee sub-linear regret. We do not mean to imply that low regret is *impossible* but simply that making an α-divergence a small constant alone is not sufficient.

At every time-step, the mapping f constructs the approximation Q_t from the posterior Π_t by moving probability mass from the region Ω_1 where $m_1 > m_2$ to the region Ω_2 where $m_2 > m_1$. Then Q_t will choose arm 2 with a constant probability at every time-step. The constant average regret per time-step is discussed in Appendix A.4.

Therefore, if we sample from $Q_t = f(\Pi_t)$ for $0.1T$ time steps and use any policy in the remaining $0.9T$ time steps, we will still incur linear regret from the $0.1T$ time-steps. On the other hand, when $\alpha \leq 0$, we show in Section 5.1 that sampling an arm uniformly at random for $\log T$ time-steps and sampling from an approximate distribution that satisfies the divergence constraint for $T - \log T$ time-steps will result in sub-linear regret.

Agrawal & Goyal (2013) show that the frequentist regret of exact Thompson sampling is $O(\sqrt{T})$ with Gaussian or Beta priors and bounded rewards. Theorem 1 implies that when the assumptions in (Agrawal & Goyal, 2013) are satisfied but there is a small constant inference error at every time-step, the regret is no longer guaranteed to be sub-linear.

If the assumption $m_1^* > m_2^*$ in Theorem 1 is satisfied with a non-zero probability ($\mathbb{P}_{\Pi_0}(M_1 > M_2) > 0$), the Bayesian regret will also be linear:

Corollary 1 (Bayesian Regret)**.** *Let $\alpha > 0$ and the number of arms be $k = 2$. Let Π_0 be a prior where $\mathbb{P}_{\Pi_0}(M_1 > M_2) > 0$ and $\mathbb{P}_{\Pi_0}(M_2 > M_1) > 0$. Then for any error threshold $\epsilon > 0$, there is a deterministic mapping $f(\Pi)$ such that for all $t \geq 0$ the two statements in Theorem 1 hold.*

Therefore sampling from Q_t for $T/10$ time-steps and using any policy for the remaining time-steps will cause linear Bayesian regret.

Russo & Roy (2016) prove that the Bayesian regret of Thompson sampling for k-armed bandits with sub-Gaussian rewards is $O(\sqrt{T})$. Corollary 1 implies that even when the assumptions in Russo & Roy (2016) are satisfied, under certain conditions and with approximation errors, the regret is no longer guaranteed to be sub-linear.

5 Regret Analysis When $\alpha < 1$

In this section we analyze the regret when $\alpha < 1$. Our result shows that for any error threshold, if the posterior Π_t places too much probability mass on the wrong arm then the approximation Q_t is allowed to avoid the optimal arm. If the sub-optimal arms do not provide information about the arms' ranking, the posterior Π_{t+1} does not concentrate. Therefore Q_{t+1} is also allowed to be close in α-divergence while avoiding the optimal arm, leading to linear regret in the long term.

Theorem 2 (Frequentist Regret)**.** *Let $\alpha < 1$, the number of arms be $k = 2$ and $m_1^* > m_2^*$. Let Π_0 be a prior where M_2 and $M_1 - M_2$ are independent. There is a deterministic mapping $f(\Pi)$ such that for all $t \geq 0$:*

1. *Sampling from $Q_t = f(\Pi_t)$ chooses arm 2 with probability 1.*
2. *For any $\epsilon > 0$, there exists $0 < z \leq 1$ such that if $\mathbb{P}_{\Pi_0}(M_2 > M_1) = z$ and arm 2 is chosen at all times before t then $D_\alpha(\Pi_t, Q_t) < \epsilon$.*
   For any $0 < z \leq 1$, there exists $\epsilon > 0$ such that if $\mathbb{P}_{\Pi_0}(M_2 > M_1) = z$ and arm 2 is chosen at all times before t then $D_\alpha(\Pi_t, Q_t) < \epsilon$.

Therefore sampling from Q_t at all time-steps results in linear frequentist regret.

We discuss why the above results are not immediately obvious. When $\alpha \to 0$, the α-divergence becomes $\mathrm{KL}(Q_t, \Pi_t)$. We might believe that the regret should be sub-linear in this case because the posterior Π_t becomes more concentrated, and so the total variation between Q_t and Π_t must decrease. For example, Ordentlich & Weinberger (2004) show the distribution-dependent Pinsker's inequality between $\mathrm{KL}(Q, P)$ and the total variation $\mathrm{TV}(P, Q)$ for discrete distributions P and Q as follows:

$$\mathrm{KL}(Q, P) \geq \phi(P) \cdot \mathrm{TV}(P, Q)^2 \ . \tag{4}$$

Here, $\phi(P)$ is a quantity that will increase to infinity if P becomes more concentrated. However, the algorithm in Theorem 2 constructs an approximation distribution that never picks the optimal arm, so the posterior Π_t can not concentrate and the regret is linear. The error threshold ϵ causing linear frequentist regret is correlated with the probability mass the prior places on the true best arm (Appendix B.4).

With some assumptions on the rewards, Gopalan et al. (2014) show that the problem-dependent frequentist regret is $O(\log T)$ for finitely-supported, correlated priors with $\pi_0(m^*) > 0$. Liu & Li (2016) study the prior-dependent frequentist regret of 2-armed-and-2-models bandits, and show that with some smoothness assumptions on the reward likelihoods, the regret is $O(\sqrt{T/\mathbb{P}_{\Pi_0}(M_2 > M_1)})$ if arm 1 is the better arm. Theorem 2 implies that when the assumptions in (Gopalan et al., 2014) or (Liu & Li, 2016) are satisfied, if M_2 and $M_1 - M_2$ are independent and there are approximation errors, the regret is no longer guaranteed to be sub-linear.

If the assumption $m_1^* > m_2^*$ in Theorem 2 is satisfied with a non-zero probability ($\mathbb{P}_{\Pi_0}(M_1 > M_2) > 0$), the Bayesian regret wil also be linear:

Corollary 2 (Bayesian Regret). *Let $\alpha < 1$ and the number of arms be $k = 2$. Let Π_0 be a prior where $\mathbb{P}_{\Pi_0}(M_1 > M_2) > 0$ and M_2 and $M_1 - M_2$ are independent. There is a deterministic mapping $f(\Pi)$ such that for all $t \geq 0$ the 2 statements in Theorem 2 hold.*

Therefore sampling from Q_t at all time-steps results in linear Bayesian regret.

Russo & Roy (2016) prove that the Bayesian regret of Thompson sampling for k-armed bandits with sub-Gaussian rewards is $O(\sqrt{T})$. Corollary 2 implies that even when the assumptions in Russo & Roy (2016) are satisfied, under certain conditions and with approximation errors, the regret is no longer guaranteed to be sub-linear.

We note that, unlike the case when $\alpha > 0$, if we use another policy in $o(T)$ time-steps to make the posterior concentrate and sample from Q_t for the remaining time-steps, the regret can be sub-linear. We provide a concrete algorithm in Section 5.1 for the case when $\alpha \leq 0$.

5.1 Algorithms with Sub-linear Regret for $\alpha \leq 0$

In the previous section, we see that when $\alpha < 1$, the approximation has linear regret because the posterior does not concentrate. In this section we show that when $\alpha \leq 0$, it is possible to achieve sub-linear regret even when ϵ is a very large constant by adding a simple exploration step to force the posterior to concentrate (the case of $\alpha > 0$ cannot be improved according to Theorem 1). We first look at the necessary and sufficient condition that will make the posterior concentrate, and then provide an algorithm that satisfies it. Russo (2016) and Qin et al. (2017) both show the following result under different assumptions:

Lemma 1 (Lemma 14 from Russo (2016)). *Let $m^* \in \mathcal{R}^k$ be the true parameter and let $a^* = A^*(m^*)$ be the true best arm. If for all arms i, $\sum_{t=1}^{\infty} P(A_t = i|H_{t-1}) = \infty$, then*

$$\lim_{t\to\infty} P(A^*(M) = a^*|H_{t-1}) = 1 \textit{ with probability } 1 \ . \tag{5}$$

If there exists arm i such that $\sum_{t=1}^{\infty} P(A_t = i|H_{t-1}) < \infty$, then $\liminf_{t\to\infty} P(A^*(M) = i|H_{t-1}) > 0$ with probability 1.

Russo (2016) make the following assumptions, which allow correlated priors:

Assumption 1. *Let the reward distributions be in the canonical one dimensional exponential family with the density: $p(y|m) = b(y)\exp(mT(y) - A(m))$ where b, T and A are known function and $A(m)$ is assumed to be twice differentiable. The parameter space $\Omega = (\overline{m}, \underline{m})$ is a bounded open hyper-rectangle, the prior density is uniformly bounded with $0 < \inf_{m\in\Omega} \pi_0(m) < \sup_{m\in\Omega} \pi_0(m) < \infty$ and the log-partition function has bounded first derivative with $\sup_{\theta\in[\overline{m},\underline{m}]} |A'(m)| < \infty$.*

Qin et al. (2017) make the following assumptions:

Assumption 2. *Let the prior be an uncorrelated multivariate Gaussian. Let the reward distribution of arm i be $\mathrm{Norm}(m_i, \sigma^2)$ with a common known variance σ^2 but unknown mean m_i.*

Even though we consider the error in sampling from the posterior distribution, the regret is a result of choosing the wrong arm. We define $\overline{\Pi}_t$ as the posterior distribution of the best arm and $\overline{Q}_t$ as the approximation of $\overline{\Pi}_t$ with the density functions

$$\overline{\pi}_t(i) = P(A^* = i|H_{t-1}) \text{ and } \overline{q}_t(i) = P(A_t = i|H_{t-1}).$$

We now define an algorithm where each arm will be chosen infinitely often, satisfying the condition of Lemma 1.

Theorem 3 (Bayesian and Frequentist Regret). *Consider the case when Assumption 1 or 2 is satisfied. Let $\alpha \leq 0$ and $p_t = o(1)$ be such that $\sum_{t=1}^{\infty} p_t = \infty$. For any number of arms k, any prior Π_0 and any error threshold $\epsilon > 0$, the following algorithm has $o(T)$ frequentist regret: at every time-step t,*

- *with probability $1-p_t$, sample from an approximate posterior Q_t such that $D_\alpha(\overline{\Pi}_t, \overline{Q}_t) < \epsilon$,*
- *with probability p_t, sample an arm uniformly at random.*

Since the Bayesian regret is the expectation of the frequentist regret over the prior, for any prior if the frequentist regret is sub-linear at all points the Bayesian regret will be sub-linear.

The following lemma shows that the error in choosing the arms is upper bounded by the error in choosing the parameters. Therefore whenever the condition $D_\alpha(\Pi_t, Q_t) < \epsilon$ is satisfied, the condition $D_\alpha(\overline{\Pi}_t, \overline{Q}_t) < \epsilon$ will be satisfied and Theorem 3 is applicable.

Lemma 2.

$$D_\alpha(\overline{\Pi}_t, \overline{Q}_t) \leq D_\alpha(\Pi_t, Q_t)\,.$$

We also note that we can achieve sub-linear regret even when ϵ is a very large constant. We revisit Eq. 4 to provide the intuition: $\mathrm{KL}(Q, P) \geq \phi(P) \cdot \mathrm{TV}(P, Q)^2$. Here, $\phi(P)$ is a quatity that will increase to infinity if P becomes more concentrated. Hence, if $KL(\overline{Q}_t, \overline{\Pi}_t) < \epsilon$ for any constant ϵ and $\overline{\Pi}_t$ becomes concentrated, the total variation $\mathrm{TV}(\overline{Q}_t, \overline{\Pi}_t)$ will decrease. Therefore, $\overline{Q}_t$ will become concentrated, resulting in sub-linear regret.

Application. Lu & Van Roy (2017) propose an approximate sampling method called Ensemble sampling where they maintain a set of $\mathcal{M}$ models to approximate the posterior and analyze its regret for the linear contextual bandits when $\mathcal{M}$ is $\Omega(\log(T))$. For the k-armed bandit problem and when $\mathcal{M}$ is $\Theta(\log(T))$, Ensemble sampling satisfies the condition $\mathrm{KL}(\overline{Q}_t, \overline{\Pi}_t) < \epsilon$ in Theorem 3 with high probability. In this case, Lu & Van Roy (2017) show a regret bound that scales linearly with T. We discuss in Appendix E how to apply Theorem 3 to get sub-linear regret with Ensemble sampling when $\mathcal{M}$ is $\Theta(\log(T))$.

6 Simulations

For each approximation method we repeat the following simulations for 1000 times and plot the mean cumulative regret, using five different policies.

1. (**Exact Thompson sampling**) Use exact posterior sampling to choose an action and update the posterior (for reference).
2. (**Approximation method**) Use the approximation method to choose an action and update the posterior. We use the approximation naively without any modification.
3. (**Forced Exploration**) With a probability (the exploration rate), choose an action uniformly at random and update the posterior. Otherwise, use the approximation method to choose an action and update the posterior. This is the method suggested by Thm. 3.
4. (**Approximate Sample**) Use the approximation method to choose an action. Use exact posterior sampling to update the posterior.
5. (**Approximate Update**) Use exact posterior sampling to choose an action. Use the approximate method to update the posterior.

The last two policies are performed to understand how the approximation affects the posterior (discussed in Section 6.3). We update the posterior using the closed-form formula when both the prior and reward distribution are Gaussian in Appendix G.

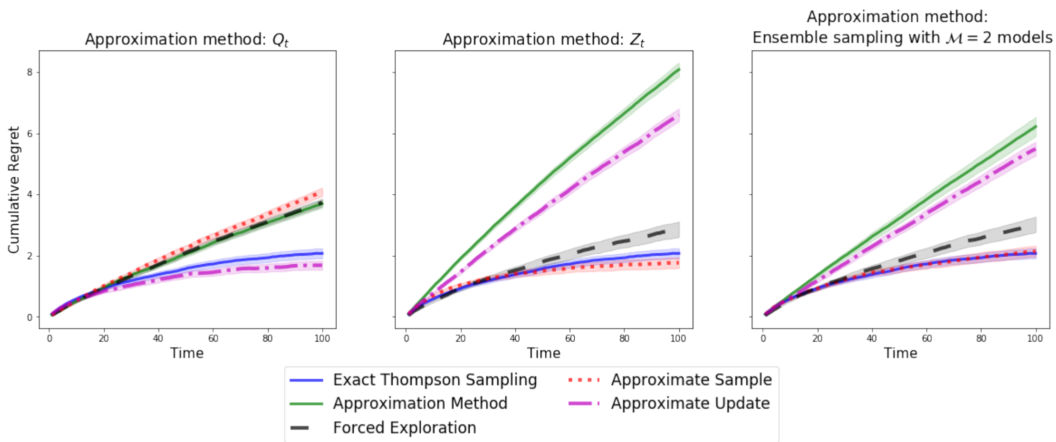
6.1 Adding Forced Exploration to the Motivating Example

In this section we revisit the example in Section 3. We apply Q_t, Z_t and Ensemble sampling with $\mathcal{M} = 2$ models to the bandit problem described in the example. We set the exploration rate at time t to be $1/t$, $T = 100$ and show the results in Figure 3a and discuss them in Section 6.3.

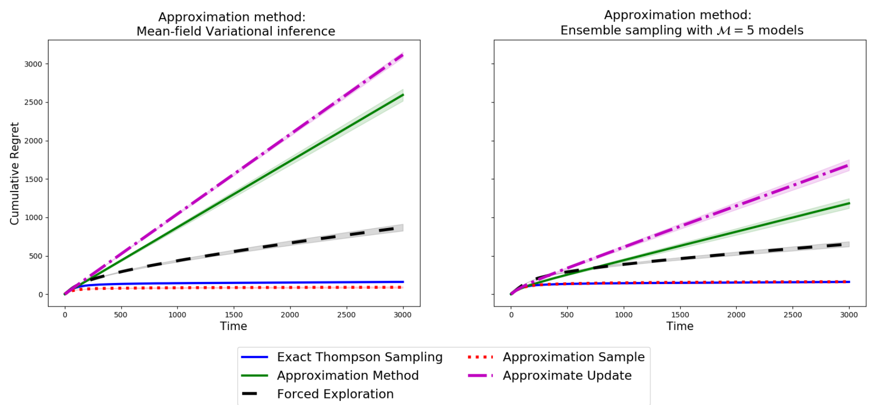
6.2 Simulations of Ensemble Sampling and Variational Inference for 50-armed bandits

Now we add forced exploration to mean-field Variational Inference (VI) and Ensemble Sampling with $\mathcal{M} = 5$ models for a 50-armed bandit instance. We generate the prior and the reward distribution as follows: the prior is $\mathrm{Norm}(\mathbf{0}, \Sigma_0)$. To generate a positive semi-definite matrix Σ_0, we generate a random matrix A of size (k, k) where entries are uniformly sampled from $[0, 1)$ and set $\Sigma_0 = A^T A/k$. The true mean m^* is sampled from the prior. The reward distribution of arm i is $\mathrm{Norm}(m^*_i, 1)$.

Mean-field VI approximates the posterior by finding an uncorrelated multivariate Gaussian distribution Q_t that minimizes $KL(\Pi_t, Q_t)$. If the posterior is $\Pi_t = \mathrm{Norm}(\mu_t, \Sigma_t)$ then Q_t has the closed-form solution $Q_t = \mathrm{Norm}(\mu_t, \mathrm{Diag}(\Sigma_t^{-1})^{-1})$, which we used to perform the simulations. We set the exploration rate at time t to be $50/t$, $T = 3000$, show the results in Figure 3b and discuss them in Section 6.3.



(a) Applying approximations Q_t, Z_t and Ensemble Sampling to the motivating example (Section 6.1).



(b) Applying mean-field Variational Inference (VI) and Ensemble sampling on a 50-armed bandit (Section 6.2).

Figure 3: Updating the posterior by exact Thompson sampling or adding forced exploration does not help the over-explored approximation Q_t, but lowers the regrets of the under-explored approximations Z_t, Ensemble sampling and mean-field VI. Shaded regions show 95% confidence intervals.

6.3 Discussion

We observe in Figure 3a that the regret of Q_t calculated from the posterior updated by exact Thompson sampling does not change significantly. Moreover, exact posterior sampling with the posterior updated by Q_t has the same regret as exact Thompson sampling. These two observations imply that Q_t has the same effect on the posterior as exact Thompson sampling. Therefore adding forced exploration is not helpful.

On the other hand, in Figures 3a and 3b the regrets of Z_t, Ensemble sampling and mean-field VI calculated from the posterior updated by exact Thompson sampling decrease significantly. Moreover, exact posterior sampling with the posterior updated by the approximations has similar regret to using the approximations. This behaviour is likely because the approximation causes the posterior to concentrate in the wrong region[1]. In combination, these two observations suggest that these methods do not explore enough for the posterior to concentrate. Therefore adding forced exploration is helpful, which is compatible with the result in Theorem 3.

[1]Note that in the case where there are 2 arms (Figure 3a), exact posterior sampling with the posterior updated by the approximate method has slightly lower regret than naively using the approximate method. This is only because there are only 2 regions, so exact posterior sampling explores more than the approximation in the other region, which happens to be the correct one.

7 Related Work

There have been many works on sub-linear Bayesian and frequentist regrets for exact Thompson sampling. We discussed relevant works in detail in Section 4 and Section 5.

Ensemble sampling (Lu & Van Roy, 2017) gives a theoretical analysis of Thompson sampling with one particular approximate inference method. Lu & Van Roy (2017) maintain a set of $\mathcal{M}$ models to approximate the posterior, and analyzed its regret for linear contextual bandits when $\mathcal{M}$ is $\Omega(\log(T))$. For the k-armed bandit problem and when $\mathcal{M}$ is $\Theta(\log(T))$, Ensemble sampling satisfies the condition $\mathrm{KL}(\overline{Q}_t, \overline{\Pi}_t) < \epsilon$ in Theorem 3 with high probability. In this case, the regret of Ensemble sampling scales linearly with T.

We show in Theorem 2 that when the constraint $\mathrm{KL}(Q_t, \Pi_t) < \epsilon$ is satisfied, which implies by Lemma 2 that $\mathrm{KL}(\overline{Q}_t, \overline{\Pi}_t) < \epsilon$ is satisfied, there can exist approximation algorithms that have linear regret in T. This result provides a linear lower bound, which is complementary with the linear regret upper bound of Ensemble Sampling in (Lu & Van Roy, 2017). Moreover, we show in Appendix E that we can apply Theorem 3 to get sub-linear regret with Ensemble sampling with $\Theta(\log(T))$ models.

In reinforcement learning, there is a notion that certain approximations are "stochastically optimistic" and that this has implications for regret (Osband et al., 2016). This is similar in spirit to our analysis in terms of α-divergence, in that the characteristics of inference errors are important.

There has been a number of empirical works using approximate methods to perform Thompson sampling. Riquelme et al. (2018) implement variational inference, MCMC, Gaussian processes and other methods on synthetic and real world data sets and measure the regret. Urteaga & Wiggins (2018) derive a variational method for contextual bandits. Kawale et al. (2015) use particle filtering to implement Thompson sampling for matrix factorization.

Finally, if exact inference is not possible, it remains an open question if it is better to use Thompson sampling with approximate inference, or to use a different bandit method that does not require inference with respect to the posterior. For example Kveton et al. (2019) propose an algorithm based on the bootstrap.

8 Conclusion

In this paper we analyzed the performance of approximate Thompson sampling when at each time-step t, the algorithm obtains a sample from an approximate distribution Q_t such that the α-divergence between the true posterior and Q_t remains at most a constant ϵ at all time-steps.

Our results have the following implications. To achieve a sub-linear regret, we can only use $\alpha > 0$ for $o(T)$ time-steps. Therefore we should use $\alpha \leq 0$ with forced exploration to make the posterior concentrate. This method theoretically guarantees a sub-linear regret even when ϵ is a large constant.

Acknowledgments

We thank Huy Le for providing the proof of Lemma 9.

References

Agrawal, S. and Goyal, N. Further optimal regret bounds for Thompson sampling. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2013)*, volume 31 of *Proceedings of Machine Learning Research*, pp. 99–107. PMLR, 2013.

Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. I. An introduction to MCMC for machine learning. *Machine Learning*, 50(1):5–43, 2003. ISSN 1573-0565. doi: 10.1023/A:1020281327116.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

Cichocki, A. and Amari, S. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy*, 12:1532–1568, 2010.

Doucet, A. and Johansen, A. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, 12:656–704, 2009.

Gopalan, A., Mannor, S., and Mansour, Y. Thompson sampling for complex online problems. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 100–108, Bejing, China, 22–24 Jun 2014. PMLR.

Kawale, J., Bui, H. H., Kveton, B., Tran-Thanh, L., and Chawla, S. Efficient Thompson sampling for online matrix-factorization recommendation. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pp. 1297–1305. Curran Associates, Inc., 2015.

Kveton, B., Szepesvari, C., Vaswani, S., Wen, Z., Lattimore, T., and Ghavamzadeh, M. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3601–3610, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Liu, C.-Y. and Li, L. On the prior sensitivity of thompson sampling. In *Algorithmic Learning Theory*, pp. 321–336, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46379-7.

Lu, X. and Van Roy, B. Ensemble sampling. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 3260–3268. Curran Associates, Inc., 2017.

Minka, T. Divergence measures and message passing. Technical Report MSR-TR-2005-173, January 2005.

Minka, T., Winn, J., Guiver, J., Zaykov, Y., Fabian, D., and Bronskill, J. /Infer.NET 0.3, 2018. Microsoft Research Cambridge. http://dotnet.github.io/infer.

Ordentlich, E. and Weinberger, M. J. A distribution dependent refinement of Pinsker's inequality. *International Symposium on Information Theory, 2004. ISIT 2004. Proceedings.*, pp. 29–, 2004.

Osband, I., Van Roy, B., and Wen, Z. Generalization and exploration via randomized value functions. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 2377–2386. JMLR.org, 2016.

Qin, C., Klabjan, D., and Russo, D. Improving the expected improvement algorithm. In *Advances in Neural Information Processing Systems 30*, pp. 5381–5391. Curran Associates, Inc., 2017.

Riquelme, C., Tucker, G., and Snoek, J. Deep Bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. In *International Conference on Learning Representations (ICLR 2018)*, 2018.

Russo, D. Simple Bayesian algorithms for best arm identification. In *29th Annual Conference on Learning Theory (COLT 2016)*, volume 49 of *Proceedings of Machine Learning Research*, pp. 1417–1418. PMLR, 2016.

Russo, D. and Roy, B. V. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016.

Russo, D. J., Roy, B. V., Kazerouni, A., Osband, I., and Wen, Z. A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018. ISSN 1935-8237. doi: 10.1561/2200000070.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016.

Tran, D., Kucukelbir, A., Dieng, A. B., Rudolph, M., Liang, D., and Blei, D. M. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*, 2016.

Urteaga, I. and Wiggins, C. Variational inference for the multi-armed contextual bandit. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics (AISTATS 2018)*, volume 84 of *Proceedings of Machine Learning Research*, pp. 698–706. PMLR, 2018.

A Proof of Theorem 1 and Corollary 1

First we will prove Theorem 1. Let $\Omega_i \subseteq \Omega$ denote the region where arm i is the best arm. Let $\Pi_{t,i}$ denote $\Pi_t(\Omega_i)$, the posterior probability that arm i is the best arm. For $r > 1$, We construct the pdf of Q_t's as follows:

$$q_t(m) = \begin{cases} \frac{1}{r}\pi_t(m), & \text{if } m_1 > m_2 \\ \frac{1-\Pi_{t,1}/r}{1-\Pi_{t,1}}\pi_t(m), & \text{otherwise.} \end{cases} \tag{6}$$

We will prove the theorem by the following steps:

- In Lemma 3 we show that Q_t's are valid distributions.
- In Lemma 4 we show that when $\alpha > 0$ the α-divergence between Q_t and Π_t can be arbitrarily small
- In Lemma 5 we show that sampling from Q_t for $\Theta(T)$ time-steps will generate linear frequentist regret, and lower bound the regret.

Since the regret is linear, in Appendix A.4 we discuss the constant average regret per time-step as a function of ϵ and α. In Appendix A.5 we provide the Bayesian regret proof for Corollary 1.

Lemma 3. *$q_t(m)$ in Eq. 6 is well-defined and if $\int \pi_t(m)dm = 1$ then:*

$$\int q_t(m)dm = 1.$$

Lemma 4. *When $\alpha > 0$, for all $\epsilon > 0$, for all Π_t, there exists $r > 1$ such that when Q_t's are constructed from r as shown in Eq. 6, $D_\alpha(\Pi_t, Q_t) < \epsilon$*

Lemma 5. *The expected frequentist regret of the policy that constructs Q_t's as in Eq. 6 and sample from Q_t for $T' = \Theta(T)$ time-steps is linear and the lower bound of the average regret per time-step is*

$$L = \begin{cases} c\Delta(1-(1-\epsilon\alpha(1-\alpha))^{\frac{1}{1-\alpha}}), & \textit{when } \alpha > 1 \textit{ and } 0 < \epsilon \\ c\Delta(1-\frac{1}{e^\epsilon}), & \textit{when } \alpha = 1 \textit{ and } 0 < \epsilon \\ c\Delta(1-(1-\epsilon\alpha(1-\alpha))^{\frac{1}{1-\alpha}}), & \textit{when } 0 < \alpha < 1 \textit{ and } 0 < \epsilon \le \frac{1}{\alpha(1-\alpha)} \end{cases},$$

where $c = \frac{T'}{T}$ is $\Theta(1)$.

A.1 Proof of Lemma 3

Proof. First we will show that $\Pi_{t,2} = 1 - \Pi_{t,1} > 0$ for all $t \geq 0$, so that $q_t(m)$ is well-defined. We have $\Pi_{0,2} = \mathbb{P}(M_2 > M_1) > 0$ by assumption. Let $S_t = \{m \in \Omega_2 : \pi_t(m) > 0\}$ be the support of Π_t in Ω_2. If $\pi_0(m) > 0$, then $\pi_t(m) > 0$ because $\pi_t(m)$ is the product of $\pi_0(m)$ and non-zero likelihoods. Therefore $S_0 \subseteq S_t$.

Since $\mathbb{P}(M_2 > M_1) = \int_{S_0} \pi_0(m)dm > 0$, $\int_{S_0} dm > 0$. Since $S_0 \subseteq S_t$, $\int_{S_t} dm > 0$. Therefore $\int_{S_t} \pi_t(m)dm > 0$ since $S_t = \{m \in \Omega_2 : \pi_t(m) > 0\}$ by definition. Then $\Pi_{t,2} = \int_{\Omega_2} \pi_t(m)dm = \int_{S_t} \pi_t(m)dm > 0$.

Assume that $\int \pi_t(m)dm = 1$, we will show that $\int q_t(m)dm = 1$:

$$
\begin{aligned}
&\int q_t(m)dm \\
&= \int_{\Omega_1} q_t(m)dm + \int_{\Omega_2} q_t(m)dm \\
&= \int_{\Omega_1} \frac{1}{r}\pi_t(m)dm + \int_{\Omega_2} \frac{1-\Pi_{t,1}/r}{1-\Pi_{t,1}}\pi_t(m)dm \\
&= \frac{1}{r}\Pi_{t,1} + \frac{1-\Pi_{t,1}/r}{1-\Pi_{t,1}}\Pi_{t,2} \\
&= \frac{1}{r}\Pi_{t,1} + \frac{1-\Pi_{t,1}/r}{1-\Pi_{t,1}}(1-\Pi_{t,1}) \\
&= 1\ .
\end{aligned}
$$

□

A.2 Proof of Lemma 4

Proof. First we calculate the α-divergence between Π_t and Q_t constructed in Eq. 6. Let $\Omega_1 \subseteq \Omega$ denote the region where $m_1 > m_2$ and $\Omega_2 \subseteq \Omega$ denote the region where $m_2 \geq m_1$.

When $\alpha > 0, \alpha \neq 1$ we have:

$$
\begin{aligned}
&D_\alpha(\Pi_t, Q_t) \\
=&\frac{1-\int\left(\frac{\pi_t(m)}{q_t(m)}\right)^\alpha q_t(m)dm}{\alpha(1-\alpha)} \\
=&\frac{1-\int_{\Omega_1}\left(\frac{\pi_t(m)}{q_t(m)}\right)^\alpha q_t(m)dm - \int_{\Omega_2}\left(\frac{\pi_t(m)}{q_t(m)}\right)^\alpha q_t(m)dm}{\alpha(1-\alpha)} \\
=&\frac{1-\int_{\Omega_1}(r)^\alpha q_t(m)dm - \int_{\Omega_2}\left(\frac{1-\Pi_{t,1}}{1-\Pi_{t,1}/r}\right)^\alpha q_t(m)dm}{\alpha(1-\alpha)} \\
=&\frac{1-Q_t(\Omega_1)(r)^\alpha - Q_t(\Omega_2)\left(\frac{1-\Pi_{t,1}}{1-\Pi_{t,1}/r}\right)^\alpha}{\alpha(1-\alpha)} \\
=&\frac{1-\frac{\Pi_{t,1}}{r}(r)^\alpha - (1-\frac{\Pi_{t,1}}{r})\left(\frac{1-\Pi_{t,1}}{1-\Pi_{t,1}/r}\right)^\alpha}{\alpha(1-\alpha)} \\
=&\frac{1}{\alpha(1-\alpha)}\left(1-\Pi_{t,1}r^{-1+\alpha} - (1-\Pi_{t,1})^\alpha(1-\frac{\Pi_{t,1}}{r})^{1-\alpha}\right)\ .
\end{aligned}
\tag{7}
$$

When $\alpha = 1$:

$$
\begin{aligned}
&D_\alpha(\Pi_t, Q_t) \\
=&\int \pi_t(m)\log\left(\frac{\pi_t(m)}{q_t(m)}\right)dm \\
=&\int_{\Omega_1} \pi_t(m)\log\frac{\pi_t(m)}{q_t(m)}dm + \int_{\Omega_2} \pi_t(m)\log\frac{\pi_t(m)}{q_t(m)}\ dm \\
=&\int_{\Omega_1} \pi_t(m)\log(r)dm \\
&+\int_{\Omega_2} \pi_t(m)\log\frac{1-\Pi_{t,1}}{1-\Pi_{t,1}/r}dm \\
=&\Pi_{t,1}\log(r) + (1-\Pi_{t,1})\log\frac{1-\Pi_{t,1}}{1-\Pi_{t,1}/r}\ .
\end{aligned}
$$

We will now upper bound the above expression. Consider 2 cases

- $\alpha = 1$: We have

$$
\begin{aligned}
&D_\alpha(\Pi_t, Q_t) \\
&= \Pi_{t,1} \log(r) + (1 - \Pi_{t,1}) \log \frac{1 - \Pi_{t,1}}{1 - \Pi_{t,1}/r} \\
&\leq \Pi_{t,1} \log(r) + (1 - \Pi_{t,1}) \log(r) \text{ because } r > 1 \\
&\leq \log(r) \ .
\end{aligned}
$$

- $\alpha > 0, \alpha \neq 1$: The following inequality is true by simple calculations when $0 < \alpha < 1$ or $\alpha > 1$:

$$
\frac{\left(\frac{1-\Pi_{t,1}}{1-\frac{\Pi_{t,1}}{r}}\right)^{\alpha-1}}{\alpha(\alpha-1)} \leq \frac{r^{\alpha-1}}{\alpha(\alpha-1)} \ . \tag{8}
$$

  Then we will have:

$$
\begin{aligned}
&D_\alpha(\Pi_t, Q_t) \\
=&\frac{\Pi_{t,1} r^{\alpha-1} + (1 - \Pi_{t,1}) \left(\frac{1-\Pi_{t,1}}{1-\frac{\Pi_{t,1}}{r}}\right)^{\alpha-1} - 1}{\alpha(\alpha-1)} \\
\leq&\frac{1}{\alpha(\alpha-1)} \left(\Pi_{t,1} r^{\alpha-1} + (1 - \Pi_{t,1}) r^{\alpha-1} - 1\right) \\
=&\frac{1}{\alpha(\alpha-1)} \left(r^{-1+\alpha} - 1\right) \ .
\end{aligned}
$$

Therefore $D_\alpha(\Pi_t, Q_t)$ is upper bounded by:

$$
\begin{cases} \frac{1-r^{\alpha-1}}{\alpha(1-\alpha)}, & \text{if } 0 < \alpha < 1 \text{ or } \alpha > 1 \\ \log(r), & \text{if } \alpha = 1 \ . \end{cases} \tag{9}
$$

Since $\lim_{r \to 1} \log(r) = 0$ and $\lim_{r \to 1} \frac{1-r^{-1+\alpha}}{\alpha(1-\alpha)} = 0$, for any $\epsilon > 0$, there exists $r > 1$ such that

$$
D_\alpha(\Pi_t, Q_t) \leq \epsilon \ .
$$

□

A.3 Proof of Lemma 5

Proof. We will now lower bound the regret as a function of ϵ.

For any posterior Π_t, since the approximate algorithm sampling from Q_t picks the optimal arm with probability at most $1/r$ it then picks arm 2 with probability at least $1 - 1/r$.

Since we sample from Q_t for T' time steps, the lower bound of the average expected regret per time step is :

$$
L = \frac{T'}{T}(m_1^* - m_2^*)(1 - 1/r) = c\Delta(1 - 1/r) \ .
$$

where $\Delta = m_1^* - m_2^*$ and $c = \frac{T'}{T}$ is $\Theta(1)$.

We calculate ϵ as a function of r from Eq. 9:

$$
\epsilon = \begin{cases} \frac{1-r^{-1+\alpha}}{\alpha(1-\alpha)}, & \text{if } \alpha \neq 1 \\ log(r), & \text{if } \alpha = 1 \ . \end{cases}
$$

The functions are continous when $r > 1$. Then by direct calculations when $r \to \infty$ and $r \to 1$, the domain of ϵ is:

$$
\begin{aligned}
&0 < \epsilon \text{ when } \alpha \geq 1 \ . \\
&0 < \epsilon < \frac{1}{\alpha(1-\alpha)} \text{ when } 0 < \alpha < 1 \ .
\end{aligned}
$$

Then

$$r = \begin{cases} (1-\epsilon\alpha(1-\alpha))^{\frac{1}{-1+\alpha}} & \text{when } \alpha > 1 \text{ and } 0 < \epsilon \\ e^{\epsilon} & \text{when } \alpha = 1 \text{ and } 0 < \epsilon \\ (1-\epsilon\alpha(1-\alpha))^{\frac{1}{-1+\alpha}} & \text{when } 0 < \alpha < 1 \text{ and } 0 < \epsilon \leq \frac{1}{\alpha(1-\alpha)}. \end{cases}$$

Therefore we can calculate the lower bound of the regret per time-step as:

$$L = \begin{cases} c\Delta(1-(1-\epsilon\alpha(1-\alpha))^{\frac{1}{1-\alpha}}), & \text{when } \alpha > 1 \text{ and } 0 < \epsilon \\ c\Delta(1-\frac{1}{e^{\epsilon}}), & \text{when } \alpha = 1 \text{ and } 0 < \epsilon \\ c\Delta(1-(1-\epsilon\alpha(1-\alpha))^{\frac{1}{1-\alpha}}), & \text{when } 0 < \alpha < 1 \text{ and } 0 < \epsilon \leq \frac{1}{\alpha(1-\alpha)}. \end{cases}.$$

We plot the lower bound of the average regret per time step when $\Delta = 0.1$ as a function of ϵ in Fig 4.

□

A.4 The Average Regret per Time-step

To understand how the constant average regret per time-step depends on ϵ and α, we plot in Figure 4 the lower bound of the average regret per time-step in Lemma 5 as a function of ϵ in the following setting of the example constructed in the proof of Theorem 1. The algorithm samples from Q_t at $T/2$ time-steps and $\Delta = 0.1$. In this case the average regret per time step is upper bounded by $\Delta/2 = 0.05$. The formula and proof are detailed in Lemma 5 in Appendix A. When α is around 1,

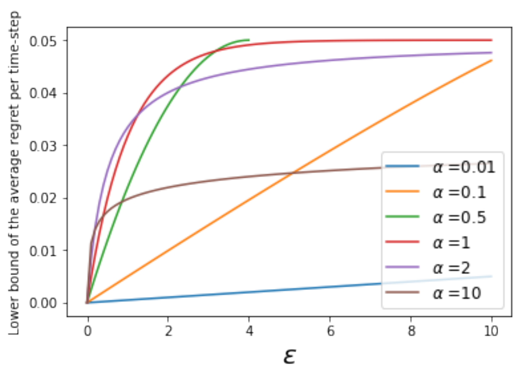


Figure 4: Lower bound of regret per time-step as a function of ϵ when $m_1^* - m_2^* = 0.1$ and we sample from the approximation for $T/2$ time-steps in the example construted in the proof of Theorem 1. When α is around 1, the lower bound converges quickly as ϵ increases.

the lower bound, and therefore the average regret per time-step, converges the fastest to $\Delta/2$ as ϵ increases. When α is very large or close to 0, the lower bound grows slowly as ϵ increases.

A.5 Proof of Corollary 1

Since $\mathbb{P}(M_1 > M_2) > 0$, there exist constants $\Delta > 0, \gamma > 0$ such that $\mathbb{P}(M_1 - M_2 \geq \Delta) = \gamma$. The probability that the assumption $m_1^* > m_2^*$ in Theorem 1 is satisfied is at least $\gamma > 0$. Therefore the expected regret over the prior is at least γ times the frequentist regret in Theorem 1, which is linear.

B Proof of Theorem 2 and Corollary 2

First we will prove Theorem 2. Let $\Pi_{t,i}$ denote $\Pi_t(\Omega_i)$. We construct the pdf of Q_t's as follows:

$$q_t(m) = \begin{cases} \frac{1}{\Pi_{t,2}}\pi_t(m), \text{ if } m_2 > m_1 \\ 0, \text{ otherwise.} \end{cases} \tag{10}$$

We will prove the theorem by the following steps:

- In Lemma 6 we show that Q_t's are valid distributions.
- In Lemma 7 we show that Q_t has linear frequentist regret, and calculate the constant average regret per time-step.
- In Lemma 8 we show that there exists a bad prior such that the α-divergence between Q_t and Π_t can be arbitrarily small.

In Appendix B.4 we discuss the prior-dependent error threshold ϵ that will cause linear regret. In Appendix B.5 we provide the Bayesian regret proof for Corollary 2.

Lemma 6. *$q_t(m)$ in Eq. 10 is well-defined and if $\int \pi_t(m)dm = 1$ then:*

$$\int q_t(m)dm = 1.$$

Lemma 7. *Q_t constructed in Eq. 10 chooses arm 2 at all time-steps. The average frequentist regret per time-step is $\Delta = m_1^* - m_2^*$.*

Lemma 8. *Let $\alpha < 1$, $M_1 - M_2$ and M_2 be independent and arm 2 be chosen at all time-steps before t.*

For any $\epsilon > 0$, there exists $0 < z \leq 1$ such that if $\Pi_{0,2} = z$ then $D_\alpha(\Pi_t, Q_t) < \epsilon$ where Q_t is constructed in Eq. 10.

For any $0 < z \leq 1$, there exists $\epsilon > 0$ such that if $\Pi_{0,2} = z$ then $D_\alpha(\Pi_t, Q_t) < \epsilon$ where Q_t is constructed in Eq. 10.

B.1 Proof of Lemma 6

Proof. Similar to the proof of Lemma 3, we have that $\Pi_{t,2} > 0$ for all $t \geq 0$.

Assume that $\int \pi_t(m)dm = 1$, we will show that $\int q_t(m)dm = 1$:

$$\begin{aligned}
&\int q_t(m)dm \\
&= \int_{\Omega_1} q_t(m)dm + \int_{\Omega_2} q_t(m)dm \\
&= 0 + \int_{\Omega_2} \frac{1}{\Pi_{t,2}} \pi_t(m)dm \\
&= \frac{1}{\Pi_{t,2}} \int_{\Omega_2} \pi_t(m)dm \\
&= 1 \; .
\end{aligned}$$

$\square$

B.2 Proof of Lemma 7

Proof. Under the approximate distribution, arm 2 is chosen with probability 1 at all times. Clearly this approximate distribution has linear regret, with $\Delta = m_1^* - m_2^*$ being the average regret per time-step. $\square$

B.3 Proof of Lemma 8

Proof. Let $D = M_1 - M_2$ which is independent of M_2 by the assumption. Let f denote the pdf. Since the algorithm always picks arm 2, H_{t-1} and M_1 are independent given M_2. Therefore for all m_1, m_2 and h, $f_{M_1|M_2,H_{t-1}}(m_1|m_2,h) = f_{M_1|M_2}(m_1|m_2)$.

Since $D = M_1 - M_2$, we have $f_{D|M_2,H_{t-1}}(m_1 - m_2|m_2,h) = f_{M_1|M_2,H_{t-1}}(m_1|m_2,h)$. Therefore for all d, m_2 and h:

$$f_{D|M_2,H_{t-1}}(d|m_2,h) = f_{M_1|M_2,H_{t-1}}(m_2 + d|m_2,h) = f_{M_1|M_2}(m_2 + d|m_2) = f_{D|M_2}(d|m_2) \; .$$

Since $f_{D|M_2,H_{t-1}}(d|m_2,h) = f_{D|M_2}(d|m_2)$ for all d, m_2 and h, D and H_{t-1} are independent given M_2. Then

$$
\begin{aligned}
& f_{D|M_2,H_{t-1}}(d|m_2,h) \\
=& f_{D|M_2}(d|m_2) \text{ because } D \text{ and } H_{t-1} \text{ are independent given } M_2 \\
=& f_D(d) \text{ because } D \text{ and } M_2 \text{ are independent.}
\end{aligned}
$$

Now we will show that D and H_{t-1} are independent. For all d and h:

$$
\begin{aligned}
& f_{D|H_{t-1}}(d|h) \\
=& \int f_{D,M_2|H_{t-1}}(d,m_2|h)dm_2 \\
=& \int f_{D|M_2,H_{t-1}}(d|m_2,h) f_{M_2|H_{t-1}}(m_2|h)dm_2 \\
=& \int f_D(d) f_{M_2|H_{t-1}}(m_2|h)dm_2 \\
=& f_D(d) \int f_{M_2|H_{t-1}}(m_2|h)dm_2 \\
=& f_D(d) \ .
\end{aligned}
$$

Since D and H_{t-1} are independent, at all times t the posterior does not concentrate:

$$
\Pi_{t,2} = \mathbb{P}(M_1 - M_2 < 0|H_{t-1}) = \mathbb{P}(M_1 < M_2) \ .
$$

For simplicity let

$$
z := \mathbb{P}(M_1 < M_2) \ .
$$

We will show that $D(\Pi_t, Q_t)$ is small if z is large enough. First we calculate the α-divergence between Π_t and Q_t constructed in Eq 10.

When $\alpha < 1, \alpha \neq 0$:

$$
\begin{aligned}
& D_\alpha(\Pi_t, Q_t) \\
=& \frac{1 - \int \left(\frac{q_t(m)}{\pi_t(m)}\right)^{1-\alpha} \pi_t(m)dm}{\alpha(1-\alpha)} \\
=& \frac{1 - \int_{\Omega_1} \left(\frac{q_t(m)}{\pi_t(m)}\right)^{1-\alpha} \pi_t(m)dm - \int_{\Omega_2} \left(\frac{q_t(m)}{\pi_t(m)}\right)^{1-\alpha} \pi_t(m)dm}{\alpha(1-\alpha)} \\
=& \frac{1 - 0 - \int_{\Omega_2} \left(\frac{1}{\Pi_{t,2}}\right)^{1-\alpha} \pi_t(m)dm}{\alpha(1-\alpha)} \text{ since } \alpha < 1 \\
=& \frac{1 - \left(\frac{1}{\Pi_{t,2}}\right)^{1-\alpha} \int_{\Omega_2} \pi_t(m)dm}{\alpha(1-\alpha)} \\
=& \frac{1 - \left(\frac{1}{\Pi_{t,2}}\right)^{1-\alpha} \Pi_{t,2}}{\alpha(1-\alpha)} \\
=& \frac{1 - (\Pi_{t,2})^\alpha}{\alpha(1-\alpha)} \\
=& \frac{1 - z^\alpha}{\alpha(1-\alpha)} \ .
\end{aligned}
$$

When $\alpha = 0$:

$$\begin{aligned}
&D_\alpha(\Pi_t, Q_t) \\
=&\int q_t(m) \log \frac{q_t(m)}{\pi_t(m)} dm \\
=&\int_{\Omega_1} q_t(m) \log \frac{q_t(m)}{\pi_t(m)} dm \\
&+ \int_{\Omega_2} q_t(m) \log \frac{q_t(m)}{\pi_t(m)} dm \\
=&\int_{\Omega_1} 0 \log(0) dm + \int_{\Omega_2} q_t(m) \log \frac{1}{\Pi_{t,2}} dm \\
=&0 + 1 \log \frac{1}{\Pi_{t,2}} = \log \frac{1}{\Pi_{t,2}} = \log \frac{1}{z} \ .
\end{aligned}$$

Note that if we don't have the condition on the prior such that picking arm 2 does not help to learn which arm is the better one, $\Pi_{t,2}$ may converge to 0, making $D_\alpha(\Pi_t, Q_t)$ goes to ∞ when $\alpha \le 0$. But since $\Pi_{t,2} = z$, we will now show that for any $\alpha < 1$, for any $\epsilon > 0$, there exists $z (0 < z < 1)$ such that

$$D_\alpha(\Pi_t, Q_t) < \epsilon \ .$$

Consider the 2 cases

- When $\alpha < 1, \alpha \neq 0$: Since

  $$\lim_{z \to 1} \frac{1 - z^\alpha}{\alpha(1 - \alpha)} = 0 \ .$$

  Then for any $\epsilon > 0$ there exists $0 < z < 1$ such that $D_\alpha(\Pi_t, Q_t) < \epsilon$. For any $0 < z < 1$ there exists $\epsilon > 0$ such that $D_\alpha(\Pi_t, Q_t) < \epsilon$.

- When $\alpha = 0$:

  $$D_\alpha(\Pi_t, Q_t) = \log \frac{1}{z} \ .$$

  Since $\lim_{z \to 1} \log(1/z) = 0$, for any $\epsilon > 0$ there exists $0 < z < 1$ such that $D_0(\Pi_t, Q_t) < \epsilon$. For any $z < 1$ there exists $\epsilon > 0$ such that $D_\alpha(\Pi_t, Q_t) < \epsilon$.

□

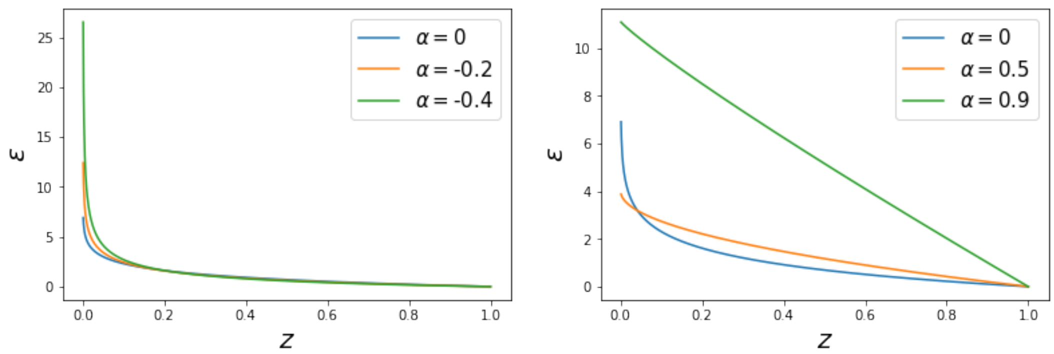
B.4 Prior-dependent Error Threshold for Linear Frequentist Regret

In the example constructed in the previous sections, the α-divergence between Π_t and Q_t can be calculated as: $\epsilon = \begin{cases} \frac{1 - z^\alpha}{\alpha(1-\alpha)}, \text{ if } 0 < \alpha < 1 \text{ or } \alpha < 0 \\ \log \frac{1}{z}, \text{ if } \alpha = 0 \end{cases}$.

In Figure 5, we show the values of ϵ as a function of z that will make the regret linear for different values of α. We can see that for both cases when $\alpha \le 0$ and $0 \le \alpha < 1$, and z is not too small, there is a threshold of ϵ for each value of z that makes the regret linear. For each value of z, if the error is smaller than the threshold we hypothesize that the regret might become sub-linear. However even if that is the case, it is not possible to calculate the exact threshold for more complicated priors. Therefore in Section 5.1 we propose an algorithm that is guaranteed to have sub-linear regret for any ϵ and any z when $\alpha \le 0$.

B.5 Proof of Corollary 2

Since $\mathbb{P}(M_1 > M_2) > 0$, there exist constants $\Delta > 0, \gamma > 0$ such that the $\mathbb{P}(M_1 - M_2 \ge \Delta) = \gamma$. The probability that the assumption $m_1^* > m_2^*$ in Theorem 2 is satisfied is at least $\gamma > 0$. Therefore the expected regret over the prior is at least γ times the frequentist regret in Theorem 2, which is linear.



(a) $D_\alpha(\Pi_t, Q_t) = \epsilon$ as a function of z when $\alpha \leq 0$. When z is very small and α is small, ϵ needs to be very large. When $z > 0.2$, there is a threshold of ϵ which is less than 8 that can cause linear regret.

(b) ϵ as a function of z when $0 \leq \alpha < 1$. There is a threshold of ϵ which is less than 8 for each value of z that can cause linear regret..

Figure 5: ϵ as a function z that makes the regret linear for different values of α for the example constructed in the proof of Theorem 2.

C Proof of Lemma 2

To convert between $D_\alpha(\Pi_t, Q_t)$ and $D_\alpha(\overline{\Pi_t}, \overline{Q_t})$ we first prove the following lemma:

Lemma 9 (Jensen's Inequality)**.** *Let $f : \mathcal{R}^2 \to \mathcal{R}$ be a convex function. Let $P : \mathcal{R}^k \to \mathcal{R}$ and $Q : \mathcal{R}^k \to \mathcal{R}$ be 2 functions. Let S is a subset of R^k, the domain of x and $|S|$ denote the volume of S. Then*

$$\begin{aligned} &\frac{1}{|S|}\int_S f(P(x), Q(x))dx \\ &\geq f\left(\frac{1}{|S|}\int_S P(x)dx, \frac{1}{|S|}\int_S Q(x)dx\right) . \end{aligned} \tag{11}$$

Proof. The multivariate Jensen's Inequality states that if X is a n-dimensional random vector and $f : \mathcal{R}^n \to \mathcal{R}$ is a convex function then

$$\mathbb{E}(f(X)) \geq f(\mathbb{E}(X)) .$$

To use the multivariate Jensen's Inequality we define the 2-dimensional random vector $X : S \to \mathcal{R}^2$ by $X(x) := (P(x), Q(x))$ and a probability distribution over S such that for all $x \in S$: $\mathbb{P}(x) = \frac{1}{|S|}$.

Then the left-hand side of Eq. 11 becomes $\mathbb{E}(f(X))$, while the right-hand side becomes $f(\mathbb{E}(X))$, and Eq. 11 follows from the multivariate Jensen's Inequality. □

Now we will prove Lemma 2.

Proof of Lemma 2. Since $D_\alpha(p, q)$ is convex (Cichocki & Amari, 2010), the following functions:

$$f(p, q) = q \log \frac{q}{p},$$
$$f(p, q) = p \log \frac{p}{q},$$
$$f(p, q) = \frac{p^\alpha q^{1-\alpha}}{\alpha(\alpha - 1)}$$

are convex, and we can apply Lemma 9:

- When $\alpha = 0$:

$$
\begin{aligned}
&D_\alpha(\Pi_t, Q_t) \\
=&\int q_t(m) \log \frac{q_t(m)}{\pi_t(m)} dm \\
=&\sum_i \int_{\Omega_i} q_t(m) \log \frac{q_t(m)}{\pi_t(m)} dm \\
\geq&\sum_i |\Omega_i| \frac{1}{|\Omega_i|} \int_{\Omega_i} q_t(m) dm \log \frac{\frac{1}{|\Omega_i|}\int_{\Omega_i} q_t(m) dm}{\frac{1}{|\Omega_i|}\int_{\Omega_i} \pi_t(m) dm} \text{ by applying Lemma 9} \\
=&\sum_i Q_{t,i} \log \frac{Q_{t,i}}{\Pi_{t,i}} \\
=&D_\alpha(\overline{\Pi_t}, \overline{Q_t}) \ .
\end{aligned}
$$

- When $\alpha = 1$:

$$
\begin{aligned}
&D_\alpha(\Pi_t, Q_t) \\
=&\int \pi_t(m) \log \frac{\pi_t(m)}{q_t(m)} dm \\
=&\sum_i \int_{\Omega_i} \pi_t(m) \log \frac{\pi_t(m)}{q_t(m)} dm \\
\geq&\sum_i |\Omega_i| \frac{1}{|\Omega_i|} \int_{\Omega_i} \pi_t(m) dm \log \frac{\frac{1}{|\Omega_i|}\int_{\Omega_i} \pi_t(m) dm}{\frac{1}{|\Omega_i|}\int_{\Omega_i} q_t(m) dm} \text{ by applying Lemma 9} \\
=&\sum_i \Pi_{t,i} \log \frac{\Pi_{t,i}}{Q_{t,i}} \\
=&D_\alpha(\overline{\Pi_t}, \overline{Q_t}) \ .
\end{aligned}
$$

- When $\alpha \neq 0, \alpha \neq 1$:

$$
\begin{aligned}
&D_\alpha(\Pi_t, Q_t) \\
=&\int \frac{\pi(x)^\alpha q(x)^{1-\alpha} - 1}{-\alpha(1-\alpha)} dx \\
=&\frac{-1}{\alpha(\alpha-1)} + \sum_i \int_{\Omega_i} \frac{\pi(x)^\alpha q(x)^{1-\alpha}}{\alpha(\alpha-1)} dx \\
\geq&\frac{-1}{\alpha(\alpha-1)} + \sum_i |\Omega_i| \frac{(\frac{\Pi_{t,i}}{|\Omega_i|})^\alpha (\frac{Q_{t,i}}{|\Omega_i|})^{1-\alpha}}{\alpha(\alpha-1)} \text{ by applying Lemma 9} \\
=&\frac{-1}{\alpha(\alpha-1)} + \sum_i \frac{\Pi_{t,i}^\alpha Q_{t,i}^{1-\alpha}}{\alpha(\alpha-1)} \\
=&D_\alpha(\overline{\Pi_t}, \overline{Q_t}) \ .
\end{aligned}
$$

□

D Proof of Theorem 3

We will prove that the frequentist regret is sub-linear for any m^*. If the algorithm has sub-linear frequentist regret for all values $M = m^*$, the Bayesian regret (which is the expected value over M) will also be sub-linear.

Without loss of generalization, let arm 1 be the best arm. From Lemma 1, since $\sum_{t=1}^{\infty} p_t = \infty$, we have for all arms i, $\sum_{t=1}^{\infty} P(A_t = i|H_{t-1}) = \infty$ and therefore with probability 1:

$$\lim_{t\to\infty} \Pi_{t,1} = \lim_{t\to\infty} \mathbb{P}(A^* = 1|H_{t-1}) = 1 \ , \tag{12}$$

which means that the posterior probability that arm 1 is the best arm converges to 1.

We will prove the theorem by proving the following steps:

- In Lemma 10 we show that if the probability that the posterior chooses the best arm tends to 1, then the probability that the approximation chooses the best arm also tends to 1
- In Lemma 11 and Lemma 12 we show that if the probability that the approximation chooses the best arm also tends to 1 almost surely, then it has sub-linear regret with probability 1. Therefore it has sub-linear regret in expectation over the history.

Lemma 10. *Let $\alpha \leq 0$ and arm 1 be the true best arm. Let $\Omega_i = \{m | m_i = max(m_1, ..., m_k)\}$ be the region where arm i is the best arm. If the posterior probability that arm 1 is the best arm converges to 1:*

$$\lim_{t \to \infty} \Pi_{t,1} = 1$$

and for all $t \geq 0$:

$$D_\alpha(\Pi_t, Q_t) < \epsilon,$$

then the sequence $\{Q_{t,1}\}_t$ where $Q_{t,1} = \int_{\Omega_1} q_t(m)dm$ converges and

$$\lim_{t \to \infty} Q_{t,1} = 1 \ .$$

Next we show that if the approximate distribution concentrates, then the probability that it chooses the wrong arm decreases as T goes to infinity.

Lemma 11. *If*

$$\lim_{t \to \infty} Q_{t,1} = 1$$

then

$$\lim_{T \to \infty} \frac{\sum_{t=1}^{T}(1 - Q_{t,1})}{T} = 0 \ .$$

From Lemma 10 and Lemma 11, since $\lim_{t \to \infty} \Pi_{t,1} = 1$ with probability 1, we have $\lim_{T \to \infty} \frac{\sum_{t=1}^{T}(1-Q_{t,1})}{T} = 0$ with probability 1. We will now show that the expected regret is sub-linear:

Lemma 12. *Let $p_t = o(1)$ be such that $\sum_{t=1}^{\infty} p_t = \infty$. For any number of arms k, any prior Π_0 and any error threshold $\epsilon > 0$, the following algorithm has $o(T)$ regret: at every time-step t,*

- *with probability $1 - p_t$, sample from an approximate posterior Q_t such that $\lim_{T \to \infty} \frac{\sum_{t=1}^{T}(1-Q_{t,1})}{T} = 0$ with probability 1, and*
- *with probability p_t, sample an arm uniformly at random.*

D.1 Proof of Lemma 10

Proof. Let $Q_{t,i} = \int_{\Omega_i} q_t(m)dm$ and $\Pi_{t,i} = \int_{\Omega_i} \pi_t(m)dm$. Then

$$\lim_{t \to \infty} \Pi_{t,1} = 1$$

and we want to show that $\{Q_{t,1}\}_t$ converges and

$$\lim_{t \to \infty} Q_{t,1} = 1 \ .$$

Since $D_\alpha(\overline{\Pi_t}, \overline{Q_t}) < \epsilon$ and $\lim \Pi_{t,1} = 1$ we want to show that $\limsup Q_{t,1} = 1$. By contradiction, assume that:

$$\limsup Q_{t,1} = c < 1 \ .$$

Then there exists a sub-sequence of $\{Q_{t,1}\}_t$, denoting $Q_{t_1,1}, Q_{t_2,1}, ..., Q_{t_n,1}, ..$ such that

$$\lim_{n\to\infty} Q_{t_n,1} = c \ . \tag{13}$$

which implies

$$0 < 1 - c = \lim_{n\to\infty} \sum_{i=2}^{k} Q_{t_n,i} \leq \sum_{i=2}^{k} \limsup_{n\to\infty} Q_{t_n,i}.$$

Therefore there exists $j \in [2, k]$ such that:

$$\limsup_{n\to\infty} Q_{t_n,j} = d > 0 \ .$$

Then there exists a sub-sequence of $\{Q_{t_n,j}\}_n$, denoting $Q_{t_{n_1},j}, Q_{t_{n_2},j}, ..., Q_{t_{n_m},j}, ..$ such that

$$\lim_{m\to\infty} Q_{t_{n_m},j} = d \ .$$

We consider the 2 cases:

- When $\alpha = 0$:

  $$D_\alpha(\overline{\Pi}_t, \overline{Q}_t) = \sum_{i=1}^{k} Q_{t,i} \log \frac{Q_{t,i}}{\Pi_{t,i}} \ .$$

  Then we have:

  $$\begin{aligned}\epsilon = &\lim_{m\to\infty} D_\alpha(\overline{\Pi}_{t_{n_m}}, \overline{Q}_{t_{n_m}}) \\ \geq &\lim_{m\to\infty} Q_{t_{n_m},1} \log \frac{Q_{t_{n_m},1}}{\Pi_{t_{n_m},1}} + \lim_{m\to\infty} Q_{t_{n_m},j} \log \frac{Q_{t_{n_m},j}}{\Pi_{t_{n_m},j}} \\ = &c \log \frac{c}{1} + d \log \frac{d}{0} \\ = &\infty \text{ since } d > 0,\end{aligned}$$

  which is a contradiction. Therefore $c = 1$.

- When $\alpha < 0$:

  $$D_\alpha(\overline{\Pi}_t, \overline{Q}_t) = \frac{\sum_{i=1}^{k} \Pi_{t,i}^{\alpha} Q_{t,i}^{1-\alpha} - 1}{\alpha(\alpha - 1)} \ .$$

  Then we have:

  $$\begin{aligned}\epsilon = &\lim_{m\to\infty} D_\alpha(\overline{\Pi}_{t_{n_m}}, \overline{Q}_{t_{n_m}}) \\ \geq &\lim_{m\to\infty} \frac{\Pi_{t_{n_m},1}^{\alpha} Q_{t_{n_m},1}^{1-\alpha} + \Pi_{t_{n_m},j}^{\alpha} Q_{t_{n_m},j}^{1-\alpha} - 1}{\alpha(\alpha - 1)} \\ = &\frac{1^\alpha c^{1-\alpha} + \frac{d^{1-\alpha}}{(0)^{-\alpha}} - 1}{\alpha(\alpha - 1)} \\ = &\infty, \text{ since } d > 0 \text{ and } \alpha < 0,\end{aligned}$$

  which is a contradiction. Therefore $c = 1$.

Similarly we will show that:

$$\liminf Q_{t,1} = 1 \ .$$

By contradiction, assume that:

$$\liminf Q_{t,1} = c' < 1 \ .$$

Then there exists a sub-sequence of $\{Q_{t,1}\}_t$, denoting $Q_{t_1,1}, Q_{t_2,1}, ..., Q_{t_{n'},1}, ..$ such that

$$\lim_{n\to\infty} Q_{t_{n'},1} = c' \ .$$

Using the same argument following Eq. 13 we will have $c' = 1$. Since $\liminf Q_{t,1} = \limsup Q_{t,1} = 1$, we have that $\{Q_{t,1}\}_t$ converges and

$$\lim Q_{t,1} = 1 \ .$$

$\square$

D.2 Proof for Lemma 11

For simplicity let x_t denote $1 - Q_{t,1}$. We want to show that if a sequence $\{x_t\}$ satisfies $x_t \geq 0 \ \forall t$ and:

$$\lim_{t\to\infty} x_t = 0,$$

then

$$\lim_{T\to\infty} S_T = 0,$$

where

$$S_T = \frac{\sum_{t=1}^{T} x_t}{T} \ .$$

Since $\lim_{t\to\infty} x_t = 0$ and $x_t \geq 0 \ \forall t$, for any $\epsilon > 0$ there exists T_0 such that for all $t > T_0$:

$$x_t < \frac{\epsilon}{2} \ .$$

Then for all $T > T_0$:

$$\begin{aligned} S_T &= \frac{x_1 + ... + x_{T_0}}{T} + \frac{x_{T_0+1} + ... + x_T}{T} \\ &\leq \frac{x_1 + ... + x_{T_0}}{T} + \frac{\frac{\epsilon}{2}T}{T} \\ &\leq \frac{x_1 + ... + x_{T_0}}{T} + \frac{\epsilon}{2} \ . \end{aligned}$$

Choose T_1 large enough such that $\frac{x_1+...+x_{T_0}}{T_1} < \frac{\epsilon}{2}$. Let $T_2 = \max(T_0, T_1)$. Then for all $T > T_2$:

$$S_T = \frac{x_1 + ... + x_{T_0}}{T} + \frac{\epsilon}{2} < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \ .$$

Therefore for any $\epsilon > 0$, there exists T_2 such that for all $T > T_2$, $S_T < \epsilon$. Since $S_T \geq 0 \ \forall T$, we have:

$$\lim_{T\to\infty} S_T = 0 \ .$$

D.3 Proof of Lemma 12

Without loss of generalization, let arm 1 be the true best arm. Let $\Delta = m_1^* - \max(m_2^*, ..., m_k^*)$ be the gap between the highest mean m_1^* and the next highest mean of the arms.

Since $p_t = o(1)$, $\sum_{t=1}^{T} p_t$ is $o(T)$. Therefore the regret from the uniform sampling steps is $o(T)$.

Since $1 - Q_{t,1}$ is the probability of choosing a sub-optimal arm by sampling from Q_t, the regret of sampling from Q_t is upper bounded by:

$$\mathbb{E} \sum_{t=1}^{T} \Delta(1 - Q_{t,1}) \ .$$

Since $\lim_{T\to\infty} \frac{\sum_{t=1}^{T}(1-Q_{t,1})}{T} = 0$ with probability 1, we have

$$\lim_{T\to\infty} \frac{\sum_{t=1}^{T} \Delta(1-Q_{t,1})}{T} = 0$$

with probability 1. Therefore

$$\lim_{T\to\infty} \mathbb{E}\frac{\sum_{t=1}^{T} \Delta(1-Q_{t,1})}{T} = 0,$$

which means that the regret of sampling from Q_t is sub-linear. Since both the expected regrets of the uniform sampling steps and of sampling from Q_t are sub-linear, the total expected regret is sub-linear.

E Ensemble Sampling and Uniform Exploration

To the best of our knowledge, (Lu & Van Roy, 2017) is the only work that provides a theoretical analysis of Thompson sampling when the sampling step is approximate. Lu & Van Roy (2017) propose an approximate sampling method called Ensemble sampling where they maintain a set of $\mathcal{M}$ models to approximate the posterior, and analyze its regret for linear contextual bandits. When the model is a k-armed bandit, the regret bound is as follow:

Lemma 13 (implied by (Lu & Van Roy, 2017))**.** *Let π^{TS} and π^{ES} denote the exact Thompson sampling and Ensemble sampling policies. Let $\Delta = max(m_1^*, ..., m_k^*) - \min(m_1^*, ..., m_k^*)$. For all $\epsilon > 0$, if*

$$\mathcal{M} \geq \frac{2k}{\epsilon^2} \log \frac{2kT}{\epsilon^2 \delta},$$

then

$$\mathrm{Regret}(T, \pi^{ES}) \leq \mathrm{Regret}(T, \pi^{TS}) + \epsilon\Delta T + \delta\Delta T \tag{14}$$

Lu & Van Roy (2017) prove the regret bound by only using the following property of the Ensemble sampling method: at time t, with probability $1-\delta$, Ensemble sampling satisfies the following constraint:

$$\mathrm{KL}(\overline{Q}_t, \overline{\Pi}_t) < \epsilon^2, \tag{15}$$

where ϵ is a constant if $\mathcal{M}$ is $\Theta(\log(T))$. If ϵ is a constant the regret will be linear because of the term $\epsilon\Delta T$.

At time t, with probability $1-\delta$, $\mathrm{KL}(\overline{Q}_t, \overline{\Pi}_t) < \epsilon^2$. The first 2 terms in the right hand side of Eq. 14 comes from the time-steps when $\mathrm{KL}(\overline{Q}_t, \overline{\Pi}_t) < \epsilon^2$, and the last term comes from the other case with probability δ.

Theorem 3 shows that applying an uniform sampling step will make the posterior concentrate. Moreover, Lemma 10 implies that if Eq. 15 is satisfied at a subset of times $\mathcal{T}_0 \subseteq [0, 1, ..., T]$, the approximation Q_t will also concentrate when $t \in \mathcal{T}_0$. Therefore the regret from the time-steps in $\mathcal{T}_0$ will be sub-linear in $\mathcal{T}_0$, which is sub-linear in T.

So if we want to maintain a small number of models $M = \Theta(\log(T))$ and achieve sub-linear regret, we can apply Theorem 3 as follow. First we choose δ to be small such that the last term in Eq. 14 $\delta\Delta T$ is $o(T)$. Then we apply the uniform sampling step as shown in Theorem 3, so that the first 2 terms in the right hand side of Eq. 14 become sub-linear. We can then achieve sub-linear regret with Ensemble sampling with a $\Theta(\log T)$ number of models.

F KL Divergence between two Gaussian Distributions

The KL divergence between two Gaussian distributions is:

$$\begin{aligned}
&\mathrm{KL}(\mathrm{Norm}(\mu_1, \Sigma_1), \mathrm{Norm}(\mu_2, \Sigma_2)) \\
&= \frac{1}{2}(\mathrm{trace}(\Sigma_2^{-1}\Sigma_1) - k \\
&+ (\mu_2 - \mu_1)^T \Sigma_2^{-1}(\mu_2 - \mu_1) + \ln \frac{\mathrm{det}\Sigma_2}{\mathrm{det}\Sigma_1})
\end{aligned}$$

G Posterior Calculation

In our simulations, when both the prior and the reward distributions are Gaussian, we calculate the true posterior using the following closed-form solution.

Let the posterior at time t be multivariate Gaussian distribution $\mathrm{Norm}(\mu_t, \Sigma_t)$ where μ_t is a $k \times 1$ vector and Σ_t is a $k \times k$ covariance matrix. Let the reward distribution of arm i be $\mathrm{Norm}(m_i^*, \sigma^2)$ where σ is known and m_i^*'s are unknown.

Let $A_t \in \{0, 1\}^k$ be a $0/1$ vector where $A_t(i) = 1$ if arm i is chosen at time t, and 0 otherwise. Let $r_t \in \mathcal{R}$ be the reward of the arm chosen at time t.

Then the posterior at time $t + 1$ is $\mathrm{Norm}(\mu_{t+1}, \Sigma_{t+1})$ where:

$$\Sigma_{t+1} = (\Sigma_t^{-1} + A_t A_t^T / \sigma^2)^{-1}$$
$$\mu_{t+1} = \Sigma_{t+1}(\Sigma_t^{-1} \mu_t + A_t r_t / \sigma^2) \ .$$