Variational inference for the multi-armed contextual bandit

Anonymous Author(s)
Affiliation
Address
email

Abstract

In many biomedical, science, and engineering problems, one must sequentially decide which action to take next so as to maximize rewards. Reinforcement learning is an area of machine learning that studies how this maximization balances exploration and exploitation, optimizing interactions with the world while simultaneously learning how the world operates. One general class of algorithms for this type of learning is the multi-armed bandit setting and, in particular, the contextual bandit case, in which observed rewards are dependent on each action as well as on given information or 'context' available at each interaction with the world. The Thompson sampling algorithm has recently been shown to perform well in real-world settings and to enjoy provable optimality properties for this set of problems. It facilitates generative and interpretable modeling of the problem at hand, though complexity of the model limits 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 approximations, applying to the reinforcement learning case advances developed for the inference case in the machine learning community over the past two decades.

1 Introduction

The multi-armed bandit problem (Sutton and Barto [1998], Ghavamzadeh et al. [2015]) is the natural abstraction for a wide variety of real-world challenges requiring learning while simultaneously maximizing reward. The goal is to decide on a series of actions under uncertainty, where each action can depend on previous rewards, actions, and contexts. Its name comes from the playing strategy one must devise when facing a row of slot machines (i.e., which arms to play), and 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 [1935, 1933] and continued with the contributions by Robbins [1952].

Interest in sequential decision making has recently intensified in both academic and industrial communities. The publication of separate works by Chapelle and Li [2011] and Scott [2015] have shown its impact in the online content management industry. This revival period has both a practical aspect (Li et al. [2010]) and a theoretical one as well (Scott [2010], Agrawal and Goyal [2011], Maillard et al. [2011]). Interestingly, most of these works have orbited around one of the oldest heuristics that address the exploration-exploitation trade-off, 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 (Agrawal and Goyal [2012a,b], Korda et al. [2013], Russo and Roy [2014, 2016]).

In this work, we are interested in extending and improving the Thompson sampling technique. Thompson sampling is applicable to restricted models of the world, as long as one can sample from the corresponding parameter posteriors and compute their expected rewards (see Scott [2010] for details). The challenge is that, for many problems of practical interest, one has partial (or no) information about the ground truth and the available models might be misspecified. In this work, we aim at extending Thompson sampling to allow for more complex and flexible reward distributions. 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 issue is how to learn such a mixture distribution within the contextual multi-armed bandit setting. We leverage the advances developed for the inference case in the last decades, and propose a variational approximation to the underlying true distribution of the environment with which one interacts. The proposed method autonomously learns the parameters of the mixture model that best approximates the true underlying reward distribution. Our contribution is unique to the 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. To the best of our knowledge, there is no other work in the literature that uses variational inference to tackle the contextual multi-armed bandit problem.

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 (arms in the bandit) and $f_a(y|x, \theta)$ the stochastic reward distribution of each arm, dependent on its 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., decide 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 this set of problems is raised when there is a lack of knowledge about the parameters. The issue amounts to the need to learn about the key properties of the environment (i.e., the parameters of 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 (Chapelle and Li [2011], Scott [2015]) and has sound theoretical bounds, for both contextual and context-free problems (Agrawal and Goyal [2012a,b]). This approach 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 are known, the above expression becomes deterministic, as one always picks the arm with the maximum expected reward

$$\begin{aligned} &\Pr\left[a = a_{t+1}^* | a_{1:t}, x_{1:t+1}, y_{1:t}, \theta\right] = \Pr\left[a = a_{t+1}^* | x_{t+1}, \theta\right] = I_a(x_{t+1}, \theta) \ , \\ &\qquad \text{with} \ \ I_a(x, \theta) = \begin{cases} 1, \ \mu_a(x, \theta) = \max\{\mu_1(x, \theta), \cdots, \mu_A(x, \theta)\} \ , \\ 0, \ \text{otherwise} \ . \end{cases} \end{aligned} \tag{3}$$

When the parameters are unknown, one needs to explore ways of computing Eqn. 3. If we model the parameters as a set of random variables, then the uncertainty over the parameters can be accounted for. Specifically, we marginalize over their probability distribution after observing rewards and actions up to time instant t, i.e.,

$$\begin{aligned}\Pr\left[a = a_{t+1}^* | a_{1:t}, x_{1:t+1}, y_{1:t}\right] &= \int f(a|a_{1:t}, x_{1:t+1}, y_{1:t}, \theta) f(\theta|a_{1:t}, x_{1:t}, y_{1:t}) \mathrm{d}\theta \\ &= \int I_a(x_{t+1}, \theta) f(\theta|a_{1:t}, x_{1:t}, y_{1:t}) \mathrm{d}\theta \ .\end{aligned} \tag{4}$$

In a Bayesian setting, if the reward distribution is known, one would assign a prior over the parameters to compute the corresponding posterior. The analytical solution to such posteriors is available for a well known set of distributions (Bernardo and Smith [2009]). 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 the variational inference methodology, which has flourished in the inference case over the past several decades, to approximate such posteriors.

3 Proposed method

The learning process, as explained in the formulation of Section 2, requires updating the posterior of the parameters at every time instant. For computation of $f(\theta|a_{1:t}, x_{1:t}, y_{1:t})$, knowledge of the reward distribution is instrumental. Broad application of the method is typically limited by the simple distributions for which sampling and calculating expectations are feasible.

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 (McLachlan and Peel [2004]). This flexibility comes at a cost, as learning the parameters of the mixture distribution becomes a challenge. In this work, we 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 $a = \{1, \cdots, A\}$, where each of the Gaussians is linearly dependent on the shared context. Formally,

$$f_a(y|x, \pi_{a,k}, w_{a,k}, \sigma_{a,k}^2) = \sum_{k=1}^{K} \pi_{a,k} \, \mathcal{N}(y|x^\top w_{a,k}, \sigma_{a,k}^2) \ , \ \text{with} \begin{cases} \pi_{a,k} \in [0,1], \ \sum_{k=1}^K \pi_{a,k} = 1 \ , \\ w_{a,k} \in \mathbb{R}^d \ , \\ \sigma_{a,k}^2 \in \mathbb{R}^+ \ . \end{cases} \tag{5}$$

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. 5 as

$$f_a(y|x, z_a, w_{a,k}, \sigma_{a,k}^2) = \prod_{k=1}^{K} \mathcal{N}(y|x^\top w_{a,k}, \sigma_{a,k}^2)^{z_{a,k}} \ , \quad \text{with } z_a \sim \mathrm{Cat}(\pi_a) \ . \tag{6}$$

Since the parameters of the mixture distribution are unknown, we consider their conjugate priors

$$\begin{aligned} f(\pi_a|\gamma_{a,0}) &= \mathrm{Dir}(\pi_a|\gamma_{a,0}) \ , \\ f(w_{a,k}, \sigma_{a,k}^2 | u_{a,k,0}, V_{a,k,0}, \alpha_{a,k,0}, \beta_{a,k,0}) &= \mathrm{NIG}(w_{a,k}, \sigma_{a,k}^2 | u_{a,k,0}, V_{a,k,0}, \alpha_{a,k,0}, \beta_{a,k,0}) \\ &= \mathcal{N}(w_{a,k}|u_{a,k,0}, \sigma_{a,k}^2 V_{a,k,0}) \Gamma^{-1}(\sigma_{a,k}^2|\alpha_{a,k,0}, \beta_{a,k,0}). \end{aligned} \tag{7}$$

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 follows

$$\begin{aligned} f(y_{1:t}, z_{a,1:t}, w_{a,k}, \sigma_{a,k}^2 | a_{1:t}, x_{1:t}) =& f(y_{1:t}|a_{1:t}, x_{1:t}, z_{a,1:t}, w_{a,k}, \sigma_{a,k}^2) f(z_{a,1:t}|\pi_a) \\ & f(\pi_a|\gamma_{a,0}) f(w_{a,k}|u_{a,k,0}, \sigma_{a,k}^2, V_{a,k,0}) f(\sigma_{a,k}^2|\alpha_{a,k,0}, \beta_{a,k,0}) \ , \end{aligned} \tag{8}$$

with

$$
\begin{aligned}
f(y_{1:t}|a_{1:t}, x_{1:t}, z_{a,1:t}, w_{a,k}, \sigma_{a,k}^2) &= \prod_t \prod_k \mathcal{N}(y_t|x_t^\top w_{a,k}, \sigma_{a,k}^2)^{z_{a,k,t}} \ , \\
f(z_{1:t}|a_{1:t}, \pi_a) &= \prod_t \prod_k \pi_{a,k}^{z_{a,k,t}} \ ,
\end{aligned}
\tag{9}
$$

and parameter priors as in Eqn. 7.

3.1 Variational approximation to the parameter posterior

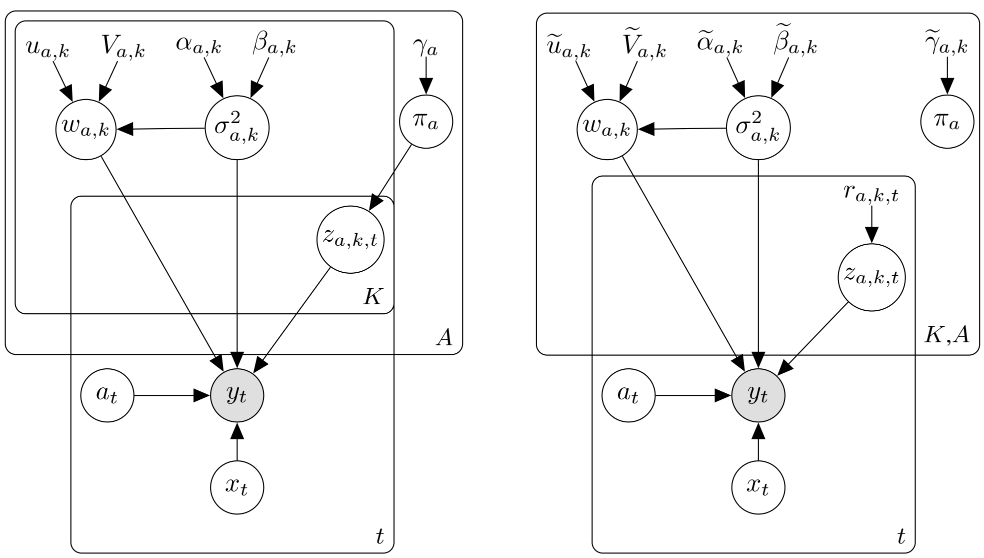
For the model as described above, the true joint posterior distribution is intractable. Under the variational framework, we consider instead a restricted family of distributions and find the one that is a locally optimal approximation to the full posterior.

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

$$
q(Z, \pi, w, \sigma^2) = q(Z) \prod_{a=1}^{A} q(\pi_a) \prod_{k=1}^{K} q(w_{a,k}, \sigma_{a,k}^2) \ , \tag{10}
$$

where we introduce notation $Z = \{z_{a,k,t}\}, \forall a, k, t$; $\pi = \{\pi_{a,k}\}, \forall a, k$; $w = \{w_{a,k}\}, \forall a, k$; and $\sigma^2 = \{\sigma_{a,k}^2\}, \forall a, k$. 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.

We illustrate the graphical model of the true and the variational bandit distributions in Fig. 1.



(a) True contextual bandit distribution.

(b) Approximating contextual bandit distribution.

Figure 1: Graphical models of the bandit distribution.

The optimal solution for each variational factor in the distribution in Eqn. 10 is obtained by computing the expectation of the log-joint true distribution with respect to the rest of the variational factor distributions (Bishop [2006]). In our setting, we compute

$$
\begin{cases}
\ln q(Z) = \mathbb{E}\left\{\ln\left[f(y_{1:t}, Z, w, \sigma|a_{1:t}, x_{1:t})\right]\right\}_{\pi, w, \sigma} + c \ , \\
\ln q(\pi_a) = \mathbb{E}\left\{\ln\left[f(y_{1:t}, Z, w, \sigma|a_{1:t}, x_{1:t})\right]\right\}_{Z, w, \sigma} + c \ , \\
\ln q(w_{a,k}, \sigma_{a,k}^2) = \mathbb{E}\left\{\ln\left[f(y_{1:t}, Z, w, \sigma|a_{1:t}, x_{1:t})\right]\right\}_{Z, \pi} + c \ .
\end{cases}
\tag{11}
$$

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

1. Given the current variational parameters, compute the responsibilities

$$\log(r_{a,k,t}) = -\frac{1}{2}\left[\ln\left(\widetilde{\beta}_{a,k}\right) - \psi\left(\widetilde{\alpha}_{a,k}\right)\right] - \frac{1}{2}\left[x_t^\top \widetilde{V}_{a,k} x_t + (y_t - x_t^\top \widetilde{u}_{a,k})^2 \frac{\widetilde{\alpha}_{a,k}}{\widetilde{\beta}_{a,k}}\right] + \left[\psi(\widetilde{\gamma}_{a,k}) - \psi\left(\sum_{k=1}^{K}\widetilde{\gamma}_{a,k}\right)\right] + c \; , \tag{12}$$

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

$$\begin{cases} \widetilde{\gamma}_{a,k} = \gamma_{a,0} + \mathrm{tr}\{R_{a,k}\} \; , \\ \widetilde{V}_{a,k}^{-1} = x_{1:t} R_{a,k} x_{1:t}^\top + V_{a,k,0}^{-1} \; , \\ \widetilde{u}_{a,k} = \widetilde{V}_{a,k}\left(x_{1:t} R_{a,k} y_{1:t} + V_{a,k,0}^{-1} u_{a,k,0}\right) \; , \\ \widetilde{\alpha}_{a,k} = \alpha_{a,k,0} + \frac{1}{2}\mathrm{tr}\{R_{a,k}\} \; , \\ \widetilde{\beta}_{a,k} = \beta_{a,k,0} + \frac{1}{2}\left(y_{1:t}^\top R_{a,k} y_{1:t} + u_{a,k,0}^\top V_{a,k,0}^{-1} u_{a,k,0} - \widetilde{u}_{a,k}^\top \widetilde{V}_{a,k}^{-1} \widetilde{u}_{a,k}\right) \; . \end{cases} \tag{13}$$

The above iterative process 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 met.

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

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 sampling based policy.

In the multi-armed bandit, at any given time instant and based on the information available, one needs to decide which arm to play next. A randomized probability matching technique picks the arm that has the highest probability of being optimal. In its simplest form, known as Thompson sampling (Thompson [1935]), instead of computing the integral in Eqn. 4, one draws a random parameter sample from the posterior and then picks the action that maximizes the expected reward. That is,

$$a_{t+1} = \underset{a}{\mathrm{argmax}}\, \mu_a(x_{t+1}, \theta_{t+1}), \;\; \text{with} \;\; \theta_{t+1} \sim f(\theta | a_{1:t}, x_{1:t}, y_{1:t}) \; . \tag{14}$$

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 (expressed 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., $\mu_{a,t+1}$. Since we are dealing with mixture models, the following approaches can be considered:

1. Expectation with mixture assignment sampling

$$\mu_{a,t+1} = x_t^\top \widetilde{u}_{a,z_{a,k,t}} \; , \qquad \text{with} \;\; z_{a,k,t} \sim \mathrm{Cat}\left(\frac{\widetilde{\gamma}_{a,k}}{\sum_{k=1}^{K}\widetilde{\gamma}_{a,k}}\right) \; . \tag{15}$$

2. Expectation with mixture proportion sampling

$$\mu_{a,t+1} = \sum_{k=1}^{K} \pi_{a,k,t} x_t^\top \widetilde{u}_{a,k} \; , \qquad \text{with} \;\; \pi_{a,k,t} \sim \mathrm{Dir}\left(\widetilde{\gamma}_{a,k}\right) \; . \tag{16}$$

Algorithm 1 Variational Thompson sampling

Require: A, K_a and 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 $\theta_{a,k,t+1} \sim q\left(\widetilde{\gamma}_{a,k}, \widetilde{\alpha}_{a,k}, \widetilde{\beta}_{a,k}, \widetilde{u}_{a,k}, \widetilde{V}_{a,k}\right)$

    end for

    Compute $\mu_{a,t+1} = \mu_a(x_{t+1}, \theta_{a,t+1})$

  end for

  Play arm $a_{t+1} = \text{argmax}_a\, \mu_{a,t+1}$

  Observe reward y_{t+1}

  $D = D \cup \{x_{t+1}, a_{t+1}, y_{t+1}\}$

  while Variational convergence criteria not met **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

3. Expectation with mixture proportions

$$\mu_{a,t+1} = \sum_{k=1}^{K} \pi_{a,k,t} x_t^\top \widetilde{u}_{a,k} \ , \qquad \text{with } \ \pi_{a,k,t} = \frac{\widetilde{\gamma}_{a,k}}{\sum_{k=1}^{K} \widetilde{\gamma}_{a,k}} \ . \tag{17}$$

4 Evaluation

In this section, we evaluate the performance of the proposed variational Thompson sampling technique for the contextual multi-armed bandit problem. We consider the two-armed contextual linear Gaussian bandit, with a two dimensional uncorrelated uniform context $x_{i,t} \sim \mathcal{U}(0,1)$, $i \in \{1,2\}$, $t \in \mathbb{N}$.

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

$$\texttt{Scenario A} \ \begin{cases} f_0(y|x_t,\theta) = 0.5 \cdot \mathcal{N}\left(y|(0\ 0)^\top x_t, 1\right) + 0.5 \cdot \mathcal{N}\left(y|(1\ 1)^\top x_t, 1\right) \ , \\ f_1(y|x_t,\theta) = 0.5 \cdot \mathcal{N}\left(y|(2\ 2)^\top x_t, 1\right) + 0.5 \cdot \mathcal{N}\left(y|(3\ 3)^\top x_t, 1\right) \ , \end{cases} \tag{18}$$

and the second, `Scenario B`, with

$$\texttt{Scenario B} \ \begin{cases} f_0(y|x_t,\theta) = 0.5 \cdot \mathcal{N}\left(y|(1\ 1)^\top x_t, 1\right) + 0.5 \cdot \mathcal{N}\left(y|(2\ 2)^\top x_t, 1\right) \ , \\ f_1(y|x_t,\theta) = 0.3 \cdot \mathcal{N}\left(y|(0\ 0)^\top x_t, 1\right) + 0.7 \cdot \mathcal{N}\left(y|(3\ 3)^\top x_t, 1\right) \ . \end{cases} \tag{19}$$

The per-arm reward distributions of the contextual bandits in both scenarios are Gaussian mixtures with two context dependent components. However, the amount of mixture overlap and the similarity between arms differ. Recall the complexity of the reward distributions in `Scenario B`, with a significant overlap between the arms and the unbalanced nature of arm 1.

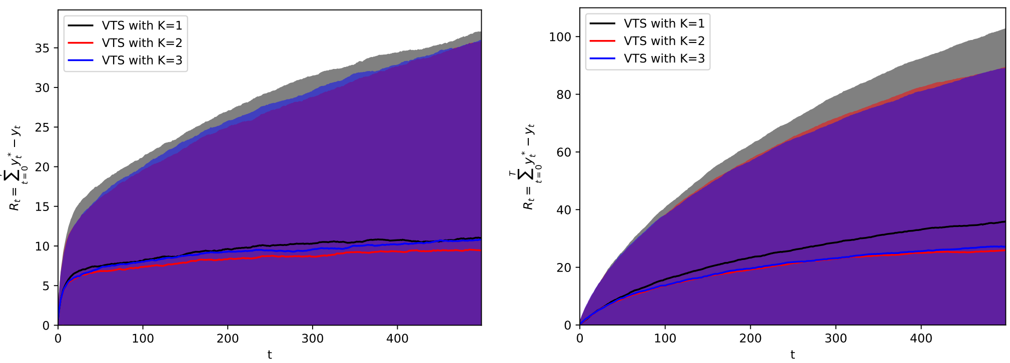
Fig. 2 shows the cumulative regret of the proposed variational Thompson sampling approach in both scenarios, when different assumptions for the variational approximating distribution are made (i.e., assumed prior K). Note that "*VTS with* $K = 1$" is equivalent to a vanilla Thompson sampling approach 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.

We define the cumulative regret as

$$R_t = \sum_{\tau=0}^{t} \mathbb{E}\left\{(y_\tau^* - y_\tau)\right\} = \sum_{\tau=0}^{t} \mu_\tau^* - \bar{y}_\tau \ , \tag{20}$$

where for each time instant t, μ_t^* denotes the expected reward of the optimal arm, and $\bar{y}_t$ the empirical mean of the observed rewards. Reported values are averages over 2000 realizations of the same set of parameters and context (with the standard deviation shown as the shaded region).

Since we have not observed 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 Fig. 2.



(a) `Scenario A`: cumulative regret.

(b) `Scenario B` cumulative regret.

Figure 2: Cumulative regret comparison.

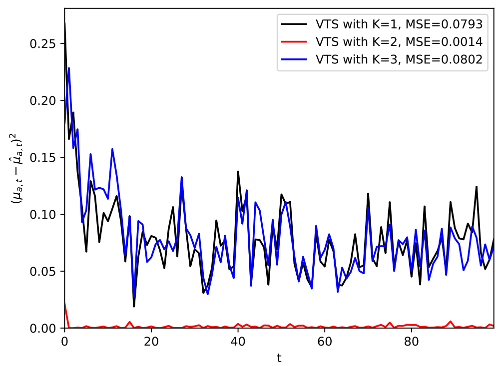
The main conclusion from the results shown in Fig. 2 is that inferring a variational approximation to the true complex reward distribution improves regret performance. Note that, for both models, the best performance is achieved when the assumed Ks match the true number of mixtures per arm (`Scenario A` and `Scenario B` are both a mixture of two Gaussians). As in any posterior sampling algorithm, the cumulative regret variability is large, but we observe a reduced regret variability for the proposed "*VTS with* $K = 2$ and $K = 3$", in comparison to the contextual linear Gaussian Thompson sampling equivalent to "*VTS with* $K = 1$".

It is interesting to observe that, for `Scenario A`, "*VTS with* $K = 1$" performs reasonably well. On the contrary, for `Scenario B`, such an assumption results in higher regret. In other words, a misspecified model performs worse than the proposed alternatives. Specially, in comparison with "*VTS with* $K = 2$", which corresponds to the true underlying mixture distributions in Eqns. 18 and 19. Precisely, the cumulative regret reduction of "*VTS with* $K = 2$" with respect to "*VTS with* $K = 1$" at $t = 500$ is of 14% for `Scenario A` and 28% for `Scenario B`. The issue of model misspecification is more evident for `Scenario B`, as the linear Gaussian contextual model fails to capture the subtleties of the unbalanced mixtures of Eqn. 19.

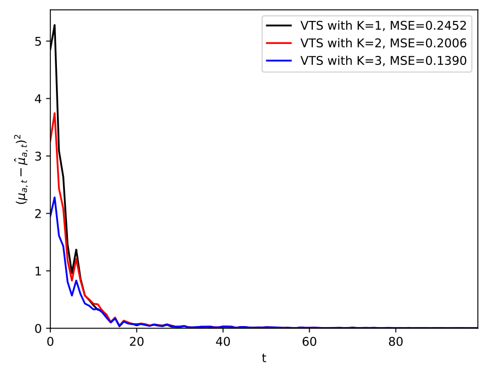
In summary, with a simplistic model assumption as in "*VTS with* $K = 1$", one can not capture the properties of the underlying reward distribution and thus, can not make well-informed decisions. However, by allowing for more complex modeling (i.e., Gaussian mixture models) and by using variational inference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

Furthermore, we highlight that even an overly complex model assumption does provide competitive performance. For both `Scenario A` and `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). 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.

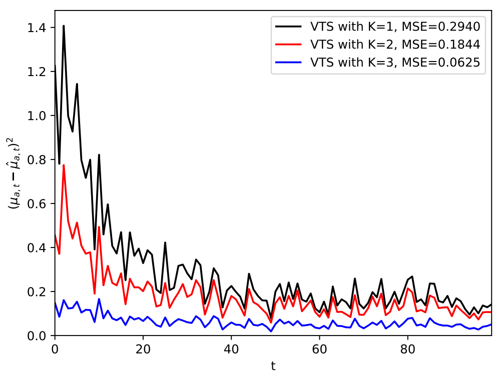
We further elaborate on the analysis of our proposed method by studying its learning accuracy. In bandit algorithms, the goal is to gather enough evidence to identify the best arm, and this can only be achieved if the learning of the arm properties is accurate. We illustrate in Fig. 3 the mean squared



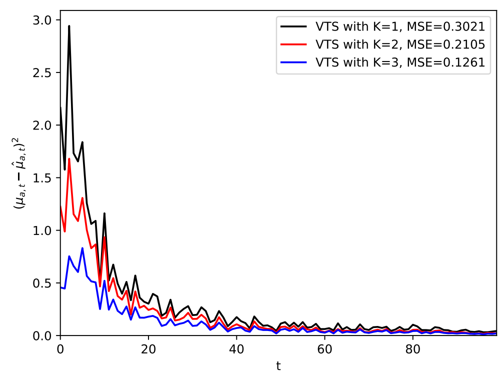
(a) MSE for `Scenario A`, arm 0.



(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 comparison.

error of the per-arm expected reward

$$MSE_a = \frac{1}{T}\sum_{t=0}^{T}\left(\mu_{a,t} - \hat{\mu}_{a,t}\right)^2 \ , \tag{21}$$

where $\hat{\mu}_{a,t}$ denotes the estimated expected reward for arm a at time t. 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* $K = 3$ can accurately estimate the expected reward of the best arm.

We emphasize the additional complexity of `Scenario B` in comparison to `Scenario A`, and its implications. 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 when facing a more complex model (as shown in Figs. 3c-3d). 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 of the presented technique. By allowing for complex modeling of the world and using variational inference to learn it, the proposed variational Thompson sampling 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 the variational inference machinery with a state of the art reinforcement learning technique. The proposed variational Thompson sampling allows for interpretable bandit modeling with complex reward functions learned from online data, extending 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 contexts when relevant attributes of items, customers, patients, or other 'examples' are unobservable, and thus the latent variables are truly 'incomplete' as in the motivating case for expectation maximization modeling (Dempster et al. [1977]).

References

Shipra Agrawal and Navin Goyal. Analysis of Thompson Sampling for the multi-armed bandit problem. *CoRR*, abs/1111.1797, 2011. URL `http://arxiv.org/abs/1111.1797`.

Shipra Agrawal and Navin Goyal. Thompson Sampling for Contextual Bandits with Linear Payoffs. *CoRR*, abs/1209.3352, 2012a. URL `http://arxiv.org/abs/1209.3352`.

Shipra Agrawal and Navin Goyal. Further Optimal Regret Bounds for Thompson Sampling. *CoRR*, abs/1209.3353, 2012b. URL `http://arxiv.org/abs/1209.3353`.

José M. Bernardo and Adrian F.M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley, 2009. ISBN 9780470317716. doi: 10.1002/9780470316870. URL `http://onlinelibrary.wiley.com/book/10.1002/9780470316870`.

Christopher Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer-Verlag New York, 2006. URL `http://www.springer.com/us/book/9780387310732`.

Olivier Chapelle and Lihong 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 `http://papers.nips.cc/paper/4321-an-empirical-evaluation-of-thompson-sampling.pdf`.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977. URL `https://www.jstor.org/stable/2984875`.

Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, and Aviv 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. URL `http://dx.doi.org/10.1561/2200000049`.

Nathaniel Korda, Emilie Kaufmann, and Rémi 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. URL `http://papers.nips.cc/paper/5110-thompson-sampling-for-1-dimensional-exponential-family-bandits.pdf`.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A Contextual-Bandit Approach to Personalized News Article Recommendation. *CoRR*, abs/1003.0146, 2010. URL `http://arxiv.org/abs/1003.0146`.

Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences. In *Conference On Learning Theory*, 2011. URL `http://hal.inria.fr/inria-00574987/fr/`.

Geoffrey McLachlan and David Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2004, 2004. ISBN 9780471654063. URL `https://books.google.cz/books?id=c2_fAox0DQoC`.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, (58):527–535, 1952. doi: https://doi.org/10.1090/S0002-9904-1952-09620-8.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014. doi: http://pubsonline.informs.org/doi/abs/10.1287/moor.2014.0650.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of Thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016. URL `http://www.jmlr.org/papers/volume17/14-087/14-087.pdf`.

Steven 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. URL `http://dx.doi.org/10.1002/asmb.874`.

Steven L. Scott. Multi-armed bandit experiments in the online service economy. *Applied Stochastic Models in Business and Industry*, 31:37–49, 2015. URL `http://onlinelibrary.wiley.com/doi/10.1002/asmb.2104/abstract`. Special issue on actual impact and future perspectives on stochastic modelling in business and industry.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press: Cambridge, MA, 1998. URL `https://mitpress.mit.edu/books/reinforcement-learning`.

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. ISSN 00063444. URL `http://www.jstor.org/stable/2332286`.

William R. Thompson. On the Theory of Apportionment. *American Journal of Mathematics*, 57(2):450–456, 1935. ISSN 00029327, 10806377. URL `http://www.jstor.org/stable/2371219`.