An Analysis of Ensemble Sampling

Chao Qin
Columbia University
cqin22@gsb.columbia.edu

Zheng Wen Xiuyuan Lu Benjamin Van Roy
DeepMind
{zhengwen,lxlu,benvanroy}@google.com

Abstract

Ensemble sampling serves as a practical approximation to Thompson sampling when maintaining an exact posterior distribution over model parameters is computationally intractable. In this paper, we establish a regret bound that ensures desirable behavior when ensemble sampling is applied to the linear bandit problem. This represents the first rigorous regret analysis of ensemble sampling and is made possible by leveraging information-theoretic concepts and novel analytic techniques that may prove useful beyond the scope of this paper.

1 Introduction

Thompson sampling (TS) is a popular heuristic for balancing between exploration and exploitation in bandit learning. In its basic form, TS operates maintains a posterior distribution over model parameters. When the prior distribution and likelihood function exhibit conjugacy properties, the posterior can be maintained through computationally tractable Bayesian inference. However, many practical contexts call for more complex models for which exact Bayesian inference becomes computationally intractable. For such contexts, ensemble sampling (ES) can serve as a practical approximation to TS [Lu and Van Roy, 2017].

Instead of the posterior distribution, ES maintains an ensemble of statistically plausible models that can be updated in an efficient incremental manner. The corresponding discrete distribution represents an approximation to the posterior distribution. At the start of each timestep, a model is sampled uniformly from the ensemble and an action is selected to optimize expected reward with respect to the sampled model. Each model is initially sampled from the prior distribution and evolves through an updating process that adapts to observations and random perturbations.

While a growing literature [Osband et al., 2016, Russo et al., 2018, Osband et al., 2018, Lu et al., 2018, Osband et al., 2019, Dwaracherla et al., 2020, Osband et al., 2022b] presents applications of ES, there has been no rigorous theory that ensures desirable behavior similar to that enjoyed by TS. A regret bound for ES applied to the linear bandit is provided by Lu and Van Roy [2017], but a flaw in the associated analysis has brought this result into question. Since the publication of that paper, several researchers have tried to remedy the analysis or otherwise establish performance guarantees for ES, but these efforts have gone without success.

We offer in this paper the first rigorous regret analysis of ES. Like Lu and Van Roy [2017], we study ES applied to the linear-Gaussian bandit. This serves as a simple sanity check for the approach. We establish a Bayesian regret bound (Theorem 1) that consists of two terms. The first term is identical to a Bayesian regret bound established for TS. The second term accounts for posterior distribution mismatch and can be made arbitrarily small by increasing the ensemble size. Our analysis leverages information-theoretic concepts and novel analytic techniques that may prove useful beyond the scope of this paper.

As a stepping stone in our analysis, we establish a general Bayesian regret bound that applies to *any* learning algorithm for a linear bandit (Theorem 2). In particular, this regret bound is based on the Hellinger distance between the action selection distribution of the learning algorithm and the

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

posterior distribution of the optimal action at each timestep. We believe that this Bayesian regret bound is of independent interest and might be useful in analyzing other learning algorithms, such as other approximate TS algorithms.

Note that for linear-Gaussian bandits considered in this paper, Kalman filtering offers a tractable means to exact inference and, as such, an approximation method like ES is not required. However, it is worth mentioning that the analysis in this paper also offers insight into more complex models that call for approximate inference. For example, consider a neural network with very wide hidden layers. A recent research thrust (e.g., Jacot et al. [2018], Lee et al. [2017]) highlights that, under suitable technical conditions, training such a neural network via stochastic gradient descent (SGD) approximates Gaussian process inference. Our result extends directly to finite-armed Gaussian process bandits. This suggests that ES applied with neural networks that are trained via SGD can be effective in complex bandit problems, though formalizing this extension requires significant technical work, which we leave for future research.

The remainder of this paper is organized as follows: Section 2 reviews related work, and Section 3 formulates the considered linear bandit problem and describes the ensemble sampling (ES) algorithm. Section 4 presents the main result of this paper: a Bayesian regret bound for ES in linear bandits. Before diving into the analysis, we define notation used in this paper in Section 5. Then, we motivate and derive the general regret bound in Section 6 and sketch the proof of the main result in Section 7. Finally, we conclude the paper and discuss future directions in Section 8. Detailed proofs are provided in the appendices.

2 Related work

Thompson sampling [Thompson, 1933, Chapelle and Li, 2011, Russo et al., 2018] is a commonly used heuristic for balancing exploration and exploitation in bandit [Lattimore and Szepesvári, 2020] and reinforcement learning [Sutton and Barto, 2018] problems. However, the basic form of Thompson sampling can be computationally intractable unless the prior distribution and likelihood function exhibit conjugacy properties. To overcome this computational challenge, variants of approximate versions of TS (see Chapter 5 of Russo et al. [2018]) have been developed, including approximate TS using Laplace approximation [Chapelle and Li, 2011], bootstrapping [Eckles and Kaptein, 2019, Kveton et al., 2019], variational inference [Urteaga and Wiggins, 2018, Yu et al., 2020], Markov chain Monte Carlo [Casella and George, 1992, Roberts and Tweedie, 1996], hypermodels [Dwaracherla et al., 2020], and optimal transport [Zhang et al., 2019].

Ensemble sampling (ES) is also an approximate version of Thompson sampling, and was formally proposed by Lu and Van Roy [2017]. It has been one of the most popular methods in approximate Bayesian inference and has wide applications in sequential decision making. For example, variations of ES have been applied to (deep) reinforcement learning [Osband et al., 2016, 2018, 2019], online recommendation [Lu et al., 2018, Hao et al., 2020, Zhu and Van Roy, 2021], multi-agent reinforcement learning [Dimakopoulou and Van Roy, 2018], behavioral sciences [Eckles and Kaptein, 2019], and marketing strategies [Yang et al., 2020]. Recently, Osband et al. [2022b] has developed a testbed to compare and rank agents for uncertainty modeling, including variants of the ensemble agents.

A lot of work has attempted to analyze ensemble sampling, but none of them has been successful. In their original paper, Lu and Van Roy [2017] try to analyze ES in linear-Gaussian bandits, but there is a mistake in the final step of the analysis where they compare the regret of ES to that of TS. In particular, in the proof of Lemma 8, the random variables A_t^{TS} are defined on histories of actions selected by ensemble sampling, instead of TS. Therefore, the sum of regret incurred by these hypothetical Thompson sampling actions A_t^{TS} is not equal to the cumulative regret of applying a TS agent throughout all timesteps. The paper incorrectly claims that they are equal.

In addition to Lu and Van Roy [2017], Phan et al. [2019] aims to analyze general approximate TS with approximation errors measured in α-divergence for K-armed bandit. To demonstrate the main result, Phan et al. [2019] uses ensemble sampling as an example, but also inherits the technical flaw in Lu and Van Roy [2017]. In addition, though Phan et al. [2019] shows that an approximate version of Thompson sampling that uses uniform sampling achieves sublinear regret, this result mainly benefits from forced exploration.

There are also papers aiming to analyze bandit algorithms with approximate inference. A recent paper [Huang et al., 2022] proposes and analyzes an upper-confidence-bound (UCB) type algorithm for bandit problems with approximate inference. Another recent paper [Ash et al., 2021] proposes and analyzes an ensemble based UCB algorithm for bandit problems. Kveton et al. [2020] studies linear bandits in the frequentist setting and provides a general regret bound that is tailored to analyzing randomized algorithms with good concentration and anti-concentration properties, but it could not be directly applied to analyzing ES due to ES's discrete and incremental update nature.

3 Preliminaries

We begin by introducing a formulation of the linear bandit and an ensemble sampling (ES) algorithm suited for this environment.

3.1 Linear bandit

The linear bandit [Abbasi-Yadkori et al., 2011, Lattimore and Szepesvári, 2020] has received much attention in the bandit learning literature. It serves as a simple didactic environment, often used to sanity-check agent designs. In this spirit, we will analyze ES as applied to the linear bandit.

We consider a linear bandit with a Gaussian prior distribution and Gaussian noise. An instance is characterized by a tuple $\mathcal{E} = (\mathcal{A}, \mu_0, \Sigma_0, \sigma^2)$, where $\mathcal{A} \subset \mathbb{R}^d$ is a finite action set, μ_0 and Σ_0 are the prior mean vector and covariance matrix, and σ^2 is the noise variance. Note that actions are vectors of dimension d. Let $K = |\mathcal{A}|$ denote the cardinality of the action set. The prior distribution over the coefficient vector $\theta \in \mathbb{R}^d$, $\mathbb{P}(\theta \in \cdot)$, is multivariate Gaussian $N(\mu_0, \Sigma_0)$. Rewards are generated by a random sequence $(R_t : t = 1, 2, \ldots)$ of K-dimensional vectors, which is i.i.d. conditioned on θ. In particular, for each t and a,

$$R_{t,a} = a^\top \theta + W_{t,a},$$

for an i.i.d. sequence $(W_t : t = 1, 2, \ldots)$ of K-dimensional vectors with each element distributed $N(0, \sigma^2 I)$.

An agent interacts with the linear bandit as follows: at each timestep $t = 0, 1, \ldots$, the agent executes an action A_t; then it observes a reward R_{t+1,A_t}. Note that only the reward associated with the executed action is observed through what is termed *bandit feedback*. The agent's experience through time t is encoded by a history $H_t = (A_0, R_{1,A_0}, \ldots, A_{t-1}, R_{t,A_{t-1}})$. The agent's objective is to maximize expected reward over some long duration T:

$$\textstyle\sum_{t=0}^{T-1} \mathbb{E}\big[R_{t+1,A_t}\big]. \tag{1}$$

This is equivalent to minimizing the Bayesian regret:

$$\mathrm{Regret}(T) = \textstyle\sum_{t=0}^{T-1} \mathbb{E}[R_{t+1,A_*} - R_{t+1,A_t}], \tag{2}$$

where $A_* \sim \mathrm{unif}\left\{\arg\max_{a \in \mathcal{A}} a^\top \theta\right\}$. Note that in both Equations (1) and (2), the expectations integrate over random reward realizations, algorithmic randomness, and the coefficient vector θ. Note that A_* is also a random variable since it depends on θ as well the uniform sampling among maximizers, which breaks ties.

3.2 Ensemble sampling

In the absence of conjugacy properties that enable efficient Bayesian inference, TS in its exact form becomes computationally infeasible. ES serves as a practical approximation to TS [Lu and Van Roy, 2017], often suitable for such contexts. The key idea behind ES is to maintain an ensemble of statistically plausible models that can be updated in an efficient incremental manner and to treat the discrete distribution represented by this set of models as an approximation to the posterior distribution. The algorithm begins by sampling M models from the prior distribution, where M is a hyperparameter. Then, at the start of each timestep, a model is sampled uniformly from this set and an action is selected to optimize the immediate expected reward with respect to the sampled model. Each model is updated incrementally based on the observed reward and a random perturbation. Algorithm 1 presents the associated pseudo-code for ES.

Algorithm 1 Ensemble Sampling

Sample: $\tilde{\theta}_{0,1}, \ldots, \tilde{\theta}_{0,M} \sim \Pr(\theta \in \cdot)$
for $t = 0, \ldots, T-1$ **do**
  Sample: $m \sim \text{unif}\{1, \ldots, M\}$
  Execute: $A_t \sim \text{unif}\left\{\arg\max_{a \in \mathcal{A}} a^\top \tilde{\theta}_{t,m}\right\}$
  Observe: R_{t+1,A_t}
  Update: $\tilde{\theta}_{t,1:M} \longrightarrow \tilde{\theta}_{t+1,1:M}$
end for

The posterior distribution at time $t+1$ is Gaussian with covariance matrix and mean vector

$$\Sigma_{t+1} = \left(\Sigma_t^{-1} + \tfrac{1}{\sigma^2} A_t A_t^\top\right)^{-1} \qquad \text{and} \qquad \mu_{t+1} = \Sigma_{t+1}\left(\Sigma_t^{-1}\mu_t + \tfrac{R_{t+1,A_t}}{\sigma^2} A_t\right). \tag{3}$$

While this highlights conjugacy properties of a sort that allow for efficient Bayesian inference and thus tractable implementation of exact TS, the linear bandit with Gaussian prior distribution and likelihood function serves as a useful context for studying the behavior of ES in relation to TS. Like [Lu and Van Roy, 2017], we consider a version of ES that updates each m-th model according to

$$\tilde{\theta}_{t+1,m} = \Sigma_{t+1}\left(\Sigma_t^{-1}\tilde{\theta}_{t,m} + \tfrac{R_{t+1,A_t} + \tilde{W}_{t+1,m}}{\sigma^2} A_t\right), \tag{4}$$

where each $\tilde{W}_{t+1} = \left(\tilde{W}_{t+1,1}, \ldots, \tilde{W}_{t+1,M}\right) \sim N(0, \sigma^2 I)$ is an independent random perturbation.

4 Regret bound for ensemble sampling

In this paper, we establish the following Bayesian regret bound for ES (Algorithm 1) when it is applied to linear bandits described in Section 3.1:

Theorem 1. *Suppose ES Algorithm (Algorithm 1) is applied to linear bandit with prior $\mathbb{P}(\theta \in \cdot) = N(\mu_0, \Sigma_0)$ and M models, then*

$$\text{Regret}(T) \leq \underbrace{\iota\sqrt{dT\mathbb{H}(A_*)}}_{(a)} + \underbrace{\kappa T\sqrt{\frac{K\log(6TM)}{M}}}_{(b)},$$

where d is the dimension of actions, T is the time horizon, K is the number of actions, $\mathbb{H}(A_*)$ is the entropy of the optimal action A_* under the prior, and

$$\iota \triangleq \sqrt{2\left(\max_{a \in \mathcal{A}} a^\top \Sigma_0 a + \sigma^2\right)} \tag{5}$$

and

$$\kappa \triangleq 2\sqrt{(4\log K + 5)\max_{a \in \mathcal{A}} a^\top \Sigma_0 a + \max_{a \in \mathcal{A}} \left(a^\top \mu_0\right)^2 + \sigma^2}. \tag{6}$$

By treating the noise variance σ^2, prior mean vector μ_0, and prior covariance matrix Σ_0 as constants, we have $\iota = O(1)$ and $\kappa = O(\sqrt{\log K})$, and thus

$$\text{Regret}(T) \leq O\left(\sqrt{dT\mathbb{H}(A_*)}\right) + \tilde{O}\left(T\sqrt{K/M}\right),$$

where $\tilde{O}$ hides the logarithmic factors.

Theorem 1 represents the first rigorously proved regret bound that ensures some degree of robustness in application of ES to a nontrivial bandit problem. One limitation is that the second term (b) in our Bayesian regret bound depends on K since our current analysis uses the method of (finite) types [Cover and Thomas, 2006]. This might not be tight for linear bandit, and we conjecture that the

second term (b) can be improved to be independent of K. Another limitation is that the result applies only to the linear bandit with Gaussian prior distribution and likelihood function, which admits tractable application of exact TS. Nevertheless, this result represents a beginning to the theory of ES, suggesting that the approach is sound and the analysis might be extended to broader problem classes in the future. In the remainder of this section, we compare our regret bound to others presented in closely related literature.

Comparison with the regret bound for TS The regret bound in Theorem 1 consists of two terms. The first term (a) is exactly the regret bound achieved by TS with exact posterior [Russo and Van Roy, 2016]. Note that $\mathbb{H}(A_*)$ is the entropy of the optimal action A_* and we always have $\mathbb{H}(A_*) \leq \log K$. As discussed in Russo and Van Roy [2016], this first term (a) is order optimal and improves upon the worst-case regret bound achieved by other algorithms such as upper confidence bound algorithms since $\mathbb{H}(A_*) \ll \log K$ when the prior is informative. On the other hand, the second term (b) is an incremental term and accounts for posterior distribution mismatch. Note that as the number of models used by ES $M \to \infty$, the second term (b) converges to 0 and the regret bound in Theorem 1 reduces to that of TS with exact posterior. Moreover, as long as $M \geq KT/d$, the regret bound in Theorem 1 has the same order as that of TS in terms of d and T (up to logarithmic factors).

Comparison with the regret bound in Lu and Van Roy [2017] In Theorem 1, the upper bound is on the Bayesian regret where the prior over θ is $N(\mu_0, \Sigma_0)$, and the regret also takes expectation over θ. Lu and Van Roy [2017] studies ES in the frequentist setting where unknown θ is fixed, and tries to upper bound the frequentist regret of ES by that of TS plus some approximation error term due to posterior mismatch, similar to the second term (b) in our Bayesian regret bound. However, there are some technical issues in the analysis, as is discussed in Section 2. Theorem 3 in Lu and Van Roy [2017] suggests that for any $\epsilon > 0$, when the number of models $M = \tilde{\Theta}\left(K/\epsilon^2\right)$ (where $\tilde{\Theta}$ hides the logarithmic factors), the approximation error term is $\epsilon\Delta(\theta)T$ where $\Delta(\theta) = \max_{a\in\mathcal{A}} a^\top\theta - \min_{a\in\mathcal{A}} a^\top\theta$, while the second term (b) in our Bayesian regret bound in Theorem 1 is linear in $\kappa = O(\sqrt{\log K})$. To make the second term (b) scale as $\tilde{O}\left(\epsilon T\right)$, the number of required models M is still $\tilde{\Theta}\left(K/\epsilon^2\right)$.

5 Notations

Before proceeding, we define some notations including Kullback–Leibler divergence, Hellinger distance, Shannon entropy and mutual information.

5.1 Kullback–Leibler divergence and Hellinger distance

For two probability distributions P and Q, the Kullback–Leibler (KL) divergence between P and Q is defined as

$$\mathbf{d}_{\mathrm{KL}}(P\|Q) \triangleq \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}P$$

if P is absolutely continuous with respect to Q; otherwise, define $\mathbf{d}_{\mathrm{KL}}(P\|Q) = +\infty$. Note that in general $\mathbf{d}_{\mathrm{KL}}(P\|Q) \neq \mathbf{d}_{\mathrm{KL}}(Q\|P)$. The Hellinger distance between P and Q is defined as

$$\mathbf{d}_{\mathrm{H}}(P\|Q) \triangleq \sqrt{\int\left(\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q}\right)^2},$$

which is symmetric in P and Q. In the special case of discrete distributions $P = (p_1, \ldots, p_n)$ and $Q = (q_1, \ldots, q_n)$, the KL divergence and Hellinger distance between P and Q can be written as

$$\mathbf{d}_{\mathrm{KL}}(P\|Q) = \textstyle\sum_{i\in[n]} p_i \log\left(p_i/q_i\right) \quad \text{and} \quad \mathbf{d}_{\mathrm{H}}(P\|Q) = \sqrt{\textstyle\sum_{i\in[n]}\left(\sqrt{p_i} - \sqrt{q_i}\right)^2}.$$

The following inequality hold for Hellinger distance and KL divergences:

Fact 1 (Lemma 2.4 in Tsybakov [2009])**.** *For two probability distributions P and Q, we have*

$$\mathbf{d}_{\mathrm{H}}^2(P\|Q) \leq \min\{\mathbf{d}_{\mathrm{KL}}(P\|Q),\, \mathbf{d}_{\mathrm{KL}}(Q\|P)\}.$$

In other words, the (squared) Hellinger distance is upper bounded by KL divergences in both directions. We will use Fact 1 in Section 6.

5.2 Measures in information theory

Consider a discrete random variable X whose support is a finite set $\mathcal{X}$ and general random variable Y and Z. The *Shannon entropy* of X is defined as

$$\mathbb{H}(X) \triangleq -\sum_{x\in\mathcal{X}} \mathbb{P}(X=x)\log\mathbb{P}(X=x).$$

The entropy of X conditional on $Y=y$ is

$$\mathbb{H}(X|Y=y) \triangleq -\sum_{x\in\mathcal{X}} \mathbb{P}\left(X=x|Y=y\right)\log\mathbb{P}(X=x|Y=y),$$

and the *conditional entropy* of X given Y is,

$$\mathbb{H}(X|Y) \triangleq \mathbb{E}_Y\left[-\sum_{x\in\mathcal{X}} \mathbb{P}\left(X=x|Y\right)\log\mathbb{P}(X=x|Y)\right],$$

where the expectation is taken with respect to the marginal distribution of Y.

The *mutual information* between X and Y

$$\mathbb{I}(X;Y) \triangleq \mathbf{d}_{\mathrm{KL}}\left(P(X,Y)\,||\,P(X)P(Y)\right) \tag{7}$$

is the KL divergence between the joint distribution of X and Y and the product of the marginal distributions, and the *conditional mutual information* between X and Y given Z is

$$\mathbb{I}(X;Y|Z) \triangleq \mathbb{E}_Z\left[\mathbf{d}_{\mathrm{KL}}\left(P(X,Y|Z)\,||\,P(X|Z)P(X|Z)\right)\right],$$

where the expectation is taken with respect to the marginal distribution of Z.

6 General regret bound

In this section, we derive a general Bayesian regret bound for *any* learning algorithm (Theorem 2), based on the Hellinger distance between the (conditional) action-selection distribution specified by the algorithm and the posterior distribution of the optimal action A_*, and use it to establish the Bayesian regret bound for ES in Theorem 1. We believe that our Bayesian regret bound is of independent interest and might be used to analyze other bandit algorithms, such as variants of approximate TS algorithms, in the future.

To simplify the exposition, we first define some shorthand notations to simplify the exposition of this paper. First, we use subscript t to denote conditioning on H_t, specifically, we define

$$\mathbb{P}_t(\cdot) \triangleq \mathbb{P}(\cdot|H_t) \quad \text{and} \quad \mathbb{E}_t[\cdot] \triangleq \mathbb{E}[\cdot|H_t].$$

In addition, we define probability distributions $\overline{p}_t$ and p_t over action set $\mathcal{A}$ as

$$\overline{p}_t(\cdot) \triangleq \mathbb{P}_t(A_t=\cdot) \quad \text{and} \quad p_t(\cdot) \triangleq \mathbb{P}_t(A_*=\cdot).$$

Given the history H_t, $\overline{p}_t$ is the conditional distribution of action A_t and p_t is the posterior distribution of the optimal action A_*. Both p_t and $\overline{p}_t$ are specified by the algorithm. Note that under TS, given the history H_t, $\overline{p}_t$ and p_t are the same, but they are not under other algorithms. We conjecture that for approximate TS algorithms, when the approximation is accurate, p_t is close to $\overline{p}_t$. We have the following Bayesian regret bound for any learning algorithm:

Theorem 2. *For linear bandits considered in Section 3.1, under any learning algorithm,*

$$\mathrm{Regret}(T) \leq \iota\sqrt{dT\mathbb{H}(A_*)} + \eta\sum_{t=0}^{T-1}\sqrt{\mathbb{E}\left[\mathbf{d}_{\mathrm{H}}^2(\overline{p}_t\|p_t)\right]}, \tag{8}$$

where ι is defined in Equation (5) *and*

$$\eta \triangleq 2\sqrt{\mathbb{E}\left[\max_{a\in\mathcal{A}}\left(a^\top\theta\right)^2\right]+\sigma^2} \tag{9}$$

where the expectation in Equation (9) *is taken with respect to the prior over θ.*

6.1 Discussion on the regret bound in Theorem 2

The regret bound in Theorem 2 holds for any learning algorithm. Note $\mathbb{H}(A_*)$ is the entropy of the optimal action A_* and we always have $\mathbb{H}(A_*) \leq \log K$. The first term matches the regret bound for TS derived in Russo and Van Roy [2016]. In the second term, $\sum_{t=0}^{T-1} \sqrt{\mathbb{E}\left[\mathbf{d}_{\mathrm{H}}^2(\overline{p}_t \| p_t)\right]}$ quantifies the cumulative distance between the conditional distribution of the action A_t and the posterior distribution the optimal action A_*. Under TS with exact posterior, $\mathbf{d}_{\mathrm{H}}^2(\overline{p}_t \| p_t) = 0$ due to the posterior matching property of TS, i.e., $\overline{p}_t = p_t$. Moreover, for approximate TS algorithms, if the approximation is accurate, $\mathbf{d}_{\mathrm{H}}^2(\overline{p}_t \| p_t)$ should be small.

Regret bound in terms of KL divergence between $\overline{p}_t$ and p_t It is sometimes more convenient to analyze the KL divergence between $\overline{p}_t$ and p_t. Since the squared Hellinger distance is symmetric and upper bound by the KL divergence (see Fact 1), we obtain the regret upper bound with the approximation error in terms of two forms of KL divergence between $\overline{p}_t$ and p_t.

Corollary 1. *For linear bandits considered in Section 3.1, under any learning algorithm,*

$$\mathrm{Regret}(T) \leq \iota \sqrt{dT\mathbb{H}(A_*)} + \eta \sum_{t=0}^{T-1} \sqrt{\mathbb{E}\left[\min\left\{\mathbf{d}_{\mathrm{KL}}(\overline{p}_t \| p_t), \mathbf{d}_{\mathrm{KL}}(p_t \| \overline{p}_t)\right\}\right]} \tag{10}$$

where ι and η are defined in Equations (5) *and* (9)*, respectively.*

We will show in Section 7 that for ensemble sampling, $\mathbf{d}_{\mathrm{KL}}(\overline{p}_t \| p_t)$ can be bounded in terms of the number of models M. It is noteworthy that in some other problems, it is more convenient to bound the other form $\mathbf{d}_{\mathrm{KL}}(p_t \| \overline{p}_t)$. One such example is the regret bound developed in Wen et al. [2022].

Upper bound on constant η in Theorem 2 and Corollary 1 The constant η in Equation (9) depends on the prior over coefficient vector θ. If the prior over θ is Gaussian random vector, so is $a^\top \theta$ for any action $a \in \mathcal{A}$. It is known that the expectation of the maximum of K (potentially correlated) Gaussian random variables is $\Theta(\sqrt{\log K})$. We show in the appendix that the expectation of the maximum of K squares of Gaussian random variables is $O(\log K)$. Please refer to Lemma 7 in Appendix B, which immediately leads to the following upper bound on η.

Corollary 2. *Suppose the prior over θ is $N(\mu_0, \Sigma_0)$. We have $\eta \leq \kappa$, where η and κ are defined in Equations* (9) *and* (6)*, respectively.*

6.2 Proof sketch of Theorem 2

The cumulative regret can be written as follows,

$$\mathrm{Regret}(T) = \sum_{t=0}^{T-1} \mathbb{E}\left[\mathbb{E}_t[R_{t+1,A_*} - R_{t+1,A_t}]\right] = \sum_{t=0}^{T-1} \mathbb{E}[G_t] + \sum_{t=0}^{T-1} \mathbb{E}[D_t] \tag{11}$$

where the per-period "main regret term" G_t and "approximation error term" D_t are defined as

$$G_t \triangleq \sum_{a \in \mathcal{A}} \sqrt{\overline{p}_t(a) p_t(a)} \left(\mathbb{E}_t[R_{t+1,a} | A_* = a] - \mathbb{E}_t[R_{t+1,a}]\right) \tag{12}$$

and

$$D_t \triangleq \sum_{a \in \mathcal{A}} \left(\sqrt{p_t(a)} - \sqrt{\overline{p}_t(a)}\right) \left(\sqrt{p_t(a)} \mathbb{E}_t[R_{t+1,a} | A_* = a] + \sqrt{\overline{p}_t(a)} \mathbb{E}_t[R_{t+1,a}]\right). \tag{13}$$

The last equality in Equation (11) is shown in the following lemma.

Lemma 1. *For $t = 0, \ldots, T-1$,*

$$\mathbb{E}_t[R_{t+1,A_*} - R_{t+1,A_t}] = G_t + D_t.$$

Proof. Fix t. Conditioned on H_t, the regret can be written as follows.

$$\begin{aligned}
&\mathbb{E}_t[R_{t+1,A_*} - R_{t+1,A_t}] \\
&= \sum_{a \in \mathcal{A}} p_t(a) \mathbb{E}_t[R_{t+1,a} | A_* = a] - \sum_{a \in \mathcal{A}} \overline{p}_t(a) \mathbb{E}_t[R_{t+1,a} | A_t = a] \\
&= \sum_{a \in \mathcal{A}} p_t(a) \mathbb{E}_t[R_{t+1,a} | A_* = a] - \sum_{a \in \mathcal{A}} \overline{p}_t(a) \mathbb{E}_t[R_{t+1,a}] \\
&= \sum_{a \in \mathcal{A}} \sqrt{\overline{p}_t(a) p_t(a)} \left(\mathbb{E}_t[R_{t+1,a} | A_* = a] - \mathbb{E}_t[R_{t+1,a}]\right) \\
&\quad + \sum_{a \in \mathcal{A}} \left(\sqrt{p_t(a)} - \sqrt{\overline{p}_t(a)}\right) \left(\sqrt{p_t(a)} \mathbb{E}_t[R_{t+1,a} | A_* = a] + \sqrt{\overline{p}_t(a)} \mathbb{E}_t[R_{t+1,a}]\right)
\end{aligned}$$

where the second equation uses $R_{t+1} = (R_{t+1,1}, \ldots, R_{t+1,K})$ is independent of the action A_t. $\square$

Now we are going to upper bound $\sum_{t=0}^{T-1} \mathbb{E}[G_t]$ and $\sum_{t=0}^{T-1} \mathbb{E}[D_t]$ in Equation (11), respectively. We first analyze $\sum_{t=0}^{T-1} \mathbb{E}[G_t]$. We upper bound each main regret term G_t by the mutual information between the optimal action A_* and action-reward pair (A_t, R_{t+1,A_t}) conditioning on a given history H_t:

Lemma 2. *For $t = 0, \ldots, T-1$, with probability 1,*

$$G_t \leq \iota \sqrt{d \cdot \mathbb{I}_t \left(A_*; (A_t, R_{t+1,A_t})\right)}$$

where ι is defined in Equation (5) *and $\mathbb{I}_t(A_*; (A_t, R_{t+1,A_t}))$ is the mutual information between A_* and (A_t, R_{t+1,A_t}) conditioning on a given history at timestep t.*

This result generalizes Proposition 5 and Corollary 1 in Russo and Van Roy [2016] for Thompson sampling to approximate Thompson sampling such that A_t and A_* are not identically distributed given the history. The proof of Lemma 2 can be found in Appendix A.

Then, similar to the analysis in Russo and Van Roy [2016], using the Cauchy-Schwarz inequality and the chain rule for mutual information, we obtain the following result.

Lemma 3. *With ι defined in Equation* (5)*, the following inequality holds.*

$$\textstyle\sum_{t=0}^{T-1} \mathbb{E}[G_t] \leq \iota \sqrt{dT\mathbb{H}(A_*)}.$$

The complete proof of Lemma 3 is in Appendix A. Note that the upper bound in the above result is the first term of the regret bound in Theorem 2.

The remaining task is to upper bound $\mathbb{E}[D_t]$ where D_t is the approximation error term defined in Equation (13). By analyzing (potentially) unbounded prior and noises, we have the following lemma.

Lemma 4. *With η defined in Equation* (9)*, for $t = 0, \ldots, T-1$,*

$$\mathbb{E}[D_t] \leq \eta \sqrt{\mathbb{E}\left[\mathbf{d}_{\mathrm{H}}^2(\overline{p}_t \| p_t)\right]}.$$

Summing over timestep t gives the second term of the regret bound in Theorem 2, i.e.,

$$\textstyle\sum_{t=0}^{T-1} \mathbb{E}[D_t] \leq \eta \sum_{t=0}^{T-1} \sqrt{\mathbb{E}\left[\mathbf{d}_{\mathrm{H}}^2(\overline{p}_t \| p_t)\right]}.$$

The proof of this lemma is provided below.

Proof of Lemma 4. Fix t.

$$\begin{aligned}
D_t =& \sum_{a \in \mathcal{A}} \left(\sqrt{p_t(a)} - \sqrt{\overline{p}_t(a)}\right) \left(\sqrt{p_t(a)}\mathbb{E}_t[R_{t+1,a}|A_* = a] + \sqrt{\overline{p}_t(a)}\mathbb{E}_t[R_{t+1,a}]\right) \\
=& \sum_{a \in \mathcal{A}} \left(\sqrt{p_t(a)} - \sqrt{\overline{p}_t(a)}\right) \sqrt{p_t(a)}\mathbb{E}_t[R_{t+1,a}|A_* = a] + \sum_{a \in \mathcal{A}} \left(\sqrt{p_t(a)} - \sqrt{\overline{p}_t(a)}\right) \sqrt{\overline{p}_t(a)}\mathbb{E}_t[R_{t+1,a}] \\
\leq& \sqrt{\sum_{a \in \mathcal{A}} \left(\sqrt{p_t(a)} - \sqrt{\overline{p}_t(a)}\right)^2} \left(\sqrt{\sum_{a \in \mathcal{A}} p_t(a) \left(\mathbb{E}_t[R_{t+1,a}|A_* = a]\right)^2} + \sqrt{\sum_{a \in \mathcal{A}} \overline{p}_t(a) \left(\mathbb{E}_t[R_{t+1,a}]\right)^2}\right) \\
\leq& \mathbf{d}_{\mathrm{H}}(\overline{p}_t \| p_t) \left(\sqrt{\sum_{a \in \mathcal{A}} p_t(a) \left(\mathbb{E}_t\left[R_{t+1,a}^2|A_* = a\right]\right)} + \sqrt{\sum_{a \in \mathcal{A}} \overline{p}_t(a) \left(\mathbb{E}_t\left[R_{t+1,a}^2\right]\right)}\right) \\
=& \mathbf{d}_{\mathrm{H}}(\overline{p}_t \| p_t) \left(\sqrt{\sum_{a \in \mathcal{A}} p_t(a)\mathbb{E}_t\left[R_{t+1,a}^2|A_* = a\right]} + \sqrt{\sum_{a \in \mathcal{A}} \overline{p}_t(a)\mathbb{E}_t\left[R_{t+1,a}^2|A_t = a\right]}\right) \\
=& \mathbf{d}_{\mathrm{H}}(\overline{p}_t \| p_t) \left(\sqrt{\mathbb{E}_t\left[R_{t+1,A_*}^2\right]} + \sqrt{\mathbb{E}_t\left[R_{t+1,A_t}^2\right]}\right)
\end{aligned}$$

where the first inequality applies Cauchy–Schwarz inequality; the second inequality uses Jensen's inequality as well as the definition of Hellinger distance; the second-to-last equality uses A_t is independent of $R_{t+1} = (R_{t+1,1}, \ldots, R_{t+1,K})$. Taking expectation on both sides, we have

$$\begin{aligned}
\mathbb{E}[D_t] &\leq \mathbb{E}\left[\mathbf{d}_{\mathrm{H}}(\overline{p}_t \| p_t)\left(\sqrt{\mathbb{E}_t\left[R_{t+1,A_*}^2\right]} + \sqrt{\mathbb{E}_t\left[R_{t+1,A_t}^2\right]}\right)\right] \\
&= \mathbb{E}\left[\mathbf{d}_{\mathrm{H}}(\overline{p}_t \| p_t)\sqrt{\mathbb{E}_t\left[R_{t+1,A_*}^2\right]}\right] + \mathbb{E}\left[\mathbf{d}_{\mathrm{H}}(\overline{p}_t \| p_t)\sqrt{\mathbb{E}_t\left[R_{t+1,A_t}^2\right]}\right] \\
&\leq \sqrt{\mathbb{E}\left[\mathbf{d}_{\mathrm{H}}^2(\overline{p}_t \| p_t)\right]}\left(\sqrt{\mathbb{E}\left[\mathbb{E}_t\left[R_{t+1,A_*}^2\right]\right]} + \sqrt{\mathbb{E}\left[\mathbb{E}_t\left[R_{t+1,A_t}^2\right]\right]}\right) \\
&= \sqrt{\mathbb{E}\left[\mathbf{d}_{\mathrm{H}}^2(\overline{p}_t \| p_t)\right]}\left(\sqrt{\mathbb{E}\left[R_{t+1,A_*}^2\right]} + \sqrt{\mathbb{E}\left[R_{t+1,A_t}^2\right]}\right) \\
&= \sqrt{\mathbb{E}\left[\mathbf{d}_{\mathrm{H}}^2(\overline{p}_t \| p_t)\right]}\left(\sqrt{\mathbb{E}\left[(A_*^\top\theta + W_{t+1,A_*})^2\right]} + \sqrt{\mathbb{E}\left[(A_t^\top\theta + W_{t+1,A_t})^2\right]}\right) \\
&= \sqrt{\mathbb{E}\left[\mathbf{d}_{\mathrm{H}}^2(\overline{p}_t \| p_t)\right]}\left(\sqrt{\mathbb{E}\left[(A_*^\top\theta)^2 + \sigma^2\right]} + \sqrt{\mathbb{E}\left[(A_t^\top\theta)^2\right] + \sigma^2}\right) \\
&\leq 2\sqrt{\mathbb{E}\left[\mathbf{d}_{\mathrm{H}}^2(\overline{p}_t \| p_t)\right]}\sqrt{\mathbb{E}\left[\max_{a\in\mathcal{A}}\left(a^\top\theta\right)^2\right] + \sigma^2}
\end{aligned}$$

where the second inequality uses Cauchy–Schwarz inequality and the last equality holds since $W_{t+1,A_*}, W_{t+1,A_t} \sim N(0, \sigma^2)$ are independent noises. □

7 Approximation error under ensemble sampling

In this section, we are going to bound the cumulative approximation error $\sum_{t=0}^{T-1}\sqrt{\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\overline{p}_t \| p_t)\right]}$ in Corollary 1 for ensemble sampling, where $\overline{p}_t$ is the conditional distribution of the chosen action A_t and p_t is the posterior distribution of the optimal action A_*. The whole section is devoted to proving the following Lemma 5 on per-period approximation error. Summing over timestep t yields the upper bound on cumulative approximation error, which implies the regret bound for ensemble sampling in Theorem 1.

Lemma 5. *Under ensemble sampling (Algorithm 1), for all $t = 0, \ldots, T-1$,*

$$\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\overline{p}_t \| p_t)\right] \leq \frac{K\log(6(t+1)M)}{M}.$$

Before proving this result, we first discuss the behavior of ensemble sampling. Ensemble sampling uniformly samples a model $m \in \{1, \ldots, M\}$, and then chooses the action A_t uniformly from the set $\tilde{\mathcal{A}}_{t,m} \triangleq \arg\max_{a\in\mathcal{A}} a^\top\tilde{\theta}_{t,m}$. We define the following approximation of $p_t(a)$:

$$\hat{p}_t(a) \triangleq \tfrac{1}{M}\sum_{m=1}^{M}\tfrac{1}{|\tilde{\mathcal{A}}_{t,m}|}\mathbb{1}\left\{a \in \tilde{\mathcal{A}}_{t,m}\right\}$$

where $|\tilde{\mathcal{A}}_{t,m}|$ is the cardinality of the set $\tilde{\mathcal{A}}_{t,m}$ (with probability 1, $|\tilde{\mathcal{A}}_{t,m}| = 1$). ES is a simple variant of approximate TS. At time t, TS would sample an action from the posterior distribution of the optimal action A_*, while ES would sample an action from the approximation $\hat{p}_t$. Notice that the history H_t does not include independent random perturbation $\tilde{W}_t \sim N(0, \sigma^2 I)$ used in updating M models (see Equation (4)). Therefore, given history H_t, $\hat{p}_t(a)$ is still a random variable, and its expectation is the conditional probability of sampling action a, i.e., $\mathbb{E}_t[\hat{p}_t(a)] = \overline{p}_t(a)$. By convexity of KL divergence, the per-period approximation error

$$\mathbf{d}_{\mathrm{KL}}(\overline{p}_t \| p_t) = \mathbf{d}_{\mathrm{KL}}\left(\mathbb{E}_t\left[\hat{p}_t\right] \| p_t\right) \leq \mathbb{E}_t\left[\mathbf{d}_{\mathrm{KL}}\left(\hat{p}_t \| p_t\right)\right].$$

Taking expectation of both sides gives $\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}(\overline{p}_t \| p_t)\right] \leq \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\hat{p}_t \| p_t\right)\right]$, and we are going to upper bound $\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\hat{p}_t \| p_t\right)\right]$. Note that the relationship between the expectation and cumulative distribution function can be written as

$$\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\hat{p}_t \| p_t\right)\right] = \int_0^\infty \mathbb{P}\left(\mathbf{d}_{\mathrm{KL}}\left(\hat{p}_t \| p_t\right) > \epsilon\right)\mathrm{d}\epsilon. \tag{14}$$

The following result in Lu and Van Roy [2017] can be used to upper bound each $\mathbb{P}\left(\mathbf{d}_{\mathrm{KL}}\left(\hat{p}_t \| p_t\right) > \epsilon\right)$ in the above integral.

Lemma 6 (Lemma 4 in Lu and Van Roy [2017])**.** *Let $\epsilon > 0$. For $t = 0, \ldots, T-1$,*

$$\mathbb{P}\left(\mathbf{d}_{\mathrm{KL}}\left(\hat{p}_t \| p_t\right) > \epsilon \,|\, \theta\right) \leq (t+1)^K (M+1)^K e^{-M\epsilon}.$$

Taking expectation over θ gives $\mathbb{P}\left(\mathbf{d}_{\mathrm{KL}}\left(\hat{p}_t \| p_t\right) > \epsilon\right) \leq (t+1)^K (M+1)^K e^{-M\epsilon}$. Consider any threshold $\delta \geq 0$. The integral in Equation (14) can be decomposed and bounded as follows.

$$\begin{aligned}
\int_0^\infty \mathbb{P}\left(\mathbf{d}_{\mathrm{KL}}\left(\hat{p}_t \| p_t\right) > \epsilon\right) \mathrm{d}\epsilon &= \int_0^\delta \mathbb{P}\left(\mathbf{d}_{\mathrm{KL}}\left(\hat{p}_t \| p_t\right) > \epsilon\right) \mathrm{d}\epsilon + \int_\delta^\infty \mathbb{P}\left(\mathbf{d}_{\mathrm{KL}}\left(\hat{p}_t \| p_t\right) > \epsilon\right) \mathrm{d}\epsilon \\
&\leq \delta + (t+1)^K (M+1)^K \int_\delta^\infty e^{-M\epsilon} \mathrm{d}\epsilon \\
&= \delta + \frac{(t+1)^K (M+1)^K e^{-M\delta}}{M}.
\end{aligned}$$

Taking derivative of RHS gives optimal $\delta^* = \frac{K[\log(t+1)+\log(M+1)]}{M}$. Plugging in δ^* yields

$$\int_0^\infty \mathbb{P}\left(\mathbf{d}_{\mathrm{KL}}\left(\hat{p}_t \| p_t\right) > \epsilon\right) \mathrm{d}\epsilon \leq \frac{K\left[\log(t+1) + \log(M+1)\right] + 1}{M} \leq \frac{K \log(6(t+1)M)}{M}.$$

This completes the proof of Lemma 5. Summing over timestep t yields the upper bound on cumulative approximation error, which implies the regret bound for ensemble sampling in Theorem 1.

8 Concluding remarks

We have presented the first rigorous analysis for ensemble sampling (ES) and developed the regret bound in Theorem 1 for linear bandits. The regret bound approaches that for Thompson sampling (TS) as the ensemble size increases. It is worth mentioning that ES is developed for broader and more challenging bandit and reinforcement learning problems, and the linear bandit serves only as a simple sanity check. In particular, ES can be used to approximate TS in the absence of conjugacy properties, when it is intractable to maintain and sample from the exact posterior. As discussed in Section 2, a substantial literature has been devoted to research in this direction. It is also noteworthy that in many practical problems, ES with a moderate number of models (e.g. ≤ 30 models) outperforms other state-of-the-art benchmarks [Lu and Van Roy, 2017, Lu et al., 2018, Osband et al., 2019].

As a stepping stone, we have also established a general regret bound in Theorem 2 that applies to *any* learning algorithm for linear bandits . This general regret bound may be of independent interest, since it is particularly useful for analyzing other varieties of approximate TS (e.g. approximate TS with epistemic neural networks [Osband et al., 2022a]). To establish a regret bound for an approximate TS algorithm based on Theorem 2, one only need to bound the Hellinger distance between the algorithm's sampling distribution and the true posterior distribution over the optimal action.

Finally, we believe that Theorem 2 could be generalized to any bandit problem based on the concept of information ratio proposed in Russo and Van Roy [2016], as long as the information ratio can be appropriately bounded. These problems include online learning with full information feedback [Shalev-Shwartz, 2012] and "graph" information feedback [Alon et al., 2015], combinatorial semi-bandits [Kveton et al., 2015, Wen et al., 2015], and pure exploration problems [Bubeck et al., 2009, Russo, 2016, Komiyama et al., 2021]. We also conjecture that the results of this paper can be extended to generalized linear bandits [Filippi et al., 2010, Dong et al., 2019] and episodic reinforcement learning [Osband et al., 2013], but leave these extensions to future work.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Conference on Learning Theory*, pages 23–35, 2015.

Jordan T. Ash, Cyril Zhang, Surbhi Goel, Akshay Krishnamurthy, and Sham Kakade. Anti-concentrated confidence bonuses for scalable exploration. *arXiv preprint arXiv:2110.11202*, 2021.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37, 2009.

George Casella and Edward I George. Explaining the Gibbs sampler. *The American Statistician*, 46 (3):167–174, 1992.

Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. A Wiley-Interscience publication. Wiley, 2006.

Maria Dimakopoulou and Benjamin Van Roy. Coordinated exploration in concurrent reinforcement learning. In *International Conference on Machine Learning*, volume 80, pages 1271–1279, 2018.

Shi Dong, Tengyu Ma, and Benjamin Van Roy. On the performance of Thompson sampling on logistic bandits. In *Conference on Learning Theory*, pages 1158–1160, 2019.

Vikranth Dwaracherla, Xiuyuan Lu, Morteza Ibrahimi, Ian Osband, Zheng Wen, and Benjamin Van Roy. Hypermodels for exploration. In *International Conference on Learning Representations*, 2020.

Dean Eckles and Maurits Kaptein. Bootstrap Thompson sampling and sequential decision problems in the behavioral sciences. *SAGE Open*, 9(2), 2019.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 23, 2010.

Botao Hao, Jie Zhou, Zheng Wen, and Will Wei Sun. Low-rank tensor bandits. *arXiv preprint arXiv:2007.15788*, 2020.

Jean Honorio and Tommi Jaakkola. Tight bounds for the expected risk of linear classifiers and PAC-Bayes finite-sample guarantees. In *International Conference on Artificial Intelligence and Statistics*, volume 33, pages 384–392, 2014.

Ziyi Huang, Henry Lam, Amirhossein Meisami, and Haofeng Zhang. Generalized Bayesian upper confidence bound with approximate inference for bandit problems. *arXiv preprint arXiv:2201.12955*, 2022.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Junpei Komiyama, Kaito Ariu, Masahiro Kato, and Chao Qin. Optimal simple regret in bayesian best arm identification. *arXiv preprint arXiv:2111.09885*, 2021.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *International Conference on Artificial Intelligence and Statistics*, volume 38, pages 535–543, 2015.

Branislav Kveton, Csaba Szepesvari, Sharan Vaswani, Zheng Wen, Tor Lattimore, and Mohammad Ghavamzadeh. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 3601–3610, 2019.

Branislav Kveton, Csaba Szepesvári, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic linear bandits. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115, pages 530–540, 2020.

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

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

Xiuyuan Lu and Benjamin Van Roy. Ensemble sampling. *Advances in Neural Information Processing Systems*, 30, 2017.

Xiuyuan Lu, Zheng Wen, and Branislav Kveton. Efficient online recommendation via low-rank ensemble sampling. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 460–464, 2018.

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

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. *Advances in Neural Information Processing Systems*, 29, 2016.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Ian Osband, Benjamin Van Roy, Daniel J. Russo, and Zheng Wen. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62, 2019.

Ian Osband, Zheng Wen, Mohammad Asghari, Morteza Ibrahimi, Xiyuan Lu, and Benjamin Van Roy. Epistemic neural networks. *arXiv preprint arXiv:2107.08924*, 2022a.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Botao Hao, Morteza Ibrahimi, Dieterich Lawson, Xiuyuan Lu, Brendan O'Donoghue, and Benjamin Van Roy. The neural testbed: Evaluating predictive distributions. *arXiv preprint arXiv:2110.04629*, 2022b.

My Phan, Yasin Abbasi Yadkori, and Justin Domke. Thompson sampling and approximate inference. *Advances in Neural Information Processing Systems*, 32, 2019.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

Daniel Russo. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418, 2016.

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

Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer series in statistics. Springer, 2009.

Iñigo Urteaga and Chris Wiggins. Variational inference for the multi-armed contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, volume 84, pages 698–706, 2018.

Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *International Conference on Machine Learning*, volume 37, pages 1113–1122, 2015.

Zheng Wen, Ian Osband, Chao Qin, Xiuyuan Lu, Morteza Ibrahimi, Vikranth Dwaracherla, Mohammad Asghari, and Benjamin Van Roy. From predictions to decisions: The importance of joint predictive distributions. *arXiv preprint arXiv:2107.09224*, 2022.

Jeremy Yang, Dean Eckles, Paramveer Dhillon, and Sinan Aral. Targeting for long-term outcomes. *arXiv preprint arXiv:2110.15835*, 2020.

Tong Yu, Branislav Kveton, Zheng Wen, Ruiyi Zhang, and Ole J. Mengshoel. Graphical models meet bandits: A variational Thompson sampling approach. In *International Conference on Machine Learning*, volume 119, pages 10902–10912, 2020.

Ruiyi Zhang, Zheng Wen, Changyou Chen, Chen Fang, Tong Yu, and Lawrence Carin. Scalable Thompson sampling via optimal transport. In *International Conference on Artificial Intelligence and Statistics*, volume 89, pages 87–96, 2019.

Zheqing Zhu and Benjamin Van Roy. Deep exploration for recommendation systems. *arXiv preprint arXiv:2109.12509*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] In this paper, we establish a Bayesian regret bound for ES in linear bandits (Theorem 1). This is based on a novel general regret bound (Theorem 2). These contributions have been accurately reflected in the abstract.
   (b) Did you describe the limitations of your work? [Yes] In particular, we have discussed the limitations of this work and future directions in Section 8.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This is a theoretical paper. We do not see any potential negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We have stated the assumptions for Theorem 1 and Theorem 2 clearly in the paper.
   (b) Did you include complete proofs of all theoretical results? [Yes] The complete proofs are provided in the appendices.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A] This paper does not include any experimental results.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendices

A Proofs of Lemmas 2 and 3 in Section 6

A.1 Proof of Lemma 2

Fix t. Recall the probability distributions $\overline{p}_t$ and p_t over action set $\mathcal{A}$:

$$\overline{p}_t(\cdot) = \mathbb{P}_t(A_t = \cdot) \quad \text{and} \quad p_t(\cdot) = \mathbb{P}_t(A_* = \cdot).$$

We first rewrite $\mathbb{I}_t\left(A_*; (A_t, R_{t+1,A_t})\right)$ as follows:

$$\begin{aligned}
\mathbb{I}_t(A_*; (A_t, R_{t+1,A_t})) &= \mathbb{I}_t(A_*; A_t) + \mathbb{I}_t(A_*; R_{t+1,A_t}|A_t) \\
&= \mathbb{I}_t(A_*; R_{t+1,A_t}|A_t) \\
&= \sum_{a\in\mathcal{A}} \overline{p}_t(a)\mathbb{I}_t(A_*; R_{t+1,A_t}|A_t = a) \\
&= \sum_{a\in\mathcal{A}} \overline{p}_t(a)\mathbb{I}_t(A_*; R_{t+1,a}) \\
&= \sum_{a\in\mathcal{A}} \overline{p}_t(a) \left(\sum_{a_*\in\mathcal{A}} p_t(a_*)\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}_t(R_{t+1,a} \in \cdot|A_* = a_*)\|\mathbb{P}_t(R_{t+1,a} \in \cdot)\right)\right) \\
&= \sum_{a,a_*\in\mathcal{A}} \overline{p}_t(a)p_t(a_*)\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}_t(R_{t+1,a} \in \cdot|A_* = a_*)\|\mathbb{P}_t(R_{t+1,a} \in \cdot)\right)
\end{aligned}$$

where the first equality uses the chain rule of mutual information (Fact 3); the second and fourth equalities follow from that conditioning on the history H_t, A_t is jointly independent of A_* and $R_{t+1} = (R_{t+1,1}, \ldots, R_{t+1,K})$; the fifth equality follows from the KL divergence form of mutual information (Fact 4).

For notational convenience, we write $\mathcal{A} = \{a_1, \ldots, a_K\}$. Define $M \in \mathbb{R}^{K\times K}$ by

$$M_{i,j} = \sqrt{\overline{p}_t(a_i)p_t(a_j)}\left(\mathbb{E}_t[R_{t+1,a_i}|A_* = a_j] - \mathbb{E}_t[R_{t+1,a_i}]\right), \quad \forall i, j \in [K].$$

For each $a_i \in \mathcal{A}$, conditional on H_t, the coefficient vector $\theta \sim N(\mu_t, \Sigma_t)$, and then

$$R_{t+1,a_i} = a_i^\top\theta + W_{t,a_i} \sim N\left(a_i^\top\mu_t, a_i^\top\Sigma_t a_i + \sigma^2\right).$$

The updating rule on covariance matrix in Equation (3) implies $\Sigma_t \preceq \Sigma_0$, and thus

$$a_i^\top\Sigma_t a_i \le a_i^\top\Sigma_0 a_i \le \max_{a\in\mathcal{A}} a^\top\Sigma_0 a.$$

Hence, for each $a_i \in \mathcal{A}$, conditional on H_t R_{t+1,a_i} is sub-Gaussian with variance proxy

$$\max_{a\in\mathcal{A}} a^\top\Sigma_0 a + \sigma^2.$$

Then by Lemma 9, we have

$$\begin{aligned}
\mathbb{I}_t(A_*; (A_t, R_{t+1,A_t})) &= \sum_{i,j\in[K]} \overline{p}_t(a_i)p_t(a_j)\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}_t(R_{t+1,a_i} \in \cdot|A_* = a_j)\|\mathbb{P}_t(R_{t+1,a_i} \in \cdot)\right) \\
&\ge \frac{1}{2\left(\max_{a\in\mathcal{A}} a^\top\Sigma_0 a + \sigma^2\right)} \sum_{i,j\in[K]} \overline{p}_t(a_i)p_t(a_j)\left(\mathbb{E}_t[R_{t+1,a_i}|A_* = a_j] - \mathbb{E}_t[R_{t+1,a_i}]\right)^2 \\
&= \frac{1}{2\left(\max_{a\in\mathcal{A}} a^\top\Sigma_0 a + \sigma^2\right)} \sum_{i,j\in[K]} M_{i,j}^2 \\
&= \frac{1}{2\left(\max_{a\in\mathcal{A}} a^\top\Sigma_0 a + \sigma^2\right)} \|M\|_{\mathrm{F}}^2.
\end{aligned}$$

On the other hand, G_t defined in Lemma 1 can be written as

$$G_t = \sum_{i\in[K]} \sqrt{\overline{p}_t(a_i)p_t(a_i)}\left(\mathbb{E}_t[R_{t+1,a_i}|A_* = a_i] - \mathbb{E}_t[R_{t+1,a_i}]\right) = \mathrm{Trace(M)}.$$

This shows, by Fact 2, that

$$\frac{G_t^2}{\mathbb{I}_t(A_*;(A_t,R_{t+1,A_t}))} \le \frac{2\left(\max_{a\in\mathcal{A}} a^\top\Sigma_0 a + \sigma^2\right)\mathrm{Trace(M)}^2}{\|M\|_{\mathrm{F}}^2} \le 2\left(\max_{a\in\mathcal{A}} a^\top\Sigma_0 a + \sigma^2\right)\mathrm{Rank}(M).$$

We now show $\mathrm{Rank}(M) \le d$. Define

$$\nu_j = \mathbb{E}_t\left[\theta|A_* = a_j\right], \quad \forall j\in[K] \quad \text{and} \quad \nu = \mathbb{E}_t\left[\theta\right],$$

and then for any $i,j\in[K]$,

$$\begin{aligned}
M_{i,j} &= \sqrt{\overline{p}_t(a_i)p_t(a_j)}\left(\mathbb{E}_t[R_{t+1,a_i}|A_* = a_j] - \mathbb{E}_t[R_{t+1,a_i}]\right)\\
&= \sqrt{\overline{p}_t(a_i)p_t(a_j)}\left(\mathbb{E}_t\left[\theta^\top a_i|A_* = a_j\right] - \mathbb{E}_t\left[\theta^\top a_i\right]\right)\\
&= \sqrt{\overline{p}_t(a_i)p_t(a_j)}\left(\nu_j - \nu\right)^\top a_i,
\end{aligned}$$

which implies

$$M = \begin{bmatrix} \sqrt{p_t(a_1)}\left(\nu_1-\nu\right)^\top \\ \vdots \\ \sqrt{p_t(a_K)}\left(\nu_K-\nu\right)^\top \end{bmatrix} \begin{bmatrix} \sqrt{\overline{p}_t(a_1)}a_1 & \cdots & \sqrt{\overline{p}_t(a_K)}a_K \end{bmatrix}.$$

Since M is the product of one $K\times d$ matrix and the other $d\times K$ matrix, $\mathrm{Rank}(M)\le d$.

A.2 Proof of Lemma 3

$$\begin{aligned}
\sum_{t=0}^{T-1}\mathbb{E}[G_t] &= \sum_{t=0}^{T-1}\mathbb{E}\left[\sqrt{\Gamma_t\mathbb{I}_t\left(A_*,(A_t;R_{t+1,A_t})\right)}\right]\\
&\le \sqrt{2\sigma^2 d}\sum_{t=0}^{T-1}\mathbb{E}\left[\sqrt{\mathbb{I}_t\left(A_*;(A_t,R_{t+1,A_t})\right)}\right]\\
&\le \sqrt{2\sigma^2 d}\sqrt{T\mathbb{E}\left[\sum_{t=0}^{T-1}\mathbb{I}_t\left(A_*;(A_t,R_{t+1,A_t})\right)\right]}\\
&= \sqrt{2\sigma^2 dT}\sqrt{\sum_{t=0}^{T-1}\mathbb{I}\left(A_*;(A_t,R_{t+1,A_t})|H_t\right)}\\
&= \sqrt{2\sigma^2 dT}\sqrt{\mathbb{H}(A_*) - \mathbb{H}(A_*|H_T)}\\
&\le \sqrt{2\sigma^2 dT\mathbb{H}(A_*)}
\end{aligned}$$

where the first and second inequalities follow from Lemma 2 and Cauchy-Schwarz inequality, respectively, and the last equality applies the chain rule for mutual information (Fact 3).

B Technical lemmas and facts

B.1 Expectation of maximum of squares of sub-Gaussian variables

We upper bound the expectation of maximum of squares of sub-Gaussian variables.

Lemma 7. *For each $a\in[K]$, let X_a be a sub-Gaussian random variable with mean μ_a and variance proxy σ_a^2. We have*

$$\mathbb{E}\left[\max_{a\in[K]} X_a^2\right] \le (4\log K + 5)\max_{a\in[K]}\sigma_a^2 + \max_{a\in[K]}\mu_a^2.$$

Proof. We begin by bounding the exponentiated quantity with the moment generating function. When $0 < t \leq \frac{1}{4\max_{a\in[K]}\sigma_a^2}$,

$$\begin{aligned}
\exp\left(t \cdot \mathbb{E}\left[\max_{a\in[K]} X_a^2\right]\right) &\leq \mathbb{E}\left[\exp\left(t \max_{a\in[K]} X_a^2\right)\right] \\
&= \mathbb{E}\left[\max_{a\in[K]} \exp(tX_a^2)\right] \\
&\leq \sum_{a\in[K]} \mathbb{E}[\exp(tX_a^2)] \\
&\leq \sum_{a\in[K]} \exp(16t^2\sigma_a^4 + t\mathbb{E}[X_a^2]) \\
&\leq K\exp\left(16t^2 \max_{a\in[K]} \sigma_a^4 + t \max_{a\in[K]} \mathbb{E}[X_a^2]\right)
\end{aligned}$$

where the first inequality uses Jensen's inequality and the second-to-last inequality applies Lemma 8 (Appendix B in Honorio and Jaakkola [2014]), which is stated after this proof. Taking the logarithm on both sides gives

$$\mathbb{E}\left[\max_{a\in[K]} X_a^2\right] \leq \frac{1}{t}\log K + 16t \max_{a\in[K]} \sigma_a^4 + \max_{a\in[K]} \mathbb{E}[X_a^2].$$

Then minimizing the RHS above over $0 < t \leq \frac{1}{4\max_{a\in[K]}\sigma_a^2}$, we have

$$\begin{aligned}
\mathbb{E}\left[\max_{a\in[K]} X_a^2\right] &\leq 4\left(\log K + 1\right) \max_{a\in[K]} \sigma_a^2 + \max_{a\in[K]} \mathbb{E}[X_a^2] \\
&\leq \left(4\log K + 5\right) \max_{a\in[K]} \sigma_a^2 + \max_{a\in[K]} \mu_a^2
\end{aligned}$$

where the last inequality uses

$$\max_{a\in[K]} \mathbb{E}[X_a^2] \leq \max_{a\in[K]} \left[\mathrm{Var}(X_a) + \mu_a^2\right] \leq \max_{a\in[K]} \sigma_a^2 + \max_{a\in[K]} \mu_a^2.$$

□

Lemma 8 (Appendix B in Honorio and Jaakkola [2014])**.** *Let X be a sub-Gaussian with mean μ and variance proxy σ^2. For $|t| \leq \frac{1}{4\sigma^2}$,*

$$\mathbb{E}\left[e^{t\left(X^2 - \mathbb{E}[X^2]\right)}\right] \leq e^{16t^2\sigma^4}.$$

B.2 Lemmas and facts from Russo and Van Roy [2016]

We present the lemmas and facts from Russo and Van Roy [2016] used in the proof of our Theorem 2.

Lemma 9 (Lemma 3 in Russo and Van Roy [2016])**.** *Suppose there is a H_t measurable random variable γ such that for each $a \in \mathcal{A}$, $R_{t+1,a}$ is γ sub-Gaussian conditional on H_t. Then for any $t \in [T]$ and $a, a_* \in \mathcal{A}$, with probability 1,*

$$\mathbb{E}_t[R_{t+1,a}|A_* = a_*] - \mathbb{E}_t[R_{t+1,a}] \leq \gamma\sqrt{2\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}_t(R_{t+1,a} \in \cdot|A_* = a_*)\|\mathbb{P}_t(R_{t+1,a} \in \cdot)\right)}.$$

Fact 2 (Fact 10 in Russo and Van Roy [2016])**.** *For any matrix $M \in \mathbb{R}^{d\times d}$,*

$$\mathrm{Trace}\left(M\right) \leq \sqrt{\mathrm{Rank}(M)}\|M\|_{\mathrm{F}}$$

where $\|M\|_{\mathrm{F}}$ is the Frobenius norm of M.

Fact 3 (Fact 5 in Russo and Van Roy [2016]: chain rule of mutual information)**.** *The mutual information between a random variable X and a collection of random variables $(Z_1, \ldots, Z_T)$ can be written as,*

$$\mathbb{I}\left(X; (Z_1, \ldots, Z_T)\right) = \mathbb{I}\left(X; Z_1\right) + \mathbb{I}\left(X; Z_2|\, Z_1\right) + \cdots + \mathbb{I}\left(X; Z_T|\, Z_1, \ldots, Z_T\right).$$

Fact 4 (Fact 6 in Russo and Van Roy [2016]: KL divergence form of mutual information)**.** *The mutual information between random variables X and Y can be written as,*

$$\begin{aligned}
\mathbb{I}\left(X; Y\right) &= \mathbb{E}_X\left[\mathbf{d}_{\mathrm{KL}}\left(P(Y|X) \,||\, P(Y)\right)\right] \\
&= \sum_{x\in\mathcal{X}} \mathbb{P}(X = x)\mathbf{d}_{\mathrm{KL}}\left(P(Y|X = x) \,||\, P(Y)\right).
\end{aligned}$$