# Nonparametric Gaussian mixture models 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

August 8, 2018

**Abstract**

The multi-armed bandit is a sequential allocation task where an agent must learn a policy that maximizes long term payoff, where only the reward of the played arm is observed at each iteration. In the stochastic setting, the reward for each action is generated from an unknown distribution, which depends on a given 'context', available at each interaction with the world. Thompson sampling is a generative, interpretable multi-armed bandit algorithm that has been shown both to perform well in practice, and to enjoy optimality properties for certain reward functions. Nevertheless, Thompson sampling requires sampling from parameter posteriors and calculation of expected rewards, which are possible for a very limited choice of distributions. We here extend Thompson sampling to more complex scenarios by adopting a very flexible set of reward distributions: nonparametric Gaussian mixture models. The generative process of Bayesian nonparametric mixtures naturally aligns with the Bayesian modeling of multi-armed bandits. This allows for the implementation of an efficient and flexible Thompson sampling algorithm: the nonparametric model autonomously determines its complexity in an online fashion, as it observes new rewards for the played arms. We show how the proposed method sequentially learns the nonparametric mixture model that best approximates the true underlying reward distribution. Our contribution is valuable for practical scenarios, as it avoids stringent model specifications, and yet attains reduced regret.

## 1 Introduction

Recent advances in reinforcement learning (13) have sparked renewed interest in sequential decision making. The aim of sequential decision making is to optimize interactions with the world (exploit) while simultaneously learning how the world operates (explore). Its origins can be traced back to the beginning of the past century, with important contributions within the field of statistics by Thompson (40) and later Robbins (28). The multi-armed bandit (MAB) problem is a natural abstraction for a wide variety of real-world challenges that require learning while simultaneously maximizing rewards.

The name "bandit" finds its origin in the playing strategy one must devise when facing a row of slot machines. The contextual setting, where at each interaction with the world side information (known as 'context') is available, is a natural extension of the bandit problem. Recently, a renaissance of the study of MAB problems has flourished (2, 22, 33). The performance of algorithms for contextual bandits with linear payoffs (10, 29) has been widely studied in the last decade (1, 7, 18). Furthermore, it has attracted interest from industry, due to its impact in digital advertising and products (8, 20).

Thompson Sampling (32, 39) and its generalization known as posterior sampling, provide an elegant approach that tackles the exploration-exploitation dilemma. It updates a posterior over expected rewards, and chooses actions based on the probability that they are optimal. It has been empirically and theoretically proven to perform competitively for many MAB models (3, 4, 8, 16, 30, 31, 34). Besides, its applicability to the more general reinforcement learning setting of Markov Decision Processes (6) has recently tracked momentum (12, 25).

Thompson sampling and the Bayesian modeling of the MAB problem facilitate not only generative and interpretable modeling, but sequential and batch processing algorithms as well. Within this framework, one only requires access to posterior samples of the model. Unfortunately, maintaining such a posterior is intractable for distributions not in the exponential family (16). As such, developing practical MAB methods to balance exploration and exploitation in complex domains remains largely unsolved.

In an effort to extend Thompson sampling to more complex scenarios, researchers have considered other flexible reward functions and Bayesian inference. For example, recent approaches have embraced approximate Bayesian neural networks for Thompson Sampling. Neural networks have proven to be powerful function approximators, and approximate Bayesian inference provides posterior uncertainty estimates. To that end, variational methods, stochastic mini-batches, and Monte Carlo techniques have been studied for uncertainty estimation of posteriors (5, 15, 19, 21, 24). In a recent benchmark of such techniques (27), it was reported that even if successful in the supervised learning setting, they under-perform in the MAB scenario. In particular, Riquelme et al. (27) emphasize the issue of adapting the slow convergence uncertainty estimates of neural net based methods to the bandit setting.

In parallel, others have focused on extending Thompson sampling by targeting alternative classes of reward functions. Some have focused on approximating the unknown bandit reward functions with Gaussian mixture models (42), while others have assumed a Gaussian process reward distribution (14, 17, 36). The latter are powerful nonparametric methods for modeling distributions over non-linear continuous functions (26). Unfortunately, standard Gaussian processes are computationally demanding, as they scale cubically in the number of observations, limiting their applicability to small datasets and the online setting (even if advancements such as pseudo-observations (35) or variational inference (41) have been proposed to mitigate these shortcomings).

In this paper, we combine the large hypothesis space of mixture models — which can approximate any continuous reward distribution — with the flexibility of Bayesian nonparametrics (11). In many contexts, a countably infinite mixture is a very realistic model to assume, and has been shown to succeed in modeling a diversity of phenomena. Within the Bayesian framework, one uses prior distributions over the mixing proportions, such as Dirichlet or Pitman-Yor processes (37), which allow for inference of the appropriate complexity of a model from observed data. These models describe mixtures in which one not only does not explicitly specify the number of mixtures, but allows the possibility of an unbounded number of mixtures. Bayesian nonparametrics support a wide class of models, yet have analytically tractable inference and online update rules.

Our contribution here is on exploiting Bayesian nonparametric mixture models for Thompson sampling to perform MAB optimization. This provides a new flexible framework for solving a rich class of MAB problems. We model the complex mapping between the observed rewards and the unknown parameters of the generating process with nonparametric Gaussian mixture models. For learning such a nonparametric distribution within the contextual multi-armed bandit setting, we leverage the advances in Markov Chain Monte Carlo methods for Bayesian nonparametric models (23).

Mixtures of distributions provide a powerful approach for nonparametric density estimation, and the generative interpretation of Bayesian nonparametric models corresponds to the sequential nature of the MAB problem as well. The proposed method learns the nonparametric mixture model that best approximates the true underlying reward distribution, adjusting its complexity as it sequentially observes additional data. To the best of our knowledge, no other work uses Bayesian nonparametric mixture models to address the contextual MAB.

We formally introduce the MAB problem and the Bayesian nonparametric framework in Section 2, before providing the description of the proposed nonparametric Thompson sampling in Section 3. We evaluate its performance in Section 4, and suggest generalizations in Section 5.

# 2 Background

## 2.1 Multi-armed bandits

A multi-armed bandit is a real time sequential decision process in which, at each iteration, an agent is asked to select an action according to a policy which maximizes the accumulated rewards over time, balancing exploitation and exploration. In the contextual case, one must decide which arm $a \in \{0, \cdots, A-1\}$ to play next (i.e., pick $a_{t+1}$), based on the available context, e.g., $x_{t+1} \in \mathbb{R}^d$. At every iteration $t$, the observed reward $y_t$ is independently drawn from the unknown reward distribution corresponding to the played arm, conditioned on the context and parameterized by unknown $\theta$; i.e., $y_t \sim p_a(y|x_t, \theta)$. Due to the stochastic nature of the bandit, one summarizes each arm's reward via its conditional expectation for that context $\mu_a(x, \theta) = \mathbb{E}_a\{y|x, \theta\}$.

When the properties of the arms (i.e., the parameters $\theta$) 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 MAB problem is not knowing the true reward parameters $\theta$ or, more generally, the lack of knowledge about the reward-generating model. Thus, one needs to simultaneously (1) learn the properties of the reward distribution and (2), decide which action to take sequentially. The next arm to play is chosen based upon the history observed, with the goal of maximizing the expected (cumulative) reward. Previous history contains the set of given contexts, played arms, and observed rewards up to time $t$, denoted as $\mathcal{H}_{1:t} = \{x_{1:t}, a_{1:t}, y_{1:t}\}$, with $x_{1:t} \equiv (x_1, \cdots, x_t)$, $a_{1:t} \equiv (a_1, \cdots, a_t)$ and $y_{1:t} \equiv (y_1, \cdots, y_t)$.

Among the many alternatives to address this class of problems, Thompson sampling is particularly appealing, due to its generative formulation and its connection with Bayesian modeling. Furthermore, it has been shown to perform empirically well and has sound theoretical bounds, for both contextual and context-free problems (3, 4, 8, 16, 30, 31).

Thompson sampling chooses what arm to play in proportion to its probability of being optimal, i.e.,

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

where $a^*_{t+1}$ is the optimal arm given the true parameters and the observed context, i.e., $a^*_{t+1} = \text{argmax}_a \, \mu_a(x_{t+1}, \theta)$. If the parameters of the model are known, the above expression becomes deterministic, as one always picks the arm with the maximum expected reward

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

where $I_a(\cdot)$ denotes the indicator function

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

When the parameters of the model are unknown, one needs to explore ways of computing the probability of each arm being optimal. In a Bayesian setting, the parameters are modeled as random variables with priors. Specifically, one marginalizes over the posterior probability distribution of the parameters, after observing rewards and actions up to time instant $t$, i.e.,

$$\Pr\left[a = a^*_{t+1} \middle| \mathcal{H}_{1:t}\right] = \int p(a = a^* | x_{t+1}, \theta, \mathcal{H}_{1:t}) p(\theta | \mathcal{H}_{1:t}) \mathrm{d}\theta = \int I_a(x_{t+1}, \theta) p(\theta | \mathcal{H}_{1:t}) \mathrm{d}\theta \ . \tag{5}$$

The above integral can not be solved exactly, even when the parameter posterior update is analytically tractable. Therefore, when reward distributions that are not within the exponential family are considered, one must resort to approximations of the posterior. In the following, we propose nonparametric mixture models as tractable yet performant reward distributions for the MAB.

## 2.2 Bayesian nonparametric mixture models

Bayesian nonparametric models provide a powerful density estimation framework that adjust model complexity in response to the data observed. The combination of mixture models with Bayesian nonparametrics embodies a large hypothesis space, which can arbitrarily approximate continuous reward distributions. Bayesian nonparametric mixture models describe countably infinite mixture distributions, which are very flexible assumptions suited for many practical settings. We refer to (11) for a detailed review of standard nonparametric models and how they can be used in practice.

A variety of Bayesian nonparametric alternatives have been studied in literature. We here focus on the Pitman-Yor model, which is a stochastic process whose sample path is a probability distribution. It is a generalization of Bayesian nonparametric models from where a drawn random sample is an infinite discrete probability distribution. In the following, we succinctly summarize the generative process and the basics for its inference.

A Pitman-Yor mixture model (37), with a discount parameter $0 \leq d < 1$ and a concentration parameter $\gamma > -d$, is described by the following generative process:

1. Mixture parameters are drawn from the Pitman-Yor process, i.e., $\theta_n \sim G$, where $G = PY(d, \gamma, G_0)$. Equivalently, the process can be described as

$$\theta_{n+1} | \theta_{1:n}, d, \gamma, G_0 \sim \sum_{k=1}^{K} \frac{n_k - d}{n + \gamma} \delta_{\theta_k} + \frac{\gamma + Kd}{n + \gamma} G_0 \ , \tag{6}$$

where $n$ refers to all the available observations, and $n_k$ to the number of observations assigned to mixture $k$.

2. The observation is drawn from the emission distribution parameterized by its corresponding parameters, i.e., $y_{n+1} \sim f(y|\theta_{n+1})$.

For parametric measures, we write $G_0(\theta) = G(\theta|\Theta_0)$ and $G_n(\theta) = G(\theta|\Theta_n)$, where $\Theta_0$ are the prior hyperparameters of the emission distribution, and $\Theta_n$ are the posterior parameters after $n$ observations.

We note that the Dirichlet process can be readily obtained from Eqn. (6) by using $d = 0$. The discount parameter gives the Pitman-Yor process more flexibility over tail behavior (the Dirichlet process has exponential tails, whereas the Pitman-Yor can have power-law tails).

For analysis and inference of these models, one incorporates auxiliary mixture variables $z_n$. These are categorical variables, where $z_n = k$, if observation $y_n$ is drawn from mixture $k$. The joint posterior of these assignments follows, for $d = 0$,

$$p(z_{1:n}|\gamma) = \prod_{i=1}^{n} p(z_i|z_{1:n-1}, \gamma) = \frac{\gamma^K \prod_{k=1}^{K}(n_k - 1)!}{\prod_{i=1}^{n}(i - 1 + \gamma)} = \frac{\Gamma(\gamma)}{\Gamma(\gamma + n)}\gamma^K \prod_{k=1}^{K} \Gamma(n_k) \,, \tag{7}$$

where $n_k$ indicates the number of observations drawn from mixture $k$ and $n = \sum_{k=1}^{K} n_k$. The full joint likelihood of assignments and observations is

$$p(y_{1:n}, z_{1:n}|\gamma, \Theta) = p(y_{1:n}|z_{1:n}, \Theta)p(z_{1:n}|\gamma) = p(y_{1:n}|z_{1:n}, \Theta)\left(\frac{\gamma^K \prod_{k=1}^{K}(n_k - 1)!}{\prod_{i=1}^{n}(i - 1 + \gamma)}\right) . \tag{8}$$

For inference of the above model given observations $y_{1:n}$, one can derive a Gibbs sampler that iterates between mixture assignment sampling and posterior updates of the emission distribution parameters (Teh and Jordan (37) provide a detailed explanation of the procedure).

The conditional distributions of observation assignments $z_n$ to already drawn mixtures $k \in \{1, \cdots, K\}$, and a new unseen mixture $k_{new}$ are

$$\begin{cases} p(z_{n+1} = k|y_{n+1}, y_{1:n}, z_{1:n}, \gamma, G_0) \propto \frac{n_k - d}{n+\gamma} \int_\theta f(y_{n+1}|\theta_{n+1}) G_n(\theta) \mathrm{d}\theta \,, \\ p(z_{n+1} = k_{new}|y_{n+1}, y_{1:n}, z_{1:n}, \gamma, G_0) \propto \frac{\gamma + Kd}{n+\gamma} \int_\theta f(y_{n+1}|\theta_{n+1}) G_0(\theta) \mathrm{d}\theta \,. \end{cases} \tag{9}$$

Given these mixture assignments, one updates the parameter posteriors conditioned on $z_{1:n}$ and observations $y_{1:n}$, based on the specific choices of emission distribution and priors: $G_n(\theta) = G(\theta|y_{1:n}, z_{1:n}, \Theta_0)$. These also determine the computation of the predictive distribution $f(y|\Theta) = \int_\theta f(y|\theta) G(\theta|\Theta) \mathrm{d}\theta$ for solving Eqn. (9). For analytical convenience, one usually resorts to emission distributions with their conjugate priors.

# 3 Proposed method

We now describe how to combine Bayesian nonparametric mixture models with Thompson sampling for the MAB setting. The graphical model of the Bayesian nonparametric MAB is rendered in Fig. 1. We consider a completely independent set of nonparametric mixture models $G_{a,0}$ per arm, with their own hyperparameters $d_a$ and $\gamma_a$.

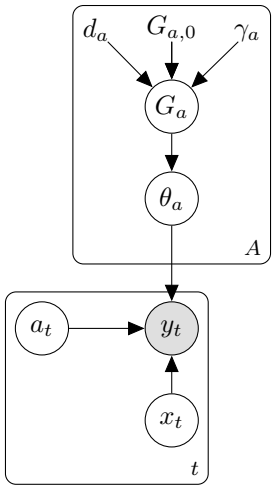



Figure 1: Graphical model of the nonparametric mixture bandit distribution.

As shown in Fig. 1, we assume complete independence of per-arm reward distributions, i.e., each arm of the bandit is allowed to follow a different family of distributions. We consider this setting to be a very powerful extension of the MAB problem, which has not attracted much interest so far.

An alternative would be to consider a hierarchical-nonparametric model (37, 38), where all arms are assumed to obey the same family of distributions, but only their mixture proportions are allowed to vary across arms. The main advantage of this alternative is that one would learn parameter posteriors from rewards of all played arms, with the disadvantage of all arms being limited to the same family of reward distributions. We illustrate this alternative hierarchical nonparametric MAB, and provide details of the model and its inference, in Appendix A.

In order to approximate any continuous reward distribution, we study nonparametric Gaussian mixtures as a flexible formulation for modeling complex MAB reward densities.

We focus on context-conditional Gaussian emission distributions $y \sim \mathcal{N}\left(y|x^\top w_{a,k}, \sigma^2_{a,k}\right)$, which are parameterized per-arm and per-mixture; i.e., $\theta_{a,k} = \{w_{a,k}, \sigma^2_{a,k}\}$. The conjugate prior for such emission distribution is a Normal-inverse Gamma, with hyperparameters $\Theta_{a,0} = \{u_{a,0}, V_{a,0}, \alpha_{a,0}, \beta_{a,0}\}$, i.e.,

$$G_{a,0}(\theta_a) = \text{NIG}\left(w_a, \sigma^2_a | u_{a,0}, V_{a,0}, \alpha_{a,0}, \beta_{a,0}\right) = \mathcal{N}\left(w_a | u_{a,0}, \sigma^2_a V_{a,0}\right) \Gamma^{-1}(\sigma^2_a | \alpha_{a,0}, \beta_{a,0}) \; . \tag{10}$$

After observing rewards $y_{1:n}$, and conditioned on assignments $z_{1:n}$, the posteriors of the parameters per arm and mixture $\theta_{a,k}$ also follow a Normal-inverse Gamma distribution.

The updated hyperparameters of such posterior depend on the number $n_{a,k}$ of rewards observed after playing arm $a$ that are assigned to mixture $k$:

$$
\begin{aligned}
G_{a,n_{a,k}}(\theta_{a,k}) &= \text{NIG}\left(w_{a,k}, \sigma^2_{a,k} | u_{a,k,n_{a,k}}, V_{a,k,n_{a,k}}, \alpha_{a,k,n_{a,k}}, \beta_{a,k,n_{a,k}}\right) \ , \\
\Theta_{a,k,n_{a,k}} &= \begin{cases}
V^{-1}_{a,k,n_{a,k}} = x_{1:n} R_{a,k} x^\top_{1:n} + V^{-1}_{a,0} \ , \\
u_{a,k,n_{a,k}} = V_{a,k,n_{a,k}} \left(x_{1:n} R_{a,k} y_{1:n} + V^{-1}_{a,0} u_{a,0}\right) \ , \\
\alpha_{a,k,n_{a,k}} = \alpha_{a,0} + \frac{1}{2} \text{tr}\left\{R_{a,k}\right\} \ , \\
\beta_{a,k,n_{a,k}} = \beta_{a,0} + \frac{1}{2}\left(y^\top_{1:n} R_{a,k} y_{1:n}\right) \\
\qquad\qquad + \frac{1}{2}\left(u^\top_{a,0} V^{-1}_{a,0} u_{a,0} - u^\top_{a,k,n_{a,k}} V^{-1}_{a,k,n_{a,k}} u_{a,k,n_{a,k}}\right) \ ,
\end{cases}
\end{aligned}
\tag{11}
$$

where $R_{a,k} \in \mathbb{R}^{n_a \times n_a}$ is a sparse diagonal matrix with elements $\left[R_{a,k}\right]_{n,n'} = \mathbb{1}[a_n = a, z_n = k]$, and $n_a$ the number of rewards observed after playing arm $a$.

Finally, the predictive emission distribution after marginalization of the parameters $\theta_{a,k}$, needed for solving Eqn. (9), follows a conditional Student-t distribution

$$
\begin{aligned}
f_{a,k}(y|x) &= \mathcal{T}\left(y | \nu_{a,k,y}, m_{a,k,y}, r_{a,k,y}\right) \\
&= \frac{\Gamma\left(\frac{\nu_{a,k,y}+1}{2}\right)}{\sqrt{\nu_{a,k,y}\pi} \cdot r_{a,k,y} \cdot \Gamma\left(\frac{\nu_{a,k,y}}{2}\right)} \cdot \left| 1 + \frac{1}{\nu_{a,k,y}} \frac{(y - m_{a,k,y})^2}{r^2_{a,k,y}} \right|^{-\frac{\nu_{a,k,y}+1}{2}} \ , \\
\text{with } &\begin{cases}
\nu_{a,k,y} = 2\alpha_{a,k} \ , \\
m_{a,k,y} = x^\top u_{a,k} \ , \\
r^2_{a,k,y} = \frac{\beta_{a,k}}{\alpha_{a,k}}(1 + x^\top V_{a,k} x) \ .
\end{cases}
\end{aligned}
\tag{12}
$$

The hyperparameters used above are those of the prior ($\Theta_{a,0}$) or the posterior ($\Theta_{a,k,n_{a,k}}$), depending on whether the predictive density refers to a new mixture $k_{new}$, or a "seen" mixture $k$ for which $n_{a,k}$ observations have been already assigned to, respectively.

Similarly, the likelihood of a set of rewards assigned to a per-arm mixture $k$, $Y_{a,k} = y_{1:n} \cdot \mathbb{1}[a_n = a, z_n = k]$, given their associated contexts $X_{a,k} = x_{1:n} \cdot \mathbb{1}[a_n = a, z_n = k]$, follows the matrix t-distribution

$$
\begin{aligned}
f(Y_{a,k} | X_{a,k}, X_{\setminus a,k}, Y_{\setminus a,k}) &= \mathcal{MT}\left(Y_{a,k} | \nu_{Y_{a,k}}, M_{Y_{a,k}}, \Psi_{Y_{a,k}}, \Omega_{Y_{a,k}}\right) \\
&= \frac{\Gamma\left(\frac{\nu_{Y_{a,k}} + n_{a,k}}{2}\right)}{\pi^{\frac{n_{a,k}}{2}} \cdot \Gamma\left(\frac{\nu_{Y_{a,k}}}{2}\right)} \cdot \left|\Omega_{Y_{a,k}}\right|^{-\frac{n_{a,k}}{2}} \cdot \left|\Psi_{Y_{a,k}}\right|^{-\frac{1}{2}} \\
&\qquad \times \left| I_{n_{a,k}} + \Psi^{-1}_{Y_{a,k}} \left(Y_{a,k} - M_{Y_{a,k}}\right) \Omega^{-1}_{Y_{a,k}} \left(Y_{a,k} - M_{Y_{a,k}}\right)^\top \right|^{-\frac{\nu_{Y_{a,k}} + n_{a,k}}{2}} \ , \\
&\qquad \text{with } \begin{cases}
\nu_{Y_{a,k}} = 2\alpha_{a,k} \ , \\
M_{Y_{a,k}} = X^\top_{a,k} u_{a,k} \ , \\
\Psi_{Y_{a,k}} = I_{n_{a,k}} + X^\top_{a,k} V_{a,k} X_{n_{a,k}} \ , \\
\Omega_{Y_{a,k}} = 2\beta_{a,k} \ .
\end{cases}
\end{aligned}
\tag{13}
$$

### 3.1 Thompson sampling for nonparametric Gaussian mixture models

We now describe our proposed Thompson sampling technique for multi-armed contextual bandits with nonparametric Gaussian mixture reward models. To that end, we leverage the Bayesian generative process described above, and infer the posteriors over the parameters, in order to implement a posterior sampling based policy (30).

In the MAB problem, the agent needs to decide which arm to play next, based on the information available at that iteration. In a randomized probability matching technique, each arm is picked based on its probability of being optimal. However, since the integral in Eqn. (5) is intractable, Thompson (40) sampling draws a random parameter sample from the posterior instead, and picks the action that maximizes the expected reward, given that parameter sample. That is,

$$a^*_{t+1} = \underset{a}{\operatorname{argmax}}\, \mu_a(x_{t+1}, \theta_{t+1}), \qquad \text{with} \qquad \theta_{t+1} \sim p(\theta|\mathcal{H}_{1:t}). \tag{14}$$

In the proposed model, we sample per-arm and per-mixture Gaussian parameters $\theta_{a,k}$ from the posterior hyperparameter distributions with updated $\Theta_{a,k,n_a}$, conditioned on the mixture assignments $z_{1:n}$ determined by the Gibbs sampler in Eqn. (9). The Gibbs sampler is run until a stopping criteria is met (i.e., the model likelihood of the sampled MCMC chain is stable within an $\epsilon$ margin between steps, or a maximum number of iterations is reached).

With the sufficient statistics of these assignments (i.e., the counts $n_{a,k}$ of rewards observed for arm $a$ and assigned to mixture $k$), and the posterior parameter samples $w_{a,k,t+1}$, one computes the expected reward for each arm of the bandit as follows:

$$\mu_{a,t+1} = \sum_{k=1}^{K_a} \frac{n_{a,k} - d_a}{n_a + \gamma_a} \cdot \left(x_{t+1}^\top w_{a,k,t+1}\right) + \frac{\gamma_a + K_a d_a}{n_a + \gamma_a} \left(x_{t+1}^\top w_{a,k,t+1}\right) \ . \tag{15}$$

This leads to the proposed nonparametric Gaussian mixture model Thompson sampling for the contextual MAB problem in Algorithm 1 .

## 4 Evaluation

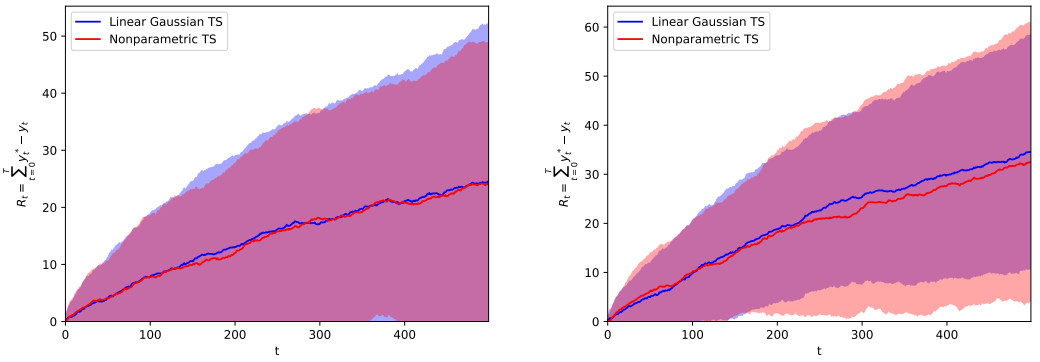
In this section, we evaluate the performance of the proposed nonparametric mixture model Thompson sampling technique. First, we validate that the method performs as expected in the simplest case, i.e., when dealing with a contextual linear Gaussian MAB.

We evaluate different parameterizations of two- and three-armed linear contextual bandits, with uniform and uncorrelated 2-dimensional context, i.e., $x_{i,t} \sim \mathcal{U}(0,1), i \in \{0,1\}, t \in \mathbb{N}$. We provide results for a specific parameterization of these contextual Gaussian bandits in Fig. 2, where we observe the flexibility of nonparametric mixture models in action (similar results were obtained for other bandit parameterizations, see Appendix B).

We show how the proposed method is able to provide as good regret performance as a Thompson Sampling method that is aware of the true underlying reward distribution. That is, the nonparametric Gaussian mixture model is able to accurately fit the mixture to the correct underlying distribution, so that the regret performance of the proposed Thompson sampling is optimal. These results serve as a validation of the quality of the nonparametric mixture model assumption, as the performance loss of the proposed bandit is negligible: the nonparametric Thompson sampling method is as good as the analytical alternative.

**Algorithm 1** Nonparametric Gaussian mixture model Thompson sampling

**Input:** Number of arms $A$ and per-arm hyperparameters $d_a$, $\gamma_a$, $\Theta_{a,0}$
$D = \emptyset$
**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 parameters from the posterior $\theta_{a,k,t+1} \sim G_{a,k,n_{a,k}}(\theta_{a,k})$
    **end for**
    Compute $\mu_{a,t+1}$ as in Eqn. (15)
  **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** NOT Gibbs convergence criteria **do**
    Update mixture assignments $z_{1:t}$ based on Eqn. (9) and compute sufficient statistics $n_{a,k}$
    Update parameter posteriors $\Theta_{a,k,n_{a,k}}$ as in Eqn. (11)
  **end while**
**end for**




(a) 2-armed contextual linear Gaussian bandit with $\theta_{0,i} = -0.1, \theta_{1,i} = 0.1, \sigma_i^2 = 1 \forall i$.

(b) 3-armed contextual linear Gaussian bandit with $\theta_{0,i} = -0.1, \theta_{0,i} = 0, \theta_{2,i} = 0.1, \sigma_i^2 = 1 \forall i$.

Figure 2: Mean regret (standard deviation shown as shaded region) for linear Gaussian bandits. The proposed nonparametric MAB method is comparable to the analytical alternative.

Furthermore, note how the Gibbs sampling inference aligns well with the online nature of the bandit, as the inference is recomputed only with the reward observed for the last played arm. Even more, because of the incremental availability of observations, the Gibbs sampler achieves convergence (as described in section 3.1) in few iterations (in our experiments, less than 5 steps where usually required to achieve a 1% loglikelihood relative difference between steps). Such a low computational burden is possible because the Gibbs sampler is run, at each interaction with the world, from a good starting point: i.e., the parameter space that describes all but this newly observed reward.

We now study more challenging cases, i.e., those were the underlying reward distributions do not fit into the exponential family assumption. We focus on the following scenarios:

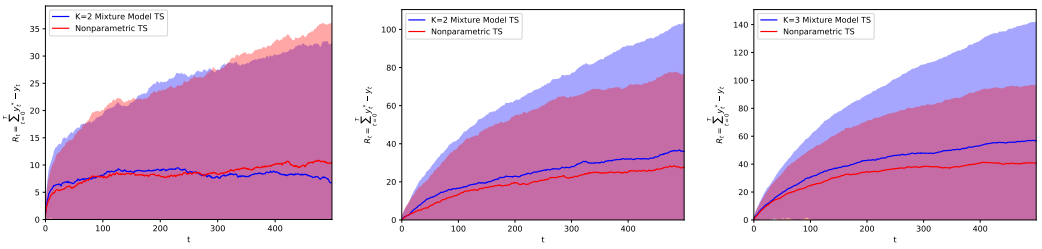
$$\texttt{Scenario A} \quad \begin{cases} f_0(y|x_t,\theta) = 0.5 \cdot \mathcal{N}\left(y|(0\ 0)^\top x_t, 1\right) + 0.5 \cdot \mathcal{N}\left(y|(1\ 1)^\top x_t, 1\right), \\ f_1(y|x_t,\theta) = 0.5 \cdot \mathcal{N}\left(y|(2\ 2)^\top x_t, 1\right) + 0.5 \cdot \mathcal{N}\left(y|(3\ 3)^\top x_t, 1\right) \ . \end{cases} \tag{16}$$

$$\texttt{Scenario B} \quad \begin{cases} f_0(y|x_t,\theta) = 0.5 \cdot \mathcal{N}\left(y|(1\ 1)^\top x_t, 1\right) + 0.5 \cdot \mathcal{N}\left(y|(2\ 2)^\top x_t, 1\right) \ , \\ f_1(y|x_t,\theta) = 0.3 \cdot \mathcal{N}\left(y|(0\ 0)^\top x_t, 1\right) + 0.7 \cdot \mathcal{N}\left(y|(3\ 3)^\top x_t, 1\right) \ . \end{cases} \tag{17}$$

$$\texttt{Scenario C} \quad \begin{cases} f_0(y|x_t,\theta) = \mathcal{N}\left(y|(1\ 1)^\top x_t, 1\right) \ , \\ f_1(y|x_t,\theta) = 0.5 \cdot \mathcal{N}\left(y|(1\ 1)^\top x_t, 1\right) + 0.5 \cdot \mathcal{N}\left(y|(2\ 2)^\top x_t, 1\right) \ , \\ f_2(y|x_t,\theta) = 0.3 \cdot \mathcal{N}\left(y|(0\ 0)^\top x_t, 1\right) \\ \qquad\qquad +0.6 \cdot \mathcal{N}\left(y|(3\ 3)^\top x_t, 1\right) + 0.1 \cdot \mathcal{N}\left(y|(4\ 4)^\top x_t, 1\right) \ . \end{cases} \tag{18}$$

The reward distributions of the contextual bandits in all the above are Gaussian mixtures dependent on a two dimensional uncorrelated uniform context, i.e., $x_{i,t} \sim \mathcal{U}(0,1)$, $i \in \{0,1\}$, $t \in \mathbb{N}$. These reward distributions are complex in that they are multi-modal and, in `Scenario B` and `Scenario C`, unbalanced. The scenarios differ in 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. Furthermore, `Scenario C` describes a MAB with different per-arm reward distributions: a linear Gaussian distribution for arm 0, a bi-modal Gaussian mixture for arm 1, and an unbalanced Gaussian mixture with three components for arm 2.

Fig. 3 shows the cumulative regret of the proposed nonparametric mixture model Thompson sampling approach in all scenarios. We compare the performance of our method to that of an oracle Thompson sampling approach that knows the true dimensionality of the problem (i.e., the number of underlying mixtures $K$). Note that this is only possible in a simulated scenario, as knowing the reward complexity of a MAB beforehand is impractical (an alternative would be to run multiple model assumptions in parallel, with a subsequent model selection).




(a) `Scenario A`. (b) `Scenario B`. (c) `Scenario C`.

Figure 3: Mean regret (standard deviation shown as shaded region) for proposed method.

Fig. 3 shows the full power and flexibility of the proposed nonparametric Gaussian mixture model based Thompson sampling. Due to the capacity of Bayesian nonparametrics to autonomously adjust the complexity of the model to the sequentially observed data, the proposed method not only fits the underlying reward function accurately, but also attains reduced regret. This is achieved in the most challenging MAB scenarios (i.e., different per-arm distributions not in the exponential family), and with no parameter tuning ($d = 0$ and $\gamma = 0.1$ have been used in this experiments).

Finally, we evaluate the proposed method in a real application, i.e., the recommendation of personalized news articles, in a similar fashion as done by Chapelle and Li (8). Online content recommendation represents an important example of reinforcement learning, as it requires efficient balancing of the exploration and exploitation tradeoff.

We use the *R6A - Yahoo! Front Page Today Module User Click Log Dataset*[1], which contains a fraction of user click log for news articles displayed in the *Featured Tab* of the *Today Module on Yahoo! Front Page* during the first ten days in May 2009. The articles to be displayed were originally chosen uniformly at random from a hand-picked pool of high-quality articles. For our evaluation, we picked 2 subsets of 20 articles shown in May 4th and 5th, with a total of 75779 and 77308 logged user interactions, respectively.

The goal is to choose the most interesting article to users, evaluated by counting the total number of clicks. In the dataset, each user is associated with six features, a bias term and 5 features that correspond to the membership features constructed via the conjoint analysis with a bilinear model described in (9).

We treat each article as an arm ($A = 20$), and the reward is whether the article is clicked or not by the user ($y_t = \{1, 0\}$). We pose the problem as a MAB, where we want to maximize the average click-through rate (CTR) on the recommended articles. We implemented both the proposed nonparametric Gaussian mixture model, and the logistic reward model as proposed by Chapelle and Li (8), with the Importance-Sampling based implementation of Urteaga and Wiggins (43).

Summary CTR results are provided in Table 1, for both evaluated reward bandit models. Observe the flexibility of the nonparametric mixture model, as it is able to attain an overall improved CTR rate.

| | | May 4th | | May 5th | |
|---|---|---|---|---|---|
| | | CTR | Normalized CTR | CTR | Normalized CTR |
| Model | Logistic | 0.0451 +/- 0.0068 | 1.0855 +/- 0.1794 | 0.0462 +/- 0.0054 | 1.0472 +/- 0.1486 |
| | Nonparametric mixture model | 0.0474 +/- 0.0044 | 1.1413 +/- 0.1381 | 0.0483 +/- 0.0038 | 1.0932 +/- 0.1098 |

Table 1: CTR results on the news article recommendation data. The normalized CTR is with respect to a random baseline.

# 5 Conclusion

With this work, we contribute to the field of reinforcement learning by proposing a nonparametric mixture model based Thompson sampling framework. We merge the advances in the field of Bayesian nonparametrics with a state of the art MAB policy (i.e., Thompson sampling), and allow its extension to complex domains.

[1] Available at https://webscope.sandbox.yahoo.com/catalog.php?datatype=r&did=49

The proposed Bayesian algorithm provides interpretable and flexible modeling of convoluted reward functions, balancing the exploration-exploitation trade-off in complex domains. Empirical results show good cumulative regret performance of the proposed nonparametric Thompson sampling in simulated challenging models, remarkably adjusting to the complexity of the underlying bandit in an online fashion. With the ability to sequentially learn the nonparametric mixture model that best approximates the true reward distribution, the proposed method attains reduced regret. Our contribution is valuable for practical scenarios, as it avoids stringent model specifications. A future application is to practical scenarios where complex models are likely to outperform simpler parameterized models in describing real data.

### 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.

## References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved Algorithms for Linear Stochastic Bandits. 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 2312–2320. Curran Associates, Inc., 2011. URL `https://papers.nips.cc/paper/4417-improved-algorithms-for-linear-stochastic-bandits`.

[2] S. Agrawal and N. Goyal. Analysis of Thompson Sampling for the multi-armed bandit problem. *CoRR*, abs/1111.1797, 2011.

[3] S. Agrawal and N. Goyal. Thompson Sampling for Contextual Bandits with Linear Payoffs. *CoRR*, abs/1209.3352, 2012.

[4] S. Agrawal and N. Goyal. Further Optimal Regret Bounds for Thompson Sampling. *CoRR*, abs/1209.3353, 2012.

[5] 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.

[6] A. N. Burnetas and M. N. Katehakis. Optimal Adaptive Policies for Markov Decision Processes. *Mathematics of Operations Research*, 22(1):222–255, 1997. doi: 10.1287/moor.22.1.222.

[7] N. Cesa-Bianchi and S. Kakade. An Optimal Algorithm for Linear Bandits. *ArXiv e-prints*, Oct. 2011.

[8] 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.

[9] W. Chu, S.-T. Park, T. Beaupre, N. Motgi, A. Phadke, S. Chakraborty, and J. Zachariah. A Case Study of Behavior-driven Conjoint Analysis on Yahoo!: Front Page Today Module. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 1097–1104, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-495-9. doi: 10.1145/1557019.1557138.

[10] W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual Bandits with Linear Payoff Functions. In G. Gordon, D. Dunson, and M. Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL `http://proceedings.mlr.press/v15/chu11a.html`.

[11] S. J. Gershman and D. M. Blei. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1 – 12, 2012. ISSN 0022-2496. doi: https://doi.org/10.1016/j.jmp.2011.08.004.

[12] A. Gopalan and S. Mannor. Thompson Sampling for Learning Parameterized Markov Decision Processes. In P. Grünwald, E. Hazan, and S. Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 861–898, Paris, France, 03–06 Jul 2015. PMLR. URL `http://proceedings.mlr.press/v40/Gopalan15.html`.

[13] A. Gosavi. Reinforcement learning: A tutorial survey and recent advances. *INFORMS Journal on Computing*, 21(2):178–192, 2009.

[14] S. Grünewälder, J.-Y. Audibert, M. Opper, and J. Shawe-Taylor. Regret Bounds for Gaussian Process Bandit Problems. In Y. W. Teh and M. Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 273–280, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL `http://proceedings.mlr.press/v9/grunewalder10a.html`.

[15] D. P. Kingma, T. Salimans, and M. Welling. Variational Dropout and the Local Reparameterization Trick. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2575–2583. Curran Associates, Inc., 2015.

[16] 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.

[17] A. Krause and C. S. Ong. Contextual Gaussian Process Bandit Optimization. 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 2447–2455. Curran Associates, Inc., 2011.

[18] J. Langford and T. Zhang. The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 817–824. Curran Associates, Inc., 2008. URL `https://papers.nips.cc/paper/3178-the-epoch-greedy-algorithm-for-multi-armed-bandits-with-side-information`.

[19] C. Li, C. Chen, D. Carlson, and L. Carin. Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 1788–1794. AAAI Press, 2016.

[20] L. Li, W. Chu, J. Langford, and R. E. Schapire. A Contextual-Bandit Approach to Personalized News Article Recommendation. *CoRR*, abs/1003.0146, 2010.

[21] Z. C. Lipton, X. Li, J. Gao, L. Li, F. Ahmed, and L. Deng. Efficient Dialogue Policy Learning with BBQ-Networks. *ArXiv e-prints*, Aug. 2016.

[22] 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.

[23] R. M. Neal. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000. ISSN 10618600.

[24] 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.

[25] Y. Ouyang, M. Gagrani, A. Nayyar, and R. Jain. Learning Unknown Markov Decision Processes: A Thompson Sampling Approach. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1333–1342. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/6732-learning-unknown-markov-decision-processes-a-thompson-sampling-approach.pdf`.

[26] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

[27] C. Riquelme, G. Tucker, and J. Snoek. Deep Bayesian Bandits Showdown: An Empirical Comparison of Bayesian Deep Networks for Thompson Sampling. In *International Conference on Learning Representations*, 2018.

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

[29] P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, May 2010. ISSN 0364-765X. doi: 10.1287/moor.1100.0446.

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

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

[32] D. J. Russo, B. V. Roy, A. Kazerouni, I. Osband, and Z. Wen. A Tutorial on Thompson Sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018. ISSN 1935-8237. doi: 10.1561/2200000070. URL `http://dx.doi.org/10.1561/2200000070`.

[33] 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.

[34] 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.

[35] E. Snelson and Z. Ghahramani. Sparse Gaussian Processes using Pseudo-inputs. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1257–1264. MIT Press, 2006.

[36] N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 1015–1022, USA, 2010. Omnipress. ISBN 978-1-60558-907-7.

[37] Y. W. Teh and M. I. Jordan. Hierarchical Bayesian nonparametric models with applications. *Bayesian nonparametrics*, 1, 2010.

[38] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. doi: 10.1198/016214506000000302.

[39] 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.

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

[41] M. Titsias. Variational Learning of Inducing Variables in Sparse Gaussian Processes. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR.

[42] I. Urteaga and C. Wiggins. Variational inference for the multi-armed contextual bandit. In A. Storkey and F. Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 698–706, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

[43] I. Urteaga and C. Wiggins. (Sequential) Importance Sampling Bandits. *ArXiv e-prints*, Sept. 2018.

# A Nonparametric Hierarchical Mixture models

The generative process of a Pitman-Yor mixture model follows:

1. $G_0 \sim PY(\eta, \gamma_0, H)$.
2. $G_a \sim PY(d, \gamma, G_0)$, for $a \in \mathcal{A}$
3. $\theta_{a,n+1} \sim G_a$, that is

$$\begin{cases} m_{a,l}|m_{a,1:l-1}, \gamma_0, H \sim \sum_{k=1}^{K} \frac{M_k - \eta}{M+\gamma_0} \delta_{\theta_k} + \frac{\gamma_0 + K\eta}{M+\gamma_0} H \\ \theta_{a,n+1}|\theta_{a,1:n_a}, d, \gamma, G_0 \sim \sum_{l=1}^{L_a} \frac{n_{a,l}-d}{n_a+\gamma} \delta_{\theta_{m_{a,l}}} + \frac{\gamma + L_a d}{n_a+\gamma} G_0 \end{cases} \tag{19}$$

4. $y_{n+1}|\theta_{a,n+1} \sim f(y|\theta_{a,n+1})$

where $m_{a,l}$ refer to the per $a \in \mathcal{A}$ assignments to local clusters $l_a \in \mathcal{L}_a$, each with mixture assignment $k \in \mathcal{K}$. For parametric measures, we write $H_0(\theta) = H(\theta|\Theta_0)$ and $H_n(\theta) = H(\theta|\Theta_n)$, where $\Theta_0$ are the prior hyperparameters of the distribution and $\Theta_n$, the posterior parameters after $n$ observations. Note again that the Hierarchical Dirichlet process is a particular case of the above with $d = 0$.

The Gibbs sampler for inference of the above model after observations $y_{1:n}$ relies on the conditional distribution of observation assignments $c_n$ to local clusters $l \in \mathcal{L}_a$,

$$\begin{cases} p(c_{a,n+1} = l|y_{a,n+1}, y_{a,1:n}, c_{a,1:n}, \gamma, \gamma_0, H) \propto \frac{n_{a,l}-d}{n_a+\gamma} \int_\theta f(y_{a,n+1}|\theta_{m_{a,l}}) H_n(\theta) \mathrm{d}\theta \\ p(c_{a,n+1} = l_{new}|y_{a,n+1}, y_{a,n}, c_{a,1:n}, \gamma, \gamma_0, H) \propto \frac{\gamma + Kd}{n_a+\gamma} \int_\theta f(y_{a,n+1}|\theta_{m_{a,l_{new}}}) H(\theta) \mathrm{d}\theta \\ \qquad \propto \frac{\gamma+Kd}{n_a+\gamma} \left[ \sum_{k=1}^{K} \frac{M_k - \eta}{M+\gamma_0} \int_\theta f(y_{a,n+1}|\theta_k) H_n(\theta) \mathrm{d}\theta + \frac{\gamma_0 + K\eta}{M+\gamma_0} \int_\theta f(y_{a,n+1}|\theta_{k_{new}}) H(\theta) \mathrm{d}\theta \right] \end{cases} \tag{20}$$

and mixture assignments for a local cluster:

$$\begin{cases} p(m_{a,l} = k|y_{1:n}, c_{n \setminus n_{a,l}}, \gamma_0, H) \propto \frac{M_k - \eta}{M+\gamma_0} \int_\theta f(Y_{a,l}|\theta_k) H_{n \setminus n_{a,l}}(\theta) \mathrm{d}\theta \\ p(m_{a,l} = k_{new}|y_{1:n}, c_{n \setminus n_{a,l}} \gamma_0, H) \propto \frac{\gamma_0 + K\eta}{M+\gamma_0} \int_\theta f(Y_{a,l}|\theta_{k_{new}}) H(\theta) \mathrm{d}\theta \end{cases} \tag{21}$$

$$\begin{cases} p(m_{a,l_{new}} = k|y_{a,n+1}, y_{a,1:n}, c_{a,1:n}, \gamma_0, H) \propto \frac{M_k - \eta}{M+\gamma_0} \int_\theta f(y_{a,n+1}|\theta_k) H_n(\theta) \mathrm{d}\theta \\ p(m_{a,l_{new}} = k_{new}|y_{a,n+1}, C_N, \gamma_0, H) \propto \frac{\gamma_0 + K\eta}{M+\gamma_0} \int_\theta f(y_{a,n+1}|\theta_{k_{new}}) H(\theta) \mathrm{d}\theta \end{cases} \tag{22}$$

where with $n \setminus n_{a,l}$ we refer to all but those assigned to local cluster $l$ in set $a$.

The alternative hierarchical nonparametric mixture model MAB is illustrated in Fig. 4.

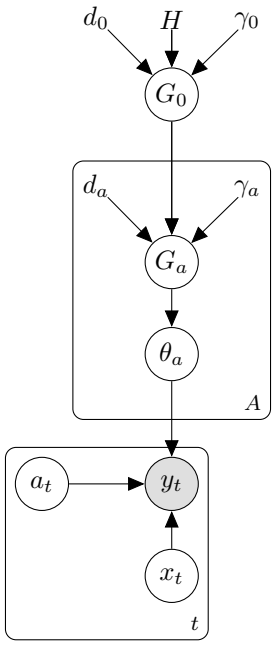
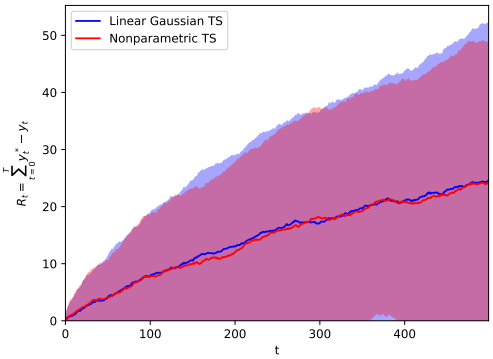
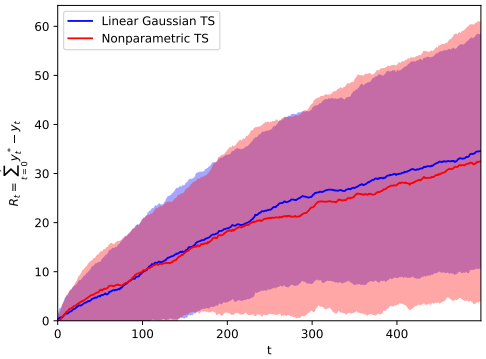



Figure 4: Graphical model of the hierarchical nonparametric mixture bandit distribution.

# B Evaluation of linear Gaussian bandits

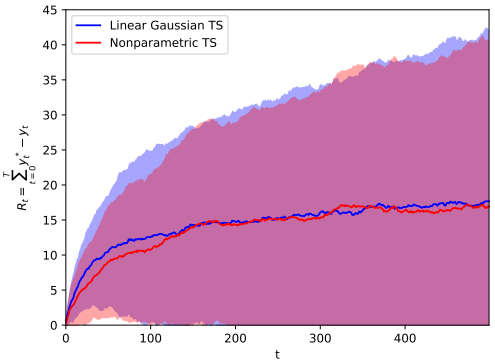



(a) 2-armed contextual linear Gaussian bandit with $\theta_{0,i} = -0.1, \theta_{1,i} = 0.1, \sigma_i^2 = 1 \forall i$.

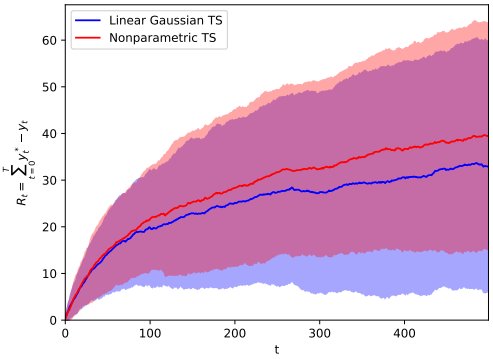



(b) 3-armed contextual linear Gaussian bandit with $\theta_{0,i} = -0.1, \theta_{0,i} = 0, \theta_{2,i} = 0.1, \sigma_i^2 = 1 \forall i$.

Figure 5: Mean regret (standard deviation shown as shaded region) for linear Gaussian bandits.

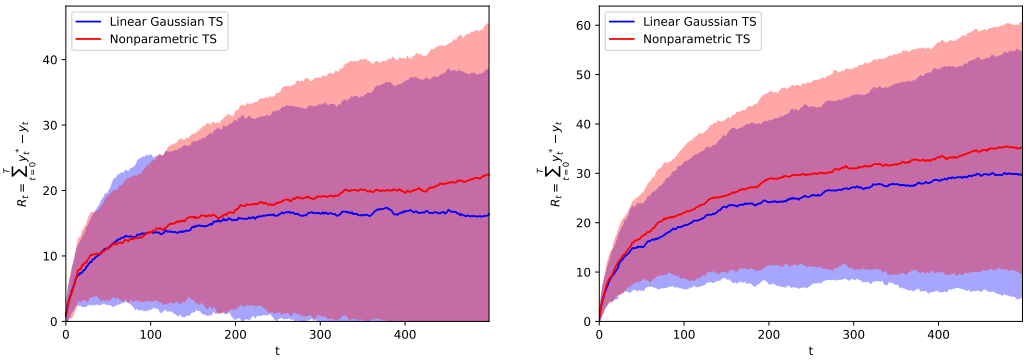



(a) 2-armed contextual linear Gaussian bandit with $\theta_{0,i} = -0.5, \theta_{1,i} = 0.5, \sigma_i^2 = 1 \forall i$.




(b) 3-armed contextual linear Gaussian bandit with $\theta_{0,i} = -0.5, \theta_{1,i} = 0, \theta_{2,i} = 0.5, \sigma_i^2 = 1 \forall i$.

Figure 6: Mean regret (standard deviation shown as shaded region) for linear Gaussian bandits.




(a) 2-armed contextual linear Gaussian bandit with $\theta_{0,i} = -1, \theta_{1,i} = 1, \sigma_i^2 = 1 \forall i$.

(b) 3-armed contextual linear Gaussian bandit with $\theta_{0,i} = 1, \theta_{1,i} = 0, \theta_{2,i} = 1, \sigma_i^2 = 1 \forall i$.

Figure 7: Mean regret (standard deviation shown as shaded region) for linear Gaussian bandits.