Leveraging priors on distribution functions for multi-arm bandits

Anonymous authors
Paper under double-blind review

Keywords: RLJ, RLC, formatting guide, style file, LaTeX template.

Summary

We introduce Dirichlet Process Posterior Sampling (DPPS), a Bayesian non-parametric algorithm for multi-arm bandits based on Dirichlet Process (DP) priors. Like Thompson-sampling, DPPS is a probability-matching algorithm, i.e., it plays an arm based on its posterior-probability of being optimal. Instead of assuming a parametric class for the reward generating distribution of each arm, and then putting a prior on the parameters, in DPPS the reward generating distribution is directly modeled using DP priors. DPPS provides a principled approach to incorporate prior belief about the bandit environment, and in the noninformative limit of the DP priors (i.e. Bayesian Bootstrap), we recover Non Parametric Thompson Sampling (NPTS), a popular non-parametric bandit algorithm, as a special case of DPPS. We employ stick-breaking representation of the DP priors, and show excellent empirical performance of DPPS in challenging synthetic and real world bandit environments. Finally, using an information-theoretic analysis, we show non-asymptotic optimality of DPPS in the Bayesian regret setup.

Contribution(s)

1. We introduce Dirichlet Process Posterior Sampling (DPPS) for multi arm bandits - a Bayesian nonparametric extension of Thompson sampling based on Dirichlet Processes that combines the strength of (Bayesian) bootstrap with a principled mechanism of *incorporating and exploiting prior information*.
   Context: Efficient performance of *parametric* Thompson sampling is limited to bandit environments wherein it's possible to have conjugate prior/posterior distributions. Besides, existing Bootstrap based algorithms cannot account for uncertainity that doesn't come from observed data (32)
2. We employ stick-breaking representation of the Dirichlet Process priors to perform numerical experiments with DPPS in both synthetic and real-world multi-arm bandit settings.
   Context: Improved performance of DPPS compared to parametric Thompson-sampling and UCB is made apparent in these simulations. Using a simple example, we also illustrate a proof-of-concept of the flexibility of DPPS in incorporating prior-knowledge about the bandit environment. Besides, Stick-Breaking implementation of DPPS provides a unified implementation for different bandit environments unlike parametric Thompson sampling whose implementation differ according to bandit environments and require careful tuning/approximations.
3. We extend the information theoretic analysis of Thompson sampling in (43) to a wider class of probability-matching algorithms that derive their posterior probability of optimal action using a valid Bayesian approach, and use this extension to establish $\sigma\sqrt{2\mathrm{TK}\log\mathrm{K}}$ non-asymptotic upper bound on the Bayesian regret of DPPS in bandit environments with σ sub-Gaussian reward noise, where T is the time horizon, and K is the number of arms.
   Context: We are unaware of any Bootstrap based bandit algorithm that enjoys the order-optimal, $\sigma\sqrt{2\mathrm{TK}\log\mathrm{K}}$, non-asymptotic regret bound in the wide class of σ-sub-Gaussian bandit environments.

Leveraging priors on distribution functions for multi-arm bandits

Anonymous authors
Paper under double-blind review

Abstract

We introduce Dirichlet Process Posterior Sampling (DPPS), a Bayesian non-parametric algorithm for multi-arm bandits based on Dirichlet Process (DP) priors. Like Thompson-sampling, DPPS is a probability-matching algorithm, i.e., it plays an arm based on its posterior-probability of being optimal. Instead of assuming a parametric class for the reward generating distribution of each arm, and then putting a prior on the parameters, in DPPS the reward generating distribution is directly modeled using DP priors. DPPS provides a principled approach to incorporate prior belief about the bandit environment, and in the noninformative limit of the DP posteriors (i.e. Bayesian Bootstrap), we recover Non Parametric Thompson Sampling (NPTS), a popular non-parametric bandit algorithm, as a special case of DPPS. We employ stick-breaking representation of the DP priors, and show excellent empirical performance of DPPS in challenging synthetic and real world bandit environments. Finally, using an information-theoretic analysis, we show non-asymptotic optimality of DPPS in the Bayesian regret setup.

1 Introduction

Multi Arm Bandits (MAB) is a paradigmatic framework to study the exploration $\sim$ exploitation dilemma in sequential decision making under uncertainty. Standard algorithms developed within this framework such as Upper-Confidence Bounds (UCB) absed algorithms (4) and Thompson sampling (TS) (47; 44) have proven to be useful in applications such as clinical trials, ad-placement strategies, etc. However, it remains difficult to apply them to more complicated real world settings such as those arising in agriculture or experimental sciences wherein the underlying uncertainty mechanism is far more sophisticated: the unknown reward distribution corresponding to each arm/action may not even conform to a parametric class of distributions such as the single-parameter exponential family, and usually exhibit characteristics such as multi-modality. With some abuse of terminology, we shall refer to this challenging setting of the MABs as *non-parametric* MABs, and we report an optimal algorithm for this setting in the current paper.

To begin with, it's worthwhile to consider the limitations of UCB and Thompson sampling algorithms in some more detail. Firstly, the efficient performance of UCB type algorithms rely on the construction of tight high-probability confidence sequences (1; 4). However, for complex problems, it becomes difficult to design such sequences, and only approximate confidence sequences can be designed, which generally tend to be statistically suboptimal (20). Next, although Thompson-Sampling (TS) (47; 26) is a neat and elegant *Bayesian* algorithm, that enjoys the flexibility of incorporating *prior* knowledge about the bandit environment, it's efficiency is limited to the regime of *conjugate* prior/posterior distributions of the relevant scalar/vector parameter, which is generally not possible beyond a few special cases of bandit environments, e.g. Bernoulli, Gaussian. In other regimes, the posterior distributions no longer exhibit a closed form, and require the application of approximate inference schemes such as Markov Chain Monte Carlo (MCMC), variational inference, etc to draw samples from the posterior distributions. This is usually computationally expensive and can easily lead to the suboptimal performance of Thompson sampling (36).

In light of the above limitations, one is tempted to look for a statistical-inference technique suitable for handling complicated real-world distribution functions, and finds the answer in Statistical Bootstrap which is a procedure for estimating the distribution of an estimator by resampling (often with replacement) one's data or a model estimated from the data. Bootstrapping has been widely used as an alternative to statistical-inference based on the assumption of a parametric-model when that assumption is in doubt, or where parametric inference is impossible or requires complicated formulas for the calculation of standard errors.

This naturally motivates the use of statistical-Bootstrap for the nonparametric setting of MAB discussed above. In fact, most of the existing algorithms for nonparametric MABs are based on different versions of the Bootstrap in one way or the other (27; 5; 34). However, these methods crucially rely on *artifical history/pseudo-rewards* to perform well, and can perform sub-optimally without a suitable mechanism to generate such artificial-history/pseudo-rewards (34). Additionally, these bootstrap sampling based algorithms cannot account for uncertainty that does not come from the observed data (32). In other words, they do not have a mechanism to incorporate *prior* knowledge about the environment which can be utilized to enhance the performance of the algorithm. This efficient harnessing of prior knowledge for improved performance is hallmark of Bayesian algorithms, and we are unaware of any bandit algorithm that enjoys the flexibility of being completely Bayesian and still efficient in the nonparametric MAB setting. Essentially, this calls for an extension of the parametric Thompson sampling, which is already Bayesian, but suffers its nemesis in the non-parametric MAB setting for reasons discussed before. Consequentially, this leads us to the following question,

Can we design a truly Bayesian algorithm that performs efficiently in the setting of nonparametric multi-arm bandits?

We answer this question in the affirmative by designing an algorithm that draws from the strengths of Bayesian Nonparametric (BN) priors. In the past, a nice line of work utilized BN priors on the *function spaces*, i.e. Gaussian Process (GP) priors, to contribute the well known GP-UCB algorithm (46), but it's not clear how this can be naturally adapted to the nonparametric MAB setting that we are interested in the current paper, and we believe that a more natural choice of BN priors in the context of multi-arm bandits would be the priors on the space of probability distributions instead of those on a much larger function space (restricted only by the choice of their smoothness) (38). Dirichlet Processes (DPs), denoted as $\mathrm{DP}(\alpha, \mathrm{F}_0)$, (where α and F_0 are the related hyperparameters, known as the concentration parameter, and the base measure respectively), fall in the category of BN priors on the space of probability distributions, and have been widely used in real world statistical applications (9; 30; 22), . We extend the strength of DPs to the multi-arm bandit setting by contributing Dirichlet Process Posterior sampling (DPPS).

DPPS directly treats reward distribution functions as *random objects*, modeling them using DP priors, and easily updating these priors utilizing the property of conjugacy of DP priors to obtain DP posteriors, and making decisions based on the the posterior probability of optimal actions induced by these DP posteriors. Since no parametric class of distribution for the arm reward distributions is assumed apriori, DPPS allows for modeling arbitrary reward distributions, and hence is amenable to the non-parametric MAB setting. This is in contrast to parametric Thompson sampling which assumes a parametric class for reward distribution apriori, and puts a prior on a scalar/vector parameter, often the sufficient-statistic of that parametric-class, thereby restricting its application to a small set of problems. Furthermore, these parametric priors do not enjoy the property of conjugacy very often, and it becomes challenging to sample from their posterior distributions even for the restricted class of problems they can model appropriately. We will illustrate this strength of DPPS in a series of numerical experiments in Section 5 for different bandit environments.

Since DPPS is a Bayesian algorithm, it provides a principled mechanism to incorporate prior knowledge about the bandit environment, specifically through the base measure of the DP priors. In fact, based on the hyperparameter, α, of the DP prior it's easy to delineate uncertainty captured in DP priors/posteriors into two parts – contributions from the observed data and contributions from the prior. In the limit of $\alpha \to 0$, one recovers the noninformative DP prior, also referred to as *Bayesian Bootstrap* which is the basis for Non Paramteric Thompson sampling introduced in (39). We discuss this in Section 4.1, and also give a proof of concept of the flexibility of DPPS to incorporate prior knowledge about bandit environment through a simple example in Section 5. Additionally, in Section 6, we extend an elegant information-theoretic analysis framework for parametric Thompson sampling to a wider set of probability matching algorithms that derive the posterior probability of optimal actions using a valid/proper Bayesian strategy. This extension, along with an important lemma on the tail of random distributions sampled from DP prior/posterior shall lead us to the result of Theorem 8 which provides an upper bound on the Bayesian regret of DPPS.

2 Problem formulation

In this section, we formalize the problem of multi-arm bandits and introduce the necessary notation. We also discuss Thompson-sampling, a Bayesian probability matching algorithm, in order to lay some ground for introducing its nonparametric counterpart, DPPS, later in this paper.

Multi-armed bandits In the K-arm bandit problem, the agent is presented with K arms/distributions/actions $\{p_k\}_{k=1}^K$. At time-steps $t = 0, 1, \ldots$, the agent executes an action $A_t \in \mathcal{A}$, $\mathcal{A}$ being the set of actions such that $|\mathcal{A}| = K$; then it observes the corresponding reward $R_{t,A_t} \in \mathcal{X}$. In this paper, we choose $\mathcal{X}$ to the set of σ-sub-Gaussian random variables, i.e. $\mathbb{E}\left[e^{(X-\mathbb{E}[X])t}\right] \le e^{\frac{\sigma^2 t^2}{2}}$, $\forall X \in \mathcal{X}$, and for all s. Let $R_t \equiv (R_{t,a})_{a\in\mathcal{A}}$ be the vector of rewards at time t. The "true reward-vector distribution" $p^\star$ is seen as a distribution over $\mathcal{X}^{|\mathcal{A}|}$ that is itself randomly drawn from the family of distributions $\mathcal{P}$. We assume that, conditioned on $p^\star$, $(R_t)_{t\in N}$ is an iid sequence with each element R_t distributed according to $p^\star$. The agent's experience through time-step t is encoded by a history $\mathcal{H}_t = (A_1, R_{1,A_1}, \ldots, A_t, R_{t,A_t})$. The action A_t is chosen based on $\mathcal{H}_t$ utilizing a sequence of deterministic functions, $\pi = (\pi_t)_{t\in N}$, so that $\pi_t(a) = \mathbb{P}(A_t = a|\mathcal{H}_t)$. π is usually referred to as randomized *policy*. The T period *regret* of the sequence of actions, $A_1, .., A_T$, induced by π, is the random variable,

$$\text{Regret}(T, \pi) = \sum_{t=1}^{T} \mathbb{E}[R_{t,A^\star} - R_{t,A_t}]$$

where $A^\star \in \mathcal{A}$ is the optimal action, i.e. $A^\star \in \underset{a\in\mathcal{A}}{\text{argmax}}\, \mathbb{E}[R_{1,a}|p^\star]$. In this paper, we study the expected regret or *Bayesian regret* given as follows,

$$\mathbb{E}\left[\text{Regret}(T, \pi)\right] = \mathbb{E}\left[\sum_{t=1}^{T} [R_{t,A^\star} - R_{t,A_t}]\right],$$

where the expectation integrates over random reward realizations, the prior distribution of $p^\star$, and algorithmic randomness.

Further notation We set $\alpha_t(a) = \mathbb{P}\left(A^\star = a|\mathcal{H}_t\right)$ to be the posterior distribution of $A^\star$. Also, we use the shorthand notation $\mathbb{E}_t[\cdot] = \mathbb{E}_t[\cdot|\mathcal{H}_t]$ for conditional expectations under the posterior distribution, and similarly write $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot|\mathcal{H}_t)$. For two probability measures P and Q over a common measurable space, if P is absolutely continuous with respect to Q, the *Kullback-Leibler divergence* between P and Q is

$$\texttt{KL}(P||Q) = \int P \log\left(\frac{dP}{dQ}\right) dP \tag{1}$$

where $\frac{dp}{dq}$ is the Radon–Nikodym derivative of p with respect to q. For a probability distribution p over a finite set $\mathcal{X}$, the *Shannon entropy* of p is defined as $\mathbb{H}(p) = -\sum_{x\in\mathcal{X}} p(x)\log(p(x))$. The *mutual information* under the posterior distribution between two random variables $X_1 : \Omega \to \mathcal{X}_1$, and $X_2 : \Omega \to \mathcal{X}_2$, denoted by

$$I_t(X_1; X_2) := \mathtt{KL}\left(\mathbb{P}\left((X_1, X_2) \in \cdot|\mathcal{H}_t\right) \;||\; \mathbb{P}\left(X_1 \in \cdot|\mathcal{H}_t\right)\mathbb{P}\left(X_2 \in \cdot|\mathcal{H}_t\right)\right), \tag{2}$$

is the Kullback-Leibler divergence between the joint posterior distribution of X_1 and X_2 and the product of the marginal distributions. Note that $I_t(X_1; X_2)$ is a random variable because of its dependence on the conditional probability measure $\mathbb{P}\left(\cdot|\mathcal{H}_t\right)$.

Thompson Sampling Thompson Sampling is a specific class of probability matching algorithms which *matches* in each round, the action-selection probability to the posterior probability-distribution of optimal action, i.e. $\mathbb{P}(A_t = a|\mathcal{H}_t) = \mathbb{P}(A^\star = a|\mathcal{H}_t)$. First, a parametric class for the reward distribution functions $\{\pi_k\}_{k=1}^K$ is assumed, such that for each arm there is a θ_a which maps the arm to a distribution in that class. Thompson sampling is a Bayesian algorithm in the sense that it considers each of these unknown θ_a, as a random variable initially distributed according to a prior distribution, i.e., $\theta_a \sim \pi_{a,0}$, and this prior evolves to a posterior distribution, $\pi_{a,t}$, in round t, through Bayes rule, as rewards are obtained in each round. At each time, a sample $\theta_{a,t}$ is drawn from each posterior $\pi_{a,t}$, and then the algorithm chooses to sample $a_t = \arg\max_{a\in\{1,\dots,K\}}\{\mu(\theta_{a,t})\}$, where $\mu(\theta_{a,t})$ represents the mean of the parametric reward distributions with parameter $\theta_{a,t}$.

3 Background on Dirichlet processes

Before discussing the main algorithm proposed in this paper, It is important to concretely discuss a few key aspects concerning Dirichlet Processes, and this is what we do in this section.

Dirichlet distribution is a multivariate generalization of the Beta distributions. We denote the Dirichlet distribution of parameters $(\alpha_1, ..., \alpha_n)$ by $\mathtt{Dir}(\alpha_1, ..., \alpha_n)$ whose density function is given by $\frac{\Gamma(\sum_{i=1}^n \alpha^i)}{\prod_{i=1}^n \Gamma(\alpha^i)} \prod_{i=1}^n w_i^{\alpha^i - 1}$ for $(w_1, ..., w_n) \in [0, 1]^n$ such that $\sum_{i=1}^n w_i = 1$

Dirichlet Processes In the Bayesian formalism (see also section A for more details), an unknown object is treated as a random variable which is then assumed to be drawn from a prior distribution. A Bayesian solution requires developing methods of computation of the posterior distribution from this prior based on available information about the unknown object. When the unknown object is a probability measure (a cumulative distribution function in the present paper, to be precise), one then faces a non-trivial question of how to even define a prior on an infinite dimensional object and also take care of the constraints of a probability measure (sum up to 1 over its support). An elegant solution was offered in (19) wherein the author introduced the idea of a Dirichlet process (DP) – a probability distribution on the space of probability measures which induces finite-dimensional Dirichlet distributions when the data are grouped. To look at it concretely, consider a random probability measure, G, on some nice (e.g. Polish) space Θ (e.g. $\mathbb{R}$). G is said to be DP distributed with base probability measure F (e.g. a Gaussian, Beta, Bernoulli, etc) and concentration parameter $\alpha \in \mathbb{R}^+$, denoted as $G \sim \mathtt{DP}(\alpha, F)$, if

$$(G(A_1), ..., G(A_r)) \sim \mathtt{Dir}(\alpha F(A_1), ..., \alpha F(A_r))$$

for every finite measurable partition $A_1, ..., A_r$ of the space Θ.

Having witnessed the construction of DP priors on the space of probability measures, one naturally wonders, how to derive posteriors from these priors, and for that we discuss the important property of *conjugacy* in some nonparametric priors.

Conjugacy In the Bayesian parametric framework, one can usually use Bayes rule for deriving posteriors for parametric models, however for non-parametric case, Bayes rule cannot be used in general (see Appendix A.1 for technical details). Posteriors for some nonparametric priors can be derived utilizing the property of conjugacy. Particularly, an observation model $M \in \mathcal{G}$, and the family of priors $\mathcal{Q}$ are called conjugate if, for any sample size n and any observation sequence $X_1, ..., X_n$, the posterior under any prior $Q \in \mathcal{Q}$ is again an element of $\mathcal{Q}$. Also, merely possessing the property of conjugacy is not enough to form a viable Bayesian prior. For example, a generalization of DPs is the so-called Neutral To The Right (NTTR) processes (14). Entire family of NTTR is known to be conjugate, but besides the specific case of DPs, there's no known explicit method of obtaining *posterior indices* in other members of the NTTR family. This leads us to discuss the form of DP posteriors next.

Dirichlet Process posteriors Let $X_1, \dots, X_n$ be a sample from an unknown real-valued distribution G_0 where $X_i \in \mathbb{R}$. To estimate G_0 from a Bayesian perspective (see Appendix A) we put a prior on the set of all distributions $\mathcal{G}$ and then we compute the posterior distribution of G_0, given $\mathbf{X}_n = (X_1, ..., X_n)$. Let's put a DP prior on the set $\mathcal{G}$. Correspondingly, Let $\texttt{DP}(\alpha, F_0)$, denote the DP prior. The distribution F_0 can be thought of as a prior guess at the true distribution G_0. The number α controls how tightly concentrated the prior is around F_0. With a DP prior on G_0, the posterior of G_0, given $\mathbf{X}_n = (X_1, ..., X_n)$, enjoys *conjugacy*, i.e, it is itself a DP given as, $\texttt{DP}(\alpha_n, \overline{F}_n)$, where, the *posterior indices*, α_n, and $\overline{F}_n$ are obtained as follows (19; 22),

$$\alpha_n = \alpha + n, \, \overline{F}_n = \frac{n}{\alpha + n} F_n + \frac{\alpha}{\alpha + n} F_0 \tag{3}$$

Here F_n is the *empirical distribution function* given $X_1, ..., X_n$, i.e., $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq x)$.

Note how the posterior index, $\overline{F}_n$, exhibited in Eq. 3 combines the information from observations (via the empirical cdf, $F_n(x)$) with that available from the prior (using F_0). This is a crucial property of DPs that our algorithm , DPPS, shall harness in order to account for information obtained via observed data, and the prior information. One can easily see that as $\alpha \to 0$, DPs can only account for uncertainty obtained via observations, with no role of prior anymore, and we discuss this next.

Bayesian Bootstrap A very useful idea in statistical inference has been that of Statistical Bootstrap (17), and a Bayesian version of Bootstrap was introduced in (41). Interestingly, this Bayesian version of Bootstrap can also be derived as a special case of the DP posteriors (23). Specifically, the weak limit, $\texttt{DP}(n, \sum_{i=1}^{n} \delta_{X_i})$, (also referred to as the *noninformed limit* sometimes) of the DP posterior, $\texttt{DP}(\alpha_n, \overline{F}_n)$, as $|\alpha| \to 0$ is known as Bayesian Bootstrap (BB), and is given as,

$$\mathrm{BB_n} := \texttt{DP}(n, \sum_{i=1}^{n} \delta_{X_i}) = \sum_{i=1}^{n} W_i \delta_{X_i} \tag{4}$$

where $\mathbf{W}_n = (W_1, ..., W_n) \sim \texttt{Dir}(1, ..., 1)$, and X_i are the observed data points. The mean of a random distribution drawn from Bayesian-Bootstrap can be easily seen to be the dot-product between the weights and the observed data-points, i.e.,

$$\mu(BB_n) = \sum_{i=1}^{n} W_i X_i = \langle \mathbf{W}_n, \mathbf{X}_n \rangle \tag{5}$$

As we shall see in Sec 4, the idea of Bayesian Bootstrap forms the basis for a bandit algorithm introduced in (39). Next we discuss an important representation of DP priors/posteriors that make them amenable to practical applications.

Stick-breaking representation of DPs With the necessary details about DP prior and posterior distributions set, one naturally asks how to draw sample from these distributions because this is necessary if one wants to do any sort of statistical inference using DPs. Particularly, the form of DP posterior (indices) in Eq.3 provide little information to answer this question. A representation of random measures sampled from DPs, reported in (45), known as Stick Breaking representation of DPs, provides an answer to this question. In general, Stick-breaking measures (25) are almost surely discrete random probability measures that can be represented as,

$$Q(\cdot) = \sum_{i=1}^{N} q_i \delta_{Z_i}(\cdot) \tag{6}$$

where δ_{Z_i} is a discrete measure concentrated at Z_i, and q_i are random weights, generated independent of Z_i, such that $q_i \in [0, 1]$, and $\sum_{i=1}^{N} q_i = 1$. As one can guess, this is analogous to breaking an actual stick into pieces, and hence the name. The author of (45) reported that if these weights, q_i, are constructed such that,

$$q_1 = V_1,\ (q_i)_{i=2}^{N-1} = V_i \prod_{j=1}^{i-1}(1 - V_j),\ q_N = \prod_{i=1}^{N}(1 - V_i) \tag{7}$$

$$V_i \overset{iid}{\sim} \mathtt{Beta}(1, \alpha),\ Z_i \overset{iid}{\sim} F,\ i = 1, 2, ...N \tag{8}$$

and N is ∞, then the generated random discrete measure, P, in Eq.7 (with N as ∞) is such that, $P \sim \mathtt{DP}(\alpha, F)$. Ofcourse, for computation one can't have N as ∞, and the infinite series is truncated at some finite N, such that a probability mass, $q_N = 1 - \sum_{i=1}^{N-1} q_i = \prod_{i=1}^{N}(1 - V_i)$, is put at the last point, Z_N, and this construction ensures that all weights, q_i sum up to one. This finite Stick-breaking representation has been widely used (25; 29) thanks to its provable optimality in closely approximating the infinite series (see also Appendix B for this and for more details on choosing finite N, etc).

Iterative form of DP posterior With the stick-breaking representation of DP priors at hand, one wonders how to compute DP posteriors in a practically feasible way, and for this, an iterative form of DP posterior comes in handy given as follows (8; 45),

$$Q_i(\cdot) \overset{d}{=} V_i \delta_{X_{i-1}} + (1 - V_i) Q_{i-1}(\cdot) \tag{9}$$

Here $V_i \sim \mathtt{Beta}(1, \alpha + i)$, and $\overset{d}{=}$ denotes equality in distribution. Beginning with a DP prior, Q_0, generated using the stick-breaking method (Eqs.7-8), the recursion in Eq.9 can be used to obtain the DP posterior, given N observations $\{X_1, ..., X_N\}$, as follows,

$$Q_N \overset{d}{=} V_N \delta_{X_N} + \sum_{i=1}^{N-1} \left[V_i \prod_{j=i+1}^{N} (1 - V_j) \right] \delta_{X_i} + \left[\prod_{i=1}^{N} (1 - V_i) \right] Q_0. \tag{10}$$

4 Dirichlet process posterior sampling

Having established the necessary background, we are now ready to introduce our algorithm, DPPS.

Algorithm 1 gives the pseudo-code for DPPS. Instead of assuming a *parametric* class for the reward generating distribution of each arm, and then putting a prior on the parameter, we model the reward generating distribution of each of the arms $\{p_k\}_{k=1}^{K}$ using a corresponding DP. In each round, DPPS operates as follows: a random distribution, D_k, is sampled from the current DP posterior for each of the K arms utilizing the stick-breaking representation of the DP posterior of Eq. 10; To select an arm, the probability matching principle is followed, that is, the arm with the highest probability of being optimal (i.e. one corresponding to the highest of the means, $\mu(D_k)$, of the random measures, D_k) in that round is pulled. It is denoted as $I(t)$. After observing the reward $R_{t,I(t)}$, the history of observed rewards, $\mathbf{R}_{I(t)}$, for this arm is updated, and the DP posterior of the pulled arm is updated using the $N_{I(t)}$ observations. Clearly, DPPS can be seen as Thompson sampling wherein the prior/posterior are nonparametric, instead of parametric[1]. As a result, most of the theoretical

[1] Note that DPPS is a (non-parametric) Bayesian algorithm that utilizes probability-matching principle for arm selection, and hence is in *exact* sense, Thompson sampling.

guarantees and proof techniques for Thompson-sampling apply to DPPS as well. An important practical advantage of DPPS is that one does not need to know the parametric-class of distribution functions. More crucially, the posteriors in parametric Thompson-sampling are often not available in exact form, and must be approximated using expensive inference techniques. This issue does not arise in DPPS, as the resulting posteriors in DPPS are always DP, and one can sample from DP posteriors utilizing their stick-breaking representation discussed in Section 3. Also, DPPS enjoys the same flexibility as that of DP posteriors in utilizing information obtained from the observed data and that from some prior knowledge. In other words it combines the (data-driven) strength of vanilla (Bayesian) Bootstrapping with the flexibility of incorporating prior beliefs.

Algorithm 1 Dirichlet Process Posterior Sampling

```
Require: Horizon $T$, number of arms $K$, arm parameters – Distribution $F_{0,k}$, constant $\alpha_{0,k}$ for
    $k \in \{1, ..., K\}$
 1: for $k = 1...K$, do
 2:     Set $\mathbf{R}_k = [\,]$, $F_k = F_{0,k}$, $\alpha_k = \alpha_{0,k}$, and $N_k = 0$
 3: end for
 4: for $t = 1...T$, do
 5:     # Sample model (a random measure):
 6:     for $k = 1...K$, do
 7:         Sample $D_k \sim \mathtt{DP}(\alpha_k, F_k)$
 8:     end for
 9:     # select and apply action:
10:     $I(t) = \text{argmax}_{k \in \{1,...,K\}}\{\mu(D_k)\}$
11:     Pull arm $I(t)$ and observe reward $R_{t,I(t)}$
12:     Update history $\mathbf{R}_{I(t)} = (\mathbf{R}_{I(t)}^\top, R_{t,I(t)})^\top$ and count $N_{I(t)} \leftarrow N_{I(t)} + 1$.
13:     # Posterior update
14:     $\alpha_{I(t)} \leftarrow \alpha_{I(t)} + 1$
15:     $F_{I(t)} = \frac{1}{\alpha_{I(t)}} \sum_{x \in \mathbf{R}_{I(t)}} \delta_x + \frac{\alpha_{0,I(t)}}{\alpha_{I(t)}} F_{0,I(t)}$
16: end for
```

Algorithm 2 Non parametric Thompson sampling (39)

```
Require: Horizon $T \geq 1$, number of arms $K \geq 1$
 1: for $k = 1...K$, do
 2:     Set $R_k := [1]$, and $N_k := 1$
 3: end for
 4: for $t = 1...T$, do
 5:     for $k = 1...K$, do
 6:         Sample $\mathbf{W}_k \sim \text{Dir}(1_{N_k})$ where $1_{N_k} = \underbrace{(1, ..., 1)}_{N_k \text{ times}}$.
 7:     end for
 8:     $I(t) := \text{argmax}_{k \in \{1,...,K\}}\{\langle \mathbf{R}_k, \mathbf{W}_k \rangle\}$
 9:     Pull arm $I(t)$ and observe reward $R_{t,I(t)}$.
10:     Update history $\mathbf{R}_{I(t)} := (\mathbf{R}_{I(t)}^\top, R_{t,I(t)})^\top$ and count $N_{I(t)} := N_{I(t)} + 1$
11: end for
```

4.1 Noninformative limit of the DPPS

In (39), authors introduced a non-parametric algorithm for multi-arm bandits, calling it Non-Parametric Thompson Sampling (NPTS), although noting that NPTS is not a Bayesian algorithm, and that it is not Thompson sampling in *strict* sense. They proved its asymptotic optimality, and showed empirically that NPTS also does well non-asymptotically. Algorithm 2 gives the pseudo-code for NPTS. In what follows, we show that NPTS is a special case of DPPS. In NPTS, the arms are selected in each-round (see lines 9-10 in Algorithm 2) based on the argmax of the weighted average of the observed rewards (weights drawn from a Dirichlet distribution). Interestingly, this is exactly the mean of a random distribution drawn from a Bayesian-Bootstrap (Eq. 5), and Bayesian-Bootstrap is a special case of Dirichlet-processes (see Eq. 4). Therefore, NPTS is a special case of DPPS, when the DP for each arm is taken to be the Bayesian-Bootstrap, and cannot account for prior knowledge (following our discussion in Section 3 on Bayesian Bootstrap and DP posteriors).

5 Numerical experiments

In this section, we exhibit empirical performance of DPPS on challenging Bernoulli bandit, Beta bandit, and a real-world agriculture dataset. In the experiments that follow, all regret plots exhibit average regret over 200 independent runs and $10\% - 90\%$ quantile levels. For Bernoulli bandits we compare DPPS with Beta-Bernoulli Thompson sampling and UCB. Whereas for the other two environemnts we compare with UCB and a generalized version of Beta/Bernoulli (3) TS because it's difficult to implement usual parametric Thompson sampling in those settings (especially for the DSSAT bandit setting). Impressive performance of DPPS in a Gaussian bandit environment (with both mean and variance unknown to the algorithmic agent) is also shown in Sec. C. A discussion on the general choice of (hyper)parameters of DP priors (α,F_0, and truncation level of DP prior) is given in Section D. Corresponding code is provided in the supplementary material.

Bernoulli and Beta bandits Here we evaluate DPPS in a 6 arm Bernoulli bandit setting with means [0.3, 0.4, 0.45, 0.5, 0.52, 0.55]. Note that all means being close to 0.5 makes it a challenging setting. We compare performance of DPPS with UCB and another algorithm which is tailor-made for Bernoulli bandit environment – Beta/Bernoulli Thompson Sampling (TS). The prior for Beta/Bernoulli TS is set as Beta(1,1) (uniform). The base measure of the DP prior is also set as Uniform distribution ($\texttt{Beta}(1, 1)$) for all the arms. Fig. 1 shows the perfomance of all the algorithms. Clearly, DPPS does as well as Beta/Bernoulli TS. This is impressive because unlike Beta/Bernoulli TS , DPPS is unaware of the parametric class of the reward distribution (Bernoulli), and still performed as well as Beta/Bernoulli TS. With the same DP priors we also run DPPS in a Beta bandit environment (with same mean as the Bernoulli bandit setting and scale factor of 5). Fig. 1 (right) also shows performance of DPPS in this setting, and clearly DPPS outperforms other baselines.

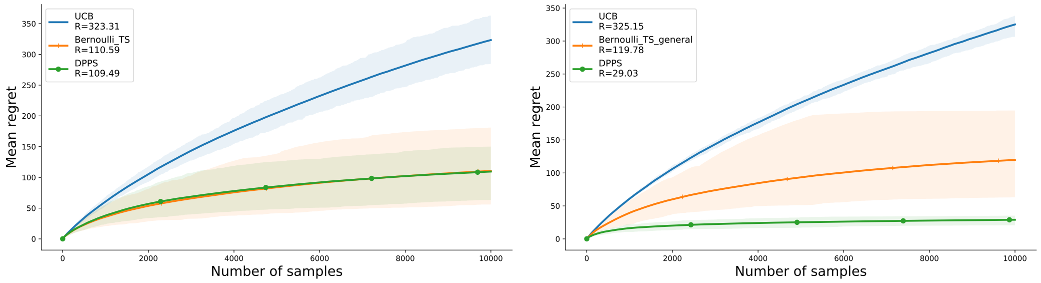

Figure 1: Comparison of average regret in the Bernoulli bandit setting (left), and Beta Bandit setting (right) discussed in the text.

DSSAT bandits Next, we illustrate the performance of DPPS on a challenging practical decision-making problem using the DSSAT-2 (Decision Support System for Agrotechnology Transfer) simulator (24; 21). Harnessing more than 30 years of expert knowledge, this simulator is calibrated on historical field data (soil measurements, genetics, planting dates, etc) and generates realistic crop yields. Such simulations can be used to explore crop management policies in silico before implementing them in the real world, where their actual effect may take months or years to manifest themselves. More specifically, we model the problem of selecting a planting date for maize grains among 7 possible options, all other factors being equal, as a 7-armed bandit. The resulting distributions incorporate historical variability as well as exogenous randomness coming from a stochastic meteorologic model. In Figure 2, we show distributions of crop yields generated from the DSSAT2 simulator. Note that these distributions are right-skewed, multimodal and exhibit a peak at zero corresponding to years of poor harvest. Given this, they hardly fit to a convenient parametric model (e.g single-parameter-exponential-family, etc). Note that, arm 3 is optimal and the distributions have bounded support and hence can be normalized to within $[0, 1]$. Like for the Bernoulli bandit case, we use DP priors with uniform base measures ($\texttt{Beta}(1, 1)$) for DPPS.

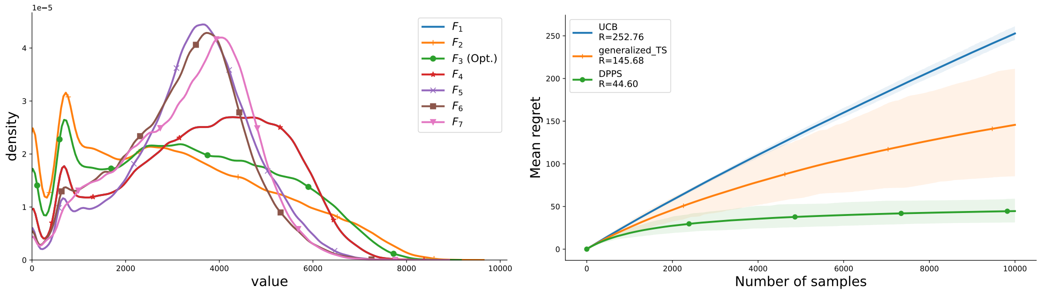

Figure 2: Reward distributions from DSSAT simulator (left) and regret performances of bandit strategies (right) in the DSSAT environment.

Since a vanilla version of Thompson sampling is no longer feasible for DSSAT environment, we instead compare DPPS against a version of Beta/Bernoulli Thompson sampling, introduced in (3), that is adapted for general stochastic rewards based on a Bernoulli trial in each round with the obtained rewards as the mean parameter of the Bernoulli random variable. The same $\texttt{Beta}(1, 1)$ prior is used for generalized TS as well. Fig.2 clearly shows DPPS outperforming generalized TS and UCB by a huge margin, and this example highlights the strength of DPPS as Bayesian nonparametric algorithm over it's closest parametric-counterpart of generalized TS. Note that so far we used agnostic base measures for the DP priors ($\texttt{Beta}(1, 1)$), i.e. these base measures (and hence the corresponding DP priors) do not convey any special knowledge about the bandit environment. However, DPPS allows for encoding this prior knowledge about the bandit environment through base-measures of the DP priors, and we illustrate this next using a simple example.

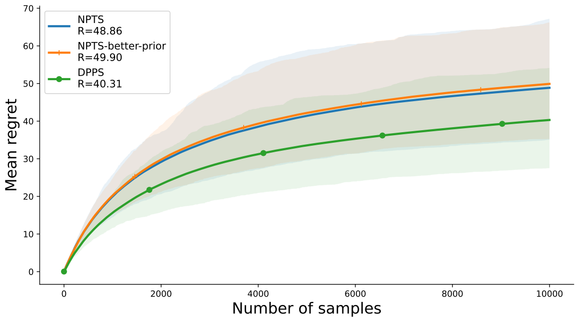

Figure 3: Average regret in the DSSAT bandit environment with beneficial priors for both NPTS and DPPS.

Incorporating prior knowledge through DPPS Recall from Sec. 4.1 that NPTS is a special case of DPPS in the Bayesian Bootstrap limit of the DP prior. Therefore, the base measure for NPTS for a particular arm is empirical CDF of the reward distributions based on current observations for that arm, beginning with some initial atomic base measure, δ_{X_k}, for each of the k-arms. Given that the base measure is an empirical CDF, in NPTS, it's not possible to utilize even some first order prior information about the bandit environment that may be available. This is, however, possible in general cases of DPPS through the continuous base measures of DP priors. This can be clearly exhibited through a simple example. We start DPPS with a more informed choice of priors, i.e.

instead of $\texttt{Beta}(1, 1)$ base measure for the DP priors for all the arms, we express more confidence in the third (optimal) arm by using $\texttt{Beta}(1, 0.1)$ as base measure for this arm. We compare this with a version of NPTS that starts with initial artificial reward observation of $X_k = 0.01$ for all but the third arm (for which it uses a value of 1). Fig. 3 confirms better performance of DPPS with this choice of DP priors, and no change in performance of NPTS even with initial condition that heavily favors the third arm.

Computational cost of DPPS Improved performance and flexibility of DPPS (and other Bootstrap based algorithms such as NPTS) does come with higher computational cost. For example, in the 6-arm Bernoulli bandit environments of horizon $T = 10000$, average run-time (over 200 independent runs) of DPPS was around 18 seconds, whereas that of parametric TS (conjugate prior/posterior) was 2-3 seconds. For the 7 arm DSSAT bandit problem, DPPS takes around 20 seconds, NPTS takes around 16 seconds. Sec. E gives a detailed overview of the computational complexity of DPPS. All this said, this run time of DPPS can be significantly brought down by utilizing *self-similarity* (22) of DP posteriors and parallel computation of DP posteriors that a construction exploiting this self-similarity would enjoy, which we plan to do in future.

6 Regret upper bounds for DPPS

In this section, we generalize the information theoretic analysis of Thompson sampling introduced in (43) to a wider class of probability matching algorithms, and then derive upper bound on Bayesian regret of DPPS. We begin by summarizing the key-steps in the original analysis (43) that are crucial for the aforesaid extension, and also include complete proofs for the sake of completion in Sec. G.

Firstly, the Bayesian regret is re-expressed in terms of the entropy of the posterior distribution of optimal action, and an upper bound on *information ratio*,

Lemma 1. *For any $T \in N$, provided that $\Gamma_t \leq \Gamma$ almost surely for each $t \in 1, .., T$, $\mathbb{E}\left[\mathrm{Regret}(T, \pi^{TS})\right] \leq \sqrt{\Gamma \mathbb{H}(\alpha_1) T}$.*

The information ratio, $\Gamma_t := \frac{(\mathbb{E}_t[R_{t,A_*} - R_{t,a}])^2}{I_t(A^\star; R_{t,a})}$ is defined as the ratio of the square of the instantaneous expected regret by choosing action a to the instantaneous *information gain* about optimal action $A^\star$ if action a is chosen. Clearly, bounding Bayesian regret of an algorithm boils down to bounding the information-ratio of that algorithm. Particularly, for Thompson-sampling, in σ-sub-Gaussian reward noise bandit setting, it's easy to obtain the following bound

Lemma 2.

$$\Gamma_t \leq 2|\mathcal{A}|\sigma^2.$$

This bound when combined with Lemma 1 and upper bound of $\log K$ for entropy of any posterior distribution of optimal action leads to the following bound on the Bayesian regret of Thompson sampling,

Theorem 3.

$$\mathbb{E}\left[\mathrm{Regret}(T, \pi^{TS})\right] \leq \sigma \sqrt{2K(\log K)T}\,,$$

The proof of Lemma 2 hinges on two crucial steps, and we highlight those referring the reader to Sec. G for more details. First, re-writing of the instantaneous per-step Bayesian regret by utilizing the probability matching property of Thompson sampling, $\mathbb{P}_t(A^\star = a) = \mathbb{P}_t(A_t = a)$, as follows,

$$\begin{aligned}\mathbb{E}_t\left[R_{t,A^\star} - R_{t,A_t}\right] &= \sum_{a \in \mathcal{A}} \mathbb{P}_t(A^\star = a)\mathbb{E}_t\left[R_{t,a}|A^\star = a\right] - \sum_{a \in \mathcal{A}} \mathbb{P}_t(A_t = a)\mathbb{E}_t[R_{t,a}|A_t = a] \quad (11)\\ &= \sum_{a \in \mathcal{A}} \mathbb{P}_t(A^\star = a)\left(\mathbb{E}_t\left[R_{t,a}|A^\star = a\right] - \mathbb{E}_t[R_{t,a}]\right).\end{aligned}$$

Second, bounding this instantaneous per-step regret by bounding ($\mathbb{E}_t\left[R_{t,a}|A^\star = a\right] - \mathbb{E}_t[R_{t,a}]$), This is done by an application of the variational formula (12) for the KL divergence, $\mathtt{KL}(P||Q)$, between two absolutely continuous measures, P and Q,

Fact 4.

$$\mathtt{KL}(P||Q) = \sup_X\{\mathbb{E}_P[X] - \log \mathbb{E}_Q[\exp\{X\}]\}.$$

If we substitute, the random variable, $X \equiv X(t) = R_{t,a} - \mathbb{E}_t[R_{t,a}]$, with $P = \mathbb{P}_t(R_a|A^\star = a)$ and $Q = \mathbb{P}_t(R_a)$ in the above variational formula, and when $X(t)$ is σ-sub-Gaussian, it's easy to obtain the following bound,

Lemma 5.

$$\mathbb{E}_t\left[R_{t,a}|A^\star = a\right] - \mathbb{E}[R_{t,a}] \leq \sigma\sqrt{2D(\mathbb{P}_t(R_{t,a}|A^\star = a)||\mathbb{P}_t(R_{t,a})}).$$

6.1 Admissible probability matching algorithms

It's easy to notice in the preceding analysis that there's no restriction on $\mathbb{P}_t(A^\star = a)$ to be derived using a Bayes-rule based posterior-distributions of arm-rewards,$\mathbb{P}_t(R_a)$ as is done in parametric Thompson sampling. This choice is rather implicit, given the decision theoretic and information theoretic *coherency* of Bayesian framework (48; 50). However, Bayesian-framework is not limited to Bayes-rule based derivation of posterior distributions. Another *valid* Bayesian approach (31; 23) for obtaining posteriors is leveraging the property of *conjugacy* as discussed in Sec 3. In particular, most *nonparametric* priors do not satisfy the necessary conditions for Bayes rule (See A.1), and one must rely on their conjugacy property to derive the corresponding posteriors. Therefore, all probability matching algorithms which derive $\mathbb{P}_t(R_a)$ (and hence $\mathbb{P}_t(A^* = a)$) using a valid Bayesian approach are *admissible* in the information theoretic analysis of (42). Additionally, these admissible algorithms would enjoy similar bounds as parametric Thompson sampling on their information-ratio (and consequently Bayesian regret), if they satisfy *auxiliary conditions* required from the original analysis.

For the case of σ-sub Gaussian reward noise discussed before, it is easy to see that we require the following auxiliary conditions: In each round t, (1) the instantaneous reward noise, $X(t)$, in Lemma 5, is σ-sub-Gaussian; (2) $\mathtt{KL}(\mathbb{P}_t(R_a|A^\star = a)||\mathbb{P}_t(R_a))$ in Lemma 5 is well defined. The second condition holds if $P_t(A^\star = a) > 0$ owing to a classical fact in conditional probability (49),

Fact 6. *For any random variable Z and event $E \subset \Omega$, where Ω is the probability space, if $\mathbb{P}_t(E) = 0$, then $\mathbb{P}_t(E|Z) = 0$ almost surely. Conversely, for any $x \in \mathcal{X}$ with $\mathbb{P}_t(X = x) > 0$, $\mathbb{P}_t(Y|X = x)$ is absolutely continuous with respect to $\mathbb{P}_t(Y)$.*

DPPS satisfies all the conditions above: It is *admissible* since it utilizes a valid Bayesian approach, i.e. conjugacy of DP priors/posteriors, to derive $\mathbb{P}_t(A^\star = a)$; Also, clearly, $\mathbb{P}_t(A^\star = a) > 0$ whenever the base measure, F_0, of the DP prior (and hence of the corresponding DP posterior), $\mathtt{DP}(\alpha, F_0)$, is non-null. Finally, the following property of the tail of DP priors/posteriors ensures σ-sub-Gaussian nature of the instantaneous reward noise, $X(t)$, whenever the base measure, F_0, of the DP prior, $\mathtt{DP}(\alpha, F_0)$, is σ-sub-Gaussian,

Fact 7 (From (15))**.** *Let $F \sim \mathtt{DP}(\alpha, F_0)$, then almost surely the tails of F and distributions sampled from the DP posterior of F, $\mathtt{DP}(\alpha + n, \overline{F_n})$, given samples $X_1, ..., X_n$, are dominated by (and are much smaller than) the tails of F_0.*

This leads us to the following upper bound on Bayesian regret of DPPS,

Theorem 8. *For the setting of σ-sub-Gaussian rewards, starting with a DP-prior with a σ sub-Gaussian base measure, the Bayesian regret of DDPS satisfies*

$$\mathbb{E}\left[\text{Regret}(T, \pi^{DPPS})\right] \leq \sigma\sqrt{2K(\log K)T}\,,$$

where the expectation is taken over the randomness in the policy and the prior of the environment.

7 Conclusions and Perspectives

In this paper, we introduced a Bayesian non parametric algorithm based on Dirichlet processes, DPPS, for multi-arm bandits that combines the strength of (Bayesian) Bootstrap with a principled mechanism of incorporating and exploiting prior information about the bandit environment. DPPS enjoys similar optimality guarantees on Bayesian regret as parametric Thompson sampling, and among other advantages of DPPS over its parametric counterpart is its *flexibility*. This is because the stick-breaking implementation of DPPS introduced in this paper can be used for different types of bandit environments, contrary to parametric Thompson sampling whose implementations differ according to the bandit environment, and can easily lead to intractable posteriors (except for a few special cases) which need to be approximated using approximate inference schemes such as MCMC, variational inference, etc, and, if not done carefully, such approximate-inference based Thompson sampling has been shown to incur sub-optimal performance, even in simple settings (36). Next, we discuss a few research directions.

Firstly, we point that DPs are not the only Bayesian nonparametric priors on the space of distribution functions, and further generalization of DPPS is possible. For example, other probability matching algorithms using Pitman-Yor (37) processes and Pólya-Tree priors (10; 9) can be useful generalizations of DPPS. Next note that, although we derived DPPS for multi-arm bandits without any structure, we believe the results in this paper could carry out on other types of online learning problems studied in (43), e.g. linear bandits. Also, since all the Bayesian regret guarantees of Thompson sampling in (43) hold for Information directed sampling (IDS) (42), we conjecture that a DPPS version of IDS may also be optimal following the arguments in our paper. This can be useful since IDS has been specifically shown to be asymptotically optimal for problems wherein Thompson sampling and UCB type algorithms fail (28) to be so. A major hurdle in IDS is however its computational-complexity, owed to intractable posteriors that result because of the use of parametric-posteriors based on Bayes-rule. It would be interesting, in future work, to study a nonparametric variant of IDS that utilizes DP posteriors as it would overcome these computational issues,

Finally, we consider DPPS as a generic *design principle*, based on Bayesian non-parametric statistics, that can be extended to the setting of Markov Decision Processes (MDPs) as well. This can be done in both model-based and model-free scenarios. In the former, a Posterior Sampling Reinforcement Learning (PSRL) (33; 18) algorithm based on Dirirchlet Process posteriors is definitely a promising direction of research. For the model-free scenario, one can extend Randomized Least Square Value Iteration (RLSVI) from its current Bayesian-Bootstrap based implementations (32; 35) to a full-fledged DP implementation to inject uncertainty that does not come from the observed data in a principled manner similar to that shown in this paper. We leave these intriguing research questions and extensions for future work.

References

[1] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26. JMLR Workshop and Conference Proceedings, 2011.

[2] Rajeev Agrawal. Sample mean based index policies by o (log n) regret for the multi-armed bandit problem. *Advances in applied probability*, 27(4):1054–1078, 1995.

[3] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

[4] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

[5] Akram Baransi, Odalric-Ambrym Maillard, and Shie Mannor. Sub-sampling for multi-armed bandits. In *Machine Learning and Knowledge Discovery in Databases: European Conference,*

ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14, pages 115–131. Springer, 2014.

[6] Dorian Baudry, Patrick Saux, and Odalric-Ambrym Maillard. From optimality to robustness: Dirichlet sampling strategies in stochastic bandits. In *NeurIPS 2021-35th International Conference on Neural Information Processing Systems*, 2021.

[7] Denis Belomestny, Pierre Menard, Alexey Naumov, Daniil Tiapkin, and Michal Valko. Sharp deviations bounds for dirichlet weighted sums with application to analysis of bayesian algorithms. *arXiv preprint arXiv:2304.03056*, 2023.

[8] David Blackwell and James B MacQueen. Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2):353–355, 1973.

[9] IsmaÃĢl Castillo. Bayesian nonparametric statistics, st-flour lecture notes. *arXiv preprint arXiv:2402.16422*, 2024.

[10] Ismaël Castillo. Pólya tree posterior distributions on densities. 2017.

[11] Murray K Clayton and Donald A Berry. Bayesian nonparametric bandits. *The Annals of Statistics*, 13(4):1523–1534, 1985.

[12] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[13] Wesley Cowan, Junya Honda, and Michael N Katehakis. Normal bandits of unknown means and variances. *Journal of Machine Learning Research*, 18(154):1–28, 2018.

[14] Jyotirmoy Dey, RV Erickson, and RV Ramamoorthi. Some aspects of neutral to right priors. *International statistical review*, 71(2):383–401, 2003.

[15] Hani Doss and Thomas Sellke. The tails of probabilities chosen from a dirichlet prior. *The Annals of Statistics*, 10(4):1302–1305, 1982.

[16] Dean Eckles and Maurits Kaptein. Thompson sampling with the online bootstrap. *arXiv preprint arXiv:1410.4009*, 2014.

[17] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

[18] Ying Fan and Yifei Ming. Model-based reinforcement learning for continuous control with posterior sampling. In *International Conference on Machine Learning*, pages 3078–3087. PMLR, 2021.

[19] Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[20] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. *Advances in neural information processing systems*, 23, 2010.

[21] Romain Gautron, Emilio J Padrón, Philippe Preux, Julien Bigot, Odalric-Ambrym Maillard, and David Emukpere. gym-dssat: a crop model turned into a reinforcement learning environment. *arXiv preprint arXiv:2207.03270*, 2022.

[22] Subhashis Ghosal. The dirichlet process, related priors and posterior asymptotics. *Bayesian nonparametrics*, 28:35, 2010.

[23] Subhashis Ghosal and Aad W van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

[24] Gerrit Hoogenboom, Cheryl H Porter, Kenneth J Boote, Vakhtang Shelia, Paul W Wilkens, Upendra Singh, Jeffrey W White, Senthold Asseng, Jon I Lizaso, L Patricia Moreno, et al. The dssat crop modeling ecosystem. In *Advances in crop modelling for a sustainable agriculture*, pages 173–216. Burleigh Dodds Science Publishing, 2019.

[25] Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American statistical Association*, 96(453):161–173, 2001.

[26] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer, 2012.

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

[28] Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR, 2017.

[29] Pietro Muliere and Luca Tardella. Approximating distributions of random functionals of ferguson-dirichlet priors. *Canadian Journal of Statistics*, 26(2):283–297, 1998.

[30] Peter Müller, Fernando Andrés Quintana, Alejandro Jara, and Tim Hanson. *Bayesian nonparametric data analysis*, volume 1. Springer, 2015.

[31] Peter Orbanz. Construction of nonparametric bayesian models from parametric bayes equations. *Advances in neural information processing systems*, 22, 2009.

[32] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29, 2016.

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

[34] Ian Osband and Benjamin Van Roy. Bootstrapped thompson sampling and deep exploration. *arXiv preprint arXiv:1507.00300*, 2015.

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

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

[37] Jim Pitman and Marc Yor. Bessel processes and infinitely divisible laws. In *Stochastic Integrals: Proceedings of the LMS Durham Symposium, July 7–17, 1980*, pages 285–370. Springer, 2006.

[38] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003.

[39] Charles Riou and Junya Honda. Bandit algorithms based on thompson sampling for bounded reward distributions. In *Algorithmic Learning Theory*, pages 777–826. PMLR, 2020.

[40] Kathryn Roeder. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85(411):617–624, 1990.

[41] Donald B Rubin. The bayesian bootstrap. *The annals of statistics*, pages 130–134, 1981.

[42] Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Advances in Neural Information Processing Systems*, 27, 2014.

[43] Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

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

[45] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

[46] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

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

[48] Abraham Wald. *Statistical decision functions*. Wiley, 1961.

[49] David Williams. *Probability with martingales*. Cambridge university press, 1991.

[50] Arnold Zellner. Optimal information processing and bayes's theorem. *The American Statistician*, 42(4):278–280, 1988.

Supplementary Materials

The following content was not necessarily subject to peer review.

A General Bayesian framework

In this section, we highlight a generalized Bayesian framework, and the conditions for existence of posteriors and, when they exist, methods of deriving posteriors from priors. Most of these results are standard in Bayesian-non-parametric statistics, and we refer the reader to (23; 31) for details.

A general Bayesian modeling problem can be formulated as follows. We choose prior Q on parameter $\Theta \in \mathbf{T}$ and the observation model M as P_Θ, observation space as $\mathbf{X}$. To summarize, both Bayesian and non-paramteric Bayesian models can be written as follows,

$$\Theta \sim Q, \tag{12}$$
$$X_1, ..., X_n|\Theta \sim P_\Theta \tag{13}$$

Whereas for Bayesian parametric models the parameter space $\mathbf{T}$ is finite-dimensional (e.g. $\mathbb{R}^d$), it's infinite for Bayesian non-parametric models. Thus in order to define a non-parametric Bayesian model, we have to define a probability distribution (the prior) on an infinite-dimensional space. A distribution on an infinite-dimensional space $\mathbf{T}$ is a stochastic process with paths in $\mathbf{T}$.

For more clarity, the DP model can be re-written in the framework of Eqs. 14 as follows,

$$\Theta \sim DP(\alpha, G_0), \tag{14}$$
$$X_1, ..., X_n|\Theta \sim \Theta \tag{15}$$

The goal in Bayesian (both parametric and nonparmetric) inference is to figure out the posterior which is a probability kernel given as,

$$q[\cdot, x] = \mathbb{P}(\Theta \in \cdot|X = x).$$

For existence of q the following is required,

Theorem 9. *If $\mathbf{T}$ is a standard Borel space, $\mathbf{X}$ a measurable space, and a Bayesian model is specified as in Eqs. 14, the posterior q exists*

Having established the existence properties, let's discuss different ways of obtaining posteriors, given observations. In Bayesian framework, there are two ways, Bayes rule and Conjugacy, and we give existence results for each of these,

A.1 Bayes-rule

It's a popular update rule, however it's not always possible to use Bayes-rule for obtaining posteriors. The following theorem makes it concrete,

Theorem 10. *(Bayes' Theorem). Let $\mathbf{M} = P(\cdot, \mathbf{T})$ be an observation model and $Q \in PM(T))$ a prior (PM denotes space of probability measures on $\mathbf{T}$). Require that there is a σ-finite measure μ on $\mathbf{X}$ such that $P(\cdot, \Theta) \ll \mu$ for every $\Theta \in \mathbf{T}$. Then the posterior under conditionally i.i.d. observations $X_1, ..., X_n$ is given as below, and $\mathbb{P}\{P(X_1, ..., X_n) \in 0, \infty\} = 0$*

$$Q(d\Theta|X_1 = x_1, ..., X_n = x_n) = \frac{\prod_{i=1}^n P(x_i|\Theta)}{P(X_1, ..., X_n)} Q(d\Theta)$$

A.2 Conjugacy

For most non-parametric priors, the important absolute continuity condition in Theorem 10 doesn't hold, and hence Bayes' rule is not applicable. For example, If $\mathbb{P}[d\Theta|X_{1:n}]$ is the posterior of a Dirichlet process, then there is no σ-finite measure ν which satisfies $\mathbb{P}[d\Theta|X_{1:n} = x_{1:n}] \ll \nu$ for all $x_{1:n}$. In particular, the prior does not, and so there is no density $P(\Theta|x_{1:n})$ (23). In order to remedy this curse on non-parametric priors, the most important alternative to Bayes theorem for computing posterior distributions is conjugacy. Suppose $\mathbf{M}$ is an observation model, and consider now a family $\mathcal{Q} \subset PM(\mathbb{T})$ of prior distributions, rather than an individual prior. We assume that the family $\mathcal{Q}$ is indexed by a parameter space $\mathbf{Y}$, that is, $\mathbf{M} = \{Q_y | y \in \mathbf{Y}\}$. Many important Bayesian models have the following two properties:

- The posterior under any prior in $\mathcal{Q}$ is again an element of $\mathcal{Q}$; hence, for any specific set of observations, there is an $y' \in \mathbf{Y}$ such that the posterior is $Q_{y'}$
- The posterior parameter y' can be computed from the data by a simple, tractable formula.

The above two points define the property of conjugacy. We saw in the main paper that DP priors enjoy conjugacy, and saw the simple update formula for the posterior, that resulted thanks to this property of conjugacy. For more details, we refer the reader to (31).

B Finite Stick breaking representation of Dirichlet Process priors

The finite stick-breaking representation of DP priors discussed in the main paper (Eqs.7-8) has been pivotal in the success of DP based Bayesian-nonparametric models. A major reason for this success is that such truncated representation is provably efficient (25). Particularly, to quantify the accuracy loss owing to truncation consider the quantities, $T_K = (\sum_K^\infty p_k)^r$ and $U_K = \sum_K^\infty p_k^r$, where K is the level at which the representation is truncated,

$$\mathbb{E}(T_K(r, a, b)) = (\frac{\alpha}{\alpha + r})^{K-1}, \tag{16}$$

$$\mathbb{E}(U_K(r, a, b)) = (\frac{\alpha}{\alpha + r})^{K-1} \frac{\Gamma(r)\Gamma(\alpha + 1)}{\Gamma(\alpha + r)} \tag{17}$$

Notice that both expressions decay exponentially fast in K, and hence good accuracy is achieved for moderate K. Fig. 4 shows an application of this scheme to sample random measures from a DP prior, $\mathrm{DP}(\alpha, F_0)$ for two different values of concentration parameter, α. In order to give more intuition to appreciate the utility of DPs for nonparametric inference, We given an example on inference on a galaxy-dataset. We also used this (and some other) benchmarks to validate the performance of our StickBreaking module for DPPS.

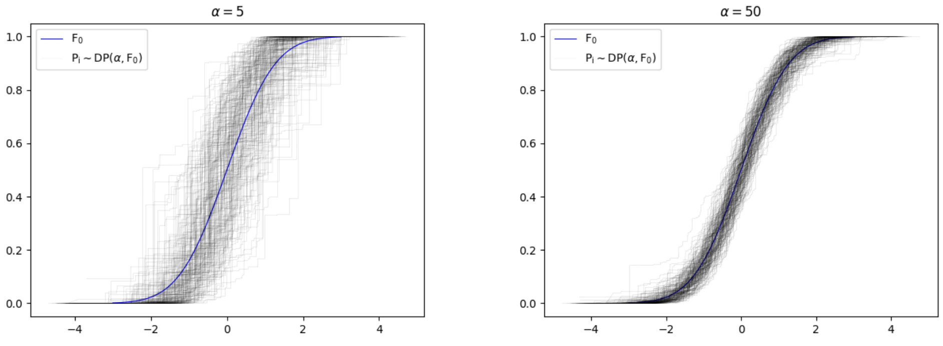

Figure 4: 200 random measures sampled from $\mathrm{DP}(\alpha, \mathrm{F}_0)$ where $\alpha = 5$ (left) and 50 (right), $\mathrm{F}_0 = N(0, 1)$

DPs for galaxy data-set We illustrate the application of Dirichlet processes for density estimation on a data set from the astronomy literature (40). The measurements are velocities at which galaxies in the Corona-Borealis region are moving away from our galaxy. If the galaxies are clustered, the velocity density will be multimodal, with clusters corresponding to modes. This happens to be the case, and the multi-modal nature is evident in the CDF of the data in Figure 5 where the left and right regions of the CDF are almost flat, and most mass resides in the center. Starting with a $\mathrm{DP}(\alpha, N(0,1))$ prior, we obtain a DP posterior, and the spread of distributions sampled from the DP posterior (not shown) can be seen as confidence-set of the density estimate through Dirichlet process.

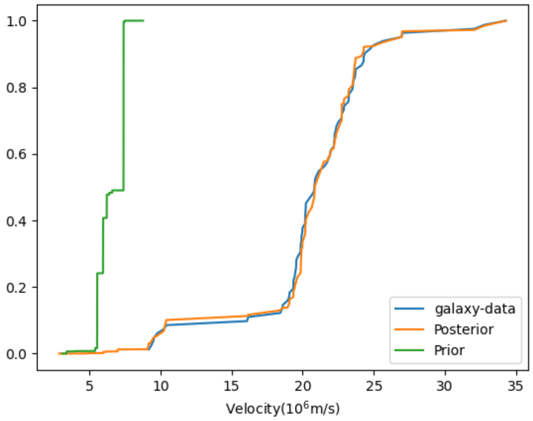

Figure 5: A random measure sampled from DP prior, DP posterior compared against original galaxy dataset distribution.

C DPPS for a Gaussian bandit

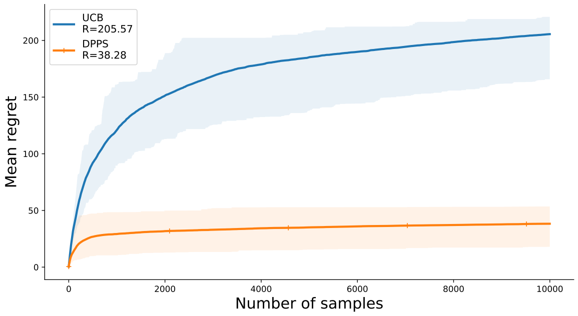

Figure 6: DPPS for a challenging Gaussian bandit setting

A challenging bandit setting is that of Gaussian bandit environment with both mean and variance of the underlying Gaussian distribution as unknown (13) to the bandit algorithm. Here we exhibit performance of DPPS in such a 7 arm Gaussian bandit environment $\{N(\mu_k, \sigma_k)\}_{k=1}^{K=7}$. The mean and variance of Gaussian bandit arms are sampled independently from a Gaussian such that $\mu_k \sim N(0, 0.5)$ and $\sigma_k = |\psi_k|, \psi_k \sim N(0, 0.5)$. Cumulative Regret averaged over 100 runs on one of the sampled instance of bandit environment is shown in Fig. 6. Excellent performance of DPPS is evident. In this experiment, we chose $\alpha = 2$, base measure of DP, F_0, as $N(0, 0.5)$.

D Choice of hyperparameters in numerical experiments

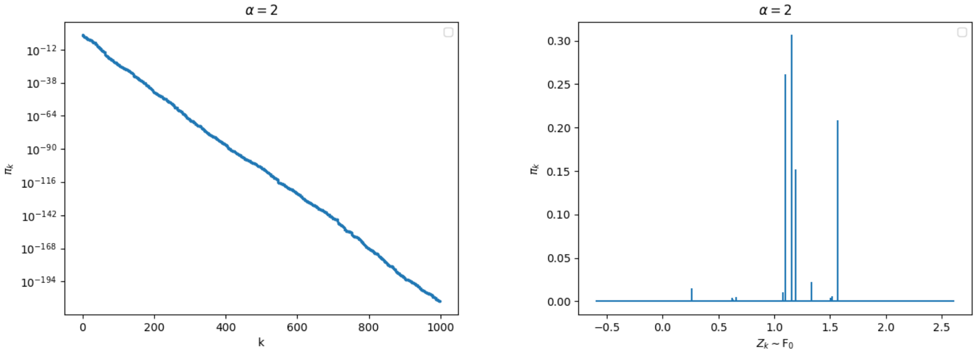

Figure 7: Plot of first 1000 stick-breaking probability measure weights, π_k, for $\mathrm{DP}(\alpha = 2, F_0)$ with k (left) and with $Z_k \sim F_0(= N(0,1)$ (right)

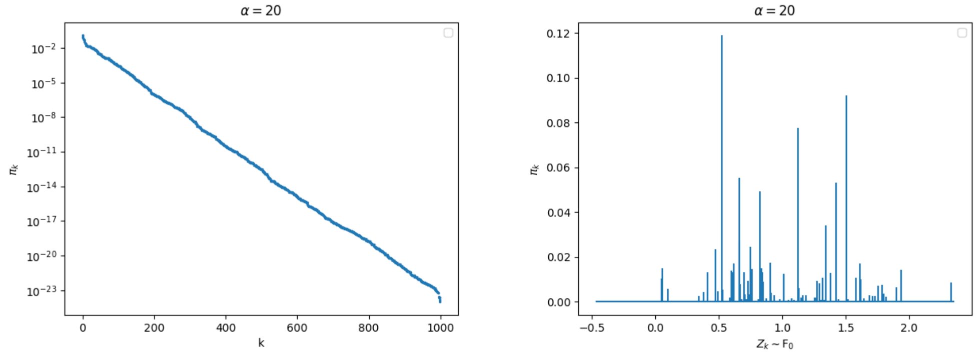

Figure 8: Plot of first 1000 stick-breaking probability measure weights, π_k, for $\mathrm{DP}(\alpha = 20, F_0)$ with k (left) and with $Z_k \sim F_0(= N(0,1)$ (right)

Two hyperparameters in DPPS are α (concentration parameter) and k_t (i.e. truncation level) in the stick breaking representation of DP prior (not the posterior), $\mathrm{DP}(\alpha, F_0)$. We used $\alpha = 2$ and $k_t = 100$ in all the experiments. Note that the choice of α directly influences the choice of k_t. This is because the number of weights q_i in the stick breaking representation, $\sum q_i \delta_{x_i}$, carrying significant probability mass increase with increase in α ($V_i \sim \mathrm{Beta}(1, \alpha)$), and for higher α one needs to increase k_t. For example, with $\alpha = 20$, we took $k_t = 300$, and we got similar results, with a slight increase in computation cost though. An easy way to determine k_t is to plot the stick breaking weights and remove stick breaking weights that are below a certain threshold (we chose 10^{-10} randomly). This relationship between α and stick breaking probability weights, q_i, can be seen in a simple example of $\mathrm{DP}(\alpha, F_0)$ as shown in figs. 7 and 8. Whereas for lower value of α only few weights have significant mass, for higher α the weights are more evenly spread compared to lower α case.

Choice of base measure, F_0, of DP prior For choosing, F_0, the tail of the underlying reward distribution and a fact on the support of DPs is important.

Lemma 11 (Support of DPs, see (22))**.** *In the weak topology, the support of DP(α, F_0) is characterized as all probability measures $P^\star$ whose supports are contained in that of F_0*

Thus, choosing Beta(1,1) for a bandit problem with $\sigma = 10$, subGaussian noise is not a good idea. Similarly, theorem 8 on Bayesian regret of DPPS, shows that choosing F_0 with σ-subGaussian tails corresponding to tails of the reward noise is optimal.

E Running costs of DPPS

Here we detail the computational costs associated to a single-arm in each round. Let n denote the number of observations for the arm. The important consideration in quantifying the running cost of DPPS is to scrutinize the posterior update step,

$$Q_n \quad = \quad V_n\delta_{X_n} + \sum_{i=1}^{n-1}\left[V_i \prod_{j=i+1}^{n}(1-V_j)\right]\delta_{X_i} + \left[\prod_{i=1}^{n}(1-V_i)\right]Q_0 \tag{18}$$

Here, one needs to sample n beta random variables and have $\mathcal{O}(n)$ multiplications of these random variables, one for each of the past observations. Thus the running cost of DPPS is $\mathcal{O}(n)$ for each arm. DPPS also incurs a fixed memory and computational cost of $\mathcal{O}(K)$, sampling a DP prior, Q_0, where K is the truncation level of the DP prior. Clearly, this additional but constant (in number of rounds and memory) cost is the difference between computational complexities of DPPS and NPTS (which needs similar $\mathcal{O}(n)$ multiplications between $\mathbf{X_n}$ and $\mathbf{W_n} \sim \mathrm{Dir}(\mathbf{n}; \mathbf{1}, ..., \mathbf{1})$ random variables), and arises because of additional flexibility of DPPS in incorporating prior knowledge.

F Further related work

To the best of our knowledge, Dirichlet Processes in the context of bandits were first used in (11) to study a version of the single-arm Gittin's index problem, when the probability distribution of the arm is assumed to be DP distributed. Use of Bootstrapping for Thompson sampling seems to have appeared first in (16), which was further improved and made more systematic in (34) where the authors also showed equivalence of Bootstrap-Thompson sampling (for Bernoulli-bandits) and Thompson sampling with Beta/Bernoulli priors in an exact sense, and speculated this equivalence for a wide class of bandit-environments if a proper mechanism for generating *artifical history* (or prior information) could be identified. As shown in the current paper, DPPS provides a neat and principled mechanism for incorporating prior information (or gnerating artificial history), and generalizes this equivalence. Non-Parametric Thompson sampling (NPTS) and Multinomial Thompson Sampling (TS) were introduced in (39) without highlighting any concrete Bayesian connection of the former algorithm. NPTS was adapted for robustness in (6). Some discussions concerning Bayesian interpretation of NPTS using DPs appeared in (7) who provided a refined analysis of Multinomial TS. Aligning towards non-Bayesian side, a sample mean based algorithm guaranteeing $O(\log N)$ instance-dependent regret appeared in (2), a sub-sampling based algorithm was reported in (5) and analyzed for a two-arm bandit setting; a nonparametric Bootstrap based algorithm was reported in (27), and regret bounds derived for a Bernoulli bandit environment.

G Technical derivations

This section gives proofs of lemmas in the main paper extracted here for completion from (43)

G.1 Proof of Fact 1

For any $T \in \mathbb{N}$, if $\Gamma_t \leq \overline{\Gamma}$ almost surely for each $t \in \{1, .., T\}$,

$$\mathbb{E}\left[\mathrm{Regret}(T, \pi^{\mathrm{TS}})\right] \leq \sqrt{\overline{\Gamma} H(\alpha_1) T}.$$

Proof. Recall that $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{H}_t]$ and we use I_t to denote mutual information evaluated under the base measure $\mathbb{P}_t$. Then,

$$\begin{aligned}
\mathbb{E}\left[\text{Regret}(T, \pi^{\text{TS}})\right] \overset{(a)}{=} \mathbb{E}\sum_{t=1}^{T}\mathbb{E}_t\left[R_{t,A^\star} - R_{t,A_t}\right] &= \mathbb{E}\sum_{t=1}^{T}\sqrt{\Gamma_t I_t\left(A^\star; (A_t, R_{t,A_t})\right)} \\
&\leq \sqrt{\overline{\Gamma}}\left(\mathbb{E}\sum_{t=1}^{T}\sqrt{I_t\left(A^\star; (A_t, R_{t,A_t})\right)}\right) \\
&\overset{(b)}{\leq} \sqrt{\overline{\Gamma} T \mathbb{E}\sum_{t=1}^{T} I_t\left(A^\star; (A_t, R_{t,A_t})\right)},
\end{aligned}$$

where (a) follows from the tower property of conditional expectation, and (b) follows from the Cauchy-Schwartz inequality. We complete the proof by showing that expected information gain cannot exceed the entropy of the prior distribution. For the remainder of this proof, let $Z_t = (A_t, R_{t,A_t})$. Then, using tower rule of conditional expectations we have,

$$\mathbb{E}_t\left[I_t\left(A^\star; Z_t\right)\right] = I\left(A^\star; Z_t | Z_1, ..., Z_{t-1}\right),$$

and therefore,

$$\begin{aligned}
\mathbb{E}\sum_{t=1}^{T} I_t\left(A^\star; Z_t\right) = \sum_{t=1}^{T} I\left(A^\star; Z_t | Z_1, ..., Z_{t-1}\right) &\overset{(c)}{=} I\left(A^\star ; Z_1, ...Z_T\right) \\
&= H(A^\star) - H(A^\star | Z_1, ...Z_T) \\
&\overset{(d)}{\leq} H(A^\star),
\end{aligned}$$

where (c) follows from the chain rule for mutual information, and (d) follows from the non-negativity of entropy. □

G.2 Proof of Fact 5

Proof. Define the random variable $X(t) = R_{t,a} - \mathbb{E}_t\left[R_{t,a}\right]$. Then, for arbitrary $\lambda \in \mathbb{R}$, applying Fact 4 to λX yields

$$\begin{aligned}
D\left(\mathbb{P}_t\left(R_{t,a}|A^\star = a^\star\right) || \mathbb{P}_t(R_{t,a})\right) &\geq \lambda\mathbb{E}_t\left[X|A^\star = a^\star\right] - \log\mathbb{E}_t\left[\exp\{\lambda X\}\right] \\
&= \lambda\left(\mathbb{E}_t[R_{t,a}|A^\star = a^\star] - \mathbb{E}_t\left[R_{t,a}\right]\right) - \log\mathbb{E}_t\left[\exp\{\lambda X\}\right] \\
&\geq \lambda\left(\mathbb{E}_t[R_{t,a}|A^\star = a^\star] - \mathbb{E}_t\left[R_{t,a}\right]\right) - (\lambda^2\sigma^2/2).
\end{aligned}$$

Maximizing over λ yields the result. □

G.3 Proof of Fact 2

Proof.

$$\begin{aligned}
\mathbb{E}_t\left[R_{t,A^\star} - R_{t,A_t}\right]^2 &\overset{(a)}{=} \left(\sum_{a\in\mathcal{A}}\mathbb{P}_t(A^\star = a)\left(\mathbb{E}_t\left[R_{t,a}|A^\star = a\right] - \mathbb{E}_t[R_{t,a}]\right)\right)^2 \\
&\overset{(b)}{\leq} |\mathcal{A}|\sum_{a\in\mathcal{A}}\mathbb{P}_t(A^\star = a)^2\left(\mathbb{E}_t\left[R_{t,a}|A^\star = a\right] - \mathbb{E}_t[R_{t,a}]\right)^2 \\
&\leq |\mathcal{A}|\sum_{a,a^\star\in\mathcal{A}}\mathbb{P}_t(A^\star = a)\mathbb{P}_t(A^\star = a^\star)\left(\mathbb{E}_t\left[R_{t,a}|A^\star = a^\star\right] - \mathbb{E}_t[R_{t,a}]\right)^2 \\
&\overset{(c)}{\leq} \frac{|\mathcal{A}|}{2}\sum_{a,a^\star\in\mathcal{A}}\mathbb{P}_t(A^\star = a)\mathbb{P}_t(A^\star = a^\star)D_{KL}\left(\mathbb{P}_t(R_{t,a}|A^\star = a^\star) \;||\; \mathbb{P}_t(R_{t,a})\right) \\
&\overset{(d)}{=} \frac{|\mathcal{A}|I(A^\star; R_{t,A_t})}{2}
\end{aligned}$$

where (b) follows from the Cauchy–Schwarz inequality, (c) follows from Fact 5, and (a) follows from Eq.11and (d) from the standard definition of mutual-information. □