Université de Lille

Thèse de doctorat

Bayesian reinforcement learning with Dirichlet processes

Apprentissage par renforcement Bayésien avec processus de Dirichlet

Thèse de doctorat de l'Université de Lille
préparée à Inria Lille Nord-Europe

École doctorale n°631 Mathématiques, sciences du numérique et de leurs interactions (MADIS)
Spécialité de doctorat: Informatique

Thèse présentée et soutenue à Villeneuve d'Ascq, le **30 January 2026**, par

SUMIT VASHISHTHA

Composition du Jury :

Ismaël Castillo Professeur, LPSM @ Sorbonne University, Paris, France	Président du Jury
Pierre Alquier Professeur, ESSEC Business School, Asia-Pacific, Singapore	Rapporteur
Shie Mannor Professeur, Technion-Israel Institute of Technology, Haifa, Israel	Rapporteur
Audrey Durand Professeur agrégé, Université Laval, Québec City, Canada	Examinatrice
Ismaël Castillo Professeur, LPSM @ Sorbonne University, Paris, France	Examinateur
Odalric-Ambrym Maillard Chargè de recherche HDR, Equipe SCOOL, Inria, Lille, France	Directeur de thèse

There's no joy in the finite; There's joy only in the infinite

-Chandogya Upanishad (8.12.6)

Truth, like gold, is to be obtained not by its growth, but by washing away from it all that is not gold

-Leo Tolstoy

Acknowledgements

I would like to express my gratitude to the following people:

- Dr. Odalric-Ambrym Maillard, my doctoral-advisor, for believing in my ideas and letting me pursue them for this thesis, and for his comments on drafts of published/in-preparation-articles/this-manuscript
- Dr. Pierre Alquier and Dr. Shie Mannor, Dr. Audrey Durand and Dr. Ismael Castillo for agreeing to be reviewers of this thesis and examiners for the defense, respectively, and for writing papers that have motivated numerous people, such as myself.
- Dr. Eric Moulines, Dr. Laetitia Jourdan, and Dr. Sylvian Salvati for being the CSI members for this thesis.
- my mentors and collaborators from previous research experiences – Dr. JA Scott Kelso and Dr. KR Sreenivasan, Dr. Mahendra K. Verma and Dr. Siddhartha Verma, Dr. Isaac Elishakoff, Dr. MF Baig, and Dr. George Karniadakis.
- Ms. Amelie Supervielle and other Inria members for helping me organize multiple travels. members of the university/CNRS-CRIStAL HR team for administrative support, and all the CROUS members and La-Famille members for working so hard to serve us delicious/nutritious meals
- members of the SCOOL team for many pétanque and table tennis and lawn-tennis games and for multiple seminars, and for get-togethers on multiple occasions.
- Dr. Hemant Tyagi (now at NTU Singapore) for many discussions and meals together. Other Inria researchers (Dr. Andrey Polyakov, Dr. Cristian Preda, Dr. Emilie Kaufmann,

Dr. Helene Le Cadre, Dr. Philippe Preux and Dr. Marc Tomassi) for discussions on multiple occasions. Organizers of "30 minutes of science" seminars at Inria, Lille. Members of point processes seminar group (in particular Dr. Mylene Maida, Dr. Pierre Chainais, and Dr. Remi Bardenet) for many seminars, hosting my seminar, and for encouragement about the ideas in this thesis.

- my parents, my sister Ms. Vaishnavi Sharma, and my grandparents for their patience, constant support, and care.

- Ms. Mathilde Raimbault (for biking adventures, meals, conversations); Ms. Moy Lim and Mr. Laurent Quentin (for hosting me in their collocation during my early days in Lille); Mr. Erwan Le Quinou (for discussions and table-tennis games)

Abstract

Reinforcement learning (RL) has shown great potential in applications ranging from board games to biology and beyond. However, it still remains to see proper light of the day in other challenging domains such as agro-ecology, the inefficiency of RL algorithms in such domains essentially owed to complicated real world data distributions that do not conform to standard statistical models such as single-parametric exponential family. Motivated by such real world problems, this thesis introduces a new framework for RL based on Bayesian nonparametric statistics, specifically Dirichlet Processes (DPs). The first contribution of the thesis is Dirichlet Process Posterior Sampling (DPPS), a provably optimal Bayesian non-parametric algorithm for multi-arm bandits based on DP priors. Essentially, DPPS is a probability-matching algorithm, and combines the strength of (Bayesian) bootstrap with a principled mechanism of incorporating and exploiting prior information. The thesis then proposes two new algorithms for reinforcement learning in Markov decision processes. The first one is Thompson sampling based deep Q learning, wherein, exploration is based on posterior sampling of action-value functions (represented through a deep neural network). In particular, the posterior sampling step in this algorithm utilizes a new DP based scheme, also introduced in this thesis, and which can be of independent interest in other applications of Bayesian neural networks. The second algorithm is a distributional RL algorithm that hinges on a distributional equality associated with the DPs.

Résumé

L'apprentissage par renforcement (RL) a montré un grand potentiel dans des applications allant des jeux de société à la biologie et au-delà. Cependant, il reste encore à voir le jour dans d'autres domaines difficiles tels que l'agro-écologie, l'inefficacité des algorithmes RL dans de tels domaines essentiellement dus à des distributions de données concrètes compliquées qui ne sont pas conformes aux modèles statistiques standard tels que la famille exponentielle monoparamétrique. Motivée par de tels problèmes du monde réel, cette thèse introduit un nouveau cadre pour RL basé sur les statistiques non paramétriques bayésiennes, en particulier les processus de Dirichlet (DP). La première contribution de la thèse est le Dirichlet Process Posterior Sampling (DPPS), un algorithme non paramétrique bayésien probablement optimal pour les bandits multi-bras basé sur les DP. Essentiellement, DPPS est un algorithme de correspondance de probabilité, et combine la force du bootstrap (bayésien) avec un mécanisme de principe d'incorporation et d'exploitation d'informations antérieures. La thèse propose alors deux nouveaux algorithmes pour l'apprentissage par renforcement dans les processus de décision de Markov. La première est l'apprentissage profond basé sur l'échantillonnage par Thompson, dans lequel, l'exploration est basée sur l'échantillonnage postérieur de fonctions de valeur d'action (représentées par un réseau de neurones profonds). En particulier, l'étape d'échantillonnage postérieure de cet algorithme utilise un nouveau schéma basé sur DP, également introduit dans cette thèse, et qui peut être d'intérêt indépendant pour d'autres applications des réseaux de neurones bayésiens. Le deuxième algorithme est un algorithme RL de distribution qui repose sur une égalité de distribution associée aux DP.

Contents

List of Figures

Chapter 1

Introduction

Reinforcement Learning (RL) has become a buzzword in machine learning, with its successful applications spanning from board games to biology and beyond. RL is a specific class of learning algorithms owing its roots to the concept of Pavlovian conditioning in psychology [Domjan, 2005], wherein the agent learns to acheive a certain end by interacting with the environment. A typical RL setting is exhibited in Figure 1.1. Whenever the agent (e.g. a neural network) takes actions which helps to attain a desired end, the agent is rewarded, whereas it is punished/not-rewarded whenever it's actions are drifted away from the attainment of the desired goal. In repeatedly doing so, the agent learns to take the suitable actions in the environment(mathematically, a dynamical system). The above colloquial description can be mathematically encapsulated in what is known as a Markov-Decision Process (MDP) [Puterman, 1990], $M = (\mathcal{X}, \mathcal{A}, R, P, \gamma)$ where $\mathcal{X}$ is the state space and $\mathcal{A}$ is the action space. At each time step t, the agent observes state $x_t \in \mathcal{X}$, takes action $a_t \in \mathcal{A}$, receives reward $r_t \sim R(x_t, a_t)$ and transitions to $x_{t+1} \sim P(x_t, a_t)$, with P being unknown, and the desired end in general being maximizing the sum of rewards received until some finite/infinite time horizon T. The goal then is to find best possible way to attain this desired end (mathematically find an optimal policy, $\pi^*(x_t, a_t)$.)

RL has seen great success in the last decade with its applications. Whether its alpha Go [Silver et al., 2017] defeating the world champion of the game of Go, or alphaFold predicting protein sequences [Jumper et al., 2021], RL has captivated the imagination of people in variegated

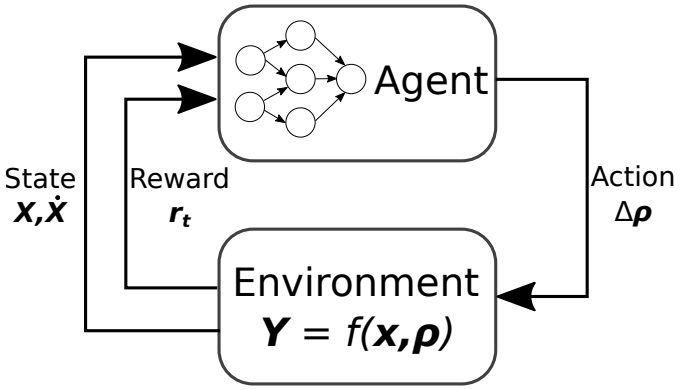

Figure 1.1: A standard RL set-up

domains of knowledge [Vashishtha and Verma, 2020, Verma et al., 2018, Degrave et al., 2022, Mankowitz et al., 2023]. However, this doesn't negate the scope for improvement in RL algorithms. A major challenge in RL has been to strike the tradeoff between learning and exploring. That is, one would like the agent to not converge to a suboptimal set of actions that may look optimal in the short run. This problem is termed as exploration$\sim$ exploitation dilemma in the reinforcement learning, and can be specifically studied in a framework knwon as Multi-Arm Bandits (MABs). In mathematical formalism, the framework of MAB is essentially an MDP with a fixed state. Numerous algorithms [Lattimore and Szepesvári, 2020] have been proposed in order to resolve this exploration$\sim$ exploitation dilemma, specifically Upper confidence bounds (UCB) based algorithm [Auer et al., 2002], Thompson sampling [Thompson, 1933], and their variants [Russo et al., 2018]. However, these algorithms tend to suffer sub-optimal performance for complicated real world problems [Phan et al., 2019, Baudry et al., 2021, Kveton et al., 2019] because of the approximations that become necessary in light of the lack of closed-form confidence-sequences (UCB) or non-conjugate posterior distributions (Thompson-sampling) in those real-world problems.

One example of complicated real-world problems arise in the field of agriculture, specifically agroecology. Consider the distributions in Fig. 1.2 for example. Each of these distributions correspond to a specific planting-date, and shows the variability (owing to exogenous randomness corresponding to weather, etc) in crop yields. Such distributions are in general unknown apriori, and in order to choose the right planting dates over time, one would typically like to run a

bandit algorithm, with each of the arms corresponding to a planting date (or the crop-yield distribution). Note that these distributions are highly complex– multi-modal for example, and in general wouldn't even correspond to broad classes of parametric distributions such as single-parameter exponential family [Wasserman, 2013]. For MAB problems corresponding to such complicated distributions, algorithms such as UCB and Thompson sampling tend to perform poorly. We will discuss this in Chapter 3, and present a new provably-optimal Bayesian algorithm based on Dirichlet Processes, which are class of Bayesian nonparametric priors on the space of probability distribution functions. We will begin with a detailed discussion of these priors in Chapter 2 wherein we will showcase their strong theoretical and simulation properties that make them amenable for complicated real-world statistical-inference problems.

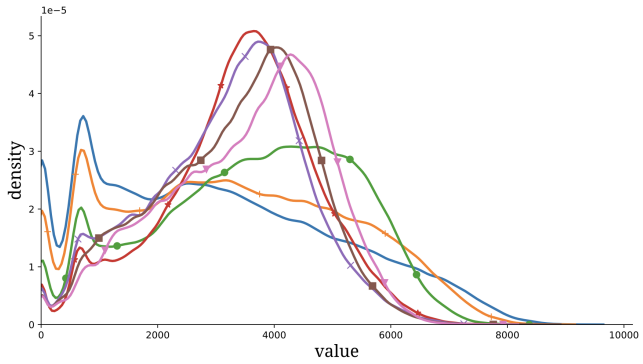

Figure 1.2: Example of distributions sampled from DSSAT simulator

In chapter 4, we will move our attention to the general framework of MDPs, and propose two new model-free reinforcement learning algorithms. The first one is Thompson sampling based deep Q learning, wherein, exploration is based on posterior sampling from action-value functions (represented through a deep neural network). In particular, the posterior sampling step in this algorithm utilizes a new DP based scheme, also introduced in this thesis, and which can be of independent interest in other applications of Bayesian neural networks. The second algorithm is a distributional RL algorithm that hinges on a distributional equality associated with the DPs.
A key feature of the algorithmic approach in this thesis is its distributional nature. Specifically, the focus shifts from learning merely mean values or some finite number of moments to iterative learning of a statistically plausible (in a Bayesian sense) discrete approximation of the true un-

derlying distribution by random-measures sampled from DPs. (We also don't resort to smoothing of these random measures and hence save on the additional computational burden of MCMC samplers [Neal, 2000].)This change in focus from mean or some moment focused learning to focusing on entire distribution is of paramount importance and underlies a promising future for statistics and fields that rely on it, e.g. machine learning, (statistical) physics, finance [Lowet et al., 2020, Bellemare et al., 2017a, Sornette, 2017, Engel et al., 2005]. We hope that this thesis makes a meaningful contribution towards this vision.

Chapter 2

Dirichlet Processes

The goal of this chapter is to review and discuss different properties of Dirichlet Processes as Bayesian nonparametric priors. In order to do that coherently, we begin by introducing a generalized Bayesian inference paradigm motivated by De-Finetti's theorem for exchangeable sequences which immediately leads us to the notion of probability distributions on the space of probability measures, and the notion of Bayesian nonparametric priors. After that, we discuss Dirichlet distribution and their infinite dimensional limit, Dirichlet Processes. We also discuss an alternate derivation of random measures distributed as a Dirichlet Processes, based on Gamma Processes, in particular this construction generalizes elegantly and enables construction of other nonparametric priors. We conclude with a very useful construction of random probability measures sampled from Dirichlet Processes, this construction makes Dirichlet Process priors highly useful in nonparametric statistical inference, and we will revisit it in the next chapters as well. We limit the exposition in this chapter to crucial properties of Dirichlet Processes, especially those utilized in the next chapters, and a more detailed account of the results for Dirichlet Processes and other Bayesian nonparametric priors can be found in [Ghosh and Ramamoorthi, 2003, Ghosal and van der Vaart, 2017, Castillo, 2024]

2.1 A general framework for Bayesian statistics

We start by considering an (ideally) infinite sequence of observations $X^{(\infty)} = (X_n)_{n\geq 1}$, defined on some probability space ($\Omega, \mathscr{F}, \mathbb{P}$) with each X_i taking values in a complete and separable metric space $\mathbb{X}$ endowed with the Borel σ-algebra $\mathscr{X}$. Throughout the thesis, as well as in the most commonly employed Bayesian models, $X^{(\infty)}$ is assumed to be exchangeable. In other terms, for any $n \geq 1$ and any permutation π of the indices $1, \ldots, n$, the probability distribution (p.d.) of the random vector ($X_1, \ldots, X_n$) coincides with the p.d. of $\left(X_{\pi(1)}, \ldots, X_{\pi(n)}\right)$. A celebrated result of De Finetti, known as De Finetti's representation theorem, states that the sequence $X^{(\infty)}$ is exchangeable if and only if it is a mixture of sequences of independent and identically distributed (i.i.d.) random variables.

Theorem 1 (De Finetti [1937])**.** *The sequence $X^{(\infty)}$ is exchangeable if and only if there exists a probability measure Q on the space $\mathcal{P}_{\mathbb{X}}$ of all probability measures on $\mathbb{X}$ such that, for any $n \geq 1$ and $A = A_1 \times \cdots \times A_n \times \mathbb{X}^{\infty}$, one has,*

$$\mathbb{P}\left[X^{(\infty)} \in A\right] = \int_{\mathcal{P}_{\mathbb{X}}} \prod_{i=1}^{n} p\left(A_i\right) Q(\,\mathrm{d}p)$$

where $A_i \in \mathscr{X}$ for any $i = 1, \ldots, n$ and $\mathbb{X}^{\infty} = \mathbb{X} \times \mathbb{X} \times \cdots$.

In the statement of the theorem, and in most of the results discussed in this thesis the space $\mathcal{P}_{\mathbb{X}}$ is equipped with the topology of weak convergence defined below,

$$p_n \to p \text{ weakly or } p_n \xrightarrow{\text{weakly}} p, \text{ if}$$
$$\int f dp_n \to \int f dp$$

for all bounded continuous functions f on $\mathbb{X}$, which makes it a complete and separable metric space (this is not the case with Hellinger or Total Variation distance [Ghosh and Ramamoorthi, 2003].)The probability Q is also termed the de Finetti measure of the sequence $X^{(\infty)}$. Such symmetry/invariance principles can be generalized to other random objects (e.g. graphs, arrays), and an excellent exposition of such matters can be found in Orbanz and Roy [2014], Kallenberg [2005].

The exchangeability assumption is usually formulated in terms of conditional independence and identity in distribution, i.e.

$$X_i \mid \tilde{p} \overset{\text{iid}}{\sim} \tilde{p},\ i \geq 1 \qquad (2.1)$$
$$\tilde{p} \sim Q.$$

Hence, $\tilde{p}^n = \prod_{i=1}^n \tilde{p}$ represents the conditional p.d. of ($X_1, \ldots, X_n$), given $\tilde{p}$. Here $\tilde{p}$ is some random probability measure defined on ($\Omega, \mathscr{F}, \mathbb{P}$) and taking values in $\mathcal{P}_{\mathbb{X}}$: its distribution Q takes on the interpretation of prior distribution for Bayesian inference. Whenever Q degenerates on a finite dimensional subspace of $\mathcal{P}_{\mathbb{X}}$, the inferential problem is usually called *parametric*. On the other hand, when the support of Q is infinite-dimensional then one typically speaks of a *nonparametric* inferential problem.

A Bayesian approach for treating nonparametric problems took some time to take off compared to its parametric counterpart. This was primarily due to two major challenges one faces in the construction of nonparametric priors (1) The support of the prior distribution should be large (i.e. should not degenerate on a finite dimensional subspace of $\mathcal{P}_{\mathbb{X}}$). (2) Posterior distributions given a sample of observations from the true probability distribution should be manageable analytically and conevenient to sample from. These properties are antagonistic in the sense that one may be obtained at the expense of the other, and this was addressed in the excellent papers of [Ferguson, 1973, 1974] who showed that the infinite extension of Dirichlet distribution priors, appropriately called Dirichlet Processes (DPs) satisfy the aforementioned conditions. Naturally then to lead us to DPs, we begin by discussing their parametric counterpart, i.e. Dirichlet distribution priors.

2.2 Dirichlet distribution and its important properties

Dirichlet distribution is known to Bayesians as the conjugate prior for the parameters of a multinomial distribution. In the formulation of the previous section, Dirichlet distribution corresponds to the priors on the space of probability measures when $\mathbb{X}$ is a simplex. Dirichlet distribution can be defined using a Gamma distribution, $\texttt{Gamma}(\alpha, \beta)$, with shape parameter $\alpha \geqq 0$, and scale parameter $\beta > 0$. For $\alpha = 0$, `Gamma` is degenerate at zero; for $\alpha > 0$, `Gamma` has density with

respect to Lebesgue measure on the real line given as follows,

$$f(z \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} e^{-z/\beta} z^{\alpha-1} I_{(0,\infty)}(z) \tag{2.2}$$

where $I_S(z)$ represents the indicator function of the set S.

Next we discuss construction of Dirichlet random (probability) vectors based on Gamma random variables,

Definition 1. *Let $Z_1, Z_2, \cdots, Z_k$ be independent random variables with $Z_j \in \texttt{Gamma}(\alpha_j, 1)$, where $\alpha_j \geqq 0$ for all j, and $\alpha_j > 0$ for some $j, j = 1, 2, \ldots, k$. The Dirichlet distribution with parameter $(\alpha_1, \cdots, \alpha_k)$, denoted by $\texttt{Dir}(\alpha_1, \cdots, \alpha_k)$, is defined as the distribution of ($Y_1, \ldots, Y_k$), where*

$$Y_j = Z_j / \sum_{i=1}^{k} Z_i \quad \text{for } j = 1, 2, \ldots, k. \tag{2.3}$$

Use of the notation $\texttt{Dir}\,(\alpha_1, \cdots, \alpha_k)$ is taken to imply that $\alpha_j \geqq 0$ for all j, and $\alpha_j > 0$ for some j. This distribution is always singular with respect to Lebesgue measure in k-dimensional space since $Y_1 + \cdots + Y_k = 1$. In addition, if any $\alpha_j = 0$, the corresponding Y_j is degenerate at zero. However, using this Gamma random variable construction and utilizing Eq. 2.2, it's easy to show that if $\alpha_j > 0$ for all j, the ($k-1$)-dimensional distribution of ($Y_1, \ldots, Y_{k-1}$) is absolutely continuous with density,

$$\begin{aligned} & f\left(y_1, \cdots, y_{k-1} \mid \alpha_1, \cdots, \alpha_k\right) \\ & \quad = \frac{\Gamma\left(\alpha_1 + \cdots + \alpha_k\right)}{\Gamma\left(\alpha_1\right) \cdots \Gamma\left(\alpha_k\right)} \left(\prod_{j=1}^{k-1} y_j^{\alpha_j - 1}\right) \left(1 - \sum_{j=1}^{k-1} y_j\right)^{\alpha_k - 1} I_{\mathrm{S}}\left(y_1, \cdots, y_{k-1}\right). \end{aligned} \tag{2.4}$$

where $\mathbb{S}$ is the simplex

$$\mathbb{S} = \left\{ (y_1, \cdots, y_{k-1}) : y_j \geqq 0, \sum_{j=1}^{k-1} y_j \leqq 1 \right\}$$

For $k = 2$, Eq. 2.4 reduces to the Beta distribution, denoted by $\texttt{Beta}\,(\alpha_1, \alpha_2)$.

We now discuss some important properties of the Dirichlet distribution useful for rest of the

chapter,

1. If $\underline{p} = (p_1, p_2, \ldots, p_k)$ is distributed as $\texttt{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_k)$, then for any partition $A_1, A_2 \ldots, A_m$ of $\mathbb{X}$, the vector $(P(A_1), P(A_2), \ldots, P(A_m)) = \left(\sum_{i \in A_1} p_i, \sum_{i \in A_2} p_i, \ldots, \sum_{i \in A_m} p_i\right)$ is a $\texttt{Dir}(\alpha_1', \alpha_2', \ldots, \alpha_k')$ where $\alpha_i' = \sum_{j \in A_i} \alpha_j$. (This follows directly from the definition of the Dirichlet distribution and the additive property of the gamma distribution: If $Z_1 \in \texttt{Gamma}(\alpha_1, 1)$, if $Z_2 \in \texttt{Gamma}(\alpha_2, 1)$, and if Z_1 and Z_2 are independent, then $Z_1 + Z_2 \in \texttt{Gamma}(\alpha_1 + \alpha_2, 1)$.)

   Remark 1. *This property suggests that it would be convenient to view the parameter* $(\alpha_1, \alpha_2, \ldots, \alpha_k)$ *as a measure* $\alpha(A) = \sum_{i \in A} \alpha_i$. *Thus every non-zero measure* α *on* $\mathbb{X}$ *defines a Dirichlet distribution and the last property takes the form*

$$(P(A_1), P(A_2), \ldots, P(A_m)) \text{ is } \texttt{Dir}(\alpha(A_1), \alpha(A_2), \ldots, \alpha(A_m))$$

   Accordingly, we will often use the convenient notation $\texttt{Dir}(\alpha)$ *to represent Dirichlet distribution with measure parameter,* α

   Now, we record the first two moments of the Dirichlet distribution.

2. If $(Y_1, \cdots, Y_k) \in \texttt{Dir}(\alpha_1, \cdots, \alpha_k)$, then

$$\begin{aligned} \mathbb{E}Y_i &= \alpha_i/\alpha \\ \mathbb{E}Y_i^2 &= \alpha_i(\alpha_i + 1)/(\alpha(\alpha + 1)) \qquad \text{and} \\ \mathbb{E}Y_iY_j &= \alpha_i\alpha_j/(\alpha(\alpha + 1)) \quad \text{for } i \neq j \end{aligned}$$

   where $\alpha = \sum_1^k \alpha_i$.

   The following is a well known Bayes property of the Dirichlet distribution,

3. If the prior distribution of $(Y_1, \cdots, Y_k)$ is $\texttt{Dir}(\alpha_1, \cdots, \alpha_k)$ and if

$$\mathbb{P}\{X = j \mid Y_1, \cdots, Y_k\} = Y_j \quad \text{a.s.} \quad \text{for } j = 1, \cdots, k,$$

   then the posterior distribution of $(Y_1, \cdots, Y_k)$ given $X = j$ is $\texttt{Dir}\left({\alpha_1}^{(j)}, \cdots, {\alpha_k}^{(j)}\right)$, where

$$\begin{aligned} {\alpha_i}^{(j)} &= \alpha_i \qquad \text{if} \quad i \neq j \\ &= \alpha_j + 1 \quad \text{if} \quad i = j. \end{aligned}$$

Contained in this property is a formula that will prove useful. Let us use $\texttt{Dir}\,(y_1, \cdots, y_k \mid \alpha_1, \cdots, \alpha_k)$ to denote the distribution function of the Dirichlet distribution, $\texttt{Dir}\,(\alpha_1, \cdots, \alpha_k)$. Then, the equality

$$\begin{aligned} &\mathbb{P}\{X = j, Y_1 \leqq z_1, \cdots, Y_k \leqq z_k\} \\ &\qquad = \mathbb{P}\{X = j\}\mathbb{P}\{Y_1 \leqq z_1, \cdots, Y_k \leqq z_k \mid X = j\} \end{aligned}$$

may be expressed in terms of $\texttt{Dir}\,(y_1, \cdots, y_k \mid \alpha_1, \cdots, \alpha_k)$, using properties 2 and 3 , as

4.

$$\begin{aligned} &\int_0^{z_1} \cdots \int_0^{z_k} y_j d\texttt{Dir}\,(y_1, \cdots, y_k \mid \alpha_1, \cdots, \alpha_k) \\ &\qquad = \frac{\alpha_j}{\alpha} \texttt{Dir}\left(z_1, \cdots, z_k \mid \alpha_1^{(j)}, \cdots, \alpha_k^{(j)}\right). \end{aligned} \tag{2.5}$$

It should be noted that this formula is true even if $\alpha_j = 0$.

5. Let α_1, α_2 be two measures on $\mathbb{X}$ and P_1, P_2 be two independent k-dimensional Dirichlet random vectors with parameters α_1, α_2. If Y independent of P_1, P_2 is distributed as beta $(\alpha_1(\mathbb{X}), \alpha_2(\mathbb{X}))$, then $YP_1 + (1-Y)P_2$ is $\texttt{Dir}\,(\alpha_1 + \alpha_2)$.

   To see this, let $Z_1, Z_2, \ldots, Z_k$ be independent random variables with $Z_i \sim \texttt{Gamma}(\alpha_1\{i\})$. Similarly for $i = 1, 2, \ldots k$ let $Z_{k+i} \sim \texttt{Gamma}\,(\alpha_2\{i\})$ be independent gamma random variables. Then

$$\frac{\sum_1^k Z_i}{\sum_1^{2k} Z_i}\left(\frac{Z_1}{\sum_1^k Z_i}, \ldots, \frac{Z_k}{\sum_1^k Z_i}\right) + \frac{\sum_{k+1}^k Z_i}{\sum_1^{2k} Z_i}\left(\frac{Z_{k+1}}{\sum_1^k Z_i}, \ldots, \frac{Z_{2k}}{\sum_1^k Z_i}\right)$$

   has the same distribution as $YP_1 + (1-Y)P_2$. But then the last expression is equal to

$$\left(\frac{Z_1 + Z_{k+1}}{\sum_1^k Z_i}, \ldots, \frac{Z_k + Z_{2k}}{\sum_1^k Z_i}\right)$$

   which is distributed as $\texttt{Dir}\,(\alpha_1 + \alpha_2)$. Note that the assertion remains valid even if some of the $\alpha_1\{i\}, \alpha_2\{j\}$ are zero. An interesting consequence is: If P is $\texttt{Dir}(\alpha)$ and Y is independent of P and distributed as $\mathrm{Beta}(c, \alpha(\mathbb{X}))$, then

$$Y\delta_{(1,0,\ldots,0)} + (1-Y)P \sim \texttt{Dir}(\alpha\{1\} + c, \alpha\{2\}, \ldots, \alpha\{k\})$$

This follows if we think of $\delta_{1,0,\ldots,0}$ as Dirichlet with parameter ($c, 0, \ldots, 0$). A corresponding statement holds if $(1, 0, \ldots, 0)$ is replaced by any vector with a 1 at one coordinate and 0 at the other coordinates.

6. Let P be distributed as $\texttt{Dir}(\alpha)$ and X independent of P be distributed as $\bar{\alpha}$. Let Y be independent of X and P, and be a $\texttt{Beta}(1, \alpha(\mathbb{X}))$ random variable. Then $Y\delta_X + (1-Y)P$ is again a $\texttt{Dir}(\alpha)$ random probability.

2.3 Dirichlet Processes

We discuss Dirichlet Process from two viewpoints – the first one is more of a measure-theoretic construction/definition that arises through Kolmogorov consistency conditions for random probabilities, and the second one is more Stochastic-Process oriented derivation based on normalized random measures derived through Gamma Processes.

2.3.1 Definition and properties motivated by Kolmogorov consistency conditions

Let $\mathbb{X}$ be a set and let $\mathscr{X}$ be a σ-field of subsets of $\mathbb{X}$. Define a random probability, P, on $(\mathbb{X}, \mathscr{X})$ by defining the joint distribution of the random variables ($P(A_1), \cdots, P(A_m)$) for every m and every sequence $A_1, \cdots, A_m$ of measurable sets ($A_i \in \mathscr{X}$ for all i). Historically speaking, the idea of Dirichlet processes naturally emerged in the attempt to prove the existence of stochastic process P by verifying the Kolmogorov consistency conditions [Kolmogorov, 2018]. We refer the reader to appendix A for more details on this, and proceed to give the definition of Dirichlet Processes that emerges out of the necessity to satisfy Condition 1 in appendix A and utilizing Property 1 of Dirichlet distributions discussed in the last section.

Definition 2. *Let α be a non-null finite measure (nonnegative and finitely additive) on $(\mathbb{X}, \mathscr{X})$. We say P is a Dirichlet process on $(\mathbb{X}, \mathscr{X})$ with parameter α if for every $k = 1, 2, \cdots$, and measurable partition $(B_1, \cdots, B_k)$ of $\mathbb{X}$, the distribution of $(P(B_1), \ldots, P(B_k))$ is Dirichlet, $\texttt{Dir}(\alpha(B_1), \ldots, \alpha(B_k))$.*

Remark 2. *Note that definition 2 only uses a measure α to define a Dirichlet Process, however the same definition can be given based on two parameters, a probability measure, $F_0(\cdot) = \frac{\alpha(\cdot)}{\alpha(\mathbb{X})}$ (also known as the base measure of the Dirichlet Process) and scalar parameter, $\alpha(\mathbb{X})$ (also known as the concentration parameter of the Dirichlet Process), with corresponding Dirichlet process denoted as $\texttt{DP}(\alpha(\mathbb{X}), F_0)$. We will heavily use this second notation in the next chapter, but mostly limit ourself to the measure based notation in this chapter.*

Next we discuss some important properties of the random probability measure P when it is a Dirichlet Process.

Property 1 (mean random-measure of the Dirichlet Process)**.** *Let P be a Dirichlet process on $(\mathbb{X}, \mathscr{X})$ with parameter α, and let $A \in \mathbb{X}$. If $\alpha(A) = 0$, then $P(A) = 0$ with probability one. If $\alpha(A) > 0$, then $P(A) > 0$ with probability one. Furthermore, $\mathbb{E}P(A) = \alpha(A)/\alpha(\mathbb{X})$.*

Proof. By considering the partition (A, A^c) of $\mathbb{X}$, one can see from definition 2 that $P(A)$ has a Beta distribution, $\texttt{Beta}\left(\alpha(A), \alpha\left(A^c\right)\right)$. The proof follows immediately. □

Now, this property may seem to suggest that α and P are mutually absolutely continuous. This interpretation is intuitive, but false; in fact, we will discuss in next section that P is essentially a discrete distribution. Thus, α and P may be mutually singular. However, α and P satisfy a weaker notion than absolute continuity, and we discuss that next

Property 2 (σ-additivity of P w.r.t α)**.** *Let P be a Dirichlet process on $(\mathbb{X}, \mathscr{X})$ with parameter α. If α is σ-additive, then so is P in the sense that for a fixed decreasing sequence of measurable sets $A_n \searrow \varnothing$, we have $P\left(A_n\right) \to 0$ with probability one.*

Proof. Since $A_n \searrow \varnothing$ and α is additive, $\alpha\left(A_n\right) \to 0$. Hence there exists a subsequence $\{n_j\}$ such that $\sum_1^\infty \alpha\left(A_{n_j}\right) < \infty$. For fixed $\varepsilon > 0$,

$$\sum_1^\infty \mathbb{P}\left\{P\left(A_{n_j}\right) > \varepsilon\right\} \leqq \sum_1^\infty \varepsilon^{-1}\mathbb{E}P\left(A_{n_j}\right) = \varepsilon^{-1}\sum_1^\infty \alpha\left(A_{n_j}\right)/\alpha(\mathbb{X}) < \infty$$

Hence, from the Borel-Cantelli lemma, $\mathbb{P}\left\{P\left(A_{n_j}\right) > \varepsilon \text{ i.o. }\right\} = 0$. This proves that $P\left(A_{n_j}\right) \to 0$ with probability one. The proof is completed by noting that $P\left(A_n\right) > P\left(A_{n+1}\right)$ with probability

one for all n, and hence, $P(A_1) > P(A_2) > \ldots$ with probability one. The converse is also true: if α is not σ-additive, then with probability one P is not σ-additive.

□

Since DPs are distributions over distributions, the support of DPs is set of distributions, and a natural and crucial question arises – what kind of distributions are DPs supported on? This is the question we address next, directly stating the result mathematically, and commenting on it subsequently.

Property 3. *[Support of the Dirichlet Process priors] Let P be a Dirichlet process on $(\mathbb{X}, \mathscr{X})$ with parameter α, and let Q be a fixed probability measure on $(\mathbb{X}, \mathscr{X})$ with $Q \ll \alpha$. Then, for any positive integer m and measurable sets $A_1, \cdots, A_m$, and $\varepsilon > 0$,*

$$\mathbb{P}\left\{|P(A_i) - Q(A_i)| < \varepsilon \text{ for } i = 1, \cdots, m\right\} > 0$$

Proof. Form $B_{\nu_1,\cdots,\nu_m}$ as in A.1 and note that

$$\mathbb{P}\left\{|P(A_i) - Q(A_i)| < \varepsilon \text{ for } i = 1, \cdots, m\right\} \tag{2.6}$$

$$\geqq \mathbb{P}\left\{\sum_{(\nu_1,\cdots,\nu_m)\partial_i=1} |P(B_{\nu_1,\cdots,\nu_m}) - Q(B_{\nu_1,\cdots,\nu_m})| < \varepsilon \text{ for } i = 1, \ldots, m\right\} \tag{2.7}$$

Therefore, it is sufficient to show

$$\mathbb{P}\left\{|P(B_{\nu_1,\cdots,\nu_m}) - Q(B_{\nu_1,\cdots,\nu_m})| < 2^{-m}\varepsilon \text{ for all } (\nu_1, \cdots, \nu_m)\right\} > 0$$

If $\alpha(B_{\nu_1,\ldots,\nu_m}) = 0$, then $Q(B_{\nu_1,\ldots,\nu_m}) = 0$ and $P(B_{\nu_1,\ldots,\nu_m}) = 0$ with probability one, so that $|P(B_{\nu_1,\cdots,\nu_m}) - Q(B_{\nu_1,\cdots,\nu_m})| = 0$ with probability one. For those $(\nu_1, \cdots, \nu_m)$ for which $\alpha(B_{\nu_1,\cdots,\nu_m}) > 0$, the distribution of the corresponding $P(B_{\nu_1,\cdots,\nu_m})$ gives positive weight to all open sets in the set

$$\sum_{(\nu_1,\cdots,\nu_m)\ni\alpha\left(B_{\nu_1},\cdots,\nu_m\right)>0} P(B_{\nu_1,\cdots,\nu_m}) = 1$$

completing the proof. □

Property 3 essentially states that in the weak topology (discussed in Section 2.1), the support of the Dirichlet process on $(\mathbb{X}, \mathscr{X})$ with parameter α contains the set of all probability measures absolutely continuous with respect to α. Conversely, any measure Q not absolutely continuous with respect to α is not in the support of P. This property is the first desirable property for nonparametric priors (huge support) mentioned in Section 2.1
We can state this more intuitively with an example of a Dirichlet Process with σ-additive α

Property 4 (Support of DP with σ-additive α)**.** *If $(\mathbb{X}, \mathscr{X})$ is the real line with the Borel sets, and if α is σ-additive, the support of P is the set of all σ-additive probability measures whose support is contained in the support of α.*

Having discussed some preliminary (but important) properties of DP priors, we now move onto discussing the nature of DP posteriors, i.e. answering the question – when utilized for statistical inference, how do sample from true probability distribution, update DP priors. To answer that, we need some definitions and some simple propositions

Definition 3. *Let P be a random probability measure on $(\mathbb{X}, \mathscr{X})$. We say that $X_1, \ldots, X_n$ is a sample of size n from P if for any $m = 1, 2, \ldots$ and measurable sets $A_1, \cdots, A_m, C_1, \cdots, C_n$,*

$$\mathbb{P}\left\{X_1 \in C_1, \cdots, X_n \in C_n \mid P(A_1), \cdots, P(A_m), P(C_1), \cdots, P(C_n)\right\} \tag{2.8}$$

$$= \prod_{j=1}^{n} P(C_j) \quad \text{a.s.} \tag{2.9}$$

Roughly, $X_1, \ldots, X_n$ is a sample of size n from P, if, given $P(C_1), \ldots, P(C_n)$, the events $\{X_1 \in C_1\}, \cdots, \{X_n \in C_n\}$ are independent of the rest of the process, and are independent among themselves, with $\mathbb{P}\{X_j \in C_j \mid P(C_1), \ldots, P(C_n)\} = P(C_j)$ a.s. for $j = 1, \ldots, n$. This definition determines the joint distribution of $X_1, \ldots, X_n, P(A_1), \cdots, P(A_m)$, once the

distribution of the process is given, since

$$P\left\{X_1 \in C_1, \cdots, X_n \in C_n, P\left(A_1\right) \leqq y_1, \cdots, P\left(A_m\right) \leqq y_m\right\} \tag{2.10}$$

may be found by integrating 2.8 with respect to the joint distribution of $P\left(A_1\right), \ldots, P\left(A_m\right), P\left(C_1\right), \cdots, P\left(C_n\right)$ over the set $[0, y_1] \times \cdots \times [0, y_m] \times [0, 1] \times \cdots \times [0, 1]$.

Proposition 1. *Let P be a Dirichlet process on $(\mathbb{X}, \mathscr{X})$ with parameter α and let X be a sample of size 1 from P. Then for $A \in \mathscr{X}$,*

$$\mathbb{P}(X \in A) = \alpha(A)/\alpha(\mathbb{X})$$

Proof. Since $\mathbb{P}(X \in A \mid P(A)) = P(A)$ a.s.,

$$\mathbb{P}(X \in A) = \mathbb{E}\mathbb{P}(X \in A \mid P(A)) \tag{2.11}$$

$$= \mathbb{E}P(A) \tag{2.12}$$

$$= \alpha(A)/\alpha(\mathbb{X}) \tag{2.13}$$

completing the proof.

□

Proposition 2. *Let P be a Dirichlet process on $(\mathbb{X}, \mathscr{X})$ with parameter α, and let X be a sample of size 1 from P. Let $(B_1, \cdots, B_k)$ be a measurable partition of $\mathbb{X}$, and let $A \in \mathscr{X}$. Then,*

$$\mathbb{P}\left\{X \in A, P\left(B_1\right) \leqq y_1, \cdots, P\left(B_k\right) \leqq y_k\right\} = \sum_{j=1}^{k} \frac{\alpha\left(B_j \cap A\right)}{\alpha\left(\mathbb{X}\right)} \mathtt{Dir}\left(y_1, \cdots, y_k \mid \alpha_1^{(j)}, \ldots, \alpha_k^{(j)}\right) \tag{2.14}$$

where $\mathtt{Dir}(y_1, \cdots, y_k \mid \alpha_1, \cdots, \alpha_k)$ is the distribution function of the Dirichlet distribution, $(\alpha_1, \cdots, \alpha_k)$, and where

$$\alpha_i^{(j)} = \alpha(B_i) \qquad \textit{if} \quad i \neq j \tag{2.15}$$

$$= \alpha(B_j) + 1 \qquad \textit{if} \quad i = j \tag{2.16}$$

Proof. Define $B_{j,1} = B_j \cap A$, and $B_{j,0} = B_j \cap A^c$ for $j = 1, \ldots, k$. Let $Y_{j,\nu} = P(B_{j,\nu})$ for $j = 1, \cdots, k$ and $\nu = 0, 1$. Then, from 2.8

$$\mathbb{P}\{X \in A \mid Y_{j,\nu}, j = 1, \cdots, k, \text{ and } \nu = 0, 1\} = \sum_{j=1}^{k} Y_{j,1} \quad \text{a.s.} \tag{2.17}$$

Hence for arbitrary $y_{j,\nu} \in [0, 1]$, for $j = 1, \ldots, k$ and $\nu = 0, 1$,

$$\mathbb{P}\{X \in A, Y_{j,\nu} \leqq y_{j,\nu} \text{ for } j = 1, \ldots, k, \text{ and } \nu = 0, 1\}$$

can be found by integrating 2.17 with respect to the distribution of the $Y_{j,\nu}$ over the set $\{Y_{j,\nu} \leqq y_{j,\nu}, j = 1, \cdots, k \text{ and } \nu = 0, 1\}$. Using property 4 of Dirichlet distribution in the last section, this integration turns out to be,

$$\sum_{j=1}^{k} \frac{\alpha(B_{j,1})}{\alpha(\mathbb{X})} \mathtt{Dir}\left(\mathbf{y} \mid \boldsymbol{\alpha}^{(j)}\right)$$

where $\mathbf{y} = (y_{1,0}, \cdots, y_{k,0}, y_{1,1}, \cdots, y_{k,1})$ and $\boldsymbol{\alpha}^{(j)} = \left(\alpha_{1,0}^{(j)}, \cdots, \alpha_{k,0}^{(j)}, \alpha_{1,1}^{(j)}, \cdots, \alpha_{k,1}^{(j)}\right)$, and where

$$\alpha_{i,\nu}^{(j)} = \alpha(B_{i,\nu}) \qquad \text{if} \quad i \neq j$$

$$= \alpha(B_{j,\nu}) + 1 \quad \text{if} \quad i = j$$

□

The conclusion of the proposition follows from this using property 1 of the Dirichlet distribution discussed in last section, since $P(B_j) = Y_{j,0} + Y_{j,1}$ a.s., and since the process of finding marginal distributions of random variables is linear.

We are now prepared to find the conditional distribution of a Dirichlet process P, given a sample $X_1, \cdots, X_n$ from P. It turns out that this conditional distribution is also a Dirichlet process.

Property 5. *[Conjugacy of Dirichlet Process posteriors] Let P be a Dirichlet process on ($\mathbb{X}, \mathscr{X}$) with parameter α, and let $X_1, \cdots, X_n$ be a sample of size n from P. Then the conditional distribution of P given $X_1, \cdots, X_n$, is a.s. a Dirichlet process with parameter $\alpha + \sum_1^n \delta_{X_i}$.*

Proof. It is sufficient to prove the theorem for $n = 1$, since the theorem would then follow by induction upon repeated application of the case $n = 1$. Let $(B_1, \cdots, B_k)$ be a measurable partition of $\mathbb{X}$ and let $A \in \mathscr{X}$. It is easy to check that the marginal distributions of a conditional distribution of a process are identical to the conditional distributions of the marginals. Hence, we must show that the conditional distribution of $P(B_1), \cdots, P(B_k)$ given X, a sample of size one from P, has distribution function

$$\texttt{Dir}\left(y_1, \cdots, y_k \mid \alpha(B_1) + \delta_X(B_1), \cdots, \alpha(B_k) + \delta_X(B_k)\right) \tag{2.18}$$

This may be done by showing that the integral of 2.18 with respect to the marginal distribution of X over A is equal to the probability 2.14. Using the marginal distribution of X as found in Proposition 1, we compute

$$\int_A \texttt{Dir}\left(y_1, \cdots, y_k \mid \alpha(B_1) + \delta_x(B_1), \cdots, \alpha(B_k) + \delta_x(B_k)\right) d\alpha(x)/\alpha(\mathscr{X}) \tag{2.19}$$

$$= \sum_{j=1}^{k} \int_{B_j \cap A} \texttt{Dir}\left(y_1, \cdots, y_k \mid \alpha_1^{(j)}, \cdots, \alpha_k^{(j)}\right) d\alpha(x)/\alpha(\mathscr{X}) \tag{2.20}$$

$$= \sum_{j=1}^{k} \frac{\alpha(B_j \cap A)}{\alpha(\mathbb{X})} \texttt{Dir}\left(y_1, \cdots, y_k \mid \alpha_1^{(j)}, \cdots, \alpha_k^{(j)}\right) \tag{2.21}$$

completing the proof. □

Property 5 is a crucial result, specifically it also forms the second desirable property for non-parametric priors stated in Sec 2.1 – the posterior distribution given a sample of observations should be manageable analytically and easy to sample from. The second aspect of this property (easy to sample from) is the subject of the last section, but for now we see from Property 5 that

posteriors of Dirichlet Porcess priors are themselves Dirichlet processes. This property is also known as *conjugacy*.

2.3.2 Derivation through Gamma processes based completely random measures

We consider an alternate construction of Dirichlet Processes in this section and the generalized notion of *completely random measures* that this construction leads to eventually. The basic idea is that since the Dirichlet distribution is definable, as in 2.3 of Section 2.2, as the joint distribution of a set of independent gamma variables divided by the sum, so also should the Dirichlet process be definable as a Gamma process with independent "increments" divided by the sum.
Essentially, using a representation of a process with independent increments as a sum of a countable number of jumps of random height at a countable number of random points, as found in [Ferguson and Klass, 1972], we may divide by the total heights of the jumps and obtain a discrete probability measure, which should be distributed as a Dirichlet process. As one can guess, an immediate consequence of this construction is that it shows implicitly that random measures sampled from a DP are essentially a.s. discrete. Next, we discuss this construction begining with the characteristic function of gamma distribution, $\texttt{Gamma}(\alpha, 1), \alpha > 0$,

$$\varphi(a) = (1 - it)^{-\alpha} \tag{2.22}$$

$$= \exp \int_0^\infty \left(e^{iux} - 1\right) dN(x) \tag{2.23}$$

where

$$N(x) = -\alpha \int_x^\infty e^{-y} y^{-1} dy \quad \text{for } 0 < x < \infty \tag{2.24}$$

We define the distribution of random variables $J_1, J_2, \cdots$ as follows.

$$\mathbb{P}\left(J_1 \leqq x_1\right) = e^{N(x_1)} \quad \text{for } x_1 > 0 \tag{2.25}$$

and for $j = 2, 3, \cdots$

$$\mathbb{P}\left(J_j \leqq x_j \mid J_{j-1} = x_{j-1}, \cdots, J_1 = x_1\right) = \exp\left[N\left(x_j\right) - N\left(x_{j-1}\right)\right] \tag{2.26}$$

for $0 < x_j < x_{j-1}$.

In other words, the distribution function of J_1 is $\exp N\left(x_1\right)$ and for $j = 2, 3, \ldots$, the distribution of J_j given $J_{j-1}, \cdots, J_1$, is the same as the distribution of J_1 truncated above at J_{j-1}. Next, we present a useful theorem for rest of the exposition concerning Gamma Processes based derivation of DP random measures.

Theorem 2 ([Ferguson and Klass, 1972]). *Let $G(t)$ be a distribution function on $[0, 1]$. Let*

$$Z_t = \sum_{j=1}^{\infty} J_j I_{[0, G(t))}\left(U_j\right) \tag{2.27}$$

where (i) the distribution of $J_1, J_2, \cdots$ is given in (3) and (4), and
(ii) $U_1, U_2 \cdots$ are independent identically distributed variables, uniformly distributed on $[0, 1]$, and independent of $J_1, J_2, \cdots$. Then, with probability one, Z_t converges for all $t \in [0, 1]$ and is a gamma process with independent increments, with $Z_t \in \texttt{Gamma}(\alpha G(t), 1)$.

In particular, $Z_1 = \sum_1^{\infty} J_j$ converges with probability one and $Z_1 \in \texttt{Gamma}(\alpha, 1)$. If we define

$$P_j = J_j / Z_1 \tag{2.28}$$

then $P_j \geqq 0$ and $\sum_1^{\infty} P_j = 1$ with probability one.

We now define a discrete random measure with weights P_j as in Eq. 2.28. As before, let ($\mathbb{X}, \mathscr{X}$) be a measurable space, and let $\alpha(\cdot)$ be a finite non-null measure on $\mathbb{X}$. Let $V_1, V_2, \ldots$ be a sequence of i.i.d. random variables with values in $\mathbb{X}$, and with probability measure Q, where $Q(A) = \alpha(A)/\alpha\left(\mathbb{X}\right)$.

$$P(A) = \sum_{j=1}^{\infty} P_j \delta_{V_j}(A) \tag{2.29}$$

Theorem 3. *The random probability measure defined by 2.29 is a Dirichlet process on* $(\mathbb{X}, \mathscr{X})$ *with parameter* α*.*

Proof. Let $(B_1, \cdots, B_k)$ be a measurable partition of $\mathbb{X}$. Then

$$(P(B_1), \cdots, P(B_k)) = \frac{1}{Z_1} \sum_{j=1}^{\infty} J_j (\delta_{Vj}(B_1), \cdots, \delta_{Vj}(B_k))$$

From the assumption on the distribution of $V_1, V_2, \cdots$,

$$\mathbf{M}_j = \left(\delta_{V_j}(B_1), \cdots, \delta_{V_j}(B_k)\right)$$

are independent identically distributed random vectors having a multinomial distribution with probability vector $(Q(B_1), \cdots, Q(B_k))$. Hence, the distribution of $\sum_1^{\infty} J_j \mathbf{M}_j$ must be the same as the distribution of

$$\left(Z_{1/k}, Z_{2/k} - Z_{1/k}, \cdots, Z_1 - Z_{(k-1)/k}\right)$$

where Z_t is the gamma process defined by (5) with $G(t)$ chosen so that

$$G\left(\frac{j}{k}\right) - G\left(\frac{j-1}{k}\right) = Q(B_j) \quad j = 1, \ldots, k$$

Hence, $\sum_{j=1}^{\infty} J_j \delta_{V_j}(B_i)$ are, for $i = 1, \cdots, k$, independent random variables, with $\sum_{j=1}^{\infty} J_j \delta_{V_j}(B_i) \in \texttt{Gamma}(\alpha(B_i), 1)$. Since Z_1 is the sum of these independent gamma variables, $(P(B_1), \ldots, P(B_k)) \in \texttt{Dir}(\alpha(B_1), \ldots, \alpha(B_k))$ from the definition of the Dirichlet distribution. Thus, P satisfies the definition of the Dirichlet process.

□

The random measures derived through independent increment processes (Levy-Flights), for example using Gamma distributions above in 2.27 , are referred to as Completely random measures (CRMs) because the partitions they induce are independent. CRMs were first introduced in an excellent paper [Kingman, 1967], and can be defined as follows,

Definition 4. *Let $\tilde{\mu}$ be a measurable mapping from ($\Omega, \mathscr{F}, \mathbb{P}$) into ($\mathcal{M}_{\mathbb{X}}, \mathscr{M}_{\mathbb{X}}$) and such that for any $A_1, \ldots, A_n$ in $\mathscr{X}$, with $A_i \cap A_j = \varnothing$ for any $i \neq j$, the random variables $\tilde{\mu}(A_1), \ldots, \tilde{\mu}(A_n)$ are mutually independent. Then $\tilde{\mu}$ is termed completely random measure (CRM).*

As is evident from DP case in 2.27, an important property of CRMs is their **almost sure discreteness** [Kingman, 1967], which means that their realizations are discrete measures with probability 1 . A CRM on $\mathbb{X}$ can always be represented as the sum of two components: a completely random measure $\tilde{\mu}_c = \sum_{i=1}^{\infty} J_i \delta_{X_i}$, where both the positive jumps J_i 's and the $\mathbb{X}$-valued locations X_i 's are random, and a measure with random masses at fixed locations. Accordingly

$$\tilde{\mu} = \tilde{\mu}_c + \sum_{i=1}^{M} V_i \delta_{x_i} \tag{2.30}$$

where the fixed jump points $x_1, \ldots, x_M$, with $M \in \{1, 2, \ldots\} \cup \{\infty\}$, are in $\mathbb{X}$, the (non-negative) random jumps $V_1, \ldots, V_M$ are mutually independent and they are independent from $\tilde{\mu}_c$. Finally, $\tilde{\mu}_c$ is characterized by the Lévy-Khintchine representation which states that

$$\mathbb{E}\left[\mathrm{e}^{-\int_{\mathbb{X}} f(x)\tilde{\mu}_c(\,\mathrm{d}x)}\right] = \exp\left\{-\int_{\mathbb{R}^+\times\mathbb{X}} \left[1-\mathrm{e}^{-sf(x)}\right]\nu(\mathrm{d}s,\,\mathrm{d}x)\right\} \tag{2.31}$$

where $f : \mathbb{X} \to \mathbb{R}$ is a measurable function such that $\int |f| \mathrm{d}\tilde{\mu}_c < \infty$ (almost surely) and ν is a measure on $\mathbb{R}^+ \times \mathbb{X}$ such that

$$\int_B \int_{\mathbb{R}^+} \min\{s, 1\}\nu(\mathrm{d}s,\,\mathrm{d}x) < \infty$$

for any B in $\mathscr{X}$. The measure ν characterizing $\tilde{\mu}_c$ is referred to as the Lévy intensity of $\tilde{\mu}_c$: it contains all the information about the distributions of the jumps and locations of $\tilde{\mu}_c$. ν can be expressed by separating the location and jump parts as follows,

$$\nu(\mathrm{d}s,\,\mathrm{d}x) = \rho_x(\,\mathrm{d}s)\alpha(\mathrm{d}x) \tag{2.32}$$

where α is a measure on $(\mathbb{X}, \mathscr{X})$ and ρ a transition kernel on $\mathbb{X} \times \mathscr{B}(\mathbb{R}^+)$, i.e. $x \mapsto \rho_x(A)$ is $\mathscr{X}-$ measurable for any A in $\mathscr{B}(\mathbb{R}^+)$and ρ_x is a measure on $(\mathbb{R}^+, \mathscr{B}(\mathbb{R}^+))$for any x in $\mathbb{X}$. If $\rho_x = \rho$ for

any x, then the distribution of the jumps of $\tilde{\mu}_c$ is independent of their location and both ν and $\tilde{\mu}_c$ are termed homogeneous. Otherwise, ν and $\tilde{\mu}_c$ are termed non-homogeneous.

CRMs provide a generalized recipe to construct Bayesian nonparametric priors beyond Dirichlet Processes [Lijoi et al., 2010]. By normalizing CRMs (e.g. In 2.29 for DPs) one obtains what are known as Normalized random measures of independent increments (NRMIs) processes. These are characterized by different alpha-stable distributions and their mixtures. In particular, one can define a ν and obtain a corresponding NRMI. Clearly, for the Dirichlet Process above, this ν was,

$$\nu(\mathrm{d}s,\ \mathrm{d}x) = \frac{\mathrm{e}^{-s}}{s}\ \mathrm{d}s\alpha(\ \mathrm{d}x)$$

Many other priors such as Beta-Stacey, dependent-Dirichlet processes Lijoi et al. [2010] have been constructed utilizing CRMs/NRMIs. They have also served useful for constructing other random-objects such as sparse-graphs [Caron and Fox, 2017]

2.4 Moments and tails of distributions sampled from a DP

The Gamma-process based discrete representation of random measures sampled from a Dirichlet processes can be quite useful to study some of their properties. A natural property of probability measures is their moments and tails, and we next discuss this aspect of DP random measures.

Property 6. *(Moments of DP random measures) Let P be the Dirichlet process defined by Eq. 2.29, and let Z be a measurable real valued function defined on ($\mathbb{X}, \mathscr{X}$). If $\int |Z| d\alpha < \infty$, then $\int |Z| dP < \infty$ with probability one, and*

$$\mathbb{E}\int ZdP = \int Zd\mathbb{E}P = \alpha(\mathbb{X})^{-1}\int Zd\alpha$$

Proof. From Eq. 2.29,

$$\int |Z| dP = \sum_{j=1}^{\infty} |Z\left(V_j\right)|\, P_j$$

so that the monotone convergence theorem gives,

$$\mathbb{E}\int |Z|dP = \sum_{j=1}^{\infty} \mathbb{E}\left|Z\left(V_j\right)\right| P_j \tag{2.33}$$

$$= \alpha(\mathbb{X})^{-1} \int |Z|d\alpha \sum_{j=1}^{\infty} \mathbb{E}P_j \tag{2.34}$$

$$= \alpha(\mathbb{X})^{-1} \int |Z|d\alpha \tag{2.35}$$

where we have used the independence of the V_j and the P_j. Therefore, $\int |Z|dP$ is finite with probability one, and hence

$$\int ZdP = \sum_{j=1}^{\infty} Z\left(V_j\right) P_j$$

is absolutely convergent with probability one. Since this series is bounded by (8), which is integrable, the bounded convergence theorem implies

$$\mathbb{E}\int ZdP = \sum_{j=1}^{\infty} \mathbb{E}Z\left(V_j\right) \mathbb{E}P_j \tag{2.36}$$

$$= \alpha(\mathbb{X})^{-1} \int Zd\alpha \tag{2.37}$$

□

Property 6 emphasizes the close relationship between α and the random probability measure P. Essentially, It implies that if α has a finite k th moment, then, on-average, with probability one P has a finite k th moment. However, a stronger result can be derived,

Property 7. *[Doss and Sellke, 1982][Tails of random measures ampled from a DP] Tails of random measures sampled from a DP prior with measure-parameter α, and corresponding DP posterior are much smaller then the tail of the base measure, G, of the DP prior where $G(\cdot) = \frac{\alpha(\cdot)}{\alpha(\mathbb{X})}$*

The proof is again based on the Gamma process construction of the DP and on some properties of alpha-stable random variables, and we state the main mathematical result. As in Property 7, Let G represent the base-measure of the DP prior with parameter α,and $M = \alpha(\mathbb{X})$, and let F

be an arbitrary random measure sampled from the DP prior (or posterior) then it holds, for any $r > 1$, that

$$\exp\left(-\frac{r\log|\log M(1-G(x))|}{M(1-G(x))}\right) \leq 1 - F(x) \leq \exp\left(-\frac{1}{M(1-G(x))|\log M(1-G(x))|^r}\right) \tag{2.38}$$

The best way to illustrate Property 7 is for a Cauchy distribution (which has no mean). The upper tail of the standard Cauchy distribution G satisfies $1 - G(x) \sim \pi^{-1}x^{-1}$, for $x \to \infty$. Since r is arbitrary, we can absorb any constant in powers of $\log x$. Inequalities 2.38 imply, for fixed M as $x \to \infty$,

$$\exp\left[-r\pi M^{-1}x\log\log x\right] \leq 1 - F(x) \leq \exp\left[-x/|\log x|^r\right]$$

Thus the tails of F are almost exponential, even though G has no mean. This tail-property of random measures sampled from DP priors/posteriors will be very useful in proving regret bounds for the algorithm presented in the next chapter.

Now we turn to one of the important properties desirable of nonparametric priors discussed in Sec 2.1, i.e. Is it easy to sample from Dirichlet Processes?

2.5 Sethuraman construction of the Dirichlet Process

We saw in Section 2.3.2 the almost surely discrete nature of the random measures distributed according to a Dirichlet Process. Sethuraman [1994] introduced an elegant construction for these random measures, and this will be quite useful in the next chapters as well, specifically the distributional equation 2.40. We discuss this construction next, and prove that random measure following Sethuraman construction is distributed as a DP.

Let α be a finite measure and $\bar{\alpha}(\cdot) = \frac{\alpha(\cdot)}{\alpha(\mathbb{X})}$, and let $\theta_1, \theta_2, ...$ be $\texttt{Beta}(1, \alpha(\mathbb{X}))$ random variables and $Y_1, Y_2, \ldots$ are i.i.d. $\bar{\alpha}$ and independent of the $\theta_i s$ Set $p_1 = \theta_1$ and for $n \geq 2$, let $p_n = \theta_n \prod_1^{n-1}(1-\theta_i)$. Easy computation shows that $\sum_1^\infty p_n = 1$ almost surely. Now define a $\mathcal{P}_\mathbb{X}$

valued random measure by,

$$P(\cdot) = \sum_{1}^{\infty} p_n \delta_{Y_n(}(\cdot) \tag{2.39}$$

Since $\sum_1^\infty p_n = 1$, $P(\cdot)$ takes values in $\mathcal{P}_{\mathbb{X}}$, and is a random discrete measure that puts weight p_i on Y_i.

Property 8. *[Sethuraman representation of DP random measures] The random measure, $P(\cdot)$, in Eq. 2.39, is distributed as a DP with parameter measure α.*

Proof. A natural approach to show this is to resort to definition of DPs as in definition 2 and then showing that for every partition $B_1, B_2, \ldots, B_k$ of $\mathbb{X}$ by Borel sets, $(P(B_1), P(B_2), \ldots, P(B_k))$ is distributed as $\texttt{Dir}(\alpha(B_1), \ldots, \alpha(B_k))$. Denote by $\delta^k_{Y_i}$ the element of S_k given by $(I_{B_1}(Y_i), I_{B_2}(Y_i), \ldots, I_{B_k}(Y_i))$. Then ,$(P(B_1), P(B_2), \ldots, P(B_k))$ can be written as $\sum_1^\infty p_i \delta^k_{Y_i}$. Let P be an S_k valued random variable, independent of the Ys and θs, and distributed as $\texttt{Dir}(\alpha(B_1), \ldots, \alpha(B_k))$. Now, consider the S_k valued random variable

$$P^1 = p_1 \delta^k_{Y_1} + (1 - p_1) P \tag{2.40}$$

where $Y_1 \in B_i, \delta^k_{Y_i}$ is the vector with a 1 in the ith coordinate and 0 elsewhere. Hence by property 5 from Section 2.2, given $Y_1 \in B_i, P^1$ is distributed as a Dirichlet with parameter $(\alpha(B_1), \ldots, \alpha(B_i) + 1, \ldots, \alpha(B_k))$. Since $\mathbb{P}(Y_1 \in B_i) = \alpha(B_i)$, by property 6 in Section 2.2, P^1 is distributed as $\texttt{Dir}(\alpha(B_1), \ldots, \alpha(B_k))$.

It follows by easy induction that for all $n, 1 - \sum_1^n p_i = \prod_1^n (1 - \theta_i)$. Using this fact, a bit of algebra gives

$$\sum_{1}^{n} p_i \delta^k_{Y_i} + \left(1 - \sum_{1}^{n} p_i\right) P \tag{2.41}$$

$$= \sum_{1}^{n-1} p_i \delta^k_{Y_i} + \left(1 - \sum_{1}^{n-1} p_i\right) \left(\theta_n \delta^k_{Y_n} + (1 - \theta_n) P\right) \tag{2.42}$$

Since our earlier argument showed that $\theta_n \delta^k_{Y_n} + (1-\theta_n) P$ has the same distribution as P, a simple induction argument shows that, for all n,

$$\sum_1^n p_i \delta^k_{Y_i} + \left(1 - \sum_1^n p_i\right) P \tag{2.43}$$

is distributed as $\texttt{Dir}(\alpha(B_1), \ldots, \alpha(B_k))$. Letting $n \to \infty$ and observing that $(1 - \sum_1^n p_i)$ goes to 0 , we get the result.

□

Chapter 3

Dirichlet process posterior sampling for multi-arm bandits

In this chapter, we introduce Dirichlet Process Posterior Sampling (DPPS), a Bayesian nonparametric 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 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. Our contributions can be summarized more concretely as follows,

- 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. 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 [Osband et al., 2018]

- 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. 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.

- We extend the information theoretic analysis of Thompson sampling in Russo and Van Roy [2016] 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. 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.

This chapter is based on the published article, [Vashishtha and Maillard, 2025], and the corresponding Python code for all the figures in this paper can be found at this web-link. We begin by laying down the motivation for DPPS and a colloquial discussion about it.

3.1 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) based algorithms [Auer et al., 2002] and Thompson sampling (TS) [Thompson, 1933, Russo et al., 2018] 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 [Abbasi-Yadkori and Szepesvári, 2011, Auer et al., 2002]. 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 [Filippi et al., 2010]. Next, although Thompson-Sampling (TS) [Thompson, 1933, Kaufmann et al., 2012] 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 [Phan et al., 2019].

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 soon finds the answer in Statistical Bootstrap [Efron, 1992, Efron and Tibshirani, 1994] which is a procedure for estimating the distribution of an estimator by resampling (often with replacement) 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 [Kveton et al., 2019, Baransi et al., 2014, Osband and Van Roy, 2015]. 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 [Osband and Van Roy, 2015]. Additionally, these bootstrap sampling based algorithms cannot account for uncertainty that does not come from the observed data [Osband et al., 2016]. 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 [Srinivas et al., 2009], 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) [Rasmussen, 2003]. 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 [Castillo, 2024, Müller et al., 2015, Ghosal, 2010], . 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 3.6 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 the so-called non parametric Thompson sampling introduced in Riou and Honda [2020]. We discuss this in Section 3.5, 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 3.6. Additionally, in Section 3.7, 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 7 which provides an upper bound on the Bayesian regret of DPPS.

3.2 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 $R_{t,A_t} \in \chi$. In this paper, we choose χ to the set of σ-sub-Gaussian random variables, i.e. $\mathbb{E}\left[e^{(X-\mathbb{E}[X])t}\right] \leq e^{\frac{\sigma^2 t^2}{2}}$, $\forall X \in \chi$, 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 $\chi^{|\mathcal{A}|}$ 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

$$\mathtt{KL}(P||Q) = \int P \log\left(\frac{dP}{dQ}\right) dP \tag{3.1}$$

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 $\mathbb{H}(p) = -\sum_{x\in\mathcal{X}} p(x)\log\left(p(x)\right)$. 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

$$I_t(X_1; X_2) := \mathtt{KL}\left(\mathbb{P}\left((X_1, X_2) \in \cdot|\mathcal{H}_t\right) \;||\; \mathbb{P}\left(X_1 \in \cdot|\mathcal{H}_t\right)\mathbb{P}\left(X_2 \in \cdot|\mathcal{H}_t\right)\right), \tag{3.2}$$

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 $a_t = \arg\max_{a\in\{1,\ldots,K\}}\{\mu(\theta_{a,t})\}$, where $\mu(\theta_{a,t})$ represents the mean of the parametric reward distributions with parameter $\theta_{a,t}$.

3.3 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. 3.12; 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 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 3.4. 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.

[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.

Algorithm 1 Dirichlet Process Posterior Sampling

Require: Horizon T, number of arms K, arm parameters – Distribution $F_{0,k}$, constant $\alpha_{0,k}$ for $k \in \{1, ..., K\}$

for $k = 1...K$, **do**
  Set $\mathbf{R}_k = [\,]$, $F_k = F_{0,k}$, $\alpha_k = \alpha_{0,k}$, and $N_k = 0$
end for
for $t = 1...T$, **do**
  # Sample model (a random measure):
  for $k = 1...K$, **do**
    Sample $D_k \sim \texttt{DP}(\alpha_k, F_k)$
  end for
  # select and apply action:
  $I(t) = \text{argmax}_{k \in \{1,...,K\}} \{\mu(D_k)\}$
  Pull arm $I(t)$ and observe reward $R_{t,I(t)}$
  Update history $\mathbf{R}_{I(t)} = (\mathbf{R}_{I(t)}^\top, R_{t,I(t)})^\top$ and count $N_{I(t)} \leftarrow N_{I(t)} + 1$.
  # Posterior update
  $\alpha_{I(t)} \leftarrow \alpha_{I(t)} + 1$
  $F_{I(t)} = \frac{1}{\alpha_{I(t)}} \sum_{x \in \mathbf{R}_I(t)} \delta_x + \frac{\alpha_{0,I(t)}}{\alpha_{I(t)}} F_{0,I(t)}$
end for

3.4 Sampling random measures from Dirichlet Process priors and posteriors

Note that DPPS involves sampling random measures from DP Priors and DP Posteriors. Natural questions arise here : How to actually sample (finite-representations of) these random-measures from DP priors/posteriors? What do they look like? How to use them for statistical-inference problems? We answer these questions in this section, by introducing finite stick-breaking construction of random measures sampled from a DP prior. This is based on the Sethuraman construction discussed in the Section 2.5. Following this, we derive a new distributional equality for the random measures sampled from Dirichlet Process posteriors, and then exhibit their use for statistical inference on a real-world data-set. We begin by revisiting Dirichlet Process posteriors.

3.4.1 Dirichlet Process posteriors

Let X_1, ... , 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, 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. As also discussed in Section 2.3, 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 [Ferguson, 1973, Ghosal, 2010],

$$\alpha_n = \alpha + n,\ \overline{F}_n = \frac{n}{\alpha + n} F_n + \frac{\alpha}{\alpha + n} F_0 \tag{3.3}$$

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. 3.3 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.

3.4.2 Bayesian Bootstrap

A very useful idea in statistical inference has been that of Statistical Bootstrap [Efron and Tibshirani, 1994], and a Bayesian version of Bootstrap was introduced in [Rubin, 1981]. Interestingly, this Bayesian version of Bootstrap can also be derived as a special case of the DP posteriors [Ghosal and van der Vaart, 2017]. 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,

$$\mathrm{BB_n} := \texttt{DP}(n, \sum_{i=1}^{n} \delta_{X_i}) = \sum_{i=1}^{n} W_i \delta_{X_i} \tag{3.4}$$

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.,

$$\mu(BB_n) = \sum_{i=1}^{n} W_i X_i = \langle \mathbf{W}_n, \mathbf{X}_n \rangle \tag{3.5}$$

As we shall see in Sec 3.3, the idea of Bayesian Bootstrap forms the basis for a bandit algorithm introduced in [Riou and Honda, 2020]. Next we discuss an important representation of DP priors/posteriors that make them amenable to practical applications.

3.4.3 Finite stick-breaking construction of DP prior random measures

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.3.3 provide little information to answer this question. The Sethuraman construction discussed in Section 2.5 also known as Stick Breaking representation of DPs, provides an answer to this question. In general, these measures can be generalized and are known as Stick-breaking measures [Ishwaran and James, 2001], and are almost surely discrete random probability measures that can be represented as,

$$Q(\cdot) = \sum_{i=1}^{N} q_i \delta_{Z_i}(\cdot) \tag{3.6}$$

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. As discussed in Section 2.5, if these weights, q_i, are constructed such that,

$$q_1 = V_1,\ (q_i)_{i=2}^{N-1} = V_i \prod_{j=1}^{i-1}(1 - V_j),\ q_N = \prod_{i=1}^{N}(1 - V_i) \tag{3.7}$$

$$V_i \overset{iid}{\sim} \texttt{Beta}(1, \alpha),\ Z_i \overset{iid}{\sim} F,\ i = 1, 2, ...N \tag{3.8}$$

and N is ∞, then the generated random discrete measure, P, in Eq.3.7 (with N as ∞) is such that, $P \sim \texttt{DP}(\alpha, F)$. 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 [Ishwaran and James, 2001, Muliere and Tardella, 1998] thanks to its provable optimality in closely approximating the infinite series [Ishwaran and James, 2001]. Particularly, to quantify the accuracy loss owing to truncation [Ishwaran and James, 2001], consider the quantities, $T_N = (\sum_N^\infty q_i)^r$ and $U_N = \sum_N^\infty q_i^r$, where N is the level at which the representation is truncated,

$$\mathbb{E}(T_N(r, a, b)) = \left(\frac{\alpha}{\alpha + r}\right)^{N-1}, \tag{3.9}$$

$$\mathbb{E}(U_N(r, a, b)) = \left(\frac{\alpha}{\alpha + r}\right)^{N-1} \frac{\Gamma(r)\Gamma(\alpha + 1)}{\Gamma(\alpha + r)} \tag{3.10}$$

Notice that both expressions decay exponentially fast in N, and hence good accuracy is achieved for moderate K. Fig. 3.1 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.

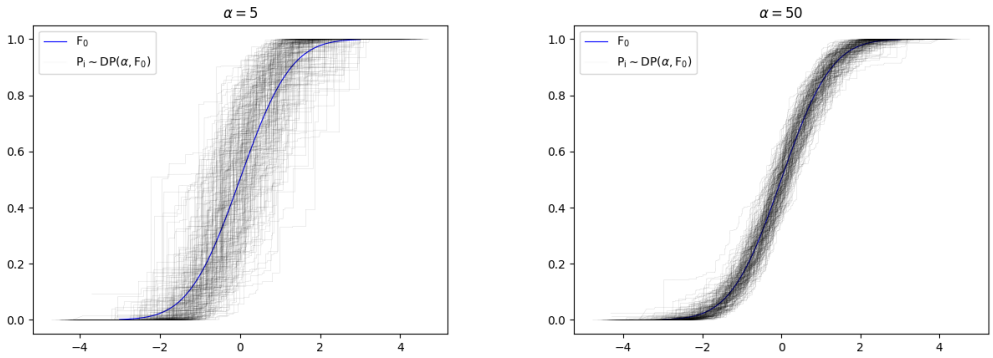

Figure 3.1: 200 random measures sampled from $\mathrm{DP}(\alpha, \mathrm{F}_0)$ where $\alpha = 5$ (left) and 50 (right), $\mathrm{F}_0 = N(0, 1)$

3.4.4 Sampling from DP posteriors with a new distributional equation

With the finite stick-breaking representation of DP priors at hand, one comes to wonder how to sample from DP posteriors (i.e. $\texttt{DP}(\alpha_n, \overline{F}_n)$) in a practically feasible way and, for this, we utilize the distributional equation, Eq. 2.40, associated to the DPs, which we restate here with an iterative flavor,

$$Q_i(\cdot) \stackrel{d}{=} V_i \delta_{X_{i-1}} + (1 - V_i) Q_{i-1}(\cdot) \tag{3.11}$$

Thus, if Q_{i-1} is a random measure sampled from $\texttt{DP}(\alpha_{i-1}, F_{i-1})$, then Q_i is a random measure sampled from DP posterior with one observation X_{i-1}, i.e. $Q_i \sim \texttt{DP}(\alpha_i, F_i)$, where from Eq. 3.3, $\alpha_i = \alpha_{i-1} + 1$, $F_i = \frac{1}{\alpha_{i-1}+1}\delta_{X_{i-1}} + \frac{\alpha_{i-1}}{\alpha_{i-1}+1} F_{i-1}$, and correspondingly to utilize Eq. 3.11, $V_i \sim \texttt{Beta}(1, \alpha_{i-1} + 1)$. Beginning with a random measure sampled from the DP prior, $Q_0 \sim \texttt{DP}(\alpha_0, F_0)$, generated using the finite stick-breaking method Eqs. 3.7-3.8, we apply the recursion in Eq.3.11 n times (corresponding to n observations $\{X_1, ..., X_n\}$) to obtain a random measure, Q_n, sampled from the DP posterior with n observations (i.e. $\texttt{DP}(\alpha_n, \overline{F}_n)$),

$$Q_n \stackrel{d}{=} V_n \delta_{X_n} + \sum_{i=1}^{n-1} \left[V_i \prod_{j=i+1}^{n} (1 - V_j) \right] \delta_{X_i} + \left[\prod_{i=1}^{n} (1 - V_i) \right] Q_0. \tag{3.12}$$

With this new distributional equation we can readily sample random measures from from $\texttt{DP}(\alpha_n, \overline{F}_n)$. In order to give more intuition to appreciate the utility of DP random measures for Bayesian nonparametric inference, we give an example of inference on a real world data-set. We also used this (and some other) benchmarks to validate the performance of our StickBreaking Python module available at this web-link

3.4.5 Statistical inference with DPs on a real-world data-set

We illustrate the application of Dirichlet processes for density estimation on a data set from the astronomy literature [Roeder, 1990]. The measurements are velocities at which galaxies in the Corona-Borealis region are moving away from our galaxy. If the galaxies are clustered,

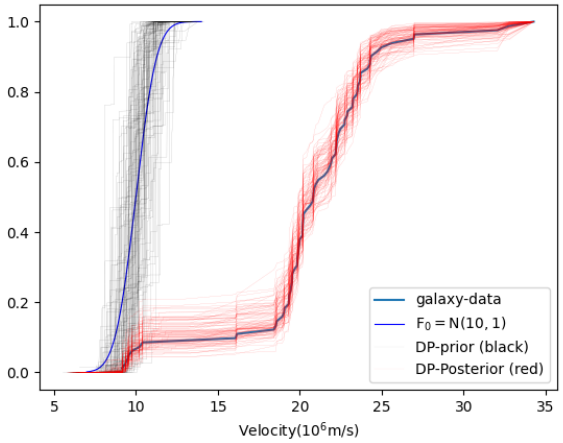

Figure 3.2: DP prior and DP posterior compared against original galaxy dataset distribution.

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 3.2 where the left and right regions of the CDF are almost flat, and most mass resides in the center. We estimate the CDF of this density using DP priors. Starting with a $\mathrm{DP}(\alpha = 2, \mathrm{F}_0 = \mathrm{N}(10, 1))$ DP prior, we obtain a DP posterior, and the spread of distributions sampled from the DP posterior can be seen as confidence-region of the true CDF underlying the galaxy-data.

3.5 Noninformative limit of the DPPS

In Riou and Honda [2020], 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. 3.5), and Bayesian-Bootstrap is a special case of Dirichlet-processes

(see Eq. 3.4). 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 Sections 3.4.2 and 3.4.1 on Bayesian Bootstrap and DP posteriors).

Algorithm 2 Algorithm in Riou and Honda [2020]

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

3.6 Numerical experiments

With the procedure of sampling from DP priors and Posteriors established in section 3.4, we are now ready to run DPPS and exhibit its empirical performance 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 [Agrawal and Goyal, 2013] 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. 3.6.4. Corresponding code can be found at this web-link. All simulations were generated using the Python 3 language, and made use of the NumPy [Harris et al., 2020], SciPy [Virtanen et al., 2020], and Matplotlib [Hunter, 2007] libraries.

3.6.1 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. 3.3 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. 3.3 (right) also shows performance of DPPS in this setting, and clearly DPPS outperforms other baselines.

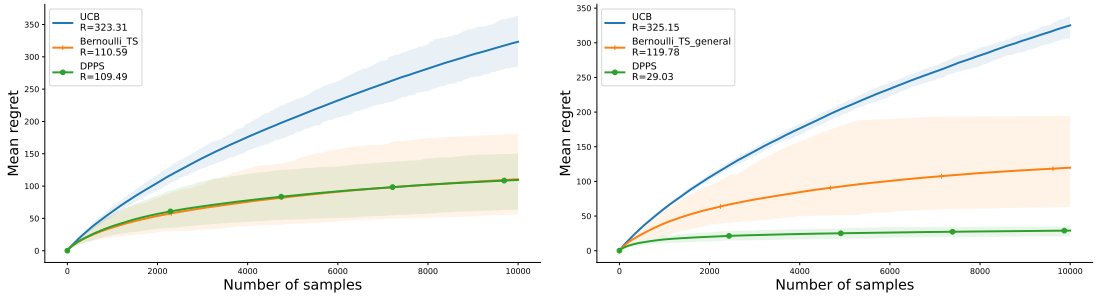

Figure 3.3: Comparison of average regret in the Bernoulli bandit setting (left), and Beta Bandit setting (right) discussed in the text.

3.6.2 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 [Hoogenboom et al., 2019, Gautron et al., 2022]. 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 3.4, 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 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.

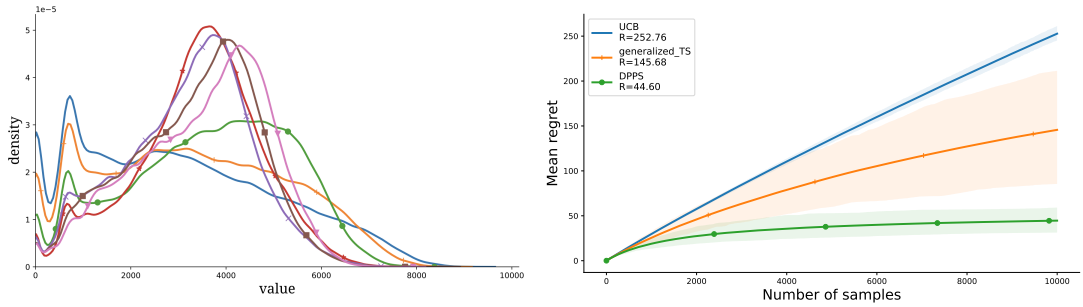

Figure 3.4: Reward distributions from DSSAT simulator (left) and regret performances of bandit strategies (right) in the DSSAT environment.

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 Agrawal and Goyal [2013], 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.3.4 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.

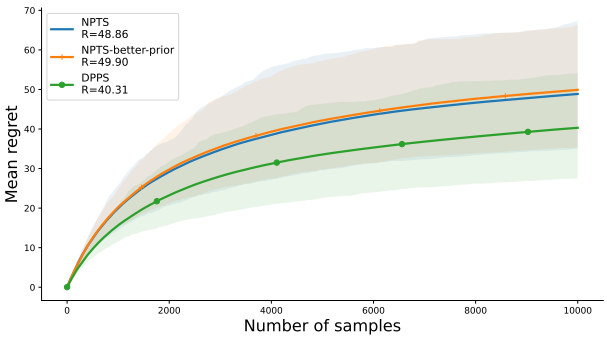

Figure 3.5: Average regret in the DSSAT bandit environment with beneficial priors for both NPTS and DPPS.

3.6.3 Incorporating prior knowledge through DPPS

Recall from Sec. 3.5 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 *pseudo-rewards/artificial-history* 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 pseudo-rewards of $X_k = 0.01$ for all but the third arm (for which it uses a value of 1). Fig. 3.5 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.

3.6.4 Gaussian bandit with unknown means and variances

A challenging bandit setting is that of Gaussian bandit environment with both mean and variance of the underlying Gaussian distribution as unknown [Cowan et al., 2018] 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

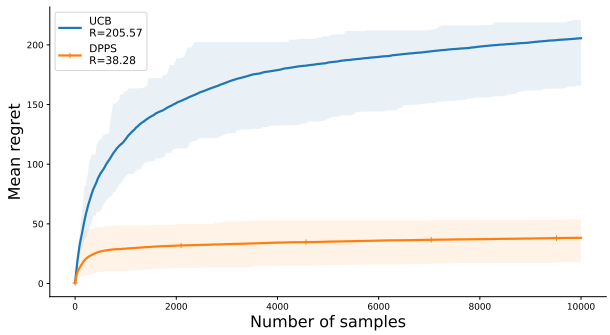

Figure 3.6: DPPS for a challenging Gaussian bandit setting

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. 3.6. Excellent performance of DPPS is evident. In this experiment, we chose $\alpha = 2$, base measure of DP, F_0, as $N(0, 0.5)$.

3.6.5 Choice of DP parameters in numerical experiments

Two parameters in DPPS are α (concentration parameter) and k_t (i.e. truncation level) in the stick breaking representation of DP prior (not the posterior), DP(α, 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 ($V_i \sim$ Beta$(1, \alpha)$), and for higher α one needs to increase k_t.

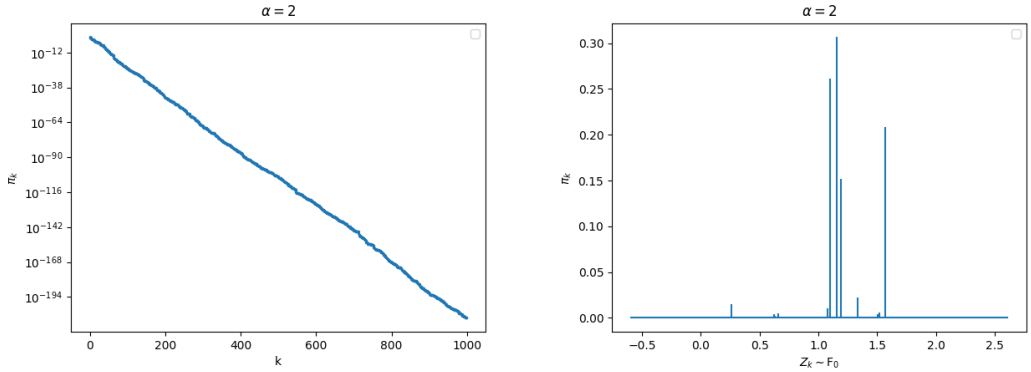

Figure 3.7: Plot of first 1000 stick-breaking probability measure weights, π_k, for $\mathrm{DP}(\alpha = 2, F_0)$ with k (left) and with $Z_k \sim F_0 (= N(0, 1)$ (right)

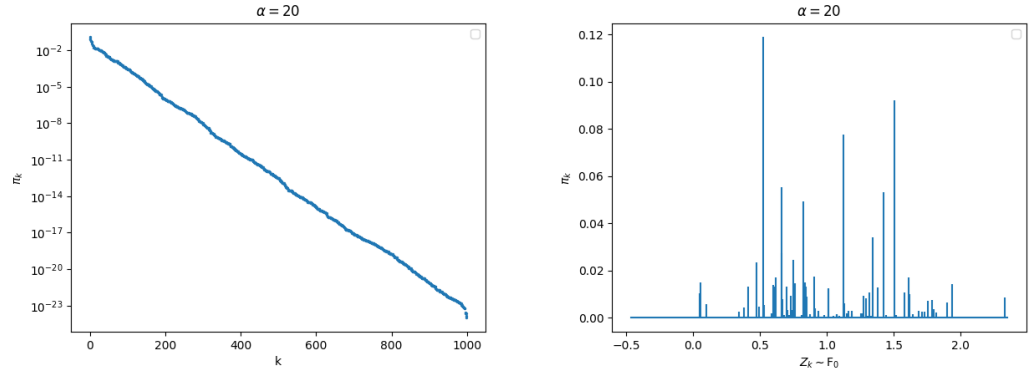

Figure 3.8: Plot of first 1000 stick-breaking probability measure weights, π_k, for $\mathrm{DP}(\alpha = 20, F_0)$ with k (left) and with $Z_k \sim F_0 (= N(0, 1)$ (right)

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. 3.7 and 3.8. 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 discussed in Section 2.3 is important, and we restate it here with notation consistent for this chapter,

Lemma 1. *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 7 on Bayesian regret of DPPS, shows that choosing F_0 with σ-subGaussian tails corresponding to tails of the reward noise guarantees order optimal regret bounds.

3.6.6 Computational costs 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. 3.6.6 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* [Ghosal, 2010] 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.

Next, 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,

$$Q_n \quad = \quad V_n \delta_{X_n} + \sum_{i=1}^{n-1} \left[V_i \prod_{j=i+1}^{n} (1 - V_j) \right] \delta_{X_i} + \left[\prod_{i=1}^{n} (1 - V_i) \right] Q_0 \tag{3.13}$$

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.

3.7 Information theoretic analysis of DPPS

In Section 3.3, we saw that, like Thompson sampling, DPPS is a probability matching algorithm. In this section, we utilize this property of DPPS to derive upper bounds on its Bayesian-regret. Particularly, we generalize the information theoretic analysis of Thompson sampling introduced in Russo and Van Roy [2016] to a wider class of probability matching algorithms, and show that DPPS is encapsulated in that generalization. We begin by summarizing the key-steps in the analysis of Russo and Van Roy [2016] and also include complete proofs for the sake of completion in Sec. B.

3.7.1 Information theoretic analysis of Thompson sampling

Russo and Van Roy [2016] introduced an elegant framework for deriving upper bounds on Bayesian regret of Thompson sampling. 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*,

Theorem 4. *[Russo and Van Roy, 2016] 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 then boils down to bounding the information-ratio of that algorithm. Particularly, for Thompson-sampling, in σ-sub-Gaussian reward noise bandit setting, one obtains the following bound,

Lemma 2. *[Russo and Van Roy, 2016]*

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

This bound when combined with Theorem 4 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 5. *[Russo and Van Roy, 2016] For a K arms bandit problem with σ sub-Gaussian reward-distributions of the arms, the Bayesian regret after T rounds is bounded as,*

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

The proof of Lemma 2 hinges on two crucial steps, and we highlight those, referring the reader to Sec. B for other 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,

$$\begin{aligned}\mathbb{E}_t\left[R_{t,A^\star} - R_{t,A_t}\right] &= \sum_{a\in\mathcal{A}} \mathbb{P}_t(A^\star = a)\mathbb{E}_t\left[R_{t,a}|A^\star = a\right] - \sum_{a\in\mathcal{A}} \mathbb{P}_t(A_t = a)\mathbb{E}_t[R_{t,a}|A_t = a] \\ &= \sum_{a\in\mathcal{A}} \mathbb{P}_t(A^\star = a)\left(\mathbb{E}_t\left[R_{t,a}|A^\star = a\right] - \mathbb{E}_t[R_{t,a}]\right).\end{aligned} \tag{3.14}$$

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 [Cover, 1999] for the KL divergence, $\mathtt{KL}(P||Q)$, between two absolutely continuous measures, P and Q,

Fact 1. *[Cover, 1999]*

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

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

Lemma 3. *[Russo and Van Roy, 2016]*

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

Substituting the result of Lemma 3 in Eq. 3.14, and utilizing the definition of information gain, $I_t(A^\star; R_{t,a})$, and information ratio, Γ_t, yields the bound in Lemma 2.

3.7.2 A generalization of the information theoretic analysis

Choice of $\mathbb{P}_t(A^\star = a)$ It's easy to notice in the preceding analysis (specifically Eq. 3.14) that in the analysis of Russo and Van Roy [2016] there's explicitly no restriction on $\mathbb{P}_t(A^\star = a)$ for it to be derived using a Bayes-rule based posterior-distributions of arm-rewards, $\mathbb{P}_t(R_{t,a})$, as is done in parametric Thompson sampling. However, this choice is rather implicit, given the decision theoretic and information theoretic *coherency* of Bayesian framework [Wald, 1961, Zellner, 1988]. Moreover, Bayesian-framework is not limited to Bayes-rule based derivation of posterior distributions. Another *valid* Bayesian approach [Orbanz, 2009, Ghosal and van der Vaart, 2017] for obtaining posteriors is leveraging the property of *conjugacy* as discussed in Sec 3.4. In particular, most nonparametric priors do not satisfy the necessary conditions for Bayes rule [Orbanz, 2009], and one relies on their conjugacy property to derive the corresponding posteriors. An elegant way to characterize the class of valid Bayesian approaches is based on martingales of probability measures, and we refer the interested reader to papers concerning *predictive-Bayes* [Holmes and Walker, 2023, Fong et al., 2023, Fortini and Petrone, 2025] for details.

Admissible probability matching algorithms From the preceding discussion, we can conclude that all probability matching algorithms which derive $\mathbb{P}_t(R_{t,a})$ (and hence $\mathbb{P}_t(A^* = a)$) using a valid Bayesian approach are *admissible* in the information theoretic analysis of Russo and Van Roy [2016]. Additionally, these admissible algorithms would enjoy same bounds on their information-ratio and (consequently) Bayesian regret as those for parametric Thompson sampling (i.e. Lemma 2 and Theorem 5), if they satisfy certain *auxiliary conditions* required from the analysis. For the case of σ-sub Gaussian reward noise discussed before, we list these next.

Auxiliary conditions It is easy to see that we require the following auxiliary conditions for an admissible probability matching algorithm to enjoy same bounds on Bayesian regret as those for parametric Thompson sampling in [Russo and Van Roy, 2016]: In each round t,

1. The instantaneous reward noise, $X(t)$, in Lemma 3, is a σ-sub-Gaussian random variable,
2. $\texttt{KL}(\mathbb{P}_t(R_{t,a}|A^\star = a)||\mathbb{P}_t(R_{t,a}))$ in Lemma 3 is well defined.

The second condition holds if $\mathbb{P}_t(A^\star = a) > 0$ owing to a classical fact in conditional probability,

Fact 2. *[Williams, 1991] 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 $\mathbb{P}_t(Y)$.*

Next, we show that DPPS is an admissible probability matching algorithm that satisfies the auxiliary conditions listed above.

3.7.3 Bayesian regret of DPPS

That DPPS is an admissible probability matching algorithm is easy to see: DPPS utilizes a valid Bayesian approach, i.e. conjugacy of DP priors/posteriors, to derive $\mathbb{P}_t(A^\star = a)$; Also DPPS satisfies both the auxiliary conditions discussed in previous sub-section; For condition (2), clearly, $\mathbb{P}_t(A^\star = a) > 0$ for DPPS, whenever the base measure, F_0, of the DP prior (and hence of the corresponding DP posterior), $\texttt{DP}(\alpha, F_0)$, is non-null. For condition (1), the following property of the tail of random measures sampled from DP priors/posteriors (also see Section 2.4 for a more detailed discussion on this) ensures σ-sub-Gaussian nature of the instantaneous reward noise, $X(t)$, whenever the base measure, F_0, of the DP prior, $\texttt{DP}(\alpha, F_0)$, is σ-sub-Gaussian,

Theorem 6. *[Doss and Sellke [1982]] Let $F \sim \texttt{DP}(\alpha, F_0)$, then almost surely the tails of F and distributions sampled from the DP posterior of F, $\texttt{DP}(\alpha + n, \overline{F_n})$, given samples $X_1, ..., X_n$, are dominated by (and are much smaller than) the tails of F_0.*

Therefore, DPPS is an admissible probability matching algorithm that satisfies the necessary auxiliary conditions under specific constraints on the DP priors, and hence whenever those constraints on the DP priors are satisfied, DPPS enjoys the same upper-bounds on Bayesian regret as those for parametric Thompson sampling. More concretely,

Theorem 7. *For the setting of σ-sub-Gaussian arm reward distributions, starting with a DP-prior, $DP(\alpha, F_0)$, with F_0 as a σ sub-Gaussian distribution, the Bayesian regret of DPPS satisfies,*

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

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

3.8 More related works

To the best of our knowledge, Dirichlet Processes in the context of bandits were first used in [Clayton and Berry, 1985] 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 Eckles and Kaptein [2014], which was further improved and made more systematic in [Osband and Van Roy, 2015] 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 generating artificial history), and generalizes this equivalence. Non-Parametric Thompson sampling (NPTS) and Multinomial Thompson Sampling (TS) were introduced in [Riou and Honda, 2020] without highlighting any concrete Bayesian connection of the former algorithm. NPTS was adapted for robustness in [Baudry et al., 2021]. Some discussions concerning Bayesian interpretation of NPTS using DPs appeared in Belomestny et al. [2023] who provided a refined analysis of Multinomial TS. In many inference problems, mixture and diffusion priors can be useful to approximate the performance of Bayesian nonparametric priors Müller and Mitra [2013], and a nice line of works performs Thompson sampling utilizing such (parametric) priors [Hong et al., 2022, 2020, Kveton et al., 2024]. Aligning towards non-Bayesian side, a sample mean based algorithm guaranteeing $O(\log N)$ instance-dependent regret appeared in [Agrawal, 1995], a sub-sampling

based algorithm was reported in [Baransi et al., 2014] and analyzed for a two-arm bandit setting; a nonparametric Bootstrap based algorithm was reported in [Kveton et al., 2019], and regret bounds derived for a Bernoulli bandit environment. Finally, we would like to comment here that Information theoretic analysis presented in this paper can also be seen as a special case of well established PAC-Bayes approach, and we refer the reader to an excellent article (and references therein) for details [Alquier, 2024].

3.9 Conclusions and Perspectives

In this chapter, 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 [Phan et al., 2019]. 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 [Pitman and Yor, 2006] processes and Pólya-Tree priors [Castillo, 2017, 2024] can be useful generalizations of DPPS.

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 as well, especially in the model-free scenario, and in Chapter 4 we introduce two new reinforcement learning algorithms based on Dirichlet processes for Markov decision processes.

Chapter 4

Dirichlet process based RL algorithms for Markov decision processes

This chapter introduces new nonparametric reinforcement learning algorithms for Markov decision processes (MDPs). We begin by highlighting the importance of Bayesian neural networks, and the challenges of sampling from the posteriors over neural-networks. Following that, we introduce a new scheme based on Dirichlet Processes for sampling from (Bayesian) posteriors over neural-networks. We utilize this scheme to design a deep reinforcement learning algorithm for Markov decision processes. This algorithm is essentially Thompson sampling algorithm with action-value function represented by a neural network model, and the DP based posterior-sampling scheme utilized for sampling from the posterior over the neural network models. Finally, we introduce a principled Bayesian nonparametric solution to distributional reinforcement learning, and propose a new algorithm that hinges on a crucial distributional equality for the Dirichlet Processes. The final section concludes the thesis highlighting ongoing and future works.

4.1 Bayesian neural networks and uncertainty in model predictions

Deep learning has garnered great attention from researchers in fields such as physics, biology, and manufacturing, to name a few [Baldi et al., 2014, Bergmann et al., 2014, Anjos et al., 2015]. Tools such as feedforward neural networks(NNs), convolutional neural networks, LSTMs, transformers, among others are used extensively. However, plain versions of neural networks are prone to overfitting [Ghahramani, 2015, Gal and Ghahramani, 2016, Krzywinski and Altman, 2013]. When applied to supervised or reinforcement learning problems these networks are also often incapable of correctly assessing the uncertainty in the training data and hence make overly confident decisions about the correct class, prediction or action [Ghahramani, 2015]. This is not good for applications that rely on robustness of the confidence of predictions, e.g. nuclear reactors [Linda et al., 2009], weather-forecasting [Price et al., 2023], etc, and hence representing model uncertainty is of crucial importance. Bayesian neural networks (BNNs) [Neal, 2012, MacKay, 2003] provide a principled approach to achieve this goal. BNNs elicit a prior distribution over neural network parameters θ. Given n observations $z_{1:n}$, one estimates a posterior distribution over θ, which in turn captures uncertainty in predictions. However, exact posterior inference for θ in BNNs is typically intractable, requiring approximations, popular examples being those using variational-inference, MCMC among others [Blundell et al., 2015, Arbel et al., 2023]. The fact that the neural networks are over-parameterized models, and can have millions of parameters in real-world applications makes matters worse. Additionally, it can be challenging to design meaningful priors over the neural network parameter space [Arbel et al., 2023], major reason being that in general it is easier to know some prior information about the generating distribution of the training data, and to transform that into corresponding (often product) parametric priors on millions of parameters θ is not straightforward. To overcome these limitations, we propose a feasible approach for sampling from posteriors over neural network models (or any other statistical model in general) by utilizing the uncertainty about the underlying generative distribution of the data, and treating that distribution as random variable distributed according to a Dirichlet Process (DP) prior.

4.2 a new Dirichlet process based scheme for sampling from posterior distribution over neural-network models

We present a new nonparametric methodology for sampling neural-network models from their posteriors and correspondingly capturing (Bayesian) uncertainty in their predictions. The core underlying idea is that uncertainty in the predictions stems from limited knowledge about the true underlying distribution of the data. This can be understood from a simple example of a neural network with parameters θ and loss $L_\theta(\cdot)$. If one knows the distribution. $G^*(\cdot)$, that generates the training data (say $\{z_i = (x_i, y_i)\}_{i=1}^n$, with x_i the input and y_i the class label belonging to one of the K classes.), one can obtain the parameters θ^* corresponding to G^* as follows,

$$\theta^* = \arg\min_\theta \mathbb{E}_{z\sim G^*}[L_\theta(z)] = \arg\min_\theta \int L_\theta(z) dG^*(z) \tag{4.1}$$

However, in reality we only have finite number of data points, and G^* is of course unknown, and therefore the uncertainty in θ and consequently in the predictions of the (neural-network) model emerge.

Recall from previous chapters that a DP prior, $\texttt{DP}(\alpha, F_0)$ serves as a distribution over the set of distributions whose support lies within the support of F_0 (Theorem 6, Lemma 1, also see Sections 2.3 and 2.4). Correspondingly, we begin by assuming that $G^* \sim \texttt{DP}(\alpha, F_0)$, choosing F_0 such that it encapsulates the support (which we assume to be known) of the true-underlying distribution, and incorporates any other available information about G^*. Next, based on the finite number of data-set, $\{z_i\}_{i=1}^n$, available (training set), we obtain the corresponding DP posterior, $\texttt{DP}(\alpha_n, \overline{F_n})$, where,

$$\alpha_n = \alpha + n,\ \overline{F}_n = \frac{n}{\alpha+n} F_n + \frac{\alpha}{\alpha+n} F_0 \tag{4.2}$$

where F_n is the empirical distribution. After this, we sample a number of i.i.d. random measures, $P_j \sim \texttt{DP}(\alpha_n, \overline{F_n})$, $j = 1, ..., J$ posterior utilizing the finite stick breaking formulation formula for DP-posterior developed in Chapter 3, Eq. 3.12.

$$P_j(\cdot) \stackrel{d}{=} V_n \delta_{z_n}(\cdot) + \sum_{i=1}^{n-1} \left[V_i \prod_{j=i+1}^{n} (1 - V_j) \right] \delta_{z_i}(\cdot) + \left[\prod_{i=1}^{n} (1 - V_i) \right] P_0(\cdot) \tag{4.3}$$

where, $V_i \sim \texttt{Beta}(1, \alpha + i)$, and $P_0 \sim \texttt{DP}(\alpha, F_0)$, and P_0 is again sampled using the finite stick breaking formulation for DP prior (Chapter 3, Eq. 3.7),

$$P_0(\cdot) = \sum_{i=n}^{n+T} p_i \delta_{z_i}(\cdot) \tag{4.4}$$

where T is the level at which the infinite series is truncated, and the weights p_i can be calculated as follows,

$$p_n = V_n,\ (p_i)_{i=n+1}^{n+T-1} = V_i \prod_{j=1}^{i-1} (1 - V_j),\ p_{n+T} = \prod_{i=n}^{n+T} (1 - V_i) \tag{4.5}$$

$$V_i \stackrel{iid}{\sim} \texttt{Beta}(1, \alpha),\ z_i \stackrel{iid}{\sim} F_0,\ i = n, ..., n + T \tag{4.6}$$

Substituting Eq. 4.4 into 4.3 one can write P_j as,

$$P_j(\cdot) = \sum_{i=1}^{n+T} p'_i \delta_{z_i}(\cdot) \tag{4.7}$$

where the first n probability weights, $\{p'_i\}_{i=1}^{n}$ correspond to the training data points, calculated as in Eq. 4.3, and the other weights, $\{p'_i\}_{i=n+1}^{n+T}$, correspond to those generated through DP prior (or base measure), derived as in Eqs. 4.5-4.6 (and multiplied by the multiplicant corresponding to $P_0(\cdot)$ in Eq. 4.3)

For each P_j we can obtain a neural-network model, θ^j, i.e. a sample from the posterior over neural-network models using Eq.4.1. It's easy to see that the discrete representation for P_j in Eq. 4.7 simplifies Eq. 4.1 to,

$$\theta^j = \arg\min_{\theta} \sum_{j=1}^{n+T} p_j L_\theta(z_j) \tag{4.8}$$

The complete algorithm is summarized in Algorithm 3. In order to compute variations in the predictions and confidence intervals, one can compute predictions for each of the sampled model.

Finally, recall from previous chapters, and observe again from Eq 4.7 that a random measure from DP posterior is supported not merely at original data-set points, but also those generated from the DP-prior (or correspondingly from the base measure F_0). When $\alpha \to 0$ one gets Bayesian Bootstrap version and gets random measures supported only at the data points. In other words, this DP based posterior-sampling scheme for neural-networks provides the flexibility to incorporate available prior information about true underlying distribution generating the training data, through the base measure of the DP prior, F_0.

Algorithm 3 DP based scheme for sampling from posterior over neural-network models

1: **Input:** Observations $z_{1:n}$, DP-prior base measure F_0 and concentration parameter α, DP prior truncation parameter T, Number of posterior samples J, neural-network architecture, neural-network loss $L_\theta(\cdot)$
2: **Output:** samples, $\{\theta^{(j)}\}_{j=1}^{J}$, from posterior over neural-network models
3: **Compute:** DP posterior, DP$(\alpha_n, \overline{F_n})$, with observations, $z_{1:n}$
4: **for** $j = 1, ..., J$ **do**
5: sample a random measure, $P_j \sim \texttt{DP}(\alpha_n, \overline{F_n})$, Eq. 4.7
6: compute $\theta^{(j)} = \arg\min_\theta \int L_\theta(z) dP_j(z)$ using Eq. 4.8
7: **end for**

Based on this scheme, we can naturally perform Thompson sampling with neural networks for reinforcement learning .

4.3 Thompson sampling with deep neural networks

Thompson Sampling [Thompson, 1933, Russo et al., 2018, Riquelme et al., 2018] provide an elegant approach that tackles the exploration-exploitation dilemma by maintaining a posterior over models and choosing actions in proportion to the probability that they are optimal. In general

these models are taken to a be a parametric class of distributions, and priors and posteriors defined over the parameter of that class (e.g. sufficient statistic of an exponential family). However, nothing stops one to use other models in principle, and perform Thompson sampling with those models if it is feasible to sample from their posterior-distributions [Riquelme et al., 2018]. In this section, we take Neural networks, instead of the parametric distributions such as exponential families to be the statistical models, given their rich universal function approximation property, and utilize the DP based posterior sampling scheme developed in the previous section for sampling from their posteriors to propose Thompson sampling based deep reinforcement learning algorithm for Markov decision processes. We refer to Thompson sampling with deep neural networks as *deep Thompson-sampling*

4.4 Dirichlet process based deep Thompson sampling for Markov decision processes

We begin by formalizing the setting of Markov decision processes, and application of neural networks for generalization across continuous state-action spaces in reinforcement learning, and then introducing a DP-based deep Thompson sampling algorithm for MDPs.

4.4.1 Markov decision processes

MDPs model stochastic, discrete-time and finite action space control problems [Puterman, 1990]. An MDP is a tuple $M = (\mathcal{X}, \mathcal{A}, R, P, \gamma)$ where $\mathcal{X}$ is the state space, $\mathcal{A}$ the action space, R the reward function, $\gamma \in]0, 1[$ the discount factor and P a stochastic kernel modeling the one-step Markovian dynamics ($P(y \mid s, a)$ is the probability of transitioning to state y by choosing action a in state s). A stochastic policy π maps each state to a distribution over actions $\pi(\cdot \mid s)$ and gives the probability $\pi(a \mid s)$ of choosing action a in state s. The quality of a policy π is assessed by the action-value function Q^π defined as:

$$Q^{\pi}(s,a) := \mathbb{E}Z^{\pi}(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^{t}R\left(s_{t},a_{t}\right)\right] \tag{4.9}$$

where $\mathbb{E}$ is the expectation over the distribution of the admissible trajectories ($s_0, a_0, s_1, a_1, \ldots$) obtained by executing the policy π starting from $s_0 = s$ and $a_0 = a$, and therefore the quantity $Q^{\pi}(s,a)$ represents the expectation of random return $Z^{\pi}(s,a)$: γ-discounted cumulative reward collected by executing the policy π starting from s and a. Fundamental to reinforcement learning is the use of Bellman's equation, easily derived from Eq. 4.9, to describe the value function:

$$Q^{\pi}(s,a) = \mathbb{E}R(s,a) + \gamma\mathbb{E}Q^{\pi}\left(s',a'\right). \tag{4.10}$$

In reinforcement learning we are typically interested in acting so as to maximize the return, and one utilizes the optimality equation corresponding to Eq. 4.10,

$$Q^{*}(s,a) = \mathbb{E}R(s,a) + \gamma\mathbb{E}\max_{a'\in\mathcal{A}}Q^{*}\left(s',a'\right). \tag{4.11}$$

This equation has a unique fixed point Q^*, the optimal value function, corresponding to the set of optimal policies Π^* (π^* is optimal if $\mathbb{E}_{a\sim\pi^*}Q^*(s,a) = \max_a Q^*(s,a)$). Q-learning is the standard algorithm to compute optimal value functions in an iterative manner, with their updates given as follows,

$$Q_{t+1}\left(s_t,a_t\right) = Q_t\left(s_t,a_t\right) + \alpha_t\left(s_t,a_t\right)\Big(r_t + \gamma\max_{a}Q_t\left(s_{t+1},a\right) - Q_t\left(s_t,a_t\right)\Big) \tag{4.12}$$

In this equation, $Q_t(s,a)$ gives the value of the action a in state s at time t. The learning rate $\alpha_t(s,a) \in [0,1]$ ensures that the update averages over possible randomness in the rewards and transitions in order to converge in the limit to the optimal action value function. The discount factor $\gamma \in [0,1)$ has two interpretations. First, it can be seen as a property of the problem that is to be solved, weighing immediate rewards more heavily than later rewards. Second, in non-episodic

tasks, the discount factor makes sure that every action value is finite and therefore well defined. It has been proven that Q-learning reaches the optimal value function Q^* with probability one in the limit under some mild conditions on the learning rates and exploration policy [Puterman, 1990].

4.4.2 deep reinforcement Learning

Deep Reinforcement Learning uses deep neural networks as function approximators for RL methods. Deep Q-Networks (DQN)[Mnih et al., 2015], Dueling architecture [Wang et al., 2016], Trust Region Policy optimization [Schulman et al., 2015] are examples of such algorithms. They frame the RL problem as the minimization of a loss function $L(\theta)$, where θ represents the parameters of the neural network. We will focus specifically on DQN. DQN [Mnih et al., 2015] uses a neural network as an approximator for the action-value function of the optimal policy $Q^\star(x, a)$. DQN's estimate of the optimal action-value function, $Q(x, a)$, is found by minimising the following loss (motivated by Q-learning update rule (Eq.4.12)) with respect to the neural network parameters θ,

$$L(\theta) = \mathbb{E}_{(x,a,r,y)\sim D}\left[\left(r + \gamma \max_{b\in A} Q\left(y, b; \theta^-\right) - Q(x, a; \theta)\right)^2\right] \tag{4.13}$$

where D is a distribution over transitions $e = (x, a, r = R(x, a), y \sim P(\cdot \mid x, a))$ drawn from a replay buffer of previously observed transitions. Here θ^-represents the parameters of a fixed and separate target network which is updated ($\theta^- \leftarrow \theta$) regularly to stabilize the learning. An ϵ-greedy policy is used to pick actions greedily according to the action-value function Q or, with probability ϵ, a random action is taken.

4.4.3 a new deep reinforcement learning algorithm based on Dirichlet Processes

Note that although deep RL algorithms such as DQN have attained superhuman performance [Mnih et al., 2015], they can still fail at simple tasks that require efficient exploration [Osband et al., 2016]. This is because they use statistically inefficient exploration schemes such as the one

stated above for DQN, ϵ-greedy, which approximate the value of an action by a single number, without accounting for the uncertainty in their estimates. Finally, note that theoretical RL literature does offer a variety of provably-efficient approaches to efficient exploration, e.g. UCRL [Auer et al., 2008], but most of these are designed for MDPs with small finite state spaces, while others require solving computationally intractable planning tasks [Brafman and Tennenholtz, 2002], and these algorithms are not practical in complex environments, where an agent must utilize highly nonlinear parametrized models such as neural networks for generalization in the continuous state-action spaces that such environments come with.

Algorithm 4 DP based deep Thompson sampling for DQN

Input: neural network for Q value function, DP prior (DP(α,F_0))
Let B be the replay buffer storing training experiences of the agent
for every episode **do**
    Obtain initial state s_0 from environment
    Sample a Q neural-network model, θ^t, from its posterior using Algorithm 3
    for time $t = 1, .., \tau$ **do**
        Select action $a_t = \arg\max_a Q_t(s_t, a; \theta^t)$
        Observe reward r_t, transition to s_{t+1}
        Add $(s_t, a_t, r_{t+1}, s_{t+1})$ to replay buffer B
    end for
end for

DP based deep Thompson sampling can be utilized to remedy this, because here one can utilize (Bayesian) uncertainty estimates (instead of a point estimate) over the action-value functions (each sample from posterior over neural network models for action-value function would correspond to one estimate) for efficient exploration. We illustrate the algorithmic principle with an example for DQN, with corresponding pseudocode in Algorithm 4. Instead of plain deep neural network, we begin with a Bayesian neural network for representing the action-value function in DP based deep DQN. Like in vanilla DQN, in each episode we sample the training data for the action-value neural network from the replay-buffer B, and sample a posterior for this neural network with this training data using algorithm 3 with the loss for neural network as DQN loss in Eq. 4.13 . The sampled neural network model is then used to make decisions in the entire episode [Osband et al., 2013] as highlighted in algorithm 4.

How to choose the DP priors for the DP based DQN of algorithm 4?

Note that, in the beginning episodes, there are no/very-less observations in the replay buffer, and the contribution of the (synthetic) data generated through DP prior would be significant in making decisions, and since the choices made initially can have significantly varying delayed consequences in RL, a proper choice of the DP prior (correspondingly the base measure of the DP prior) is crucial. A natural way to choose the base measure for the DP prior is to use product base measures with $F_0 = F_{0x}F_{0a}F_{0r}$. F_{0x} and F_{0r} would heavily hinge on the support of the MDP transition kernel $P(\cdot|x, a)$, and reward distribution R respectively – One should choose F_{0x} such that its support encapsulates the support of $P(y|x, a)$ and same holds for F_{0r} wrt R (See Theorem 6 and Lemma 1). For example, if R is known to be σ-subgaussian then one should choose F_{0r} to be a σ_0-subGaussian (e.g. $N(\mu, \sigma_0)$, $\sigma_0 \geq \sigma$, but not, for example, a Beta distribution because Beta has bounded support). The action space may be continuous for the MDP setting, and we should choose F_{0a} based on the nature of the action-space, i.e. categorical distribution with $|\mathcal{A}|$ entries if $|\mathcal{A}|$ is discrete and finite, else choose F_{0a} such that it encapsulates the bounds of the action-space (say bounded between 0 and 1 if on the real line, etc).

4.5 Bayesian distributional RL with Dirichlet processes

Most advances in reinforcement learning have largely revolved around the Bellman equation, Eq. 4.10. However, Bellman equation does have certain caveats. Firstly, the exponential discounting of the form γ^t in Eq. 4.9, that the Bellman equation is build upon, doesn't cohere well with observations in neuroscience, psychology, and economics [Schultheis et al., 2022, Fedus et al., 2019]. Next, approaches such as Q-learning, Eq. 4.12, learn point estimates of value-functions, whereas a growing literature points in the favor of learning a distribution that is trained to minimize a loss against the distribution of data it observes, under the umbrella framework of distributional reinforcement learning [Bellemare et al., 2017a, Dabney et al., 2020, Bellemare et al., 2017b], and the focus shifts to the distributional counterpart of the Bellman equation,

$$Z(s,a) \stackrel{D}{=} R(s,a) + \gamma Z\left(s', a'\right). \tag{4.14}$$

Besides improving stability and performance in deep RL benchmarks, the distribution of returns learned in distributional-RL can also be utilized for risk-sensitive RL [Morimura et al., 2010]. Various algorithms for distributional RL have been proposed [Bellemare et al., 2023], however they tend to rely on dithering strategies such as ϵ greedy for exploration part without being able to utilize the learned distributions for efficient-exploration. This is not surprising because efficient-exploration relies on Bayesian uncertainty in beliefs (epistemic uncertainty) [Kearns and Singh, 2002], whereas distributional RL algorithms focus on distributional-noise/aleatoric uncertainty by construction. Besides note that Eq. 4.14 still relies on exponential version of discounting. Therefore, we seek a generic RL approach that resolves these issues. Essentially, we ask the following question,

Can we design a fundamentally Bayesian framework for RL that,

- performs *natural discounting*, i.e. discounting that emerges as a natural consequence of the framework being Bayesian, rather than being explicitly designed (exponential, hyperbolic, etc), also corroborating growing evidence towards the Bayesian brain hypothesis [Friston, 2012]
- induces a notion of action-value distributions which are learned in an iterative Bayesian (nonparametric) manner, and hence these action-value distributions can be utilized for efficient exploration as well.

To begin with, instead of the distributional Bellman equation Eq. 4.14, consider the distributional equation,

$$Z(s,a) \stackrel{d}{=} V\delta_{r(s,a)} + (1-V)Z(s',a') \tag{4.15}$$

where $Z(s,a)$ represents action-value distribution, and V is a random variable (unlike γ in

Bellman equation which is a fixed constant.) Naturally, multiple possibilities arise as to how to choose V, and we choose it such that the distributional equation above is Bayes plausible. In particular, we choose $V \sim \texttt{Beta}(1, \alpha(s,a))$ which reduces it to the distributional equation corresponding to Dirichlet Process posterior, Eq. 3.11, and hence enables one to learn $Z(s,a)$ in an iterative Bayesian manner, and we discuss an algorithm for this next.

Algorithm 5 Bayesian distributional RL with DPs for tabular MDPs

Input: DP prior ($Z(s,a) \sim \texttt{DP}(\alpha_{s,a}, F_{0s,a})$) $\forall$ s,a pairs
History of rewards $H(s,a) = ()$ for all s, a pairs
for every episode **do**
    Obtain initial state s_0 from environment
    for time $t = 1, .., \tau$ **do**
        Select action $a_t = \arg\max_a \phi_1(Z(s_t,a) \sim \texttt{DP}(\alpha(s_t,a), \overline{F}(s_t,a))$
        Observe reward r_t, transition to s_{t+1}
        $a' = \arg\max_a \phi_2(Z(s_{t+1},a) \sim \texttt{DP}(\alpha(s_{t+1},a), \overline{F}(s_{t+1},a))$
        Set $H(s_t,a_t) = H(s_{t+1},a') = H(s_t,a_t) \cup H(s_{t+1},a')$
    end for
end for

Algorithm 5 exhibits the pseudo-code to learn $Z(s,a)$ in a Bayesian non-parametric manner. Following the choice of V discussed in previous paragraph, the agent essentially begins by assuming the action value distributions, $Z(s,a)$, to be distributed as a Dirichlet Process, $\texttt{DP}(\alpha_0(s,a), F_0(s,a))$. In a state, s_t, the agent takes action, a_t based on a statistical functional ϕ_1 of the random action-value distribution, i.e. $a_t = \arg\max_a \phi_1(Z(s_t,a) \sim \texttt{DP}(\alpha(s_t,a), \overline{F}(s_t,a))$, where $\alpha(s_t,a)$ and $\overline{F}(s_t,a)$ are concentration parameter and base measure of the DP posterior, and are calculated based on the history of rewards, $H(s_t,a)$, and are given as follows,

$$\alpha(s_t,a) = \alpha_0(s_t,a) + \text{length}(H(s_t,a)),\ \overline{F}(s_t,a) = \frac{n}{\alpha+n} F(H(s_t,a)) + \frac{\alpha}{\alpha+n} F_0(s_t,a) \quad (4.16)$$

Here $F(H(s_t,a))$ represents empirical distribution function corresponding to the elements of the rewards history list, $H(s_t,a)$. Following this the agent recieves a reward, $r(s_t,a_t)$, transitions to the next state s_{t+1}, and updates the distributional equation, Eq. 4.15, with $s' = s_{t+1}$, and $a' = \arg\max_a \phi_2(Z(s_{t+1},a) \sim \texttt{DP}(\alpha(s_{t+1},a), \overline{F}(s_{t+1},a))$. Then a crucial step involves synthesizing the histories, $H(s_t,a_t)$, and $H(s_{t+1},a')$, this step is necessary to ensure that the distributional equation, Eq. 4.15, is satisfied. After, a number of episodes the DP posterior for each $Z(s,a)$

contracts sufficiently, and the agent essentially converges to a unique action-value distribution. One can utilize existing results on contraction rates of DP posteriors [Castillo, 2024, Ghosal and van der Vaart, 2017], and/or max-type recursive distributional equations [Aldous and Bandyopadhyay, 2005] to estimate asymptotic/non-asymptotic convergence rates towards these unique action-value distributions.

4.6 Ongoing and future work

Computational analysis of the algorithms proposed in this chapter is ongoing work, and forms the basis for manuscripts to be submitted for peer-review.

Some interesting ideas follow for future work. Firstly, for DPPS in chapter 3, we focused on Bayesian-regret as the performance metric, and given the fact that we have access to random cumulative distribution functions in DPPS, instead of simply mean or variance, we believe that DPPS can be extended to other performance metrics, such as Conditional Value at Risk (CVaR) [Tamar et al., 2015]. Another promising area of research is designing non-parametric best arm identification algorithms based on Dirichlet Processes and their utility in nonparametric hypothesis testing [Gopalan and Berry, 1998, Ferguson, 1973, 1974]. Furthermore, we believe that one can derive refined properties for DP random measures (e.g. their tails, CDF confidence-sequences around the true underlying distribution as exhibited in Fig 3.2) by utilizing the (sub/super)-martingales framework, essentially synthesizing their Gamma process construction with likelihood (super/sub) ratio martingales [Neiswanger and Ramdas, 2021]. Among other things, this can be quite useful for conformal prediction [Barber et al., 2023]

Next, neural network based weather forecasting has shown great success in the recent years Price et al. [2023], and requires efficient scheme for capturing uncertainty in predictions. A promising research direction is to utilize our DP based scheme (Algorithm 3) for capturing this uncertainty in predictions in neural-network based weather forecasting instead of ensemble based techniques that current state of the art models rely on [Price et al., 2023, Lakshminarayanan et al., 2017]. In particular, the DP based scheme can also harness prior information, which can naturally be quite useful for enhancing the accuracy of weather forecasting. One can also fuse together numerical weather prediction based on physics informed neural networks [Raissi et al.,

2019] along with the DP based scheme for uncertainty quantification.

Finally, we would like to highlight some of our new ideas extending the power of DPs to statistical physics/dynamical systems, particularly for the last unsolved mystery of classical physics– "Turbulence" [Frisch, 1995, Falkovich and Sreenivasan, 2006]. A better understanding/modeling of such flows would benefit a wide range of fields, e.g. Astrophysics, Climate Sciences, Nuclear fusion. A remarkable and challenging characteristic of well-developed turbulent flows is their multiscale nature which is sustained through an elegant cascading phenomenon wherein large-scale flow structures (large eddies) break down and transfer their kinetic-energy to smaller ones and so forth until sufficiently small eddies are obtained where this kinetic energy is dissipated into thermal energy by the action of viscosity of the fluid. Furthermore, in an intermediate range of flow scales (known as inertial subrange), a universal statistical scaling of kinetic-energy with flow-scales ($E(k) \sim k^{-5/3}$, where $E(k)$ is the kinetic energy corresponding to flow-scale k (Fourier mode)) emerges [Obukhov, 1941, Kolmogorov, 1968]. This phenomenon of cascade has been studied widely with various models proposed [Mandelbrot, 1974, Frisch, 1995, Meneveau and Sreenivasan, 1987](it even found its place in the "Starry-nights" of Van Gogh [Ma et al., 2024]) but not within a principled statistical framework. We have realized that the phenomenon of cascading can be revisited through a Bayesian nonparametric lens, specifically it is akin to Stick breaking construction of random measures sampled from a DP. Utilizing self-similarity property of DPs, it was straightforward to further realize that turbulence can be studied as an *emergent* phenomenon, emerging out of local interactions in a lattice of particles, where each particle is a Dirichlet Process (i.e. 'particle' now is represented by a set of infinitely many possible random manifestations) and interactions between particles boil down to local sampling operations between each of the neighboring particles. This lattice based framework can be seen as a nonparametric version of interacting particle systems [Liggett and Liggett, 1985] where nonparametric aspect is introduced through the Dirichlet Processes. We believe that development of such a DP based interacting-particle system framework would go on a long way in enhancing our hitherto limited understanding about numerous complex phenomenon in statistical-physics, turbulence being one of them, but others included, e.g. Spin-glasses [Parisi, 2023], with interesting ramifications for neuroscience [Kelso, 1995] and complex systems [Mitchell, 2009] in general.

Appendix A

Kolmogorov consistency conditions for random probabilities

Let $\mathbb{X}$ be a set and let $\mathscr{X}$ be a σ-field of subsets of $\mathbb{X}$. We define below a random probability, P, on $(\mathbb{X}, \mathscr{X})$ by defining the joint distribution of the random variables ($P(A_1), \cdots, P(A_m)$) for every m and every sequence $A_1, \cdots, A_m$ of measurable sets ($A_i \in \mathscr{X}$ for all i). We then verify the Kolmogorov consistency conditions to show there exists a probability, $\mathbb{P}$, on $\left([0,1]^{\mathscr{X}}, B\mathscr{F}^{\mathscr{X}}\right)$ yielding these distributions. Here, $[0,1]^{\mathscr{X}}$ represents the space of all functions from $\mathscr{X}$ into $[0,1]$, and $B\mathscr{F}^{\mathscr{X}}$ represents the σ-field generated by the field of cylinder sets [Kolmogorov, 2018]

Note that, it is more convenient to define the random probability, P, by defining the joint distribution of $(P(B_1), \ldots, P(B_k))$ for all k and all measurable partitions $(B_1, \cdots, B_k)$ of $\mathscr{C}$. (We say $(B_1, \cdots, B_k)$ is a measurable partition of $\mathbb{X}$ if $B_i \in \mathbb{X}$ for all $i, B_i \cap B_j = \varnothing$ for $i \neq j$, and $\bigcup_{j=1}^{k} B_j = \mathbb{X}$.) From these distributions, the joint distribution of $(P(A_1), \ldots, P(A_m))$ for arbitrary measurable sets $A_1, \cdots, A_m$ may be defined using the seives schemes we discuss below [Kolmogorov, 2018]

Given arbitrary measurable sets $A_1, \cdots, A_m$, form the $k = 2^m$ sets obtained by taking intersections of the A_i and their complements; that is, define $B_{\nu_1, \cdots, \nu_m}$ for each $\nu_j = 0$ or 1 as

$$B_{\nu_1,\cdots,\nu_m} = \bigcap_{j=1}^{m} {A_j}^{\nu_j} \tag{A.1}$$

where ${A_j}^1$ is interpreted as A_j, and ${A_j}^0$ is interpreted as ${A_j}^c$, the complement of A_j. Thus the $\{B_{\nu_1,\cdots,\nu_m}\}$ form a measurable partition of $\mathscr{X}$. If we are given the joint distribution of

$$\{P\left(B_{\nu_1,\cdots,\nu_m}\right); \nu_j = 0 \text{ or } 1, j = 1,\cdots,m\} \tag{A.2}$$

then we may define the joint distribution of $\left(P\left(A_1\right),\ldots,P\left(A_m\right)\right)$ to be that obtainable from A.1 upon defining for $i = 1,\cdots,m$

$$P\left(A_i\right) = \sum_{(\nu_1,\cdots,\nu_m)\nu_{\nu_i}=1} P\left(B_{\nu_1,\cdots,\nu_m}\right) \tag{A.3}$$

We note that if $(A_1,\cdots,A_m)$ was a measurable partition to start with, then this does not lead to contradictory definitions of the distribution of $\left(P\left(A_1\right),\ldots,P\left(A_m\right)\right)$ provided

$$P(\varnothing) \text{ is degenerate at } 0. \tag{A.4}$$

Under assumption A.4, the distribution of $\left(P\left(A_1\right),\ldots,P\left(A_m\right)\right)$ for arbitrary measurable $A_1,\cdots,A_m$ is defined uniquely, once the distributions of $\left(P\left(B_1\right),\cdots,P\left(B_k\right)\right)$ are given for arbitrary measurable partitions $\left(B_1,\ldots,B_k\right)$.

If we are given a system of distributions of $\left(P\left(B_1\right),\ldots,P\left(B_k\right)\right)$ for all k and all measurable partitions $\left(B_1,\cdots,B_k\right)$, there is one consistency criterion we would certainly like to have satisfied; namely,

Condition 1. *If $(B_1{}',\cdots,B_{k'}')$ and $(B_1,\cdots,B_k)$ are measurable partitions, and if $\left(B_1{}',\cdots,B_{k'}\right)$ is a refinement of $\left(B_1,\cdots,B_k\right)$ with $B_1 = \bigcup_1^{r_1} B_i{}', B_2 = \bigcup_{r_1+1}^{r_2} B_i{}',\cdots,B_k = \bigcup_{r_{k-1}+1}^{k'} B_i{}'$,*

then the distribution of

$$\left(\sum_{1}^{r_1} P\left(B_i{}'\right), \sum_{r_1+1}^{r_2} P\left(B_i{}'\right), \cdots, \sum_{r_{k-1}+1}^{k'} P\left(B_i{}'\right)\right)$$

as determined from the joint distribution of $\left(P\left(B_1{}'\right), \ldots, P\left(B'_{k'}\right)\right)$*, is identical to the distribution of* $\left(P\left(B_1\right), \ldots, P\left(B_k\right)\right)$.

The following lemma shows that this condition is sufficient for the validity of the Kolmogorov consistency conditions for the distributions of $\left(P\left(A_1\right), \ldots, P\left(A_m\right)\right)$. We will say that P is a random probability measure on $(\mathbb{X}, \mathscr{X})$, if above condition 1 is satisfied, if $P(A)$ takes values only in $[0,1]$, and if $P\left(\mathbb{X}\right)$ is degenerate at 1.

Lemma 4. *If a system of joint distributions of (* $P\left(B_1\right), \cdots, P\left(B_k\right)$ *) for all* k *and measurable partitions (* $B_1, \cdots, B_k$ *) is defined satisfying Condition 1 , and if for arbitrary measurable sets* $A_1, \cdots, A_m$*, the distribution of* $\left(P\left(A_1\right), \cdots, P\left(A_m\right)\right)$ *is defined as in A.1, A.2, A.3, then there exists a probability* $\mathbb{P}$ *on* $\left([0,1]^{\mathscr{X}}, B\mathscr{F}^{\mathscr{X}}\right)$ *yielding these distributions.*

Proof. Since $\mathscr{X} \cup \varnothing = \mathscr{X}$, it follows from Condition 1 that $P(\varnothing)$ is degenerate at zero, and hence the distribution of $\left(P\left(A_1\right), \ldots, P\left(A_m\right)\right)$ is welldefined by A.2 and A.3. To check the **Kolmogorov consistency conditions**, we must show that, for arbitrary m and measurable sets $A_1, \cdots, A_m$, the marginal distribution of $\left(P\left(A_1\right), \cdots, P\left(A_{m-1}\right)\right)$ derived from the distribution of $\left(P\left(A_1\right), \cdots,\right.$ $\left.P\left(A_m\right)\right)$ is identical to the defined distribution of $\left(P\left(A_1\right), \ldots, P\left(A_{m-1}\right)\right.$).

The marginal distribution of $\left(P\left(A_1\right), \cdots, P\left(A_{m-1}\right)\right)$ derived from the distribution of $\left(P\left(A_1\right), \cdots, P\left(A_m\right)\right)$ is identical to the distribution of

$$\left(\sum_{\left(\nu_1, \cdots, \nu_m\right), \nu_1=1} P\left(B_{\nu_1, \cdots, \nu_m}\right), \cdots, \sum_{\left(\nu_1, \cdots, \nu_m\right), \nu_{m-1}=1} P\left(B_{\nu_1, \cdots, \nu_m}\right)\right) \tag{5}$$

derived from the distribution of (2). The distribution of ($P\left(A_1\right), \ldots, P\left(A_{m-1}\right)$) is defined as the distribution of

$$\left(\sum_{(\nu_1,\cdots,\nu_{m-1}),\nu_1=1} P\left(B_{\nu_1,\cdots,\nu_{m-1}}\right), \cdots, \sum_{(\nu_1,\cdots,\nu_{m-1})',\nu_{m-1}=1} P\left(B_{\nu_1,\cdots,\nu_{m-1}}\right) \right) \tag{6}$$

derived from the distribution of $\left\{P\left(B_{\nu_1,\cdots,\nu_{m-1}}\right); \nu_i = 0 \text{ or } 1, i = 1,\ldots,m-1\right\}$ where

$$B_{\nu_1,\cdots,\nu_{m-1}} = \bigcup_{j=1}^{m-1} {A_j}^{\nu_j}$$

Since $B_{\nu_1,\cdots,\nu_{m-1}} = B_{\nu_1,\cdots,\nu_{m-1},1} \cup B_{\nu_1,\cdots,\nu_{m-1},0}$, Condition 1 implies that the distribution of $\left\{P\left(B_{\nu_1,\cdots,\nu_{m-1}}\right); \nu_j = 0 \text{ or } 1, j = 1,\cdots,m-1\right\}$ is identical to the distribution of $\left\{P\left(B_{\nu_1,\cdots,\nu_{m-1},1}\right) + P\left(B_{\nu_1,\cdots,\nu_{m-1},0}\right); \nu_j = 0 \text{ or } 1, j = 1,\ldots,m-1\right\}$ as determined from the distribution of A.2. Thus, the distribution of 6 can also be found from the distribution of A.2 upon replacing $P\left(B_{\nu_1,\cdots,\nu_{m-1}}\right)$ by $P\left(B_{\nu_1,\cdots,\nu_{m-1},1}\right) + P\left(B_{\nu_1,\cdots,\nu_{m-1},0}\right)$. With this replacement, 6 becomes formally identical to 5, proving that their distributions are identical.

□

Appendix B

Additional proofs

This section gives proofs skipped in Chapter 3

Proof of Theorem 4

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,

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

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,

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

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

Proof of Lemma 3

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 1 to λX yields

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

Maximizing over λ yields the result. □

Proof of Lemma 2

Proof.

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

where (b) follows from the Cauchy–Schwarz inequality, (c) follows from Fact 3, and (a) follows from Eq.3.14and (d) from the standard definition of mutual-information. □

Bibliography

Michael Domjan. Pavlovian conditioning: A functional perspective. *Annu. Rev. Psychol.*, 56(1): 179–206, 2005.

Martin L Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Sumit Vashishtha and Siddhartha Verma. Restoring chaos using deep reinforcement learning. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(3), 2020.

Siddhartha Verma, Guido Novati, and Petros Koumoutsakos. Efficient collective swimming by harnessing vortices through deep reinforcement learning. *Proceedings of the National Academy of Sciences*, 115(23):5849–5854, 2018.

Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897): 414–419, 2022.

Daniel J Mankowitz, Andrea Michi, Anton Zhernov, Marco Gelmi, Marco Selvi, Cosmin Paduraru, Edouard Leurent, Shariq Iqbal, Jean-Baptiste Lespiau, Alex Ahern, et al. Faster sorting algorithms discovered using deep reinforcement learning. *Nature*, 618(7964):257–263, 2023.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

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

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.

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.

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

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.

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.

Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

Radford M Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

Adam S Lowet, Qiao Zheng, Sara Matias, Jan Drugowitsch, and Naoshige Uchida. Distributional reinforcement learning in the brain. *Trends in neurosciences*, 43(12):980–997, 2020.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. PMLR, 2017a.

Didier Sornette. *Why stock markets crash: critical events in complex financial systems*. Princeton university press, 2017.

Yaakov Engel, Shie Mannor, and Ron Meir. Reinforcement learning with gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pages 201–208, 2005.

Jayanta K Ghosh and RV Ramamoorthi. *Bayesian nonparametrics*. Springer, 2003.

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

Ismaël Castillo. Bayesian nonparametric statistics, st-flour lecture notes. *arXiv preprint arXiv:2402.16422*, 2024.

Bruno De Finetti. La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68, 1937.

Peter Orbanz and Daniel M Roy. Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE transactions on pattern analysis and machine intelligence*, 37(2): 437–461, 2014.

Olav Kallenberg. *Probabilistic symmetries and invariance principles*. Springer, 2005.

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

Thomas S Ferguson. Prior distributions on spaces of probability measures. *The annals of statistics*, 2(4):615–629, 1974.

Andre Nikolaevich Kolmogorov. *Foundations of the theory of probability: Second English Edition*. Courier Dover Publications, 2018.

Thomas S Ferguson and Michael J Klass. A representation of independent increment processes without gaussian components. *The Annals of Mathematical Statistics*, 43(5):1634–1643, 1972.

John Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967.

Antonio Lijoi, Igor Prünster, et al. Models beyond the dirichlet process. *Bayesian nonparametrics*, 28(80):342, 2010.

François Caron and Emily B Fox. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(5):1295–1366, 2017.

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

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

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

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

Sumit Vashishtha and Odalric-Ambrym Maillard. Leveraging priors on distribution functions for multi-arm bandits. *Reinforcement Learning Journal*, 6:1600–1623, 2025.

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.

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

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.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer, 1992.

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

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.

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

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

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.

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

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

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

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

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

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

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

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.

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

Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9 (03):90–95, 2007.

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.

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.

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.

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

Abraham Wald. *Statistical decision functions*. Wiley, 1961.

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

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

Chris C Holmes and Stephen G Walker. Statistical inference with exchangeability and martingales. *Philosophical Transactions of the Royal Society A*, 381(2247):20220143, 2023.

Edwin Fong, Chris Holmes, and Stephen G Walker. Martingale posterior distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1357–1391, 2023.

Sandra Fortini and Sonia Petrone. Exchangeability, prediction and predictive modeling in bayesian statistics. *Statistical Science*, 40(1):40–67, 2025.

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

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

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

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.

Peter Müller and Riten Mitra. Bayesian nonparametric inference–why and how. *Bayesian analysis (Online)*, 8(2):10–1214, 2013.

Joey Hong, Branislav Kveton, Manzil Zaheer, Mohammad Ghavamzadeh, and Craig Boutilier. Thompson sampling with a mixture prior. In *International Conference on Artificial Intelligence and Statistics*, pages 7565–7586. PMLR, 2022.

Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, Amr Ahmed, and Craig Boutilier. Latent bandits revisited. *Advances in Neural Information Processing Systems*, 33:13423–13433, 2020.

Branislav Kveton, Boris Oreshkin, Youngsuk Park, Aniket Anand Deshmukh, and Rui Song. Online posterior sampling with a diffusion prior. *Advances in Neural Information Processing Systems*, 37:130463–130484, 2024.

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.

Pierre Alquier. User-friendly introduction to pac-bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2):174–303, 2024.

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.

Ismaël Castillo. Pólya tree posterior distributions on densities. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 53(4):2074 – 2102, 2017. doi: 10.1214/16-AIHP784. URL https://doi.org/10.1214/16-AIHP784.

Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5(1):4308, 2014.

Sören Bergmann, Sören Stelzer, and Steffen Strassburger. On the use of artificial neural networks in simulation-based manufacturing control. *Journal of Simulation*, 8(1):76–90, 2014.

Ofélia Anjos, Carla Iglesias, Fatima Peres, Javier Martínez, Angela Garcia, and Javier Taboada. Neural networks applied to discriminate botanical origin of honeys. *Food chemistry*, 175: 128–136, 2015.

Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553): 452–459, 2015.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Martin Krzywinski and Naomi Altman. Points of significance: Importance of being uncertain. *Nature methods*, 10(9), 2013.

Ondrej Linda, Todd Vollmer, and Milos Manic. Neural network based intrusion detection system for critical infrastructures. In *2009 international joint conference on neural networks*, pages 1827–1834. IEEE, 2009.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Gencast: Diffusion-based ensemble forecasting for medium-range weather. *arXiv preprint arXiv:2312.15796*, 2023.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015.

Julyan Arbel, Konstantinos Pitas, Mariia Vladimirova, and Vincent Fortuin. A primer on bayesian neural networks: review and debates. *arXiv preprint arXiv:2309.16314*, 2023.

Carlos Riquelme, George Tucker, and Jasper Snoek. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. *arXiv preprint arXiv:1802.09127*, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, pages 1995–2003. PMLR, 2016.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

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

Matthias Schultheis, Constantin A Rothkopf, and Heinz Koeppl. Reinforcement learning with non-exponential discounting. *Advances in neural information processing systems*, 35:3649–3662, 2022.

William Fedus, Carles Gelada, Yoshua Bengio, Marc G Bellemare, and Hugo Larochelle. Hyperbolic discounting and learning over multiple horizons. *arXiv preprint arXiv:1902.06865*, 2019.

Will Dabney, Zeb Kurth-Nelson, Naoshige Uchida, Clara Kwon Starkweather, Demis Hassabis, Rémi Munos, and Matthew Botvinick. A distributional code for value in dopamine-based reinforcement learning. *Nature*, 577(7792):671–675, 2020.

Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017b.

Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 799–806, 2010.

Marc G Bellemare, Will Dabney, and Mark Rowland. *Distributional reinforcement learning*. MIT Press, 2023.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49:209–232, 2002.

Karl Friston. The history of the future of the bayesian brain. *NeuroImage*, 62(2):1230–1233, 2012.

David J Aldous and Antar Bandyopadhyay. A survey of max-type recursive distributional equations. 2005.

Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Policy gradient for coherent risk measures. *Advances in neural information processing systems*, 28, 2015.

Ramanan Gopalan and Donald A Berry. Bayesian multiple comparisons using dirichlet process priors. *Journal of the American Statistical Association*, 93(443):1130–1139, 1998.

Willie Neiswanger and Aaditya Ramdas. Uncertainty quantification using martingales for misspecified gaussian processes. In *Algorithmic learning theory*, pages 963–982. PMLR, 2021.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Uriel Frisch. *Turbulence: the legacy of AN Kolmogorov*. Cambridge university press, 1995.

Gregory Falkovich and Katepalli R Sreenivasan. Lessons from hydrodynamic turbulence. *Physics Today*, 59(4):43–49, 2006.

Alexander Obukhov. Spectral energy distribution in a turbulent flow. *Izv. Akad. Nauk. SSSR. Ser. Geogr. i. Geofiz*, 5:453–466, 1941.

Andrei Nikolaevich Kolmogorov. Local structure of turbulence in an incompressible viscous fluid at very high reynolds numbers. *Soviet Physics Uspekhi*, 10(6):734, 1968.

Benoit B Mandelbrot. Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier. *Journal of fluid Mechanics*, 62(2):331–358, 1974.

C Meneveau and KR Sreenivasan. Simple multifractal cascade model for fully developed turbulence. *Physical review letters*, 59(13):1424, 1987.

Yinxiang Ma, Wanting Cheng, Shidi Huang, François G Schmitt, Xin Lin, and Yongxiang Huang. Hidden turbulence in van gogh's the starry night. *Physics of Fluids*, 36(9), 2024.

Thomas Milton Liggett and Thomas M Liggett. *Interacting particle systems*, volume 2. Springer, 1985.

Giorgio Parisi. Nobel lecture: Multiple equilibria. *Reviews of Modern Physics*, 95(3):030501, 2023.

JA Scott Kelso. *Dynamic patterns: The self-organization of brain and behavior*. MIT press, 1995.

Melanie Mitchell. *Complexity: A guided tour*. Oxford university press, 2009.