Variational inference for the multi-armed contextual bandit

Iñigo Urteaga and Chris H. Wiggins
{inigo.urteaga, chris.wiggins}@columbia.edu

Department of Applied Physics and Applied Mathematics
Data Science Institute
Columbia University
New York City, NY 10027

May 3, 2021

Abstract

In many biomedical, science, and engineering problems, one must sequentially decide which action to take next so as to maximize rewards. One general class of algorithms for optimizing interactions with the world, while simultaneously learning how the world operates, is the multi-armed bandit setting and, in particular, the contextual bandit case. In this setting, for each executed action, one observes rewards that are dependent on a given 'context', available at each interaction with the world. The Thompson sampling algorithm has recently been shown to enjoy provable optimality properties for this set of problems, and to perform well in real-world settings. It facilitates generative and interpretable modeling of the problem at hand. Nevertheless, the design and complexity of the model limit its application, since one must both sample from the distributions modeled and calculate their expected rewards. We here show how these limitations can be overcome using variational inference to approximate complex models, applying to the reinforcement learning case advances developed for the inference case in the machine learning community over the past two decades. We consider contextual multi-armed bandit applications where the true reward distribution is unknown and complex, which we approximate with a mixture model whose parameters are inferred via variational inference. We show how the proposed variational Thompson sampling approach is accurate in approximating the true distribution, and attains reduced regrets even with complex reward distributions. The proposed algorithm is valuable for practical scenarios where restrictive modeling assumptions are undesirable.

1 Introduction

Reinforcement learning is an area of machine learning that studies optimizing interactions with the world while simultaneously learning how the world operates. The multi-armed bandit problem (9, 21) is a natural abstraction for a wide variety of such real-world challenges that require learning while simultaneously maximizing rewards. The goal is to decide on a series of actions under uncertainty, where each action can depend on previous rewards, actions, and contexts, aiming at balancing exploration and exploitation.

The name “bandit” finds its origin in the playing strategy one must devise when facing a row of slot machines (i.e., which arms to play). The setting is more formally referred to as the theory of sequential decision processes. Its foundations in the field of statistics began with the work by Thompson (22, 23) and continued with the contributions by Robbins (16). Interest in sequential decision making has recently intensified in both academic and industrial communities. The publication of separate works by Chapelle and Li (7), and Scott (20) have shown its impact in the online content management industry. This renaissance period of the multi-armed bandit problem has both a practical aspect (11) and a theoretical one as well (1, 13, 19).

Interestingly, most of these works have orbited around one of the oldest heuristics that address the exploration-exploitation tradeoff, i.e., Thompson sampling. It has been empirically proven to perform satisfactorily, and to enjoy provable optimality properties, both for problems with and without context (1, 2, 3, 10, 17, 18).

In this work, we are interested in extending and improving Thompson sampling. In its standard form, it is applicable to restricted models of the world, as one needs to sample from the corresponding parameter posteriors and compute their expected rewards: see (19) for details. The issue is that, for many problems of practical interest, one has partial (or no) knowledge about the ground truth, and the available models might be misspecified.

We aim at extending Thompson sampling to allow for more complex and flexible reward distributions. We target a richer class of bandits than in the most recent literature, where the posterior is usually assumed to be from the exponential family of distributions (10).

We model the convoluted relationship between the observed variables (rewards), and the unknown parameters governing the underlying process by mixture models, a large hypothesis space which for many components can accurately approximate any continuous reward distribution. The main challenge is how to learn such a mixture distribution within the contextual multi-armed bandit setting.

To that end, we leverage the advances developed for statistical inference in the last decades, and propose a variational approximation to the underlying true distribution of the environment with which one interacts. Variational inference is a principled framework, with roots in statistical physics and widely applied in the machine learning community (5).

Approximation of Bayesian models by variational inference has already attracted interest within the reinforcement learning community, e.g., to learn a probability distribution on the weights of a neural network (6). Thompson sampling has also been applied in the context of deep Q-networks, e.g., (12) and (15). Nevertheless, our focus here is (a) not on Q-learning but on bandit problems, and (b), the variational inference is for a hierarchical Bayesian mixture model approximation to the true reward distribution. We show that variational inference allows for Thompson sampling to be applicable for complex reward models.

Our contribution is unique to the contextual multi-armed bandit setting in that (a) we approximate unknown bandit reward functions with Gaussian mixture models, and (b) we provide variational mean-field parameter updates for the distribution that minimizes its divergence (in the Kullback-Leibler sense) to the mixture model reward approximation.

The proposed method autonomously learns, in the contextual bandit setting, the variational parameters of the mixture model that best approximates the true underlying reward distribution. It attains reduced cumulative regrets when operating under complex reward models, and is valuable when restrictive modeling assumptions are undesirable. To the best of our knowledge, no other work uses variational inference to address the contextual multi-armed bandit setting.

We formally introduce the contextual multi-armed bandit problem in Section 2, before providing a description of our proposed variational Thompson sampling method in Section 3. We evaluate its performance in Section 4, and we conclude with final remarks in Section 5.

2 Problem formulation

The contextual multi-armed bandit problem is formulated as follows. Let $a \in \{1, \cdots, A\}$ be any possible action to take (arms in the bandit), and $f_a(y|x, \theta)$ the stochastic reward distribution of each arm, dependent on its intrinsic properties (i.e., parameters θ) and context $x \in \mathbb{R}^d$. For every time instant t, the observed reward y_t is independently drawn from the reward distribution corresponding to the played arm, parameterized by θ and the applicable context; i.e., $y_t \sim f_a(y|x_t, \theta)$. We denote a set of given contexts, played arms, and observed rewards up to time instant t as $x_{1:t} \equiv (x_1, \cdots, x_t)$, $a_{1:t} \equiv (a_1, \cdots, a_t)$ and $y_{1:t} \equiv (y_1, \cdots, y_t)$, respectively.

In the contextual multi-armed bandit setting, one must decide which arm to play next (i.e., pick a_{t+1}), based on the context x_{t+1}, and previously observed rewards $y_{1:t}$, played arms $a_{1:t}$, and contexts $x_{1:t}$. The goal is to maximize the expected (cumulative) reward. We denote each arm's expected reward as $\mu_a(x, \theta) = \mathbb{E}_a\{y|x, \theta\}$.

When the properties of the arms (i.e., their parameters) are known, one can readily determine the optimal selection policy as soon as the context is given, i.e.,

$$a^*(x, \theta) = \underset{a}{\operatorname{argmax}}\, \mu_a(x, \theta) \; . \tag{1}$$

The challenge in the contextual multi-armed bandit problem is raised when there is a lack of knowledge about the model. The issue amounts to the need to learn about the key properties of the environment (i.e., the reward distribution), as one interacts with the world (i.e., takes actions sequentially).

Amongst the many alternatives to address this class of problems, the randomized probability matching is particularly appealing. In its simplest form, known as Thompson sampling, it has been shown to perform empirically well (7, 20) and has sound theoretical bounds, for both contextual and context-free problems (1, 2, 3). It plays each arm in proportion to its probability of being optimal, i.e.,

$$a_{t+1} \sim \Pr\left[a = a_{t+1}^* | a_{1:t}, x_{1:t+1}, y_{1:t}, \theta\right] \; . \tag{2}$$

If the parameters of the model are known, the above expression becomes deterministic, as one always picks the arm with the maximum expected reward

$$\Pr\left[a = a_{t+1}^* | a_{1:t}, x_{1:t+1}, y_{1:t}, \theta\right] = \Pr\left[a = a_{t+1}^* | x_{t+1}, \theta\right] = I_a(x_{t+1}, \theta) \; , \tag{3}$$

where we define the indicator function $I_a(\cdot)$ as

$$I_a(x, \theta) = \begin{cases} 1, \; \mu_a(x, \theta) = \max\{\mu_1(x, \theta), \cdots, \mu_A(x, \theta)\} \; , \\ 0, \; \text{otherwise} \; . \end{cases} \tag{4}$$

In practice, since the parameters of the model are unknown, one needs to explore ways of computing the probability of each arm being optimal. If the parameters are modeled as a set of random variables, then the uncertainty over the parameters can be accounted for.

Specifically, we marginalize over the posterior probability distribution of the parameters after observing rewards and actions up to time instant t, i.e.,

$$\begin{aligned}\Pr\left[a=a_{t+1}^*|a_{1:t},x_{1:t+1},y_{1:t}\right] &= \int f(a|a_{1:t},x_{1:t+1},y_{1:t},\theta)f(\theta|a_{1:t},x_{1:t},y_{1:t})\mathrm{d}\theta \\ &= \int I_a(x_{t+1},\theta)f(\theta|a_{1:t},x_{1:t},y_{1:t})\mathrm{d}\theta \ .\end{aligned} \tag{5}$$

In a Bayesian setting, if the reward distribution is known, one would assign a prior over the parameters to compute the corresponding posterior $f(\theta|a_{1:t},x_{1:t},y_{1:t})$. The analytical solution to such posterior is available for a well known set of distributions (4). Nevertheless, when reward distributions beyond simple well known cases (e.g., Bernoulli, Gaussian, etc.) are considered, one must resort to approximations of the posterior.

In this work, we leverage variational inference to approximate such posteriors, which was founded within the discipline of statistical physics and has flourished over the past several decades in the machine learning community.

3 Proposed method

The learning process in the multi-armed bandit, as explained in the formulation of Section 2, requires updating the posterior of the reward model parameters at every time instant. For computation of $f(\theta|a_{1:t},x_{1:t},y_{1:t})$ in Eqn. (5), knowledge of the reward distribution is instrumental. Typically, bandit algorithms are applied to simple distributions for which sampling and calculating expectations are feasible (e.g., the exponential family (10)).

In this work, we study finite mixture models as reward functions of the multi-armed bandit. Mixture models allow for the statistical modeling of a wide variety of stochastic phenomena; e.g., Gaussian mixture models can approximate arbitrarily well any continuous distribution and thus, provide a useful parametric framework to model unknown distributional shapes (14). This flexibility comes at a cost, as learning the parameters of the mixture distribution becomes a challenge.

We here use and empirically validate variational inference to approximate underlying Gaussian mixture models in the contextual bandit case.

For the rest of the paper, we consider a mixture of K Gaussian distributions per arm $a\in\{1,\cdots,A\}$, where each of the Gaussians is linearly dependent on the shared context. Formally,

$$f_a(y|x,\pi_{a,k},w_{a,k},\sigma_{a,k}^2)=\sum_{k=1}^{K}\pi_{a,k}\,\mathcal{N}\left(y|x^\top w_{a,k},\sigma_{a,k}^2\right) \ , \tag{6}$$

with per-arm mixture weights $\pi_{a,k}\in[0,1]$, $\sum_{k=1}^{K}\pi_{a,k}=1$ and Gaussian sufficient statistics, $w_{a,k}\in\mathbb{R}^d$ and $\sigma_{a,k}^2\in\mathbb{R}^+$.

For our analysis, we incorporate an auxiliary mixture indicator variable z_a. These are 1-of-K encoded vectors, where $z_{a,k}=1$, if mixture k is active; $z_{a,k}=0$, otherwise.

One can now rewrite Eqn. (6) as

$$f_a(y|x,z_a,w_{a,k},\sigma_{a,k}^2)=\prod_{k=1}^{K}\mathcal{N}\left(y|x^\top w_{a,k},\sigma_{a,k}^2\right)^{z_{a,k}} \ , \tag{7}$$

where $z_a\sim\mathrm{Cat}\left(\pi_a\right)$.

We consider conjugate priors for the unknown parameters of the mixture distribution

$$\begin{aligned}
f(\pi_a|\gamma_{a,0}) &= \mathrm{Dir}\left(\pi_{\mathrm{a}}|\gamma_{\mathrm{a},0}\right) \ , \\
f(w_{a,k}, \sigma_{a,k}^2|u_{a,k,0}, V_{a,k,0}, \alpha_{a,k,0}, \beta_{a,k,0}) &= \mathrm{NIG}\left(w_{a,k}, \sigma_{a,k}^2|u_{a,k,0}, V_{a,k,0}, \alpha_{a,k,0}, \beta_{a,k,0}\right) \\
&= \mathcal{N}\left(w_{a,k}|u_{a,k,0}, \sigma_{a,k}^2 V_{a,k,0}\right) \Gamma^{-1}\left(\sigma_{a,k}^2|\alpha_{a,k,0}, \beta_{a,k,0}\right) .
\end{aligned} \tag{8}$$

Given a set of contexts $x_{1:t}$, played arms $a_{1:t}$, mixture assignments $z_{a,1:t}$, and observed rewards $y_{1:t}$, the joint distribution of the model follows

$$\begin{aligned}
f(y_{1:t}, z_{a,1:t}, w_{a,k}, \sigma_{a,k}^2|a_{1:t}, x_{1:t}) = f(y_{1:t}|a_{1:t}, x_{1:t}, z_{a,1:t}, w_{a,k}, \sigma_{a,k}^2) \cdot f(z_{a,1:t}|\pi_a) \\
\cdot f(\pi_a|\gamma_{a,0}) \cdot f(w_{a,k}, \sigma_{a,k}^2|u_{a,k,0}, V_{a,k,0}, \alpha_{a,k,0}, \beta_{a,k,0}) \ ,
\end{aligned} \tag{9}$$

with

$$\begin{aligned}
f(y_{1:t}|a_{1:t}, x_{1:t}, z_{a,1:t}, w_{a,k}, \sigma_{a,k}^2) &= \prod_t \prod_k \mathcal{N}\left(y_t|x_t^\top w_{a,k}, \sigma_{a,k}^2\right)^{z_{a,k,t}} \ , \\
f(z_{1:t}|a_{1:t}, \pi_a) &= \prod_t \prod_k \pi_{a,k}^{z_{a,k,t}} ,
\end{aligned} \tag{10}$$

and parameter priors as in Eqn. (8).

3.1 Variational parameter inference

For the model as described above, the true joint posterior distribution is intractable. Under the variational framework, we consider instead a restricted family of distributions, and find the one that is a locally optimal approximation to the full posterior.

We do so by minimizing the Kullback-Leibler divergence between the true distribution $f(\cdot)$, and our approximating distribution $q(\cdot)$. We here consider a set of parameterized distributions with the following mean-field factorization over the variables of interest

$$q(Z, \pi, w, \sigma^2) = q(Z) \prod_{a=1}^{A} q(\pi_a) \prod_{k=1}^{K} q(w_{a,k}, \sigma_{a,k}^2) \ , \tag{11}$$

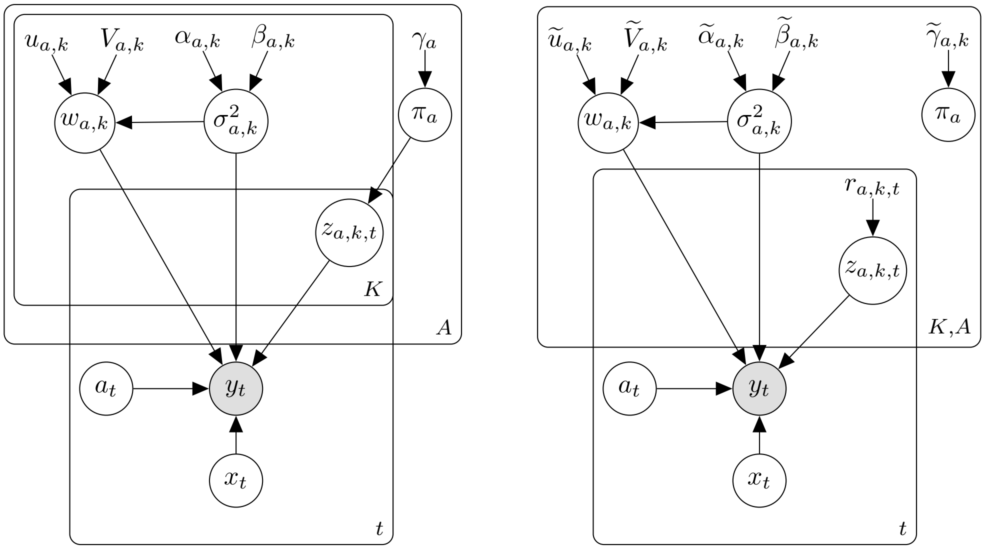
where we introduce notation $Z = \{z_{a,k,t}\}$, $\forall a, k, t$, for all latent variables; and similarly $\pi = \{\pi_{a,k}\}$, $\forall a, k$; $w = \{w_{a,k}\}$, $\forall a, k$; and $\sigma^2 = \{\sigma_{a,k}^2\}$, $\forall a, k$; for parameters. We illustrate the graphical model of the true and the variational bandit distributions in Fig. 1.

We place no restriction on the functional form of each distributional factor, and we seek to optimize the Kullback-Leibler divergence between this and the true distribution.

The optimal solution for each variational factor in the distribution in Eqn. (11) is obtained by computing the expectation of the log-joint true distribution with respect to the rest of the variational factor distributions, as explained in (5).

In our setting, we compute

$$\begin{aligned}
\ln q(Z) &= \mathbb{E}\left\{\ln\left[f(y_{1:t}, Z, w, \sigma|a_{1:t}, x_{1:t})\right]\right\}_{\pi,w,\sigma} + c \ , \\
\ln q(\pi_a) &= \mathbb{E}\left\{\ln\left[f(y_{1:t}, Z, w, \sigma|a_{1:t}, x_{1:t})\right]\right\}_{Z,w,\sigma} + c \ , \\
\ln q(w_{a,k}, \sigma_{a,k}^2) &= \mathbb{E}\left\{\ln\left[f(y_{1:t}, Z, w, \sigma|a_{1:t}, x_{1:t})\right]\right\}_{Z,\pi} + c \ .
\end{aligned} \tag{12}$$



(a) True contextual bandit distribution.

(b) Variational contextual bandit distribution.

Figure 1: Graphical models of the bandit distribution.

The resulting solution to the variational parameters that minimize the divergence iterates over the following two steps:

1. Given the current variational parameters, compute the responsibilities

$$\begin{aligned}\log(r_{a,k,t}) = &-\frac{1}{2}\left[\ln\left(\widetilde{\beta}_{a,k}\right) - \psi\left(\widetilde{\alpha}_{a,k}\right)\right] - \frac{1}{2}\left[x_t^\top \widetilde{V}_{a,k} x_t + (y_t - x_t^\top \widetilde{u}_{a,k})^2 \frac{\widetilde{\alpha}_{a,k}}{\widetilde{\beta}_{a,k}}\right] \\ &+ \left[\psi(\widetilde{\gamma}_{a,k}) - \psi\left(\sum_{k=1}^{K} \widetilde{\gamma}_{a,k}\right)\right] + c \ ,\end{aligned} \tag{13}$$

   with $\sum_{k=1}^{K} r_{a,k,t} = 1$. These responsibilities correspond to the expected value of assignments, i.e., $r_{a,k,t} = \mathbb{E}\left\{z_{a,k,t}\right\}_Z$.

2. Given the current responsibilities, we define $R_{a,k} \in \mathbb{R}^{t \times t}$ as a sparse diagonal matrix with diagonal elements $\left[R_{a,k}\right]_{t,t'} = r_{a,k,t} \cdot \mathbb{1}[a_t = a]$, and update the variational parameters

$$\begin{aligned}\widetilde{\gamma}_{a,k} &= \gamma_{a,0} + \mathrm{tr}\left\{R_{a,k}\right\} \ , \\ \widetilde{V}_{a,k}^{-1} &= x_{1:t} R_{a,k} x_{1:t}^\top + V_{a,k,0}^{-1} \ , \\ \widetilde{u}_{a,k} &= \widetilde{V}_{a,k}\left(x_{1:t} R_{a,k} y_{1:t} + V_{a,k,0}^{-1} u_{a,k,0}\right) \ , \\ \widetilde{\alpha}_{a,k} &= \alpha_{a,k,0} + \frac{1}{2}\mathrm{tr}\left\{R_{a,k}\right\} \ , \\ \widetilde{\beta}_{a,k} &= \beta_{a,k,0} + \frac{1}{2}\left(y_{1:t}^\top R_{a,k} y_{1:t}\right) + \frac{1}{2}\left(u_{a,k,0}^\top V_{a,k,0}^{-1} u_{a,k,0} - \widetilde{u}_{a,k}^\top \widetilde{V}_{a,k}^{-1} \widetilde{u}_{a,k}\right) \ .\end{aligned} \tag{14}$$

Note that, for simplicity, we have considered the same number of mixtures per arm K. Nevertheless, the above expressions are readily generalizable to differing per-arm number of mixtures K_a, for $a \in \{1, \cdots, A\}$.

The iterative procedure presented above is repeated until a convergence criterion is met. Usually, one iterates until the optimization improvement is small (relative to some prespecified ϵ) or a maximum number of iterations is executed.

3.2 Variational Thompson sampling

We now describe our proposed variational Thompson sampling (VTS) technique for the multi-armed contextual bandit problem, which leverages the variational distribution in subsection 3.1 and implements a posterior sampling based policy (17).

In the multi-armed bandit setting, at any given time and based on the information available, one needs to decide which arm to play next. A randomized probability matching technique picks each arm based on its probability of being optimal. In its simplest form, known as Thompson sampling (23), instead of computing the integral in Eqn. (5), one draws a random parameter sample from the posterior, and then picks the action that maximizes the expected reward. That is,

$$a_{t+1}^* = \underset{a}{\operatorname{argmax}}\, \mu_a(x_{t+1}, \theta_{t+1}), \qquad \text{with} \qquad \theta_{t+1} \sim f(\theta | a_{1:t}, x_{1:t}, y_{1:t}). \tag{15}$$

In a pure Bayesian setting, one deals with simple models that allow for analytical computation (and sampling) of the posterior. Here, as we allow for more realistic and complex modeling of the world that may not result in closed-form posterior updates, we propose to sample the parameters from the variational approximating distributions computed in subsection 3.1.

We describe the proposed variational Thompson sampling technique in Algorithm 1, for a general Gaussian mixture model with context.

An instrumental step in the proposed algorithm is to compute the expected reward for each arm, i.e., $\mu_{a,t+1}$, for which we need both the per-arm and per-mixture parameters $\{w_{a,k}, \sigma_{a,k}^2\}$ and the mixture assignments $\pi_{a,k}$. To compute the expected reward for each arm, we propose to draw per-arm and per-mixture posterior parameters from their updated variational posteriors, i.e., $(w_{a,k,t}, \sigma_{a,k,t}^2) \sim q\left(w_{a,k}, \sigma_{a,k}^2 | \widetilde{\alpha}_{a,k}, \widetilde{\beta}_{a,k}, \widetilde{u}_{a,k}, \widetilde{V}_{a,k}\right)$, and consider the following mixture expectation alternatives:

1. Expectation with mixture assignment sampling

$$\mu_{a,t+1} = x_t^\top w_{a,z_{a,k,t},t}, \quad z_{a,k,t} \sim \text{Cat}\left(\frac{\widetilde{\gamma}_{a,k}}{\sum_{k=1}^K \widetilde{\gamma}_{a,k}}\right) . \tag{16}$$

2. Expectation with mixture proportion sampling

$$\mu_{a,t+1} = \sum_{k=1}^K \pi_{a,k,t} x_t^\top w_{a,k,t}, \quad \pi_{a,k,t} \sim \text{Dir}\left(\widetilde{\gamma}_{\text{a,k}}\right) . \tag{17}$$

3. Expectation with mixture proportions

$$\mu_{a,t+1} = \sum_{k=1}^K \pi_{a,k,t} x_t^\top w_{a,k,t}, \quad \pi_{a,k,t} = \frac{\widetilde{\gamma}_{a,k}}{\sum_{k=1}^K \widetilde{\gamma}_{a,k}} . \tag{18}$$

Algorithm 1 Variational Thompson sampling

```
Require: Model description $A$, $K_a$
Require: Parameters $\gamma_{a,0}$, $u_{a,k,0}$, $V_{a,k,0}$, $\alpha_{a,k,0}$, $\beta_{a,k,0}$
$D = \emptyset$
Initialize $\widetilde{\gamma}_{a,k} = \gamma_{a,0}$, $\widetilde{\alpha}_{a,k} = \alpha_{a,k,0}$, $\widetilde{\beta}_{a,k} = \beta_{a,k,0}$, $\widetilde{u}_{a,k} = u_{a,k,0}$, $\widetilde{V}_{a,k} = V_{a,k,0}$
for $t = 1, \cdots, T$ do
  Receive context $x_{t+1}$
  for $a = 1, \cdots, A$ do
    for $k = 1, \cdots, K_a$ do
      Draw new parameters $\theta_{a,k,t} := \{z_{a,k,t}, \pi_{a,k,t}, w_{a,k,t}, \sigma_{a,k,t}\}$
        $$\theta_{a,k,t} \sim q\left(z_{a,k}, \pi_{a,k}, w_{a,k}, \sigma_{a,k} \middle| \widetilde{\gamma}_{a,k}, \widetilde{\alpha}_{a,k}, \widetilde{\beta}_{a,k}, \widetilde{u}_{a,k}, \widetilde{V}_{a,k}\right)$$
    end for
    Compute $\mu_{a,t+1} = \mu_a(x_{t+1}, \theta_{a,t})$
  end for
  Play arm $a_{t+1} = \mathrm{argmax}_a \, \mu_{a,t+1}$
  Observe reward $y_{t+1}$
  $D = D \cup \{x_{t+1}, a_{t+1}, y_{t+1}\}$
  while NOT Variational convergence criteria do
    Compute $r_{a,k,t}$
    Update $\widetilde{\gamma}_{a,k}, \widetilde{\alpha}_{a,k}, \widetilde{\beta}_{a,k}, \widetilde{u}_{a,k}, \widetilde{V}_{a,k}$
  end while
end for
```

4 Evaluation

In this section, we evaluate the performance of the proposed variational Thompson sampling technique for the contextual multi-armed bandit problem.

We focus on two illustrative scenarios: the first, referred to as `Scenario A`, with per-arm reward distributions

$$\texttt{Scenario A} \begin{cases} f_0(y|x_t,\theta) = 0.5 \cdot \mathcal{N}\left(y|(0\;0)^\top x_t, 1\right) + 0.5 \cdot \mathcal{N}\left(y|(1\;1)^\top x_t, 1\right) , \\ f_1(y|x_t,\theta) = 0.5 \cdot \mathcal{N}\left(y|(2\;2)^\top x_t, 1\right) + 0.5 \cdot \mathcal{N}\left(y|(3\;3)^\top x_t, 1\right) \; , \end{cases} \tag{19}$$

and the second, `Scenario B`, with

$$\texttt{Scenario B} \begin{cases} f_0(y|x_t,\theta) = 0.5 \cdot \mathcal{N}\left(y|(1\;1)^\top x_t, 1\right) + 0.5 \cdot \mathcal{N}\left(y|(2\;2)^\top x_t, 1\right) \; , \\ f_1(y|x_t,\theta) = 0.3 \cdot \mathcal{N}\left(y|(0\;0)^\top x_t, 1\right) + 0.7 \cdot \mathcal{N}\left(y|(3\;3)^\top x_t, 1\right) \; . \end{cases} \tag{20}$$

The reward distributions of the contextual bandits in both scenarios are Gaussian mixtures with two context dependent components. These reward distributions are complex in that they are multimodal and, in `Scenario B`, unbalanced. Furthermore, they depend on a two dimensional uncorrelated uniform context, i.e., $x_{i,t} \sim \mathcal{U}\left(0,1\right)$, $i \in \{1,2\}$, $t \in \mathbb{N}$.

The key difference between the scenarios is the amount of mixture overlap and the similarity between arms. Recall the complexity of the reward distributions in `Scenario B`, with a significant overlap between arm rewards and the unbalanced nature of arm 1.

We evaluate variational Thompson sampling in terms of its cumulative regret, defined as

$$R_t = \sum_{\tau=0}^{t} \mathbb{E}\left\{(y^*_\tau - y_\tau)\right\} = \sum_{\tau=0}^{t} \mu^*_\tau - \bar{y}_\tau \; , \tag{21}$$

where for each time instant t, μ^*_t denotes the true expected reward of the optimal arm, and $\bar{y}_t$ the empirical mean of the observed rewards.

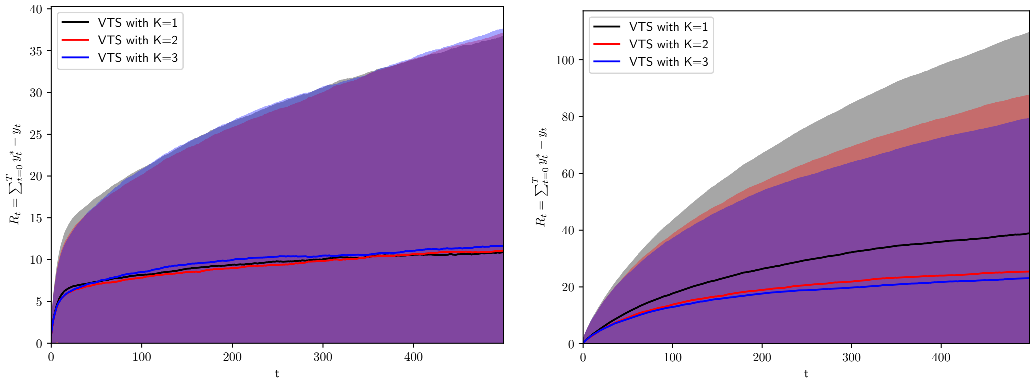
Since we have not noticed significant cumulative regret differences between the three approaches to computing the expected reward $\mu_{a,t+1}$ described in subsection 3.2, we avoid unnecessary clutter and do not plot them in the figures below. All reported values are averaged over 5000 realizations of the same set of parameters and context (with the standard deviation shown as the shaded region in the figures).

Fig. 2 shows the cumulative regret of the proposed variational Thompson sampling approach in both scenarios, when different assumptions for the variational approximating distribution are made (i.e., assumed number of components K).

Note that "*VTS with* $K=1$" is equivalent to a vanilla Thompson sampling approach with a linear contextual Gaussian model assumption. Since $r_{a,k=1,t}=1$ for all a and t, the variational update equations match the corresponding Bayesian posterior updates for Thompson sampling. We are thus effectively comparing the performance of the proposed method to the Thompson sampling benchmark, as in (3).

The main conclusion from the results shown in Fig. 2 is that inferring a variational approximation to the true complex reward distribution attains satisfactory regret performance.

For `Scenario A`, the regret performance of the proposed VTS with mixture of Gaussians is equivalent to "*VTS with* $K=1$" (i.e., vanilla Thompson sampling). On the contrary, for `Scenario B`, our flexible approach attains considerably lower regret. As in any posterior sampling bandit algorithm, the variance of the cumulative regret is large for all methods.



(a) `Scenario A`: cumulative regret.

(b) `Scenario B` cumulative regret.

Figure 2: Cumulative regret comparison.

Nevertheless, we observe a reduction in both mean regret and its variability for the proposed "*VTS with* $K = 2$ and $K = 3$" cases, in comparison to the contextual linear Gaussian Thompson sampling case (i.e., "*VTS with* $K = 1$"), for the challenging `Scenario B` illustrated in Fig. 2b.

In other words, a misspecified (and simplified) model performs worse than the proposed (more complex) alternatives. Precisely, the cumulative regret reduction of "*VTS with* $K = 2$" (which corresponds to the true underlying mixture distributions in Eqn. (20)) with respect to "*VTS with* $K = 1$" at $t = 500$ is of 35%. The issue of model misspecification is evident for `Scenario B`, as the linear Gaussian contextual model fails to capture the subtleties of the unbalanced mixtures of Eqn. (20).

In summary, with a simplistic model assumption as in "*VTS with* $K = 1$", one can not capture the properties of the underlying complex reward distributions and thus, can not make well-informed decisions. On the contrary, by considering more complex models (i.e., Gaussian mixture models), and by using variational inference for learning its parameters, the proposed technique attains reduced regret.

Furthermore, we highlight that even an overly complex model assumption does provide competitive performance. For both `Scenario A` and `Scenario B`, the regret of the variational approximation with $K = 3$ is similar to that of the true model assumption $K = 2$, ("*VTS with* $K = 3$" and "*VTS with* $K = 2$" in Fig. 2, respectively). For the challenging `Scenario B`, the cumulative regret reduction of "*VTS with* $K = 3$" with respect to the "*VTS with* $K = 1$" benchmark at $t = 500$ is of 40%.

The explanation relies on the flexibility provided by the variational machinery, as the learning process adjusts the parameters to minimize the divergence between the true and the variational distributions. Nonetheless, one must be aware that this flexibility comes with an additional computational cost, as more parameters need to be learned.

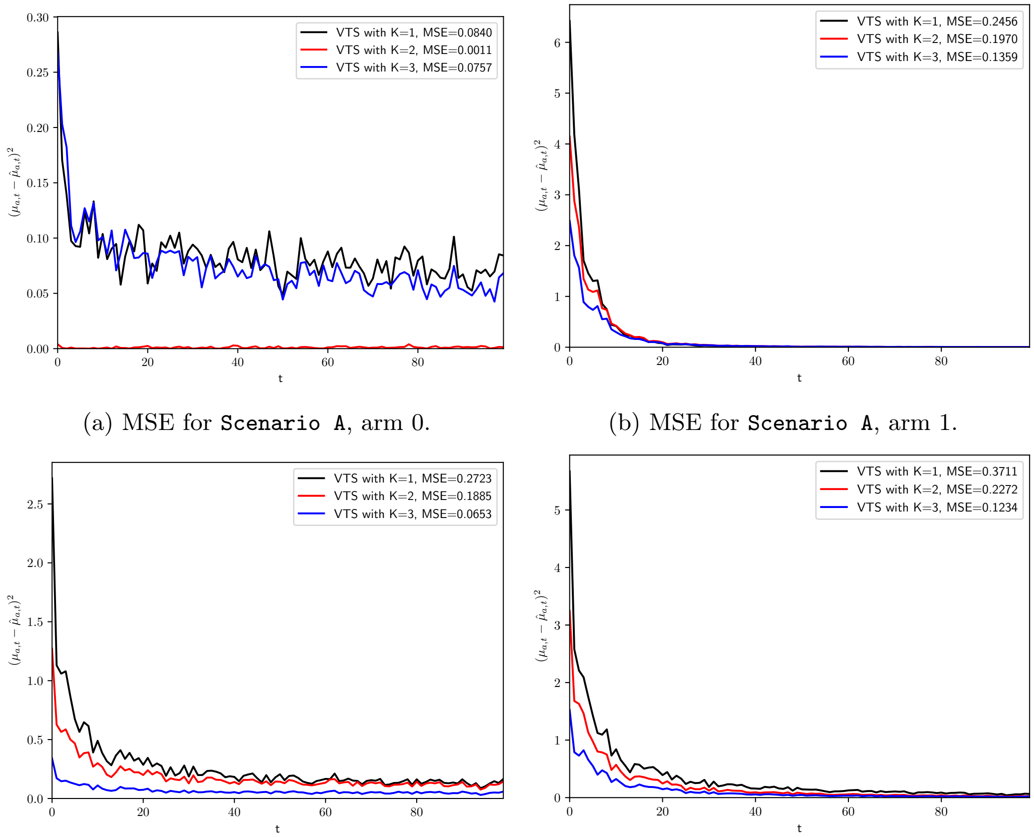
We further elaborate on the analysis of our proposed variational Thompson sampling method by studying its learning accuracy.

In bandit algorithms, the goal is to gather enough evidence to identify the best arm (in terms of expected reward), and this can only be achieved if the arm properties (i.e., the reward distributions) are learned accurately; their expectation being the most important sufficient statistic.

We illustrate in Fig. 3 the mean squared error of the variational per-arm expected reward estimation

$$MSE_a = \frac{1}{T} \sum_{t=0}^{T} \left(\mu_{a,t} - \hat{\mu}_{a,t}\right)^2 \ , \tag{22}$$

where $\hat{\mu}_{a,t}$ denotes the estimated expected reward for arm a at time t.



(a) MSE for `Scenario A`, arm 0.

(b) MSE for `Scenario A`, arm 1.

(c) MSE for `Scenario B`, arm 0.

(d) MSE for `Scenario B`, arm 1.

Figure 3: Expected reward estimation accuracy.

We show that the learning is faster and more accurate when the approximating mixture model has flexibility to adapt. That is, both "*VTS with* $K = 2$" and "*VTS with* $K = 3$" accurately estimate the expected reward of the best arm.

We once again recall the complexity of the model in `Scenario B` in comparison to that of `Scenario A`, and more importantly, its implications for a bandit algorithm. In Figs. 3a-3b, the simplest model that assumes a single Gaussian distribution ("*VTS with* $K = 1$") is able to quickly and accurately estimate the expected reward. In contrast, its estimation accuracy is the worst (as shown in Figs. 3c-3d) when facing a more complex model with overlapping and unbalanced arm rewards. Note how, for all results in Fig. 3, the most complex model (i.e., "*VTS with* $K = 3$") fits the expected reward best.

These observations reinforce our claims on the flexibility and applicability of the presented technique. By allowing for complex modeling of the world and using variational inference to learn it, the proposed variational Thompson sampling technique can provide improved performance (in the sense of regret) for the contextual multi-armed bandit problem.

5 Conclusion

We have presented variational Thompson sampling, a new algorithm for the contextual multi-armed bandit setting, where we combine variational inference machinery with a state of the art reinforcement learning technique. The proposed algorithm allows for interpretable bandit modeling with complex reward functions, learned from online data. We extend the applicability of Thompson sampling by accommodating more realistic and complex models of the world. Empirical results show a significant cumulative regret reduction when using the proposed algorithm in simulated models. A natural future application is to scenarios when relevant context (attributes of items, customers or patients) are unobservable, and thus the latent variables are truly 'incomplete' as in the motivating case for expectation maximization modeling (8).

5.1 Software and Data

The implementation of the proposed method is available in this public repository. It contains all the software required for replication of the findings of this study.

Acknowledgments

This research was supported in part by NSF grant SCH-1344668. We thank Shipra Agrawal, David Blei and Daniel J. Hsu for discussions that helped motivate this work. We also thank Hang Su, Edward Peng Yu, and all the reviewers for their feedback and comments.

References

[1] S. Agrawal and N. Goyal. Analysis of Thompson Sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012.

[2] S. Agrawal and N. Goyal. Further Optimal Regret Bounds for Thompson Sampling. In *Artificial Intelligence and Statistics*, pages 99–107, 2013.

[3] S. Agrawal and N. Goyal. Thompson Sampling for Contextual Bandits with Linear Payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

[4] J. M. Bernardo and A. F. Smith. *Bayesian Theory.* Wiley Series in Probability and Statistics. Wiley, 2009. ISBN 9780470317716. doi: 10.1002/9780470316870.

[5] C. Bishop. *Pattern Recognition and Machine Learning.* Information Science and Statistics. Springer-Verlag New York, 2006.

[6] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight Uncertainty in Neural Networks. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1613–1622. JMLR.org, 2015.

[7] O. Chapelle and L. Li. An Empirical Evaluation of Thompson Sampling. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2249–2257. Curran Associates, Inc., 2011. URL `https://papers.nips.cc/paper/4321-an-empirical-evaluation-of-thompson-sampling`.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[9] M. Ghavamzadeh, S. Mannor, J. Pineau, and A. Tamar. Bayesian Reinforcement Learning: A Survey. *Foundations and Trends® in Machine Learning*, 8(5-6):359–483, 2015. ISSN 1935-8237. doi: 10.1561/2200000049.

[10] N. Korda, E. Kaufmann, and R. Munos. Thompson Sampling for 1-Dimensional Exponential Family Bandits. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1448–1456. Curran Associates, Inc., 2013.

[11] L. Li, W. Chu, J. Langford, and R. E. Schapire. A Contextual-Bandit Approach to Personalized News Article Recommendation. In *Proceedings of the 19th international conference on World wide web*, volume abs/1003.0146, pages 661–670, 2010.

[12] Z. C. Lipton, X. Li, J. Gao, L. Li, F. Ahmed, and L. Deng. BBQ-Networks: Efficient Exploration in Deep Reinforcement Learning for Task-Oriented Dialogue Systems. In *AAAI*, 2018.

[13] O.-A. Maillard, R. Munos, and G. Stoltz. Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences. In *Conference On Learning Theory*, 2011.

[14] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2004, 2004. ISBN 9780471654063.

[15] I. Osband, C. Blundell, A. Pritzel, and B. V. Roy. Deep Exploration via Bootstrapped DQN. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4026–4034. Curran Associates, Inc., 2016.

[16] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, (58):527–535, 1952.

[17] D. Russo and B. V. Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[18] D. Russo and B. V. Roy. An information-theoretic analysis of Thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

[19] S. L. Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010. ISSN 1526-4025. doi: 10.1002/asmb.874.

[20] S. L. Scott. Multi-armed bandit experiments in the online service economy. *Applied Stochastic Models in Business and Industry*, 31:37–49, 2015. Special issue on actual impact and future perspectives on stochastic modelling in business and industry.

[21] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction.* MIT Press: Cambridge, MA, 1998.

[22] W. 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. ISSN 00063444.

[23] W. R. Thompson. On the Theory of Apportionment. *American Journal of Mathematics*, 57(2):450–456, 1935. ISSN 00029327, 10806377.