Lévy processes in Bayesian nonparametric models

Candidate number: 1055360

University of Oxford

A thesis submitted for the degree of

M.Sc. in Mathematical Sciences

Trinity Term 2021

Abstract

A major point of Bayesian nonparametrics has been the problem of clustering, or partitioning, where each data point is modeled as belonging to one or in a multiple of clusters or partition blocks. The underlying mathematical mechanism that drives Bayesian nonparametric models are certain combinatorial stochastic processes, such as the Dirichlet process, Chinese restaurant process and Indian buffet process.

In this dissertation, we provide a detailed description of these processes. Specifically, we present their stick-breaking constructions in order to link them to Lévy processes through the gamma and beta subordinators and the theory of exchangeability, which lead to the study of the gamma and beta process. Finally, completely random measures and their inference in Bayesian nonparametrics are presented, as they arise through the modelling of jump sizes of subordinators as weights of atoms at locations drawn independently from the base distributions of processes such as the beta process.

Word count: 7221 (Document → Word Count feature in LaTeX)

Contents

0 Abbreviations

The following abbreviations will be used throughout this dissertation:

PPP	Poisson point process
BeS	Beta subordinator
BP	Beta process
BeP	Bernoulli process
GaS	Gamma subordinator
GaP	Gamma process
DP	Dirichlet process
CRP	Chinese restaurant process
IBP	Indian buffet process
CRM	Completely random measure
BNP	Bayesian nonparametrics

1 Introduction

Probabilistic models are used throughout machine learning to model distributions over empirical data. Standard parametric models use a fixed and finite number of parameters to arrive at a solution to a particular problem. This can result into problems such as under- or over-fitting when the complexity of the model does not align with the amount of data available. Thus, choosing a model with the correct number of parameters, a problem known as model selection, can be a difficult task. The alternative way of setting up a model that can sidestep this problem is through the Bayesian nonparametric approach. Using a model of infinite complexity, or an infinite number of parameters, there is a greater degree of flexibility in describing the underlying structure of the data, and the problem of model selection is avoided.

The Bayesian paradigm for learning from data $\mathcal{D}$ is simple and intuitive: if we want to reason about some variables or parameters θ, and we have started with some prior knowledge about these $p(\theta)$, we can update that knowledge to add more information from the data. Specifically, we characterise how likely different values of θ are given the observed data using a likelihood function $p(\mathcal{D}|\theta)$. This is known as Bayesian inference and relies on the classical Bayes' rule

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}.$$

The denominator $p(\mathcal{D})$ is a normalisation constant known as the marginal likelihood and is used to ensure that $p(\theta|\mathcal{D})$ is a valid probability distribution (or probability density). Lévy processes have been used as priors in the Bayesian context, as they provide analytical representations for the posterior. Examples include the gamma process, beta process, negative binomial process, while the prevailing Lévy process underlying most stochastic processes in this setting is the Poisson point process.

The problem of assigning datapoints to clusters in Bayesian nonparametrics has been under major investigation in the modern era. It refers to assigning datapoints to one or simultaneously many clusters, given a dataset. In this context, flexible data structures meet tractable probabilistic inference through models such as the Dirichlet process, Chinese restaurant process (each datapoint belongs to one cluster) and Indian Buffet process (each datapoint belongs to many clusters).

Section 2 introduces Lévy processes and subordinators through the Lévy Khintchine formula, Lévy-Ito decomposition for describing the path properties of Lévy processes, as well the Poisson point process.

Section 3 describes three models used for modeling the clustering problem in Bayesian nonparametric inference: the Dirichlet process, Chinese Restaurant process and Indian Buffet process. We are able to combine information from different sources, in order to describe the rich mathematical structure of these processes and the ones underlying them, such as the gamma and beta processes. Specifically, the CRP and

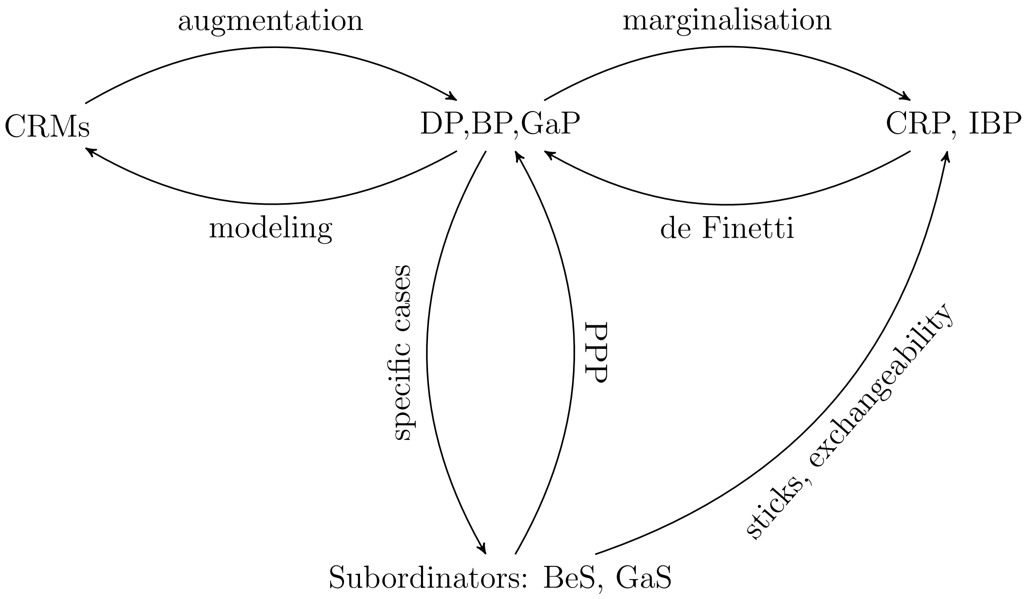

Figure 1: Overall description of notions in the dissertation.

IBP are connected to the gamma and beta subordinator respectively using their stick-breaking representations and the theory of exchangeability, while further connections are made for the CRP to the DP and gamma process, and the IBP to the beta process.

As the sequence of subordinator jumps in order of appearance are shown to form random measures (a normalised random measure for the DP and completely random measures for the beta, gamma processes) it is natural to study completely random measures in Section 4. Furthermore, a thorough description for the use of CRMs in the BNP setting is given, the problem of inference for CRMs is addressed and connections to the processes studied in Section 2 are presented.

Finally, a recap of the aim of the dissertation is given in the Conclusions 4, while the Appendix 6 offers information about the Dirichlet distribution, its stick-breaking representation and various results needed throughout proofs in Section 3. Figure 1 is a graphical representation of how notions throught this dissertation interplay.

2 Lévy processes and preliminaries

Lévy processes are stochastic processes that can be used as mathematical models for the description of random phenomena that are dynamic over time, such as the movement of stocks in finance, an insurer's vulnerability to insolvency/ruin in risk theory or, as priors in Bayesian nonparametrics.

Lévy processes are a natural generalisation of Brownian motion: they disregard the Gaussianity assumption, modify the continuity to càdlàg and keep the stationary and independent increments.

Definition 1. *(Lévy Process) A real-valued process $(X_t)_{t\geq 0}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a Lévy Process, if*

(i) The paths of X_t, $t \geq 0$ are $\mathbb{P}$-almost surely right-continuous with left limits,

(ii) For every $s, t \geq 0$, $X_{t+s} - X_t$ is equal in distribution to X_s,

(iii) For every $s, t \geq 0$, $X_{t+s} - X_s$ is independent of $\{X_u : u \leq s\}$.

(iv) $\mathbb{P}(X_0 = 0) = 1$.

A family of distributions with which a Lévy process is closely related to are infinitely divisible distributions. Such distributions can be split into n i.i.d. parts for every natural number n.

Definition 2. *(Infinite divisible distribution) A real-valued random variable Y has an infinitely divisible distribution if for each $n = 1, 2, \ldots,$ there exists a sequence of i.i.d. random variables $Y_{1,n}, \ldots, Y_{n,n}$ such that*

$$Y \overset{d}{=} Y_{1,n} + Y_{2,n} + \cdots + Y_{n,n}.$$

Lévy processes are infinitely divisible, as they can be split into n i.i.d. pieces using (i), (ii) of Definition 1

$$X_s = X_{s/n} + X_{2s/n} - X_{s/n} + \cdots + X_s - X_{(s-1)/n}.$$

Infinite divisibility induces a characteristic triplet, by which every Lévy process can be uniquely identified, i.e. by its Lévy-Khintchine representation. The notation $\Lambda(dx) = g(x)dx$ for some function g is used.

Theorem 1. *(Lévy-Khintchine formula) [1, Theorem 1.3] A probability law μ of a real-valued random variable is infinitely divisible with characteristic exponent Φ,*

$$\int_{\mathbb{R}} e^{i\theta x} \mu(dx) = e^{-\Phi(\theta)}, \text{ for } \theta \in \mathbb{R}, \tag{1}$$

if and only if there exists a unique triple $(\alpha, \sigma^2, \Lambda)$, with $\alpha \in \mathbb{R}$, $\sigma^2 \geq 0$ and Λ a measure concentrated on $\mathbb{R}\backslash\{0\}$ satisfying $\int_{\mathbb{R}}(1 \wedge x^2)\Lambda(dx) < \infty$, such that

$$\Phi(\theta) = -i\alpha\theta + \frac{1}{2}\sigma^2\theta^2 + \int_{\mathbb{R}}(1 - e^{i\theta x} + i\theta x \mathbb{1}_{|x|<1})\Lambda(dx), \text{ for every } \theta \in \mathbb{R}. \tag{2}$$

Definition 3. *The measure Λ from Theorem 1 is called the Lévy (characteristic) measure.*

The reverse is also true: for a chosen triple as above, under (2) we can guarantee the existence of a Lévy process with Φ as its characteristic exponent.

Theorem 2. *[1, Theorem 1.6] (Lévy-Khintchine formula for Lévy processes) Suppose $\alpha \in \mathbb{R}$, $\sigma^2 \geq 0$ and Λ is a measure concentrated on $\mathbb{R}\backslash\{0\}$ such that $\int_{\mathbb{R}}(1 \wedge x^2)\Lambda(dx) < \infty$. For Φ given by (2), there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, on which a Lévy process $X = (X_t)_{t\geq 0}$ is defined, such that X_t has characteristic exponent Φ.*

An essential feature of Lévy Processes are their jumps, and thus an appropriate mathematical mechanism is needed to describe the jump structure. That mechanism is a Poisson random measure, or equivalently a Poisson point process.

Definition 4. *[2] (Poisson counting measure) Let $D \subset \mathbb{R}^d$, $\lambda : D \to [0, \infty)$ be a locally integrable intensity function. A set function $A \to N(A)$ that satisfies*

1. *for all $n \geq 1$ and disjoint $A_1, \ldots, A_n \subset D$, the random variables $N(A_1), \ldots, N(A_n)$ are independent,*
2. *$N(A) \sim$ Poisson$(\int_A(\lambda(x)dx))$,*

is called a Poisson counting measure with intensity function $\lambda(x)$.

Definition 5. *[2] (Poisson point process) Let g be locally integrable on $D_0 \subset \mathbb{R}^{d-1}\backslash\{0\}$ (or ν locally finite). A process $(\Delta_t)_{t\geq 0}$ in $D_0 \cup \{0\}$ such that*

$$N((a,b] \times A_0) = \#\{t \in (a,b] : \Delta_t \in A_0\}, \quad 0 \leq a < b, A_0 \subset D_0 \text{ (measurable)},$$

is a Poisson counting measure with intensity $\Lambda((a,b] \times A_0) = (b-a)\int_{A_0} g(x)dx$ (or $\Lambda((a,b] \times A_0) = (b-a)\nu(A_0)$), is called a Poisson point process with intensity g (or intensity measure ν).

2.1 Path properties

We can understand path-related properties of Lévy processes by studying the characteristic exponent further. The characteristic exponent Φ can be decomposed into 3 parts

$$\Phi(\theta) = \underbrace{-i\alpha\theta + \frac{1}{2}\sigma^2\theta^2}_{\text{Brownian motion}} + \underbrace{\Lambda(\mathbb{R}\backslash\{(-1,1)\} \int_{\mathbb{R}\backslash\{(-1,1)\}} (1 - e^{i\theta x})\Lambda(dx)/\Lambda(\mathbb{R}\backslash\{(-1,1)\}}_{\text{compound Poisson process}}$$
$$+ \underbrace{\int_{0<|x|<1} (1 - e^{i\theta x} + i\theta x)\Lambda(dx).}_{\text{Lévy martingale}} \tag{3}$$

Each part is itself a characteristic exponent for a Brownian motion, compound Poisson process and a Lévy martingale respectively. This is captured in the Lévy-Itô decomposition.

Theorem 3. *[1, Theorem 2.1] (Lévy-Itô decomposition) Given any characteristic triplet $(\alpha, \sigma^2, \Lambda)$, with the measure Λ concentrated on $\mathbb{R}\backslash\{0\}$ such that*

$$\int_{\mathbb{R}} (1 \wedge x^2)\Lambda(dx) < \infty,$$

there exists a probability space on which three independent Lévy processes $X^{(1)}$,$X^{(2)}$ and $X^{(3)}$ exist. These are defined for every $t \geq 0$, as

$$X_t^{(1)} = \sigma B_t + \alpha t, \; X_t^{(2)} = \sum_{i=1}^{N_t} \xi_i,$$

where B_t is a standard Brownian Motion, $(N_t)_{t\geq 0}$ is a Poisson process with rate $\Lambda(\mathbb{R}\backslash(-1,1))$ and $\{\xi_i\}_{i\geq 1}$ are independent and identically distributed with common distribution $\Lambda(dx)/\Lambda(\mathbb{R}\backslash(-1,1))$ restricted to $\{x : |x| \geq 1\}$, and if $\Lambda(\mathbb{R}\backslash(-1,1)) = 0$, $X^{(2)} = 0$ identically. Furthermore, $X^{(3)}$ is a square-integrable martingale with an almost surely countable number of path discontinuities on each finite time interval, of magnitude less than one. Taking $X = X^{(1)} + X^{(2)} + X^{(3)}$, Theorem 2 holds and ensures the existence of a probability space on which a Lévy process is defined with characteristic exponent given by the Lévy-Khintchine formula (2).

The first two parts of the decomposition can be identified as the characteristic components of a linear Brownian motion and a compound Poisson process of "big" jumps. To identify the third component, a Lévy process of "small" jumps is constructed, defined via compound Poisson processes with drift.

We will focus on a specific class of Lévy processes called subordinators [1, 3].

Definition 6. *A subordinator $(T_t)_{t\geq 0}$ is a non-negative increasing Lévy process.*

Subordinators are interesting for our purposes, as they can be decomposed in two parts: a deterministic part (drift) and a random part (Poisson point process). The subordinator decomposition is presented in the following theorem. The Lévy-Khintchine formula of the subordinator is presented.

Theorem 4. *[2] Let $c \geq 0$, and let $(\Delta)_{t\geq 0}$ be a Poisson point process with intensity measure ν on $(0, \infty)$ such that*

$$\int_0^\infty (1 \wedge x)\nu(dx) < \infty. \tag{4}$$

Then, the process $T_t = ct + \sum_{s\leq t} \Delta_s$ is a subordinator with characteristic exponent

$$\Phi(u) = cu + \int_0^\infty (e^{ux} - 1)\nu(dx)$$

Remark. *Every subordinator $(T_s, s \geq 0)$ can be written as*

$$T_s = cs + \sum_{k=1}^{\infty} x_k \mathbb{1}\{\psi \leq s\}, \tag{5}$$

for some $c \geq 0$, and $\{(x_k, \psi_k)\}_k$ is a countable set of points of a Poisson point process with intensity measure $\nu(dx, d\psi) = \Lambda(dx)d\psi$, for a Lévy measure Λ such that Eq. (4) holds.

Theorem 5. *[4] Let $(T_s, s \geq 0)$ be a subordinator, and $u \geq 0$. Then*

$$\mathbb{E}[e^{-uT_s}] = e^{-s\Phi(u)}, \qquad \Phi(u) = cu + \int_0^\infty (1 - e^{-ux})\Lambda(dx), \tag{6}$$

where $c \geq 0$ is the drift constant and Λ is a non-negative Lévy measure on $(0, \infty)$.

In the next section, we will see how subordinators are used in the problem of clustering or feature allocation. The subordinators that we will use are the following. Different notation is used to distinguish general results from these specific subordinators.

Definition 7. *(Gamma process subordinator) The gamma process subordinator (GaS) is characterized by its drift c and Lévy measure Λ*

$$c = 0, \quad \text{and} \quad \Lambda(d\theta) = \alpha\theta^{-1}e^{-b\theta}d\theta,$$

for α, $b > 0$.

Definition 8. *(Beta process subordinator) The beta process subordinator (BeS) is characterized by its drift c and Lévy measure Λ*

$$c = 0, \quad \text{and} \quad \Lambda(d\theta) = \alpha\theta^{-1}(1-\theta)^{\alpha-1}d\theta,$$

for $\alpha > 0$ its concentration parameter.

2.2 Exchangeability and de Finetti's theorem

The processes to be studied will be connected via the subordinators presented through the theory of exchangeability, and specifically de Finetti's theorem [5].

Definition 9. *(Exchangeable distribution) Consider a finite sequence $X_i, i = 1, \dots, n$ random variables, and let their joint distribution be $p(x_1, \dots, x_n)$. The distribution $(X_1, \dots, X_n)$ is called exchangeable (or the random variables $X_1, \dots, X_n$ are exchangeable) if for any permutation σ of the integers $\{1, \dots, n\}$, the joint distribution of the random variables is unchanged*

$$p(x_1, \dots, x_n) = p(x_{\sigma(1)}, \dots, x_{\sigma_{(n)}}).$$

Theorem 6. *[6] (de Finetti's theorem) Let $(X_i)_{i=1}^{\infty}$ be a sequence of Bernoulli(θ) distributed, finitely exchangeable random variables, i.e. for every $n > 0$ each finite sub-sequence $\{X_i\}_{i=1}^{\infty}$ is exchangeable. Then, there exists a random variable Θ and a distribution function $\mathcal{F}(\theta)$, such that*

$$\begin{aligned} &p\Big(\lim_{n\to\infty} \frac{\sum X_i}{n} = \Theta\Big) = 1 \;\; \textit{with}\;\; \Theta \sim \mathcal{F}(\theta) \\ &p(x_1, \dots, x_n) = \int_0^1 \Big[\prod_i \theta^{x_i}(1-\theta)^{1-x_i}\Big] d\mathcal{F}(\theta). \end{aligned} \tag{7}$$

Remark. *There exists a generalisation of the above theorem, that holds for any finite sequence of exchangeable random variables, i.e. not only for Bernoulli random variables. It takes the form*

$$p(x_1, \dots, x_n) = \int \left[\prod_{i=1}^{n} \mathbb{P}(x_i|G)\right] dP(G). \tag{8}$$

$P(G)$ will be thought of as the "de Finetti mixing distribution". De Finetti's theorem implies that if the prior is exchangeable, then one can equivalently assume that the random variables are independent conditional on an underlying probability distribution G on the space of outcomes.

3 Models in Machine Learning

In Bayesian nonparametrics (BNP), finite-dimensional priors are replaced by stochastic processes. In early and late work, such as [7, 8, 9, 10] the definition of Lévy processes used is an older one than the one introduced in the previous chapter. Under this older definition, Lévy processes are purely independent increment processes and do not necessarily exhibit stationary increments. The result of this definition is that such processes have fixed times of discontinuity, while incorporating the stationarity requirement would generate jumps that happen at random times of a Poisson point process. However, for specific cases of some processes in BNP (e.g. beta process, gamma process), connections to Lévy processes can be made through their corresponding subordinators and the theory of exchangeability [4].

A strong point of connection between Lévy processes and independent increment processes is the Poisson point process. Many stochastic processes in BNP use the Poisson point process as a starting point from which points and locations will be drawn, and either the weights at these points will be transformed into weights of the new process (gamma process), or only the location of the weights will be kept, and new weights will be generated according to a different rate measure (beta process, negative binomial process, Bernoulli process).

The foundation of BNP inference was laid by two technical reports [7, 8]. The Dirichlet process was introduced in [7], while the more general approach of neutral to the right processes was presented in [8]. Due to the simplicity in the calculation of the posterior of the DP via parameter updating, it has been under great exploration and use. The connection between stochastic processes such as the CRP and IBP used in BNP and subordinators through their stick-breaking representations is found in [4].

3.1 Dirichlet process-Chinese restaurant process

A partition π_n of $[n]$ is a collection of mutually exclusive, exhaustive and nonempty subsets of $[n]$, called blocks. A generalisation of a partition introduced in [4] is a feature allocation f_n of $[n]$, defined as a multiset of nonempty subsets of $[n]$, called blocks, such that each index n can belong to any finite number of blocks. Blocks of a partition are called clusters, while blocks of a feature allocation are called features.

Subordinators can be used for feature allocation: generate feature membership from a subordinator by sampling Bernoulli draws at each of its jumps, with probability of success equal to the jump size. Every jump has positive size, and thus the feature associated will eventually score a Bernoulli success for some index n with probability one. Then, the jumps can be enumerated in order of appearance: at the n-th iteration, enumerate all features in which index n appears, but not previous indexes. In particular, in the case of a subordinator being finite up to time t, its

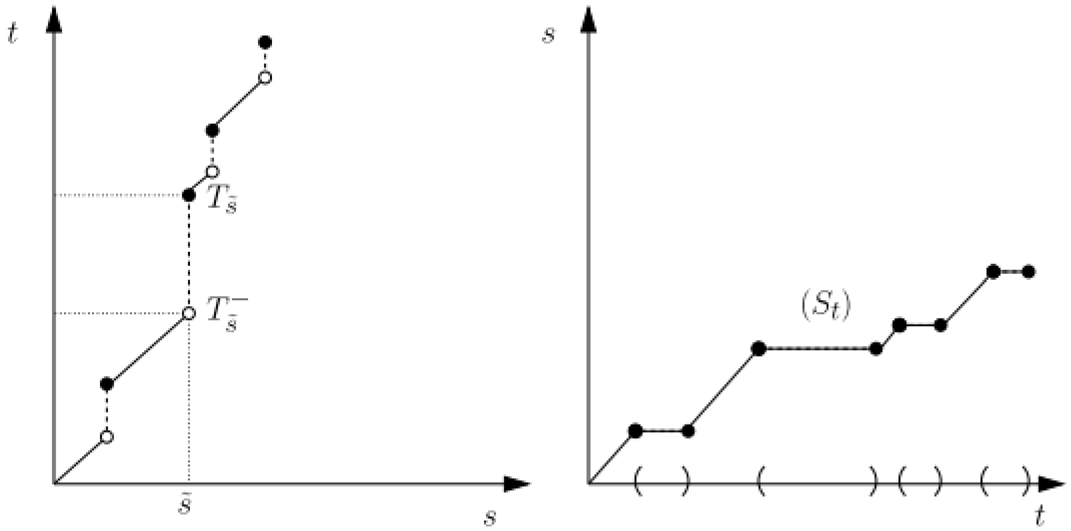

Figure 2: Sample path T_s of a subordinator, with $T_{\tilde{s}}^-$ the left limit of T_s at $s = \tilde{s}$ (Left) and the right continuous inverse S_t of the same subordinator, with open intervals in the t axis denoting the subordinator jumps (Right).

jumps up to time t can be used as feature block frequencies if they have support on [0,1) or via normalization as partition block frequencies. This can be done using the right-continuous inverse S_t of a subordinator T_s: $S_t := \inf\{s : T_s > t\}$, with open intervals along t-axis corresponding to the jumps of the subordinator T_s. For the jumps of the subordinator to partition intervals of the form $[0, t)$, we need the drift component to be zero. For an illustration see Figure 2, found in [4].

The Chinese restaurant process was conceived by Lester Dubins and Jim Pitman in the 1980's as a way of constructing random permutations and random partitions. It is a stochastic process with index set the set of permutations of $[n] = \{1, \ldots, n\}$. It turns out that the probability of a resulting configuration of the CRP is an exchangeable partition probability function, which we introduce [11].

Definition 10. *A random partition Π_n of $[n]$ is called exchangeable if its distribution is invariant under the natural action on partitions of $[n]$ by the symmetric group of permutations of $[n]$. Equivalently, for each partition $\{A_1, \ldots, A_n\}$ of $[n]$,*

$$\mathbb{P}(\Pi_n = \{A_1, \ldots, A_k\}) = p(|A_1|, |A_2|, ..., |A_k|),$$

for some symmetric function p of compositions $(n_1, \ldots, n_k)$ of $[n]$. The function p is called the exchangeable partition probability function (EPPF) of Π_n.

Let a sequence of random permutations $(\sigma_n, n \geq 1)1$ be such that (i) σ_n is a uniformly distributed random permutation of $[n]$ for each n and (ii) for each n if σ_n

is written as a product of cycles, then σ_{n-1} is derived from σ_n by deletion of element n from its cycle. These determine a unique distribution for the sequence σ_n, which has an elegant way of being described: an initially empty restaurant has an unlimited number of circular tables numbered $1, 2, \ldots$, each with sitting capacity of an infinite number of customers. Customers are numbered $1, 2, \ldots$, arrive one by one and are seated according to the following plan.

The first customer is seated at the first table. After that, customer $n+1$ is placed

- at an occupied tabled j with probability $p(\ldots, n_j + 1, \ldots)/p(n_1, ..., n_k)$,
- at a new table with probability $p(n_1, \ldots, n_k, 1)/p(n_1, \ldots, n_k)$.

Chinese restaurant process (CRP): Imagine a restaurant with an infinite number of tables, and a sequence of customers entering the restaurant one-by-one. The first customer sits at some unoccupied table. A "dish" is set out at the new table, and dish "1" is called the first dish. The customer is assigned to their table: $Z_1 = 1$. Recursively, for a restaurant with concentration parameter $\alpha > 0$, the n-th customer joins an occupied table with probability proportional to the number of previous customers sitting there, or opens up a new table with probability proportional to α. In the first case, Z_n takes the value of the existing dish at the table, or in the other case, the value of the next available dish k (equal to the number of already available dishes plus one) and $Z_n = k$. Summing over all possibilities of the arrival of the n-th customer, and normalizing via that term we get for $K_n := \max\{Z_1, \ldots, Z_n\}$ that the distribution of the table assignments for the n-th customer is

$$\mathbb{P}(Z_n = k | Z_1, \ldots, Z_{n-1}) = (n - 1 + \alpha)^{-1} \begin{cases} N_k \text{ if } k \leq K_{n-1} \\ \alpha \text{ otherwise,} \end{cases} \tag{9}$$

where $N_k = \#\{m \in \mathbb{N} : m < n, Z_m = k\}$. The following result is taken from [12].

Theorem 7. *The probability of a resulting configuration π_n of the CRP is*

$$\mathbb{P}(\Pi_N = \pi_N) = \frac{\alpha^{K-1} \prod_{k=1}^{K} (N_k - 1)!}{\prod_{n=1}^{N} (n - 1 + \alpha)}, \tag{10}$$

for N_k the number of customers assigned to the k-th group and is an EPPF.

Proof. We first find the probability of the partition induced by considering the collection of indices sitting at each table as a block in the partition. Suppose that I_k is the set of indices of customers assigned to the k-th group and N_k is the number of individuals sitting at the k-th table, with a total number of non-zero cardinality of table occupancies $N := \sum_{k=1}^{K} N_k$. That is, consider the case when N customers have entered the restaurant and sat at K different tables in the specified configuration.

From Eq. (9), we will get for table k a product of terms

$$\frac{\alpha \cdot 1 \cdot 2 \dots (N_k - 1)}{(I_{k,1} - 1 + \alpha) \dots (I_{k,N_k} - 1 + \alpha)}, \tag{11}$$

grouping over K groups, the fact that each customer belongs to one table and using the chain rule of probabilities we get

$$p(\Pi_N = \pi_N) = \prod_{k=1}^{K} \frac{\alpha (N_k - 1)!}{(I_{k,1} - 1 + \alpha) \dots (I_{k,N_k} - 1 + \alpha)} = \frac{\alpha^k \prod_{k=1}^{K} (N_k - 1)!}{\prod_{n=1}^{N} (n - 1 + \alpha)}, \tag{12}$$

denoting with I_k the set of indices of customers assigned to the k-th group and N_k the number of customers assigned to the k-th group. Note that Eq. (12) depends only on the block sizes and not the order of arrival of the customers or dishes at tables. Thus, the partition generated by the CRP is exchangeable, with probability partition function symmetric, thus it is an EPPF. □

Connection to the gamma subordinator

The CRP EPPF is connected to the normalized jumps of the gamma subordinator in order of appearance through its EPPF and its stick-breaking representation. First we develop the stick-breaking argument for the CRP and then state the results connecting it to the gamma subordinator. We explicitly follow [4].

The **stick-breaking representation of the CRP** is shown using an argument of urns. The original exchangeability result of de Finetti [5] (Theorem 6, for binary random variables), and can be used in conjunction with the Pólya urn scheme [13], to result in the stick-breaking construction of the CRP.

In the Pólya urn scheme, an urn starts with G_0 gray balls and W_0 white balls. At each round N, we draw a ball from the urn, replace it, and add κ balls of the same color of the ball to the urn. At the end of the round, the urn contains G_N gray balls and W_N white balls. Using Eq. (9) which defines the CRP, we can see that the coloring of gray (all customers sitting at the first table) and white (all the other customers) customer assignments starting with the second customer has the same distribution as a sequence of balls from a Pólya urn with $G_{1,0} = 1$ initial gray balls, $W_{1,0} = \alpha$ initial white balls and $\kappa_1 = 1$ replacement ball. Let $G_{1,N}$ and $W_{1,N}$ be the number of gray and white balls in the urn after N rounds. Then, there exists some $V \sim \text{Beta}(G_0/\kappa, W_0/\kappa)$ such that $\kappa^{-1}(G_{N+1} - G_N) \overset{iid}{\sim} \text{Bernoulli}(V)$ for all N. For the above case, we get $G_{1,N+1} - G_{1,N} \overset{iid}{\sim} \text{Bernoulli}(V_k)$ is distributed according to a Bernoulli r.v. with probability of success if a customer sits at the first table (and the corresponding random variable taking value 1 in the case of success, zero otherwise).

Looking at the sequence of customers sitting at the second and following tables, we condition on customers not sitting at the first table (or equivalently on $G_{1,N+1} - G_{1,N} = 0$). The first customer then sits at the first table, by the CRP. Coloring customers sitting at the second table gray and all others (sitting at tables $3, 4, \dots$) white, we use Eq. (9) to see that customer color assignments follow a Pólya urn scheme with $G_{2,0} = 1$ initial gray balls, $W_{2,0} = \alpha$ initial white balls and $\kappa_2 = 1$ replacement balls. Then, the N-th customer sits at the second table, given that they don't sit at the first, with iid distribution Bernoulli(V_2), $V_2 \sim Beta(1, \alpha)$, resulting in the independence of V_2 from V_1.

Recursively, the N-th customer, conditional on not sitting on the first $K-1$ tables (for $K \geq 1$), sits at the K-th table with iid distribution Bernoulli(V_2), $V_2 \sim \mathrm{Beta}(1, \alpha)$ with V_K independent of the previous $V_1, \dots, V_{K-1}$.

Theorem 8. *[4] The EPPF for partition block membership chosen according to the normalized jumps π_k of the gamma subordinator (Eq. (7)) with parameter α is the CRP EPPF (Eq. (10)).*

Proof. Using Theorem 19 we can find all order derivatives of the Laplace exponent Ψ (which exist), and then calculate the EPPF for the partitions generated with frequencies of the normalized jumps of the gamma subordinator. Specifically, the first order derivative for the gamma subordinator is

$$\Psi'(u) = \int_0^\infty \theta e^{-u\theta} \alpha \theta^{-1} e^{-b\theta} d\theta = \frac{\alpha}{u+b},$$

which yields using $\Psi(0) = 0$

$$\Psi(u) = \theta \mathrm{log}(u+b) - \alpha \mathrm{log}(b).$$

Then the general form of the derivatives of the Laplace exponent become

$$\Psi^{(n)}(u) = (-1)^{n-1} \frac{(n-1)!\alpha}{(u+b)^n},\ n \geq 1.$$

Using the above expression, the general formula of Theorem 19 for the EPPF becomes

$$\begin{aligned} p(N_1, \dots, N_k) &= \frac{(-1)^{n-1}}{(N-1)!} \int_0^\infty u^{N-1} (u+b)^{-\theta} b^\alpha \prod_{k=1}^K \frac{(-1)^{N_k-1}\alpha}{(u+b)^{N_k}} du \\ &= \frac{b^\alpha \alpha^K (-1)^{2(N-K)}}{(N-1)!} \prod_{k=1}^K (N_k-1)! \int_0^\infty b^{-\alpha} x^{N-1} (x+1)^{-N-\alpha} dx \\ &= \alpha^K \Big(\prod_{k=1}^K (N_k-1)! \Big) \frac{\Gamma(N)\Gamma(\alpha)}{\Gamma(N+\alpha)} = \frac{\alpha^K \big(\prod_{k=1}^K (N_k-1)!}{\prod_{n=1}^N (n-1+\alpha)}, \end{aligned}$$

which is the CRP EPPF Eq. (10). □

Theorem 9. *[4] The normalized subordinator jumps (π_k) in order of appearance of the gamma subordinator with concentration parameter α and arbitrary parameter $b > 0$ have the same distribution as the stick-breaking representation of the CRP.*

Proof. Let $T = \sum_k p_k$ be the sum over all the jumps of the gamma subordinator and $T_k = T - \sum_{j=1}^{k} p_j$, the total sum minus the first k elements in order of appearance and note that $T = T_0$. Also, denote $W_k = T_k/T_{k-1}$ and $V_k = 1 - W_k$. Then

$$\pi_k = V_k \prod_{j=1}^{k-1}(1 - V_j) = (1 - \frac{T_k}{T_{k-1}}) \prod_{j=1}^{k-1} \frac{T_j}{T_{j-1}} = \frac{T_{k-1} - T_k}{T_0} = \frac{p_k}{T} = \pi_k.$$

It remains to show that $V_k \overset{iid}{\sim} \mathrm{Beta}(1, \alpha)$, or equivalently $W_k \overset{iid}{\sim} \mathrm{Beta}(\alpha, 1)$ since $V_k = 1 - W_k$. Using Theorem 20 and a change of variables, we can derive the desired result. Specifically, we use the bijection $\{W_1, \dots, \tau\} \to \{\tau_0, \dots, \tau_k\}$ defined by $\tau_k = \prod_{j=1}^{k} W_j$ and the function f of Theorem 20 defined for the gamma process as $f(t) = \mathrm{Gamma}(t|\alpha, b) = b^\alpha \Gamma(\alpha)^{-1} t^{\alpha-1} e^{-bt}$. Then, calculating the Jacobian of the transformation, we get

$$\mathbb{P}(W_1 \in dw_1, \dots, W_k \in d_k, \tau \in dt_0) \propto t_k^{\alpha-1} e^{-bt_0} = t_0^{\alpha-1} e^{bt_0} \prod_{j=1}^{k} w_j^{\alpha-1}.$$

From this expression we get $W_k \overset{iid}{\sim} \mathrm{Beta}(\alpha, 1)$, with W_k being independent of each other and of τ. □

Dirichlet process

The following presentation of the Dirichlet process is followed from [6].

Definition 11. *(Dirichlet process) Let Ψ be a probability space, G be a random distribution over Ψ, H be a distribution over Ψ and α a positive real number. Then, for any finite measurable partition $A_1, \dots, A_r$ of Ψ, G is a Dirichlet process (DP) distributed with base distribution H and concentration parameter α, written $G \sim DP(\alpha, H)$ if*

$$(G(A_1), \dots, G(A_r)) \sim Dir(\alpha H(A_1), \dots, \alpha H(A_r)), \tag{13}$$

for every measurable partition $A_1, \dots, A_r$ of Ψ.

The parameters H and α have the following roles in the definition of the DP. The mean of the DP is the base distribution: for any measurable set $A \subseteq \Psi$ we get $\mathbb{E}(G(A)) = H(A)$, while the concentration parameter scales inversely to the variance: $\mathbb{V}(G(A)) = H(A)(1 - H(A))/(\alpha + 1)$. The larger the α, the smaller the variance, and thus the DP will concentrate more around its mean.

Theorem 10. *[14] Let $G \sim DP(\alpha, H)$ be a Dirichlet process over Ψ and $\psi_1, \dots, \psi_n$ be a sequence of independent draws from G. Also, let $A_1, \dots, A_r$ be a finite measurable partition of Ψ and $n_k = \#\{i : \psi_i \in A_k\}$ be the number of observed values in A_k. Then*

$$(G(A_1), \dots, G(A_r))|\psi_1, \dots, \psi_n \sim Dir(\alpha H(A_1) + n_1, \dots, \alpha H(A_r) + n_r). \tag{14}$$

Since this is true for all measurable partitions of Ψ, the posterior distribution over G is a DP with concentration parameter $\alpha + n$ and base distribution $\frac{\alpha H + \sum_{i=1}^{n} \delta_{\psi_i}}{\alpha+n}$, where δ_{ψ_i} is a point mass located at ψ_i and $n_k = \sum_{i=1}^{n} \delta_i(A_k)$. That is

$$G|\psi_1, \ldots, \psi_n \sim DP(\alpha + n, \frac{\alpha}{\alpha+n} H + \frac{\alpha}{\alpha+n} \frac{\sum_{i=1}^{n} \delta_{\psi_i}}{n}). \tag{15}$$

Predictive distribution and Blackwell-MacQueen Urn scheme

[14]Consider $G \sim DP(\alpha, H)$, and an i.i.d. sequence $\psi_1, \psi_2, \cdots \sim G$. Then, using the conditional mean given the first n observations with G marginalised out in Eq. (15) we get

$$\mathbb{P}(\psi_{n+1} \in A|\psi_1, \ldots, \psi_n) = \frac{1}{\alpha+n}\Big(\alpha H(A) + \sum_{i=1}^{n} \delta_{\psi_i}(A)\Big). \tag{16}$$

The sequence of predictive distributions Eq. (16) for $\psi_1, \psi_n, \ldots$ is called the Blackwell-MacQueen urn scheme [15] and is defined as follows. Each value in Ψ is a unique color, and draws $\psi \sim G$ are balls with the drawn value being the color of the ball. Also, there is an urn with previously observed balls. In the beginning, when the urn is empty, we pick a color from H, i.e. $\psi_1 \sim H$, paint a ball with that color and drop it in the urn. In the $n+1$-th step, we either pick a new color $\psi_{n+1} \sim H$ with probability $\frac{\alpha}{\alpha+n}$, paint a ball with that color and drop it in the urn, or with reach into the urn to pick a random ball, paint a new ball with that color and drop both balls in the urn, i.e. draw ψ_{n+1} from the empirical distribution, with probability $\frac{n}{\alpha+n}$.

Theorem 11. *[14] The Dirichlet process exists.*

Proof. Using Eq. (16) we can generate a distribution over sequences $\psi_1, \psi_2, \ldots$ by drawing each ψ_i given the previous values and denote the joint distribution over $n \geq 1$ observations by

$$P(\psi_1, \ldots, \psi_n) = \prod_{i=1}^{n} \mathbb{P}(\psi_i|\psi_1, \ldots, \psi_{i-1}). \tag{17}$$

Since the sequence of ψ_i is infinitely exchangeable, i.e. for any finite number of elements $\psi_1, \ldots, \psi_n$ the joint distribution over that sequence and any sequence that has been reordered is unchanged. This can be observed via Eq. (16). Then, using de Finetti's Theorem 6, we conclude by identifying that the prior random distribution $P(G)$ in

$$P(\psi_1, \ldots, \psi_n) = \int \prod_{i=1}^{n} G(\psi_i) dP(G)$$

is the Dirichlet process $DP(\alpha, H)$. □

Connection to the Chinese restaurant process

Eq. (16) implies a clustering property of the DP. Assuming that H is smooth, so that repeated values arise from the discreteness of the DP. Denote $\psi_1^*, \ldots, \psi_m^*$ the unique values amongst $\psi_1, \ldots, \psi_n$, and n_k the number of repeats of ψ_k^*. The predictive distribution then is

$$\psi_{n+1}|\psi_1, \ldots, \psi_n \sim \frac{1}{\alpha + n}\Big(aH + \sum_{k=1}^{m} n_k \delta_{\psi_k^*}\Big). \tag{18}$$

That is, ψ_k^* is repeated by ψ_{n+1} with probability proportional to the number of times it has already been observed, while the larger the n_k, the higher the probability that cluster k (the set of ψ_i with identical values ψ_k^*) will grow.

The clusters of the DP are induced in the following way. The unique values of $\psi_1, \ldots, \psi_n$ partition the set of integers $[n] = \{1, \ldots, n\}$ into clusters, such that within every cluster k, the ψ_i take the same value of ψ_k^*, which induces a random partition of $[n]$, which is the CRP (Eq. (9)).

Remark. *Two properties of the Dirichlet process were proven in [7]. The first one is that a realization of the Dirichlet process (a Dirichlet measure) can take the discrete form*

$$P = \sum_{k=1}^{\infty} p_k \delta_{\psi_k}, \tag{19}$$

for p_k the probability assigned to the k-th atom, and ψ_k is the location of that atom. These values are drawn independently from the base distribution H. Thus, the distribution of the DP only has support for atomic distributions with infinite atoms, and zero probability for any non-atomic distribution or distributions of finite atoms.

The second property is one that connects the DP to the CRP. Consider a random distribution drawn from a DP, followed by repeated draws from the random distribution Eq. (19),

$$G \sim DP(\alpha, H),\ \psi_i \sim G,\ i = 1, \ldots, n. \tag{20}$$

The joint distribution of $\psi_{1:n}$ is derived by marginalizing out G:

$$p(\psi_1, \ldots, \psi_n|\alpha, H) = \int \left(\prod_{i=1}^{n} p(\psi_i|G)\right) dP(G|\alpha, H). \tag{21}$$

Under this joint distribution, the ψ_i will share repeated values with positive probability. The structure of these values defines a partitions of the integers $\{1, \ldots, n\}$ and the distribution of this partition is a CRP with parameter α.

Samples from a DP

A useful way in which we can generate samples from a DP is via the stick-breaking construction and, analogously to the Dirichlet distribution, via the gamma process [7].

Dirichlet realizations are characterized by Eq. (13). We are able to characterize the distribution of weights and locations, which will enable us to sample from the DP. The stick-breaking construction of the Dirichlet process is due to [16]. Consider a stick of unit length and iterate the following: break off a random Beta$(1, \beta)$ fraction of what is left of the stick and assign this to a point mass at a location drawn randomly from H.

Theorem 12. *[14] (Stick-breaking DP) For H a base distribution, consider the following construction*

$$\begin{aligned} V_k &\overset{iid}{\sim} Beta(1, \alpha), & \psi_k^* &\overset{iid}{\sim} H \\ p_k &= V_k \prod_{l=1}^{k-1}(1 - V_l), & G &= \sum_{k=1}^{\infty} p_k \delta_{\psi_k^*}. \end{aligned} \tag{22}$$

Then, the resulting distribution $G \sim DP(\alpha, H)$.

Theorem 13. *[7] Consider the weights $\{p_k\}_{k=0}^{\infty}$ of the gamma process and set*

$$D = \sum_{k=0}^{\infty} \pi_k \delta_{\psi_k}, \qquad \pi_k = p_k / \sum_{i=0}^{\infty} p_i, \tag{23}$$

The resulting normalized random measure D is distributed according to a DP(α_0, H_0), where $\alpha_0 = G_0(\Psi)$ and $H_0 = G_0/\alpha_0$.

Remark. *The above procedure is natural and analogous to the simulation of the Dirichlet distribution from gamma random variables. Note that the stick-breaking representation of the DP via Theorem 12, enables us to construct a random measure of the form*

$$G = \sum_{k=1}^{\infty} p_k \delta_{\psi_k^*}, \tag{24}$$

where $\delta_{\psi_k^*}$ is a unit point mass located at ψ_k^*. An equivalent way to describe Eq. (24) is through a subordinator in $[0, 1]$

$$T_s = \sum_{k=1}^{\infty} p_k \mathbb{1}\{\psi_k^* \leq s\}.$$

In that case, the atoms of G are in one to one correspondence with the jumps of a driftless subordinator T_s (using Theorem 4).

3.2 Beta process-Indian buffet process

The beta process originates from [9], while the following two definitions are found in [10].

Definition 12. *(Beta process) A beta process $B \sim BP(c, B_0)$ is an independent increment process with rate measure depending on two parameters: c is a positive function over Ψ that is called the concentration function, and B_0 is a fixed measure on Ψ, called the base measure. Two cases are identified.*

1. *If B_0 is continuous, the rate measure takes the form*

$$\nu(d\theta, d\psi) = c(\psi)p^{-1}(1-p)^{c(\psi)-1}dpB_0(d\psi), \textit{ on } \Psi \times [0,1]. \tag{25}$$

2. *If B_0 is discrete and takes the form $B_0 = \sum_i q_i \delta_{\psi_i}$, then B has atoms at the same locations but with different weights*

$$B = \sum_i p_i \delta_{\psi_i},\ p_i \sim Beta(c(\psi_i)q_i, c(\psi_i)(1-q_i)),\ q_i \in [0,1]. \tag{26}$$

Definition 13. *(Bernoulli process) A Bernoulli process X with a measure B on Ψ, called hazard measure, is written $X \sim BeP(B)$ and is an independent increment process with rate measure*

$$\mu(dp, d\psi) = \delta_1(dp)B(d\psi). \tag{27}$$

Taking two cases for B we have:

1. *If B is continuous, X is a Poisson process with intensity B: $X = \sum_{i=1}^{N} \delta_{\psi_i}$ with $N \sim Poisson(B(\Psi))$ and $\psi_i \overset{indep}{\sim} B/B(\Psi)$.*

2. *If B is discrete and of the form $B = \sum_i p_i \delta_{\psi_i}$, then $X = \sum_i b_i \delta_{\psi_i}$, where $b_i \overset{indep}{\sim} Bernoulli(p_i)$.*

A model in which the underlying number of features isn't known a priori is the Indian buffet process (IBP) [17], which is a generative model for a random feature allocation that can be viewed analogously to the Chinese restaurant process. As the CRP, the name metaphor forms an equivalence between customers and indices n that will be partitioned. "Dishes" now correspond to feature labels, just as they corresponded to partition lebels for the CRP. The major difference is that in the IBP, each customer can try multiple dishes.

Indian buffet process (IBP)

Consider a sequence of customers tasting dishes at an infinite buffet. The first customer enters the buffet and chooses to try a Poisson(α) number of dishes. For $i > 1$, customer i tastes dish k that has previously been sampled from any customer $1, \ldots, n-1$ with probability m_k/i, in which case $Z_{i,k} = 1$ and $Z_{i,k} = 0$ otherwise. Thus $Z_{i,k} \sim$ Bernoulli(m_k/i). Also, customer i tastes a number of new dishes given by a draw from a Poisson(α/i) distribution. This procedure defines a stochastic process called the Indian buffet process.

Theorem 14. *[10, Sections 3 and 4] Let $B \sim BP(c, B_0)$, and let $X_i|B \overset{indep}{\sim} BeP(B)$ for $i = 1, \ldots, n-1$. For $X_{1\ldots n-1}$ equal to the set of observations $\{X_1, \ldots, X_{n-1}\}$, we have the following.*

1. *Conjugacy: The posterior $B|X_{1\ldots n-1}$ is a beta process*

$$B|X_{1\ldots n-1} \sim BP\left(c+n-1, \frac{c}{c+n-1}B_0 + \frac{1}{c+n-1}\sum_{i=1}^{n-1} X_i\right). \tag{28}$$

2. *The underlying de Finetti mixing distribution of the IBP is the beta process, i.e. for $m_{n-1,j}$ the number of customers amongst $X_{1\ldots n-1}$ having tried dish ψ_j, we have*

$$\begin{aligned} X_n|X_{1-1} &\sim BeP\left(\frac{c}{c+n-1}B_0 + \frac{1}{c+n-1}\sum_{i=1}^{n} X_i\right) \\ &= BeP\left(\frac{c}{c+n-1}B_0 + \sum_{j=1}^{n-1} \frac{m_{n,j}}{c+n-1}\delta_{\psi_j}\right) \end{aligned} \tag{29}$$

Proof. The first result is derived by applying Corollary 4.1 of [9]. For the second result, we have the following. Marginalizing out B, we obtain the marginal distribution of X_1. As independence on disjoint intervals is preserved, X_1 is a Lévy process. Since it is a Bernoulli process, its Lévy measure is of the form Eq. (27), so X_1 is still a Bernoulli process. It has expectation $\mathbb{E}(X_1) = \mathbb{E}[\mathbb{E}(X_1|B)] = \mathbb{E}(B) = B_0$, see Appendix [10], so its hazard measure is B_0. We then have

$$\begin{aligned} p(X_n|X_{1\ldots n-1}) &= \int p(X_n|B)P(B|X_{1\ldots n-1})dB \\ &\sim BeP\left(\frac{c}{c+n-1}B_0 + \frac{1}{c+n-1}\sum_{i=1}^{n-1} X_i\right) \\ &= BeP\left(\frac{c}{c+n-1} + \sum_{j} \frac{m_{n-1,j}}{c+n-1}\delta_{\psi_j}\right). \end{aligned} \tag{30}$$

Assume c is a constant and B_0 is continuous with total finite mass $B_0(\Psi) = \gamma$. By sequentially generating $X_{1\ldots n}$ from Eq. (30) we have

1. For $n = 1$: $X_1 \sim BeP(B_0)$ and B_0 is continuous, X_1 is a Poisson process with intensity B_0. In particular, for the whole space, we have $X_1(\Psi) \sim \text{Poisson}(\gamma)$. Thus, the first customer tries a Poisson(γ) number of dishes.

2. For $n \geq 2$: we separate the base measure in Eq. (30) into discrete and continuous parts, and see that $X_n = U + V$, with $U \sim BeP(\sum_j \frac{m_{n-1,j}}{c+n-1}\delta_{\psi_j})$ having an atom at ψ_j with independent probability $m_{n-1,j}/(c+n-1)$ and $V \sim \text{Poisson}(\frac{c}{c+n-1}B_0)$, with a total number of values in the space of $\text{Poisson}(\frac{c\gamma}{c+n-1})$ (which corresponds to the number of new dishes). Let $c = 1, \gamma = \alpha$ and we recover the IBP.

□

Stick-breaking IBP

The IBP with concentration parameter a admits the following stick-breaking representation [10], as seen in the proof of Theorem 14. We follow [4].

In the first round of the IBP, $M_1^+ \sim \text{Poisson}(\gamma)$ features are chosen to contain dish 1. For a dish k, each future customer N tastes dish j with probability $m_{N-1,j}/(\alpha + N - 1)$. Thus, we model the sequence after the first data point as a Pólya urn with $G_{j,0} = 1$ gray, $W_{j,0} = \alpha$ white balls and $\kappa_j = 1$ replacement ball. As there exists a r.v. $V_j \sim \text{Beta}(1, \alpha)$, such that customer N tastes dish j iid across N, as Bernoulli(V_j). The Bernoulli draws conditional on the previous draws are independent across k, and the V_k independent from each other.

For any round N, $M_N^+ \sim \text{Poisson}(\gamma\alpha/(\alpha + N - 1))$ new dishes are sampled, with probability $m_{N-1,j}/(\alpha+N-1)$. Thus, we model the sequence after the n-th customer enters with a Pólya urn with $G_{j,0} = 1$ initial gray and $W_{j,0} = \alpha + n - 1$ initial white balls, $\kappa_j = 1$ replacement balls. So, there exists a r.v. $V_j \sim \text{Beta}(1, \alpha + n - 1)$ such customer N tastes dish j iid across all N, as Bernoulli(V_j).

Using [10, 4], the model for tasting dishes by customer $n = 1, \ldots, N$ is

$$\begin{aligned}
&M_n^+ \overset{indep}{\sim} Poisson\Big(\frac{\gamma\alpha}{\alpha+n-1}\Big), && M_n = M_{n-1} + M_n^+ \\
&M_n = M_{n-1} + M_n^+ && \\
&V_j \overset{indep}{\sim} Beta(1, \alpha+n-1), && j = M_{n-1}+1, \ldots, M_n, \\
&I_{n,j} \overset{indep}{\sim} Bernoulli(V_j), && j = 1, \ldots, M_n,
\end{aligned} \tag{31}$$

where $I_{n,j}$ is an indicator r.v. for whether customer n tastes dish j.

The following theorem, analogously to the theorem connecting distribution of jumps from the gamma subordinator to the CRP sticks, does the same for the beta process subordinator and the sequence of IBP sticks.

Theorem 15. *[4] Generate a feature allocation from a beta process subordinator with Lévy measure given by 8. Then, the sequence of subordinator jumps (p_k), indexed in order of appearance, has the same distribution as the sequence of IBP sticks (V_k) given by Eqs.* (31).

Proof. The theorem will be proven recursively. We need to show that the IBP sticks in Eqs. (31) agree with the total number of dishes and the distribution of the atom weights of the Levy measure

$$\Lambda_n(d\theta)\gamma\alpha\theta^{-1}(1-\theta)^{\alpha+n-1}d\theta,$$

and that the recursion holds.

In the n-th round, we choose dishes with probability equal to their atom weight. Thus, a thinned Poisson process (3) is formed of rate measure $\theta\Lambda(d\theta)$ and has total mass

$$\int_0^1 \theta\Lambda_{n-1}(d\theta) = \frac{\gamma\alpha}{\alpha+n-1} := \gamma_{n-1}.$$

So, the new number of dishes chosen follows a Poisson($\frac{\gamma\alpha}{\alpha+n-1}$) which agrees with Eq. (31). The atom weights are distributed according to

$$\gamma_{n-1}^{-1}\gamma\alpha\theta^{-1}(1-\theta)^{\alpha+n-1-1}d\theta = \text{Beta}(\theta|1,\alpha+n-1),$$

as in Eq. (31).

Lastly, the recursion holds, as sticks that remain are chosen in the event of Bernoulli failure (probability equal to one minus their weight). Thus, the next round of the rate measure is

$$(1-\theta)\gamma\alpha\theta^{-1}(1-\theta)^{\alpha+n-1-1} = \Lambda_n.$$

□

Stick-breaking Beta process

An alternative formulation of the beta process is the following [18]. Let B_0 be a continuous measure on $(\Psi, \mathcal{B})$, where $\mathcal{B}$ is a σ-algebra on Ψ, and $B_0(\Psi) = \gamma$. Also, let α be a positive scalar and define the truncated beta process B_K

$$B_K = \sum_{i=1}^{K} p_i\delta_{\psi_i},\ p_i \overset{iid}{\sim} \text{Beta}\Big(\frac{\alpha\gamma}{K}, \alpha(1-\frac{\gamma}{K})\Big),\ \psi_i \overset{iid}{\sim} B_0. \tag{32}$$

Then, as $K \to \infty$, $B_K \to B$, where $B \sim BP(\alpha, B_0)$. [18] presented the following construction for beta processes based on a notion of stick-breaking.

$$\begin{gathered} B = \sum_{i=1}^{\infty}\sum_{j=1}^{C_j} V_{ij}^{(i)} \prod_{l=1}^{i-1}(1-V_{ij}^{(l)})\delta_{\psi_{ij}}, \\ C_i \overset{iid}{\sim} \text{Poisson}(\gamma),\ V_{ij}^{(l)} \overset{iid}{\sim} \text{Beta}(1,c),\ \psi_{ij} \overset{iid}{\sim} \frac{1}{\gamma}B_0. \end{gathered} \tag{33}$$

This construction can be rewritten [19] in the following way. Specifically, let V_i be i.i.d. Beta$(1, c)$ and let $f(V_{1:i-1}) := \prod_{j<i}(1 - V_j)$. If $T \sim Gamma(1 - i, c)$ then $f(V_{1:i-1}) \overset{d}{=} e^{-T}$. Therefore Eq. (33) is equivalent to

$$\begin{aligned} &B = \sum_{i=1}^{C_1} V_{1j}\delta_{\psi_{1j}} + \sum_{i=2}^{\infty}\sum_{j=1}^{C_i} V_{ij}e^{-T_{ij}}\delta_{\psi_{ij}},\ C_i \overset{iid}{\sim} \text{Poisson}(\gamma), \\ &V_{ij} \overset{iid}{\sim} \text{Beta}(1, c),\ T_{ij} \overset{indep}{\sim} \text{Gamma}(i - 1, c),\ \psi_{ij} \overset{iid}{\sim} \frac{1}{\gamma}B_0. \end{aligned} \tag{34}$$

Theorem 16. *[19, Theorems 1, 2]*

1. *The construction of B given by Eq.* (34) *has an underlying Poisson process.*

2. *The construction given by Eq.* (33) *is that of a beta process with concentration parameter c and base measure B_0.*

Proof. 1. Let $\pi_{1j} := V_{1j}$ and $\pi_{ij} := V_{ij}e^{-T_{ij}}$, $i > 1$. Also, let $B_i := \sum_{j=1}^{C_i} \pi_{ij}\delta_{\psi_{ij}}$ and therefore $B = \sum_{i=1}^{\infty} H_i$. Noting that $C_i \sim$ Poisson$(\mu(\Psi))$, each set of atoms $\{\psi_{ij}\}$ of H_i forms a Poisson process Π^* on Ψ with mean measure B_0. Every atom ψ_{ij} is marked by a weight $\pi_{ij} \in [0, 1]$ and has a probability measure λ_i. By the marked pp Lemma 1, each B_i is characterized by a Poisson process $\Pi_i = \{(\psi_{ij}, \pi_{ij})\}$ on $\Psi \times [0, 1]$ with Lévy measure $B_0 \times \lambda$. Using the superposition Lemma 2, it follows that B is characterized by a countable superposition of Poisson processes $\Pi = \cup_{i=1}^{\infty}\Pi_i$.

2. The goal is to show that the Lévy measure of Π is of the form of the beta process (Eq. (25)) with parameters c and B_0. This will be shown by calculating each Lévy measure and then summing them up as follows.

1. For the first term of B in Eq. (34), consider $B_1 = \sum_{j=1}^{C_1} \pi_{1j}\delta_{\psi_{1j}}$ for $\pi_{1j} := V_{1j}$ which is characterized by a Poisson process $\Pi_1 = \{(\psi_{1j}, \pi_{1j}\}$ of Lévy measure $\nu_1 = B_0 \times \lambda_1$ by Lemma 1, due to the choice of V_{1j} in Eq. (34). Then, $\lambda(d\pi) = c(1 - \pi)^{c-1}d\pi$, with density $f_1(\pi|c) = c(1 - \pi)^{c-1}$.

2. To calculate the density for terms $i > 1$ in Eq. (34), again use Lemma 1 to get that each B_i is characterized by a Poisson process $\Pi = \{(\psi_{ij}, \pi_{ij}\}$ with mean measure $B_0 \times \lambda_i$, with λ_i determining the probability π_{ij}. The measure is written as $\lambda_i(d\pi) = f_i(\pi|c)$, where $f_i(\pi|c)$is the density of π_{ij}, meaning of the i-th break from a Beta$(1, c)$ stick-breaking process. Using the exact form of π_{ij} define the random variables $W_{ij} := e^{-T_{ij}}$ and change variables to get $p_W(w|i, c) = \frac{c^{i-1}}{(i-2)!}w^{c-1}(-\ln w)^{i-1}$. Then, the density $\pi_{ij} = V_{ij}W_{ij}$ can be found to be

$$f_i(\pi|c) = \int_{\pi}^{1} w^{-1}p_v(\frac{\pi}{w}|c)p_W(w|i, c)dw = \frac{c^i}{(i-2)!}\int_{\pi}^{1} w^{c-2}(\ln\frac{1}{w})^{i-2}(1-\frac{\pi}{w})^{c-1}dw.$$

Now, using the fact that

$$\begin{aligned}\sum_{i=2}^{\infty} f_i(\pi|c) &= \sum_{i=2}^{\infty} \frac{c^i}{(i-2)!}\int_{\pi}^{1} w^{c-2}(\ln\frac{1}{w})^{i-2}(1-\frac{\pi}{w})^{c-1}dw \\ &= c^2\int_{\pi}^{1} w^{c-2}\sum_{i=2}^{\infty}\frac{c^{i-2}}{(i-2)!}(\ln\frac{1}{w})^{i-2}(1-\frac{\pi}{w})^{c-1}dw \\ &= c^2\int_{\pi}^{1} w^{c-2}\frac{1}{w^c}(1-\frac{\pi}{w})^{c-1}dw \\ &= \frac{c(1-\pi)^c}{\pi}.\end{aligned} \tag{35}$$

Then, add this result with the first density f_1, to get the wanted density of the Beta process

$$\nu(d\theta, d\pi) = \sum_{i=1}^{\infty}\nu_i(d\theta, d\pi) = B_0(d\theta)\sum_{i=1}^{\infty} f_i(\pi|c) = B_0(d\theta)c\pi^{-1}(1-p)^{c-1}d\pi, \tag{36}$$

which completes the proof.

□

Gamma process

Equivalently to the continuous beta process, the gamma process can be defined, and is shown to admit a stick-breaking representation [20].

Definition 14. *(Gamma process) A gamma process $\Gamma \sim GaP(c, G_0)$ is an independent increment process, with Lévy measure ν*

$$\nu(d\theta, d\psi) = c\theta^{-1}e^{-c\theta}d\theta G_0(d\psi),\ \theta \in [0,\infty) \tag{37}$$

depending on two parameters: the concentration parameter c and base measure G_0.

Theorem 17. *[20] A gamma process $G \sim GaP(c, G_0)$ with positive concentration parameter c and finite base measure G_0 can be constructed through*

$$\begin{aligned} &G = \sum_{i=1}^{\infty}\sum_{j=1}^{C_i} E_{ij}e^{-T_{ij}}\delta_{\psi_{ij}}, \qquad E_{ij} \overset{iid}{\sim} Exp(c), \\ &T_{ij} \sim Gamma(i, a) \qquad C_i \overset{iid}{\sim} Poisson(\gamma), \quad \psi_{ij} \overset{iid}{\sim} \frac{1}{\gamma}G_0. \end{aligned} \tag{38}$$

Proof. The proof follows analogously to the proof of the stick-breaking construction of the beta process. □

Remark. *Note that this is indeed a stick breaking construction, if one decouples the variables $E_{ij} \sim Exp(c)$ into the product of a $Beta(1, c)$ and a $Gamma(\alpha{+}1, c)$ random variables. Then, Eq.* (38) *takes the form*

$$\begin{aligned} G = \sum_{i=1}^{\infty} \sum_{j=1}^{C_i} G_{ij}^i V_{ij}^i \prod_{l=1}^{i} (1 - V_{ij}^l) \delta_{\psi_{ij}}, \qquad & V_{ij}^i \overset{iid}{\sim} Beta(1, \alpha), \\ T_{ij} \sim Gamma(i, a), \qquad C_i \overset{iid}{\sim} Poisson(\gamma), \qquad & \psi_{ij} \overset{iid}{\sim} \frac{1}{\gamma} G_0. \end{aligned} \tag{39}$$

4 Completely random measures

Completely random measures are a family of random measures introduced by Kingman [21]. The term "random measure" can be thought of as a random variable M assigning a measure M_ω to every outcome $\omega \in \Omega$.

Definition 15. *[22, Chapter IV] Let $(E, \mathcal{E})$ be a measurable space. A mapping $M : \Omega \times \mathcal{E} \mapsto \bar{\mathbb{R}}_+$ is a random measure if $\omega \mapsto M(\omega, A)$ is a random variable for each $A \in \mathcal{E}$ and if $A \mapsto M(\omega, A)$ is a measure on $(E, \mathcal{E})$ for each $\omega \in \Omega$.*

As described in Remark , there is an equivalency between random measures and representations in terms of a subordinator without drift. Specifically, for any chosen Levy measure, as long as the condition in Theorem 4 holds, then a representation of the subordinator in terms of CRMs is possible.

Definition 16. *[21] Let Σ_ψ be a σ-algebra of subsets of Ψ, where all singletons are included. A completely random measure (CRM) Θ is a random measure such that for any disjoint, finite collection of measurable sets $A_1, \dots, A_K \in \Sigma_\psi$, the random variables $\Theta(A_1), \dots, \Theta(A_K)$ are independent.*

CRMs can admit a decomposition [21]

$$\Theta = \Theta_{det} + \Theta_{fix} + \Theta_{ord}, \tag{40}$$

into a deterministic component Θ_{det}, one concentrated on a fixed set of atoms Θ_{fix} and one concentrated on a random set of atoms Θ_{ord}, thus (except for a possibly deterministic part) a CRM is purely atomic. Θ_{det} is any deterministic measure, while Θ_{fix} is the fixed-location component, i.e. it is constructed from a set of random weights at fixed locations. That is

$$\Theta_{fix} = \sum_{k=1}^{K_{fix}} \theta_{fix,k} \delta_{\psi_{fix,k}}, \tag{41}$$

for K_{fix} the number of fixed locations (finite or infinite), $\psi_{fix,k}$ deterministic and $\theta_{fix,k}$ a non-negative random variable taking values in $\mathbb{R}$. We will assume that the locations $\psi_{fix,k}$ are distinct, and by the independence assumption of CRMs, the independence of the random variables $\theta_{fix,k}$ over k follows.

We can think of Ψ as a space of "traits", the frequency of the trait represented by ψ_k be $\theta_k \in \mathbb{R}_+$, where k indexes countably many traits. We will formally introduce these notions in Section 4.1.

The ordinary component Θ_{ord} is constructed from a random and countable set of points $\Pi = \{(\theta_{ord,k}, \psi_{ord,k})\}_{k=1}^{K_{ord}}$, yielded by a Poisson point process on $\mathbb{R}_+ \times \Psi$, of rate maeasure $\nu(d\theta \cdot d\psi)$. For $\theta_{ord,k}$ the weight of the atom located at $\psi_{ord,k}$ we construct

$$\Theta_{ord} = \sum_{i=1}^{K_{ord}} \theta_{ord,k} \delta_{\psi_{ord,k}}, \tag{42}$$

where K_{ord} is finite or infinite. Hence the significance of the Poisson process in CRMs is obvious.

4.1 Inference

Bayesian nonparametric priors will be viewed as models for the allocation of data points to traits. The processes that build these priors, such as the Dirichlet process, the gamma process and the beta process, produce pairs of traits together with frequencies with which the traits occur [23]. Thus, Bayesian nonparametric models are seen to be composed of a collection of pairs together with their frequencies, and an allocation to different traits for each data point. This sections follows [23].

Let each trait be ψ belonging to a space of traits Ψ, the frequency of the trait represented by ψ_k be $\theta_k \in \mathbb{R}_+$, where k indexes countably many traits. To represent the collection of pairs of traits with their frequencies, we use the discrete measure on Ψ

$$\Theta = \sum_{i=1}^{K} \theta_k \delta_{\psi_k}, \tag{43}$$

where K may be infinite.

The data point X_n is viewed as a discrete measure

$$X_n = \sum_{i=1}^{K_n} x_{n,k} \delta_{\psi_{n,k}}, \tag{44}$$

where each atom of X_n is a pair of a trait to which the nth individual is allocated and a degree to which this happens. Note that $\psi_{n,k} \in \Psi$ is a trait, $x_{n,k}$ represents the degree to which the nth data belongs to trait $\psi_{n,k}$, and K_n is the total number of traits the nth data belongs to.

For our observed dataset $X_{1:N} = \{X_n : n \in [N]\}$, we want to specify a Bayesian model and need to further specify a prior and a likelihood. A prior distribution for the random measure Θ as well as a likelihood for each random measure $X_n|\Theta$. Σ_ψ is a σ-algebra of subsets of Ψ, where all singletons are included. The random measures Θ, X_n we will consider have values in Ψ, and for any random measure Θ and measurable set $A \in \Sigma_\psi$, $\Theta(A)$ represents a random variable.

From Eqs. (43), (44), it is seen that a distribution on random measures is needed that produces discrete measures and for this reason completely random measures are introduced.

We note that to produce almost surely discrete structures from CRMs, we need to eliminate the deterministic component from our random measures. That is because from Eq.s (41), (42) we can guarantee the a.s. discreteness, while Θ_{fix} will be set to 0, as its' structure can be incorporated into Θ_{fix}. Then $\Theta = \Theta_{fix} + \Theta_{ord}$.

Prior, likelihood and assumptions

We will consider Θ as our parameters, and the prior that we will assume that these follow will be a general CRM (without any deterministic component), with an additional assumption on the rate measure of the ordinary component. Θ has a fixed-location component with K_{fix} atoms, where the kth atom follows a distribution $F_{fix,k} : \theta_{fix,k} \sim F_{fix,k}(d\theta)$ such that $K_{fix} \leq \infty$ and Θ_{ord}, the ordinary component of Θ, is characterized by rate measure $\nu(d\theta \times d\psi)$. The extra assumption to be made is that

$$\nu(d\theta \times d\psi) = \nu(d\theta) \cdot G(d\psi), \tag{45}$$

meaning that the weights of the ordinary component are assumed to be decomposable with respect to the distribution of the locations, for ν any σ-finite, deterministic measure on $\mathbb{R}_+$ and any proper distribution G on Ψ. To generate the ordinary component of Θ, let $\{\theta_{ord}\}_{k=1}^{K_{ord}}$ be the points of a Poisson point process generated on $\mathbb{R}_+$ with rate ν. Then, draw the locations of these points according to G: $\{\psi_{ord}\}_{k=1}^{K_{ord}} \overset{i.i.d}{\sim} G(d\theta)$. Now each atom k in Θ_{ord} is represented as $\theta_{ord,k}\delta_{\psi_{ord,k}}$, and we can shift our focus towards trait frequencies rather that trait locations. Lastly, we will assume that Θ_{ord} has atoms in a.s. distinct locations, which are a.s. distinct from the locations of the fixed atoms.

Recall from Eqs. (43), (44) that

$$\Theta = \sum_{i=1}^{K} \theta_k \delta_{\psi_k},$$

where $K := K_{ord} + K_{fix}$, and ψ_k is a.s. unique.

To specify the likelihood we need to specify how to generate data points X_n given Θ. Assuming that each point X_n is generated i.i.d. given Θ across n, such that from Eq. (44) each point is a CRM with only a fixed-location component given Θ

$$X_n = \sum_{i=1}^{K} x_{n,k} \delta_{\psi_{n,k}}.$$

Each $x_{n,k}$ is drawn independently for all n, k from a distribution H that may take the weight θ_k of Θ at location ψ_k as a parameter

$$x_{n,k} \overset{indep}{\sim} H(dx|d\theta_k). \tag{46}$$

Next, we will specify a number of restrictions needed to form a Bayesian nonparametric model in this context. Firstly, as a feature of Bayesian nonparametrics is that the number of traits represented in some data can grow as the number of data points grows, we need to consider modeling a countable infinity of traits. To achieve this, we need the prior to have a countable infinity of atoms, which can belong to either the fixed-location component or the ordinary component. Fixed-location atoms could represent some known traits, which must be finite and known to begin with, while ordinary component atoms could represent traits that are unknown and whose number could be ever growing. So, the number of fixed-location atoms in Θ should be finite and the ordinary component should contain a countable infinity of atoms.

A1: The number of fixed-location atoms in Θ is finite.

A2: The ordinary component contains a countable infinity of atoms iff $\nu(\mathbb{R}_+) = \infty$.

Furthermore, we need to restrict Eq. (46), meaning the distribution of values of X, such that each data point is allocated to a finite number of traits. Using the previous two assumptions made, we need to ensure that only finitely many such values are nonzero. This can be achieved by enforcing $H(dx|\theta)$ to exhibit an atom at zero.

Assuming that $H(dx|\theta)$ is discrete with support on $\mathbb{Z}_* = \{0, 1, 2, \dots\}$, let $h(x|\theta)$ be the p.m.f. of x given θ. Thus we need the number of atoms of X_n to take values in $\{1, 2, \dots\}$ to be finite. As by construction $\{(\theta_{ord,k}, x_{ord,k})\}_{k=1}^{K_{ord}}$ form a marked Poisson point process with rate measure $\nu_{mark}(d\theta \times dx) := \nu(d\theta)h(x|\theta)$. Also, the pairs $x_{ord,k} = x \in \mathbb{Z}_+$ form a thinned Poisson point process of rate measure $\nu_x(d\theta) := \nu(d\theta)h(x|\theta)$. Specifically the number of atoms of X with weight $x \sim \text{Poisson}(\nu_x(\mathbb{R}_+))$. So, the number of atoms of X is finite if and only if

$$\text{A3: } \sum_{x=1}^{\infty} \nu_x(\mathbb{R}_+) < \infty, \text{ for } \nu_x := \int_{\theta \in \mathbb{R}_+} \nu(d\theta)h(x|\theta).$$

Posterior

Based on the previous setup, assuming we have a general CRM as the prior, and a "finite" CRM for the likelihood, we can derive the posterior that model and it will be a CRM Θ given multiple data-points $X_1, \dots, X_N$ that have been generated i.i.d.

Theorem 18. *[23, Corollary 3.2] Let Θ be a CRM that satisfies the assumptions A1, A2; Θ is a CRM with a finite number K_{fix} of fixed atoms, such that the kth atom can be written as $\theta_{fix,k}\delta_{\psi_{fix,k}}$ with $\theta_{fix,k} \overset{indep}{\sim} F_{fix,k}(d\theta)$, for a proper distribution $F_{fix,k}$ and deterministic $\psi_{fix,k}$. Let the ordinary component of Θ have rate measure which decomposes according to*

$$\nu(d\theta \times d\psi) = \nu(d\theta) \cdot G(d\psi),$$

where G is a proper distribution and $\nu(\mathbb{R}_+) = \infty$. Write $\Theta = \sum_{k=1}^{\infty} \theta_k \delta_{\psi_k}$, and let $X_1, \dots, X_n$ be generated i.i.d. conditional on Θ according to $X = \sum_{k=1}^{\infty} x_{n,k}\delta_{\psi_{n,k}}$ with $x_{n,k} \overset{indep}{\sim} h(x|\theta_k)$ for a proper, discrete probability mass function h. Suppose that X_1 and Θ jointly satisfy assumption A3, so

$$\sum_{x=1}^{\infty} \int_{\theta \in \mathbb{R}_+} \nu(d\theta)h(x|\theta) < \infty.$$

Then, let Θ_{post} be a random measure characterised by the distribution of $\Theta|X_{1:N}$. Θ_{post} is a CRM made of 3 parts.

1. *For $K \in [K_{fix}]$, Θ_{post} has a fixed-location atom at $\psi_{fix,k}$ with weight $\theta_{post,fix,k}$ distributed according to the finite-dimensional posterior $F_{post,fix,k}(d\theta)$ that comes from prior $F_{fix,k}$, likelihood h and observation $X(\{\psi_{fix,k}\})$.*

2. *Let $\{\psi_{new,k}\} \in [K_{new}]$ be the union of atom locations across $X_1, \dots, X_N$ that are not at fixed locations in the prior of Θ. K_{new} is finite by A3. Let $x_{new,k}$ be the wright of the atom in X_n located at $\psi_{new,k}$. Then Θ_{post} has a fixed-location atom at $x_{new,k}$ with random weight $\theta_{post,new,k}$, whose distribution $F_{post,new,k}(d\theta)$ is proportional to*
$$\nu(d\theta)\prod_{n=1}^{N} h(x_{new,n,k}|\theta).$$

3. *The ordinary component of Θ_{post} has rate measure*
$$\nu_{post}(d\theta) := \nu(d\theta)(h(0|\theta))^n.$$

The above theorem characterizes the posterior of Θ, given multiple data points. In [23], the authors are able to derive a theorem for characterizing the posterior in the same setting for a single data point, and then iteratively apply it using Bayes rule to arrive at the above result. In particular, fixed-location points of the posterior are characterised by locations of fixed-location points from the prior with new weights, and from new fixed-location points and weights, that aren't present in the prior. Also, the rate measure of Θ_{post} is characterised, so we arrive at the specification of the posterior.

4.2 Connection from subordinators

The processes that have been presented, such as the beta, gamma and binomial process, can be represented in the form of CRMs [4]. Specifically, a collection of points $\{(\theta_k, \psi_k)\}_{k=1}^{\infty}$ are drawn according to a Poisson point process with rate measure specific to the aforementioned process, and infinite atomic sums form these processes.

The beta process B can be formed from the beta subordinator in the following way. For a continuous base measure B_0, use the Levy measure ν of the beta process, draw points from the PPP with that Levy measure and form the corresponding CRM
$$\nu(d\theta, d\psi) = \gamma\alpha\theta^{-1}(1-\theta)^{\alpha-1}d\theta B_0(d\psi), \qquad \{(\theta_k, \psi_k)\}_{k=1}^{\infty} \sim PPP(\nu)$$
$$B = \sum_{k=1}^{\infty} \theta_k \delta_{\psi_k}.$$

5 Conclusion

The main aim of this dissertation was to highlight the mathematical structure of Bayesian nonparametric models, such as the Chinese Restaurant process, Indian Buffet process and Dirichlet process, and processes underlying them such as the Poisson point process, beta process, gamma process and Bernoulli process, starting from the point of view of Levy processes and specifically subordinators. We expaned on proofs showcasing conjugacy results, building general inference models and showed arguments that connect BNP models to subordinators, heavily relying on the Poisson point process.

Creating flexible probabilistic models to address the problem of clustering is of great importance, especially in the modern era. Further literature includes hierarchical extensions of models, such as for the DP [14, 24], Markov Chain Monte Carlo algorithms for sampling and inference for CRMs [25], variational inference for the beta, gamma process and IBP [26, 27, 28], applied to problems in document classification, topic modelling and recently network theory [29, 30]. We provide further information about the Dirichlet distribution, and technical lemmas used in the Appendix.

6 Appendix

Dirichlet distribution

The Dirichlet distribution was firstly defined by [7], while its stick-breaking representation is found in [16].

Definition 17. *Let $Q = (Q_1, \dots, Q_k)$ be a random probability mass function and suppose that $\alpha = (\alpha_1, \dots, \alpha_k)$ with $\alpha_i > 0$ for every i and $\alpha_0 = \sum_{i=1}^k \alpha_i$. Then, Q follows a Dirichlet distribution with parameter α, denoted by $Q \sim Dir(\alpha)$, if its density is written as*

$$f(y; a) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k q_i^{\alpha_i - 1}, \; q_i \geq 0, \; \sum_{i=1}^n q_i = 1, \tag{47}$$

and zero otherwise.

Stick-breaking

1. Simulate $u_1 \sim \text{Beta}(\alpha_1, \sum_{i=2}^K \alpha_i)$, and set $q_1 = u_1$. This is the first piece of the stick. The remaining has length $1 - u_1$.
2. For $2 \leq j \leq k-1$ pieces, if $j-1$ pieces $u_1, \dots, u_{j-1}$ have been broken off, the remaining length is $\prod_{i=1}^{j-1}(1-u_i)$ and we can simulate $u_j \sim \text{Beta}\Big(\alpha_j, \sum_{i=j+1}^k \alpha_i\Big)$ and set $q_j = u_j \prod_{i=1}^{j-1}(1-u_i)$.
3. The length of the remaining piece is $q_k = \prod_{i=1}^j (1-u_i)$. The remaining stick is $\prod_{i=1}^j (1-u_j) - u_j \prod_{i=1}^{j-1}(1-u_i) = \prod_{i=1}^j (1-u_i)$.

Gamma random variables

The following procedure for $i = 1, \dots, k$ generates samples from the Dirichlet distribution $q = (q_1, \dots, q_k) \sim \text{Dir}(\alpha)$

$$z_i \sim \text{Gamma}(\alpha_i, 1), \qquad q_i = \frac{z_i}{\sum_{j=1}^n z_j}.$$

Poisson point process

The following results for the Poisson process (marking, superpotition) PPP(thinning) are found in [31].

Lemma 1. *Let Π^* be a Poisson process on Ω with mean measure μ. For each $\theta \in \Pi^*$ associate a random variable π drawn independently with probability measure λ on $[0, 1]$. Then, the set $\Pi = \{(\theta, \pi)\}$ is a Poisson process on $\Omega \times [0, 1]$ with mean measure $\mu \times \lambda$.*

Lemma 2. *Let $\Pi_1, \Pi_2, \dots$ be a countable collection of independent Poisson processes on $\Omega \times [0,1]$ and let Π_n have mean measure μ_n for every n. Then, the superposition $\Pi = \cup_{n=1}^{\infty} \Pi_n$ is a Poisson process with mean measure $\mu = \sum_{n=1}^{\infty} \mu_n$.*

Lemma 3. *Suppose that a Poisson point process with intensity function ν generates points with values x. Then, let us keep each point x with probability $h(x)$. The resulting set of poits is also a Poisson point process, with intensity function $\nu'(A) = \int_A \nu(dx)h(x)dx$.*

Pitman theorems

Theorem 19. *(Corollary 6 [32]) Let μ be a probability measure by normalizing jumps of the subordinator with Laplace exponent Φ and (Π_n) be a consistent set of exchangeable partitions induced from iid draws from μ. For each exchangeable partition $\pi_N = \{A_1, \dots, A_K\}$ of $[N]$ with $N_k := |A_k|$ for each k*

$$\mathbb{P}(\Pi_N = \pi_N) = p(N_1, \dots, N_k) = \frac{(-1)^{N-K}}{(N-1)!} \int_0^{\infty} u^{N-1} e^{-\Phi(u)} \prod_{k=1}^{K} \Phi^{N_k}(u) du, \tag{48}$$

where $\Phi^{N_k}(u)$ is the N_k-th derivative of the Laplace exponent Φ calculated at u.

Theorem 20. *[3] Consider a subordinator with Levy measure Λ, jumps p_k. Let τ equal to the sum of all jumps of the subordinator $\tau = \sum_k p_k$ and $\tau_k = \tau - \sum_{j=1}^{k} p_j$. Also, denote ρ the density of Λ with respect to the Lebesgue measure and f the density of the distribution of τ with respect to the Lebesgue measure. Then*

$$\mathbb{P}(\tau_0 \in dt_0, \dots, \tau_k \in dt_k) = f(t_k)dt_k \Big(\prod_{j=0}^{k-1} \frac{(t_j - t_{j+1})\rho(t_j - t_{j+1})}{t_j} dt_j \Big).$$

Bibliography

[1] A. E. Kyprianou, *Fluctuations of Lévy processes with applications: Introductory Lectures.* Springer Science & Business Media, 2014.

[2] M. Winkel, "Lévy processes and finance," *Unpublished Lecture notes. Available at http://www.stats.ox.ac.uk/ winkel/ms3b.html Oxford University, UK*, 2010.

[3] J. Pitman, *Combinatorial Stochastic Processes: Ecole d'Eté de Probabilités de Saint-Flour XXXII-2002.* Springer, 2006.

[4] T. Broderick, M. I. Jordan, and J. Pitman, "Cluster and feature modeling from combinatorial stochastic processes," *Statistical Science*, pp. 289–312, 2013.

[5] B. De Finetti, "Sul significato soggettivo della probabilita," *Fundamenta mathematicae*, vol. 17, no. 1, pp. 298–329, 1931.

[6] Y. W. Teh, "Dirichlet Process.," 2010.

[7] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, pp. 209–230, 1973.

[8] K. Doksum, "Tailfree and neutral random probabilities and their posterior distributions," *The Annals of Probability*, pp. 183–201, 1974.

[9] N. L. Hjort *et al.*, "Nonparametric Bayes estimators based on beta processes in models for life history data," *The Annals of Statistics*, vol. 18, no. 3, pp. 1259–1294, 1990.

[10] R. Thibaux and M. I. Jordan, "Hierarchical beta processes and the Indian buffet process," in *Artificial Intelligence and Statistics*, pp. 564–571, PMLR, 2007.

[11] J. Pitman, "Exchangeable and partially exchangeable random partitions," *Probability theory and related fields*, vol. 102, no. 2, pp. 145–158, 1995.

[12] S. J. Gershman and D. M. Blei, "A tutorial on Bayesian nonparametric models," *Journal of Mathematical Psychology*, vol. 56, no. 1, pp. 1–12, 2012.

[13] G. Pólya, "Sur quelques points de la théorie des probabilités," in *Annales de l'institut Henri Poincaré*, vol. 1, pp. 117–161, 1930.

[14] Y. W. Teh, "A hierarchical Bayesian language model based on Pitman-Yor processes," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 985–992, 2006.

[15] D. Blackwell, J. B. MacQueen, *et al.*, "Ferguson distributions via Pólya urn schemes," *The Annals of Statistics*, vol. 1, no. 2, pp. 353–355, 1973.

[16] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica sinica*, pp. 639–650, 1994.

[17] T. L. Griffiths and Z. Ghahramani, "Infinite latent feature models and the Indian buffet process," in *NIPS*, vol. 18, pp. 475–482, 2005.

[18] J. W. Paisley, A. K. Zaas, C. W. Woods, G. S. Ginsburg, and L. Carin, "A stick-breaking construction of the beta process," in *ICML*, 2010.

[19] J. Paisley, D. Blei, and M. Jordan, "Stick-breaking beta processes and the Poisson process," in *Artificial Intelligence and Statistics*, pp. 850–858, PMLR, 2012.

[20] A. Roychowdhury and B. Kulis, "Gamma processes, stick-breaking, and variational inference," in *Artificial Intelligence and Statistics*, pp. 800–808, PMLR, 2015.

[21] J. Kingman, "Completely random measures," *Pacific Journal of Mathematics*, vol. 21, no. 1, pp. 59–78, 1967.

[22] E. Çınlar, *Probability and stochastics*, vol. 261. Springer Science & Business Media, 2011.

[23] T. Broderick, A. C. Wilson, and M. I. Jordan, "Posteriors, conjugacy, and exponential families for completely random measures," *Bernoulli*, vol. 24, no. 4B, pp. 3181–3221, 2018.

[24] J. Paisley, C. Wang, D. M. Blei, and M. I. Jordan, "Nested hierarchical Dirichlet processes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 2, pp. 256–270, 2014.

[25] P. Zhu, A. Bouchard-Côté, and T. Campbell, "Slice sampling for general completely random measures," in *Conference on Uncertainty in Artificial Intelligence*, pp. 699–708, PMLR, 2020.

[26] J. W. Paisley, L. Carin, and D. M. Blei, “Variational Inference for Stick-Breaking Beta Process Priors.,” in *ICML*, pp. 889–896, Citeseer, 2011.

[27] F. Doshi, K. Miller, J. Van Gael, and Y. W. Teh, “Variational inference for the Indian buffet process,” in *Artificial Intelligence and Statistics*, pp. 137–144, PMLR, 2009.

[28] T. Campbell, J. Straub, J. W. Fisher III, and J. P. How, “Streaming, distributed variational inference for Bayesian nonparametrics,” *arXiv preprint arXiv:1510.09161*, 2015.

[29] D. Cai and T. Broderick, “Completely random measures for modeling power laws in sparse graphs,” *arXiv preprint arXiv:1603.06915*, 2016.

[30] F. Caron and E. B. Fox, “Sparse graphs using exchangeable random measures,” *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, vol. 79, no. 5, p. 1295, 2017.

[31] J. F. C. Kingman, “Poisson processes,” *Encyclopedia of Biostatistics*, vol. 6, 2005.

[32] J. Pitman, “Poisson-Kingman partitions,” *Lecture Notes-Monograph Series*, pp. 1–34, 2003.