# 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

April 12, 2021

**Abstract**

We here adopt Bayesian nonparametric mixture models to extend multi-armed bandits in general, and Thompson sampling in particular, to scenarios where there is reward model uncertainty. In the stochastic multi-armed bandit, where an agent must learn a policy that maximizes long term payoff, the reward for the selected action is generated from an unknown distribution. Thompson sampling is a generative and 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 knowledge of the true reward model, for calculation of expected rewards and sampling from its parameter posterior. In this work, we extend Thompson sampling to complex scenarios where there is model uncertainty, by adopting a very flexible set of reward distributions: Bayesian nonparametric Gaussian mixture models. The generative process of Bayesian nonparametric mixtures naturally aligns with the Bayesian modeling of multi-armed bandits: the nonparametric model autonomously determines its complexity as new rewards are observed for the played arms. By characterizing each arm's reward distribution with independent nonparametric mixture models, the proposed method sequentially learns the model that best approximates the true underlying reward distribution, achieving successful performance in complex —not in the exponential family— bandits. Our contribution is valuable for practical scenarios, as it avoids stringent case-by-case model specifications and hyperparameter tuning, yet attains reduced regret in diverse bandit settings.

## 1 Introduction

Sequential decision making aims to optimize interactions with the world (exploit), while simultaneously learning how the world operates (explore). The origins of the study of the exploration-exploitation trade-off can be traced back to the beginning of the past century, with important contributions within the field of statistics by Thompson [1935] and later Robbins [1952]. The multi-armed bandit (MAB) is a natural abstraction for a wide variety of real-world challenges that require learning while simultaneously maximizing rewards [Lattimore

and Szepesvári, 2020]. The name 'bandit' finds its origin in the playing strategy one must devise when facing a row of slot machines [Lai and Robbins, 1985]. The contextual MAB, 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 algorithms has flourished [Agrawal and Goyal, 2012, Maillard et al., 2011], attracting interest from industry as well, due to its impact in digital advertising and products [Li et al., 2010].

Thompson [1933] sampling provides an elegant approach that tackles the exploration-exploitation dilemma in MABs. It updates a posterior over expected rewards for each arm, and chooses actions based on the probability that they are optimal. It has been empirically and theoretically proven to perform competitively for MAB models within the exponential family [Agrawal and Goyal, 2013b,a, Korda et al., 2013]. Its applicability to the more general reinforcement learning setting of Markov decision processes [Burnetas and Katehakis, 1997] has recently tracked momentum as well [Gopalan and Mannor, 2015, Ouyang et al., 2017].

Thompson sampling, and the Bayesian approach to the MAB problem, facilitate not only generative and interpretable modeling, but sequential and batch processing as well. A Thompson sampling policy requires access to posterior samples of the model. Unfortunately, maintaining such posterior is intractable for distributions not in the exponential family [Russo et al., 2018]. Therefore, developing practical MAB methods to balance exploration and exploitation in real-life domains that might not pertain to such reward family remains largely unsolved.

In an effort to extend Thompson sampling to more complex scenarios, researchers have considered other flexible reward functions and Bayesian inference. Recent approaches have embraced Bayesian neural networks and approximate inference for Thompson sampling. Variational methods, stochastic mini-batches, and Monte Carlo techniques have been studied for uncertainty estimation of reward posteriors [Blundell et al., 2015, Kingma et al., 2015, Lipton et al., 2018, Osband et al., 2016, Li et al., 2016].

Riquelme et al. [2018] have benchmarked such techniques and reported that neural networks with approximate inference, even if successful for supervised learning, underperform in the MAB setting. In particular, they emphasize the issue of adapting the slow convergence uncertainty estimates of neural network based methods to MABs. In parallel, others have focused on extending Thompson sampling by targeting alternative classes of reward functions, such as approximating the unknown bandit reward functions with Gaussian mixture models [Urteaga and Wiggins, 2018]; or maintaining and incrementally updating an ensemble of plausible models that approximates the (otherwise intractable) posterior distribution of interest [Lu and Roy, 2017].

An alternative view of Thompson sampling relies on the notion that posterior sampling can be viewed as a perturbation scheme that is sufficiently optimistic. This point was noted by Agrawal and Goyal [2013a,b], and several authors have analyzed how estimating the bandit reward means with a follow-the-perturbed-leader exploration approach can be successful in the bandit setting. Bootstrapping techniques that use a combination of observed and artificially generated data have been introduced for the multi-armed bandit and reinforcement learning problems by Osband and Roy [2015], Eckles and Kaptein [2019]. Bootstrapping over artificial data induces a prior distribution that is critical for effective exploration. Recently, pseudo-rewards based bootstrapping has also been studied for the multi-armed bandit setting [Kveton et al., 2019b,a], where the pseudo-rewards are used to increase the variance of the bootstrap mean, leading to exploration. Kveton et al. [2019b] show how these pseudo-rewards introduce bias that has to be controlled, which their proposed algorithm achieves, resulting in sublinear regret for Bernoulli bandits.

In our work, we explore a different route, in which instead of following the perturbation scheme view of posterior sampling, we focus on the Bayesian generative modeling view of Thompson sampling. Even if, for bandit regret minimization, proper modeling of the full reward distributions may not be in general necessary, we defend that a statistical modeling-based approach, which leverages the advances on nonparametric density estimation within statistics, can be performant in the multi-armed bandit setting.

We argue that modeling bandit reward distributions via nonparametric Bayesian mixtures, which adjust to the complexity of the underlying reward model, can provide successful bandit performance. Our contribution is on exploiting Bayesian nonparametric mixture models for Thompson sampling to perform MAB optimization. To that end, we propose to combine Thompson Sampling with nonparametric Bayesian mixture models that can accommodate continuous reward functions, and develop a Thompson sampling algorithm that —without incurring on model misspecification— adapts to a wide variety of complex bandits.

Bayesian nonparametrics have been considered for MAB problems to accommodate continuous actions via Gaussian processes (GPs) [Srinivas et al., 2010, Grünewälder et al., 2010, Krause and Ong, 2011], or to allow for an unknown yet countable number of actions via hierarchical Pitman-Yor processes [Battiston et al., 2018]. GPs are powerful nonparametric methods for modeling distributions over continuous functions [Rasmussen and Williams, 2005], and have been used to model a continuum of MAB actions [Krause and Ong, 2011]. Exact inference with GPs is computationally demanding —it scales cubically in the number of observations— limiting their applicability to the online setting, even if advancements such as pseudo-observations [Snelson and Ghahramani, 2006] or variational inference [Titsias, 2009] can mitigate these shortcomings. Alternatively, Battiston et al. [2018] consider MABs with a discrete but unknown action space, and propose a hierarchical Pitman-Yor process for the unknown populations, with per-arm Bernoulli reward distributions. In this work, we are not interested in a nonparametric prior over arms (with specific per-arm reward distributions), but in MABs with a discrete set of actions, for which there is uncertainty on the per-arm reward model.

We propose to account for reward model uncertainty by combining the flexibility of Bayesian nonparametrics with the large hypothesis space of mixture models. 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 [Gershman and Blei, 2012]. Nonparametric processes are useful priors for Bayesian density estimation. Within such framework, one uses nonparametric prior distributions over the mixing proportions, such as Dirichlet or Pitman-Yor processes [Teh and Jordan, 2010].

These models do not only avoid explicitly specifying the number of mixtures, but allow for an unbounded number of mixtures to appear as data are observed. The important issue of nonparametric posterior consistency, with convergence guarantees for a wide class of mixture models, has already been settled [Ghosal et al., 1999, Ghosal and van der Vaart, 2001, Lijoi et al., 2004, Ghosal and van der Vaart, 2007].

We here model each of the MAB arm reward functions with per-arm nonparametric mixture models, i.e., the complex unknown mapping of the observed per-arm rewards is estimated with nonparametric Gaussian mixture models. By means of a Bayesian nonparametric model, we can accurately approximate continuous reward distributions, yet have analytically tractable inference and online update rules, which allow for sequential adjustment of the complexity of the model to the observed data. For learning such a nonparametric distribution within the MAB setting, we leverage the well-established advances in Markov chain Monte Carlo methods for Bayesian nonparametric models [Neal, 2000].

It is both the combination of nonparametric Bayesian mixture models with Thompson sampling (i.e., merging statistical advances with a state-of-the art bandit algorithm), as well as the resulting flexibility and generality (i.e., avoiding model misspecification) that is novel in this work. We note that the generative interpretation of Bayesian nonparametric processes aligns well with the sequential nature of the MAB problem. To the best of our knowledge, no other work uses Bayesian nonparametric mixtures to model per-arm reward functions in contextual MABs.

Our specific contributions are:

1. To propose a unique, yet flexible Thompson sampling-based bandit method that learns the Bayesian nonparametric mixture model that best approximates the true, but unknown, underlying reward distribution per-arm, adjusting its complexity as it sequentially observes data.

2. An asymptotic regret bound for the proposed Thompson sampling algorithm, which assumes a Dirichlet process Gaussian mixture model prior, of order $O(|\mathcal{A}| \log^{\kappa} T\sqrt{T})$; where $|\mathcal{A}|$ denotes the number of bandit arms, $T$ the number of agent iterations with the environment, and the constant $\kappa \geq 0$ depends on the tail behavior of the true reward distribution and the priors of the Dirichlet process.

3. To demonstrate empirically that the proposed nonparametric Thompson sampling method:

   (a) attains reduced regret in complex MABs —with different unknown per-arm distributions not in the exponential family— when compared to state-of-the art baseline bandit algorithms; and

   (b) is as good as an Oracle (i.e., one that knows the true underlying model class) that implements a Thompson sampling policy, i.e., the proposed per-arm nonparametric posterior densities quickly converge to the true unknown distributions, incurring in minimal additional bandit regret.

These contributions are valuable for bandit scenarios in the presence of model uncertainty, i.e., in real-life. The same algorithm —which automatically adjusts its complexity to the observed bandit data— is run for complex (not in the exponential family) multi-armed bandits. The proposed Thompson sampling method avoids hyperparameter tuning and case-by-case reward model design choices (bypassing model mispecification) yet attains reduced regret.

# 2 Background

## 2.1 Multi-armed bandits

A multi-armed bandit (MAB) is a real time sequential decision process in which, at each interaction with the world, an agent selects an action (i.e., arm) $a \in \mathcal{A}$, where $\mathcal{A}$ is the set of arms of the bandit, according to a policy targeted to maximize cumulative rewards over time. The rewards observed by the agent are independent and identically distributed (i.i.d.) from the true outcome distribution: $Y \sim p^*(Y) = \mathbb{P}(Y = y)$[1] is a stochastic reward, where

[1]The argument of a distribution denotes the random variable, which is capitalized, its realizations are denoted in lower-case.

$p^*(Y)$ is the joint probability distribution of rewards, itself randomly drawn from a family of distributions $\mathcal{P}$. We denote with $p_a^*(Y) = p^*(Y|a)$ the conditional reward distribution of arm $a$, from which outcomes $Y_{t,a}$ are drawn: $Y_{t,a} \sim p^*(Y|a) = \mathbb{P}\left(Y = y|a_t = a\right)$.

These distributions are often parameterized by $\theta \in \Theta$, i.e., $\mathcal{P} = \{p(Y|\theta)\}_{\theta\in\Theta}$, where the true reward distribution corresponds to a unique $\theta^* \in \Theta$, i.e., $p^*(Y) = p(Y|\theta^*)$. Without loss of generality, we relate to the parametric notation hereafter, and in a Bayesian view of MABs, specify a prior with hyperparameter $\Phi$ over the parameter distribution $p(\theta|\Phi)$ when necessary.

In the contextual MAB, one must decide which arm $a_t$ to play at each time $t$, based on the available context (i.e., $x_t \in \mathcal{X}$) where the observed reward for the played arm $y_{t,a_t}$ is drawn from the unknown reward distribution of arm $a_t$ conditioned on the context,

$$Y_{t,a_t} \sim p(Y|a_t, x_t, \theta^*) \ . \tag{1}$$

Given the true model $p(Y|x_t, \theta^*)$, the optimal action is to select

$$a_t^* = \underset{a'\in\mathcal{A}}{\operatorname{argmax}}\, \mu_{t,a'}(x_t, \theta^*) \ , \tag{2}$$

where

$$\mu_{t,a}(x_t, \theta^*) = \mathbb{E}_{p(Y|a,x_t,\theta^*)}\left\{Y\right\} \tag{3}$$

is the conditional expectation of rewards with respect to the true distribution $p(Y|a, x_t, \theta^*)$ of each arm $a$, given the context $x_t$ at time $t$, and true parameter $\theta^*$.

The challenge in (contextual) MABs is the lack of knowledge about the reward-generating distribution, i.e., uncertainty about $\theta^*$ induces uncertainty about the true optimal action $a_t^*$. One needs to simultaneously learn the properties of the reward distribution —its expected value, at a minimum— and sequentially decide which action to take next.

We use $\pi(A)$ to denote a bandit policy, which is in general stochastic (i.e., $A$ is a random variable) on its choices of arms: $\pi(A) = \mathbb{P}\left(A = a\right), \forall a \in \mathcal{A}$. MAB policies choose the next arm to play towards maximizing (expected) rewards, based upon the history observed. 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 $y_{1:t} \equiv (y_{1,a_1}, \cdots, y_{t,a_t})$.

The goal of a policy is to maximize its cumulative reward, or equivalently, to minimize the cumulative regret (the loss incurred due to not knowing the best arm $a_t^*$ at each time $t$), i.e., $r_T = \sum_{t=1}^{T}\left(y_{t,a_t^*} - y_{t,a_t}\right)$, where $a_t$ denotes the arm picked by the policy $\pi(A)$ at time $t$. In the stochastic MAB setting, we study the expected cumulative *frequentist* regret at time horizon $T$,

$$R_T = \mathbb{E}\left\{\sum_{t=1}^{T}\left(Y_{t,A_t^*} - Y_{t,A_t}\right)\right\} = \mathbb{E}_{p(Y|\theta^*),\pi(A_t^*),\pi(A_t)}\left\{\sum_{t=1}^{T} Y_{t,A_t^*} - Y_{t,A_t}\right\} \ , \tag{4}$$

where the expectation is taken over the randomness of the outcomes $Y$, for a given true parametric model $p(Y|\theta^*)$, and the arm selection policies $\pi(\cdot)$: $\pi(A_t^*) = \mathbb{P}\left(A_t^* = a_t^*\right)$ denotes the optimal policy, $\pi(A_t) = \mathbb{P}\left(A_t = a_t\right)$ denotes an stochastic bandit policy. For clarity of notation, we drop the dependency on context $x_t$ from $\pi(\cdot) = \pi(\cdot|x_t)$ and $p(Y|\theta^*) = p(Y|x_t, \theta^*)$, as these are fixed and observed for all $t = \{1, \cdots, T\}$.

A related notion of regret, where the uncertainty in the true bandit model is averaged over an assumed prior $\theta^* \sim p(\theta^*|\Phi)$, is known as the expected cumulative *Bayesian* regret at time horizon $T$,

$$\begin{aligned}\mathbb{E}_{p(\theta^*|\Phi)}\left\{R_T\right\} &= \mathbb{E}_{p(\theta^*|\Phi)}\left\{\mathbb{E}\left\{\sum_{t=1}^{T}\left(Y_{t,A_t^*} - Y_{t,A_t}\right)\right\}\right\} \\ &= \mathbb{E}_{p(\theta^*|\Phi)}\left\{\mathbb{E}_{p(Y|\theta^*),\pi(A_t^*),\pi(A_t)}\left\{\sum_{t=1}^{T} Y_{t,A_t^*} - Y_{t,A_t}\right\}\right\} , \end{aligned} \tag{5}$$

and has been considered by many for the analysis of Thompson sampling [Bubeck and Liu, 2013, Russo and Roy, 2014, 2016]. Note that, as pointed out by [Agrawal and Goyal, 2013a], a regret bound on the frequentist sense implies the same bound on Bayesian regret, but not vice-versa.

## 2.2 Thompson sampling

In this work, we focus on Thompson sampling (TS) [Thompson, 1933, Russo et al., 2018], a stochastic policy that chooses what arm to play next in proportion to its probability of being optimal, given the history up to time $t$, i.e.,

$$\pi\left(A_t\right) = \pi_p\left(A_t|x_t, \mathcal{H}_{1:t-1}\right) = \mathbb{P}_p\left(A_t = a_t^*|x_t, \mathcal{H}_{1:t-1}\right) . \tag{6}$$

We specifically denote with a subscript $_p$ the parametric model class $p = p(Y|\theta)$ assumed by a Thompson sampling policy $\pi_p\left(\cdot\right)$. In a Bayesian view of MABs, the uncertainty over the reward model —the unknown parameter $\theta$— is accounted for by modeling it as a random variable with an appropriate prior $p(\theta|\Phi)$ with hyperparameters $\Phi$ (we will omit the hyperparameters of the prior when it is clear from context).

The goal in Thompson sampling is to compute the probability of an arm being optimal by marginalizing over the posterior probability distribution of the model parameter $\theta$ after observing history $\mathcal{H}_{1:t}$,

$$\begin{aligned}\pi_p\left(A_t|x_t, \mathcal{H}_{1:t-1}\right) &= \mathbb{P}_p\left(A_t = a_t^*|x_t, \mathcal{H}_{1:t-1}\right) \\ &= \int \mathbb{P}_p\left(A_t = a_t^*|x_t, \mathcal{H}_{1:t-1}, \theta\right) p(\theta|\mathcal{H}_{1:t-1})\mathrm{d}\theta \\ &= \int \mathbb{1}\left[A_t = \underset{a' \in \mathcal{A}}{\operatorname{argmax}}\, \mu_{t,a'}(x_t, \theta)\right] p(\theta|\mathcal{H}_{1:t-1})\mathrm{d}\theta . \end{aligned} \tag{7}$$

With the above integral, the uncertainty over the parameter posterior of the assumed model class given history $\mathcal{H}_{1:t-1}$ is marginalized. However, the challenge with the integral in Eqn. (7) is that it cannot be solved exactly, even when the parameter posterior $p(\theta|\mathcal{H}_{1:t-1})$ is analytically tractable over time.

Instead, Thompson sampling draws a random parameter sample $\theta^{(t)}$ from the updated posterior $p(\theta|\mathcal{H}_{1:t-1})$, and picks the arm that maximizes the expected reward given such drawn parameter sample,

$$\pi_p\left(A_t|x_t, \mathcal{H}_{1:t-1}\right) = \mathbb{1}\left[A_t = \underset{a' \in \mathcal{A}}{\operatorname{argmax}}\, \mu_{t,a'}(x_t, \theta^{(t)})\right] , \theta^{(t)} \sim p(\theta|\mathcal{H}_{1:t-1}) . \tag{8}$$

Computing the reward expectations above, as well as drawing posterior parameters, is attainable in closed form for reward models $p(Y|\theta)$ within the exponential family [Korda et al., 2013, Russo et al., 2018].

In practice however, knowledge of the true reward model is illusory. In the following, we propose Bayesian nonparametric mixture models per-arm, as tractable yet performant distributions for estimating unknown reward densities in MAB settings where there is uncertainty about the true reward model.

## 2.3 Bayesian nonparametric mixture models

A Bayesian nonparametric model is a Bayesian model on an infinite-dimensional parameter space, typically chosen as the set of possible solutions for a learning problem of interest [Müller et al., 2015]. For instance, in regression problems, the parameter space can be the set of continuous functions —e.g., specified via a prior correlation structure in Gaussian process regression [Rasmussen and Williams, 2005]; and in density estimation problems, the hypothesis space can consist of all the densities with continuous support —e.g., a Dirichlet Gaussian mixture model prior [Escobar and West, 1995].

A Bayesian nonparametric model uses only a finite subset of the available parameter dimensions to explain a finite sample of observations, with the set of dimensions adjusted according to the observed sample, such that the effective complexity of the model (as measured by the number of dimensions used) adapts to the data. In Gaussian process regression, the correlation structure or kernel function is refined as we observe more samples; in density estimation, Dirichlet process mixtures adapt the number of mixands to the complexity of the observed data. Therefore, classic adaptive problems, such as nonparametric estimation and model selection, can be formulated as Bayesian inference problems. Here, we leverage Bayesian nonparametric mixture models as a powerful density estimation framework that adjust model complexity in response to the observed data [Ghosal and der Vaart, 2017].

Bayesian nonparametric mixture models describe countably infinite mixture distributions, characterizing a very flexible model class suited for many practical settings [Ghosal and der Vaart, 2017]. These models provide a natural and flexible approach to density estimation, where the data are modeled as samples from (potentially infinite) mixtures of densities. The combination of mixture models with Bayesian nonparametric priors embodies a large hypothesis space, which can arbitrarily approximate continuous distributions [Ghosal et al., 1999, Ghosal and van der Vaart, 2001, Lijoi et al., 2004, Ghosal and van der Vaart, 2007].

A nonparametric Bayesian approach to density estimation starts with a prior on densities. As in a Parzen window-based density estimator, a useful and desirable property of a nonparametric Bayesian density estimation technique is the smoothness of the resulting empirical density, which requires a prior that can generate smooth posterior distributions. In Dirichlet mixture processes [Escobar and West, 1995], one leverages a mixing distribution that is random (e.g., the Dirichlet process) as a prior for the mixtures, inducing a nonparametric posterior that is flexible for density estimation [Ghosal, 2010]. Pólya Trees [Mauldin et al., 1992] are another set of priors on probability distributions that can generate both discrete and piecewise continuous densities, depending on the choice of parameters (the Dirichlet process is a special parameterization of a Pólya tree). Another generalization of the Dirichlet process, called the Pitman-Yor process, has been successful in modeling power-law data. We refer to [Gershman and Blei, 2012] for a detailed literature review of a variety of Bayesian nonparametric alternatives, and how they can be used in practice.

A Pitman-Yor process is a stochastic process whose sample path is a probability distribu-

tion, i.e., it is a Bayesian nonparametric model from where a drawn random sample is an infinite discrete probability distribution. We succinctly summarize the generative process and the basics for its inference here, and refer the interested reader to Teh and Jordan [2010] for further details.

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

- Parameters are drawn from the Pitman-Yor process, i.e.,

$$\varphi_n \sim G = PY(d, \gamma, G_0) \ , \tag{9}$$

  where $G_0$ is the base measure. We write $G_0(\varphi) = G(\varphi|\Phi_0)$ and $G_n(\varphi) = G(\varphi|\Phi_n)$ for the prior and posterior distributions of the parameter set $\varphi$, respectively. $\Phi_0$ are the prior hyperparameters of the base emission distribution, and $\Phi_n$ the posterior hyperparameters, after $n$ observations.

  The Pitman-yor process gives rise to a discrete random measure $G$, with ties among the $\varphi_n$s, which naturally define parameter groupings (clusters). Consequently, the Pitman-Yor process can be equivalently described as

$$\varphi_{n+1}|\varphi_{1:n}, d, \gamma, G_0 \sim \sum_{k=1}^{K} \frac{n_k - d}{n+\gamma}\delta_{\varphi_k} + \frac{\gamma + Kd}{n+\gamma}G_0 \ , \tag{10}$$

  where $\delta_{\varphi_k}$ is the Dirac delta function located at parameter atom $\varphi_k$, $n_k$ refers to the number of observations assigned to mixture component $k$, and $n = \sum_{k=1}^{K} n_k$. After $n$ observations, there are $K$ already '*seen*' clusters, and a non-zero probability $\frac{\gamma+Kd}{n+\gamma}$ of observing a '*new*' mixture component $k_{new}$ drawn from the base measure $G_0$.

- The $n+1$th observation $y_{n+1}$ is drawn from the emission distribution parameterized by the parameters of its corresponding mixture component, i.e., $Y_{n+1} \sim p(Y|\varphi_{n+1})$.

The Pitman-Yor process is a generalization of the well studied Dirichlet process, which can be readily obtained from Eqn. (10) by using $d = 0$. The discount parameter $d$ 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 Bayesian nonparametric models, one incorporates auxiliary latent variables $z_n$. These are $K+1$ dimensional categorical variables, where $z_n = k$ if observation $y_n$ is drawn from mixture component $k$.

The joint posterior of this auxiliary assignment variables $z_{1:n}$ factorizes as

$$p(z_{1:n}|\gamma) = \prod_{i=1}^{n} p(z_i|z_{1:i-1}, \gamma) \ . \tag{11}$$

To compute the full joint likelihood of the Pitman-Yor process assignments and observations, one must consider its emission distribution with hyperparameters $\Phi$, which factorizes as

$$p(y_{1:n}, z_{1:n}|\gamma, \Phi) = p(y_{1:n}|z_{1:n}, \Phi)p(z_{1:n}|\gamma) \ . \tag{12}$$

For inference, given observations $y_{1:n}$, of the unknown latent variables and parameters, one derives a Gibbs sampler that iterates between sampling mixture assignments $z_{1:n}$, and updating the emission distribution parameter posterior $G_n(\varphi)$.

The conditional distributions of observation assignment $z_n$ to already drawn mixture components $k \in \{1, \cdots, K\}$, and a new '*unseen*' mixand $k_{new}$ follow

$$\begin{cases} p(z_{n+1} = k | y_{n+1}, y_{1:n}, z_{1:n}, \gamma, G_0) \propto \frac{n_k - d}{n+\gamma} \int_\varphi p(y_{n+1}|\varphi) G_n(\varphi) \mathrm{d}\varphi \ , \\ p(z_{n+1} = k_{new} | y_{n+1}, y_{1:n}, z_{1:n}, \gamma, G_0) \propto \frac{\gamma + Kd}{n+\gamma} \int_\varphi p(y_{n+1}|\varphi) G_0(\varphi) \mathrm{d}\varphi \ . \end{cases} \tag{13}$$

Note that, after $n$ observations $y_{1:n}$, the nonparametric posterior contains $K$ already '*seen*' clusters, and accommodates a non-zero probability $\frac{\gamma + Kd}{n+\gamma}$ of a new mixture component $k_{new}$ (drawn from the base measure $G_0$) that may explain the complexity of the newly observed data $y_{n+1}$ best.

Given these mixture assignments, one updates the parameter posteriors conditioned on $z_{1:n}$ and observations $y_{1:n}$, based on the specific emission distributions and priors $G_n(\varphi) = G(\varphi | y_{1:n}, z_{1:n}, \Phi_0)$. For analytical convenience, one often resorts to emission distributions and their conjugate priors. These determine the computation of the predictive distribution $p(Y|\Phi) = \int_\varphi p(Y|\varphi) G(\varphi|\Phi) \mathrm{d}\varphi$ involved in solving Eqn. (13). [Teh and Jordan, 2010] provide a detailed explanation of the Gibbs sampling based inference procedure.

# 3 Bayesian nonparametric Thompson sampling

We combine Bayesian nonparametric mixture models with Thompson sampling for MABs under model uncertainty. We consider an independent set of nonparametric mixture models $G_a$ per-arm (with their own hyperparameters $d_a$,$\gamma_a$, and base measure $G_{a,0}$) allowing for flexible, potentially different, reward distributions for each arm $a \in \mathcal{A}$ of the MAB. The graphical model of the Bayesian nonparametric bandit is rendered in Figure 1, where we assume complete independence of each arm's reward distribution.[2]

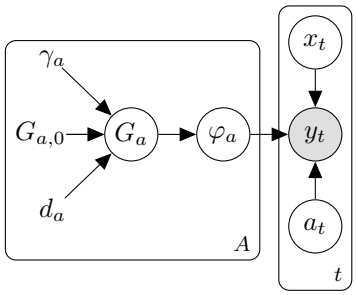



Figure 1: The Bayesian nonparametric mixture bandit, as a probabilistic graphical model.

By characterizing each arm of the bandit with a different nonparametric model, we enjoy full flexibility to estimate each per-arm distribution independently, covering MAB cases with distinct reward model classes per-arm. This setting is a very powerful extension of the MAB problem, which has not attracted interest so far, yet can circumvent model misspecification.

At every interaction of the MAB agent with the environment, reward $y_{t,a_t}$ is i.i.d. drawn from a context dependent unknown distribution $Y_{t,a_t} \sim p(Y|a_t, x_t, \theta^*)$ of the played arm $a_t$, which we here approximate via Bayesian nonparametric mixture models [Ghosal and der Vaart, 2017].

[2] An alternative model would be to consider a hierarchical nonparametric model [Teh et al., 2006, Teh and Jordan, 2010], where all arms are assumed to obey the same family of distributions, but only their mixture proportions vary across arms. We provide details of this alternative model in Section A of the Appendix.

Specifically, we model context-conditional reward densities with nonparametric Gaussian mixtures per-arm, i.e.,

$$Y_{t,a} \sim p(Y|a, x_t, \varphi_a) = \sum_{k=1}^{K_a} \frac{n_{a,k} - d_a}{n_a + \gamma_a} \cdot \mathcal{N}\left(Y|x_t^\top w_{a,k}, \sigma_{a,k}^2\right) + \frac{\gamma_a + K_a d_a}{n_a + \gamma_a} \mathcal{N}\left(Y|x_t^\top w_{a,k_{new}}, \sigma_{a,k_{new}}^2\right) , \quad (14)$$

where the number of mixands $K_a$ is determined by independent per-arm Pitman-Yor processes: $n_{a,k}$ refers to the rewards observed after playing arm $a$ that are assigned to mixture $k$, and $n_a = \sum_{k=1}^{K_a} n_{a,k}$. After $n_a$ observations for arm $a$, there are $K_a$ already '*seen*' mixtures, and a probability of $\frac{\gamma_a + K_a d_a}{n_a + \gamma_a}$ of incorporating a new mixand $k_{new}$ to the mixture.

Eqn. (14) describes per-arm nonparametric Gaussian mixture densities, with a Pitman-Yor nonparametric prior as described in Eqn. (10). Each per-arm distribution is modeled independently with per-arm specific parameterizations: $d_a$, $\gamma_a$, $\varphi_{a,k} = \{w_{a,k}, \sigma_{a,k}^2\}$, for $k = 1, \cdots, K_a$.

The proposed contextual nonparametric model relies on leveraging context-conditional Gaussian models (with an expected value that is linearly dependent on the context at time $t$, i.e., $\mu_{t,a,k} = x_t^\top w_{a,k}$), and extending them to a potentially infinite mixture. As the number of mixands $K_a$ grows, the nonparametric distribution can be non-linear in the context.

With the proposed per-arm nonparametric mixture of Gaussian densities, we make a very flexible reward model assumption that automatically adjusts to the observed data: we are nonparametrically estimating complex, unknown per-arm continuous reward densities. We leverage the well known linear Gaussian model and allow for the nonparametric model to accommodate as many mixands as necessary to best describe the observed bandit data.

The proposed Bayesian nonparametric model provides a flexible approach to density estimation, which can arbitrarily approximate continuous distributions. The Bayesian nonparametric literature has already established strong convergence results on the density estimation properties of these models: for a wide class of continuous distributions, the nonparametric posterior converges to the true data-generating density, under mild regularity conditions [Ghosal et al., 1999, Lijoi et al., 2004, Tokdar, 2006, Ghosal and van der Vaart, 2007, Bhattacharya and Dunson, 2010, Pati et al., 2013].

In theory, the proposed Bayesian nonparametric model is on an infinite-dimensional parameter space (i.e., the Pitman-Yor process can accommodate countably infinite mixands). In practice, the model as in Eqn. (14) will use a finite subset of the available parameter dimensions to explain a finite sample of observations: i.e., it sets the number of mixands per-arm $K_a$ according to the observed per-arm rewards. Consequently, the effective complexity of the resulting model (i.e., the dimensionality $K_a$ in Eqn. (14)) adapts to the observed data.

## 3.1 Nonparametric context-conditional Gaussian mixture model posterior

We now derive the procedure for inference of the per-arm, context-dependent reward posterior density of the proposed Bayesian nonparametric Gaussian mixture model. As outlined in Section 2.3, we rely on auxiliary latent variables per-arm $z_{1:n_a}$, and implement a Gibbs sampler that iterates between sampling mixture assignments $z_{1:n_a}$, and updating the emission distribution parameter posterior $G_{n_{a,k}}(\varphi_{a,k})$ for each arm and mixture.

We start with the derivation of the parameter posteriors. Per-arm and per-mixand emission distributions in Eqn. (14) are context-conditional Gaussian densities

$$\mathcal{N}\left(Y|x^\top w_{a,k}, \sigma^2_{a,k}\right) \ , \tag{15}$$

where $x^\top w_{a,k}$ and $\sigma^2_{a,k}$ are the means and variances, respectively, of the $k$-th mixand of arm $a$ in round $t$. The conjugate prior of each of the mixands is a Normal-inverse Gamma,

$$G_{a,0}(\varphi_a) = \mathcal{NIG}\left(w_a, \sigma^2_a|U_{a,0}, V_{a,0}, \alpha_{a,0}, \beta_{a,0}\right) \ , \tag{16}$$

with hyperparameters $\Phi_{a,0} = \{U_{a,0}, V_{a,0}, \alpha_{a,0}, \beta_{a,0}\}$.

After observing rewards $y_{1:n}$, and conditioned on the auxiliary assignment variables $z_{1:n_a}$, the posteriors of per-arm and mixand parameters $\varphi_{a,k}$ follow a Normal-inverse Gamma distribution with updated hyperparameters $\Phi_{a,k,n_{a,k}}$:

$$\begin{aligned} G_{a,n_{a,k}}(\varphi_{a,k}) &= \mathcal{NIG}\left(w_{a,k}, \sigma^2_{a,k}|\Phi_{a,k,n_{a,k}}\right) \ , \\ \Phi_{a,k,n_{a,k}} &= \{U_{a,k,n_{a,k}}, V_{a,k,n_{a,k}}, \alpha_{a,k,n_{a,k}}, \beta_{a,k,n_{a,k}}\} \ , \end{aligned} \tag{17}$$

that depend on the number $n_{a,k}$ of rewards observed after playing arm $a$ that are assigned to mixand $k$. Specifically,

$$\begin{cases} V^{-1}_{a,k,n_{a,k}} = x_{1:n} R_{a,k} x^\top_{1:n} + V^{-1}_{a,0} \ , \\ U_{a,k,n_{a,k}} = V_{a,k,n_{a,k}} \left(x_{1:n} R_{a,k} y_{1:n} + V^{-1}_{a,0} U_{a,0}\right) \ , \\ \alpha_{a,k,n_{a,k}} = \alpha_{a,0} + \frac{1}{2}\mathrm{tr}\left\{R_{a,k}\right\} \ , \\ \beta_{a,k,n_{a,k}} = \beta_{a,0} + \frac{1}{2}\left(y^\top_{1:n} R_{a,k} y_{1:n}\right) + \frac{1}{2}\left(U^\top_{a,0} V^{-1}_{a,0} U_{a,0} - U^\top_{a,k,n_{a,k}} V^{-1}_{a,k,n_{a,k}} U_{a,k,n_{a,k}}\right) \ , \end{cases} \tag{18}$$

where $R_{a,k} \in \mathbb{R}^{n_a \times n_a}$ is a sparse diagonal matrix with elements $[R_{a,k}]_{i,i} = \mathbb{1}[a_i = a, z_i = k]$ for $i = \{0, \cdots, n_a\}$, and $n_a = \sum_{k=1}^{K_a} n_{a,k}$ is the number of rewards observed after playing arm $a$. The number of mixands per-arm $K_a$ of the bandit is independently drawn from its own Pitman-Yor process. Note that the above expression can be computed sequentially as data are observed for the played arm.

The predictive emission distribution after marginalization of the parameters $\varphi_{a,k}$, needed for solving Eqn. (13), follows a conditional Student-t distribution

$$\begin{aligned} p_{a,k}(Y|a, x, \Phi_{a,k,,n_{a,k}}) &= \mathcal{T}\left(Y|\nu_{a,k,n_{a,k}}, m_{a,k,n_{a,k}}, r_{a,k,n_{a,k}}\right) \ , \\ \text{with } \Phi_{a,k,,n_{a,k}} &= \begin{cases} \nu_{a,k,n_{a,k}} = 2\alpha_{a,k} \ , \\ m_{a,k,n_{a,k}} = x^\top U_{a,k} \ , \\ r^2_{a,k,n_{a,k}} = \frac{\beta_{a,k}}{\alpha_{a,k}}(1 + x^\top V_{a,k} x) \ . \end{cases} \end{aligned} \tag{19}$$

The hyperparameters $\Phi_{a,k,,n_{a,k}} = \{\nu_{a,k,n_{a,k}}, m_{a,k,n_{a,k}}, r_{a,k,n_{a,k}}\}$ above are those of the prior $\Phi_{a,0}$, or the posterior $\Phi_{a,k,n_{a,k}}$, depending on whether the predictive density refers to a '*new*' mixand $k_{new}$ with $n_{a,k=k_{new}} = 0$, or a '*seen*' mixand $k$, for which $n_{a,k} \geq 0$ observations have been already assigned to, respectively.

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

follows the matrix t-distribution

$$p(Y_{a,k}|X_{a,k}, X_{\setminus a,k}, Y_{\setminus a,k}, \Phi_{a,k}) = \mathcal{MT}\left(Y_{a,k}|\nu_{Y_{a,k}}, M_{Y_{a,k}}, \Psi_{Y_{a,k}}, \Omega_{Y_{a,k}}\right) ,$$
$$\text{with } \begin{cases} \nu_{Y_{a,k}} = 2\alpha_{a,k} , \\ M_{Y_{a,k}} = X_{a,k}^{\top} U_{a,k} , \\ \Psi_{Y_{a,k}} = I_{n_{a,k}} + X_{a,k}^{\top} V_{a,k} X_{n_{a,k}} , \ \Omega_{Y_{a,k}} = 2\beta_{a,k} . \end{cases} \tag{20}$$

With parameter posteriors as in Eqns. (19) and (20), we implement a Gibbs sampler to infer the mixture assignments $z_{1:n}$, based on the assignment probabilities described in Eqn. (13), for per-arm already drawn mixture components $k_a \in \{1, \cdots, K_a\}$, and a new '*unseen*' mixand $k_{a,new}$. Therefore, the proposed Gibbs sampler adjusts the nonparametric posterior's complexity (i.e., number of mixands $K_a$) according to the observed per-arm rewards distribution.

## 3.2 Nonparametric Gaussian mixture model based Thompson sampling

We leverage the nonparametric context-conditional Gaussian mixture model described above, and combine it with a posterior sampling MAB policy, i.e., Thompson sampling [Russo et al., 2018]. The proposed Thompson sampling technique for contextual bandits with nonparametric Gaussian mixture reward models is presented in Algorithm 1.

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

**Input:** Number of arms $|\mathcal{A}|$
**Input:** Per-arm hyperparameters $d_a$, $\gamma_a$, $\Phi_{a,0}$
**Input:** Gibbs convergence criteria $\epsilon$, $Gibbs_{max}$
$\mathcal{H}_1 = \emptyset$
**for** $t = 1, \cdots, T$ **do**
  Receive context $x_t$
  **for** $a = 1, \cdots, |\mathcal{A}|$ **do**
    Draw parameters from the posterior $\varphi_{a,k}^{(t)} \sim G_{a,k,n_{a,k}}(\Phi_{a,k}), \forall k$, as in Eqn. (18)
    Compute $\mu_{t,a}(x_t, \varphi_a^{(t)})$ as in Eqn. (21)
  **end for**
  Play arm $a_t = \text{argmax}_{a' \in \mathcal{A}} \mu_{t,a'}(x_t, \varphi_{a'}^{(t)})$
  Observe reward $y_t$
  $\mathcal{H}_{1:t} = \mathcal{H}_{1:t-1} \cup \{x_t, a_t, y_t\}$
  **while** NOT Gibbs convergence criteria **do**
    Update mixture assignments $z_{1:n}$ based on Eqn. (13)
    Compute sufficient statistics $n_{a,k}$
    Update parameter posteriors $\Phi_{a,k,n_{a,k}}$ based on Eqn. (18)
  **end while**
**end for**

At each interaction with the world, the proposed Thompson sampling decides which arm to play next based on a random parameter sample, drawn from the posterior nonparametric distribution updated with all the information available at time $t$.

The parameters' posterior distributions for the proposed nonparametric Gaussian mixture model are presented in Section 3.1. Specifically, for nonparametric models as in Eqn (14), one

draws per-arm and per-mixand Gaussian parameters $\varphi_{a,k}$ from the posterior distributions with updated hyperparameters $\Phi_{a,k,n_{a,k}}$ in Eqn. (18), conditioned on the mixture assignments $z_{1:n}$ determined by the Gibbs sampler in Eqn. (13), with marginalized emission densities provided in Eqns. (19) and (20).

Given the inferred sufficient statistics of the assignments (i.e., the counts $n_{a,k}$ of rewards observed for arm $a$ and assigned to mixand $k$), and the drawn posterior parameter samples $w_{a,k}^{(t)}$, one computes the expected reward for each arm of the nonparametric bandit, i.e.,

$$\mu_{t,a}(x_t, \varphi_a^{(t)}) = \sum_{k=1}^{K_a} \frac{n_{a,k} - d_a}{n_a + \gamma_a} \left( x_t^\top w_{a,k}^{(t)} \right) + \frac{\gamma_a + K_a d_a}{n_a + \gamma_a} \left( x_t^\top w_{a,k_{new}}^{(t)} \right) . \tag{21}$$

The proposed Thompson sampling policy $\pi_{\tilde{p}} \left( \tilde{A}_t | x_t, \mathcal{H}_{1:t-1} \right)$, with assumed per-arm nonparametric distribution $\tilde{p}(Y|a, x_t, \varphi_a)$ in Eqn (14), picks the arm that maximizes the above expected reward, i.e.,

$$\begin{aligned} \pi_{\tilde{p}} \left( \tilde{A}_t | x_t, \mathcal{H}_{1:t-1} \right) &= \pi_{\tilde{p}} \left( \tilde{A}_t | x_t, \varphi_a^{(t)} \right) \\ &= \mathbb{1} \left[ \tilde{A}_t = \underset{a' \in \mathcal{A}}{\operatorname{argmax}} \, \mu_{t,a} \left( x_t, \varphi_a^{(t)} \right) \right] , \varphi_a^{(t)} \sim p(\varphi_a | \mathcal{H}_{1:t-1}) , \end{aligned} \tag{22}$$

with updated hyperparameters for $\tilde{p}(\varphi_a|\mathcal{H}_{1:t-1})$ as in Eqn. (18).

### 3.2.1 Regret bound

We leverage asymptotic posterior converge rates —the rate at which the distance between two densities becomes small as the number of observation grows— to asymptotically bound the regret of the proposed nonparametric Thompson sampling algorithm.

A Thompson sampling-based policy operates according to the probability of each arm being optimal. This probability is equivalent to the expectation with respect to the joint posterior distribution of the expected rewards given history and context, $p(\mu_t|x_t, \mathcal{H}_{1:t-1})$, of the optimal arm indicator function, i.e.,

$$\pi_p \left( A_t | x_t, \mathcal{H}_{1:t-1} \right) = \mathbb{P}_p \left( A_t = \underset{a' \in \mathcal{A}}{\operatorname{argmax}} \, \mu_{t,a'} \right) = \mathbb{E}_p \left\{ \mathbb{1} \left[ A_t = \underset{a' \in \mathcal{A}}{\operatorname{argmax}} \, \mu_{t,a'} \right] \right\} .$$

Note that the indicator function $\mathbb{1} \left[ A_t = \operatorname{argmax}_{a' \in \mathcal{A}} \mu_{t,a'} \right]$ for each arm requires the posterior over all arms $a' \in \mathcal{A}$ as input. That is, the posterior $p(\mu_t|x_t, \mathcal{H}_{1:t-1})$ is the joint posterior distribution over the expected rewards of all arms: $\mu_t = \{\mu_{t,a}\}, \forall a \in \mathcal{A}$; i.e., it is a $|\mathcal{A}|$ dimensional multivariate distribution over all arms of the bandit.

We now present our first lemma, with the proof provided in Section B of the Appendix, which is key to the cumulative regret theorem that follows.

**Lemma 3.1.** *The difference in action probabilities between two Thompson sampling policies, given the same history and context up to time t, is bounded by the total-variation distance $\delta_{TV}(p_t, q_t)$ between the posterior distributions of their expected rewards at time t, $p_t = p(\mu_t|x_t, \mathcal{H}_{1:t-1})$ and $q_t = q(\mu_t|x_t, \mathcal{H}_{1:t-1})$, respectively,*

$$\pi_{p_t} \left( A_t = a \right) - \pi_{q_t} \left( A_t = a \right) \leq \delta_{TV}(p_t, q_t) .$$

*The **total variation distance** $\delta_{TV}(p,q)$ between distributions $p$ and $q$ on a sigma-algebra $\mathcal{F}$ of subsets of the sample space $\Omega$ is defined as*

$$\delta_{TV}(p,q) = \sup_{B \in \mathcal{F}} |p(B) - q(B)| \ , \tag{23}$$

*which is properly defined for both discrete and continuous distributions (see details in Section B of the Appendix).*

We make use of Lemma 3.1 to asymptotically bound the cumulative regret of the proposed Thompson sampling with Dirichlet process priors (i.e., $d_a = 0, \forall a$) and Gaussian emission distributions, for bandits with true reward densities that meet certain regularity conditions.

**Theorem 3.2.** *The expected cumulative regret at time $T$ of a Dirichlet process Gaussian mixture model based Thompson sampling algorithm is asymptotically bounded by*

$$R_T \leq \mathcal{O}\left(|\mathcal{A}| \log^{\kappa} T \sqrt{T}\right) \quad \text{as } T \to \infty \ .$$

*We use big-O notation $\mathcal{O}(\cdot)$ for the asymptotic regret bound, as it bounds from above the growth of the cumulative regret over time for large enough bandit interactions, i.e.,*

$$\lim_{T \to \infty} \frac{R_T}{|\mathcal{A}| \log^{\kappa} T \sqrt{T}} \leq \mathcal{O}(1) \ . \tag{24}$$

*We note that this bound holds both in a frequentist and Bayesian view of expected regret.*

The proof of Theorem 3.2, provided in Section B of the Appendix, consists of bounding the regret introduced by two factors: the first, related to the use of Thompson sampling (i.e., a policy that does not know the true parameters of the reward distribution, but has knowledge of the true reward model class); and the second, a term that accounts for the convergence of the posterior of a nonparametric model to that of the true data generating distribution.

The logarithmic term $\log^{\kappa} T$ in the bound appears due to the convergence rate of the nonparametric density estimation, where the exponent $\kappa \geq 0$ depends on the tail behavior of the base measure and the priors of the Dirichlet process —see Section B of the Appendix, and references therein, for details on density convergence and its impact on the exponent $\kappa \geq 0$.

### 3.2.2 Computational complexity

The Gibbs sampler in the proposed nonparametric Thompson sampling (lines 14-18 within Algorithm 1) is run $Gibbs_{steps}$ until a stopping criteria is met: either the model likelihood of the sampled chain is stable within an $\epsilon$ likelihood margin between steps, or a maximum number of iterations $Gibbs_{max}$ is reached.

As new rewards $y_{t,a_t}$ are acquired, updates to assignments $z_{t',a_t}$ are computed sequentially within the Gibbs sampler for $t' = \{1, \cdots, t | a_{t'} = a_t\}$; i.e., only the posterior over the last played armed $a_t$ is recomputed. Since Eqn. (18) can be sequentially computed for each per-arm observation, the computational cost of the Gibbs sampler grows with the number of available observation of the played arm. Therefore, the overall computational cost is upper-bounded by $\mathcal{O}(T \cdot Gibbs_{steps})$ per-interaction with the world, i.e., per newly observed reward $y_{t,a_t}$.

Due to the sequential acquisition of observations in the bandit setting, and the need to only update the posterior for the played arm, the Gibbs sampler is *warm-started* at each bandit interaction, and good convergence can be achieved in few iterations per observed reward. In practice, and because of the *warm-start*, one can limit the number of Gibbs sampler iterations per-bandit interaction to upper-bound the algorithm's complexity to $O(T \cdot Gibbs_{max})$ per interaction, yet achieve satisfactory performance —empirical evidence of this claim is provided in Section 4.2.2. Due to the *warm-start*, the Gibbs sampler is run from a good starting point: the per-arm parameter space that describes all but this newly observed reward $y_{t,a_t}$.

We emphasize that we propose a Gibbs sampler that runs until convergence, but suggest to limit the number of Gibbs iterations as a practical recommendation with good empirical regret performance, yet upper-bounded $\mathcal{O}(T \cdot Gibbs_{max})$ computational complexity per MAB interaction with the environment.

# 4 Evaluation

We evaluate the proposed nonparametric Gaussian mixture model based Thompson sampling (i.e., `Nonparametric TS` as in Algorithm 1) in diverse and complex synthetic datasets, for which practical methods that balance exploration and exploitation remain elusive. Our goal is two-pronged:

1. To compare the proposed `Nonparametric TS` algorithm to state-of-the-art alternatives (described in Section 4.1), showcasing its flexibility, generality and how it attains reduced regret in complex multi-armed bandits —Section 4.2.

2. To demonstrate that its performance is equivalent to an Oracle (i.e., one that knows the true underlying model class) that implements a Thompson sampling policy — Section 4.3.

## 4.1 Thompson sampling-based state-of-the-art baselines

Thompson sampling provides a bandit framework that requires access to posterior samples of the reward model. An approach to extend its applicability to complex domains is to leverage the advances in Bayesian neural networks and to merge them with approximate Bayesian inference methods. The algorithms listed below are state-of-the art techniques that provide different computations of (or approximations to) reward posteriors that can be combined with Thompson sampling to address contextual multi-armed bandits.

1. `LinearGaussian TS`: A powerful (not neural network based) Thompson sampling baseline, which assumes a contextual linear Gaussian reward function,

$$Y_{t,a} \sim \mathcal{N}\left(Y|x_t^\top \theta_a, \sigma_a^2\right) \ .$$

   In its simplest setting, when the reward variance is known, the resulting Thompson sampling implementation follows that by Agrawal and Goyal [2013b]. In our experiments, in a similar fashion as done by Riquelme et al. [2018], we model the joint distribution of $\theta_a$ and $\sigma_a^2$, $\forall a \in \mathcal{A}$, which allows the method to adaptively adjust to the observed reward noise. We leverage the Normal-inverse Gamma conjugate prior, i.e.,

$$\left(\theta_a, \sigma_a^2\right) \sim \mathcal{NIG}\left(\theta_a, \sigma_a^2 | m_{a,0}, \Sigma_{a,0}, \alpha_{a,0}, \beta_{a,0}\right) \ , \tag{25}$$

of the Gaussian contextual model to derive the exact Bayesian posterior. We use an uninformative prior for the variance ($\alpha_{a,0} = 1, \beta_{a,0} = 1, \forall a$) and a standard uncorrelated Gaussian for the mean prior ($m_{a,0} = 0, \Sigma_{a,0} = I, \forall a$).

2. `MultitaskGP`: This is an alternative and popular Bayesian nonparametric technique that models context to reward mappings via Gaussian processes [Srinivas et al., 2010, Grünewälder et al., 2010, Krause and Ong, 2011]. This implementation regresses the expected reward of different context-actions pairs by fitting a multitask Gaussian process given observed bandit data.

3. `NeuralLinear`: This algorithm, introduced by Riquelme et al. [2018], operates by learning a neural network that maps contexts to rewards for each action, and simultaneously, updates a Bayesian linear regression in the network's last layer. The last layer maps a learned representation $z$ linearly to the rewards $y$. The corresponding Thompson sampling draws the learned linear regression parameters $\theta_a$ for each arm, but keeps the representation $z$ output by the learned network. In our experiments, the representation network and the Bayesian linear posterior are retrained and updated at every MAB interaction.

4. `NeuralBootstrapped`: This algorithm is based on [Osband et al., 2016], which trains simultaneously (in parallel) $Q$ neural networks based on different bootstrapped bandit histories $\mathcal{H}_{1:t}^{(1)}, \cdots, \mathcal{H}_{1:t}^{(Q)}$. These are generated by adding each newly observed evidence $(x_t, a_t, y_t)$ to each history $\mathcal{H}_{1:t-1}^{(q)}$, $q = \{1, \cdots, Q\}$, independently and with probability $p \in (0, 1]$. In order to choose an action for a given context, one of the $q$ networks is selected with uniform probability $(1/Q)$, and the best action according to the selected network is played.

5. `NeuralRMS`: This is a simple bandit benchmark that trains a neural network to map contexts to rewards. At each time $t$, it acts $\epsilon$-greedily according to the current model, which due to the Stochastic Gradient Descent (SGD) algorithm used for training (i.e., the RMSProp optimizer in this implementation), captures randomness in its output.

6. `NeuralDropoutRMS`: Dropout is a neural network training technique where the output of each neuron is independently zeroed out with a given probability in each forward pass. Once the network is trained, dropout can also be used to obtain a distribution of predictions for a specific input. By choosing the best action with respect to the random dropout prediction, an implicit form of Thompson sampling is implemented.

7. `NeuralParamNoise`: An approach to approximate a distribution over neural networks consists in randomly perturbing the point estimates attained by SGD on the available data Plappert et al. [2018]. In this case, the model uses a heuristic to control the amount of i.i.d. noise it adds to the parameters of the neural network, which is used for making a decision, but not for training steps: SGD re-starts from the last, noiseless parameter value.

8. `BNNVariationalGaussian`: This algorithm is based on ideas presented in [Blundell et al., 2015] that combine stochastic variational inference and Bayes by backpropagation. It implements a Bayesian neural network by modeling each individual weight posterior as a univariate Gaussian distribution. Thompson sampling then draws a network at each time step by sampling each weight independently. The variational approach

consists in maximizing a proxy for the maximum likelihood estimates of the network weights given observed data, to fit the unknown parameters of the variational posterior.

9. `BNNAlphaDiv`: This technique leverages expectation-propagation and Black-box alpha-divergence minimization as in Hernandez-Lobato et al. [2016] to approximate the unknown reward distribution. The algorithm iteratively approximates the posterior of interest by updating a single approximation factor at a time, which usually corresponds to the likelihood of one data point. The implementation adopted here optimizes the global objective directly via stochastic gradient descent.

10. `Optimal`: When possible (i.e., for simulated bandits for which there is ground truth), we implement an optimal multi-armed bandit policy that selects the best arm for the given context, based on the known true expected reward of each arm.

As reported by Riquelme et al. [2018], deep learning methods are very sensitive to the selection of a wide variety of hyperparameters, and these hyperparameter choices are known to be highly dataset dependent. Besides, in bandit scenarios, we do not have access to each problem to perform any a-priori tuning. Therefore, we resort to the default settings provided by the authors in their implementation.

A full description of these algorithms and all implementation details can be found in the original deep bandit showdown work. For completeness, we present the set of hyperparameters used for our experiments in Section D of the Appendix. In addition, and following the insights from Riquelme et al. [2018] that partially optimized uncertainty estimates can lead to catastrophic decisions with neural networks, we retrain the neural network models at every iteration of the multi-armed bandit —note that since stochastic gradient descent is used for training, randomness is incorporated in all implementations.

## 4.2 Complex contextual bandits: reduced regret under model uncertainty

Our proposed nonparametric Thompson sampling is valuable for MAB scenarios in the presence of model uncertainty. In the following, we show that `Nonparametric TS`, while avoiding case-by-case reward model design choices and bypassing model mispecification, attains reduced regret across a variety of bandit scenarios.

To that end, we evaluate its performance under different reward settings: in Section 4.2.1, contextual linear Gaussian bandits [3], and in Section 4.2.2 contextual bandits with rewards not in the exponential family.

### 4.2.1 Contextual linear Gaussian bandits

We first investigate the performance of our proposed method in the classic contextual linear Gaussian MAB setting: i.e., for bandits where rewards are drawn from linearly parameterized arms, where the mean reward for context $x_t$ and arm $a$ is given by $x_t^\top \theta_a$, for some unknown latent parameter $\theta_a$.

We evaluate different parameterizations of such bandits, both for non-sparse and sparse settings (where only few components of $\theta_a$ are non-zero for each arm). Specifically, we draw a 10 dimensional $\theta_a$ uniformly at random (with only 3 components left as non-zero for each

[3] Results for non-contextual Gaussian bandits are provided in Section C of the Appendix, showcasing how the proposed method attains satisfactory performance.

arm in the sparse setting), with contexts distributed as a 10-dimensional standard Gaussian. In all cases, Gaussian noise (with a minimum standard deviation of 0.01) is added to the rewards of each arm, for $|\mathcal{A}| = 8$.

Figures 2 and 3 showcase the five top performing algorithms in non-sparse and sparse linear contextual bandits respectively, with cumulative regret results (and attained relative regret improvements) detailed in Tables 1 and 2 for all the evaluated methods. The corresponding reward results are provided in Section E of the Appendix.

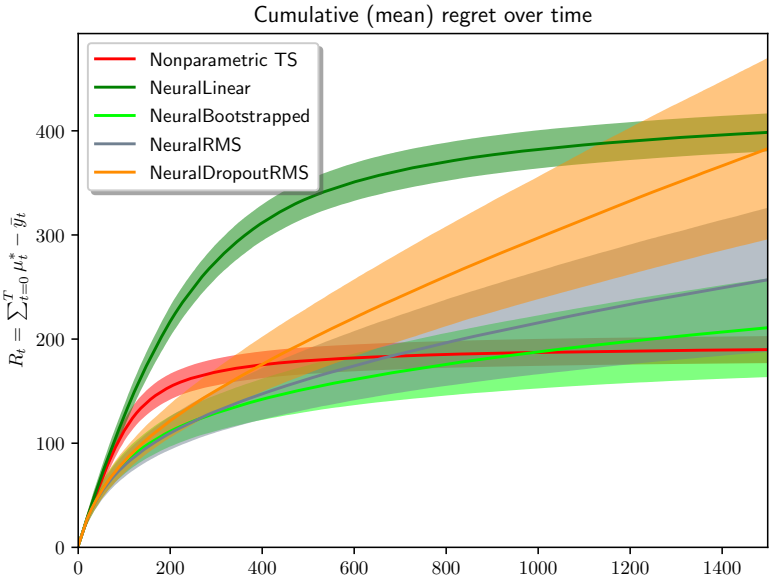



Figure 2: Mean regret (standard deviation shown as shaded region) for $R = 500$ realizations of the presented methods in contextual linear Gaussian MABs.

Table 1: Cumulative regret at $t = 1500$ for $R = 500$ realizations of contextual linear Gaussian MABs. The second column showcases the additional relative cumulative regret incurred by each algorithm when compared to `Nonparametric TS` at $t = 1500$.

| Algorithm | Cumulative regret | Relative cumulative regret |
|---|---|---|
| Nonparametric TS | 189.889 | %0.000 |
| NeuralLinear | 398.485 | %109.852 |
| NeuralBootstrapped | 210.914 | %11.073 |
| NeuralRMS | 256.897 | %35.288 |
| NeuralDropoutRMS | 382.870 | %101.629 |
| NeuralParamNoise | 241.039 | %26.937 |
| MultitaskGP | 213.648 | %12.512 |
| BNNVariationalGaussian | 827.369 | %335.712 |
| BNNAlphaDiv | 793.118 | %317.675 |

We observe a reduced cumulative regret of our proposed nonparametric Thompson sampling when compared to all state-of-the-art Thompson sampling alternatives described in Section 4.1. The proposed `Nonparametric TS` method achieves logarithmic regret, even as it estimates the true form of the underlying unknown reward function, resulting in considerably smaller regret than the alternatives —designed to be flexible for complex bandit scenarios— in the contextual linear Gaussian setting.

The neural network based linear Thompson sampling method (i.e., `NeuralLinear`) achieves logarithmic regret as well, but suffers an important loss: an average cumulative regret that doubles that of `Nonparametric TS` at $t = 1500$.

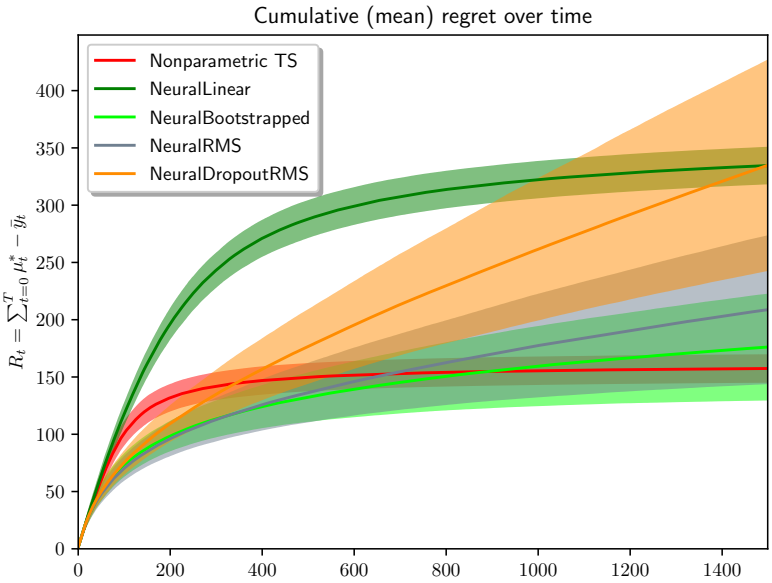



Figure 3: Mean regret (standard deviation shown as shaded region) for $R = 500$ realizations of the presented methods in sparse contextual linear Gaussian MABs.

Table 2: Cumulative regret at $t = 1500$ for $R = 500$ realizations of sparse contextual linear Gaussian MABs. The second column showcases the additional relative cumulative regret incurred by each algorithm when compared to `Nonparametric TS` at $t = 1500$.

| Algorithm | Cumulative regret | Relative cumulative regret |
|---|---|---|
| Nonparametric TS | 157.406 | %0.000 |
| NeuralLinear | 334.569 | %112.551 |
| NeuralBootstrapped | 176.136 | %11.899 |
| NeuralRMS | 208.738 | %32.611 |
| NeuralDropoutRMS | 334.870 | %112.742 |
| NeuralParamNoise | 199.122 | %26.502 |
| MultitaskGP | 186.023 | %18.180 |
| BNNVariationalGaussian | 762.321 | %384.301 |
| BNNAlphaDiv | 695.097 | %341.594 |

Bootstrapped and RMS based alternatives fail to achieve a good exploration-exploitation trade-off (their cumulative regret does not plateau by $t = 1500$), with cumulative regrets that are at least 10% and 30% higher in all contextual MABs. In addition, their performance is highly volatile across realizations of the same problem (recall their wide shaded regions in Figures 2 and 3). Specially concerning is the poor performance of all the Bayesian Neural Network (BNN) based baselines, which incur in more than %300 cumulative regret increase with respect to `Nonparametric TS`.

These results support our claim that a nonparametric Bayesian based Thompson sampling is a flexible alternative that can provide reduced regret when compared to state-of-the-art baselines.

#### 4.2.2 Contextual bandits not in the exponential family

We now study a set of more challenging contextual MABs, i.e., those where the underlying reward distributions do not fit into the exponential family assumption. Specifically, we study the following contextual bandit scenarios, where the context is randomly drawn from a two dimensional uniform distribution, i.e., $x_{i,t} \sim \mathcal{U}(0,1)$, $i \in \{0,1\}$, $t \in \mathbb{N}$.

`Scenario A`:

$$\begin{cases} p_1(y|x_t,\theta) = 0.5 \cdot \mathcal{N}\left(y|x_t^\top (0\ 0), 1\right) + \ 0.5 \cdot \mathcal{N}\left(y|x_t^\top (1\ 1), 1\right) \\ p_2(y|x_t,\theta) = 0.5 \cdot \mathcal{N}\left(y|x_t^\top (2\ 2), 1\right) + \ 0.5 \cdot \mathcal{N}\left(y|x_t^\top (3\ 3), 1\right) \end{cases}$$

`Scenario B`:

$$\begin{cases} p_1(y|x_t,\theta) = 0.5 \cdot \mathcal{N}\left(y|x_t^\top (1\ 1), 1\right) + \ 0.5 \cdot \mathcal{N}\left(y|x_t^\top (2\ 2), 1\right) \\ p_2(y|x_t,\theta) = 0.3 \cdot \mathcal{N}\left(y|x_t^\top (0\ 0), 1\right) + \ 0.7 \cdot \mathcal{N}\left(y|x_t^\top (3\ 3), 1\right) \end{cases}$$

`Scenario C`:

$$\begin{cases} p_1(y|x_t,\theta) = \mathcal{N}\left(y|x_t^\top (1\ 1), 1\right) \ , \\ p_2(y|x_t,\theta) = 0.5 \cdot \mathcal{N}\left(y|x_t^\top (1\ 1), 1\right) + \ 0.5 \cdot \mathcal{N}\left(y|x_t^\top (2\ 2), 1\right) \\ p_3(y|x_t,\theta) = 0.3 \cdot \mathcal{N}\left(y|x_t^\top (0\ 0), 1\right) + \ 0.6 \cdot \mathcal{N}\left(y|x_t^\top (3\ 3), 1\right) + \ 0.1 \cdot \mathcal{N}\left(y|x_t^\top (4\ 4), 1\right) \end{cases}$$

`Scenario D`:

$$\begin{cases} p_1(y|x_t,\theta) = 0.75 \cdot \mathcal{N}\left(y|x_t^\top (0\ 0), 1\right) + \ 0.25 \cdot \mathcal{N}\left(y|x_t^\top (0\ 0), 10\right) \\ p_2(y|x_t,\theta) = 0.75 \cdot \mathcal{N}\left(y|x_t^\top (2\ 2), 1\right) + \ 0.25 \cdot \mathcal{N}\left(y|x_t^\top (2\ 2), 10\right) \end{cases}$$

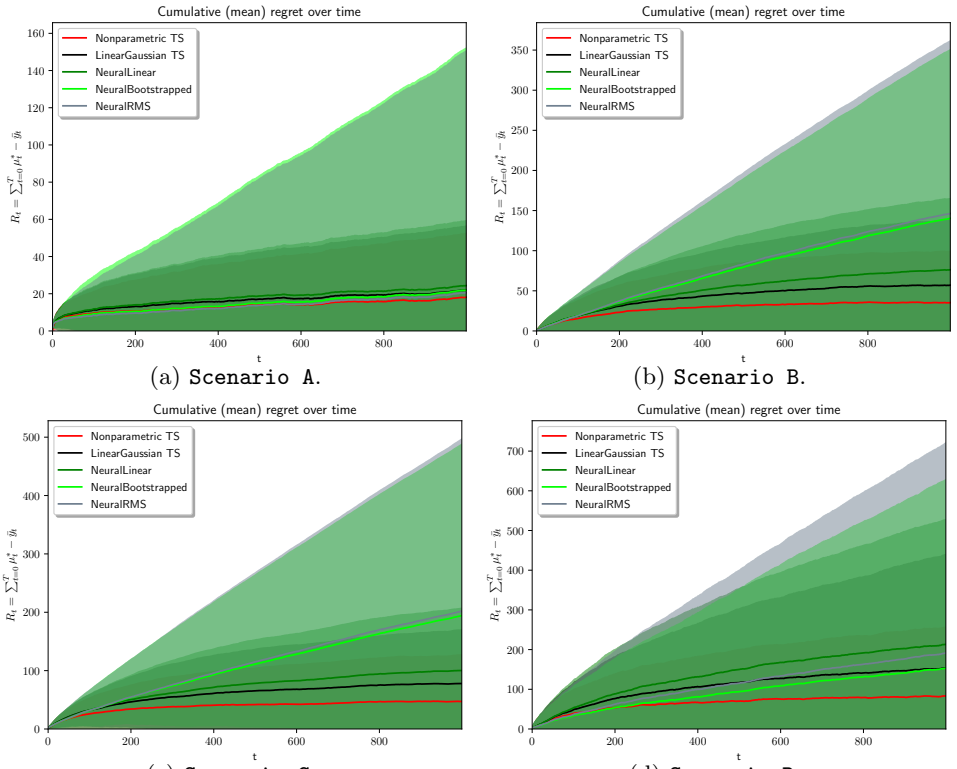
The reward distributions of these contextual bandits are all Gaussian mixtures, which differ in the amount of mixture overlap and the similarity between arms. In these scenarios, the reward functions are not within the exponential family of distributions: they are all multi-modal, unbalanced in `Scenarios B` and `C`, and with heavy tails in `Scenario D`.

`Scenario A` is a balanced mixture of two Gaussian distributions, with rewards easily separable per-arm. On the contrary, there is a significant overlap between arm rewards in `Scenario B`, with quite unbalanced mixtures for arm 2: rewards from a mixand of low expected value are drawn with probability 0.3, and higher rewards are expected with probability 0.7. `Scenario C` describes a MAB with different per-arm reward distributions: a linear Gaussian distribution for arm 1, a bi-modal Gaussian mixture for arm 2, and an unbalanced Gaussian mixture with three components for arm 3. Finally, `Scenario D` models heavy-tailed distributions, where the bandit is subject to outlier rewards.

We show in Figure 4 how our proposed method adjusts to the underlying reward model complexity, attaining reduced cumulative regret when compared to other Thompson sampling-based alternatives, in all the studied scenarios.

In `Scenario A`, even if all shown methods attain logarithmic regret (i.e., they are able to find the right exploitation-exploration balance) as shown in Figure 4a, our proposed method attains meaningful regret reductions. The attained cumulative regret results (with relative regret improvements) are detailed in Table 3, and the corresponding reward results are provided in Section F of the Appendix.

In `Scenario A`, the second best performing `NeuralRMS` baseline incurs an additional 13.5% cumulative regret when compared to `Nonparametric TS`, `LinearGaussian TS` is 21.3% worse,




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

(c) `Scenario C`. (d) `Scenario D`.

Figure 4: Mean cumulative regret (standard deviation shown as shaded region) for $R = 500$ realizations of the top-performing methods in all studied scenarios.

Table 3: Mean cumulative regret at $t = 1000$ for $R = 500$ realizations of the studied methods in all scenarios. We indicate in parentheses the additional relative cumulative regret incurred by each algorithm when compared to `Nonparametric TS`.

| Algorithm | `Scenario A` | `Scenario B` | `Scenario C` | `Scenario D` |
|---|---|---|---|---|
| Nonparametric TS | 18.105 (%0.000) | 35.268 (%0.000) | 47.172 (%0.000) | 83.464 (%0.000) |
| LinearGaussian TS | 21.957 (%21.276) | 57.04 (%61.733) | 77.921 (%65.184) | 153.955 (%84.456) |
| NeuralLinear | 24.358 (%34.538) | 76.312 (%116.377) | 100.202 (%112.415) | 213.073 (%155.286) |
| NeuralBootstrapped | 21.942 (%21.195) | 140.302 (%297.816) | 194.198 (%311.676) | 153.315 (%83.690) |
| NeuralRMS | 20.559 (%13.551) | 146.276 (%314.756) | 201.818 (%327.829) | 191.552 (%129.501) |
| NeuralDropoutRMS | 26.191 (%44.663) | 146.182 (%314.490) | 202.297 (%328.846) | 170.3 (%104.039) |
| NeuralParamNoise | 23.17 (%27.976) | 129.758 (%267.921) | 170.385 (%261.197) | 177.931 (%113.182) |
| MultitaskGP | 114.79 (%534.020) | 145.883 (%313.643) | 245.441 (%420.305) | 667.407 (%699.631) |
| BNNVariationalGaussian | 27.643 (%52.682) | 172.461 (%389.001) | 217.891 (%361.903) | 340.92 (%308.462) |
| BNNAlphaDiv | 67.969 (%275.416) | 196.028 (%455.824) | 306.14 (%548.980) | 247.998 (%197.130) |

and a regret increase of 21.2% and 34.5% is observed for the `NeuralBootstrapped` and `NeuralLinear` baselines, respectively.

`Nonparametric TS` provides important regret savings when compared to the best performing alternative in each scenario: at least an additional 61.7%, 65.2% and 83.69% cumulative regret is incurred by other baselines at the end of `Scenario B`, `Scenario C`, and `Scenario D`, respectively. The proposed Bayesian nonparametric mixture model Thompson sampling clearly outperforms all the alternatives in the studied scenarios. Besides, the volatility in

regret performance of `Nonparametric TS` is the smallest of all the studied alternatives across all scenarios.

Even if each scenario is a different MAB setting, `Nonparametric TS` is readily applicable to all, without any per-case specific fine-tuning: even in `Scenario C`, where there are different model classes per-arm. Overall, `Nonparametric TS` attains reduced regret in all the studied MABs —with different unknown per-arm distributions not in the exponential family.

On the contrary, we observe a poor performance of neural network and approximate Bayesian inference based alternatives. Specifically, we find that the linear neural alternative takes longer to reach the exploration-exploitation tradeoff. Riquelme et al. [2018] argued that `NeuralLinear` is able to simultaneously learn a good latent data representation, and to accurately quantify the uncertainty over linear models, to explain the observed rewards in terms of these learned representation. This competitive advantage allowed Riquelme et al. [2018] to successfully solve problems that require non-linear representations where linear approaches fail. However, we here observe a significant regret increase in comparison to our proposed nonparametric method for bandits with reward functions not in the exponential family: additional 34.54%, 116.377%, 112.415% and 155.29% regret is incurred in `Scenarios A`, `B`, `C`, and `D`, respectively.

In addition, bootstrapped and RMS based neural Thompson sampling techniques struggle to find a good exploration-exploitation balance, incurring in additional (and very volatile) regret performance, as shown in Figure 4. As noted by Riquelme et al. [2018], (*i*) bootstrapping incurs in a heavy computational overhead, depending on the number of networks considered; (*ii*) a noise-injection based method is hard to tune and sensitive to the heuristic controlling the injected noise-level; and (*iii*) dropout algorithms heavily depend on their hyperparameters, and it is unclear how to disentangle better training from better exploration in these models. On the contrary, `Nonparametric TS` achieves reduced regret across all scenarios without any case-specific fine-tuning.

Alternative nonparametric methods, such as those based on Gaussian processes also suffer in all the studied scenarios (see Table 3). We argue that the low performance of this method is explained by model misspecification: Gaussian process regression methods require knowledge of the true underlying model class (i.e., what mean and kernel functions to use) —if the reward model class of the MAB they are targeting is not correctly specified, then increased regret is attained. Note that, in `Nonparametric TS`, no per-MAB fine tuning is required.

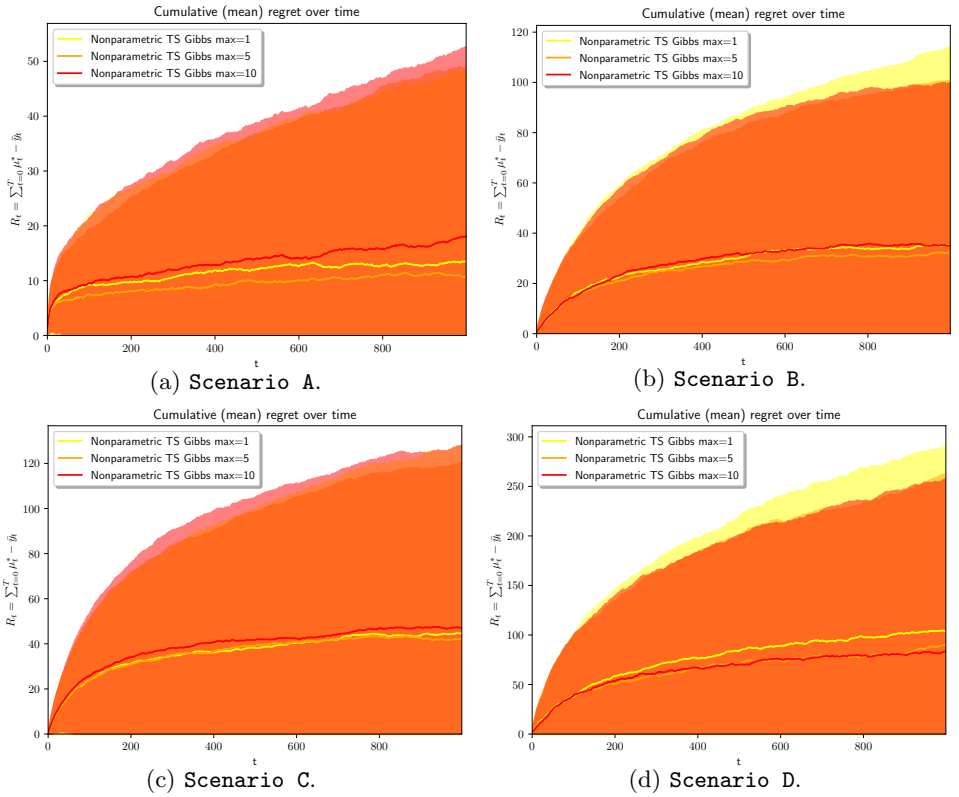
Variational inference and expectation-propagation based algorithms also perform poorly in all the studied scenarios (see Table 3). As reported by Riquelme et al. [2018], while optimizing to convergence at every interaction with the world incurs increased computational cost, results suggest that it may not be sufficient to partially optimize the variational parameters in bandit settings. Similarly, for Black-Box $\alpha$-divergence, partial convergence may be the cause of poor performance. Overall, there is a need for investigating how to sidestep the slow convergence of the uncertainty estimates in bandit settings for these neural network based methods.

The volatility of all the evaluated neural network based alternatives is noteworthy, and worrisome: their performance is highly variable when compared to the proposed nonparametric method (reward standard deviation details are provided in Section F of the Appendix.).

On the contrary, the performance of `Nonparametric TS` is much less volatile. For real-life bandit applications, the variance of an algorithm's performance is critical: even if expected reward guarantees are in place, the understanding of how volatile a bandit algorithm's decisions are, has undeniable real-world impact.

All in all, we conclude that our proposed nonparametric Thompson sampling outperforms (both in averaged cumulative regret, and in regret volatility) other alternatives across all the studied scenarios. Due to the capacity of Bayesian nonparametrics to autonomously adjust the complexity of the model to the sequentially observed data, the proposed method can, without any per-scenario tuning ($Gibbs_{max} = 10$, $d_a = 0$, $\gamma_a = 0.1$, $\forall a$, in all the experiments), readily target all of them, and attain considerable regret reductions.

Important savings are achieved for MAB settings not in the exponential family (i.e., unbalanced and heavy tailed distributions) and for bandits with different per-arm reward distributions, where we observe that alternative methods struggle. The competitive advantage lies on the capacity of Bayesian nonparametrics to adjust the complexity of the posterior density to the sequentially observed bandit data.




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

(c) `Scenario C`. (d) `Scenario D`.

Figure 5: Mean regret (standard deviation shown as shaded region) for $R = 500$ realizations of the proposed `Nonparametric TS` method with different $Gibbs_{max}$ in all scenarios.

Finally, we investigate the *warm-start* effect in the proposed algorithm's Gibbs sampling procedure, and how the practical recommendations on limiting the number of Gibbs iterations of Section 3.2.2 impact regret performance.

In general, and because of the incremental availability of observations in the bandit setting, we observe that the proposed Gibbs sampler achieves quick convergence: in all our experiments, a 1% log-likelihood relative difference between iterations is usually achieved within $Gibbs_{max} \leq 10$ iterations. We show in Figure 5 that no significant regret performance improvement is achieved by letting the sampler run for a high number of iterations.

*A good enough* posterior convergence at a limited computational budget —at most $Gibbs_{max}$ updates over all $t_a$ observations for the played arm— is possible because the Gibbs sampler is run, at each interaction with the world, from a good starting point: the per-arm parameter space that describes all but this newly observed reward.

Therefore, when computational constraints are appropriate, e.g., for real-time bandits, limiting the number of Gibbs iterations allows for a drastic reduction in execution-time (see run-time details in Section G of the appendix), yet still achieving satisfactory cumulative regret.

## 4.3 Nonparametric TS compared to Oracle TS

In the previous section, we showed that the proposed `Nonparametric TS` outperforms state-of-the-art Thompson sampling alternatives. Our aim now is to determine how optimal the proposed nonparametric Bayesian based technique is.

To fully reveal the flexibility the proposed method, we scrutinize the `Nonparametric TS` algorithm by comparing it to `Oracle TS` algorithms: i.e., Oracles that know the true per-arm reward distributions of the bandit they are targeted to.

This is an unrealistic setting in practice, yet possible in a simulated environment, as knowing the reward complexity of a MAB beforehand is impractical [4].

We implement `Oracle TS` algorithms for each simulated contextual bandit setting. Results below demonstrate how, due to the flexible and general density estimation technique provided by Bayesian nonparametric models, the per-arm nonparametric posterior densities converge to the true unknown distribution, allowing for our `Nonparametric TS` method to incur in minimal additional regret when compared to `Oracle TS` alternatives in all the studied scenarios.

### 4.3.1 Contextual linear Gaussian bandits: Oracle TS

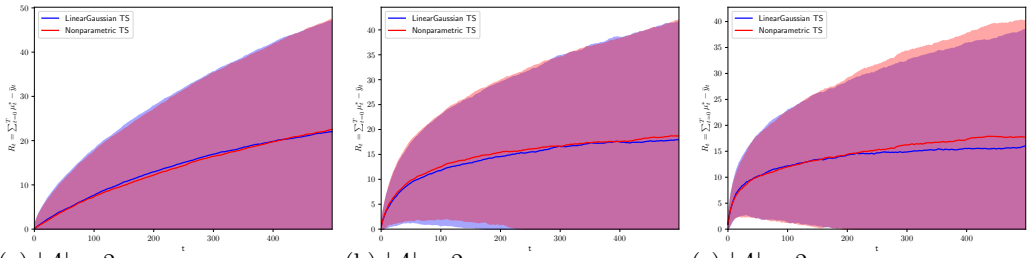
We showcase the flexibility of our proposed method in the contextual linear Gaussian MAB setting first, where we can readily compare its performance to a well studied `Oracle TS`: i.e., the linear Gaussian Thompson sampling in Agrawal and Goyal [2013b]. In this set-up, `LinearGaussian TS` correctly assumes the true underlying contextual linear Gaussian model $Y_{t,a} \sim \mathcal{N}\left(Y|x_t^\top \theta_a, \sigma_a^2\right)$, and can compute posterior updates in closed form.

In Figure 6, we show the mean cumulative regret of various parameterizations of multi-armed ($\mathcal{A} = \{2, 3, 4, 5\}$) contextual linear Gaussian bandits, with two-dimensional contexts randomly drawn from a uniform distribution, i.e., $x_{i,t} \sim \mathcal{U}(0, 1)$, $i \in \{0, 1\}$, $t \in \mathbb{N}$.

The proposed `Nonparametric TS` attains cumulative regret comparable to that of `LinearGaussian TS`. That is, the proposed method matches the performance of the analytical linear Gaussian posterior Thompson sampling, even as it estimates the true form of the underlying reward function.

The proposed nonparametric Thompson sampling is almost as good as the analytical alternative when there is no model mismatch, as in this setting: the per-arm nonparametric posterior density quickly converges to the true unknown distribution, incurring in minimal additional regret when compared to the analytical posterior based `Oracle TS` alternative.

[4] An alternative would be to run multiple model assumptions in parallel, with a subsequent model selection.



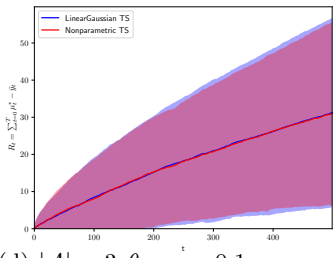

(a) $|\mathcal{A}| = 2$, $\theta_{0,i} = -0.1$, $\theta_{1,i} = 0.1$.



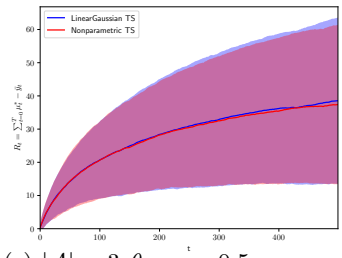

(b) $|\mathcal{A}| = 2$, $\theta_{0,i} = -0.5$, $\theta_{1,i} = 0.5$.



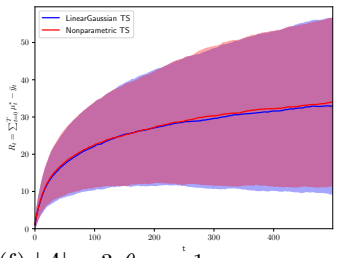

(c) $|\mathcal{A}| = 2$, $\theta_{0,i} = -1$, $\theta_{1,i} = 1.0$.



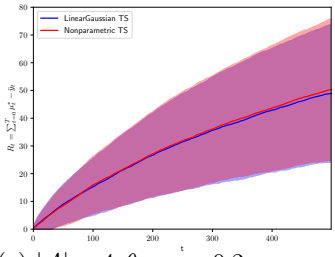

(d) $|\mathcal{A}| = 3, \theta_{0,i} = -0.1$, $\theta_{0,i} = 0, \theta_{2,i} = 0.1$.



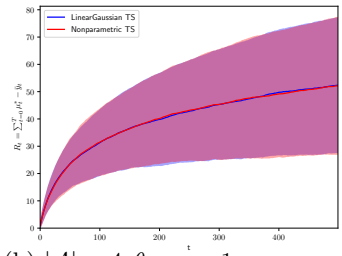

(e) $|\mathcal{A}| = 3, \theta_{0,i} = -0.5$, $\theta_{1,i} = 0, \theta_{2,i} = 0.5$.



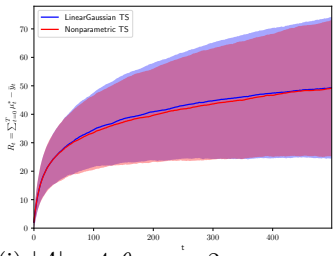

(f) $|\mathcal{A}| = 3, \theta_{0,i} = 1$, $\theta_{1,i} = 0, \theta_{2,i} = 1$.



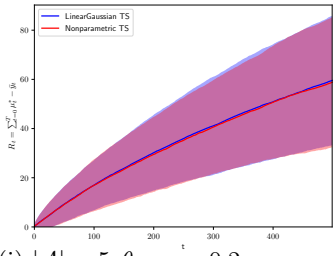

(g) $|\mathcal{A}| = 4, \theta_{0,i} = -0.2$, $\theta_{1,i} = -0.1, \theta_{2,i} = 0.1$, $\theta_{3,i} = 0.1$.



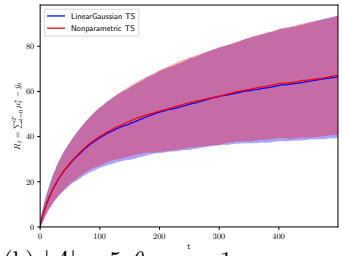

(h) $|\mathcal{A}| = 4, \theta_{0,i} = -1$, $\theta_{1,i} = -0.5, \theta_{2,i} = 0.5$, $\theta_{3,i} = 1$.



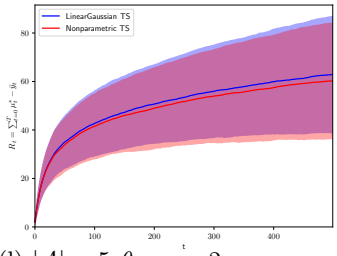

(i) $|\mathcal{A}| = 4, \theta_{0,i} = -2$, $\theta_{1,i} = -1, \theta_{2,i} = 1$, $\theta_{3,i} = 2$.




(j) $|\mathcal{A}| = 5, \theta_{0,i} = -0.2$, $\theta_{1,i} = -0.1$, $\theta_{2,i} = 0$, $\theta_{3,i} = 0.1, \theta_{4,i} = 0.1$.




(k) $|\mathcal{A}| = 5, \theta_{0,i} = -1$, $\theta_{1,i} = -0.5$, $\theta_{2,i} = 0$, $\theta_{3,i} = 0.5, \theta_{4,i} = 1$.




(l) $|\mathcal{A}| = 5, \theta_{0,i} = -2$, $\theta_{1,i} = -1$, $\theta_{2,i} = 0$, $\theta_{3,i} = 1, \theta_{4,i} = 2$.

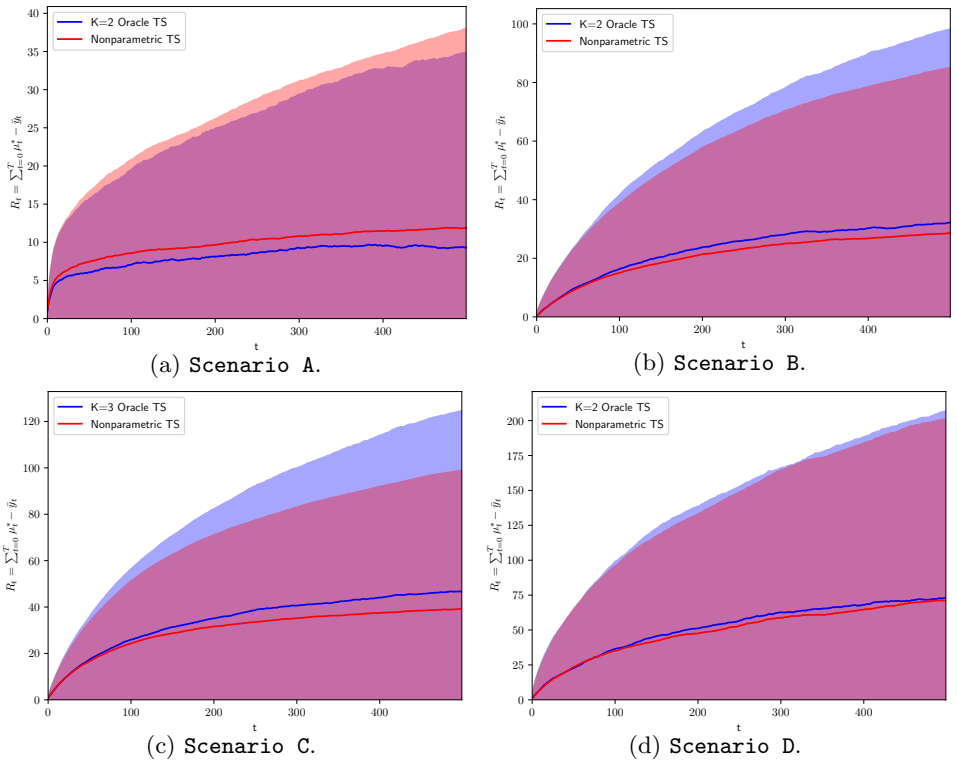
Figure 6: Mean cumulative regret (and standard deviation shown as shaded region) for $R = 1000$ realizations of different $\mathcal{A} = \{2, 3, 4, 5\}$ armed contextual linear Gaussian bandits, with $\sigma_a^2 = 1 \forall a$.

### 4.3.2 Contextual bandits not in the exponential family: Oracle TS

We further scrutinize Algorithm 1 by comparing it to `Oracle TS` algorithms for `Scenarios A`, `B`, `C` and `D`. We implement separate `Oracle TS` algorithms for each scenario, via a Dirichlet prior distribution that has knowledge of the true underlying dimensionality $K_a$ per-arm.

This approach is similar to Urteaga and Wiggins [2018], where mixtures of Gaussian distributions model per-arm reward functions. To provide a fair comparison to our proposed method, and instead of the variational inference-based original approach of Urteaga and Wiggins [2018], we implement a Gibbs sampler as described in Section 3 where for each `Oracle TS`, the correct $K_a$ per-scenario is known (instead of the Dirichlet process prior assumed by `Nonparametric TS`).

We compare the performance of our proposed nonparametric Thompson sampling to that of each per-scenario `Oracle TS` in Figure 7. The proposed nonparametric Thompson sampling provides satisfactory performance across all studied scenarios, when compared to an unrealistic `Oracle TS` that knows the true number of underlying mixtures of the problem it is targeted to.




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

(c) `Scenario C`. (d) `Scenario D`.

Figure 7: Mean regret (standard deviation shown as shaded region) for $R = 3000$ realizations of the proposed and `Oracle TS` methods.

The `Nonparametric TS` fits the underlying reward function accurately in all cases, attaining comparable regret in all scenarios. We emphasize that the `Nonparametric TS` method does not demand any per-scenario hyperparameter tuning, and avoids model misspecification: i.e., the same algorithm is used for all scenarios, while scenario specific `Oracle TS` methods are required.

These results demonstrate the competitive advantage of Bayesian nonparametrics to adjust the complexity of the reward model to the sequentially observed bandit data. The proposed nonparametric generative modeling provides per-arm reward understanding (by plotting or computing figures of merit from these distributions), as the learned per-arm posteriors converge to the true posteriors. We note that nonparametric posterior density convergence does not imply that it is consistent in $K$ —on data from a finite mixture, nonparametric posteriors do not necessarily concentrate at the true number of components [Miller and Harrison, 2014]. Nevertheless, we argue that density convergence suffices in the bandit setting.

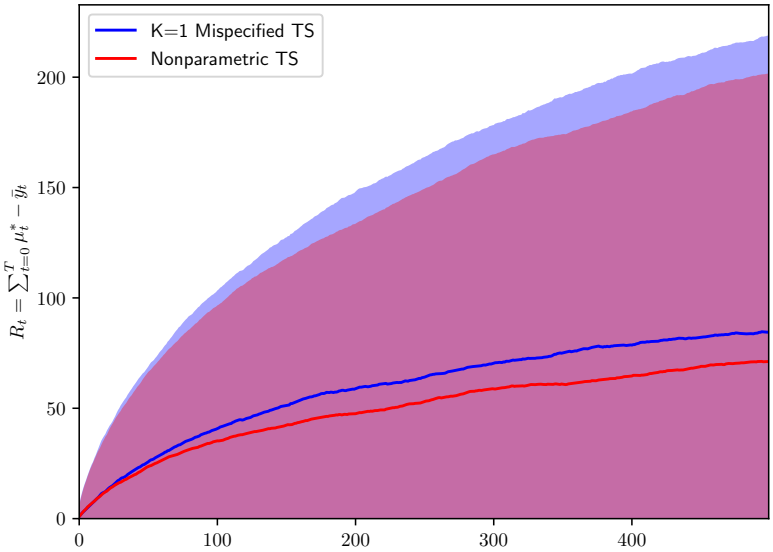



Figure 8: Mean regret (standard deviation shown as shaded region) for $R = 3000$ realizations of the proposed and `Oracle TS` methods in `Scenario D` under model misspecification.

We further highlight the built-in flexibility of the proposed nonparametric method by showing in Figure 8 how Thompson sampling with a mispecified model (i.e., fitting a unimodal Gaussian distribution to the heavy-tailed `Scenario D`) suffers in comparison to the proposed nonparametric method: a %18 cumulative regret reduction is attained by `Nonparametric TS`. With this, we reiterate the robustness and generality of our proposed nonparametric method in avoiding model-misspecification, a significant advantage for real-life bandit settings.

We conclude by emphasizing that the proposed `Nonparametric TS` avoids stringent case-by-case model assumptions for each specific MAB setting, yet attains competitive regret when faced with distinct, complex MAB reward distributions: the same algorithm (with no hyperparameter tuning) is run for contextual Gaussian bandits and other complex (not in the exponential family) multi-armed bandits.

# 5 Conclusion

We contribute to the field of sequential decision processes by proposing a Bayesian nonparametric mixture model based Thompson sampling. We merge advances in the field of Bayesian nonparametrics with a state-of-the art MAB policy (i.e., Thompson sampling), allowing for its extension to complex multi-armed bandit domains where there is model uncertainty.

The proposed algorithm provides flexible modeling of convoluted reward functions with convergence guarantees, and attains the exploration-exploitation trade-off in complex MABs with minimal assumptions.

We provide an asymptotic upper bound for the expected cumulative regret of the proposed Dirichlet process Gaussian mixture model based Thompson sampling.

In addition, empirical results show improved cumulative regret performance of the proposed nonparametric Thompson sampling in challenging domains —where there is model uncertainty— remarkably adjusting to the complexity of the underlying bandit in an online fashion —bypassing model mispecification and hyperparameter tuning.

Important savings are attained for complex bandit settings (e.g., unbalanced and heavy tailed reward distributions, and bandits with different per-arm reward distributions), where alternative methods struggle.

The competitive advantage lies on the capacity of Bayesian nonparametrics to adjust the complexity of the posterior density to the sequentially observed bandit data. With the ability to sequentially learn the Bayesian nonparametric mixture model that best approximates the true reward distribution —not necessarily in the exponential family— the proposed method can be applied to diverse MAB settings without stringent model specifications and attain reduced regret.

A future direction is to tighten the presented regret bound, as well as to apply the proposed method to real-life MAB applications where complex models are likely to outperform simpler ones.

# References


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

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

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

M. Battiston, S. Favaro, and Y. W. Teh. Multi-Armed Bandit for Species Discovery: A Bayesian Nonparametric Approach. *Journal of the American Statistical Association*, 113 (521):455–466, 2018. doi: 10.1080/01621459.2016.1261711. URL `https://doi.org/10.1080/01621459.2016.1261711`.

A. Bhattacharya and D. B. Dunson. Nonparametric Bayesian density estimation on manifolds with applications to planar shapes. *Biometrika*, 97(4):851–865, 2010.

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.

S. Bubeck and C.-Y. Liu. Prior-free and prior-dependent regret bounds for Thompson Sampling. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 638–646. Curran Associates, Inc., 2013. URL `http://papers.nips.cc/paper/`

5108-prior-free-and-prior-dependent-regret-bounds-for-thompson-sampling.pdf.

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.

D. Eckles and M. Kaptein. Bootstrap Thompson Sampling and Sequential Decision Problems in the Behavioral Sciences. *SAGE Open*, 9(2):2158244019851675, 2019. doi: 10.1177/2158244019851675. URL https://doi.org/10.1177/2158244019851675.

M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

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.

S. Ghosal. The Dirichlet process, related priors and posterior asymptotics. *Bayesian nonparametrics*, 28:35, 2010.

S. Ghosal and A. V. der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

S. Ghosal and A. van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, 35(2):697–723, 04 2007. doi: 10.1214/009053606000001271.

S. Ghosal and A. W. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29(5):1233–1263, 10 2001. doi: 10.1214/aos/1013203452. URL https://doi.org/10.1214/aos/1013203452.

S. Ghosal, J. K. Ghosh, and R. Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics*, 27(1):143–158, 1999. URL http://repository.ias.ac.in/22510/1/308.pdf.

S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 04 2000. doi: 10.1214/aos/1016218228. URL https://doi.org/10.1214/aos/1016218228.

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.

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.

J. Hernandez-Lobato, Y. Li, M. Rowland, T. Bui, D. Hernandez-Lobato, and R. Turner. Black-box alpha divergence minimization. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning,*, volume 48 of *Proceedings of Machine Learning Research*, pages 1511–1520, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `http://proceedings.mlr.press/v48/hernandez-lobatob16.html`.

I. Ibragimov and R. Has'minskii. *Statistical Estimation: Asymptotic Theory*. Springer, 1981. doi: 10.1007/978-1-4899-0027-2.

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.

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.

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.

B. Kveton, C. Szepesvari, M. Ghavamzadeh, and C. Boutilier. Perturbed-History Exploration in Stochastic Multi-Armed Bandits. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 2786–2793. International Joint Conferences on Artificial Intelligence Organization, 7 2019a. doi: 10.24963/ijcai.2019/386.

B. Kveton, C. Szepesvari, S. Vaswani, Z. Wen, T. Lattimore, and M. Ghavamzadeh. Garbage In, Reward Out: Bootstrapping Exploration in Multi-Armed Bandits. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3601–3610, Long Beach, California, USA, 09–15 Jun 2019b. PMLR. URL `http://proceedings.mlr.press/v97/kveton19a.html`.

T. L. Lai and H. Robbins. Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6(1):4–22, mar 1985. ISSN 0196-8858. doi: 10.1016/0196-8858(85)90002-8.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

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.

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.

A. Lijoi, I. Prünster, and S. G. Walker. Extending Doob's consistency theorem to nonparametric densities. *Bernoulli*, 10(4):651–663, 2004.

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.

X. Lu and B. V. Roy. Ensemble sampling. In *Advances in Neural Information Processing Systems*, pages 3258–3266, 2017.

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.

R. D. Mauldin, W. D. Sudderth, and S. C. Williams. Polya trees and random distributions. *The Annals of Statistics*, pages 1203–1221, 1992.

J. W. Miller and M. T. Harrison. Inconsistency of Pitman-Yor Process Mixtures for the Number of Components. *Journal of Machine Learning Research*, 15:3333–3370, 2014. URL `http://jmlr.org/papers/v15/miller14a.html`.

P. Müller, F. A. Quintana, A. Jara, and T. Hanson. *Bayesian nonparametric data analysis*. Springer, 2015. doi: 10.1007/978-3-319-18968-0.

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.

I. Osband and B. V. Roy. Bootstrapped Thompson sampling and deep exploration. *arXiv preprint arXiv:1507.00300*, 2015.

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.

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`.

D. Pati, D. B. Dunson, and S. T. Tokdar. Posterior consistency in conditional distribution estimation. *Journal of multivariate analysis*, 116:456–472, 2013.

M. Plappert, R. Houthooft, P. Dhariwal, S. Sidor, R. Y. Chen, X. Chen, T. Asfour, P. Abbeel, and M. Andrychowicz. Parameter Space Noise for Exploration. In *International Conference on Learning Representations*, 2018.

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

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.

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

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

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

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`.

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.

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.

Y. W. Teh and M. I. Jordan. Hierarchical Bayesian nonparametric models with applications. *Bayesian nonparametrics*, 1:158–207, 2010.

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.

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.

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

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.

S. T. Tokdar. Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics*, pages 90–110, 2006.

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.

# A Nonparametric hierarchical mixture model bandit

An alternative MAB model, where each arm is drawn from the same base distribution, is to consider a hierarchical Pitman-Yor mixture model. The generative process of a hierarchical 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. $\varphi_{a,n+1} \sim G_a$, that is

$$\begin{cases} m_{a,l}|m_{a,1:l-1}, \gamma_0, H \sim \sum_{k=1}^{K} \frac{M_k - \eta}{M+\gamma_0} \delta_{\varphi_k} + \frac{\gamma_0 + K\eta}{M+\gamma_0} H \ , \\ \varphi_{a,n+1}|\varphi_{a,1:n_a}, d, \gamma, G_0 \sim \sum_{l=1}^{L_a} \frac{n_{a,l} - d}{n_a+\gamma} \delta_{\varphi_{m_{a,l}}} + \frac{\gamma + L_a d}{n_a+\gamma} G_0 \end{cases} \tag{26}$$

   where $m_{a,l}$ refers to the per-arm $a \in \mathcal{A}$ assignments to local mixands $l_a \in \mathcal{L}_a$, each with mixture assignment $k \in \mathcal{K}$, now shared across arms. That is, there is a global mixture with $K$ mixands for the bandit, but each per-arm distribution consists of a subset of $\mathcal{L}_a \in K$ mixands.

4. The $n+1$th observation $y_{n+1}$ is drawn from the emission distribution parameterized by the parameters of its corresponding mixture component $Y_{n+1}|\varphi_{a,n+1} \sim p(Y|\varphi_{a,n+1})$.

For parametric measures, we write $H_0(\varphi) = H(\varphi|\Phi_0)$ and $H_n(\varphi) = H(\varphi|\Phi_n)$, where $\Phi_0$ are the prior hyperparameters of the emission distribution, and $\Phi_n$ the posterior parameters after $n$ observations, respectively. 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 $c_{a,n}$ to local mixands $l \in \mathcal{L}_a$,

$$\begin{cases} p(c_{a,n+1} = l|y_{a,n+1}, y_{a,1:n}, c_{a,1:n}, \gamma, \gamma_0, H) \\ \qquad \propto \frac{n_{a,l} - d}{n_a+\gamma} \int_\varphi p(y_{a,n+1}|\varphi_{m_{a,l}}) H_n(\varphi) \mathrm{d}\varphi \\ p(c_{a,n+1} = l_{new}|y_{a,n+1}, y_{a,n}, c_{a,1:n}, \gamma, \gamma_0, H) \\ \qquad \propto \frac{\gamma + Kd}{n_a+\gamma} \int_\varphi p(y_{a,n+1}|\varphi_{m_{a,l_{new}}}) H(\varphi) \mathrm{d}\varphi \\ \qquad \propto \frac{\gamma + Kd}{n_a+\gamma} \left[ \sum_{k=1}^{K} \frac{M_k - \eta}{M+\gamma_0} \int_\varphi p(y_{a,n+1}|\varphi_k) H_n(\varphi) \mathrm{d}\varphi \right. \\ \qquad\qquad \left. + \frac{\gamma_0 + K\eta}{M+\gamma_0} \int_\varphi p(y_{a,n+1}|\varphi_{k_{new}}) H(\varphi) \mathrm{d}\varphi \right] \end{cases} \tag{27}$$

and mixture assignments $m_{a,l}$ for a local mixand $l \in \mathcal{L}_a$:

$$\begin{cases} p(m_{a,l} = k|y_{1:n}, c_{n \setminus n_{a,l}}, \gamma_0, H) \propto \frac{M_k - \eta}{M+\gamma_0} \int_\varphi p(Y_{a,l}|\varphi_k) H_{n \setminus n_{a,l}}(\varphi) \mathrm{d}\varphi \\ p(m_{a,l} = k_{new}|y_{1:n}, c_{n \setminus n_{a,l}} \gamma_0, H) \propto \frac{\gamma_0 + K\eta}{M+\gamma_0} \int_\varphi p(Y_{a,l}|\varphi_{k_{new}}) H(\varphi) \mathrm{d}\varphi \end{cases} \tag{28}$$

$$\begin{cases} p(m_{a,l_{new}} = k|y_{a,n+1}, y_{a,1:n}, c_{a,1:n}, \gamma_0, H) \\ \qquad \propto \frac{M_k - \eta}{M+\gamma_0} \int_\varphi p(y_{a,n+1}|\varphi_k) H_n(\varphi) \mathrm{d}\varphi \\ p(m_{a,l_{new}} = k_{new}|y_{a,n+1}, y_{a,1:n}, c_{a,1:n}, \gamma_0, H) \\ \qquad \propto \frac{\gamma_0 + K\eta}{M+\gamma_0} \int_\varphi p(y_{a,n+1}|\varphi_{k_{new}}) H(\varphi) \mathrm{d}\varphi \end{cases} \tag{29}$$

where $n \backslash n_{a,l}$ refers to all observations but those assigned to local mixand $l$ in arm $a$, $M_k$ are the number of local mixands assigned to global mixture component $k$, and $M = \sum_{k=1}^{K} M_k$.

The alternative nonparametric MAB considers the above hierarchical nonparametric model, where all arms are assumed to obey the same family of distributions, but only their mixture proportions vary across arms, as illustrated in Figure 9. The main advantage of this alternative is that one learns per-mixture parameter posteriors based on rewards of all played arms, with the disadvantage of all arms of the bandit being of the same family of reward distributions.

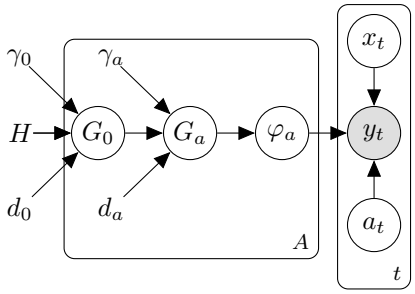



Figure 9: Graphical model of the hierarchical nonparametric mixture bandit distribution.

# B Asymptotic regret bound for nonparametric mixture based Thompson sampling

We start by clarifying the notation we use in the sequel:

- We denote the distribution $p(\Omega)$ of the random variable $\Omega$ for the probability of a random event $\omega$ with $\mathbb{P}_p(\Omega = \omega)$.
- We specify the distribution $p(\cdot)$ of the random variable within an expectation with a subscript, $\mathbb{E}_p\{\cdot\}$.
- We use $\mu_a = \mathbb{E}_p\{Y_a\}$ to indicate the expectation under some distribution $p$ of the reward for each arm $a \in \mathcal{A}$.
- We use $\mu = \{\mu_{t,a}\}, \forall a \in \mathcal{A}$ for the set of all per-arm expected values.
- We define the union of the context at time $t$ and history up to $t-1$ with $h_{1:t} = \{x_t, \mathcal{H}_{1:t-1}\}$.
- We use $(\mu_{t,a}|h_{1:t}) = \mathbb{E}_p\{Y_a|h_{1:t}\}$ to indicate the expectation under the posterior of the reward distribution $p$ of each arm $a$ given context and history $h_{1:t}$ up to time $t$.
- We denote stochastic policies with $\pi_p(\cdot)$, where the subscript makes explicit the assumed reward model class $p(\cdot)$.

- For Thompson sampling policies, we may interchangeably write

$$\begin{aligned}\pi_p\left(A_t\right) &= \pi_p\left(A_t|h_{1:t}\right) \\ &= \mathbb{P}_p\left(A_t = a_t^*|h_{1:t}\right) \\ &= \mathbb{P}_p\left(A_t = \underset{a'\in\mathcal{A}}{\operatorname{argmax}}\left(\mu_{t,a'}\middle|h_{1:t}\right)\right) \\ &= \mathbb{E}_p\left\{\mathbb{1}\left[A_t = \underset{a'\in\mathcal{A}}{\operatorname{argmax}}\left(\mu_{t,a'}\middle|h_{1:t}\right)\right]\right\} \ .\end{aligned} \tag{30}$$

- **The total variation distance** $\delta_{TV}(p,q)$ between distributions $p$ and $q$ on a sigma-algebra $\mathcal{F}$ of subsets of the sample space $\Omega$ is defined as

$$\delta_{TV}(p,q) = \sup_{B\in\mathcal{F}} |p(B) - q(B)| \ . \tag{31}$$

  When $\Omega$ is countable,

$$\delta_{TV}(p,q) = \sup_{B\in\mathcal{F}} |p(B) - q(B)| = \frac{1}{2}\sum_{\omega\in\Omega} |p(\omega) - q(\omega)| \ , \tag{32}$$

  which is directly related to the $L1$ norm

$$\delta_{TV}(p,q) = \frac{1}{2}\sum_{\omega\in\Omega} |p(\omega) - q(\omega)| = \frac{1}{2}\|p-q\|_1 \ . \tag{33}$$

  More broadly, if $p$ and $q$ are both absolutely continuous with respect to some base measure $\mu$,

$$\delta_{TV}(p,q) = \sup_{B\in\mathcal{F}} |p(B) - q(B)| = \frac{1}{2}\int_\Omega \left|\frac{\mathrm{d}p}{\mathrm{d}\mu} - \frac{\mathrm{d}q}{\mathrm{d}\mu}\right| \mathrm{d}\mu \ , \tag{34}$$

  where $\frac{\mathrm{d}p}{\mathrm{d}\mu}$ and $\frac{\mathrm{d}q}{\mathrm{d}\mu}$ are the Radon-Nikodyn derivatives of $p$ and $q$ with respect to $\mu$.

We now re-state and proof Lemma 3.1:

**Lemma 3.1**: The difference in action probabilities between two Thompson sampling policies, given the same history and context up to time $t$, is bounded by the total-variation distance $\delta_{TV}(p_t, q_t)$ between the posterior distributions of their expected rewards at time $t$, $p_t = p(\mu_t|h_{1:t})$ and $q_t = q(\mu_t|h_{1:t})$, respectively:

$$\pi_{p_t}\left(A_t = a\right) - \pi_{q_t}\left(A_t = a\right) \leq \delta_{TV}(p_t, q_t) \ . \tag{35}$$

The proof of Lemma 3.1 consists on showing that the difference between the expected values of a function of a random variable is bounded by the total variation distance between the corresponding distributions.

*Proof.* Let us define a linear function $l : \Omega \to [-1/2, 1/2]$ of a bounded function $g(\omega)$:

$$l(\omega) = \frac{g(\omega) - \inf_{\omega\in\Omega} g(\omega)}{\sup_{\omega\in\Omega} g(\omega) - \inf_{\omega\in\Omega} g(\omega)} - \frac{1}{2} \ . \tag{36}$$

Then,

$$\begin{aligned}\delta_{TV}(p,q) &= \frac{1}{2}\int_{\Omega}\left|\frac{\mathrm{d}p}{\mathrm{d}\mu}-\frac{\mathrm{d}q}{\mathrm{d}\mu}\right|\mathrm{d}\mu \geq \frac{1}{2}\int_{\Omega}\left|2l\left(\frac{\mathrm{d}p}{\mathrm{d}\mu}-\frac{\mathrm{d}q}{\mathrm{d}\mu}\right)\right|\mathrm{d}\mu \\ &\geq \int_{\Omega} l\left(\frac{\mathrm{d}p}{\mathrm{d}\mu}-\frac{\mathrm{d}q}{\mathrm{d}\mu}\right)\mathrm{d}\mu \geq \int_{\Omega} l\cdot \mathrm{d}p - \int_{\Omega} l\cdot \mathrm{d}q \\ &\geq \mathbb{E}_p\left\{l(\omega)\right\} - \mathbb{E}_q\left\{l(\omega)\right\} = \frac{\mathbb{E}_p\left\{g(\omega)\right\}-\mathbb{E}_q\left\{g(\omega)\right\}}{\sup_{\omega\in\Omega} g(\omega) - \inf_{\omega\in\Omega} g(\omega)}\ . \end{aligned} \tag{37}$$

We now recall that we can write the difference between two Thompson sampling policies as

$$\begin{aligned}\pi_{p_t}(A) - \pi_{q_t}(A) = \mathbb{E}_{p_t}&\left\{\mathbb{1}\left[A = \underset{a'\in\mathcal{A}}{\operatorname{argmax}}\left(\mu_{t,a'}\middle|h_{1:t}\right)\right]\right\} \\ &- \mathbb{E}_{q_t}\left\{\mathbb{1}\left[A = \underset{a'\in\mathcal{A}}{\operatorname{argmax}}\left(\mu_{t,a'}\middle|h_{1:t}\right)\right]\right\}\ . \end{aligned} \tag{38}$$

Let us define $g(\mu_t) = \mathbb{1}\left[A = \operatorname{argmax}_{a'\in\mathcal{A}}\left(\mu_{t,a'}\middle|h_{1:t}\right)\right]$, which is bounded in $[0,1]$:

$$\begin{cases}\inf_{\mu_t} g(\mu_t) = 0\ , \\ \sup_{\mu_t} g(\mu_t) = 1\ . \end{cases} \tag{39}$$

Direct substitution in Equation (37) results in

$$\begin{aligned}\delta_{TV}(p_t,q_t) &\geq \mathbb{E}_{p_t}\left\{\mathbb{1}\left[A = \underset{a'\in\mathcal{A}}{\operatorname{argmax}}\left(\mu_{t,a'}\middle|h_{1:t}\right)\right]\right\} \\ &\quad - \mathbb{E}_{q_t}\left\{\mathbb{1}\left[A = \underset{a'\in\mathcal{A}}{\operatorname{argmax}}\left(\mu_{t,a'}\middle|h_{1:t}\right)\right]\right\} \\ &\geq \pi_{p_t}(A) - \pi_{q_t}(A)\ , \end{aligned} \tag{40}$$

which concludes the proof. □

We make use of Lemma 3.1 to bound the asymptotic expected cumulative regret of the proposed Thompson sampling with a Dirichlet process (i.e., $d_a = 0, \forall a$) Gaussian mixture prior.

To that end, let us define the following Thompson sampling policies:

- The optimal Thompson sampling policy, $\pi_{p^*}(\cdot)$, which chooses the optimal arm given the true reward model $p^* = p(Y|\theta^*)$,

$$\begin{aligned}\pi_{p^*}(A_t^*|h_{1:t}) &= \mathbb{P}_{p^*}\left(A_t^* = \underset{a'\in\mathcal{A}}{\operatorname{argmax}}\left(\mu_{t,a'}\middle|h_{1:t}\right)\right) \\ &= \mathbb{E}_{p^*}\left\{\mathbb{1}\left[A_t^* = \underset{a'\in\mathcal{A}}{\operatorname{argmax}}\left(\mu_{t,a'}\middle|h_{1:t}\right)\right]\right\}\ . \end{aligned} \tag{41}$$

- A parametric Thompson sampling policy, $\pi_p(\cdot)$, which knows the true reward distribution model class $p = p(Y|\theta)$, but not the true parameter $\theta^*$,

$$\begin{aligned}\pi_p(A_t|h_{1:t}) &= \mathbb{P}_p\left(A_t = \underset{a'\in\mathcal{A}}{\operatorname{argmax}}\left(\mu_{t,a'}\middle|h_{1:t}\right)\right) \\ &= \mathbb{E}_p\left\{\mathbb{1}\left[A_t = \underset{a'\in\mathcal{A}}{\operatorname{argmax}}\left(\mu_{t,a'}\middle|h_{1:t}\right)\right]\right\}\ . \end{aligned} \tag{42}$$

The actions of this Thompson sampling policy, denoted as $A_t \sim \pi_p\left(A_t|h_{1:t}\right)$, are stochastic due to the uncertainty on the parameter $\theta$ of the true density $p(Y|\theta)$.

- A nonparametric Thompson sampling policy, $\pi_{\tilde{p}}\left(\cdot\right)$, which estimates the unknown true reward distribution with a nonparametric density $\tilde{p} = \tilde{p}(Y|\varphi)$,

$$\begin{aligned}\pi_{\tilde{p}}\left(\tilde{A}_t|h_{1:t}\right) &= \mathbb{P}_{\tilde{p}}\left(\tilde{A}_t = \underset{a' \in \mathcal{A}}{\operatorname{argmax}}\left(\mu_{t,a'}\middle|h_{1:t}\right)\right) \\ &= \mathbb{E}_{\tilde{p}}\left\{\mathbb{1}\left[\tilde{A}_t = \underset{a' \in \mathcal{A}}{\operatorname{argmax}}\left(\mu_{t,a'}\middle|h_{1:t}\right)\right]\right\} \ . \end{aligned} \tag{43}$$

The actions of this Thompson sampling policy, denoted as $\tilde{A}_t \sim \pi_{\tilde{p}}\left(\tilde{A}_t|h_{1:t}\right)$, are stochastic due to the uncertainty on the parameter $\varphi$ of the nonparametric density $\tilde{p}(Y|\theta)$.

**Theorem 3.2**: The expected cumulative regret at time $T$ of a Dirichlet process Gaussian mixture model based Thompson sampling algorithm is asymptotically bounded by

$$R_T = \mathbb{E}\left\{\sum_{t=1}^{T} Y_{t,A_t^*} - Y_{t,\tilde{A}_t}\right\} \leq \mathcal{O}\left(|\mathcal{A}| \log^{\kappa} T\sqrt{T}\right) \ \text{ as } T \to \infty \ , \tag{44}$$

where the expectations are taken over the random rewards $Y_t \sim p^* = p(Y|x_t, \theta^*)$ and the random actions of the stochastic policies $\pi_{p^*}\left(A_t^*\right)$ and $\pi_{\tilde{p}}\left(\tilde{A}_t\right)$.

This expected regret bound holds in the frequentist sense.

We use big-O notation $\mathcal{O}(\cdot)$ as it bounds from above the growth of the cumulative regret over time for large enough input sizes, i.e.,

$$\lim_{T \to \infty} \frac{R_T}{|\mathcal{A}| \log^{\kappa} T\sqrt{T}} \leq \mathcal{O}(1) \ . \tag{45}$$

In the following, we avoid notation clutter and denote $p^* = p^*(Y) = p(Y|\theta^*)$ for the true reward distribution given the true parameters $\theta^*$, and drop the dependency over the observed history $h_{1:t}$ at time $t$ in the considered Thompson sampling policies: $\pi_{p^*} = \pi_{p^*}\left(A_t^*|h_{1:t}\right)$, for the optimal Thompson sampling policy with knowledge of the true reward model $p^* = p(Y|\theta^*)$; $\pi_p = \pi_p\left(A_t|h_{1:t}\right)$, for a Thompson sampling policy with knowledge of the true reward distribution model class $p = p(Y|\theta)$ —but not the true parameter $\theta^*$; and $\pi_{\tilde{p}} = \pi_{\tilde{p}}\left(\tilde{A}_t|h_{1:t}\right)$, for a nonparametric Thompson sampling policy that estimates the unknown true reward distribution with a nonparametric density $\tilde{p} = \tilde{p}(Y|\varphi)$.

*Proof.*

$$R_T = \mathbb{E}\left\{\sum_{t=1}^{T} Y_{t,A_t^*} - Y_{t,\tilde{A}_t}\right\} \tag{46}$$

$$= \mathbb{E}_{\pi_{p^*},\pi_{\tilde{p}}}\left\{\mathbb{E}_{p^*}\left\{\sum_{t=1}^{T} Y_{t,A_t^*} - Y_{t,\tilde{A}_t}\right\}\right\} \tag{47}$$

$$= \sum_{t=1}^{T} \mathbb{E}_{\pi_{p^*},\pi_{\tilde{p}}}\left\{\mathbb{E}_{p^*}\left\{Y_{t,A_t^*} - Y_{t,\tilde{A}_t}\right\}\right\} \tag{48}$$

$$= \sum_{t=1}^{T} \mathbb{E}_{\pi_{p^*},\pi_{\tilde{p}},\pi_p}\left\{\mathbb{E}_{p^*}\left\{Y_{t,A_t^*} - Y_{t,A_t} + Y_{t,A_t} - Y_{t,\tilde{A}_t}\right\}\right\} \tag{49}$$

$$= \sum_{t=1}^{T} \mathbb{E}_{\pi_{p^*},\pi_p}\left\{\mathbb{E}_{p^*}\left\{Y_{t,A_t^*} - Y_{t,A_t}\right\}\right\} + \sum_{t=1}^{T} \mathbb{E}_{\pi_p,\pi_{\tilde{p}}}\left\{\mathbb{E}_{p^*}\left\{Y_{t,A_t} - Y_{t,\tilde{A}_t}\right\}\right\} \tag{50}$$

$$= \sum_{t=1}^{T} \mathbb{E}_{\pi_{p^*},\pi_p}\left\{\mu_{t,A_t^*} - \mu_{t,A_t}\right\} + \sum_{t=1}^{T} \mathbb{E}_{\pi_p,\pi_{\tilde{p}}}\left\{\mu_{t,A_t} - \mu_{t,\tilde{A}_t}\right\}, \tag{51}$$

where we have split the expected cumulative regret of Equation 46 in two terms.

The first term in the RHS of Equation 51 relates to the regret between the optimal policy $A_t^* \sim \pi_{p^*}\left(A_t^*|h_{1:t}\right)$ and a Thompson sampling policy that knows the true model class $A_t \sim \pi_p\left(A_t|h_{1:t}\right)$; and the second term in the RHS of Equation 51 accommodates the regret between a Thompson sampling policy that knows the true model class $A_t \sim \pi_p\left(A_t|h_{1:t}\right)$, and a Thompson sampling that estimates reward functions via nonparametric processes $\tilde{A}_t \sim \pi_{\tilde{p}}\left(\tilde{A}_t|h_{1:t}\right)$.

Let us first work on the first term in the RHS of Equation 51:

$$\sum_{t=1}^{T} \mathbb{E}_{\pi_{p^*},\pi_p} \left\{ \mu_{t,A_t^*} - \mu_{t,A_t} \right\} \tag{52}$$

$$= \sum_{t=1}^{T} \left[ \left( \sum_{a_t^* \in \mathcal{A}} \mu_{t,a_t^*} \pi_{p^*} \left( A_t^* = a_t^* | h_{1:t} \right) \right) \right. \tag{53}$$

$$\left. - \left( \sum_{a_t \in \mathcal{A}} \mu_{t,a_t} \pi_p \left( A_t = a_t | h_{1:t} \right) \right) \right] \tag{54}$$

$$= \sum_{t=1}^{T} \left( \sum_{a \in \mathcal{A}} \mu_{t,a} \left[ \pi_{p^*} \left( A_t^* = a | h_{1:t} \right) - \pi_p \left( A_t = a | h_{1:t} \right) \right] \right) \tag{55}$$

$$\leq \sum_{t=1}^{T} \left( \sum_{a \in \mathcal{A}} C_A \left[ \pi_{p^*} \left( A_t^* = a | h_{1:t} \right) - \pi_p \left( A_t = a | h_{1:t} \right) \right] \right) \tag{56}$$

$$\leq \sum_{t=1}^{T} \left( \sum_{a \in \mathcal{A}} C_A \delta_{TV} \left( p^*(\mu_t | h_{1:t}), p(\mu_t | h_{1:t}) \right) \right) \tag{57}$$

$$\leq \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} C_A C_p t^{-1/2} \tag{58}$$

$$\leq C_A C_p \sum_{a \in \mathcal{A}} \left( \sum_{t=1}^{T} t^{-1/2} \right) \tag{59}$$

$$\leq C_A C_p \sum_{a \in \mathcal{A}} \left( \int_{t=1}^{T} t^{-1/2} \mathrm{d}t \right) \tag{60}$$

$$\leq C_A C_p \sum_{a \in \mathcal{A}} (2\sqrt{T} - 2) \tag{61}$$

$$\leq 2 C_A C_p |\mathcal{A}| \sqrt{T} \,, \tag{62}$$

where

- in Equation 56: we define $C_A \coloneqq \max_{a \in \mathcal{A}} \mu_{a,t}, \forall t$, i.e., it is an upper bound on the expected rewards of the bandit.
- in Equation 57: by direct application of Equation 35 in Lemma 3.1: $\pi_{p^*} \left( A_t^* = a | h_{1:t} \right) - \pi_p \left( A_t = a | h_{1:t} \right) \leq \delta_{TV} \left( p^*(\mu_t | h_{1:t}), p(\mu_t | h_{1:t}) \right)$.

  That is, the difference in probabilities of playing each arm $a$ are bounded by the total variation distance between the posterior distributions of the expected rewards for each policy.

  For the optimal Thompson sampling policy, the parameters of the reward distribution

are known, i.e., the posterior is a delta at the true $\theta^*$ value:

$$\begin{aligned} p^*(\mu_t|h_{1:t}) &= \int_{\theta^*} p^*(\mu_t|\theta^*)p(\theta^*|h_{1:t})\mathrm{d}\theta^* \\ &= \int_{\theta^*} p^*(\mu_t|\theta^*)\delta(\theta^*)\mathrm{d}\theta^* \\ &= p^*(\mu_t|\theta^*) \ . \end{aligned}$$

For the Thompson sampling policy that knows the true model class, the parameters of the reward distribution are updated as history $h_{1:t}$ is observed:

$$p(\mu_t|h_{1:t}) = \int_{\theta} p(\mu_t|\theta)p(\theta|h_{1:t})\mathrm{d}\theta \ .$$

- in Equation 58: $\delta_{TV}\left(p^*(\mu_t|h_{1:t}), p(\mu_t|h_{1:t})\right) \sim C_p t^{-1/2}$, as $t \to \infty$, where $C_p$ is a constant that depends on the properties of the parameterized distributions, and does not depend on the amount of observed data.
  As explained in [Ghosal et al., 2000], for a class of parameterized distributions $\mathcal{P} = \{p(Y|\theta)\}_{\theta\in\Theta}$ and a prior constructed by putting a measure on the parameter set $\Theta$, it is well known that the posterior distribution of $\theta$ asymptotically achieves the optimal rate of convergence under mild regularity conditions —i.e., $\Theta$ is subset of a finite-dimensional Euclidean space, and the prior and model dependence is sufficiently regular [Ibragimov and Has'minskii, 1981]. In particular, and according to the Bernstein-von Mises theorem, if the model $p(Y|\theta)$ is suitably differentiable, then the convergence rate of the posterior mean $p(\mu_t|h_{1:t})$ and $p^*(\mu_t)$ is of order $t^{-1/2}$, where $t$ indicates the amount of i.i.d. data drawn from the true distribution $p^*(Y)$.

  Note that the true $p^*(\mu_t)$ and the posterior $p(\mu_t|h_{1:t})$ are over the expected rewards of all arms. Therefore, $t = \sum_{a\in\mathcal{A}} t_a$, where $t_a$ indicates the number of observations for each arm, is the number of times all arms $\forall a \in \mathcal{A}$ have been pulled. Consequently, the total variation Equation 58 depends on the total number of observations $t$ across all arms $a$.

- in Equation 60: $\sum_{t=1}^{T} t^{-1/2} = \mathbb{H}^{1/2}(T) \leq \int_{t=1}^{T} t^{-1/2}\mathrm{d}t$, where $\mathbb{H}$ is the generalized harmonic number of order $1/2$ of $T$.

This concludes the proof of the bound of the first term in the RHS of Equation 51.

We now bound the second term in the RHS:

$$\sum_{t=1}^{T} \mathbb{E}_{\pi_p, \pi_{\tilde{p}}} \left\{ \mu_{t,A_t} - \mu_{t,\tilde{A}_t} \right\} \tag{63}$$

$$= \sum_{t=1}^{T} \left[ \left( \sum_{a_t \in \mathcal{A}} \mu_{t,a_t} \pi_p \left( A_t = a_t | h_{1:t} \right) \right) \right. \tag{64}$$

$$\left. - \left( \sum_{\tilde{a}_t \in \mathcal{A}} \mu_{t,\tilde{a}_t} \pi_{\tilde{p}} \left( \tilde{A}_t = \tilde{a}_t | h_{1:t} \right) \right) \right] \tag{65}$$

$$= \sum_{t=1}^{T} \left( \sum_{a \in \mathcal{A}} \mu_{t,a} \left[ \pi_p \left( A_t = a | h_{1:t} \right) - \pi_{\tilde{p}} \left( \tilde{A}_t = a | h_{1:t} \right) \right] \right) \tag{66}$$

$$\leq \sum_{t=1}^{T} \left( \sum_{a \in \mathcal{A}} C_A \left[ \pi_p \left( A_t = a | h_{1:t} \right) - \pi_{\tilde{p}} \left( \tilde{A}_t = a | h_{1:t} \right) \right] \right) \tag{67}$$

$$\leq \sum_{t=1}^{T} \left( \sum_{a \in \mathcal{A}} C_A \delta_{TV} \left( p(\mu_t | h_{1:t}), \tilde{p}(\mu_t | h_{1:t}) \right) \right) \tag{68}$$

$$\leq \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} C_A C_{\tilde{p}} t^{-1/2} (\log t)^{\kappa} \tag{69}$$

$$\leq C_A C_{\tilde{p}} \sum_{a \in \mathcal{A}} \left( \sum_{t=1}^{T} t^{-1/2} (\log T)^{\kappa} \right) \tag{70}$$

$$\leq C_A C_{\tilde{p}} \sum_{a \in \mathcal{A}} (\log T)^{\kappa} \left( \int_{t=1}^{T} t^{-1/2} \mathrm{d}t \right) \tag{71}$$

$$\leq C_A C_{\tilde{p}} \sum_{a \in \mathcal{A}} (\log T)^{\kappa} (2\sqrt{T} - 2) \tag{72}$$

$$\leq 2 C_A C_{\tilde{p}} |\mathcal{A}| (\log T)^{\kappa} \sqrt{T} \,, \tag{73}$$

where

- in Equation 67: $C_A \coloneqq \max_{a \in \mathcal{A}} \mu_{a,t}, \forall t$, as above.
- in Equation 68: by direct application of Equation 35 in Lemma 3.1: $\pi_p \left( A_t = a | h_{1:t} \right) - \pi_{\tilde{p}} \left( \tilde{A}_t = a | h_{1:t} \right) \leq \delta_{TV} \left( p(\mu_t | h_{1:t}), \tilde{p}(\mu_t | h_{1:t}) \right)$.

  That is, the difference in probabilities of playing each arm $a$ are bounded by the total variation distance between the posterior distributions of the expected rewards for each policy.

  For the Thompson sampling policy that knows the true model class, the parameters of the reward distribution are updated as history $h_{1:t}$ is observed:

$$p(\mu_t | h_{1:t}) = \int_{\theta} p(\mu_t | \theta) p(\theta | h_{1:t}) \mathrm{d}\theta \,.$$

For the Thompson sampling that estimates reward functions via nonparametric model $\tilde{p}(Y_t|\varphi)$, the parameters $\varphi$ of the nonparametric reward distribution are updated as history $h_{1:t}$ is observed:

$$\tilde{p}(\mu_t|h_{1:t}) = \int_{\varphi} \tilde{p}(\mu_t|\varphi)\tilde{p}(\varphi|h_{1:t})\mathrm{d}\varphi \ .$$

- in Equation 69: $\delta_{TV}\left(p(\mu_t|h_{1:t}), \tilde{p}(\mu_t|h_{1:t})\right) \sim C_{\tilde{p}} t^{-1/2}(\log t)^{\kappa}$, as $t \to \infty$; where $C_{\tilde{p}}$ is a constant that depends on the properties of both the true parametric posterior distribution and the nonparametric prior model, but does not depend on the amount of observed data. We asymptotically bound the total variation distance between the true parametric posterior distribution and a nonparametric model-based posterior distribution, leveraging state-of-the-art results.

  Note that the posterior $p(\mu_t|h_{1:t})$ is over the expected rewards over all arms. Therefore, Equation 69 depends on the total number of observations across all arms $t = \sum_{a\in\mathcal{A}} t_a$, where $t_a$ indicates the number of observations observed for each arm $\forall a \in \mathcal{A}$.

  The behavior of posterior distributions for infinite dimensional models has been thoroughly studied at the beginning of this century, with work by Ghosal and van der Vaart [2001, 2007] providing posterior convergence rates of Dirichlet process Gaussian mixtures to different mixture distributions.

  For example, for a mixture of normals with standard deviations bounded by two positive numbers, Ghosal and van der Vaart [2001] show that the Hellinger distance between the nonparametric posterior given $n$ data samples and the true distribution is asymptotically bounded,

$$d(\tilde{p}, p^*) \leq M n^{-1/2}(\log n)^{\kappa} \ , \tag{74}$$

  where the value $\kappa \geq 0$ depends on the choices of priors over the location and scale of the mixtures, and data is drawn from the true distribution $p^*$. Since $\|p - q\|_1 \leq 2d(p, q)$, bounds in Hellinger distance apply to total variation distance as well. Note that the convergence of the posterior at such rate also implies that there exist estimators, such as the posterior mean, that converge at the same rate in the frequentist sense.

  Technical details of the bound in Equation (74) can be found in [Ghosal and van der Vaart, 2001], where they consider Gaussian location mixtures and location-scale mixtures, assumed the standard deviations to be bounded away from zero and infinity, and that the true mixing distribution of the location is compactly supported or has sub-Gaussian tails.

  A rate with $\kappa = 1$ is obtained when a compactly supported base measure is used for the location prior (and the scale prior has a continuous and positive density on an interval containing the true scale parameter). For the commonly used normal base measure, the bound yields a rate $O(n^{-1/2}(\log n)^{3/2})$. When the base measure is the product of a normal distribution with a distribution supported within the range of the true scale, such that the density is positive on a rectangle containing the true location-scale space, the rate results in $O(n^{-1/2}(\log n)^{7/2})$.

  Later work by Ghosal and van der Vaart [2007] provides new posterior convergence rates for densities that are twice continuously differentiable, where under some regularity conditions, the posterior distribution based on a Dirichlet mixture of normal prior

attains a convergence rate of $O(n^{-2/5}(\log n)^{4/5})$. As such, it seems reasonable that the power of the logarithm, i.e., $\kappa$ in Equation (74), can be improved. Ghosal and van der Vaart [2007] argue that, by using a truncated inverse gamma prior on the scale of the Gaussian mixtures, a nearly optimal convergence rate is obtained —for which one would need to extend the Gibbs sampler with an additional accept-reject step to take care of the scale truncation.

All these bounds would not be directly applicable if the true data generating density would not be part of the model classes considered. However, Ghosal and van der Vaart [2001] argue that a rate for these cases may as well be calculated, but since they may not be close to the optimal rate, have not been pursued yet.

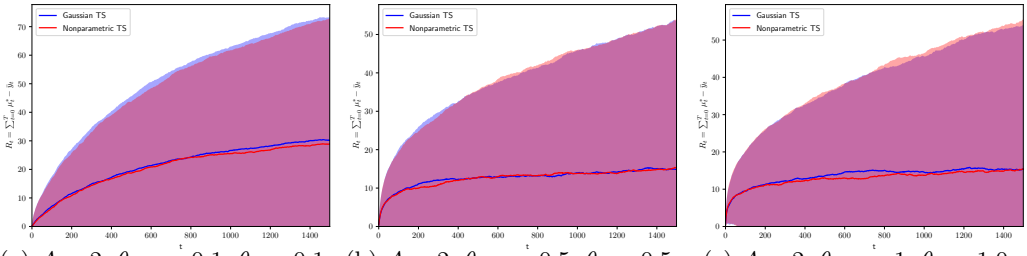
- in Equation 70: $(\log t)^\kappa \leq (\log T)^\kappa, \forall 1 \leq t \leq T, \kappa \geq 0$.
- in Equation 71: $\sum_{t=1}^{T} t^{-1/2} = \mathbb{H}^{1/2}(T) \leq \int_{t=1}^{T} t^{-1/2} \mathrm{d}t$, where $\mathbb{H}$ is the generalized harmonic number of order $1/2$ of $T$.

Combining the above results, we can now bound the asymptotic cumulative regret in Equation 46, for a nonparametric Thompson sampling policy with Dirichlet process Gaussian mixtures, with priors and data-generating densities that meet the necessary regularity conditions:

$$
\begin{aligned}
R_T = & \sum_{t=1}^{T} \mathbb{E}_{\pi_{p^*}, \pi_p} \left\{ \mathbb{E}_{p^*} \left\{ Y_{t,A_t^*} - Y_{t,A_t} \right\} \right\} \\
& + \sum_{t=1}^{T} \mathbb{E}_{\pi_p, \pi_{\tilde{p}}} \left\{ \mathbb{E}_{p^*} \left\{ Y_{t,A_t} - Y_{t,\tilde{A}_t} \right\} \right\} && (75) \\
& \leq \mathcal{O}\left( 2C_A C_p |\mathcal{A}| \sqrt{T} + 2C_A C_{\tilde{p}} |\mathcal{A}| (\log T)^\kappa \sqrt{T} \right) && (76) \\
& \leq \mathcal{O}\left( 2C_A |\mathcal{A}| \sqrt{T} (C_p + C_{\tilde{p}} (\log T)^\kappa) \right) && (77) \\
& \leq \mathcal{O} |\mathcal{A}| \sqrt{T} (\log T)^\kappa \ . && (78)
\end{aligned}
$$

We note that this bound holds both in a frequentist and Bayesian view of expected cumulative regret. □

# C Non-contextual Gaussian bandits: comparison to the Oracle TS




(a) $A = 2$, $\theta_1 = -0.1$, $\theta_2 = 0.1$. (b) $A = 2$, $\theta_1 = -0.5$, $\theta_2 = 0.5$. (c) $A = 2$, $\theta_1 = -1$, $\theta_2 = 1.0$.

Figure 10: Mean regret (standard deviation shown as shaded region) for 1000 independent realizations of different two-armed Gaussian bandits, with $\sigma_a^2 = 1 \forall a$.

We show in Figure 10 how our proposed nonparametric Thompson sampling method achieves regret comparable to that of the non-contextual Gaussian Thompson sampling as in Agrawal and Goyal [2012] for diverse parameterizations of such Gaussian bandits.

Recall that the non-contextual bandit scenario is seamlessly accommodated by our proposed `Nonparametric TS` algorithm by assuming a constant context, i.e., $x_t = \mathbb{1}$.

# D Thompson sampling baseline hyperparameters

We here collect the specific hyperparameters used for the results presented in this work. These were selected based on the default suggested values in https://github.com/tensorflow/models/tree/master/research/deep contextual_bandits.

We first describe in Table 4 the neural network hyperparameters, shared across all the studied alternatives but the `linearGaussian TS` and the `MultitaskGP` baselines.

The specific details for each neural network based baseline are summarized in Table 5, with details for the Gaussian process based baseline in Table 6.

Table 4: Shared neural network hyperparameters.

| Hyperparameter | Value |
|---|---|
| training freq | 1 |
| training epochs | 100 |
| activation | tf.nn.relu |
| layer size | 50 |
| batch size | 512 |
| init scale | 0.3 |
| optimizer | 'RMS' |
| initial pulls | 2 |
| activate decay | True |
| max grad norm | 5.0 |
| initial lr | 0.1 |
| reset lr | True |
| lr decay rate | 0.5 |
| show training | False |
| freq summary | 1000 |

Table 5: Shared neural network hyperparameters.

| Algorithm | Baseline details |
|---|---|
| `NeuralLinear` | The network and the posterior parameter of the last layer are updated at every bandit iteration; Prior over linear parameters is $a_0 = 6$, $b_0 = 6$, $\lambda_0 = 0.25$ |
| `NeuralRMS` | Neural network learned with RMS optimizer with default parameters |
| `NeuralBootstrapped` | $q = 3$ networks and datasets for bootstrapping, with $p = 0.95$ |
| `NeuralParamNoise` | The i.i.d. noise added to parameters follow $\mathcal{N}(0, \sigma = 0.05)$, and an $\epsilon = 0.1$ greedy is implemented with 300 samples |
| `NeuralDropoutRMS` | Dropout with parameter 0.8 is used for training neural networks with RMS optimizer |
| `BNNVariationalGaussian` | Variational inference over Gaussian independent weight noises with sigma exponential transform and noise $\sigma = 0.1$; 100 initial training steps and 10 cleared times used in training. |
| `BNNAlphaDiv` | The Black-Box method is used with $\alpha = 1$, noise $\sigma = 0.1$ and $k = 20$, with sigma exponential transform and prior variance $\sigma^2 = 0.1$; 100 initial training steps and 20 cleared times used in training. |

Table 6: Gaussian Process hyperparameters.

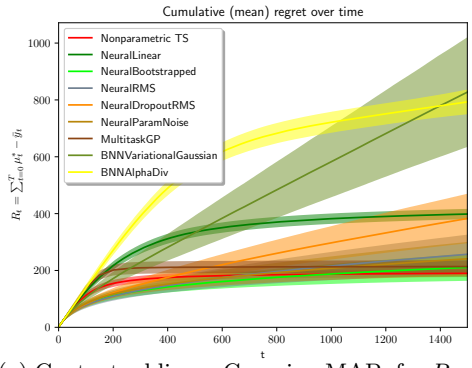
| Hyperparameter | Value |
|---|---|
| training freq | 50 |
| training epochs | 100 |
| learn embeddings | True |
| task latent dim | 5 |
| max num points | 1000 |
| batch size | 512 |
| optimizer | 'RMS' |
| initial pulls | 2 |
| lr | 0.01 |
| initial lr | 0.001 |
| lr decay rate | 0.0 |
| reset lr | False |
| activate decay | False |
| keep fixed after max obs | True |
| show training | False |
| freq summary | 1000 |

# E Contextual linear Gaussian bandits: baselines

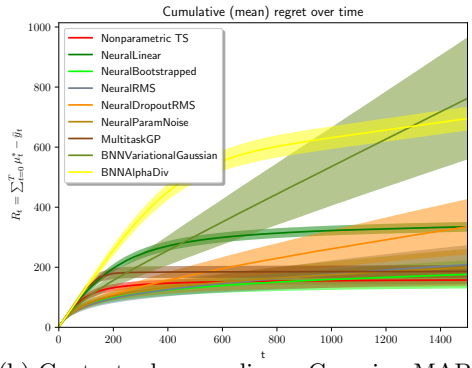
We show in Figure 11 the mean cumulative regret (and its standard deviation as the shaded region) of all the studied multi-armed bandit algorithms for diverse contextual linear Gaussian bandit parameterizations —per-algorithm cumulative reward results are shown in Table 7.

Table 7: Cumulative reward (mean and variance) at $t = 1500$ for $R = 500$ realizations of contextual linear Gaussian MABs.

| Algorithm | Final cumulative reward (linear) | Final cumulative reward (sparse linear) |
|---|---|---|
| Optimal | 2064.985 ± 4502.593 | 1960.926 ± 6363.005 |
| Nonparametric TS | 1875.096 ± 66.837 | 1803.519 ± 79.203 |
| NeuralLinear | 1666.499 ± 64.629 | 1626.356 ± 78.577 |
| NeuralBootstrapped | 1854.071 ± 77.620 | 1784.790 ± 87.437 |
| NeuralRMS | 1808.088 ± 93.845 | 1752.188 ± 97.987 |
| NeuralDropoutRMS | 1682.115 ± 100.281 | 1626.056 ± 113.981 |
| NeuralParamNoise | 1823.946 ± 83.025 | 1761.804 ± 96.642 |
| MultitaskGP | 1851.337 ± 64.938 | 1774.902 ± 75.291 |
| BNNVariationalGaussian | 1237.616 ± 203.954 | 1198.605 ± 215.776 |
| BNNAlphaDiv | 1271.867 ± 66.406 | 1265.829 ± 71.714 |




(a) Contextual linear Gaussian MAB, for $R = 500$ realizations.




(b) Contextual sparse linear Gaussian MAB, for $R = 500$ realizations.

Figure 11: Mean regret (standard deviation shown as shaded region) for 1000 independent realizations of the presented methods in all scenarios.

# F Contextual bandits not in the exponential family

We show in Table 8 the mean (and standard deviation) cumulative regret of all the studied multi-armed bandit algorithms in the proposed complex scenarios described in Section 4.2.2.

Table 8: Final (at $t = 1000$) cumulative reward (mean and standard deviation) for $R = 500$ realizations of the studied methods in all scenarios.

| Algorithm | Scenario A | Scenario B | Scenario C | Scenario D |
|---|---|---|---|---|
| Optimal | 2603.637 ± 43.871 | 2703.374 ± 48.053 | 2920.616 ± 47.294 | 3511.531 ± 137.860 |
| Nonparametric TS | 2481.226 ± 46.969 | 2064.933 ± 70.063 | 2155.669 ± 86.414 | 1916.523 ± 176.516 |
| LinearGaussian TS | 2477.374 ± 46.967 | 2043.161 ± 87.453 | 2124.920 ± 98.306 | 1846.032 ± 289.280 |
| NeuralLinear | 2474.973 ± 47.411 | 2023.889 ± 95.728 | 2102.640 ± 111.766 | 1786.915 ± 316.087 |
| NeuralBootstrapped | 2477.389 ± 134.222 | 1959.900 ± 212.630 | 2008.643 ± 296.424 | 1846.672 ± 478.053 |
| NeuralRMS | 2478.773 ± 134.110 | 1953.925 ± 218.489 | 2001.024 ± 299.355 | 1808.436 ± 532.736 |
| NeuralDropoutRMS | 2473.140 ± 165.730 | 1954.019 ± 215.282 | 2000.544 ± 302.480 | 1829.687 ± 495.924 |
| NeuralParamNoise | 2476.161 ± 134.467 | 1970.443 ± 196.946 | 2032.456 ± 262.984 | 1822.056 ± 513.216 |
| MultitaskGP | 2384.541 ± 56.206 | 1954.318 ± 71.298 | 1957.400 ± 87.940 | 1332.580 ± 227.129 |
| BNNVariationalGaussian | 2471.688 ± 158.345 | 1927.741 ± 176.984 | 1984.950 ± 250.108 | 1659.067 ± 388.108 |
| BNNAlphaDiv | 2431.362 ± 52.776 | 1904.173 ± 66.993 | 1896.702 ± 83.000 | 1751.989 ± 206.198 |

# G Computational complexity and run-times of the evaluated Thompson sampling algorithms

Computational cost is an important metric in real-life bandit scenarios, which motivated us to provide a computational analysis of our proposed algorithm in Section 3.2.2. As explained there, in general, the computational complexity is upper bounded by $\mathcal{O}(T \cdot Gibbs_{steps})$, which depends on the convergence criteria: i.e., either the model likelihood of the sampled chain is stable within an $\epsilon$ likelihood margin between steps, or a maximum number of iterations $Gibbs_{max}$ is reached. As such, tweaking these two values controls the resulting run-times.

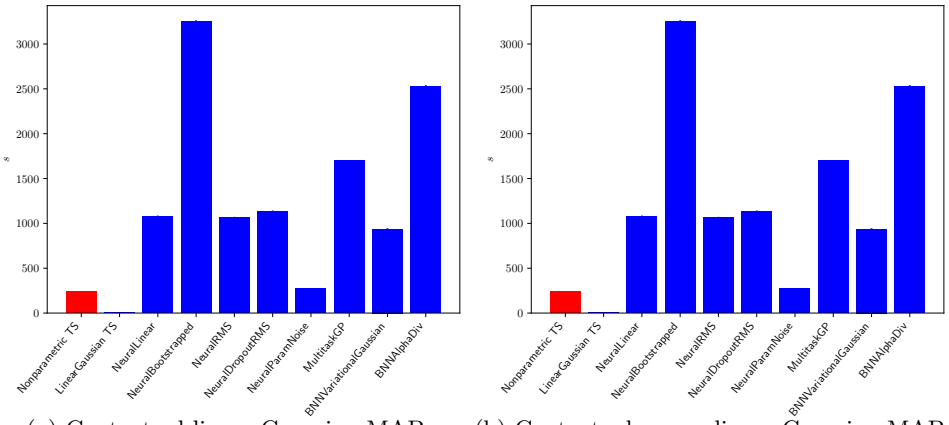
We provide below a comparison of the run-times incurred in the set of experiments described in the manuscript, for which we would like to raise two cautionary disclaimers:

- We compare **algorithms with different implementations**:
  The proposed `Nonparametric TS` is implemented with standard python libraries (i.e., numpy, scipy), while the rest of the algorithms are implemented in Tensorflow, as provided in the Deep Contextual bandit implementation by Riquelme et al. [2018]. Our goal here is to introduce a new bandit algorithm, and improving the efficiency of our implementation (or coding it in Tensorflow) is out of the scope of this work.

- Both our algorithm and those in the deep contextual bandits showdown require **updates at every time instant that depend on the number of observations per-arm** $t_a$:
  Performance and running-time differences can be achieved if one tweaks each algorithm's settings for model updates. As explained in Riquelme et al. [2018], a key question is how often and for how long models are updated, as these will determine their running-times in practice. Even if ideally, one would like to re-train models after each new observation (and for as long as possible), this may limit the applicability of the algorithms in online scenarios. In this work, we have executed all baselines based on the default hyperparameters suggested by Riquelme et al. [2018] —which limits the retraining process per interaction to 100 epochs, upper bounding the execution time per bandit interaction— and argue that tweaking the hyperparameters of such algorithms to reduce running-times is out of the scope of this work.

Nevertheless, and as illustrative examples, we show in Figure 13 the running times

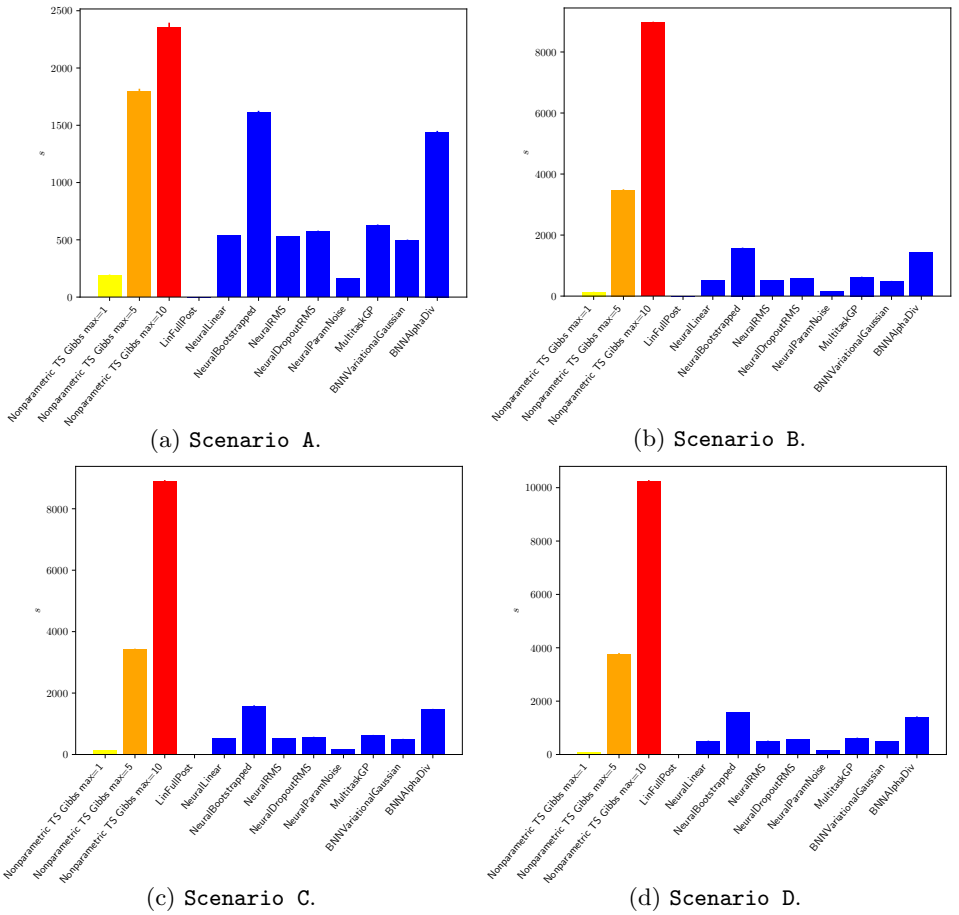
of all algorithms (averaged across realizations). First, we note that `LinearGaussian TS`, due to its conjugacy-based posterior updates that can be computed sequentially, is the fastest algorithm in all scenarios. Second, we observe that the algorithms in Riquelme et al. [2018] have a similar running-time across all scenarios, expected due to the suggested configuration that limits per-interaction run-time to 100 epochs. Third, the run-times of our `Nonparametric TS` algorithm vary across scenarios, as updating the nonparametric posterior model takes more or less time depending on the complexity of the true reward model: it shows low computational complexity in linear Gaussian scenarios, while incurring in higher computational cost when fitting the most challenging `Scenarios B, C`, and `D`. However, as demonstrated in Figures 13a, 13b, 13c and 13d, we can drastically reduce the run-time of `Nonparametric TS` by upper-bounding the number of Gibbs iterations, yet achieve satisfactory performance, as demonstrated in Figure 5 of the manuscript.




(a) Contextual linear Gaussian MAB. (b) Contextual sparse linear Gaussian MAB.

Figure 12: Mean run-time (standard deviation shown as error bars) in seconds for $R = 500$ realizations of the studied methods in linear contextual multi-armed bandits.

We conclude by reiterating that, in general, we recommend to run the algorithm until full convergence, but suggest to limit the number of Gibbs iterations as a practical recommendation with good empirical regret performance —analogous to the suggestion by Riquelme et al. [2018] to limit the number of per-iteration epochs for neural network based algorithms.




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

(c) `Scenario C`. (d) `Scenario D`.

Figure 13: Mean run-time (standard deviation shown as error bars) in seconds for $R = 500$ realizations of the studied methods in all scenarios.