A Gentle Introduction to the Dirichlet Process, the Beta Process and Bayesian Nonparametrics

[DRAFT: Please Do Not Distribute]

Michael I. Jordan & Yee Whye Teh

May 20, 2015

Contents

1 Introduction

In this article we aim to provide a gentle introduction to the field of Bayesian nonparametric modeling and inference. The basic motivation for nonparametrics is that in many statistical inference problems we expect that structures or patterns will continue to emerge as data accrue, perhaps ad infinitum, and that when we find ourselves in such situations we may wish to consider a modeling framework that supplies a growing, unbounded number of degrees of freedom to the data analyst. Of course, as in all statistical inference problems, if we allow degrees of freedom to accrue too quickly, we risk finding structures that are statistical artifacts; that is, we will "overfit." This is a serious problem, and it motivates the "Bayesian" aspect of the Bayesian nonparametric programme. While Bayesian inference is by no means immune to overfitting, its use of integration over probability measures as the core methodology does provide a natural resilience to overfitting, one that is suggested by terminology such as "Ockham's razor," "parsimony" and "shrinkage" (Bernardo and Smith, 1994; Jeffreys and Berger, 1992).

Let us emphasize from the outset that "nonparametric" does not mean "no parameters." Rather, it means "not parametric," which has the interpretation that we do not assume a parametric model in which the number of parameters is fixed once and for all. Thus, Bayesian nonparametrics is not opposed to parameters; quite to the contrary, the framework can be viewed as allowing an infinite number of parameters. Another important point to make is that it is quite possible to treat some of the parameters in a Bayesian nonparametric model as classical parameters with a fixed meaning, such as "treatment effect" or "rate of decay." Such models, which blend parametric modeling and nonparametrics, are often referred to as "semiparametric." For example, as we will discuss, one important branch of Bayesian nonparametrics aims to find clusterings in data, with the number of clusters unknown a priori. We can imagine situations in which two heterogeneous populations are being studied, one of which has been exposed to a treatment and other which serves as a control, such that our model contains both a classical parameter characterizing the treatment effect and an open-ended number of parameters for clustering. The nonparametric clustering machinery is able to capture the heterogeneity in the two populations, thereby allowing us to obtain a more precise estimate of the fixed parameter. Finally, note that it usually makes little sense to ask whether a specific model, for a specific fixed number of data points, is parametric or nonparametric. The distinction really refers to our attitude with respect to growing amounts of data.

In the Bayesian approach to statistical inference, parameters are treated as random variables. To obtain models comprising open-ended, potentially infinite numbers of parameters, we require infinite collections of random variables. That is, we require *stochastic processes*. Indeed, Bayesian nonparametrics can be viewed as the branch of Bayesian analysis in which general stochastic processes replace the fixed-dimensional prior distributions of classical Bayesian analysis. The familiar Bayesian process of updating a prior distribution into a posterior distribution via the likelihood becomes the notion of updating a prior stochastic process into a posterior stochastic process. Building a branch of Bayesian analysis on this idea requires not only addressing the mathematical issues that arise in making such updating rigorous, but also discovering classes of stochastic processes that are both useful in describing real-world phenomena and are computationally tractable.

There are two major areas of Bayesian nonparametrics that have seen the most development to date. The first area is built around the *Gaussian process* and involves nonparametric forms of regression, survival analysis and other methods in which the object of inference is a continuous function. The Gaussian process allows such continuous functions to be treated nonparametrically as random functions. Textbook treatments of this area and its applications can be found in Rasmussen and Williams (2006) and Stein (1999). Our focus is on the second major area, in which the random objects of interest are not continuous functions but are discrete, combinatorial objects. As we will see, this area is built around a family of stochastic processes known as *completely random measures*. Examples of completely random measures include the *beta process* and the *gamma process*. Another important stochastic process, the *Dirichlet process*, is obtained by normalizing a completely random measure. From these basic stochastic processes we can obtain distributions on combinatorial objects such as partitions, trees and discrete-valued features. Such random objects can be used to provide Bayesian nonparametric treatments of structural aspects of models.

We aim to provide a treatment of Bayesian nonparametrics that can be read by anyone with an undergraduate-level understanding of probability theory and a minimal exposure to Bayesian statistics. Indeed, we provide a self-contained treatment of core ideas, spelling out the mathematical details in the main text and the appendices. That said, it must be noted that Bayesian nonparametrics is fundamentally a topic that requires a somewhat sophisticated mathematical treatment. Nontrivial ideas from probability theory—such as de Finetti's theorem, Poisson random measures and the Lévy-Khinchin theorem—will make their appearance in our treatment. Anyone wanting to work in the field of Bayesian nonparametrics will eventually need to feel comfortable with such ideas. And yet, Bayesian nonparametrics is a branch of statistics, where the fundamental goal is to solve applied problems. Although our treatment is not an applied one, it is still true that the concerns of applications are never too far below the surface, and our real hope is that readers will make use of the mathematical ideas presented here to find creative solutions to problems that arise in real-world applications.

The article is organized as follows. Part I considers the clustering problem, with the Dirichlet process and the Pitman-Yor process providing the probabilistic underpinnings. We build up to these stochastic processes slowly, starting with the Chinese restaurant process and urn models, introducing exchangeability and de Finetti's theorem and then deriving stick-breaking representations for the Dirichlet process and Pitman-Yor process. We then indulge in a further level of abstraction, introducing the framework of completely random measures and showing how the Dirichlet process can be obtained from this framework. We also discuss hierarchical Dirichlet processes. Part II considers featural representations of objects. In this case we start from the underlying completely random measure (the beta process) and head towards the combinatorial representation (the Indian buffet process). In Part III we discuss the class of normalized completely random measures and in Part IV we present some pointers to further reading. The appendices contain mathematical background and various detailed derivations that support the main text.

Part I: The Dirichlet Process

2 Clustering and Partitions

The Dirichlet process has been a centerpiece of Bayesian nonparametrics since its introduction in a seminal paper by Ferguson (1973). Our aim in Part I is to provide an elementary—but relatively thorough—treatment of this important stochastic process. We do so within an applied context, that of the problem of clustering, where the goal is to discover a set of clusters underlying a data set and to assign each of N data points to exactly one cluster.

The Dirichlet process can be viewed as an infinite-dimensional analog of the classical Dirichlet distribution. This latter distribution plays an important role in Bayesian statistics as a conjugate distribution to the multinomial.[1] For example, a standard Bayesian model for clustering involves assuming that each data point is assigned to one of K clusters, with the assignment to cluster k occurring with probability w_k, for $k = 1, 2, \ldots, K$. We assume that $\sum_{k=1}^{K} w_k = 1$, and place a Dirichlet prior placed on the probabilities $\{w_k\}$. One can attempt to arrive at the Dirichlet process by taking K to infinity in this setup. This is *not* the approach that we take here. Instead we take a combinatorial approach to the clustering problem, treating the problem as one of inferring the partition underlying the data. From a Bayesian point of view this requires placing probability distributions on partitions. We introduce a particular probability distribution on partitions known as the *Chinese restaurant process*, and show that an understanding of the properties of the Chinese restaurant process, most notably its *exchangeability*, lead to the Dirichlet process.

Recall that a *partition* is a set of non-empty subsets of a set of basic entities, such that the subsets are non-overlapping and each entity is contained in exactly one subset. If the entities are data points, denoted $(x_1, x_2, \ldots, x_N)$, where each x_i is a p-dimensional vector, then a partition is often referred to as a *clustering*.

In working towards a probabilistic framework for clustering based on probability distributions on partitions, we find it useful to first consider a deterministic methodology. In particular, let us consider the classical K-means algorithm (MacQueen, 1967). This algorithm is neither Bayesian nor nonparametric, but it will provide us with a useful point of departure. Indeed, it will help us to better understand the motivations for being Bayesian and nonparametric in the setting of clustering.

In K-means clustering, the goal is to assign each data point to one (and only one) of a set of K clusters, where K is assumed fixed and known. Let z_i be an allocation variable that ranges over $\{1, 2, \ldots, K\}$, denoting the cluster assignment for the ith data point. Let $C_k = \{i \ : \ z_i = k\}$ denote the set of indices of the data points forming the kth cluster. Finally, to each cluster we also associate a p-dimensional parameter vector, μ_k, which lives in the same space as the data points and can be viewed as a "prototype" for the kth cluster. The goal of the algorithm is to determine values for the vectors $z = (z_1, z_2, \ldots, z_N)$ and the cluster prototypes $\mu = (\mu_1, \mu_2, \ldots, \mu_K)$. This is done iteratively. Using a superscript t to denote the value of a variable at the tth iteration, which begins at $t = 0$, and initializing

[1] We provide an overview of the Dirichlet distribution in Appendix B, and in particular Dirichlet-multinomial conjugacy is discussed in XXX.

$z^{(0)} = (z_1^{(0)}, z_2^{(0)}, \ldots, z_N^{(0)})$ randomly, the algorithm is as follows:

$$\mu_k^{(t+1)} = \frac{1}{|C_k^{(t)}|} \sum_{i \in C_k^{(t)}} x_i \tag{1}$$

$$z_i^{(t+1)} = \underset{k=1,\ldots,K}{\operatorname{argmin}} \|x_i - \mu_k^{(t+1)}\|, \tag{2}$$

where $|C_k^{(t)}|$ denotes the cardinality of the set $C_k^{(t)}$. (If this cardinality is zero, then we simply set $\mu_k^{(t+1)}$ to zero). Eq. (1) updates the means for each cluster to be the average of the points currently assigned to the cluster, and Eq. (2) updates the assignments by assigning each point to the nearest cluster.

In this algorithmic description of a clustering procedure there is no role for probabilities, and as we move towards a probabilistic framework it is useful to step back and first consider why we should do so. One answer is that probabilities arise naturally in the *analysis* of K-means and other clustering algorithms. In particular, although we do not think it likely that K-means would work well on all possible data sets, is it likely to work well on "most" data sets? Quantifying this generally involves putting a probability measure on data sets. Moreover, if we begin to consider such probability measures as reflective of an underlying population from which the data arise, we can begin to talk about a broader range of inference problems, such as the prediction problem of assigning new data points to existing clusters. Finally, some of the qualitative aspects of probability can provide insight; as we will talk about below, it is natural to make an "exchangeability" assumption for clustering, in which the probability of a data set is invariant to the ordering of the data points. If this is the case, then there is a theorem (de Finetti's theorem) that encourages us to seek additional probabilistic structure in the problem.

Once probability has entered into the problem via the analysis of the algorithm, it is then natural to allow the probabilities to also enter into the design of the algorithm. In particular, it is natural to consider replacing each of the K parameter vectors μ_k with a probabilistic model, $p(x_i \,|\, \mu_k)$. For example, if this model is a Gaussian with mean μ_k and unit variance, then the density $p(x_i \,|\, \mu_k)$ is a simple monotone function of the Euclidean distance $\|x_i - \mu_k\|$ employed by K-means. In general, using probability models allows us to be creative in defining the "distance" function used in clustering.

Having motivated the use of probabilities—and therefore headed in a Bayesian direction—what about nonparametrics? Clearly, it would be desirable to remove the assumption that K is known a priori. Although there are various ways to do this, in many problems it is natural to imagine new clusters arising as we collect more data; moreover, we may want to model the growth rate of new clusters. In the following section we begin our discussion of the Bayesian nonparametric approach to achieving this goal.

3 Clustering via the Chinese Restaurant Process

K-means exemplifies an algorithmic approach to clustering where one specifies a procedure that performs clustering and then provides a theoretical analysis of the procedure. The Bayesian approach is somewhat more indirect. Bayesian methods are model-based methods,

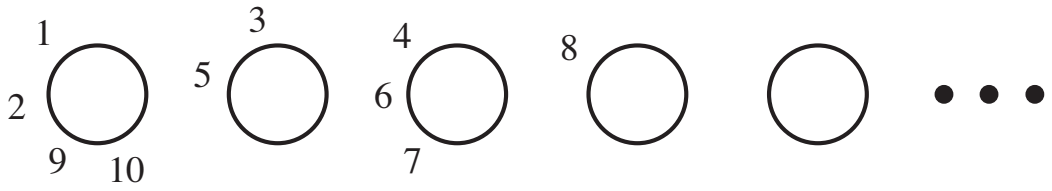

Figure 1: A depiction of a possible state of the Chinese restaurant process after ten customers have arrived.

in that they first specify a model by which the data are assumed to be generated. The clustering algorithm arises as a procedure for computing posterior probabilities under this model.

Thus we begin our discussion of Bayesian nonparametric clustering by specifying a model by which the data are assumed to be generated. We do this in stages, first concentrating on the core of the clustering problem—the partitioning of the data into disjoint subsets. Later we will discuss how to go beyond the partition to provide probabilities for the observed data points given the partition.

3.1 The Chinese restaurant process

Let us denote a partition of N points as $\boldsymbol{\pi}_{[N]}$. Recall that this is a set of subsets of the N points, where each point belongs to exactly one subset. Thus, for example, if the points are the integers one through ten, a possible partition is $\boldsymbol{\pi}_{[10]} = \{\{3,5\},\{1,2,9,10\},\{4,6,7\},\{8\}\}$. We refer to the subsets as *clusters*. Note that the ordering of the subsets and the ordering of points within subsets is arbitrary.

The *Chinese restaurant process* (CRP) is a probability distribution on partitions. The distribution is built up in a sequential manner, where one point at a time is added to an existing set of clusters. The CRP describes this process using the metaphor of a restaurant, with points corresponding to customers and clusters corresponding to tables. Customers arrive at the restaurant one at a time. The first customer is seated alone. Each subsequent customer is either seated at one of the already occupied tables, with probability proportional to the number of customers sitting at the table, or, with probability proportional to a fixed constant, α, the customer starts a new table. Consider, for example, the configuration shown in Fig. (1), where after ten customers have arrived the seating pattern is $\{\{3,5\},\{1,2,9,10\},\{4,6,7\},\{8\}\}$. If we set $\alpha = 1$, the eleventh customer joins customers 3 and 5 with probability 2/11, joins customers 1, 2, 9, and 10 with probability 4/11, and starts a new table with probability 1/11.

The tables are to be viewed as unordered, and to avoid introducing labels that suggest an ordering, we refer to a table by the subset of customers sitting at the table. In particular, we use the same symbol, c, to refer to either a cluster or a table, and let $|c|$ denote the cardinality of the cluster c as well as the number of customers sitting at table c. With this

notation, we write the probabilistic rule characterizing the CRP as follows:

$$P(\text{customer } n+1 \text{ joins table } c \,|\, \boldsymbol{\pi}_{[n]}) = \begin{cases} \dfrac{|c|}{\alpha+n} & \text{if } c \in \boldsymbol{\pi}_{[n]}, \\ \dfrac{\alpha}{\alpha+n} & \text{otherwise.} \end{cases} \tag{3}$$

Note that we can think of there being an infinite number of unlabeled tables in the restaurant at any given point in time, and when a customer is assigned to a new table, one of the unlabeled tables is chosen arbitrarily. In particular, this rule applies to the first customer.

After N customers have arrived, their seating pattern defines a set of clusters and thus a partition. This partition is random, and thus the CRP defines a distribution on partitions. We denote this distribution as follows:

$$\boldsymbol{\pi}_{[N]} \sim \mathrm{CRP}(\alpha, N). \tag{4}$$

As suggested by our notation, the CRP is a family of distributions, one for each value of N.

Although the CRP is specified using an ordering of the customers, it turns out that the distribution on partitions defined by the CRP is invariant to the ordering, in the sense that it is only the size of the clusters that matters in determining the probability of the partition, not the identities of the specific customers forming the clusters. This property is known as *exchangeability*, and it will play an essential role in our development. As an example, let $\alpha = 1$, and consider the probability that customers 1 and 2 will be found sitting together at the same table after N customers have entered the restaurant. This probability is $1/2$—customer 1 sits at an arbitrary table and customer 2 joins customer 1 with probability $1/2$. Now, by exchangeability, this probability doesn't change if the customers were to enter the restaurant in a different order. Put differently, the probability of any two customers i and j sitting at the same table is $1/2$.

We can prove the exchangeability of the CRP by computing explicitly the probability of a partition under the CRP. Consider a partition $\boldsymbol{\pi}_{[N]}$ composed of K clusters, $c_1, c_2, \ldots, c_K$. The probability of forming such a partition is obtained from the conditional probabilities in Eq. (3). Each customer contributes one factor and the overall probability is a product of these factors. Consider, for example, the following sequence of events: customer 1 starts a new table, customer 2 joins customer 1, customer 3 sits at a new table, customer 4 joins customer 3, customer 5 joins 1 and 2, and then customer 6 sits at a new table. We have:

$$P(\{\{1,2,5\},\{3,4\},\{6\}\}) = \left(\frac{\alpha}{\alpha}\right)\left(\frac{1}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right)\left(\frac{1}{\alpha+3}\right)\left(\frac{2}{\alpha+4}\right)\left(\frac{\alpha}{\alpha+5}\right). \tag{5}$$

In general, as we proceed through N customers under the CRP model, the denominators simply increment by one. For every new table that is started we obtain a factor of α, so if at the end K clusters are present, we obtain a factor of α^K. Finally, for each cluster c that is present at the end, each time a customer is added to the table the numerator is simply the number of current customers at the table, so that overall we obtain a factor of $1 \cdot 2 \cdot \cdots \cdot (|c|-1) = (|c|-1)!$. Putting this all together, we obtain:

$$P(\boldsymbol{\pi}_{[N]}) = \frac{\alpha^K}{\alpha^{(N)}} \prod_{c \in \boldsymbol{\pi}_{[N]}} (|c|-1)!, \tag{6}$$

where $\alpha^{(N)} := \alpha(\alpha+1)\cdots(\alpha+N-1)$. This equation shows that the probability of a partition is a function only of the sizes of the clusters forming the partition. Thus, if we relabel the customers we obtain new clusters, but the set of cluster sizes does not change and therefore neither does the probability. This establishes the exchangeability of the CRP.

We can also turn the argument around and recover the CRP update rule in Eq. (3) from the partition probability in Eq. (6). Let $\boldsymbol{\pi}'_{[n+1]}$ be obtained from $\boldsymbol{\pi}_{[n]}$ by adding customer $n+1$ to table c. We have:

$$\begin{aligned} P(\text{customer } n+1 \text{ joins table } c \,|\, \boldsymbol{\pi}_{[n]}) &= \frac{P(\boldsymbol{\pi}'_{[n+1]})}{P(\boldsymbol{\pi}_{[n]})} \\ &= \frac{\alpha^{(n)}}{\alpha^{(n+1)}}|c| \\ &= \frac{|c|}{\alpha+n}, \end{aligned} \tag{7}$$

where the first equality follows from the fact that the event that $\boldsymbol{\pi}'_{[n+1]}$ is the partition after $n+1$ customers have arrived logically implies the event that $\boldsymbol{\pi}_{[n]}$ is the partition after n customers have arrived. If instead we let $\boldsymbol{\pi}'_{[n+1]}$ be obtained from $\boldsymbol{\pi}_{[n]}$ by placing customer $n+1$ at a new table, which we again denote by c, we have:

$$\begin{aligned} P(\text{customer } n+1 \text{ starts new table } c \,|\, \boldsymbol{\pi}_{[n]}) &= \frac{P(\boldsymbol{\pi}'_{[n+1]})}{P(\boldsymbol{\pi}_{[n]})} \\ &= \left(\frac{\alpha^{K+1}}{\alpha^{(n+1)}}\right)\left(\frac{\alpha^{(n)}}{\alpha^{K}}\right) \\ &= \frac{\alpha}{\alpha+n}, \end{aligned} \tag{8}$$

where we use $(|c|-1)! = (1-1)! = 0! = 1$, when c is a new table. We thus recover the two cases of Eq. (3).

Exchangeability is a natural property for the clustering of data. Indeed, many algorithms for clustering, including K-means, are invariant to the ordering of the data points (indeed, if such invariance is broken in the implementation of the algorithm this is generally viewed as undesirable). From a probabilistic point of view, such algorithms can be viewed as making an implicit assumption of exchangeability. Note also that although we are focusing on CRP for didactic reasons, there are other distributions on partitions that are exchangeable. Indeed, a significant line of research focuses on characterizing general "exchangeable probability partition functions" (EPPFs), of which Eq. (6) is an example (Pitman, 2006). We will see another example of an EPPF in Sec. (7).

3.2 The CRP mixture model

The CRP provides us with a vocabulary for talking about probabilities and partitions but leaves us short of a model for generating data points. To take the next step, we recall that the data points in clustering problems generally have a representation as points in a vector space (or some other metric space). Distances between the points in this space play an essential role in algorithms for determining a clustering. K-means exemplifies one way in

which this is done: each of the K clusters is represented by a vector μ_k that is embedded in the same space as the data, and the distances $\|x_i - \mu_k\|$ are used to determine the assignment of points x_i to clusters. We would like to do something similar in a model-based framework.

Let $x = (x_1, x_2, \ldots, x_N)$ denote the observed data. Let us treat the data points as customers in the CRP by identifying data point x_i with its index i. Thus we say that data point x_i sits at table c if $i \in c$. To each table $c \in \boldsymbol{\pi}_{[N]}$ we assign a *parameter vector* ϕ_c and we assume that the data points at table c are generated independently from a common probability distribution indexed by ϕ_c. Thus, for $i \in c$, we let $f(x_i \,|\, \phi_c)$ denote the probability density for generating data point x_i and we take the product over $i \in c$ to obtain the total probability of generating the data associated with table c. Finally, the overall conditional probability of the data (given the parameters $\{\phi_c\}$ and the partition $\boldsymbol{\pi}_{[N]}$) is the product over clusters and over data points within clusters:

$$p(x \,|\, \phi, \boldsymbol{\pi}_{[N]}) = \prod_{c \in \boldsymbol{\pi}_{[N]}} \prod_{i \in c} f(x_i \,|\, \phi_c), \tag{9}$$

where $\phi = (\phi_1, \ldots, \phi_K)$. Viewed as a function of ϕ and $\boldsymbol{\pi}_{[N]}$ for fixed x, this probability density is known as the *likelihood function*.

To complete the probability model, we need to specify a distribution for the parameters ϕ. Let us simply assume that these parameters are drawn independently (across the tables) from a distribution G_0. Putting together the pieces we obtain the following model for generating data points:

$$\boldsymbol{\pi}_{[N]} \sim \mathrm{CRP}(\alpha, N) \tag{10}$$

$$\phi_c \,|\, \boldsymbol{\pi}_{[N]} \overset{iid}{\sim} G_0, \qquad \text{for } c \in \boldsymbol{\pi}_{[N]}, \tag{11}$$

$$x_i \,|\, \phi, \boldsymbol{\pi}_{[N]} \overset{ind}{\sim} F(\phi_c), \qquad \text{for } c \in \boldsymbol{\pi}_{[N]},\ i \in c. \tag{12}$$

where the notation "$\overset{iid}{\sim}$" means that the draws are assumed to be independent and identically distributed, where "$\overset{ind}{\sim}$" means that the draws are assumed to be independent and where $F(\phi_c)$ is the distribution with density $f(\cdot \,|\, \phi_c)$. These linked conditional probabilities yield a joint probability distribution on the collection of variables $(x, \phi, \boldsymbol{\pi}_{[N]})$. As we discuss in the following section, Bayesian inference can then be invoked to obtain various posterior probabilities of interest, in particular the probability of $\boldsymbol{\pi}_{[N]}$ given x, which serves as a Bayesian clustering procedure.

We will refer to the model specification in Eq. (12) as a *CRP mixture model*. In general, a mixture model is a probability model in which each data point is generated from one of a set of "mixture components," and the choice of mixture component is made randomly for each data point. In our case, the choice of ϕ defines the mixture components, the choice of $\boldsymbol{\pi}_{[N]}$ selects randomly among the mixture components (by choosing a table at which to seat each data point), and a data point is generated from the selected mixture component via the draw from $F(\phi_c)$.

Let us make two additional comments before turning to a description of a clustering procedure based on the CRP mixture model. First, it is important to note that although we have begun to use the word "parameter," our framework is definitely a nonparametric

one. In particular, as N grows, the number of clusters grows (at rate $O(\log N)$, as we discuss in Sec. (7)), and therefore the number of parameter vectors grows. We are not in a parametric situation in which the number of parameter vectors is fixed once and for all. Second, in writing down the probability model, several assumptions of independence were introduced without comment, and we need to provide some justification for these assumptions. This requires us to delve more deeply into the notion of exchangeability, a task that we undertake beginning in Sec. (4).

3.3 Gibbs sampling based on the CRP mixture model

In general, the Bayesian approach to statistical inference is based on the use of Bayes' theorem to compute the probability of a parameter of interest given the data. That is, given a parameter θ, a likelihood, $p(x \mid \theta)$, and a prior, $p(\theta)$, inferences are based on the *posterior probability*:

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{\int p(x \mid \theta)p(\theta)d\theta}. \tag{13}$$

Often one is interested in only a subset of the components of the vector θ; the components that are not of inferential interest are marginalized out of the posterior probability. That is, if $\theta = (\theta_1, \theta_2)$, and only θ_2 is of inferential interest, then we would form the marginal posterior: $p(\theta_2 \mid x) = \int p(\theta_1, \theta_2 \mid x)d\theta_1$.

In our setting, the overall "parameter" is the pair $(\phi, \boldsymbol{\pi}_{[N]})$.[2] While in some cases we might be interested in inferring both ϕ and $\boldsymbol{\pi}_{[N]}$, the main goal of clustering is often that of finding the partition. We thus focus on the problem of computing the posterior probability of $\boldsymbol{\pi}_{[N]}$ given the data x, marginalizing over the vector ϕ.

There are two problems that arise in attempting to use Eq. (13) to derive an inference algorithm for clustering. The first is that $\boldsymbol{\pi}_{[N]}$ is a discrete structure and our use of probability densities in Eq. (13) needs some interpretation. Here the interpretation is easily provided, because $\boldsymbol{\pi}_{[N]}$ ranges over a finite set and for each value of $\boldsymbol{\pi}_{[N]}$ there is a well-defined density on ϕ. (The issue is worth raising because in our later work this issue will become more problematic, due to the fact that the objects of interest will often not range over finite sets.) The second issue is the more serious one: it is infeasible to compute the posterior probability in Eq. (13). The integral in the denominator involves a sum over all possible partitions, the number of which (known as the Bell number) grows at a rate of $O(N^N)$.

The standard response to this problem in Bayesian statistics is to make use of sampling-based frameworks such as Markov chain Monte Carlo (MCMC), importance sampling and sequential Monte Carlo to approximate the posterior probability (see, e.g., Gilks et al., 1996). All of these frameworks have indeed been applied to Bayesian nonparametric inference problems, including the CRP-based clustering problem that is our focus here. Specifically, within the MCMC framework, Gibbs sampling and Metropolis-Hastings algorithms have been explored for Bayesian nonparametric clustering (Neal, 2000). In this section we

[2] We have put "parameter" in quotes to again recall our earlier discussion contrasting "parametric" and "nonparametric." But, given that the number of data points are being viewed as fixed in this section, the distinction becomes less essential than before, and henceforth we will no longer use the quotes.

present a particular example of a Gibbs sampling algorithm. Our choice of algorithm is based mainly on its simplicity—as we will see, it can be derived easily given the tools that we have developed thus far, and it has a simple intuitive interpretation. But we hasten to note that there are a variety of competing algorithms that may be better in various practical situations. We will present some of these algorithms later, after the necessary mathematical groundwork has been laid. We will also, in Sec. (**??**), provide pointers to literature on inference algorithms, where more systematic treatments of inference algorithms are provided than we wish to provide in this paper.

The basic idea of Gibbs sampling is to perform a kind of stochastic "hill-climbing" or "coordinate ascent," where we consider one variable at a time and sample a new value of that variable while keeping all other variables fixed. Thus at each step we sample from a conditional probability. In our setting, the variables of interest are the cluster assignments of each of the data points. To implement a classical Gibbs sampler we need to sample the cluster assignment of a given data point while holding fixed the cluster assignments of all other data points. However, rather than explicitly introducing indicator variables to denote cluster assignments, we will instead work directly in the space of partitions, defining our sampler via "hill-climbing" steps in this space.

The prior distribution and the likelihood for our problem are provided by the set of conditionals in Eq. (12). Multiplying these conditionals we obtain the overall joint probability density:

$$p(\boldsymbol{\pi}_{[N]}, \phi, x) = \frac{\alpha^K}{\alpha^{(N)}} \prod_{c \in \boldsymbol{\pi}_{[N]}} \left((|c| - 1)!\ g_0(\phi_c) \prod_{i \in c} f(x_i \,|\, \phi_c) \right), \tag{14}$$

where for simplicity we assume that G_0 has a probability density that we denote by g_0. The overall posterior is proportional to this joint probability density.

Given that our interest is in marginal posterior $p(\boldsymbol{\pi}_{[N]} \,|\, x)$, which is proportional to $p(\boldsymbol{\pi}_{[N]}, x)$, the first step is to integrate out ϕ. Given the independence of the components of ϕ, we obtain:

$$p(\boldsymbol{\pi}_{[N]}, x) = \frac{\alpha^K}{\alpha^{(N)}} \prod_{c \in \boldsymbol{\pi}_{[N]}} (|c| - 1)!\ f(x_c), \tag{15}$$

where we have defined $x_c := (x_i : i \in c)$, and where

$$f(x_c) = \int \left(\prod_{i \in c} f(x_i | \phi_c) \right) g_0(\phi_c) d\phi_c \tag{16}$$

is the marginal probability density of the data points associated with cluster c.

We will assume that the integral in Eq. (16) can be performed analytically. This can be done in particular when the prior g_0 is *conjugate* to the likelihoods $f(x_i|\phi_c)$. There exist conjugate priors for many widely-used distributions; in particular for likelihoods in the exponential family there exist conjugate priors (Bernardo and Smith, 1994). That said, one may not want to use a conjugate prior in a given situation, and in that case, the algorithm that we present in this section cannot be used. (See Sec. (**??**) for references to papers that treat inference in the non-conjugate setting.)

We can now derive a Gibbs sampler. To lighten our notation, we will drop the subscript "$[N]$" in denoting the partition $\boldsymbol{\pi}_{[N]}$. The state space of the sampler is the set of partitions $\Pi_{[}N]$ of the items $[N]$. Given the current state $\boldsymbol{\pi}$, we consider a single index i, either chosen at random or according to a fixed order, and consider the set of neighboring partitions obtained by reassigning the index i. Let $\boldsymbol{\pi}_{-i}$ denote the partition of the set $[N]\backslash\{i\}$ obtained by removing reference to the index i in $\boldsymbol{\pi}$. If i is reassigned to an existing cluster, $c \in \boldsymbol{\pi}_{-i}$, then the resulting partition will be $\boldsymbol{\pi}^+ = \boldsymbol{\pi}_{-i} - c + (c \cup \{i\})$.[3] If instead item i is reassigned to a new cluster not already in $\boldsymbol{\pi}_{-i}$, then the new value of $\boldsymbol{\pi}^+$ will be $\boldsymbol{\pi}_{-i} + \{i\}$. The probabilities of these choices are given by the joint density Eq. (15), so that:

$$p(\boldsymbol{\pi}^+ \,|\, x) \propto \begin{cases} \dfrac{\alpha^K}{\alpha^{(N)}} (|c|)!\; f(x_{c\cup\{i\}}) \displaystyle\prod_{\substack{c' \in \boldsymbol{\pi}_{-i} \\ c' \neq c}} (|c'|-1)!\; f(x_{c'}) & \text{for } \boldsymbol{\pi}^+ = \boldsymbol{\pi}_{-i} - c + (c \cup \{i\}), c \in \boldsymbol{\pi}_{-i}, \\ \dfrac{\alpha^{K+1}}{\alpha^{(N)}} f(x_i) \displaystyle\prod_{c' \in \boldsymbol{\pi}_{-i}} (|c'|-1)!\; f(x_{c'}) & \text{for } \boldsymbol{\pi}^+ = \boldsymbol{\pi}_{-i} + \{i\}, \end{cases}$$

where the proportionality comes from the denominator, $p(x)$, which is common to both cases. We now cancel all of the factors that appear in both of these equations, as it is only the relative values of the probabilities that matter. In doing so, we retain a common denominator factor of $\alpha + N - 1$ that helps—as we will show—with interpretability. This yields:

$$\propto \begin{cases} \frac{|c|}{\alpha+N-1} f(x_i \,|\, x_c) & \text{for } \boldsymbol{\pi} = \boldsymbol{\pi}_{-i} - c + (c \cup \{i\}), c \in \boldsymbol{\pi}_{-i}, \\ \frac{\alpha}{\alpha+N-1} f(x_i) & \text{for } \boldsymbol{\pi} = \boldsymbol{\pi}_{-i} + \{i\}, \end{cases} \tag{17}$$

where $f(x_i \,|\, x_c) = f(x_{c\cup\{i\}})/f(x_c)$ is the conditional probability of x_i under cluster c which currently contains data points x_c.

The pair of equations in Eq. (17) defines a Gibbs sampling algorithm. This algorithm was originally proposed by MacEachern (1994) and Neal (1992); the derivation that we have presented here, based on the EPPF of the Chinese restaurant process, is streamlined relative to the derivations presented by those authors.

The Gibbs sampler in Eq. (17) has a simple intuitive interpretation. Consider the Nth data point, x_N. From the point of view of the CRP prior, this data point is the last to arrive in the restaurant, and it sits at an existing table c with probability proportional to $|c|$, or starts a new table with probability proportional to α. These prior probabilities are multiplied by marginalized likelihoods, either the likelihood $f(x_N \,|\, x_c)$ associated with an existing table, or the likelihood $f(x_N)$ associated with a new table, to form the conditional probabilities of the data point sitting at each table while accounting for the similarity of x_N with the other data points currently sitting at each cluster, as defined by the likelihood model.

Now consider the ith data point, for $i \neq N$. It might seem that our simple argument would break down, because the Gibbs sampler needs to consider all other data points,

[3] Recall that the partition $\boldsymbol{\pi}$ is a set of sets, and we use the notation "+" and "−" to denote the addition and removal of sets from the partition.

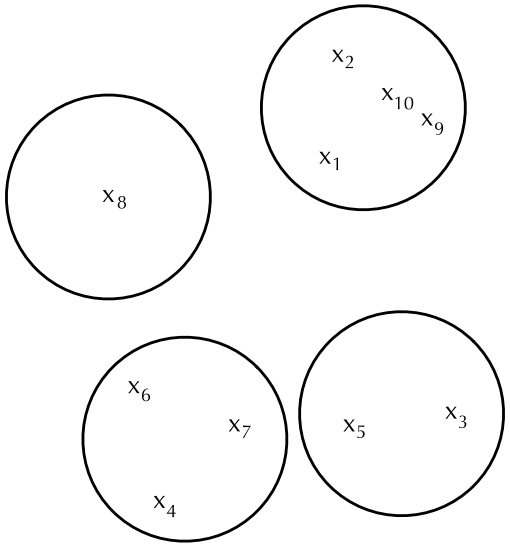

Figure 2: Gibbs sampling example.

and the CRP prior involves only the previous data points. But exchangeability comes to the rescue. By exchangeability, the joint probability is invariant to permutation, and thus the conditional for any given data point given the other data points is the same as the conditional for the last data point given the previous data points. Thus the overall Gibbs sampling algorithm in Eq. (17) has the same simple form for all data points.

As a concrete example, consider a two-dimensional clustering problem where we assume that the component density, $f(x_i \,|\, \phi_c)$, is given by a Gaussian distribution, with ϕ_c the mean vector and with the covariance matrix assumed to be a fixed constant matrix $\sigma^2 I$, with σ^2 known:

$$f(x_i \,|\, \phi_c) = \frac{1}{2\pi\sigma} \exp\left(-\frac{1}{2\sigma^2}\|x_i - \phi_c\|\right). \tag{18}$$

If we place a conjugate Gaussian prior on ϕ_c, then the marginal distributions $f(x_i)$ and $f(x_i \,|\, x_c)$ will also be Gaussian. Suppose that there are ten data points, and that at a given moment the state of the Gibbs sampler is $\boldsymbol{\pi} = \{\{3,5\},\{1,2,9,10\},\{4,6,7\},\{8\}\}$ (see Fig. (2)). Suppose that we now wish to reassign data point x_5, which is currently clustered with data point x_3. There is pressure from the CRP prior to move this data point to the larger clusters. This is balanced by the pressure from the likelihood to move the data point to a nearby cluster, where "nearby" is defined by the exponential of the negative Euclidean distance. As shown in the figure, data point x_5 is nearby the cluster containing $\{x_4, x_6, x_7\}$ (i.e., it has high probability under the posterior defined by those points), and overall it is likely that the point will be moved to that cluster.

It is interesting to compare this Gibbs sampling algorithm to the K-means algorithm. Both algorithms are based on the Euclidean distance. But where the K-means places a data point in a cluster based solely on the Euclidean distance between the point and a cluster centroid, the Gibbs sampler makes use of the CRP prior, favoring the growth of large

clusters and not requiring a fixed choice of K. Moreover, the Gibbs sampler makes use of an integrated notion of "distance," via the factor $f(x_i \mid x_c)$ which is computed as an integral over ϕ_c. The major difference between these ideas, however, is in the kinds of generalizations that they inspire. As we will see, the Bayesian nonparametric framework provides natural upgrade paths, via different choices of mixture components, different stochastic process priors and the use of hierarchical modeling concepts.

4 Urn Models and Exchangeability

In this section and the next several sections our goal is to provide a deeper mathematical understanding of the particular Bayesian nonparametric approach to clustering that has been our focus thus far. This understanding will culminate in the presentation of several important concepts: stick-breaking, the Dirichlet process and completely random measures. There are three major motivations that we have in mind in embarking on this discussion. First, in developing the CRP mixture model we made several conditional independence assumptions. We would like to justify these assumptions, in particular by tying them to properties of the observed data. Such justifications are essential when trying to understand when the CRP mixture model is appropriate for practical data analysis problems and when it is not. Second, mathematial concepts such as stick-breaking, the Dirichlet process and completely random measures provide an abstract understanding of clustering, and as such they provide a platform for the development of other Bayesian nonparametric models, in the clustering domain and beyond. Finally, the various representations that we develop can provide new variables with which to express posterior inference algorithms. Indeed, we will present several such algorithms along the way.

4.1 The Pólya urn and the Blackwell-MacQueen urn

As the point of departure for our further development we return to the notion of exchangeability. Thus far we have exploited exchangeability in only a limited way, as a tool for computing conditional probabilities in the context of the Gibbs sampler. Moreover, exchangeability has been developed for random partitions, which are rather specialized mathematical objects. To make further progress, we need to study exchangeability for random variables, which are more widely useful mathematical objects. To convert from random partitions to random variables, we change our focus from the Chinese restaurant process, which is a distribution on partitions, to the closely-related Pólya urn, which is a distribution on (sequences of) random variables.

The Pólya urn model is defined as follows (Johnson and Kotz, 1977). Consider an urn in which there are b_0 black balls and w_0 white balls. Choose a ball at random, and put two balls of the same color back in the urn. That is, letting b_n and w_n denote the number of black and white balls in the urn at step n, pick a black ball with probability $b_n/(b_n + w_n)$ and a white ball with probability $w_n/(b_n + w_n)$, setting $b_{n+1} = b_n + 1$ and $w_{n+1} = w_n$ if a black ball is chosen, and conversely if a white ball is chosen.

The Pólya urn is closely related to the CRP. Indeed, suitably generalized, it is identical to the first two stages (Eq. (10) and Eq. (11)) of the CRP mixture model that we discussed in Sec. (3.2). To make the connection, note first that if we can allow the initial values b_0

and w_0 to be real numbers instead of integers. The resulting urn model is still well defined mathematically; we can still compute $b_n/(b_n + w_n)$ and $w_n/(b_n + w_n)$ at each step and augment the value associated with the winning color by one. Second, it is straightforward to generalize to any finite number of colors. Third, we can actually let the number of colors be open-ended with the following trick. Let black be a special color such that if black is chosen, we generate a new color (from some continuous palette), label a ball with the new color and place it in the urn. The "number" of black balls (which need not be an integer) stays fixed, and we designate that fixed value by the constant α. We thus generate a new color at each step with probability proportional to α, just as the CRP chooses a new table with probability proportional to α. Indeed, if the initial state of the urn is such that it only contains the color black, at the first step we choose a new color and assign it the value one, just as the CRP seats the first customer at a new table, and all subsequent steps of choosing colors are isomorphic to the seating decisions of the CRP. This generalization of the Pólya urn has been studied by a number of authors, including Blackwell and MacQueen (1973) and Hoppe (1984). We will refer to it as the *Blackwell-MacQueen urn*.

Finally, the "palette" does not need to be the real line, as is suggested by the word "color," but it can be any infinite-dimensional space, where the process of choosing a new value is implemented by drawing from a diffuse distribution on that space (so that unique values are chosen with probability one). Denoting this distribution by G_0 and equating parameter vectors with colors, we obtain a model which is isomorphic with the model specification in Eq. (10) and Eq. (11), where the CRP places customers at tables, and each table is labeled with a parameter vector, which is inherited by each of the customers sitting at that table.

Let the parameter vector ("color") chosen at the nth step be denoted θ_n and let the space from which θ_n is drawn be denoted Θ. The Blackwell-MacQueen urn can be written as follows:

$$\theta_{n+1} \,|\, \theta_1, \ldots, \theta_n \sim \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{\theta_i}, \tag{19}$$

where δ_{θ_i} denotes an atom at location θ_i (i.e., a probability distribution whose mass is concentrated at θ_i). The factor $\alpha + n$ ensures that the right-hand side is a probability distribution (i.e., that it assigns a total mass of one to Θ). Letting $\theta = (\theta_1, \theta_2, \ldots, \theta_N)$, we write:

$$\theta \sim \mathrm{BM}(\alpha, G_0, N) \tag{20}$$

to denote the draw of an entire sequence under the Blackwell-MacQueen urn model.

Having drawn a sequence $(\theta_1, \theta_2, \ldots, \theta_N)$, we can form the set of unique values, and identify the unique values with the parameters $\{\phi_c\}$ that label a set of tables in a CRP. Assigning each index $n \in \{1, \ldots, N\}$ to a table according to the equality of the parameter θ_n to a label ϕ_c, we can map a draw from the Blackwell-MacQueen urn onto a draw from the CRP.

As should be clear intuitively from the representation of this process in terms of an underlying CRP, there must be some sense in which the distribution we obtain on sequences of random vectors is exchangeable. If the indices of the customers are irrelevant in determining the partition, then they should also be irrelevant in determining the vectors that label

the customers, if these vectors are chosen independently given the partition. In fact, while partitions are somewhat exotic objects in probability theory, sequences of random vectors are garden-variety objects, and, as we discuss in the following section, there is a classical notion of exchangeability for such objects, one that leads to an important theorem.

4.2 Exchangeability and the de Finetti theorem

An infinite sequence of random vectors, $(\theta_1, \theta_2, \ldots)$, is said to be *infinitely exchangeable* if the distribution of any finite subsequence is invariant to permutation. That is, if σ is a permutation of the integers 1 through N, we require that

$$P(\theta_1 \in A_1, \theta_2 \in A_2, \ldots, \theta_N \in A_N) = P(\theta_{\sigma(1)} \in A_1, \theta_{\sigma(2)} \in A_2, \ldots, \theta_{\sigma(N)} \in A_N), \tag{21}$$

for arbitrary N and arbitrary sets $(A_1, \ldots, A_N)$. Intuitively, infinitely exchangeable sequences are just those for which the ordering doesn't matter.[4]

In 1931, de Finetti proved a fundamental theorem for exchangeable sequences (de Finetti, 1931). De Finetti's original theorem was for binary random variables, a special case that will be of interest to us, but it has been generalized far beyond that setting. Here is such a generalization:

Theorem 1. *The infinite sequence of random vectors, $(\theta_1, \theta_2, \ldots)$, is infinitely exchangeable if and only if the joint distribution of any N elements can be written as follows:*

$$P(\theta_1 \in A_1, \theta_2 \in A_2, \theta_N \in A_N) = \int \left(\prod_{n=1}^{N} G(A_n) \right) Q(dG) \tag{22}$$

for some random probability measure G with distribution Q.

If the notation "$Q(dG)$" is unfamiliar, please see Appendix A.

This theorem establishes an equivalence between exchangeability and the notion of "conditionally independent and identically distributed (iid)." That is, conditioning on G, the sequence $(\theta_1, \theta_2, \ldots, \theta_N)$ is iid; that is the meaning of the product in the integrand. In one direction, the theorem is obvious: if $(\theta_1, \theta_2, \ldots, \theta_N)$ has a representation as on the right-hand side of Eq. (22), then we clearly have invariance to permutation (because the product is invariant to permutation). The deep result is that the converse holds.

The notion of a *random probability measure* will seem mysterious to the uninitiated, and it is our purpose in the next two sections to make it seem natural. As a first step, let us consider the special case of Bernoulli random variables. In this case, a probability measure can be encoded by a single number $\eta \in (0, 1)$ giving the probability of heads. If this number η is itself a random variable, then it can be viewed as encoding a random probability measure. In this setting, we write de Finetti's theorem in a simplified form. For a Bernoulli sequence $(Z_1, Z_2, \ldots)$, and binary values $(z_1, z_2, \ldots, z_N)$, we have exchangeability if and

[4] There is also a notion of finite exchangeability, but, given that the focus of Bayesian nonparametrics is sequences with an open-ended number of elements, the notion of infinite exchangeability is more suited to our purposes.

only if:

$$P(Z_1 = z_1, Z_2 = z_2, \ldots, Z_N = z_N) = \int \left(\prod_{n=1}^{N} \eta^{z_n} (1-\eta)^{1-z_n} \right) Q(d\eta), \tag{23}$$

for some distribution Q which is the distribution of η. In some (but not all) cases this distribution can be obtained from a density, $p(\eta)$, in which case we can substitute $p(\eta)d\eta$ for $P(d\eta)$ in the integral.

As a concrete example, let us consider the case in which P is the *beta distribution*. The beta distribution has a density of the form:

$$q(\eta) = \frac{1}{B(\alpha_1, \alpha_2)} \eta^{\alpha_1 - 1} (1-\eta)^{\alpha_2 - 1}, \tag{24}$$

where $B(\alpha_1, \alpha_2)$ is the normalizing constant:

$$B(\alpha_1, \alpha_2) = \int \eta^{\alpha_1 - 1} (1-\eta)^{\alpha_2 - 1} d\eta. \tag{25}$$

If we plug this density into the right-hand side of Eq. (23) and define $s = \sum_{n=1}^{N} z_n$ as the number of heads, we obtain:

$$\begin{aligned} P(Z_1 = z_1, Z_2 = z_2, \ldots, Z_N = z_N) &= \int \left(\prod_{n=1}^{N} \eta^{z_n} (1-\eta)^{1-z_n} \right) \frac{1}{B(\alpha_1, \alpha_2)} \eta^{\alpha_1 - 1} (1-\eta)^{\alpha_2 - 1} d\eta \\ &= \frac{1}{B(\alpha_1, \alpha_2)} \int \eta^{s + \alpha_1 - 1} (1-\eta)^{N - s + \alpha_2 - 1} d\eta \\ &= \frac{B(s + \alpha_1, N - s + \alpha_2)}{B(\alpha_1, \alpha_2)} \\ &= \frac{\Gamma(s + \alpha_1)\Gamma(N - s + \alpha_2)\Gamma(\alpha_1 + \alpha_2)}{\Gamma(N + \alpha_1 + \alpha_2)\Gamma(\alpha_1)\Gamma(\alpha_2)}. \end{aligned} \tag{26}$$

We see that the result is invariant to the ordering of the random variables Z_n; it is a function only of the total count s. This confirms exchangeability.

On the other hand, if someone had handed us a distribution on binary sequences that is proportional to $\Gamma(s + \alpha_1)\Gamma(N - s + \alpha_2)$, we would know by de Finetti's theorem that there must exist an underlying random variable η such that we obtain this distribution upon integrating out η. We would set off on a hunt for such a random variable, and (hopefully) would soon discover that η is a beta random variable.

It should now be clear why we have introduced de Finetti's theorem into our discussion. The Blackwell-MacQueen urn is exchangeable,[5] and thus, by de Finetti, there must exist some underlying random measure G such that if we condition on a specific instantiation of G, then the sequence obtained from the Blackwell-MacQueen urn can be mimicked by drawing vectors iid from G. We are motivated to hunt for such a G.

[5] We have not proved the exchangeability of the Blackwell-MacQueen urn, but as we have alluded to, it follows from the exchangeability of the random partition provided by the CRP and the isomorphism of the Blackwell-MacQueen urn to the CRP combined with the iid labeling process.

In setting off on this hunt, we note a few facts. First, G cannot be a diffuse distribution (i.e., a distribution with no atoms). This is clear from the fact that the Blackwell-MacQueen urn can output the same vector multiple times. Second, G cannot have a finite support. This is clear from the fact that the Blackwell-MacQueen urn can continue to generate new vectors, with no upper bound. Thus, G is not the kind of object that is studied in an elementary probability class. Moreover, we are not interested in only a single G, but in a distribution on G (not a single measure, but a random measure). This all seems rather daunting. As we will see, however, the hunt can be successfully carried out, and the prey—the *Dirichlet process*—is an elegant and relatively simple object.

4.3 Gibbs sampling based on the Blackwell-MacQueen urn representation

Before turning to the Dirichlet process, we wish to provide a concrete example of the use of the Blackwell-MacQueen urn representation in the context of Gibbs sampling. Whereas Gibbs sampling in the CRP representation integrates over parameter vectors (the ϕ_c), in the Blackwell-MacQueen urn representation the Gibbs sampler instantiates parameter vectors (the θ_n) explicitly (Escobar, 1994; Escobar and West, 1995).

We note at the outset that this Gibbs sampler is *not* a very effective sampler, tending to mix slowly. We present it in part for historical reasons (it was among the first Markov chain Monte Carlo samplers to be developed for Bayesian nonparametric clustering), but also because it is useful to understand why the sampler mixes slowly.

The model on which the sampler is based is the following:

$$\begin{aligned} \theta &\sim \mathrm{BM}(\alpha, G_0, N) \\ x_i \,|\, \theta_i &\overset{ind}{\sim} F(\theta_i) \qquad\qquad \text{for } i = 1, \ldots, N. \end{aligned} \tag{27}$$

From a generative point of view, this model can be viewed as a notational variant of the CRP mixture model in Eq. (12). In particular, given that the sequence $\theta = (\theta_1, \theta_2, \ldots)$ is exchangeable, the sequence $x = (x_1, x_2, \ldots)$ is also exchangeable, and the marginal distribution obtained for x is the same as that in Eq. (12). Inferentially, however, the different model specification yields a quite different algorithm.

We consider a Gibbs sampler in which the state is the vector θ. Consider updating a component θ_i given the observations x and the other parameters, $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_N)$. The conditional distribution of θ_i can be obtained from the normalized product of the conditional prior of θ_i given θ_{-i} and the likelihood $f(x_i \,|\, \theta_i)$. Exploiting the exchangeability of θ under the Blackwell-MacQueen urn scheme, we can treat θ_i as the last variable in the sequence, so that the conditional prior is simply:

$$\theta_i \,|\, \theta_{-i} \sim \frac{\alpha}{N-1+\alpha} G_0 + \sum_{j \neq i} \frac{1}{N-1+\alpha} \delta_{\theta_j} = \frac{\alpha}{N-1+\alpha} G_0 + \sum_{k=1}^{K} \frac{n_k}{N-1+\alpha} \delta_{\phi_k}, \tag{28}$$

where $\phi_1, \ldots, \phi_K$ are the K unique values among θ_{-i} and where n_k is the number of components θ_i equal to ϕ_k. (Note that slight change in notation—we now index the parameter vectors ϕ_k by integers k rather than by clusters c as in the CRP where we wrote ϕ_c. One can map between these indices by letting the tables in the CRP be numbered according to

the order of their occupancy.) Multiplying by the likelihood $f(x_i \mid \theta_i)$ and normalizing gives the conditional distribution of θ_i given θ_{-i}. As the conditional prior in Eq. (28) has both a diffuse component G_0 and atoms δ_{ϕ_k} at the current values of θ_{-i}, the conditional posterior will also have these components. Thus there are $K+1$ options for θ_i, with probabilities given as follows:

$$P(\theta_i \in \Theta \backslash \{\phi_k\}_{k=1}^K \mid x, \theta_{-i}) \propto \frac{\alpha}{N-1+\alpha} \int f(x_i \mid \phi) g_0(\phi) d\phi$$
$$P(\theta_i = \phi_k \mid x, \theta_{-i}) \propto \frac{n_k}{N-1+\alpha} f(x_i \mid \phi_k) \qquad \text{for each } k = 1, \ldots, K, \tag{29}$$

where the constant of proportionality is obtained by normalizing across these $K+1$ alternatives. To sample θ_i, we first sample from Eq. (29) to determine if the new value of θ_i will equal some ϕ_k. If not, then θ_i takes on a value distinct from those in θ_{-i}, and we obtain that value by sampling from its conditional distribution with density:

$$p(\theta_i \mid x, \theta_{-i}, \theta_i \notin \{\phi_k\}_{k=1}^K) \propto f(x_i \mid \theta_i) g_0(\theta_i). \tag{30}$$

As we noted earlier, this Gibbs sampler tends to mix slowly. To understand the problem, consider a situation in which the Gibbs sampler has correctly identified a cluster c of similar observations, but that the parameter of the cluster is not sharply determined in the posterior, so that the Gibbs sampler should integrate over the likely values for the parameter. Suppose the current value is ϕ, but that there is another value ϕ^* which has higher likelihood, and consider the chance that the Gibbs sampler will update the parameter value from ϕ to ϕ^*. Note that under the Blackwell-MacQueen urn representation there is no single explicit random variable that represents a cluster or the corresponding parameter; rather, the cluster is represented implicitly as a maximal set of indices c such that all $\theta_i = \phi$ for each $i \in c$. For the Gibbs sampler to update the parameter of the cluster to ϕ^* would require each individual θ_i to be updated to the new value ϕ^* in turn, conditioned on the current values of θ_j for $j \in c \backslash \{i\}$. Some of these may have been updated to ϕ^* already, while some have not been updated so still have the value ϕ. This means that on the way to the state in which all θ_i equal the new value ϕ^*, the Gibbs sampler has to pass through a state in which the single cluster is split into two clusters, one with the previous parameter value ϕ and one with the new value ϕ^*. Such a state generally has low probability, so that the chance of the Gibbs sampler successfully updating the parameter of the cluster to ϕ^* is low.

Another problematic aspect of this Gibbs sampler is the computation of integrals needed to sample from Eq. (29) and Eq. (30). This is generally possible only when G_0 is conjugate to the likelihood $f(x \mid \phi)$, or when the parameter space is low dimensional so that numerical integration is feasible.

While the mixing problem can be alleviated via an improved scheme in which the values of all values of θ_i for each $i \in c$ are updated together (MacEachern, 1994; Bush and MacEachern, 1996), the difficulty of evaluating the needed integrals remains. This problem is mitigated by the samplers that we discuss in later sections. We also refer to Neal (2000), who presents samplers based on Metropolis-Hastings and augmentation techniques that do not require conjugacy.

5 Stick-Breaking and the Dirichlet Process

We return to the main thread of our discussion, the attempt to identify the random measure G underlying the Blackwell-MacQueen urn model. We begin this section with a short tutorial on measures and atoms, including a discussion of random measures, and then derive the Dirichlet process from the Blackwell-MacQueen urn model.

Recall that a *measure* μ on a set Θ is a function from subsets of Θ to the nonnegative real numbers that satisfies countable additivity:[6] for disjoint subsets, $(A_1, A_2, \ldots)$, we have $\mu(\cup_n A_n) = \sum_n \mu(A_n)$. As a simple example of a measure, consider the *atomic measure*, $\mu = \delta_\theta$, which places a unit mass at the point $\theta \in \Theta$. We will refer to such a point as an *atom*. For any subset A, we have $\delta_\theta(A) = 1$ if the atom θ lies in A and zero otherwise. This function clearly satisfies countable additivity.

Atomic measures can be used to build up general discrete measures. Note in particular that multiplying δ_θ by a nonnegative scalar yields a new measure. Moreover, countable sums of such measures are measures, so the following object is a measure:

$$\mu = \sum_{k=1}^{\infty} w_k \delta_{\phi_k}, \tag{31}$$

for nonnegative "weights" $\{w_k\}$. For $A \subseteq \Theta$, we have

$$\mu(A) = \sum_{k=1}^{\infty} w_k \delta_{\phi_k}(A) = \sum_{k:\phi_k \in A} w_k, \tag{32}$$

which is the sum of the weights of the atoms falling in A.

We now take the step to a *random measure*. We do this by making the object in Eq. (31) random, doing so in two ways. First, we let the $\{w_k\}$ be nonnegative random variables. Second, the atoms $\{\phi_k\}$ are also chosen randomly. Specific ways of making these random choices will define particular families of random measures. We write

$$G = \sum_{k=1}^{\infty} w_k \delta_{\phi_k} \tag{33}$$

to denote the random measure; this is the same as Eq. (31) but the $\{w_k\}$ and the $\{\phi_k\}$ are now random, and we have written G instead of μ to remind us that the measure is random.[7]

If we fix a particular subset A of Θ and compute $G(A)$, the result is no longer a number as in the case of $\mu(A)$, but a random variable. Note in particular that we can apply G to Θ. If $G(\Theta) = 1$, with probability one, then we refer to G as a *random probability measure*.[8]

[6] For completeness, let us note that the set of subsets of Θ on which the measure is defined needs to be a *sigma algebra*, but it will not be necessary to know what a sigma algebra is to proceed.

[7] We will forgo an attempt to define a random measure more formally, but for readers wishing to pursue the matter, let us note in passing that the formal definition of a random measure is as a "transition kernel." See, e.g., Kallenberg (2002). See also Sec. (6.1) for a discussion that links random measures to stochastic processes.

[8] A statement such as "$G(\Theta) = 1$" is a probabilistic statement, given that $G(\Theta)$ is a random variable, and in referring to G as a random probability measure, we require this statement to hold only with probability one with respect to the underlying randomness that makes G a random measure.

We can achieve this by requiring $\sum_{k=1}^{\infty} w_k = 1$ to hold (with probability one), because

$$G(\Theta) = \sum_{k=1}^{\infty} w_k \delta_{\phi_k}(\Theta) = \sum_{k=1}^{\infty} w_k. \tag{34}$$

Returning to the Blackwell-MacQueen urn, it is reasonable to expect that the underlying measure promised by the de Finetti theorem can be written in the form of Eq. (33), given that we require a countably infinite number of atoms. Indeed, thinking in terms of the CRP representation, it seems reasonable that we should associate one atom with each table in the CRP, and require $\sum_{k=1}^{\infty} w_k = 1$ to capture the fact that the CRP chooses one and only one table at each step. The problem is to uncover the specific way in which we must turn the $\{w_k\}$ and the $\{\phi_k\}$ into random variables for the case of the Blackwell-MacQueen urn.

Let us focus first on the $\{w_k\}$. For these numbers to sum to one, they must decay as i goes to infinity. This corresponds to the occupancy of the tables in the CRP—those tables that are occupied early in the process are likely to continue to remain the most highly occupied, and should thus correspond stochastically to the larger values of w_k. Let us therefore consider treating w_1 as the probability associated with the first occupied table in the CRP, and see if we can deduce the distribution of w_1 from the CRP dynamics.

Just after the moment in time in which the first table is occupied in the CRP, we have one customer at that table. Future customers can either sit at that table or sit at some other table. Let us therefore consider a Pólya urn in which "one" denotes sitting at the identified table and "zero" denotes sitting at some other table. Let Z_i denote the binary indicator associated with the ith customer after the first customer (thus, Z_1 denotes the indicator associated with the second customer). We compute the probability of observing a sequence of N ones under this Pólya urn:

$$\begin{aligned} P(Z_1 = 1, Z_2 = 1, \ldots, Z_N = 1) &= \left(\frac{1}{\alpha+1}\right)\left(\frac{2}{\alpha+2}\right)\cdots\left(\frac{N}{\alpha+N}\right) \\ &= \frac{\Gamma(N+1)}{(\alpha+1)^{(N)}}. \end{aligned} \tag{35}$$

On the other hand, from the Bernoulli version of de Finetti's theorem in Eq. (23), we have:

$$P(Z_1 = 1, Z_2 = 1, \ldots, Z_N = 1) = \int \eta^N P(d\eta), \tag{36}$$

for some random variable η. This latter expression is the Nth moment of η. The set of moments of a random variable uniquely identify the distribution of that random variable, and so the problem is to find a random variable whose Nth moment is $\Gamma(N+1)/(\alpha+1)^{(N)}$. But we have already found this random variable: from Eq. (26) if we substitute $s = N$, and let $\alpha_1 = 1$ and $\alpha_2 = \alpha$, we obtain:

$$P(Z_1 = 1, Z_2 = 1, \ldots, Z_N = 1) = \frac{\Gamma(N+1)\Gamma(1+\alpha)}{\Gamma(N+1+\alpha)} \tag{37}$$

$$= \frac{\Gamma(N+1)}{(\alpha+1)^{(N)}}, \tag{38}$$

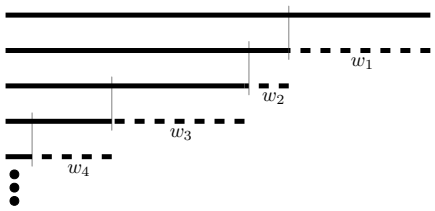

Figure 3: Stick-breaking construction for the Dirichlet process. We start with a stick of length 1, and recursively break off pieces of sticks of lengths $w_1, w_2, \ldots$.

which is the same as Eq. (35). Thus we see that $w_1 \sim \text{Beta}(1, \alpha)$.

The same thought experiment applies to the determination of w_2. Having allocated probability w_1 to the event of sitting at the first occupied table, we have probability $1 - w_1$ to allocate to further seating decisions. We again consider a Pólya urn, in which "one" denotes sitting at the second occupied table, and "zero" denotes sitting at some heretofore unoccupied table (i.e., neither the first or the second). This latter event still occurs with probability proportional to α, and so we have the identical urn as before. Thus the probability of sitting at the second occupied table, given that we do not sit at the first occupied table, is drawn from $\text{Beta}(1, \alpha)$. Denoting that draw by β_2, we have $w_2 = \beta_2(1 - w_1)$.

In general, we can generate the values w_k as follows:

$$\begin{aligned} \beta_k &\overset{iid}{\sim} \text{Beta}(1, \alpha) && k = 1, 2, \ldots \\ w_k &= \beta_k \prod_{j<k} (1 - \beta_j) && k = 1, 2, \ldots . \end{aligned} \tag{39}$$

This procedure is referred to as *stick-breaking*, and the distribution that it defines on the infinite sequence $w = (w_1, w_2, \ldots)$ is known as the *GEM distribution*:

$$w \sim \text{GEM}(\alpha). \tag{40}$$

As shown in Fig. (3), the values w can be viewed as fragments of a unit-length stick, obtained by breaking off a fraction of the remainder of the stick after the preceding fragments have been removed.

We also need to specify the random mechanism behind the choice of the atoms ϕ_k in Eq. (33). Given that the stick-breaking process has captured the CRP dynamics, all that is left to do is to draw the vectors ϕ_k independently:

$$\phi_k \overset{iid}{\sim} G_0 \qquad k = 1, 2, \ldots . \tag{41}$$

Thus, G as defined in Eq. (33) is indeed random in two ways: the weights w are chosen randomly by stick-breaking, and the atoms ϕ_k are chosen randomly by iid draws.

We show some examples of draws of the random measure G in Fig. (4). Note that these draws are all discrete, and they are centered around the underlying measure G_0. The heights of the atoms are determined independently of the locations by the stick-breaking process.

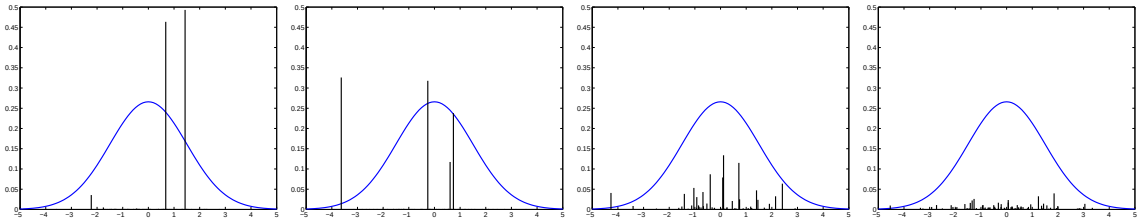

Figure 4: Examples of draws from the Dirichlet process. From left to right are 4 independent draws G, with concentration parameters $0.5, 1, 10$ and 100 respectively. The blue curve is the density of the base distribution, while each vertical black line corresponds to an atom (with mass given on y axis) in G.

The random measure that we have constructed has a name: it is the *Dirichlet process*. We explain the name—in what sense our random measure is a "process" and where the "Dirichlet" comes from—in Sec. (6). We denote a draw from the Dirichlet process as follows:

$$G \sim \mathrm{DP}(\alpha, G_0), \tag{42}$$

where we refer to α as the *concentration parameter* and G_0 as the *base measure* or *base distribution*. That "concentration parameter" is a reasonable name for α can be seen from Fig. (4), where we see that large values of α lead to draws that are concentrated around the base measure G_0 and small values of α lead to draws that are more variable. This point will be made more precise later in Eq. (51) where we compute the variance of the Dirichlet process, finding that α appears in the denominator of the variance. The intuitive reason for the behavior should be clear from considering the role of α in the CRP; a small value of α leads to draws G that have their support on a small number of atoms at random locations. Note also that the Dirichlet process is sometimes denoted as $\mathrm{DP}(\alpha G_0)$, where αG_0 is an unnormalized measure and where the concentration parameter α can be recovered from the parameter αG_0 by normalization.

Finally, we define the *Dirichlet process mixture model*, a full-blown model specification where we connect the Dirichlet process to a likelihood and thereby provide a Bayesian generative model for data. The overall model specification is as follows:

$$G \sim \mathrm{DP}(\alpha, G_0), \tag{43}$$

$$\theta_i \,|\, G \overset{iid}{\sim} G \qquad \text{for } i = 1, 2, \ldots, N, \tag{44}$$

$$x_i \,|\, \theta_i \overset{ind}{\sim} F(\theta_i) \qquad \text{for } i = 1, 2, \ldots, N, \tag{45}$$

where $F(\theta_i)$ is the distribution corresponding to the density $f(\cdot \,|\, \theta_i)$. This model should be viewed as an augmentation of the model presented earlier in Eq. (27) to include the underlying random measure G.

5.1 Gibbs sampling based on stick-breaking

The stick-breaking representation of the DP opens up new opportunities for sampling-based posterior inference algorithms. In particular, we can now design algorithms that explicitly instantiate the random measure G, thereby decoupling (i.e., rendering conditionally independent) the parameters $\{\theta_i\}$. Samplers that explicitly instantiate G are referred to as "conditional samplers." The samplers that we have discussed in previous sections, which operate on a model in which G has been integrated out, are referred to as "marginalized samplers." Conditional samplers are often simpler to implement and more readily parallelizable than marginalized samplers. Further, with conditional samplers we can estimate various functionals of the posterior of G, such as medians and quantiles, that are not available with marginalized samplers.

The main difficulty one faces in the design of conditional samplers comes from the fact that the random measure G has its support on infinitely many atoms, as seen in Eq. (33). Representing such a random measure exactly would be impossible on a computer with a finite amount of memory. We present two approaches to dealing with this issue. The first involves an approximation in which the infinite sum in Eq. (33) is truncated to a finite sum, exploiting the fact that the fact that the masses w_k decrease to zero exponentially quickly. This can be viewed as the analog of the floating-point representation of real numbers on computers, where we simply keep track of the most significant digits of each real number (its mantissa), along with an exponent. The second involves no approximation, but instead exploits the fact that only a finite number of the atoms comprising G are used in the generation of data points; these atoms can be generated on the fly as needed. We present the first approach in this section and present the latter in Sec. (6.4).

In the truncation approach (Ishwaran and James, 2001), we truncate the representation of G by picking a value $K_{\max}$ and retaining only atoms ϕ_k for which $k \leq K_{\max}$:

$$G = \sum_{k=1}^{K_{\max}} w_k \delta_{\phi_k},$$

where

$$\begin{aligned} \phi_k &\sim G_0 & w_k &= \beta_k \prod_{j<k}(1-\beta_j), & k &= 1, \ldots, K_{\max}, \\ \beta_{K_{\max}} &= 1, & \beta_k &\sim \text{Beta}(1, \alpha), & k &= 1, \ldots, K_{\max} - 1. \end{aligned} \tag{46}$$

Note that we set $\beta_{K_{\max}} = 1$ so that G has a total mass of one and is thus still a random probability measure even after truncation.

The choice of $K_{\max}$ is obviously important here, and Ishwaran and James (2001) present theoretical results capturing the error induced in the prior distribution by the truncation. This error in the prior does not translate immediately into control over error in the posterior, but it does serve as a guide. In practice, one often sets $K_{\max}$ to a value that is $4-5$ times larger than a subjectively chosen value of the number of clusters in the data, runs the algorithm discussed below, and assesses whether the posterior appears to place most of its mass significantly below $K_{\max}$. If not, one runs the algorithm again with a larger $K_{\max}$.

To specify a sampling algorithm based on the truncated model, we introduce cluster assignment variables, z_i, which indicate the index $k \in \{1, \ldots, K_{\max}\}$ to which each observation x_i is assigned in the generative model. The conditional distributions are:

$$\begin{aligned} z_i \,|\, G &\sim \text{Discrete}(w_1, \ldots, w_{K_{\max}}) && i = 1, \ldots, N \\ x_i \,|\, z_i = k, G &\sim F(\phi_k) && i = 1, \ldots, N. \end{aligned} \tag{47}$$

We now define a Gibbs sampler with state space $\{\beta_k, \phi_k\}_{k=1}^{K_{\max}}$ and $\{z_i\}_{i=1}^n$. The conditional distributions are straightforward to derive, giving:

$$\begin{aligned} p(z_i = k \,|\, x, z_{-i}, \beta, \phi) &\propto w_k f(x_i \,|\, \phi_k) && i = 1, \ldots, N \\ p(\phi_k \,|\, x, z, \beta, \phi_{-k}) &\propto g_0(\phi_k) \prod_{i: z_i = k} f(x_i \,|\, \phi_k) && k = 1, \ldots, K_{\max} \\ p(\beta_k \,|\, x, z, \beta_{-k}, \phi) &\propto \beta_k^{n_k} (1 - \beta_k)^{\alpha + n_{>k} - 1} && k = 1, \ldots, K_{\max} - 1 \end{aligned} \tag{48}$$

where $n_k = \sum_{i=1}^n \mathbb{1}(z_i = k)$ is the number of observations in cluster k and where $n_{>k} = \sum_{i=1}^n \mathbb{1}(z_i > k)$ is the number of observations in clusters $\ell > k$. We see that the conditional distribution of β_k is just $\text{Beta}(1 + n_k, \alpha + n_{>k})$. The conditional distribution of ϕ_k is also obtained in a closed form if G_0 is conjugate to the likelihood; if not we can employ a Metropolis-Hastings step to update ϕ_k (or some other ergodic MCMC update which leaves the conditional distribution invariant).

An advantage of this Gibbs sampler relative to the samplers that we have discussed earlier is that the introduction of G in the sampler (i.e., the use of $\{\beta_k, \phi_k\}$) exposes the inherent independence between the cluster assignment variables, making it possible to sample these values in parallel. Moreover, updates to β and ϕ can also be sampled in parallel. Note also that the conditional probability of the cluster assignment variables does not involve the evaluation on an integral, diminishing the need for conjugacy in this approach.

6 Properties of the Dirichlet Process and its Posterior

The Dirichlet process is our first example of a random measure. As we proceed we will see other random measures. These random measures play a key role in Bayesian nonparametrics, by introducing an infinite number of degrees of freedom into a model and thereby freeing us from parametric restrictions. To be able to exploit this flexibility, we need to show that the standard operations of Bayesian inference are feasible within a model that includes random measures. In particular, we need to show that a formula akin to Bayes' theorem allows us to compute conditional probabilities in a model that includes random measures as first-class citizens in the model. In this section we demonstrate this computation in the simple setting of a hierarchical model in which vectors $\theta = (\theta_1, \ldots, \theta_N)$ are drawn independently from a random measure G and the problem is to compute the posterior of G given θ.

Before turning to this problem we provide some further background on basic properties of the Dirichlet process (DP). We first explain the name—in particular we explain the sense in which the DP is a "process," and we explain the meaning of "Dirichlet." We show how to compute moments of the DP. We then turn to the key issue of obtaining the posterior

distribution of a DP. From the posterior representation of the DP we show how to derive the Chinese restaurant process, bringing our chain of arguments full circle. Finally, we illustrate the use of the posterior representation in the design of a slice sampling algorithm for posterior inference.

6.1 Marginals and moments

A stochastic process is an indexed collection of random variables that obey certain consistency conditions. Often the index set is the real line (e.g., in the case of Brownian motion) or some other Euclidean space. However, other index sets can be considered and indeed the right way to think about the DP (and other random measures) is that its index set is a set of sets. Let us slow down and understand this perspective.

Recall that if we evaluate G on a fixed subset A, the resulting object $G(A)$ is a random variable. Ranging over a set of such subsets, we obtain a collection of random variables. Indeed, applying G to a collection of sets, $(A_1, A_2, \ldots, A_K)$, we obtain a random vector, $(G(A_1), G(A_2), \ldots, G(A_K))$. If G is to be a random probability measure, then we require the joint probability distribution of this random vector to satisfy certain consistency conditions as we range over choices of subsets $(A_1, A_2, \ldots, A_K)$. These consistency conditions are easy to understand if we specialize to consider sets of sets that are finite *partitions* of the underlying space, and, in fact, a small amount of measure-theoretic thinking reveals that it suffices partitions. Thus, again denoting the underlying space by Θ, let $(A_1, A_2, \ldots, A_K)$ be a partition of Θ, and consider all such partitions, for all values of K. If G is to be a random probability measure, then we require $G(\Theta) = 1$ and we require that the following aggregation property holds: the random vector $(G(A_1), \ldots, G(A_i) + G(A_{i+1}), \ldots, G(A_K))$ must have the same distribution as the random vector $(G(A_1), \ldots, G(A_i \cup A_{i+1}), \ldots, G(A_K))$, for all choices of i.

One well-known distribution that satisfies these properties is the *Dirichlet distribution*. (See Appendix B for an overview of the Dirichlet distribution.) Thus, if we have in hand a putative random measure G that satisfies the following finite-dimensional distributional requirement:

$$(G(A_1), G(A_2), \ldots, G(A_K)) \sim \mathrm{Dir}(\alpha G_0(A_1), \alpha G_0(A_2), \ldots, \alpha G_0(A_K)), \tag{49}$$

for all partitions $(A_1, A_2, \ldots, A_K)$ and for all K, then we certainly satisfy the consistency conditions associated with G being a random measure.

In fact, the random probability measure G that we defined in Eq. (33) turns out to satisfy Eq. (49). It is for this reason that the distribution of the random probability measure G is known as the "Dirichlet process." We summarize Eq. (49) by saying that "the Dirichlet process has Dirichlet marginals." We provide a direct proof of this fact, starting with Eq. (33), in Appendix C. We also provide a second path to understanding how Dirichlet marginals arise in Sec. (9), via the theory of completely random measures and the gamma process.

The Dirichlet process first appeared in a paper by Ferguson (1973), and in that paper Eq. (49) was treated as the definition of the process. Such a definition leaves open the question as to whether such a process actually exists, and Ferguson made an appeal

to a general theorem of Kolmogorov on the existence of stochastic processes, given collections of finite-dimensional distributions that satisfy consistency conditions. As emphasized by Sethuraman (1994), such an appeal runs into a measure-theoretic difficulty that requires certain topological conditions to be placed on Θ. We have followed Sethuraman in treating Eq. (33) as the definition of the Dirichlet process. This approach avoids the need to place any conditions on Θ. Moreover, under the definition, G is clearly a probability measure (under the event that the weights w_k sum to one, which happens with probability one).

Although Eq. (33) has virtues as a definition, it makes computations difficult which are easy under Eq. (49). In particular, a simple consequence of Eq. (49) is that we can readily compute the mean and variance of $G(A)$, for any $A \in \mathcal{A}$:

$$\mathbb{E}(G(A)) = G_0(A) \tag{50}$$

$$\mathsf{Var}(G(A)) = \frac{G_0(A)(1 - G_0(A))}{1 + \alpha}. \tag{51}$$

This follows directly from Eq. (271) and Eq. (277) in Appendix B. In the following section we will see another important application of Eq. (49).

6.2 The posterior Dirichlet process

We now turn to the important problem of obtaining the posterior distribution of a DP in models in which a draw from a DP is one component of a more elaborate model. Let us consider the following simple hierarchical model:

$$G \sim \mathrm{DP}(\alpha, G_0) \tag{52}$$

$$\theta_i \,|\, G \overset{iid}{\sim} G \qquad\qquad \text{for } i = 1, \ldots, N, \tag{53}$$

where a set of variables, $\theta = (\theta_1, \ldots, \theta_N)$, are drawn independently from the random probability measure G. The random measure G is a draw from the Dirichlet process. The problem is to determine the posterior distribution of G given θ.

We can obtain a hint as to what this posterior distribution should be by considering the finite-dimensional Dirichlet distribution. A draw from the Dirichlet distribution can be considered as a probability measure over a discrete set of values. Fixing this probability measure and drawing from it repeatedly corresponds to multinomial sampling. It is well known that the posterior distribution in such a setting is Dirichlet, with the Dirichlet parameters updated by adding the observed counts (this is often expressed by saying that the Dirichlet is conjugate to the multinomial). Our problem is essentially the same problem, with the difference that G provides a countable infinity of options. Selecting a value θ_i from G corresponds to picking one of these options, which is the infinite-dimensional analog of multinomial sampling. This line of reasoning suggests that the posterior distribution under the model in Eq. (53) should itself be a Dirichlet process.

To turn this intuition into a rigorous argument, we make use of the connection induced between the Dirichlet process and the Dirichlet distribution when considering finite partitions of Θ. Given such a finite partition, $(A_1, \ldots, A_K)$, consider the vector $(\delta_{\theta_i}(A_1), \ldots, \delta_{\theta_i}(A_K))$ for fixed i. Because $(A_1, \ldots, A_K)$ is a partition, exactly one of the entries of this vector is equal to one with the rest being equal to zero. Further, since θ_i

is distributed according to G, the probability that $\delta_{\theta_i}(A_j) = 1$, that is, the probability that $\theta_i \in A_j$, is simply $G(A_j)$. Thus, $(\delta_{\theta_i}(A_1), \ldots, \delta_{\theta_i}(A_K))$ has a multinomial distribution with parameter $(G(A_1), \ldots, G(A_K))$. This is true for each $i = 1, \ldots, N$; moreover, the corresponding multinomial vectors are independent, given that the θ_i are independent. We thus use the fact that the Dirichlet distribution is conjugate to the multinomial and thereby obtain the posterior distribution of $(G(A_1), \ldots, G(A_K))$ given the multinomial vectors $\{(\delta_{\theta_i}(A_1), \ldots, \delta_{\theta_i}(A_K))\}$:

$$(G(A_1), \ldots, G(A_K)) \,|\, \{(\delta_{\theta_i}(A_1), \ldots, \delta_{\theta_i}(A_K))\} \sim \mathrm{Dir}\left(\alpha G_0(A_1) + \sum_{i=1}^{N} \delta_{\theta_i}(A_1), \ldots, \alpha G_0(A_K) + \sum_{i=1}^{N} \delta_{\theta_i}(A_K)\right), \tag{54}$$

where $\sum_{i=1}^{N} \delta_{\theta_i}(A_j)$ is the count of the number of times that θ_i lies in the subset A_j for each $j = 1, \ldots, K$. Since this result holds for any partition $(A_1, \ldots, A_K)$, and for any value of K, we have shown that the posterior distribution of G—given the multinomial vectors $\{(\delta_{\theta_i}(A_1), \ldots, \delta_{\theta_i}(A_K))\}$—is the Dirichlet process. However, we want to obtain the posterior of G given θ, rather than given the multinomial vectors, and so this is not quite what we need. Indeed, each θ_i in principle contains more information than $(\delta_{\theta_i}(A_1), \ldots, \delta_{\theta_i}(A_K))$, which only tells us in which subset θ_i lies, not *where* in the subset θ_i lies. It turns out, however, that these two conditional distributions are the same; the extra information about where θ_i lies within each subset is not relevant for the conditional distribution. This fact, which is referred to as the "tail-free property of the Dirichlet process," is established in Appendix D. Assuming the tail-freeness, we obtain our expected conclusion: the posterior of G given θ is itself a Dirichlet process.

We can now remove reference to the specific partition and write our result simply in terms of the concentration parameter and the base measure:

$$G \,|\, \theta \sim \mathrm{DP}\left(\alpha + N, \frac{\alpha}{\alpha + N} G_0 + \frac{1}{\alpha + N} \sum_{i=1}^{N} \delta_{\theta_i}\right). \tag{55}$$

We see that the concentration parameter of the posterior DP is $\alpha + N$, the prior concentration plus the number N of observed values of θ_i, while the updated base distribution is $\frac{\alpha}{\alpha+N} G_0 + \frac{1}{\alpha+N} \sum_{i=1}^{N} \delta_{\theta_i}$. This updated base distribution is a convex combination of the original base distribution, G_0, and the empirical distribution $\frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_i}$. Thus we obtain the standard Bayesian compromise between our prior as given by the base distribution and the "data," as given by the vector θ. Further, as the number of observations increase, and recalling Eq. (51), the posterior distribution over G concentrates around its posterior mean, which, recalling Eq. (50), becomes increasingly dominated by the empirical distribution.

A second representation for the posterior DP can be derived from Eq. (55) that will prove useful in our subsequent work. Let $\theta_1^*, \ldots, \theta_K^*$ be the K unique values among the components of θ. Consider the partition of Θ into $K + L$ subsets $(\{\theta_1^*\}, \ldots, \{\theta_K^*\}, A_1, \ldots, A_L)$, where $(A_1, \ldots, A_L)$ forms a partition of $\Theta \backslash \theta$. Since the posterior distribution of G given θ is a DP, it has posterior marginals given by Eq. (49):

$$(G(\{\theta_1^*\}), \ldots, G(\{\theta_K^*\}), G(A_1), \ldots, G(A_L)) | \theta \sim \mathrm{Dir}(n_1, \ldots, n_K, \alpha G_0(A_1), \ldots, \alpha G_0(A_L)) \tag{56}$$

where $n_k = \sum_{i=1}^{N} \mathbb{1}(\theta_i = \theta_k^*)$ is the number of occurrences of θ_k^* among the components of θ. Using Proposition 7 from Appendix B, we can decompose this random probability vector of length $K + L$ into two random probability vectors, of lengths $K + 1$ and L, respectively, that are conditionally independent given θ:

$$(G(\{\theta_1^*\}), \ldots, G(\{\theta_K^*\}), G(\Theta \backslash \theta)) \perp\!\!\!\perp \frac{1}{G(\Theta \backslash \theta)} (G(A_1), \ldots, G(A_L)), \,|\, \theta \tag{57}$$

with conditional distributions:

$$(G(\{\theta_1^*\}), \ldots, G(\{\theta_K^*\}), G(\Theta \backslash \theta)) \,|\, \theta \sim \mathrm{Dir}(n_1, \ldots, n_K, \alpha G_0(\Theta \backslash \theta)) \tag{58}$$

$$\frac{1}{G(\Theta \backslash \theta)} (G(A_1), \ldots, G(A_L)) \,|\, \theta \sim \mathrm{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_L)). \tag{59}$$

We now use the fact that G_0 is diffuse and θ is finite, such that $G_0(\Theta \backslash \theta) = 1$. We see that the posterior DP has the following representation:

$$\begin{aligned} G \,|\, \theta &= \sum_{k=1}^{K} w_k \delta_{\theta_k^*} + w' G' \\ (w_1, \ldots, w_K, w') \,|\, \theta &\sim \mathrm{Dir}(n_1, \ldots, n_K, \alpha) \\ G' \,|\, \theta &\sim \mathrm{DP}(\alpha, G_0) \end{aligned} \tag{60}$$

In words, the posterior distribution of G given θ is described as a collection of weighted atoms at the K unique values of θ, and an additional collection of atoms from an independently drawn DP G', where the weights associated with these atoms are Dirichlet distributed and independent of G'.

6.3 From the Dirichlet process to the Chinese restaurant process

In this section we bring our chain of arguments full circle, showing that the Chinese restaurant process can be derived from the Dirichlet process.

Recall that the CRP can be obtained by working with the Blackwell-MacQueen urn model and grouping together values of θ_i that are equal, viewing them as customers at a table in the restaurant. Our derivation actually works with the Blackwell-MacQueen urn representation.

The Blackwell-MacQueen urn model generates a sequence $(\theta_1, \theta_2, \ldots, \theta_N)$. We wish to compute the conditional probability distribution of θ_{N+1} given $(\theta_1, \theta_2, \ldots, \theta_N)$. To do so, we let A be a subset of Θ, and calculate as follows:

$$\begin{aligned} P(\theta_{N+1} \in A \,|\, \theta_1, \ldots, \theta_N) &= \mathbb{E}[P(\theta_{N+1} \in A \,|\, \theta_1, \ldots, \theta_N, G) \,|\, \theta_1, \ldots, \theta_N] \\ &= \mathbb{E}[P(\theta_{N+1} \in A \,|\, G) \,|\, \theta_1, \ldots, \theta_N] \\ &= \mathbb{E}[G(A) \,|\, \theta_1, \ldots, \theta_N] \\ &= \frac{\alpha}{\alpha + N} G_0(A) + \frac{1}{\alpha + N} \sum_{i=1}^{N} \delta_{\theta_i}(A), \end{aligned} \tag{61}$$

where the first equality is the tower property of conditional expectation, the second equality is the conditional independence assumption from Eq. (53), the third equality is also obtained from Eq. (53), and the final equality is obtained from Eq. (50) using the posterior base measure from Eq. (55).

This final equation is just the Blackwell-MacQueen urn. To see this, first consider the case in which A is equal to the singleton $\{\theta_i\}$. We see that in this case the conditional probability of obtaining another label $\{\theta_i\}$ is proportional to the previous number of draws with the label $\{\theta_i\}$. Now consider the probability of obtaining a new label. This probability is obtained by letting $A = \Theta \backslash \{\theta_1, \ldots, \theta_N\}$. We obtain a probability that is proportional to $\alpha G_0(A) = \alpha G_0(\Theta) = \alpha$, where the first equality comes from the assumption that G_0 is diffuse.

6.4 Conditional slice sampler for DP mixture models

So far we have described three distinct samplers for different mixture models: a CRP mixture model in Section 3.3, a Pólya urn mixture model in Section 4.3, and a stick-breaking mixture model in Section 5.1. Further, we have seen that all three models are intimately tied to the Dirichlet process. In this section we will describe our last sampler, derived for the *DP mixture model* as given in the following:

$$G \sim \mathrm{DP}(\alpha, G_0) \tag{62}$$

$$\theta_i \,|\, G \overset{iid}{\sim} G \qquad \text{for } i = 1, \ldots, N, \tag{63}$$

$$x_i \,|\, \theta_i \overset{ind}{\sim} F(\theta_i) \qquad \text{for } i = 1, \ldots, N. \tag{64}$$

In this section, we develop a second conditional sampler for DP mixture models, based on the posterior DP representation in Eq. (60). The sampler also introduces a second idea for truncation, based on slice sampling, which does not introduce any approximation errors (Walker, 2007). In particular, we augment the state space with additional variables $s = (s_1, \ldots, s_N)$ which are independent, with s_i uniformly distributed between 0 and $G(\{\theta_i\})$; i.e., the mixing proportion of the cluster to which data point x_i is currently assigned to:

$$s_i \,|\, G, \theta_i \sim \mathcal{U}[0, G(\{\theta_i\})]. \tag{65}$$

We now consider a sampler which alternates between sampling G and s together given θ and sampling θ given G and s. Let $\theta_1^*, \ldots, \theta_K^*$ be the unique values among θ. These form the parameters of the K components in the mixture model currently associated with data. Let z_i denote the component that x_i belongs to, i.e. $\theta_i = \theta_{z_i}^*$. We can update each θ_k^* using any ergodic MCMC update which has invariant distribution given by:

$$p(\theta_k^* | x, z) \propto g_0(\theta_k^*) \prod_{i : z_i = k} f(x_i \,|\, \theta_k^*) \tag{66}$$

In the first phase, the conditional distribution of G given θ is given by Eq. (60). Making use of the stick-breaking representation for $G' = \sum_{\ell=1}^{\infty} u_\ell \delta_{\theta'_\ell}$,

$$G \,|\, \theta = \sum_{k=1}^{K} w_k \delta_{\theta_k^*} + w' \sum_{\ell=1}^{\infty} u_\ell \delta_{\theta'_\ell}, \tag{67}$$

where

$$(w_1, \ldots, w_K, w') \,|\, \theta \sim \text{Dir}(n_1, \ldots, n_K, \alpha) \tag{68}$$

$$(u_1, u_2, \ldots) \,|\, \theta \sim \text{GEM}(\alpha) \tag{69}$$

$$\theta'_\ell \,|\, \theta \overset{iid}{\sim} G_0 \qquad \text{for } \ell = 1, 2, \ldots. \tag{70}$$

For $i = 1, \ldots, N$, let $z_i \in [K]$ denote the cluster index such that $\theta_i = \theta^*_{z_i}$, so that $G(\{\theta_i\}) = w_{z_i}$. Now given G and θ each s_i is conditionally independent of the others and have distribution:

$$s_i \,|\, G, \theta \sim \mathcal{U}[0, w_{z_i}]. \tag{71}$$

Note that the s_i's depend only on the masses $w_1, \ldots, w_K$ of the atoms $\theta^*_1, \ldots, \theta^*_K$ associated with the data points, and not on those in G' which are not associated with data. Thus they can be sampled after $(w_1, \ldots, w_K, w')$ but before G'. Finally G' is independent of the data and can be simulated using the stick-breaking representation. Note that this technically requires simulating all infinitely many atoms in G'. However we will see next that only a finite number of atoms in G' are needed.

In the second phase of the sampler, conditioned on G and s, the θ_i's are sampled independently with distributions given by:

$$p(\theta_i \,|\, G, s_i, x_i) \propto \begin{cases} w_k \cdot p(s_i \,|\, G, \theta_i) p(x_i \,|\, \theta_i) & \text{for } \theta_i = \theta^*_k, \text{ for some } k = 1, \ldots, K, \\ w' u_\ell \cdot p(s_i \,|\, G, \theta_i) p(x_i \,|\, \theta_i) & \text{for } \theta_i = \theta'_\ell, \text{ for some } \ell = 1, 2, \ldots. \end{cases} \tag{72}$$

$$\propto \begin{cases} w_k \cdot \frac{1}{w_k} f(x_i \,|\, \theta^*_k) & \text{for } \theta_i = \theta^*_k, \text{ for some } k = 1, \ldots, K \text{ with } w_k > s_i, \\ w' u_\ell \cdot \frac{1}{w' u_\ell} f(x_i \,|\, \theta'_\ell) & \text{for } \theta_i = \theta'_\ell, \text{ for some } \ell = 1, 2, \ldots \text{ with } w' u_\ell > s_i, \\ 0 & \text{otherwise.} \end{cases} \tag{73}$$

$$\propto \begin{cases} f(x_i \,|\, \theta^*_k) & \text{for } \theta_i = \theta^*_k, \text{ for some } k = 1, \ldots, K \text{ with } w_k > s_i, \\ f(x_i \,|\, \theta'_\ell) & \text{for } \theta_i = \theta'_\ell, \text{ for some } \ell = 1, 2, \ldots \text{ with } w' u_\ell > s_i, \\ 0 & \text{otherwise.} \end{cases} \tag{74}$$

In particular, note that we only need the atoms θ'_ℓ in G' with mass u_ℓ greater than $S = \min_i s_i / w'$, since all the other atoms will have probability 0 for being used in this second phase. Since $w' \leq 1$ and each s_i is positive with probability 1, we have that with probability 1 as well S is positive and there are only a finite number of atoms L in G' with mass greater than S. We can enumerate all these atoms by generating the atoms of G' iteratively using the stick-breaking representation, stopping once the total left over mass is less than S. While the number of atoms generated L is potentially unbounded, since the stick-breaking weights decrease exponentially quickly, in practice the number of such atoms is small. In summary, Algorithm 1 summarizes one iteration of the sampler.

7 Pitman-Yor Processes

While the Dirichlet process generates an infinite number of atoms, the rate at which new atoms are generated is relatively slow. Many real-world phenomena are characterized by

Algorithm 1 Conditional Slice Sampler for DP Mixture Model

1: Set $\theta_1^*, \ldots, \theta_K^*$ to be the collection of unique values among $\theta_1, \ldots, \theta_N$.
2: Sample each component parameter $\theta_1^*, \ldots, \theta_K^*$ independently using an ergodic MCMC update with the invariant distribution in Eq. (66).
3: Sample the weights $(w_1, \ldots, w_K, w')$ from Eq. (68).
4: Sample each slice variable $s_1, \ldots, s_N$ independently from Eq. (71).
5: Compute the minimum slice level $S = \min_i s_i$.
6: Sample the number L and weights $u_1, u_2, \ldots, u_L$ of atoms in G' using the stick-breaking construction, Eq. (69), where L is the smallest index with $1 - \sum_{\ell=1}^{L} u_\ell < S$.
7: Sample the parameters $\theta'_1, \ldots, \theta'_L$ in G' independently from G_0.
8: Sample each cluster assignment $\theta_1, \ldots, \theta_N$ independently from Eq. (74).

faster growth, namely that characterized by power law distributions (Newman, 2005). As we discuss in this section, while the Dirichlet process cannot generate such power laws, a generalization of the Dirichlet process known as the *Pitman-Yor process* can generate power laws (Pitman and Yor, 1997). Like the Dirichlet process, the Pitman-Yor process is a random probability measure that induces marginal distributions characterized by a (generalized) Chinese restaurant process and a Blackwell-MacQueen urn. In this section we build on our previous work and show how these three perspectives play out in the case of the Pitman-Yor process.

One way to understand the slow rate of growth of new atoms in the Dirichlet process is to consider the Chinese restaurant process in Eq. (3). We see that the probability of the $N+1$st customer selecting a new table is $\alpha/(\alpha+N)$ while the probability of selecting an occupied table goes as $N/(\alpha+N)$. The latter quickly dominates the former and few new tables emerge as N becomes large. Indeed the expected number of tables occupied by N customers is

$$\sum_{n=1}^{N} \frac{\alpha}{\alpha+n} \asymp \alpha \log(N) \tag{75}$$

which grows slowly with N. Similarly, in the stick-breaking formulation in Eq. (39), each weight w_k is obtained by multiplying the remaining stick length by a $\text{Beta}(1, \alpha)$ random variable, which has a large expected value (e.g., $1/2$ for the case $\alpha = 1$) and which quickly chews up the stick. To obtain a larger number of new tables, or a slower decay in the size of the sticks, we need to let α grow as the process proceeds. The issue is how to do this while retaining exchangeability.

The Pitman-Yor process (PYP) achieves this growth in the rate of generating new tables via a third parameter σ, known as the *discount parameter*, in addition to the concentration parameter α and the base distribution G_0. The discount parameter σ takes values in the range $[0, 1)$, while the range of the concentration parameter α is expanded to $(-\sigma, \infty)$. When $\sigma = 0$ the PYP reduces to the DP.

Let us first consider a generalized form of the Chinese restaurant process that makes use of the discount parameter. As before, let $\boldsymbol{\pi}_{[n]}$ denote a partition based on the first n customers, and consider the probabilities associated with customer $n+1$. In the Pitman-Yor

generalization we replace the conditional probabilities in Eq. (3) with:

$$P(\text{customer } n+1 \text{ joins table } c \,|\, \boldsymbol{\pi}_{[n]}) = \begin{cases} \dfrac{|c| - \sigma}{\alpha + n} & \text{if } c \in \boldsymbol{\pi}_{[n]}, \\ \dfrac{\alpha + \sigma K_n}{\alpha + n} & \text{otherwise,} \end{cases} \tag{76}$$

where K_n is the number of clusters in the partition $\boldsymbol{\pi}_{[n]}$. Note that the probability of joining an existing table is reduced by an amount proportional to σ relative to the CRP. The reductions are added to the probability of starting a new table, leading to an overall increase proportional to σK_n. This shift in probabilities allows the resulting number of occupied tables in the generalized CRP to increase, with the amount of increase larger for larger values of σ. Moreover, the reduction of the probabilities for sitting at existing tables is the same (proportional to σ) for all tables. This affects smaller tables more than large ones, so that in the generalized CRP there is a preponderance of more tables of smaller sizes, since these have a smaller chance of growing large. As we discuss in Section 7.3, this leads to the power law behavior of the PYP.

This Pitman-Yor CRP (PYCRP) retains the exchangeability property of the original CRP. To see this, we again calculate the probability of a partition $\boldsymbol{\pi}_{[N]}$ under the PYCRP by multiplying together the conditional probabilities Eq. (76) for $n = 1, \ldots, N$, giving:

$$P(\boldsymbol{\pi}_{[N]}) = \frac{\alpha(\alpha+\sigma)\cdots(\alpha+\sigma(K_N-1))}{\alpha^{(N)}} \prod_{c \in \boldsymbol{\pi}_{[N]}} (1-\sigma)(2-\sigma)\cdots(|c|-1-\sigma) \tag{77}$$

Comparing to Eq. (6), the factor $\alpha^{(N)} := \alpha(\alpha+1)\cdots(\alpha+N-1)$ again arises from the denominators in the table assignment probabilities, incrementing by one irrespective of whether an old table or a new table is selected. The factor $(1-\sigma)(2-\sigma)\cdots(|c|-1-\sigma)$ corresponds to the factor $(|c|-1)!$ from before, arising from the probabilities of customers sitting at table c, and the factor $\alpha(\alpha+\sigma)\cdots(\alpha+\sigma(K_N-1))$ corresponds to the factor α^{K_N} from before, arising from the decisions to start new tables. We see that the overall probability of a partition under the PYCRP depends only on the number of clusters and on the cluster sizes; thus the PYCRP defines an exchangeable distribution on partitions.

We can also develop a Blackwell-MacQueen urn model from the PYCRP in the same way as before, associating an iid sequence of random vectors $\{\phi_c\}$ with the tables in the PYCRP, where $\phi_c \sim G_0$. This defines a sequence $(\theta_1, \theta_2, \ldots, \theta_N)$, where $\theta_i = \phi_c$ if customer i sits at table c. From this construction we obtain an infinitely exchangeable sequence of random variables. We can therefore invoke de Finetti's theorem, and we can conclude that there exists a unique random probability measure G such that each θ_i is independently and identically distributed according to G. It is this random probability measure that we refer to as the Pitman-Yor process, which we denote as $\text{PYP}(\alpha, \sigma, G_0)$.

Finally, we can also develop a stick-breaking representation of the random probability measure G. By analogy with the earlier derivation of a stick-breaking representation for the DP in Section 5, suppose that the first customer has just been seated at a table and define a Pólya urn scheme for subsequent customers in which "one" denotes sitting at that table, and "zero" denotes sitting at some other table. Let Z_i denote the binary indicator associated with the ith subsequent customer. At the outset the two alternatives have weight

$1-\sigma$ and $\alpha+\sigma$, and thus the normalization is initially $\alpha+1$. The probability of observing a sequence of N ones under this Pólya urn is thus as follows:

$$P(Z_1 = 1, Z_2 = 1, \ldots, Z_N = 1) = \frac{(1-\sigma)(2-\sigma)\cdots(N-\sigma)}{(\alpha+1)(\alpha+2)\cdots(\alpha+N)}. \tag{78}$$

Now recall Eq. (26), where the de Finetti mixing measure is a Beta(α_1, α_2) distribution. Substituting $s = N$, and letting $\alpha_1 = 1-\sigma$ and $\alpha_2 = \alpha+\sigma$, we match the probability in Eq. (78). This holds for all N, and thus the first stick-breaking weight is a Beta$(1-\sigma, \alpha+\sigma)$ random variable.

Next, suppose that $k-1$ tables have been occupied and consider the moment just after the kth table has been occupied. We focus only on customers that do not sit at one of the first $k-1$ tables, and define a Pólya urn in which $\{Z_i = 1\}$ denotes sitting at the kth table, and $\{Z_i = 0\}$ denotes sitting at some subsequent table. The initial weights associated with these two events are $1-\sigma$ and $\alpha+k\sigma$, where the latter reflects the additional boost given to selecting a new table given that k tables are currently occupied. The normalization is thus initially $\alpha+1+(k-1)\sigma$. The probability of observing a sequence of N ones is thus:

$$P(Z_1 = 1, Z_2 = 1, \ldots, Z_N = 1) = \frac{(1-\sigma)(2-\sigma)\cdots(N-\sigma)}{(\alpha+1+(k-1)\sigma)(\alpha+2+(k-1)\sigma)\cdots(\alpha+N+(k-1)\sigma)}. \tag{79}$$

This probability is matched by choosing $s = N$, $\alpha_1 = 1-\sigma$ and $\alpha_2 = \alpha+k\sigma$ in Eq. (26). Thus the kth stick-breaking weight is obtained by multiplying the remaining stick length after the first $k-1$ breaks by a Beta$(1-\sigma, \alpha+\sigma k)$ random variable.

Summarizing, the stick-breaking representation for the PYP is given as follows:

$$\beta_k \overset{ind}{\sim} \text{Beta}(1-\sigma, \alpha+\sigma k) \qquad \text{for } k = 1, 2, \ldots \tag{80}$$

$$w_k = \beta_k \prod_{j=1}^{k-1}(1-\beta_j) \qquad \text{for } k = 1, 2, \ldots \tag{81}$$

$$\phi_k \overset{iid}{\sim} G_0 \qquad \text{for } k = 1, 2, \ldots \tag{82}$$

$$G = \sum_{k=1}^{\infty} w_k \delta_{\phi_k}. \tag{83}$$

This representation, described by Perman et al. (1992) and Pitman and Yor (1997), is the analog of the GEM distribution for the case of the PYP. We refer to it as the *Perman-Pitman-Yor* (PPY) distribution:

$$w \sim \text{PPY}(\alpha, \sigma), \tag{84}$$

where $w = (w_1, w_2, \ldots)$. We can now write an explicit formula for a draw from the Pitman-Yor process. Letting $\phi = (\phi_1, \phi_2, \ldots)$ denote an iid sequence of draws from the base measure G_0, we have:

$$G = \sum_{k=1}^{\infty} w_k \delta_{\phi_k}, \tag{85}$$

as the random probability measure promised by de Finetti's theorem for the PYCRP.

7.1 Posterior Pitman-Yor process

As in the case of the DP, our ability to make use of the Pitman-Yor process as an ingredient in Bayesian models requires an understanding of the posterior posterior distribution over the random probability measure given observations. We thus consider the following familiar hierarchical model:

$$\begin{aligned} G &\sim \mathrm{PYP}(\alpha, \sigma, G_0) \\ \theta_i \,|\, G &\overset{iid}{\sim} G \qquad\qquad \text{for } i = 1, \ldots, N. \end{aligned} \tag{86}$$

Since G is an atomic probability measure, it is possible for multiple θ_i's to take on the same value; again let $\theta_1^*, \ldots, \theta_K^*$ be the K unique observed values with $n_1, \ldots, n_K$ being their frequencies among $\theta = (\theta_1, \ldots, \theta_N)$. Conditioned on θ, we again expect G to contain an atom at each of the observed values $\theta_1^*, \ldots, \theta_K^*$. In addition, it still has an infinite number of atoms located elsewhere in the space Θ, hence we are led to write the posterior G as:

$$G \,|\, \theta = \sum_{k=1}^{K} w_k \delta_{\theta_k^*} + w' G' \tag{87}$$

for some random probability vector $(w_1, \ldots, w_K, w')$ and random atomic probability measure G'. The vector $(w_1, \ldots, w_K, w')$ and G' can be shown to be independent with the following conditional distributions:

$$(w_1, \ldots, w_K, w') \,|\, \theta \sim \mathrm{Dir}(n_1 - \sigma, \ldots, n_K - \sigma, \alpha + \sigma K) \tag{88}$$

$$G' \,|\, \theta \sim \mathrm{PYP}(\alpha + \sigma K, \sigma, G_0) \tag{89}$$

This posterior representation directly extends the posterior representation in Eq. (60) for the DP. The distribution over the vector $(w_1, \ldots, w_K, w')$ is Dirichlet, with parameter $n_k - \sigma$ for the component associated with θ_k^*, while the parameter for the component associated with the other atoms in G' is $\alpha + \sigma K$. These parameters are the exactly those appearing in the numerator of the probabilities of the PYCRP in Eq. (76), and indeed can be derived from the PYCRP.

In particular, the clustering structure for the observations θ correspond to a state of the PYCRP in which the first N customers in the restaurant sit at K tables, with $n_1, \ldots, n_K$ being the numbers of customers at these tables. Subsequently, consider an infinite sequence of customers enter the restaurant, and their occupancy is captured by the atoms in G'.

Define an urn scheme with $K+1$ possible colors, where the first K colors correspond to the K extant tables with initial weights $n_1 - \sigma, \ldots, n_K - \sigma$, while the last color corresponds to the event that a subsequent customer sits at some table other than the first K. With each subsequent customer one of these events occur, and the corresponding weight is incremented by one: for the first K colors the increment corresponds to an additional customer at the table, while for the last color either an extant table beyond the first K gets an additional customer, or a new table is occupied; in the latter case the initial weight is $1 - \sigma$ but the weight associated with occupying new tables is incremented by σ, so that the total weight for the last color is incremented by one also in this case. This urn scheme generalizes the binary Pólya urn scheme, and it is unsurprising that its de Finetti measure is the Dirichlet distribution with parameters given by its initial weights, as shown in Eq. (88).

We can also understand the distribution of G' in Eq. (89) via the PYCRP, specifically by relating it to the distribution of seating arrangements of customers around the tables of the Chinese restaurant beyond the first K tables. In particular, consider the subsequence of customers that do not sit at the first K tables. The initial probability of one such customer sitting at a new table is proportional to $\alpha + \sigma K$. With each new table occupied, its initial probability of being picked again is proportional to $1 - \sigma$, and this is incremented by one with each subsequent customer sitting at the table. Further, with each new table occupied, the weight of sitting at new tables is incremented by σ. Hence the seating arrangement among these tables is described precisely by a PYCRP with parameters $\alpha + \sigma K$ and σ, and so the corresponding random probability measure G' must have the distribution in Eq. (89) given by $\mathrm{PYP}(\alpha + \sigma K, \sigma, G_0)$.

7.2 Posterior sampling algorithms for PYP mixture models

The characterizations of the PYP and its posterior distribution that we have presented lead to a variety of Markov chain Monte Carlo algorithms for inference in PYP-based models. For example, Ishwaran and James (2001) developed a Gibbs sampler for the PYP mixture model based on the truncated stick-breaking construction, while the PYCRP representation leads to a direct generalization of the Gibbs sampler of Section 3.3. In this section we present a third example of a posterior inference algorithm, one that extends the conditional slice sampler of Section 6.4 to the PYP mixture model.

We consider the following model:

$$G \sim \mathrm{PYP}(\alpha, \sigma, G_0) \tag{90}$$

$$\theta_i \,|\, G \overset{iid}{\sim} G \qquad \text{for } i = 1, \ldots, N, \tag{91}$$

$$x_i \,|\, \theta_i \overset{ind}{\sim} F(\theta_i) \qquad \text{for } i = 1, \ldots, N. \tag{92}$$

As before, we augment the state space with additional slice variables $s = (s_1, \ldots, s_N)$, which are conditionally independent with s_i uniformly distributed between 0 and $G(\{\theta_i\})$:

$$s_i \,|\, G, \theta_i \sim \mathcal{U}[0, G(\{\theta_i\})]. \tag{93}$$

Let $\theta_1^*, \ldots, \theta_K^*$ be the unique values among $\theta = (\theta_1, \ldots, \theta_N)$, with each unique value corresponding to a component in the mixture model. Let z_i denote the component that x_i belongs to; i.e., $\theta_i = \theta_{z_i}^*$. The component parameters can be updated individually using any ergodic MCMC update with invariant distribution given by:

$$p(\theta_k^* \,|\, x, z) \propto g_0(\theta_k^*) \prod_{i:z_i=k} f(x_i \,|\, \theta_k^*). \tag{94}$$

As we have seen in Sec. (7.1), given θ, the posterior G can be represented as follows:

$$G \,|\, \theta = \sum_{k=1}^{K} w_k \delta_{\theta_k^*} + w' G', \tag{95}$$

where

$$(w_1, \ldots, w_K, w') \,|\, \theta \sim \mathrm{Dir}(n_1 - \sigma, \ldots, n_K - \sigma, \alpha + K\sigma) \tag{96}$$

$$G' \,|\, \theta \sim \mathrm{PYP}(\alpha + K\sigma, \sigma, G_0) \tag{97}$$

As in Section 6.4, we can simulate G' using the stick-breaking construction for the PYP:

$$G' \,|\, \theta = \sum_{\ell=1}^{\infty} u_\ell \delta_{\theta'_\ell} \tag{98}$$

$$(u_1, u_2, \ldots) \,|\, \theta \sim \mathrm{PPY}(\alpha + K\sigma, \sigma) \tag{99}$$

$$\theta'_\ell \,|\, \theta \overset{iid}{\sim} G_0 \qquad \text{for } \ell = 1, 2, \ldots, \tag{100}$$

with the simulation truncated, after say L atoms, once all remaining atoms can be guaranteed to have mass less than the slice variables, $1 - \sum_{\ell=1}^{L} u_\ell < \min_i s_i / w'$.

Finally, conditioned on G and s, we can sample the parameters θ individually, using the conditional distributions:

$$p(\theta_i \,|\, G, s_i, x_i) \propto \begin{cases} f(x_i \,|\, \theta^*_k) & \text{for } \theta_i = \theta^*_k, \text{ for some } k = 1, \ldots, K \text{ with } w_k > s_i, \\ f(x_i \,|\, \theta'_\ell) & \text{for } \theta_i = \theta'_\ell, \text{ for some } \ell = 1, 2, \ldots \text{ with } w' u_\ell > s_i, \\ 0 & \text{otherwise.} \end{cases} \tag{101}$$

7.3 Power law properties

As alluded to in the beginning of this section, when $\sigma > 0$ the PYP yields power-law behavior (Pitman, 2002). Due to these power law properties, the PYP has had numerous applications, in particular in the modeling of various linguistic phenomena (Goldwater et al., 2006a; Teh, 2006; Cohn et al., 2010), image segmentation (Sudderth and Jordan, 2009), and PET analysis (Fall et al., 2009). The power-law nature of the PYP can be expressed in several ways, all of which depend crucially on the discount parameter σ. In this section we discuss these power laws; see Figure 5 for an illustration.

To begin, we show that under the PPY stick-breaking representation in Eq. (81) we have $\mathbb{E}[w_k] \in \mathcal{O}(k^{-1/\sigma})$ if $\sigma > 0$, which indicates that cluster sizes decay slowly according to a power law, with a slower decay when σ is larger. Indeed, noting that w_k is the product of independent beta random variables, we have:

$$\mathbb{E}[w_k] = \mathbb{E}[\beta_k] \prod_{i=1}^{k-1} \mathbb{E}[1 - \beta_i] = \frac{1-\sigma}{\alpha + 1 + (k-1)\sigma} \prod_{i=1}^{k-1} \left(1 - \frac{1-\sigma}{\alpha + 1 + (i-1)\sigma}\right) \tag{102}$$

$$\begin{aligned} \log \mathbb{E}[w_k] &= \log \frac{1-\sigma}{\alpha + 1 + (k-1)\sigma} + \sum_{i=1}^{k-1} \log \left(1 - \frac{1-\sigma}{\alpha + 1 + (i-1)\sigma}\right) \\ &\asymp -\log k + \sum_{i=1}^{k-1} \frac{1 - 1/\sigma}{\alpha/\sigma + 1/\sigma + i - 1} \asymp -\log k + (1 - 1/\sigma) \log k = -1/\sigma \log k \end{aligned} \tag{103}$$

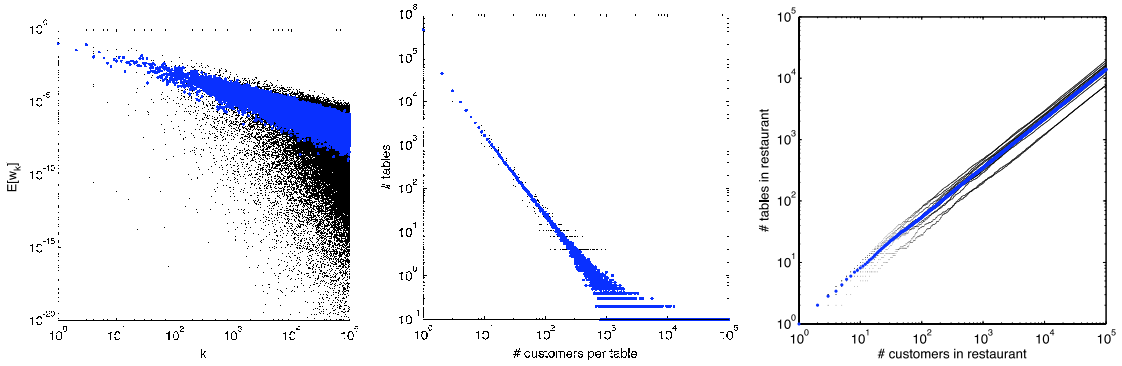

Figure 5: Power law behaviors of the Pitman-Yor process. Left: $\mathbb{E}[w_k]$ vs. k. Middle: number of tables vs. number of customers at each table. Right: total number of tables in restaurant vs. number of customers in restaurant. Each plot shows the results of 10 draws (small black dots) and their mean (large blue dots). The log-log plots are well approximated by straight lines, indicating power laws.

Thus $\mathbb{E}[w_k] \in \mathcal{O}(k^{-1/\sigma})$. By way of comparison, in the case of the DP, where $\sigma = 0$, we have $\mathbb{E}[w_k] = \frac{1}{1+\alpha}(\frac{\alpha}{1+\alpha})^{k-1}$, which decays exponentially quickly in k.

The Pitman-Yor process also yields Heaps' Law (Heaps, 1978), which, in the language of the CRP, states that the total number of tables in a restaurant with N customers scales as $\mathcal{O}(N^d)$. Denoting this number as K_N, let us derive $\mathbb{E}[K_N]$. Note that the expected number of tables when there are $N+1$ customers is the expected number when there are N customers, plus the probability that the $N+1$st customer sits at a new table. This gives a recursion relating $\mathbb{E}[K_{N+1}]$ to $\mathbb{E}[K_N]$ which can be solved. Specifically,

$$\mathbb{E}[K_{N+1}] = \mathbb{E}[K_N] + \mathbb{E}\left[\frac{\alpha + \sigma K_N}{\alpha + N}\right] = \frac{\alpha + \sigma + N}{\alpha + N}\mathbb{E}[K_N] + \frac{\alpha}{\alpha + N}. \tag{104}$$

When $\sigma = 0$, this equation simplifies to

$$\mathbb{E}[K_N] = \alpha \sum_{i=0}^{N-1} \frac{1}{\alpha + i} \asymp \alpha \log N. \tag{105}$$

On the other hand, when $\sigma > 0$ it is easy to show by induction that

$$\mathbb{E}[K_N] = \frac{\Gamma(N + \alpha + \sigma)\Gamma(\alpha + 1)}{\sigma\Gamma(\alpha + \sigma)\Gamma(N + \alpha)} - \frac{\alpha}{\sigma}. \tag{106}$$

Using Stirling's formula for large values of N,

$$\mathbb{E}[K_N] \asymp \frac{\Gamma(\alpha + 1)}{\sigma\Gamma(\alpha + \sigma)} \frac{\sqrt{2\pi(N + \alpha + \sigma - 1)}(\frac{N+\alpha+\sigma-1}{e})^{N+\alpha+\sigma-1}}{\sqrt{2\pi(N + \alpha - 1)}(\frac{N+\alpha-1}{e})^{N+\alpha-1}} \tag{107}$$

$$\asymp \frac{\Gamma(\alpha + 1)}{\sigma\Gamma(\alpha + \sigma)} N^{\sigma}. \tag{108}$$

Finally, the power law known as Zipf's law (Zipf, 1932) can be derived from the PYCRP as well. This law states that the proportion of tables with m customers scales as $\mathcal{O}(m^{-1-\sigma})$, such that most tables contain only a small number of customers. This conforms to our previous observation that the discounting by σ of the probabilities of customers joining existing tables affects smaller tables more than larger ones, so that smaller tables have less chance at growing large when $\sigma > 0$. The derivation of Zipf's law is somewhat involved and we defer it to Appendix F.

8 Hierarchical Dirichlet Processes

Our treatment of the Dirichlet process has culminated in the notion of a *random measure*. Such an object is very natural from a Bayesian perspective—a Bayesian assumes that a probability measure underlies the data that one observes, and when the measure is unknown it should be treated as random. While classical Bayesian modeling is based on treating probability measures as being indexed with finite-dimensional parameters that are then treated as random, it is quite natural to dispense with the intermediate notion of a parameter and treat the measures themselves as random. For this perspective to be useful in practice, it is necessary that random measures can be used as ingredients in building more elaborate models; in particular, it is necessary that random measures can be readily integrated with hierarchical modeling ideas.

Hierarchical modeling is a fundamental concept in Bayesian statistics. The basic idea is that parameters are endowed with distributions which may themselves introduce new parameters, and this construction recurses. We have already seen an example of hierarchical modeling in Sec. (3.2) and Sec. (6), where we discussed conditional independence hierarchies. In the conditional independence motif a set of parameters are coupled by making their distributions depend on a shared underlying parameter. These distributions are often taken to be identical, based on an assertion of exchangeability and an appeal to de Finetti's theorem. Intuitively, the rationale for introducing such coupling is to "share statistical strength"—it allows the sharpening of the posterior distribution for one parameter to be transferred to other parameters. In the Bayesian nonparametric setting, we wish to share statistical strength among multiple random measures.

Let us consider a concrete example of such sharing. We return to the clustering setting that motivated our development of the Dirichlet process, and consider the design of a fraud detection system, where the data take the form of a set of transactions and the goal is to discover "patterns" in the transaction data that may be instances of fraud and which should be flagged. It is natural to treat such a problem as a clustering problem, where a cluster corresponds to a particular type of transaction. But it is also often the case in practice that there are additional "contextual" variables available that can help sharpen the inference of clusters; for example, seasonal variables, geographic variables, etc. Conditioning on the contextual variables, it may be possible to obtain a higher-quality clustering, leading to fewer false positives in subsequent usage of the fraud detection system. Thus, instead of thinking in terms of a single clustering problem, it is natural to think in terms of multiple, context-dependent clustering problems. However, by breaking the data down in this way, some of the individual clustering problems may be associated with very few data points. Thus, while lumping the data together loses precision, separating the data can increase

variance. The goal then is to develop a method that compromises between lumping and splitting, a method in which clusters can be shared among multiple context-dependent clustering problems.[9]

We could attempt to develop such a method by starting with K-means and inventing a mechanism to share clusters across multiple instances of K-means clustering problems. But we would again encounter the problem of determining K; indeed, we would need to determine a value of K for each clustering problem. Moreover, the appropriate value should somehow reflect the sharing pattern among the clustering problems. It seems unclear as to how to develop such a method. In this section we show that the problem can instead be approached in an elegant way using the tools of Bayesian nonparametrics. In particular, we consider a simple conditional independence hierarchy in which a set of random measures are coupled via an underlying random measure. It is the atoms of the underlying random measure that are shared due to this coupling; these atoms correspond to clusters and thus we obtain the desired sharing of clusters.

Let us formalize the hierarchical model that we have just alluded to. We imagine that the data can be subdivided in a countable collection of *groups*. We model the groups of data by considering a collection of DPs, $\{G_j : j \in \mathcal{J}\}$, defined on a common space Θ, where $\mathcal{J}$ indexes the groups. We define a *hierarchical Dirichlet process* (HDP) as follows:

$$G_0 \,|\, \gamma, H \sim \mathrm{DP}(\gamma, H) \tag{109}$$

$$G_j \,|\, \alpha, G_0 \sim \mathrm{DP}(\alpha, G_0) \qquad \text{for } j \in \mathcal{J}, \tag{110}$$

where H and γ are the global base measure and concentration parameter. This hierarchical model induces sharing of atoms among the random measures G_j since each inherits its set of atoms from the same G_0. Thinking in terms of the stick-breaking construction, when a new atom is to be drawn in forming the random measure G_j, this atom is drawn from the base measure G_0, but this base measure is itself composed of atoms, being a draw from the Dirichlet process. Thus each atom in G_j is an atom of G_0, and the atoms that form G_0 are shared among all the random measures G_j.

We complete the model by generating the data in the jth group by drawing parameters from G_j and then drawing data points from a distribution indexed by these parameters:

$$\begin{aligned} \theta_{jn} \,|\, G_j &\sim G_j && \text{for } n = 1, \ldots, N_j, \\ x_{jn} \,|\, \theta_{jn} &\sim F(\theta_{jn}) && \text{for } n = 1, \ldots, N_j, \end{aligned} \tag{111}$$

where N_j is the number of data points in the jth group. Recalling that two data points are treated as belonging to the same cluster if they are associated with the same underlying atom, we see that not only can data points within a single group belong the same cluster, (i.e., $\theta_{jn} = \theta_{jn'}$), but data points between two different groups can also be assigned to the same cluster (i.e., $\theta_{jn} = \theta_{j'n'}$). See Figure 6 for a graphical representation of the overall HDP mixture model.

[9] We have chosen this example given that no specialized domain knowledge is needed to understand the problem. For a full development of a real-world application of this kind in the domain of protein structural modeling, see Ting et al. (2010). Another real-world application is the modeling of document collections via topic models; see Teh et al. (2006) and Sec. (8.4).

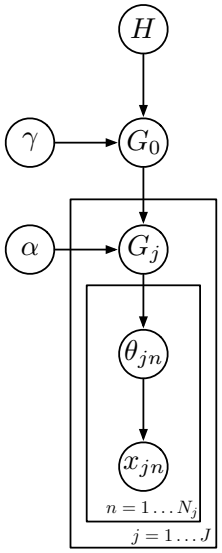

Figure 6: A graphical model representation of the HDP mixture model.

Note that the recursive construction of the HDP can be generalized to arbitrary hierarchies in the obvious way. Each G_j is given a DP prior with base measure $G_{\mathrm{pa}(j)}$, where $\mathrm{pa}(j)$ is the parent index of j in the hierarchy. As in the two-level hierarchy in Eq. (109), the set of atoms at the top level is shared throughout the hierarchy, while the multi-level hierarchy allows for a richer dependence structure on the weights of the atoms. Such representations are often natural in applications, arising when the contextual variables that define the groups of data have a nested structure.

Other ways to couple multiple Dirichlet processes have been proposed in the literature; in particular the *dependent Dirichlet process* of MacEachern et al. (2001) provides a general formalism based on the stick-breaking representation. Ho et al. (2006) gives a complementary view of the HDP and its Pitman-Yor generalizations in terms of coagulation operators.

To understand the precise nature of the sharing induced by the HDP it is helpful to consider representations akin to the stick-breaking and Chinese restaurant representations of the DP. We consider these representations in the remainder of this section.

8.1 Stick-breaking representation

In this section we develop a stick-breaking construction for the HDP. This representation provides a concrete representation of draws from an HDP and it provides insight into the sharing of atoms across multiple DPs.

We begin with the stick-breaking representation for the random base measure G_0. Given that this base measure is distributed according to $\mathrm{DP}(\gamma, H)$, we have:

$$G_0 = \sum_{k=1}^{\infty} \tau_k \delta_{\psi_k}, \tag{112}$$

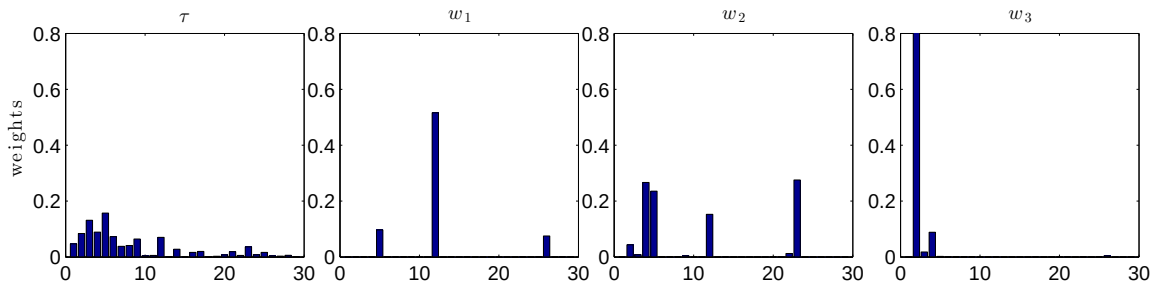

Figure 7: The HDP stick-breaking construction. The left panel depicts a draw of τ, and the remaining panels depict draws of w_1, w_2 and w_3 conditioned on τ.

where

$$v_k \,|\, \gamma \sim \text{Beta}(1, \gamma), \qquad \tau_k = v_k \prod_{l=1}^{k-1}(1 - v_l), \qquad \text{for } k = 1, 2, \ldots, \tag{113}$$

$$\psi_k \,|\, H \sim H, \qquad \text{for } k = 1, 2, \ldots.$$

The random measures G_j are also (conditionally) distributed according to a DP. Moreover, the support of each G_j is contained within the support of its base distribution G_0. Thus the stick-breaking representation for G_j is a reweighted sum of the atoms in G_0:

$$G_j = \sum_{k=1}^{\infty} w_{jk} \delta_{\psi_k}. \tag{114}$$

The problem reduces to finding a relationship between the global weights $\tau = (\tau_1, \tau_2, \ldots)$ and the group-specific weights $w_j = (w_{j1}, w_{j2}, \ldots)$. Let us interpret these weight vectors as probability measures on the discrete space of natural numbers $\{1, 2, \ldots\}$. Each partition on Θ induces a partition on the atoms of G_0, which in turn induces a partition on the natural numbers. Hence the fact that the DP has Dirichlet marginal distributions, as shown in Eq. (49), implies

$$w_j \,|\, \alpha, \tau \sim \text{DP}(\alpha, \tau). \tag{115}$$

Some algebra then readily yields the following explicit construction for w_j conditioned on τ:

$$\beta_{jk} \,|\, \alpha, \tau_1, \ldots, \tau_k \sim \text{Beta}\left(\alpha\tau_k, \alpha\left(1 - \sum_{l=1}^{k} \tau_l\right)\right), \qquad \text{for } k = 1, 2, \ldots, \tag{116}$$

$$w_{jk} = \beta_{jk} \prod_{l=1}^{k-1}(1 - \beta_{jl}).$$

Figure 7 shows a sample draw of τ along with draws from w_1, w_2 and w_3 given τ.

From Eq. (113) we see that the mean of τ_k is $\mathbb{E}[\tau_k] = \gamma^{k-1}(1 + \gamma)^{-k}$ which decreases exponentially in k. The mean for w_j is simply its base measure τ; thus $\mathbb{E}[w_{jk}] = \mathbb{E}[\tau_k] =$

$\gamma^{k-1}(1+\gamma)^{-k}$ as well. However the law of total variance shows that w_{jk} has higher variance than τ_k: $\mathsf{Var}[w_{jk}] = \mathbb{E}[\frac{\tau_k(1-\tau_k)}{1+\alpha}] + \mathsf{Var}[\tau_k] > \mathsf{Var}[\tau_k]$. The higher variance is reflected in Figure 7 by the sparser nature of w_j relative to τ.

8.2 Chinese restaurant franchise

Recall that the Chinese restaurant process (CRP) describes the marginal probabilities of the DP in terms of a random partition obtained from a sequence of customers sitting at tables in a restaurant. There is an analogous representation for the HDP which we refer to as a *Chinese restaurant franchise* (CRF). In a CRF the metaphor of a Chinese restaurant is extended to a set of restaurants, one for each DP G_j indexed by $j \in \mathcal{J}$. The customers in the jth restaurant sit at tables in the same manner as the CRP, and this is done independently in the restaurants, reflecting the fact that each G_j is a conditionally independent DP given G_0. The coupling among restaurants is achieved via a franchise-wide menu. The first customer to sit at a table in a restaurant chooses a dish from the menu and all subsequent customers who sit at that table share that dish. Dishes are chosen with probability proportional to the number of tables (franchise-wide) which have previously served that dish.

More formally, label the nth customer in the jth restaurant with a random variable θ_{jn} that is distributed according to G_j. Similarly, let ϕ_{jt} denote a random variable corresponding to the tth table in the jth restaurant; these variables are independently and identically distributed (iid) according to G_0. Finally, the dishes are iid variables ψ_k distributed according to the base measure H. We couple these variables as follows. Each customer sits at one table and each table serves one dish; let customer n in restaurant j sit at table t_{jn}, and let table t serve dish k_{jt}. Then let $\theta_{jn} = \phi_{jt_{jn}} = \psi_{k_{jt_{jn}}}$. Let N_{jtk} be the number of customers in restaurant j seated around table t and being served dish k, let M_{jk} be the number of tables in restaurant j serving dish k, and let K be the number of unique dishes served in the entire franchise. We denote marginal counts with dots; e.g., $N_{j\cdot k}$ is the number of customers in restaurant j served dish k.

To show that the CRF captures the marginal probabilities of the HDP, we integrate out the random measures G_j and G_0 in turn from the HDP. We start by integrating out the random measure G_j; this yields a set of conditional distributions for $\theta_{j1}, \theta_{j2}, \ldots$ described by a Blackwell-MacQueen urn scheme:

$$\theta_{j,n+1} \,|\, \theta_{j1}, \ldots, \theta_{jn}, \alpha, G_0 \sim \sum_{t=1}^{M_{j\cdot}} \frac{N_{jt\cdot}}{\alpha + N_{j\cdot\cdot}} \delta_{\phi_{jt}} + \frac{\alpha}{\alpha + N_{j\cdot\cdot}} G_0. \tag{117}$$

A draw from this mixture can be obtained by drawing from the terms on the right-hand side with probabilities given by the corresponding mixing proportions. If a term in the first summation is chosen then the customer sits at an already occupied table: we increment $N_{jt\cdot}$, set $\theta_{jn} = \phi_{jt}$ and let $t_{jn} = t$ for the chosen table with index t. If the second term is chosen then the customer sits at a new table, associated with a new draw from G_0: We increment $M_{j\cdot}$ by one, set $N_{jM_{j\cdot}\cdot} = 1$, draw $\phi_{jM_{j\cdot}} \sim G_0$, set $\theta_{jn} = \phi_{jM_{j\cdot}}$ and $t_{jn} = M_{j\cdot}$.

Notice that each ϕ_{jt} is drawn iid from G_0 in the Blackwell-MacQueen urn scheme in Eq. (117), and this is the only reference to G_0 in that generative process. Thus we can

readily integrate out G_0 as well, obtaining a Blackwell-MacQueen urn scheme for the ϕ_{jt}'s:

$$\phi_{j,t+1} \,|\, \phi_{11}, \ldots, \phi_{1M_1.}, \ldots \phi_{j1}, \ldots, \phi_{jt}, \gamma, H \sim \sum_{k=1}^{K} \frac{M_{.k}}{\gamma + M_{..}} \delta_{\psi_k} + \frac{\gamma}{\gamma + M_{..}} H, \tag{118}$$

where we have presumed for ease of notation that $\mathcal{J} = \{1, \ldots, |\mathcal{J}|\}$, and we have imposed an ordering among the tables of the franchise. As promised at the beginning of this section, we see that the kth dish is chosen with probability proportional to the number of tables franchise-wide that previously served that dish ($M_{.k}$).

The above approach of integrating out the random measures in turn is a hierarchical generalization of the Blackwell-MacQueen urn scheme from the DP to the HDP. Each draw from each random probability measure is associated with a parameter vector ultimately associated with draws from the global base measure H. We can map this representation back to a partition-based one by ignoring the vector values themselves and concentrating on the partitions described by the indexing variables t_{jn}'s and k_{jt}'s. Specifically, for each j we have a partition $\boldsymbol{\pi}_j$ of $[N_j]$, while the base distribution G_0 is associated with a partition of the disjoint union[10] $\sqcup_{j \in \mathcal{J}} \boldsymbol{\pi}_j$. In practice, we do not make a distinction between these two representations of the HDP.

The CRF is useful in understanding scaling properties of the clustering induced by an HDP. Recall that in a DP the number of clusters scales logarithmically (see Sec. (7.3)). Thus $M_{j.} \in \mathcal{O}(\alpha \log N_j)$ where $M_{j.}$ and N_j are respectively the total number of tables and customers in restaurant j. Since G_0 is itself a draw from a DP, we have that $K \in \mathcal{O}(\gamma \log \sum_j M_{j.}) = \mathcal{O}(\gamma \log(\alpha \sum_j \log N_j))$. If we assume that there are J groups and that the groups (the customers in the different restaurants) have roughly the same size N, $N_j \in \mathcal{O}(N)$, we see that $K \in \mathcal{O}(\gamma \log \alpha J \log N) = \mathcal{O}(\gamma \log \alpha + \gamma \log J + \gamma \log \log N)$. Thus the number of clusters scales doubly logarithmically in the size of each group, and logarithmically in the number of groups. The HDP thus expresses a prior belief that the number of clusters grows very slowly in N.

8.3 Posterior structure of the HDP

The Chinese restaurant franchise is obtained by integrating out the random measures G_j and then integrating out G_0. Integrating out the random measures G_j yields a Chinese restaurant for each group as well as a sequence of iid draws from the base measure G_0, which are used recursively in integrating out G_0. Having obtained the CRF, it is of interest to derive the conditional distributions for the random measures given the CRF. This not only illuminates the combinatorial structure of the HDP but it also prepares the ground for a discussion of posterior inference algorithms (see Sec. (**??**)), where it can be useful to instantiate the CRF and the random measures explicitly.

The state of the CRF consists of the dish labels $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_K)$, the table t_{jn} at which the nth customer sits, and the dish k_{jt} served at the tth table in the jth restaurant. As functions of the state of the CRF, we also have the numbers of customers $\mathbf{N} = (N_{jtk})$, the numbers of tables $\mathbf{M} = (M_{jk})$, the customer labels $\boldsymbol{\theta} = (\theta_{jn})$ and the table labels

[10] A disjoint union is one where elements from distinct sets are considered distinct.

$\boldsymbol{\phi} = (\phi_{jt})$. The relationship between the customer labels and the table labels is given as follows: $\phi_{jt} = \psi_{jk_{jt}}$ and $\theta_{jn} = \phi_{jt_{jn}}$.

Consider the distribution of G_0 conditioned on the state of the CRF. G_0 is independent from the rest of the CRF when we condition on the iid draws $\boldsymbol{\phi}$, because the restaurants interact with G_0 only via the iid draws. The posterior thus follows from the usual posterior for a DP given iid draws:

$$G_0 \,|\, \gamma, H, \boldsymbol{\phi} \sim \mathrm{DP}\left(\gamma + M_{\cdot\cdot}, \frac{\gamma H + \sum_{k=1}^{K} M_{\cdot k}\delta_{\psi_k}}{\gamma + M_{\cdot\cdot}}\right). \tag{119}$$

Note that values for $\mathbf{M}$ and $\boldsymbol{\psi}$ are determined given $\boldsymbol{\phi}$, since they are simply the unique values and their counts among $\boldsymbol{\phi}$[11]. A draw from Eq. (119) can be constructed using the posterior DP Eq. (60):

$$\begin{aligned}(\beta_1, \ldots, \beta_K, \beta') \,|\, \gamma, G_0, \boldsymbol{\phi} &\sim \mathrm{Dir}(M_{\cdot 1}, \ldots, M_{\cdot K}, \gamma) \\ G_0' \,|\, \gamma, H &\sim \mathrm{DP}(\gamma, H) \\ G_0 &= \sum_{k=1}^{K} \beta_k \delta_{\psi_k} + \beta' G_0'.\end{aligned} \tag{120}$$

We see that the posterior for G_0 is a mixture of atoms corresponding to the dishes and an independent draw from $\mathrm{DP}(\gamma, H)$.

Conditioning on this draw of G_0 as well as the state of the CRF, the posterior distributions for the G_j's are independent. In particular, the posterior for each G_j follows from the usual posterior for a DP as well, given its base measure G_0 and iid draws $\boldsymbol{\theta}_j$:

$$G_j \,|\, \alpha, G_0, \boldsymbol{\theta}_j \sim \mathrm{DP}\left(\alpha + N_{j\cdot\cdot}, \frac{\alpha G_0 + \sum_{k=1}^{K} N_{j\cdot K}\delta_{\psi_k}}{\alpha + N_{j\cdot\cdot}}\right). \tag{121}$$

Note that $\mathbf{N}_j$ and $\boldsymbol{\psi}$ are simply the unique values and their counts among the $\boldsymbol{\theta}_j$. Making use of the decomposition of G_0 into G_0' and atoms located at the dishes $\boldsymbol{\psi}$, a draw from Eq. (121) can thus be constructed following Eq. (60):

$$\begin{aligned}(w_{j1}, \ldots, w_{jK}, w_j') \,|\, \alpha, \boldsymbol{\theta}_j &\sim \mathrm{Dir}(\alpha\beta_1 + N_{j\cdot 1}, \ldots, \alpha\beta_K + N_{j\cdot K}, \alpha\beta') \\ G_j' \,|\, \alpha, G_0 &\sim \mathrm{DP}(\alpha\beta', G_0') \\ G_j &= \sum_{k=1}^{K} w_{jk}\delta_{\psi_k} + w_j' G_j'.\end{aligned} \tag{122}$$

We see that G_j is a mixture of atoms at $\boldsymbol{\psi}$ and an independent draw from a DP, where the concentration parameter is $\alpha\beta'$. The posterior over the entire HDP is obtained by averaging the conditional distributions of G_0 and G_j over the posterior state of the Chinese restaurant franchise given $\boldsymbol{\theta}$.

This derivation shows that the posterior for the HDP can be split into a "discrete part" and a "continuous part." The discrete part consists of atoms at the unique values $\boldsymbol{\psi}$, with

[11] Here we make the simplifying assumption that H is a continuous distribution so that draws from H are unique. If H is not continuous then additional bookkeeping is required.

different weights on these atoms for each DP. The continuous part is a separate draw from an HDP with the same hierarchical structure as the original HDP and global base distribution H, but with altered concentration parameters, and consists of an infinite series of atoms at locations drawn iid from H. Although we have presented this posterior representation for a two-level hierarchy, the representation extends immediately to general hierarchies.

8.4 Nonparametric topic modeling

To provide a concrete example of the HDP we present an application to the problem of *topic modeling*. A topic model, also known as a *mixed membership model* or an *admixture model*, is a generalization of a finite mixture model in which each data point is associated with multiple draws from a mixture model instead of a single draw (Blei et al., 2003; Erosheva, 2003; Pritchard et al., 2000). As we will see, while the DP is the appropriate tool to extend finite mixture models to the nonparametric setting, the appropriate tool for nonparametric topic models is the HDP.

To motivate the topic model formulation, consider the problem of modeling the word occurrences in a set of newspaper articles (e.g., for the purposes of classifying future articles). A simple clustering methodology might attempt to place each article in a single cluster. But it would seem more useful to be able to cross-classify articles according to "topics"; for example, an article might be mainly about Italian food, but it might also refer to health, history and the weather. Moreover, as this example suggests, it would be useful to be able to assign numerical values to the degree to which an article treats each topic.

Topic models achieve this goal as follows. Define a *topic* to be a probability distribution across a set of *words* taken from some vocabulary W. A *document* is modeled as a probability distribution across topics. In particular, let us assume the following generative model for the words in a document. First choose a probability vector π from the K-dimensional simplex, and then repeatedly (1) select one of the K topics with probabilities given by the components of π and (2) choose a word from the distribution defined by the selected topic. The vector π thus encodes the expected fraction of words in a document that are allocated to each of the K topics. In general a document will be associated with multiple topics.

To fully specify a topic model we require a distribution for π. Taking this distribution to be symmetric Dirichlet, we obtain the *latent Dirichlet allocation* (LDA) model, developed by Blei et al. (2003) and Pritchard et al. (2000) as a model for documents and admixture in genetics, respectively. This model has been widely used not only in the fields of information retrieval and statistical genetics, but also across many other fields. For example, in machine vision, a "topic" is a distribution across visual primitives, and an image can be modeled as a distribution across topics (Fei-Fei and Perona, 2005), while in network analysis, a "topic" indicates a social role and an individual's interactions with others can be modeled as a distribution over the roles that the individual plays (Airoldi et al., 2006).

Let us now turn to the problem of developing a Bayesian nonparametric version of LDA in which the number of topics is allowed to be open-ended. As we have alluded to, this requires the HDP, not merely the DP. To see this, consider the generation of a single word in a given document. According to LDA, this is governed by a finite mixture model, in which one of K topics is drawn and then a word is drawn from the corresponding topic distribution. Generating all of the words in a single document requires multiple draws

from this finite mixture. If we now consider a different document, we again have a finite mixture, with the same mixture components (the topics), but with a different set of mixing proportions (the document-specific vector π). Thus we have multiple finite mixture models with the same components shared across the mixture models.

In the nonparametric setting, a naïve approach is to replace each finite mixture model with an independent DP mixture model. As the atoms in a DP are iid draws from the base distribution, if the base distribution is smooth then the atoms across different DP mixture models will be distinct. As these correspond to the components of the mixture models, this implies that the components are not shared across different mixture models. To enforce the sharing of mixture components, we can allow the DPs to share a common base distribution G_0, which is itself DP distributed and so will be discrete. The discreteness of G_0 then implies that its atoms are shared across all the individual DPs, so that we now have the desired property that the mixture components are shared across the mixture models. We are thus led to the following model, which we refer to as HDP-LDA:

$$\begin{aligned} G_0 \,|\, \gamma, H &\sim \mathrm{DP}(\gamma, H), && \\ G_j \,|\, \alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0), && \text{for each document } j \in \mathcal{J}, \\ \theta_{jn} \,|\, G_j &\sim G_j, && \text{for each word } n = 1, \ldots, N_j, \\ x_{jn} \,|\, \theta_{jn} &\sim F(\theta_{jn}), && \end{aligned} \tag{123}$$

where x_{jn} is the nth word in document j, H is the prior distribution over topics and $F(\theta_{jn})$ is the distribution over words in the topic parametrized by θ_{jn}. The model is presented as a graphical model in Fig. (6). Note that the atoms present in the random distribution G_0 are shared among the random distributions G_j. Thus, as desired, we have a collection of tied mixture models, one for each document.

8.5 Hidden Markov models with infinite state spaces

Hidden Markov models (HMMs) are widely used to model sequential data and time series data (Rabiner, 1989). An HMM is a doubly-stochastic Markov chain in which a state sequence, $\theta_1, \theta_2, \ldots, \theta_\tau$, is drawn according to a Markov chain on a discrete state space. A corresponding sequence of observations, $x_1, x_2, \ldots, x_\tau$, is drawn conditionally on the state sequence, where for all t the observation x_t is conditionally independent of the other observations given the state θ_t, with "emission distribution" parametrized by θ_t.

In this section we describe how to use the HDP to obtain an "infinite HMM"—an HMM with a countably infinite state space (Beal et al., 2002; Teh et al., 2006). The idea is similar in spirit to the passage from a finite mixture model to a DP mixture model. However, as we show, the appropriate nonparametric tool is again the HDP, not the DP. The resulting model is thus referred to as the *hierarchical Dirichlet process hidden Markov model* (HDP-HMM). We present both the HDP formulation and a stick-breaking formulation in this section; the latter is particularly helpful in understanding the relationship to finite HMMs. It is also worth noting that a Chinese restaurant franchise (CRF) representation of the HDP-HMM can be developed, and indeed Beal et al. (2002) presented a precursor to the HDP-HMM that was based on an urn model akin to the CRF.

To understand the need for the HDP rather than the DP, note first that a classical HMM specifies a set of finite mixture distributions, one for each value of the current state

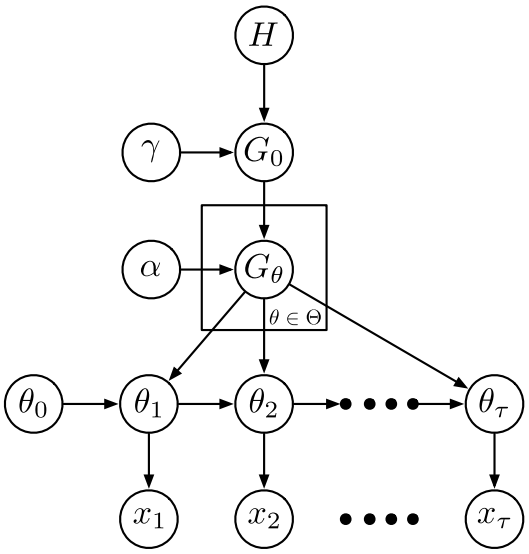

Figure 8: HDP hidden Markov model.

θ_t. Indeed, given θ_t, the observation x_{t+1} is chosen by first picking a state θ_{t+1} and then choosing x_{t+1} conditional on that state. Thus the probabilities of transiting from θ_t to θ_{t+1} play the role of the mixing proportions and the emission distributions play the role of the mixture components. It is natural to consider replacing this finite mixture model by a DP mixture model. In so doing, however, we must take into account the fact that we obtain a *set* of DP mixture models, one for each value of the current state. If these DP mixture models are not tied in some way, then the set of states accessible in a given value of the current state will be disjoint from those accessible for some other value of the current state. We would obtain a branching structure rather than a chain structure. The solution to this problem is as for topic models—we use the HDP to tie the DPs.

More formally, let us consider a collection of random transition kernels, $\{G_\theta : \theta \in \Theta\}$, drawn from an HDP:

$$\begin{aligned} G_0 \,|\, \gamma, H &\sim \mathrm{DP}(\gamma, H), \\ G_\theta \,|\, \alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0), \qquad \text{for } \theta \in \Theta, \end{aligned} \tag{124}$$

where H is a base measure on the probability space Θ. As for topic models, the random base measure G_0 allows the transitions out of each state to share the same set of next states. Let $\theta_0 = \psi_0 \in \Theta$ be a predefined initial state. The conditional distributions of the sequence of latent state variables $\theta_1, \ldots, \theta_\tau$ and observed variables $x_1, \ldots, x_\tau$ are:

$$\begin{aligned} \theta_t \,|\, \theta_{t-1}, \{G_\theta : \theta \in \Theta\} &\sim G_{\theta_{t-1}} \qquad \text{for } t = 1, \ldots, \tau, \\ x_t \,|\, \theta_t &\sim F(\theta_t). \end{aligned} \tag{125}$$

A graphical model representation for the HDP-HMM is shown in Figure 8.

We have defined a probability model consisting of an uncountable number of DPs, which should raise concerns among the more measure-theoretically minded readers. These

concerns can be dealt with, however, essentially due to the fact that the sample paths of the HDP-HMM only ever encounter a finite number of states. To see this more clearly, and to understand the relationship of the HDP-HMM to the parametric HMM, it is helpful to consider a stick-breaking representation of the HDP-HMM. This representation is obtained directly from the stick-breaking representation of the underlying HDP:

$$\begin{aligned} G_0 &= \sum_{k=1}^{\infty} \beta_k \delta_{\psi_k}, \\ G_{\psi_\ell} &= \sum_{k=1}^{\infty} w_{\ell k} \delta_{\psi_k}, \qquad \text{for } \ell = 0, 1, 2, \ldots, \end{aligned} \tag{126}$$

where

$$\begin{aligned} \psi_k \,|\, H &\sim H, \qquad \text{for } k = 1, \ldots, \infty, \\ \beta \,|\, \gamma &\sim \mathrm{GEM}(\gamma), \\ w_k \,|\, \alpha, \beta &\sim \mathrm{DP}(\alpha, \beta), \end{aligned} \tag{127}$$

where $w_k = (w_{k1}, w_{k2}, \ldots)$ are the atom masses in G_{ψ_k}. The atoms ψ_k are shared across G_0 and the transition kernels G_{ψ_ℓ}. Since all states visited by the HMM are drawn from the transition kernels, the states possibly visited by the HMM with positive probability (given G_0) will consist only of the initial state ψ_0 and the atoms $\psi_1, \psi_2, \ldots$. Relating to the parametric HMM, we see that the transition probability from state ψ_ℓ to state ψ_k is given by $w_{\ell k}$ and the distribution on the observations given the current state being ψ_k is $F(\psi_k)$.

This relationship to the parametric HMM can be seen even more clearly if we identify the state ψ_k with the integer k, for $k = 0, 1, 2, \ldots$, and if we introduce integer-valued variables z_t to denote the state at time t. In particular, if $\theta_t = \psi_k$ is the state at time t, we let z_t take on value k. The HDP-HMM can now be expressed as:

$$\begin{aligned} z_t \,|\, z_{t-1}, \{w_k : k = 0, 1, 2, \ldots\} &\sim w_{z_{t-1}}, \qquad \text{for } t = 1, \ldots, \tau, \\ x_t \,|\, z_t, \{\psi_k : k = 1, 2, \ldots\} &\sim F(\psi_{z_t}), \end{aligned} \tag{128}$$

with priors on the parameters and transition probabilities given by Eq. (128). This construction shows explicitly that the HDP-HMM can be interpreted as an HMM with a countably infinite state space.

A difficulty with the HDP-HMM as discussed thus far is that it tends to be poor at capturing state persistence; it has a tendency to create redundant states and rapidly switch among them. This may not be problematic for applications in which the states are nuisance variables and it is overall predictive likelihood that matters, but it can be problematic for segmentation or parsing applications in which the states are the object of inference and when state persistence is expected. This problem can be solved by giving special treatment to self-transitions. In particular, let G_θ denote the transition kernel associated with state θ. Fox et al. (2011) proposed the following altered definition of G_θ (compare to Eq. (124)):

$$G_\theta \,|\, \alpha, \kappa, G_0, \theta \sim \mathrm{DP}\left(\alpha + \kappa, \frac{\alpha G_0 + \kappa \delta_\theta}{\alpha + \kappa}\right), \tag{129}$$

where δ_θ is a point mass at θ and where κ is a parameter that determines the extra mass placed on a self-transition. To see in more detail how this affects state persistence, consider the stick-breaking weights w_k associated with one of the countably many states ψ_k that can be visited by the HMM. The stick-breaking representation of G_{ψ_k} is altered as follows (compare to Eq. (128)):

$$w_k \,|\, \alpha, \beta, \kappa \sim \mathrm{DP}\left(\alpha + \kappa, \frac{\alpha\beta + \kappa\delta_{\psi_k}}{\alpha + \kappa}\right). \tag{130}$$

Fox et al. (2011) further place a vague gamma prior on $\alpha+\kappa$ and a beta prior on $\kappa/(\alpha+\kappa)$. The hyperparameters of these distributions allow prior control of state persistence. See also Beal et al. (2002), who develop a related prior within the framework of their hierarchical urn scheme.

Teh and Jordan (2010) review several applications of the HDP-HMM. These include the problem of speaker diarization, where the task is that of segmenting the audio recording into time intervals associated with individual speakers (Fox et al., 2011) and the problem of word segmentation, where the task if that of identifying coherent segments of words and their boundaries in continuous speech (Goldwater et al., 2006b).

9 Completely Random Measures

Although we began our treatment of the Dirichlet process with a discussion of Chinese restaurants and urn models—concrete representations that are closely tied to the clustering problem—subsequent sections have hopefully made clear that there can be advantages to identifying the abstract objects underlying the urn models. As we have seen, random measures can play a role akin to random variables in classical Bayesian analysis, most notably forming hierarchies that allow statistical strength to be shared across components of a model.

Our repertoire of random measures is limited at present, however; the only procedure that we have discussed for construction random measures is the stick-breaking process. In the current section we show how to overcome this limitation. We do so by taking a further step in the direction of abstraction, showing how the Dirichlet process can be obtained from a relatively simple underlying object—the Poisson random measure. Once we gain the perspective provided by the Poisson random measure, it will be possible to construct a wide variety of other random measures.

To motivate the constructions that we will soon present, let us return to the Dirichlet process and reflect upon the key property identified in Sec. (6.1): for any partition, $(A_1, A_2, \ldots, A_K)$, of the space Θ, the vector of marginals, $(G(A_1), G(A_2), \ldots, G(A_K))$, has a finite-dimensional Dirichlet distribution. Given that the Dirichlet distribution can be obtained by normalizing a set of K independent gamma random variables (see Appendix B), this suggests that we investigate the possibility of constructing a "gamma process," which should presumably be a random measure G that assigns a gamma-distributed mass to any subset $A \subseteq \Theta$. That is, $G(A)$ should have a gamma distribution. Moreover, for a partition $(A_1, A_2, \ldots, A_K)$, the random variables $(G(A_1), G(A_2), \ldots, G(A_K))$ should be independent, so that we obtain a Dirichlet-distributed vector once they are normalized by the total mass

$G(\Theta)$. In other words, the normalized random measure $G/G(\Theta)$ should have the Dirichlet marginals of the Dirichlet process.

Given the hint that the Poisson random measure is involved in this construction, we might consider something like the following: Use a Poisson random measure to generate a set of atoms on the space Θ according to a base measure. Associate to each atom a weight that is an independent draw from a gamma random variable (yielding a "marked Poisson process") and define the weighted sum across these atoms to be our putative gamma process. While this construction is not too far off the mark, there are two issues that we must face that are tricky. The first is that we require a infinite number of atoms (given our interest in nonparametric models). This would seem to require that the base measure have infinite mass, which is somewhat incommensurate with its role as a prior for parameters in Bayesian models. Moreover, even if we are able to generate an infinite number of atoms, we face the second issue, which is that the sum of the resulting gamma masses needs to be finite (so that we can normalize the gamma process, and thereby construct the Dirichlet process). The construction that we describe in the remainder of this section surmounts these problems in an elegant way.

9.1 Completely random measures

We begin with a definition:

Definition 1. *A* completely random measure *(CRM) G defined on a space Θ is a random measure whose marginals, $G(A_1), G(A_2), \ldots, G(A_K)$, are independent for any sequence of disjoint subsets, $A_1, A_2, \ldots, A_K$, of Θ.*

This definition is non-constructive, and the natural questions here are: what types of mathematical objects are completely random measures, and can we characterize all completely random measures?

Before exploring these general questions, we give a first archetypal example of a completely random measure (henceforth a CRM)—the Poisson random measure. Let μ be a diffuse measure on the space Θ (not a random measure, just a measure). That is, for each subset $A \subset \Theta$, $\mu(A) \geq 0$ is simply a constant, and $\mu(\{\theta\}) = 0$ for each $\theta \in \Theta$. We say that N is a *Poisson random measure* if it is a CRM and the marginal $N(A)$ has the distribution Poisson$(\mu(A))$. That such a random measure exists is a classical fact of probability theory that we will simply accept.[12] The simplest Poisson random measure takes μ to be the uniform measure on Θ, in which case the intuition is of a process that randomly drops atoms in Θ with no preference for any particular region. Letting μ be non-uniform makes it possible to obtain higher density of atoms in some regions rather than others (as shown in Fig. (9)).

Given that a random variable drawn from the Poisson(λ) distribution has mean λ, we see that $\mathbb{E}[N(A)] = \mu(A)$. We will therefore refer to μ as the *mean measure* associated

[12] Note, by the way, that a "Poisson point process" and a "Poisson random measure" are different (although closely related) mathematical objects. A realization of a Poisson point process on a space Θ is a set of points in Θ, whereas a realization of a Poisson random measure is an atomic measure over Θ where each atom has mass one. Given a realization of a Poisson point process, we can construct a realization of a Poisson random measure by placing an atom with mass one at each point in the point set. Similarly, we can obtain a Poisson point process via the locations of the atoms of a Poisson random measure.

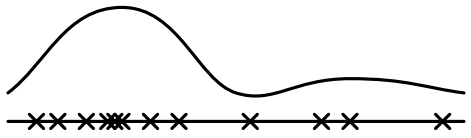

Figure 9: A depiction of a realization of a Poisson point process on the real line. Each cross is a point in the realization of the Poisson point process. The corresponding Poisson random measure places atoms of unit mass at each of the points forming the Poisson point process. The function shown is the density function $f(x)$ corresponding to the mean measure μ, that is, $\mu(A)$ is obtained as an integral, $\mu(A) = \int_A f(x)dx$, for each $A \subset \Theta$.

with the Poisson random measure N. Since N is an atomic measure where each atom has unit mass, $N(A)$ is the number of atoms lying in A, and $\mu(A)$ corresponds to the expected number of such atoms. Further, as N is completely random, the random numbers of atoms lying in disjoint subsets are independent. In fact, a straightforward argument shows that the measures formed as restrictions of N to the disjoint subsets are themselves independent.

We now turn to the construction of general completely random measures. We first note that any deterministic measure is completely random (in a trivial sense). Apart from a deterministic component, it can be shown that completely random measures are discrete (Kingman, 1967). We will be interested in the atoms that form these discrete measures, which are of two types. The first are the *fixed atoms*. These have non-random locations and random masses that are mutually independent and independent of the other randomness in the construction. The remaining, *non-fixed* atoms vary randomly both in location and in mass. We can treat these two types of atoms separately, due to the complete randomness of the measure. A beautiful and somewhat surprising fact is that all of the non-fixed atoms can be obtained from an underlying Poisson random measure (Kingman, 1967). In the remainder of this section we present the general construction of these non-fixed atoms from a Poisson random measure, to be followed by a concrete example in Sec. (9.2).

The goal is to construct a CRM consisting only of non-fixed atoms on a space Θ. We do this by considering an augmented space, in particular the product space $\Xi = \Theta \otimes \mathbb{R}^+$, where $\mathbb{R}^+$ denotes the nonnegative real numbers. On this product space $\Theta \otimes \mathbb{R}^+$ we will place a Poisson random measure N, and construct the CRM on Θ as a function of N.

Suppose that the mean measure of N on $\Theta \otimes \mathbb{R}^+$ can be written in the following product form:[13]

$$\mu = G_0 \otimes \nu, \tag{131}$$

where G_0 is a diffuse probability measure on Θ and ν is a diffuse measure on $\mathbb{R}^+$. What this notation means is that for sets $A \subset \Theta$ and $E \subset \mathbb{R}^+$, we have $\mu(A \otimes E) = G_0(A)\nu(E)$. We refer to the probability measure G_0 on Θ as the base (probability) measure, and to the component ν of the overall mean measure as the *Lévy measure* of the CRM.

We will also use the suggestive notation $\mu(d\phi, dw) = G_0(d\phi)\nu(dw)$ when we are thinking of taking integrals with respect to the measure μ.

[13] A CRM constructed using such a product mean measure is referred to as *homogeneous*. Non-homogeneous CRMs can also be constructed, using a mean measure which does not decompose as a product.

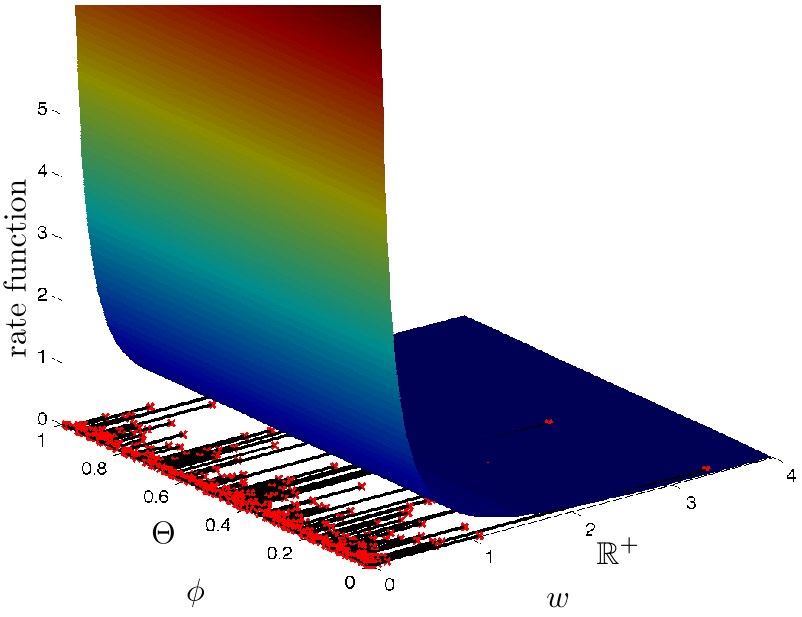

Figure 10: The construction of a completely random measure on Θ from a Poisson random measure on $\Theta \otimes \mathbb{R}^+$.

Given a Poisson random measure N on $\Theta \otimes \mathbb{R}^+$, denote by $\{(\phi_k, w_k)\}$ the set of atoms in a particular realization of N. That is, let us write the realization as follows:

$$N = \sum_k \delta_{\phi_k, w_k}, \tag{132}$$

where δ_{ϕ_k, w_k} is an atom at location (ϕ_k, w_k) in the product space. Given these atoms, we construct a measure G on Θ as a weighted sum of atoms, where ϕ_k specifies the location of the kth atom and w_k is its mass:

$$G = \sum_k w_k \delta_{\phi_k}. \tag{133}$$

Note that we have not specified the range of k as being finite or infinite, but in fact we will be able to achieve an infinite range (as desired) via the choice of ν. A depiction of this construction is shown in Fig. (10).

Clearly G is a random measure, with $G(A)$ corresponding to the total sum of masses for those atoms ϕ_k that fall in A:

$$G(A) = \sum_k w_k \delta_{\phi_k}(A) = \sum_{k:\phi_k \in A} w_k. \tag{134}$$

Moreover, because the underlying Poisson random measure N is completely random, G is completely random as well. To see this, suppose that $A_1, A_2, \ldots, A_K$ are disjoint subsets of Θ. Then the collection of subsets $\{A_j \otimes \mathbb{R}^+\}$ are disjoint as well, so that the restrictions of N to these subsets are mutually independent. Since each $G(A_j)$ is a function only of

the atoms in $\{A_j \otimes \mathbb{R}^+\}$, we see that the masses $G(A_1), G(A_2), \ldots, G(A_K)$ are mutually independent too.

Having formed a CRM G from the Poisson random measure, we are free to add to G a set of fixed atoms (where the locations of the fixed atoms are determined prior to generating G) with mutually independent masses which are independent of N as well. We can also add to this CRM a deterministic measure; the resulting object remains a CRM. In fact this fully characterizes CRMs—all CRMs can be constructed as a sum of three components in this way (Kingman, 1967).

9.2 The gamma process

To illustrate the general construction of CRMs provided in the previous section, we now define the gamma process, which is a CRM consisting only of non-fixed atoms. To do so, we consider the following mean measure defined on $\Theta \otimes \mathbb{R}^+$:

$$\mu(d\phi, dw) = G_0(d\phi)\alpha w^{-1} e^{-\beta w} dw. \tag{135}$$

where $\alpha > 0$ and $\beta > 0$ parameterize the Lévy measure ν. (Recall that this notation is simply a suggestive way to indicate how to use the product measure to compute integrals; in particular, for $A \subseteq \Theta$ and $E \subseteq \mathbb{R}^+$, we have $\mu(A, E) = G_0(A) \int_E w^{-1} e^{-\beta w} dw$.) We denote the resulting CRM—the gamma process—as follows:

$$G \sim \mathrm{GaP}(\alpha, \beta, G_0), \tag{136}$$

where the parameters α, β and G_0 are known as the concentration parameter, the inverse scale parameter and the base measure respectively.

Note that the Lévy measure for the gamma process is defined in terms of a density, $\alpha w^{-1} e^{-\beta w}$. Recalling that the density of a gamma random variable is proportional to $w^{a-1} e^{-bw}$, we see that the gamma process is based on a Lévy measure whose density has the form of the gamma density but where the shape parameter a is set equal to zero and there is an additional multiplicative constant. For this value of a the integral over $\mathbb{R}^+$ is equal to infinity. This implies that the total mean measure of the underlying Poisson random measure is infinite:

$$\mu(\Theta \otimes \mathbb{R}^+) = G_0(\Theta) \int_0^\infty \alpha w^{-1} e^{-\beta w} dw = \infty. \tag{137}$$

This is as desired—it yields a countably infinite number of atoms under the Poisson random measure and under G as well. We note that the base measure G_0 is normalized, and the infinity is placed in the Lévy measure ν instead.

What remains for us to show is that the CRM that we have defined has gamma marginals; i.e., that $G(A)$ is distributed as a gamma random variable for a fixed subset $A \subseteq \Theta$. Note in particular that this result will show that introducing an infinite mass into the Lévy measure has not caused trouble by yielding an infinite value for $G(A)$. The derivation of the marginal probability under a CRM requires a general result from probability theory known as the Lévy-Khinchin formula. Given the importance of this formula, we will provide a derivation of the formula (in the generality that we require) in the next section. A reader already knowing this formula may skip ahead to Sec. (9.4).

9.3 The Lévy-Khinchin formula

The Laplace transform of a positive random variable X is the expectation $\mathbb{E}[e^{-tX}]$ treated as a function of $t \geq 0$. The Laplace transform is related to the moment generating function of X, and all moments of X can be computed from the derivatives (with respect to t) of the Laplace transform. Consequently, we can establish that two positive random variables have the same distribution by showing that they have the same Laplace transform. In our circumstance we can show that $G(A)$ has a gamma distribution by computing its Laplace transform, and recognizing that it is the same as that for gamma random variables.

The way we can compute the Laplace transform is via a device called the Lévy-Khinchin formula, which allows us to compute the Laplace transform of random variables obtained from integrals of Poisson random measures. Deriving this formula will require the computation of the Laplace transform of the Poisson distribution. Moreover, applying the Lévy-Khinchin formula to the special case of the gamma process will require the Laplace transform of the gamma distribution. See Appendix E for the derivation of these Laplace transforms.

To work up to the Lévy-Khinchin formula, we first focus on the somewhat easier calculation of the expected value of $G(A)$. If $G(A)$ is to have a gamma distribution, Gamma(a, b), for given shape parameter a and inverse scale parameter b, then $\mathbb{E}[G(A)]$ should be equal to a/b. We begin by rewriting $G(A)$ in a slightly more convenient form. Recall that G is defined in terms of an underlying Poisson random measure N. Recalling Eq. (132), we can write the sum in the definition of G as an integral over the discrete measure N:

$$G(A) = \sum_k w_k \delta_{\phi_k}(A) = \int w \mathbb{1}_A(\phi) N(d\phi, dw), \tag{138}$$

where $\mathbb{1}_A(\phi)$ denotes an indicator function that is equal to one if $\phi \in A$ and zero otherwise.

Writing $f(\phi, w) = w\mathbb{1}_A(\phi)$, we see that $G(A)$ is an integral of the form $\int f(\phi, w) N(d\phi, dw)$. Simplifying the notation further, let $\xi = (\phi, w)$ and we can write the integral as follows:

$$G(A) = \int f(\xi) N(d\xi), \tag{139}$$

which is an integral under the Poisson random measure N of a function on the space $\Xi = \Theta \otimes \mathbb{R}^+$. To compute the expectation of such an integral, we focus first on computing the expectation of integrals of step functions and work our way towards more complex functions. Consider first a function of the form $c\mathbb{1}_C(\xi)$, for a fixed set $C \subset \Xi$ and a constant $c > 0$. For this choice of f, we have:

$$\begin{aligned} \mathbb{E}\left[\int f(\xi) N(d\xi)\right] &= \mathbb{E}\left[\int c\mathbb{1}_C(\xi) N(d\xi)\right] \\ &= c\mathbb{E}\left[N(C)\right] \\ &= c\mu(C), \end{aligned}$$

where in the last equality we have used the fact that N is a Poisson random measure with mean measure μ.

We now carry out the same computation for a general step function, $f(\xi) = \sum_{j \in J} c_j \mathbb{1}_{C_j}(\xi)$, where J is a finite index set, where the C_j's are non-overlapping and where each c_j is a positive constant. We have:

$$\begin{aligned}
\mathbb{E}\left[\int f(\xi) N(d\xi)\right] &= \mathbb{E}\left[\int \sum_{j \in J} c_j \mathbb{1}_{C_j}(\xi) N(d\xi)\right] \\
&= \sum_{j \in J} \mathbb{E}\left[\int c_j \mathbb{1}_{C_j}(\xi) N(d\xi)\right] \\
&= \sum_{j \in J} c_j \mu(C_j) \\
&= \int f(\xi) \mu(d\xi),
\end{aligned}$$

where the last equality is by the definition of the integral under μ of the step function f. Finally, we appeal to a classical result from real analysis—the monotone convergence theorem (Kallenberg, 2002)—to extend the result to general positive functions using the fact that it is possible to approximate general positive functions via an increasing sequence of step functions. Thus, for general positive functions we have:

$$\mathbb{E}\left[\int f(\xi) N(d\xi)\right] = \int f(\xi) \mu(d\xi). \tag{140}$$

In particular, this result holds for our target function $f(\phi, w) = w \mathbb{1}_A(\phi)$. It also extends to a general function $f(\xi)$ by writing it as the difference between two positive functions f_+ and f_-, $f(\xi) = f_+(\xi) - f_-(\xi)$. This result allows us to compute the first moments of a Poisson random measure, and restates the fact that the mean measure is indeed the mean of the Poisson random measure. It is referred to as Campbell's theorem, or the first moment formula for Poisson random measures.

Applying the general result in Eq. (140) to the special case of the gamma process, we compute:

$$\begin{aligned}
\mathbb{E}\left[G(A)\right] &= \mathbb{E}\left[\int w \mathbb{1}_A(\phi) N(d\phi, dw)\right] \\
&= \int w \mathbb{1}_A(\phi) \mu(d\phi, dw) \\
&= \int w \mathbb{1}_A(\phi) G_0(d\phi) \alpha w^{-1} e^{-\beta w} dw \\
&= \alpha G_0(A) \int_0^\infty e^{-\beta w} dw \\
&= \frac{\alpha G_0(A)}{\beta}.
\end{aligned} \tag{141}$$

This result is consistent with $G(A)$ being a gamma random variable, with shape parameter $\alpha G_0(A)$ and inverse scale parameter β.

Moreover, if we partition a set A into $A = A_1 \cup A_2$ for disjoint subsets A_1 and A_2, then we obtain

$$\begin{aligned} \mathbb{E}\left[G(A)\right] &= \frac{\alpha G_0(A)}{\beta} = \frac{\alpha G_0(A_1) + \alpha G_0(A_2)}{\beta} \quad (142) \\ &= \frac{\alpha G_0(A_1)}{\beta} + \frac{\alpha G_0(A_2)}{\beta}, \quad (143) \end{aligned}$$

which is consistent with the fact that the sum of independent gamma random variables with the same inverse scale parameter is a gamma random variable, with a shape parameter that is the sum of the individual shape parameters.

Although we could continue to check moments of $G(A)$ against the moments of the gamma distribution, the right thing to do at this point is to take aim at all of the moments via the Laplace transform. We thus want to compute the Laplace transform $\mathbb{E}\left[e^{-tG(A)}\right]$, which is the same as the Laplace transform of $\int f(\xi)N(d\xi)$, for $f(\xi) = w\mathbb{1}_A(\phi)$. Our strategy is once again to perform the calculation for indicator functions and general step functions, and then to appeal to the monotone convergence theorem. We thus start with $f(\xi) = c\mathbb{1}_C(\xi)$ and compute:

$$\begin{aligned} \mathbb{E}\left[e^{-t\int f(\xi)N(d\xi)}\right] &= \mathbb{E}\left[e^{-tcN(C)}\right] \\ &= \exp\left(-\mu(C)\left(1 - e^{-ct}\right)\right), \end{aligned}$$

where we have used the fact that $N(C)$ is Poisson and plugged in the form of the Laplace transform for Poisson random variables (see Appendix E).

Now for step functions $f(\xi) = \sum_{j\in J} c_j \mathbb{1}_{C_j}(\xi)$, we compute:

$$\begin{aligned} \mathbb{E}\left[e^{-t\int f(\xi)N(d\xi)}\right] &= \mathbb{E}\left[e^{-t\sum_{j\in J} c_j N(C_j)}\right] \\ &= \prod_{j\in J} \mathbb{E}\left[e^{-tc_j N(C_j)}\right] \\ &= \prod_{j\in J} \exp\left(-\mu(C_j)(1 - e^{-c_j t})\right) \\ &= \exp\left(-\sum_{j\in J} \mu(C_j)(1 - e^{-c_j t})\right) \\ &= \exp\left(-\int (1 - e^{-tf(\xi)})\mu(d\xi)\right), \end{aligned}$$

where the second equality used the complete randomness of N, and the final equality used the definition of the integral of the function $(1 - e^{-tf(\xi)})$ with respect to μ.

The monotone convergence theorem allows us to conclude that for general positive functions $f(\xi)$ the Laplace transform can be computed as follows:

$$\mathbb{E}\left[e^{-t\int f(\xi)N(d\xi)}\right] = \exp\left(-\int (1 - e^{-tf(\xi)})\mu(d\xi)\right). \quad (144)$$

This result is known as the Lévy-Khinchin formula. It highlights the key role played by the mean measure in computing probabilities under Poisson random measures.

9.4 The gamma process and the Dirichlet process

We return to the problem of computing the distribution of $G(A)$ for the special case of the gamma process. Letting $f(\xi) = w\mathbb{1}_A(\phi)$, we appeal to the Lévy-Khinchin formula in Eq. (144) and compute:

$$\begin{aligned}
\mathbb{E}\left[e^{-tG(A)}\right] &= \exp\left(-\int(1-e^{-tf(\xi)})\mu(d\xi)\right) \\
&= \exp\left(-\int_\Theta\int_0^\infty\left(1-e^{-tw\mathbb{1}_A(\phi)}\right)\alpha G_0(d\phi)\nu(dw)\right) \\
&= \exp\left(-\int_\Theta\int_0^\infty \mathbb{1}_A(\phi)(1-e^{-tw})\alpha G_0(d\phi)\nu(dw)\right) \\
&= \exp\left(-\alpha G_0(A)\int_0^\infty(1-e^{-tw})\nu(dw)\right), \qquad (145)
\end{aligned}$$

where the third equality can be verified by noting that the indicator is either one or zero and that equality holds in both cases.

The problem thus reduces to computing an integral under the Lévy measure:

$$\begin{aligned}
\int_0^\infty(1-e^{-tw})\nu(dw) &= \int_0^\infty(1-e^{-tw})\alpha w^{-1}e^{-\beta w}dw \\
&= \alpha\int_0^\infty\frac{e^{-\beta w}-e^{-(\beta+t)w}}{w}dw \\
&= \alpha\int_0^\infty\int_\beta^{\beta+t}e^{-sw}dsdw \\
&= \alpha\int_\beta^{\beta+t}\int_0^\infty e^{-sw}dwds \\
&= \alpha\int_\beta^{\beta+t}\frac{1}{s}ds \\
&= \alpha\log\left(\frac{\beta+t}{\beta}\right). \qquad (146)
\end{aligned}$$

Plugging this result back in Eq. (145) we obtain:

$$\begin{aligned}
\mathbb{E}\left[e^{-tG(A)}\right] &= \exp\left(-\alpha G_0(A)\log\left(\frac{\beta+t}{\beta}\right)\right) \\
&= \left(\frac{\beta}{\beta+t}\right)^{\alpha G_0(A)}. \qquad (147)
\end{aligned}$$

Referring to Appendix E for the Laplace transform of the gamma distribution, we see that the marginal $G(A)$ has the gamma distribution with shape parameter $\alpha G_0(A)$ and inverse scale parameter β.

Finally, we return to our original motivation for this development, the Dirichlet process. Let $(A_1, A_2, \ldots, A_K)$ denote a partition of the space Θ, and let $G \sim \mathrm{GaP}(\alpha, \beta, G_0)$ denote a gamma process on Θ. We now know that the marginals $G(A_k)$ are gamma random variables,

and, given that the gamma process is a CRM, these random variables are independent. We can thus form the normalized marginals:

$$\tilde{G}(A_k) = \frac{G(A_k)}{\sum_{j=1}^{K} G(A_j)}, \tag{148}$$

which have the Dirichlet distribution with parameters $(\alpha G_0(A_1), \ldots, \alpha G_0(A_K))$ (by definition; see Appendix B). Given that this result holds for an arbitrary partition, we see that the Dirichlet process, $\mathrm{DP}(\alpha, G_0)$, can be represented as a normalization of the gamma process, $\mathrm{GaP}(\alpha, \beta, G_0)$. It is also worth noting that the inverse scale parameter β of the gamma process has no impact on the Dirichlet process. This is to be expected since the normalization of the gamma process removes the effect of the overall scale of the gamma process.

Although we have engaged in a rather considerable effort to obtain this further representation of the Dirichlet process, the payoff is quite significant. Indeed, the framework that we have introduced yields a variety of interesting new random measures beyond the Dirichlet process, including random measures that are CRMs and random measures obtained from CRMs via normalization. The rest of the article will take CRMs as the point of departure.

10 Discussion

Review the various representations.

Note that each of the representations can be used as a point of departure for variations. The CRM framework will be our main tool going forward.

There are a *lot* of issues that we haven't addressed. Our treatment should be viewed as a jumping-off point rather than a review.

* how to set alpha * empirical Bayes for alpha, and for G_0 * DDP * conditional stick-breaking (logistic, kernel, probit) * variational inference * posterior checking (or put this later) * theory (or put this later)

product partition models: BARRY, D. & HARTIGAN, J. A. (1992). Product partition models for change point problems. Ann. Statist. 20, 26079. QUINTANA, F. A. & IGLESIAS, P. L. (2003). Bayesian clustering and product partition models. J. R. Statist. Soc. B 65, 55774.

Part II: The Beta Process

11 The Beta Process and the Indian Buffet Process

The DP mixture model embodies the assumption that the data can be partitioned or clustered into discrete classes. This assumption is made particularly clear in the Chinese restaurant representation, where the table at which a data point sits indexes the class (the mixture component) to which it is assigned. If we represent the restaurant as a binary matrix in which the rows are the data points and the columns are the tables, we obtain a matrix with a single one in each row and all other elements equal to zero.

A different assumption that is natural in many settings is that objects can be described in terms of a collection of binary *features* or *attributes*. For example, we might describe a set of animals with features such as `diurnal/nocturnal`, `avian/non-avian`, `cold-blooded/warm-blooded`, etc. Forming a binary matrix in which the rows are the objects and the columns are the features, we obtain a matrix in which there are multiple ones in each row. We will refer to such a representation as a *featural representation*.

A featural representation can of course be converted into a set of clusters if desired: if there are K binary features, we can place each object into one of 2^K clusters. In so doing, however, we lose the ability to distinguish between classes that have many features in common and classes that have no features in common. Also, if K is large, it may be infeasible to consider models with 2^K parameters. Using the featural representation, we might hope to construct models that use on the order of K parameters to describe 2^K classes.

In this section we discuss a Bayesian nonparametric approach to featural representations. In essence, we replace the Dirichlet/multinomial probabilities that underlie the Dirichlet process with a collection of beta/Bernoulli draws. This is achieved via the *beta process*, a stochastic process whose realizations provide a countably infinite collection of coin-tossing probabilities. We also discuss some other representations of the beta process that parallel those for the DP. In particular we describe a stick-breaking construction as well as an analog of the Chinese restaurant process known as the *Indian buffet process*.

11.1 The beta process and the Bernoulli process

The beta process is an instance of a general class of stochastic processes known as completely random measures, which was described in Sec. (9). The key property of completely random measures is that the random variables obtained by evaluating a random measure on disjoint subsets of the probability space are mutually independent. Moreover, draws from a completely random measure are discrete (up to a fixed deterministic component). Thus we can represent such a draw as a weighted collection of atoms on some probability space, as we do for the DP.

Applications of the beta process in Bayesian nonparametric statistics have mainly focused on its use as a model for random hazard functions (Hjort, 1990). In this case, the probability space is the real line and it is the cumulative integral of the sample paths that is of interest (yielding a random, nondecreasing step function). In the application of the beta process to featural representations, on the other hand, it is the realization itself that is of interest and the underlying space is no longer restricted to be the real line.

Following Thibaux and Jordan (2007), let us thus consider a general probability space Θ

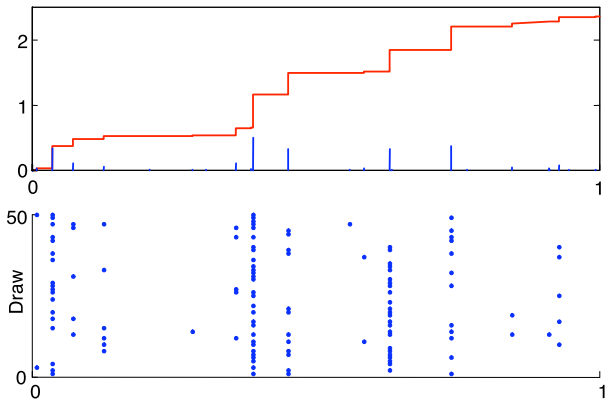

Figure 11: (a) A draw $B \sim \mathrm{BP}(1, U[0,1])$. The set of blue spikes is the sample path and the red curve is the corresponding cumulative integral $\int_{-\infty}^{x} B(d\phi)$. (b) 100 samples from $\mathrm{BeP}(B)$, one sample per row. Note that a single sample is a *set* of unit-weight atoms.

endowed with a finite *base measure* B_0 (note that B_0 is not a probability measure; it does not necessarily integrate to one). Intuitively we wish to partition Θ into small regions, placing atoms into these regions according to B_0 and assigning a weight to each atom, where the weight is a draw from a beta distribution. A similar partitioning occurs in the definition of the DP, but in that case the aggregation property of Dirichlet random variables immediately yields a consistent set of marginals. Because the sum of two beta random variables is not a beta random variable, the construction is somewhat less straightforward in the beta process case.

The general machinery of completely random processes deals with this issue in an elegant way. Consider first the case in which B_0 is diffuse and define the Lévy measure on the product space $\Theta \otimes [0,1]$ in the following way:

$$\mu(d\phi, dw) = cw^{-1}(1-w)^{c-1} dw B_0(d\phi), \tag{149}$$

where $c > 0$ is a *concentration parameter*. Now let N be a Poisson random process with μ as its mean measure. This yields a set of atoms at locations $\{(\phi_1, w_1), (\phi_2, w_2), \ldots\}$. Define a realization of the beta process as:

$$B = \sum_{k=1}^{\infty} w_k \delta_{\phi_k}, \tag{150}$$

where δ_{ϕ_k} is an atom at ϕ_k with w_k its mass in B. We denote this stochastic process as $B \sim \mathrm{BP}(c, B_0)$. Figure 11(a) provides an example of a draw from $\mathrm{BP}(1, U[0,1])$, where $U[0,1]$ is the uniform distribution on $[0,1]$.

We obtain a countably infinite set of atoms from this construction because the Lévy measure in Eq. (149) has infinite mass. Further, the sum of the atom masses is finite with probability one, since $\int w\mu(d\phi, dw) < \infty$.

If B_0 contains atoms, then these are treated separately. In particular, denote the measure of the kth atom as q_k (assumed to lie in $(0,1)$). The realization B necessarily contains that

atom, with the corresponding weight w_k defined as an independent draw from $\text{Beta}(cq_k, c(1-q_k))$. The overall realization B is a sum of the weighted atoms coming from the continuous component and the discrete component of B_0.

Let us now define a *Bernoulli process* $\text{BeP}(B)$ with an atomic base measure B as a stochastic process whose realizations are collections of atoms of unit mass on Θ. Atoms can only appear at the locations of atoms of B. Whether or not an atom appears is determined by independent tosses of a coin, where the probability of success is the corresponding weight of the atom in B. After n draws from $\text{BeP}(B)$ we can fill a binary matrix that has n rows and an infinite number of columns (corresponding to the atoms of B arranged in some order). Most of the entries of the matrix are zero while a small (finite) number of the entries are equal to one. Figure 11(b) provides an example.

The beta process and the Bernoulli process are *conjugate*. Consider the specification:

$$\begin{aligned} B \,|\, c, B_0 &\sim \text{BP}(c, B_0) \\ Z_i \,|\, B &\sim \text{BeP}(B), \qquad\qquad \text{for } i = 1, \ldots, n, \end{aligned} \tag{151}$$

where $Z_1, \ldots, Z_n$ are conditionally independent given B. The resulting posterior distribution is itself a beta process, with updated parameters:

$$B \,|\, Z_1, \ldots, Z_n, c, B_0 \sim \text{BP}\left(c + n, \frac{c}{c+n} B_0 + \frac{1}{c+n} \sum_{i=1}^{n} Z_i\right). \tag{152}$$

This formula can be viewed as an analog of standard finite-dimensional beta/Bernoulli updating. Indeed, given a prior $\text{Beta}(a, b)$, the standard update takes the form $a \to a + \sum_i z_i$ and $b \to b + n - \sum_i z_i$. In Eq. (152), c plays the role of $a + b$ and cB_0 is analogous to a.

11.2 The Indian buffet process

Recall that the Chinese restaurant process can be obtained by integrating out the Dirichlet process and considering the resulting distribution over partitions. In the other direction, the Dirichlet process is the random measure that is guaranteed (by exchangeability and De Finetti's theorem) to underlie the Chinese restaurant process. In this section we discuss the analog of these relationships for the beta process.

We begin by defining a stochastic process known as the *Indian buffet process* (IBP). The IBP was originally defined directly as a distribution on (equivalence classes of) binary matrices by Griffiths and Ghahramani (2006) and Ghahramani et al. (2007). The IBP is an infinitely exchangeable distribution on these equivalence classes, thus it is of interest to discover the random measure that must underlie the IBP according to De Finetti's Theorem. Thibaux and Jordan (2007) showed that the underlying measure is the beta process; that is, the IBP is obtained by integrating over the beta process B in the hierarchy in Eq. (151).

The IBP is defined as follows. Consider an Indian buffet with a countably-infinite number of dishes and customers that arrive in sequence in the buffet line. Let Z^* denote a binary-valued matrix in which the rows are customers and the columns are the dishes, and where $Z^*_{nk} = 1$ if customer n samples dish k. The first customer samples $\text{Poisson}(\alpha)$ dishes, where $\alpha = B_0(\Theta)$ is the total mass of B_0. A subsequent customer n samples dish k with probability $\frac{m_k}{c+n-1}$, where m_k is the number of customers who have previously sampled dish

k; that is, $Z^*_{nk} \sim \text{Bernoulli}(\frac{m_k}{c+n-1})$. Having sampled from the dishes previously sampled by other customers, customer n then goes on to sample an additional number of new dishes determined by a draw from a $\text{Poisson}(\frac{c}{c+n-1}\alpha)$ distribution.

To derive the IBP from the beta process, consider first the distribution Eq. (152) for $n = 0$; in this case the base measure is simply B_0. Drawing from $B \sim \text{BP}(B_0)$ and then drawing $Z_1 \sim \text{BeP}(B)$ yields atoms whose locations are distributed according to a Poisson process with rate B_0; the number of such atoms is $\text{Poisson}(\alpha)$. Now consider the posterior distribution after $Z_1, \ldots, Z_{n-1}$ have been observed. The updated base measure is $\frac{c}{c+n-1}B_0 + \frac{1}{c+n-1}\sum_{i=1}^{n-1} Z_i$. Treat the discrete component and the continuous component separately. The discrete component, $\frac{1}{c+n-1}\sum_{i=1}^{n-1} Z_i$, can be reorganized as a sum over the unique values of the atoms; let m_k denote the number of times the kth atom appears in one of the previous Z_i. We thus obtain draws $w_k \sim \text{Beta}((c+n-1)q_k, (c+n-1)(1-q_k))$, where $q_k = \frac{m_k}{c+n-1}$. The expected value of w_k is $\frac{m_k}{c+n-1}$ and thus (under Bernoulli sampling) this atom appears in Z_n with probability $\frac{m_k}{c+n-1}$. From the continuous component, $\frac{c}{c+n-1}B_0$, we generate $\text{Poisson}(\frac{c}{c+n-1}\alpha)$ new atoms. Equating "atoms" with "dishes," and rows of Z^* with draws Z_n, we have obtained exactly the probabilistic specification of the IBP.

11.3 Stick-breaking constructions

The stick-breaking representation of the DP is an elegant constructive characterization of the DP as a discrete random measure (Sec. (5)). This construction can be viewed in terms of a metaphor of breaking off lengths of a stick, and it can also be interpreted in terms of a size-biased ordering of the atoms. In this section, we consider analogous representations for the beta process. Draws $B \sim \text{BP}(c, B_0)$ from the beta process are discrete with probability one, which gives hope that such representations exist. Indeed, we will show that there are two stick-breaking constructions of B, one based on a size-biased ordering of the atoms (Thibaux and Jordan, 2007), and one based on a stick-breaking representation known as the inverse Lévy measure (Wolpert and Ickstadt, 1998).

The size-biased ordering of Thibaux and Jordan (2007) follows straightforwardly from the discussion in Section 11.2. Recall that the Indian buffet process is defined via a sequence of draws from Bernoulli processes. For each draw, a Poisson number of new atoms are generated, and the corresponding weights in the base measure B have a beta distribution. This yields the following truncated representation:

$$B_N = \sum_{n=1}^{N}\sum_{k=1}^{K_n} w_{nk}\delta_{\phi_{nk}}, \tag{153}$$

where

$$\begin{aligned} K_n \,|\, c, B_0 &\sim \text{Poisson}(\tfrac{c}{c+n-1}\alpha) && \text{for } n = 1, \ldots, \infty \\ w_{nk} \,|\, c &\sim \text{Beta}(1, c+n-1) && \text{for } k = 1, \ldots, K_n \\ \phi_{nk} \,|\, B_0 &\sim B_0/\alpha. \end{aligned} \tag{154}$$

It can be shown that this size-biased construction B_N converges to B with probability one. The expected total weight contributed at step N is $\frac{c\alpha}{(c+N)(c+N-1)}$, while the expected total

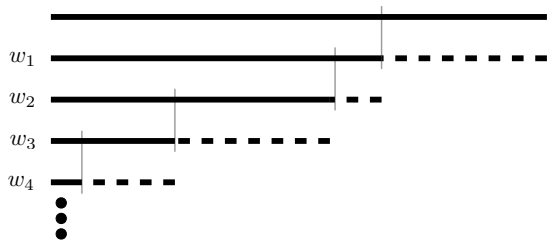

Figure 12: Stick-breaking construction for the one-parameter beta process. A stick of length 1 is recursively broken into pieces of sticks (dashed lines) whose lengths sum to one. These lengths are the atom masses in the DP, while the lengths of the left-over piece of stick after each break form the atom masses for the BP.

weight remaining, in $B - B_N$, is $\frac{c\alpha}{c+N}$. The expected total weight remaining decreases to zero as $N \to \infty$, but at a relatively slow rate. Note also that we are not guaranteed that atoms contributed at later stages of the construction will have small weight—the sizes of the weights need not be in decreasing order.

The stick-breaking construction of Teh et al. (2007) can be derived from the inverse Lévy measure algorithm of Wolpert and Ickstadt (1998). This algorithm starts from the Lévy measure of the beta process, and generates a sequence of weights of decreasing size using a nonlinear transformation of a one-dimensional Poisson process to one with uniform rate. In general this approach does not lead to closed forms for the weights; inverses of the incomplete Beta function need to be computed numerically. However for the one-parameter beta process (where $c = 1$) we do obtain a simple closed form:

$$B_K = \sum_{k=1}^{K} w_k \delta_{\phi_k}, \tag{155}$$

where

$$\begin{aligned} v_k \,|\, \alpha &\sim \text{Beta}(1, \alpha) \qquad\qquad \text{for } k = 1, \ldots, \infty \\ w_k &= \prod_{l=1}^{k} (1 - v_l) \\ \phi_k \,|\, B_0 &\sim B_0/\alpha. \end{aligned} \tag{156}$$

Again $B_K \to B$ as $K \to \infty$, but in this case the expected weights decrease exponentially to zero. Further, the weights are generated in strictly decreasing order, so we are guaranteed to generate the larger weights first.

The stick-breaking construction for the one-parameter beta process has an intriguing connection to the stick-breaking construction for the DP. In particular, both constructions use the same beta-distributed breakpoints v_k; the difference is that for the DP we use the lengths of the sticks just broken off as the weights while for the beta process we use the remaining lengths of the sticks. This is depicted graphically in Fig. (12).

11.4 Hierarchical beta processes

Recall the construction of the hierarchical Dirichlet process: a set of Dirichlet processes are coupled via a random base measure. A similar construction can be carried out in the case of the beta process: let the common base measure for a set of beta processes be drawn from an underlying beta process (Thibaux and Jordan, 2007). Under this hierarchical Bayesian nonparametric model, the featural representations that are chosen for one group will be related to the featural representations that are used for other groups.

We accordingly define a *hierarchical beta process* (HBP) as follows:

$$\begin{aligned} B_0 \,|\, \kappa, A &\sim \mathrm{BP}(\kappa, A) && \\ B_j \,|\, c, B_0 &\sim \mathrm{BP}(c, B_0) && \text{for } j \in \mathcal{J} \\ Z_{ji} \,|\, B_j &\sim \mathrm{BeP}(B_j) && \text{for } i = 1, \ldots, n_j, \end{aligned} \tag{157}$$

where $\mathcal{J}$ is the set of groups and there are n_j individuals in group j. The hyperparameter c controls the degree of coupling among the groups: larger values of c yield realizations B_j that are closer to B_0 and thus a greater degree of overlap among the atoms chosen in the different groups.

As an example of the application of the HBP, Thibaux and Jordan (2007) considered the problem of document classification, where there are $|\mathcal{J}|$ groups of documents and where the goal is to classify a new document into one of these groups. In this case, Z_{ji} is a binary vector that represents the presence or absence in the ith document of each of the words in the vocabulary Θ. The HBP yields a form of regularization in which the group-specific word probabilities are shrunk towards each other. This can be compared to standard Laplace smoothing, in which word probabilities are shrunk towards a fixed reference point. Such a reference point can be difficult to calibrate when there are rare words in a corpus, and Thibaux and Jordan (2007) showed empirically that the HBP yielded better predictive performance than Laplace smoothing.

11.5 Applications of the beta process

In the following sections we describe a number of applications of the beta process to hierarchical Bayesian featural models. Note that this is a rather different class of applications than the traditional class of applications of the beta process to random hazard functions.

11.5.1 Sparse latent variable models

Latent variable models play an essential role in many forms of statistical analysis. Many latent variable models take the form of a regression on a latent vector; examples include principal component analysis, factor analysis and independent components analysis. Paralleling the interest in the regression literature in sparse regression models, one can also consider sparse latent variable models, where each observable is a function of a relatively small number of latent variables. The beta process provides a natural way of constructing such models. Indeed, under the beta process we can work with models that define a countably-infinite number of latent variables, with a small, finite number of variables being *active* (i.e., non-zero) in any realization.

Consider a set of n observed data vectors, $x_1, \ldots, x_n$. We use a beta process to model a set of latent features, $Z_1, \ldots, Z_n$, where we capture interactions among the components of these vectors as follows:

$$\begin{aligned} B \,|\, c, B_0 &\sim \mathrm{BP}(c, B_0) \\ Z_i \,|\, B &\sim \mathrm{BeP}(B) \qquad\qquad \text{for } i = 1, \ldots, n. \end{aligned} \tag{158}$$

As we have seen, realizations of beta and Bernoulli processes can be expressed as weighted sums of atoms:

$$\begin{aligned} B &= \sum_{k=1}^{\infty} w_k \delta_{\phi_k} \\ Z_i &= \sum_{k=1}^{\infty} Z_{ik}^* \delta_{\phi_k}. \end{aligned} \tag{159}$$

We view ϕ_k as parametrizing feature k, while Z_i denotes the features that are active for item i. In particular, $Z_{ik}^* = 1$ if feature k is active for item i. The data point x_i is modeled as follows:

$$\begin{aligned} y_{ik} \,|\, H &\sim H \qquad\qquad \text{for } k = 1, \ldots, \infty \\ x_i \,|\, Z_i, \boldsymbol{\theta}, \mathbf{y}_i &\sim F_{\{\phi_k, y_{ik}\}_{k:Z_{ik}^*=1}}, \end{aligned} \tag{160}$$

where y_{ik} is the value of feature k if it is active for item i, and the distribution $F_{\{\phi_k, y_{ik}\}_{k:Z_{ik}^*=1}}$ depends only on the active features, their values, and their parameters.

Note that this approach defines a latent variable model with an infinite number of sparse latent variables, but for each data item only a finite number of latent variables are active. The approach would often be used in a predictive setting in which the latent variables are integrated out, but if the sparseness pattern is of interest per se, it is also possible to compute a posterior distribution over the latent variables.

There are several specific examples of this sparse latent variable model in the literature. One example is an independent components analysis model with an infinite number of sparse latent components (Knowles and Ghahramani, 2007; Teh et al., 2007), where the latent variables are real-valued and x_i is a noisy observation of the linear combination $\sum_k Z_{ik}^* y_{ik} \phi_k$. Another example is the "noisy-or" model of Wood et al. (2006), where the latent variables are binary and are interpreted as presence or absence of diseases, while the observations x_i are binary vectors indicating presence or absence of symptoms.

11.5.2 Relational models

The beta process has also been applied to the modeling of relational data (also known as dyadic data). In the relational setting, data are relations among pairs of objects (Getoor and Taskar, 2007); examples include similarity judgments between two objects, protein-protein interactions, user choices among a set of options, and ratings of products by customers.

We first consider the case in which there is a single set of objects and relations are defined among pairs of objects in that set. Formally, define an observation as a relation x_{ij}

between objects i and j in a collection of n objects. Each object is modeled using a set of latent features as in Eq. (158) and Eq. (159). The observed relation x_{ij} between objects i and j then has a conditional distribution that is dependent only on the features active in objects i and j. For example, Navarro and Griffiths (2007) modeled subjective similarity judgments between objects i and j as normally distributed with mean $\sum_{k=1}^{\infty} \phi_k Z^*_{ik} Z^*_{jk}$; note that this is a weighted sum of features active in both objects. Chu et al. (2006) modeled high-throughput protein-protein interaction screens where the observed binding affinity of proteins i and j is related to the number of overlapping features $\sum_{k=1}^{\infty} Z^*_{ik} Z^*_{jk}$, with each feature interpreted as a potential protein complex consisting of proteins containing the feature. Görür et al. (2006) proposed a nonparametric *elimination by aspects* choice model where the probability of a user choosing object i over object j is modeled as proportional to a weighted sum, $\sum_{k=1}^{\infty} \phi_k Z^*_{ik}(1 - Z^*_{jk})$, across features active for object i that are not active for object j. Note that in these examples, the parameters of the model, ϕ_k, are the atoms of the beta process.

Relational data involving separate collections of objects can be modeled with the beta process as well. Meeds et al. (2007) modeled movie ratings, where the collections of objects are movies and users, and the relational data consists of ratings of movies by users. The task is to predict the ratings of movies not yet rated by users, using these predictions to recommend new movies to users. These tasks are called *recommender systems* or *collaborative filtering*. Meeds et al. (2007) proposed a featural model where movies and users are modeled using separate IBPs. Let Z^* be the binary matrix of movie features and Y^* the matrix of user features. The rating of movie i by user j is modeled as normally distributed with mean $\sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \phi_{kl} Z^*_{ik} Y^*_{jl}$. Note that this dyadic model cannot be represented using two independent beta processes, since there is a parameter ϕ_{kl} for each combination of features in the two IBPs. The question of what random measure underlies this model is an interesting one.

Part III: Normalized Completely Random Measures

12 Normalized Completely Random Measures

In Sec. (9) we introduced the notion of a completely random measure, and showed the last of many representations of the Dirichlet process, as a normalized gamma process. In this section, we pick up on this development by extending the construction of random probability measures to normalizing completely random measures from a wider class.

The construction is in fact straightforward. Suppose G is a completely random measure defined as in Eq. (133), with the underlying Poisson random measure N having a mean measure $\mu = G_0 \otimes \nu$, where G_0 is the base probability measure, and ν is the Lévy intensity. Both G_0 and N are discrete measures, say with atoms

$$\mathsf{N} = \sum_{k=1}^{\infty} \delta_{\phi_k, W_k} \qquad\qquad G = \sum_{k=1}^{\infty} W_k \delta_{\phi_k}. \tag{161}$$

Define

$$T = G(\Theta) = \sum_{k=1}^{\infty} W_k \tag{162}$$

to be the total mass of the atoms in G. Supposing that T is positive and finite, we can define a random probability measure by normalizing the completely random measure as:

$$\tilde{G} = \frac{G}{T}. \tag{163}$$

The Dirichlet process is such an example, obtained by normalizing a gamma process. The popularity of the Dirichlet process is partly because of the fact that besides this representation, it also has a whole host of alternative and equivalent representations, giving different useful perspectives and better understanding, and leading to alternative inference algorithms. In this section we will endeavor to derive similar levels of understanding for the larger class of normalized completely random measures.

As a running example, we will consider an important class of normalized completely random measures called normalized generalized gamma processes (NGGPs). Generalized gamma processes are completely random measures where the Lévy measure,

$$\nu_{\alpha,\beta,\sigma}(dw) = \frac{\alpha}{\Gamma(1-\sigma)} w^{-1-\sigma} e^{-\beta w} dw, \tag{164}$$

is parameterized by a concentration parameter $\alpha > 0$, a rate parameter $\beta \geq 0$ and an index parameter $0 \leq \sigma < 1$. We recover the gamma process when $\sigma = 0$ and $\beta > 0$, whence in this case the normalized generalized gamma process encompasses the Dirichlet process. Another subclass, called the normalized stable process, is obtained by setting $\beta = 0$ and $\alpha = \sigma > 0$. This class is also a subclass of the Pitman-Yor process, obtained when the concentration parameter of the Pitman-Yor process is set to 0. A third subclass, called the normalized inverse Gaussian process, is obtained by setting $\sigma = 1/2$. In the following, quantities subscripted by α, β, σ will refer to the normalized generalized gamma process, while those without will refer to the whole class of normalized completely random measures.

Note that there is in fact a degree of redundancy in the parameterization of the NGGP. Specifically, if we rescale all the atom masses $W_k \mapsto cW_k$ by a positive constant c, the

resulting random measure is still completely random. The transformed Lévy measure can be obtained by a change of variable,

$$\frac{\alpha}{\Gamma(1-\sigma)}(w/c)^{-1-\sigma}e^{-\beta w/c}\frac{1}{c}dw, \tag{165}$$

which can be seen as the Lévy measure for a generalized gamma process with parameter $(\alpha c^{\sigma}, \beta/c, \sigma)$. Rescaling the atom masses does not affect the distribution of the normalized measure, so that $\tilde{G}_{\alpha,\beta,\sigma}$ and $\tilde{G}_{\alpha c^{\sigma},\beta/c,\sigma}$ have the same distribution for each $c > 0$. In other words the space of NGGPs is two-dimensional. In the above we used a parameterization with three parameters to make the connections to other random probability measures more explicit.

12.1 Lévy Measures for Completely Random Measures

A first question we seek an answer to is the following: What conditions do the Lévy measure ν have to satisfy in order for our construction to be valid, and what conditions will guarantee that the total mass T is positive and finite, so that the normalization operation is well-defined? Firstly, for the CRM to not be degenerate, we require the Laplace transform of the total mass to be well-defined. In particular, the integral in the Lévy-Khinchin formula Eq. (144) should be finite when $f(\xi) = w$:

$$\int_{\Theta}\int_{0}^{\infty}(1-e^{-sw})\mu(d\xi) = \int_{0}^{\infty}(1-e^{-sw})\nu(dw) < \infty, \tag{166}$$

for each $s > 0$. In fact, it is easy to see that if the integral is finite for any s, it is finite for all $r > 0$: Suppose that r is another real number with $r > s$. Then,

$$1-e^{-sw} < 1-e^{-rw} < \frac{s}{r}(1-e^{-sw}) \tag{167}$$

so that both integrals are finite simultaneously.

For positiveness, it is sufficient to guarantee that there is at least one atom in G, since this atom will have positive mass. The number of atoms in G is the same as that for N, which is Poisson distributed with mean given by:

$$\mu(\Theta \otimes \mathbb{R}^{+}) = G_0(\Theta)\int_{0}^{\infty}\nu(dw) = \int_{0}^{\infty}\nu(dw), \tag{168}$$

the mass of the Lévy measure ν. If this is finite, then there is positive probability for N to have no atoms. On the other hand, if this is infinite, then N will almost surely have an infinite number of atoms. Thus the second condition on ν is for it to have infinite mass,

$$\int_{0}^{\infty}\nu(dw) = \infty. \tag{169}$$

For finiteness of T, one possibility is to require that T has a finite expectation, which is sufficient to guarantee finiteness. Referring to Eq. (140),

$$\mathbb{E}[T] = \int w\mu(d\xi) = G_0(\Theta)\int_{0}^{\infty}w\nu(dw) = \int_{0}^{\infty}w\nu(dw) < \infty, \tag{170}$$

which gives the condition on ν for T to have finite expectation. However, it is possible for a random variable to be finite while at the same time have an infinite expectation. For example, when $\nu(dw) = (1+w)^{-1.5}dw$, we have $\int_0^\infty \nu(dw) < \infty$ so that G has a finite number of atoms and hence T is finite. But $\int_0^\infty w\nu(dw) = \infty$ and T has infinite expectation. This example shows that a small (finite) number of atoms with very large masses could contribute significantly to the total mass T, such that the expectation of T becomes infinite.

For a more precise condition on ν for T to be finite, we can thus treat the atoms with large masses separately from those with small masses. Let r be a positive number, for example, $r = 1$. The total mass T is the sum of the total mass $T_{>r}$ of those atoms with mass larger than r, and of the total mass $T_{\leq r}$ among those with mass below r, and T is finite if and only if both $T_{>r}$ and $T_{\leq r}$ are. Since each atom whose mass contributes to $T_{>r}$ has mass at least $r > 0$, a necessary and sufficient condition for $T_{>r}$ to be finite is for the number of contributing atoms to be finite. This is equivalent to the condition:

$$\int_0^\infty \mathbb{1}_{(r,\infty)}(w)\nu(dw) < \infty. \tag{171}$$

On the other hand, since the total number of atoms has to be infinite, the number below r has to be infinite. Following previous discussion, a sufficient condition for $T_{\leq r}$ to be finite is for its expectation to be, which is equivalent to the condition:

$$\int_0^\infty w\mathbb{1}_{(0,r)}(w)\nu(dw) < \infty. \tag{172}$$

We can combine both conditions together into one by adding r times Eq. (171) to Eq. (172), which simplifies to:

$$\int_0^\infty \min(w,r)\nu(dw) < \infty. \tag{173}$$

This gives a third condition on the Lévy measure ν. Up to a multiplicative constant, the function $\min(w,r)$ has the same asymptotic behavior as $1-e^{-tw}$ in Eq. (166): it asymptotes to a constant for large w, and linear in w as $w \to 0$, and it can be shown that Eq. (166) is finite exactly when Eq. (173) is, so that this third condition is in fact the same as for the first condition Eq. (166).

In summary, we require that the Lévy measure ν satisfy the two conditions Eq. (166) and Eq. (169). We will write $\tilde{G} \sim \mathrm{NRM}(\nu, G_0)$ for the random probability measure $\tilde{G}$ constructed as in Eq. (163), and call it a normalized random measure (NRM). We will also write $G \sim \mathrm{CRM}(\nu, G_0)$ for the completely random measure whose normalization gives $\tilde{G}$.

It is straightforward to verify that the Lévy measure of the generalized gamma process given in Eq. (164) satisfies the two conditions set out above. Firstly, $\nu_{\alpha,\beta,\sigma}$ has infinite mass since its density grows at a scale of $w^{-1-\sigma}$ as $w \to 0$. Secondly, the integral in the

Lévy-Khinchin formula can be calculated by integration by parts,

$$\begin{aligned}
&\int_0^\infty (1-e^{-sw})\frac{\alpha}{\Gamma(1-\sigma)}w^{-1-\sigma}e^{-\beta w}dw \\
=&\frac{\alpha}{\Gamma(1-\sigma)}\int_0^\infty (e^{-\beta w}-e^{-(s+\beta)w})w^{-1-\sigma}dw \\
=&\frac{\alpha}{\Gamma(1-\sigma)}\left(\left[e^{-\beta w}-e^{-(s+\beta)w})(-\sigma)^{-1}w^{-\sigma}\right]_0^\infty - \int_0^\infty (-\beta e^{-\beta w}+(s+\beta)e^{-(s+\beta)w})(-\sigma)^{-1}w^{-\sigma}dw\right) \\
=&\frac{\alpha}{\Gamma(1-\sigma)}\left(\frac{s+\beta}{\sigma}\int_0^\infty e^{-(s+\beta)w}w^{-\sigma}dw - \frac{\beta}{\sigma}\int_0^\infty e^{-\beta w}w^{-\sigma}dw\right) \\
=&\frac{\alpha}{\Gamma(1-\sigma)}\left(\frac{s+\beta}{\sigma}\frac{\Gamma(1-\sigma)}{(s+\beta)^{1-\sigma}} - \frac{\beta}{\sigma}\frac{\Gamma(1-\sigma)}{(\beta)^{1-\sigma}}\right) \\
=&\frac{\alpha}{\sigma}((s+\beta)^\sigma-\beta^\sigma), \qquad (174)
\end{aligned}$$

which we see is finite. Using L'Hopital's rule, we also see that the integral approaches that of the gamma process in Eq. (146) as $\sigma \to 0$. The above gives the Laplace transform of the total mass $T_{\alpha,\beta,\sigma}$ of the generalized gamma process,

$$\mathbb{E}[e^{-sT_{\alpha,\beta,\sigma}}] = \exp\left(-\frac{\alpha}{\sigma}((s+\beta)^\sigma-\beta^\sigma)\right), \qquad (175)$$

which will be an oft-used result in the following.

13 The Posterior Normalized Random Measure

Suppose G is a CRM with diffuse base distribution G_0 and Lévy measure ν satisfying the two integral conditions (166) and (169), and $\tilde{G}$ is the random probability measure defined by normalizing G via Eq. (163). Just as for the DP, we can elucidate the properties of the NRM by treating it as a distribution from which we can draw iid samples,

$$\theta_n \,|\, \tilde{G} \overset{iid}{\sim} \tilde{G} \qquad \text{for } n = 1, 2, \ldots, N. \qquad (176)$$

As $\tilde{G}$ is discrete by construction, there will be repeated values among $\theta_1, \ldots, \theta_N$, which induces a random exchangeable partition of $[N]$. Let us denote the random partition by $\boldsymbol{\pi}$, and denote the unique value corresponding to each cluster c in $\boldsymbol{\pi}$ by ϕ_c. We will now study the structure of these random objects as a way to understand the NRM.

The starting point of our study will be in understanding the posterior distribution over both the normalized random measure $\tilde{G}$ and the unnormalized measure G conditioned on the iid draws from it, using the calculus we have developed for Poisson random measures in Sec. (9). We will need an additional tool for Poisson random measures called the Palm formula, which we will develop along the way. In the next section, using our newfound understanding, we can construct a mixture model where the mixing measure has a NRM prior, and discuss posterior inference schemes in this NRM mixture model.

We start with the conditional distribution of $\boldsymbol{\pi}$ and $(\phi_c)_{c\in\boldsymbol{\pi}}$ given the random measure. Let $\varphi_c \in \Theta$ be a value for each $c \in \boldsymbol{\pi}$. We can write the conditional distribution as follows:

$$P(\boldsymbol{\pi} = \pi, \text{ and } \phi_c \in d\varphi_c \text{ for } c \in \pi \,|\, G) = \prod_{c\in\pi}\tilde{G}(d\varphi_c)^{|c|} = T^{-N}\prod_{c\in\pi}G(d\varphi_c)^{|c|}, \qquad (177)$$

where for a value $\varphi \in \Theta$, $d\varphi$ denotes a small neighbourhood around φ, and $\phi \in d\varphi$ denotes the event that ϕ is in the neighbourhood. Since $\tilde{G}$ is a discrete probability measure, we have that $\tilde{G}(d\varphi)$ equals 0 if none of its atoms equal φ, and equals the mass of the atom if an atom is located precisely at φ.

Consider the problem of determining the posterior distribution of G given an observation that $\boldsymbol{\pi} = \pi$ and $\phi_c \in d\varphi_c$ for $c \in \pi$. Intuitively, we can compute a posterior by "multiplying a prior with a likelihood and renormalizing". The likelihood, given in Eq. (177), consists of $K+1$ terms, a term for each of the $K = |\pi|$ unique values and T^{-N}. The T^{-N} term turns out to be the more difficult term to work with as T is a sum of the masses of all the atoms in G, so that the term couples all the atoms of G in the posterior in a non-trivial manner.

Fortunately, we can perform an uncoupling by augmenting the system with an additional positive-valued random variable, which we denote by U. Specifically, let U be a gamma random variable with shape parameter N and rate parameter T which is conditionally independent of $\boldsymbol{\pi}$ and $(\phi_c)_{c\in\boldsymbol{\pi}}$ given G. The conditional distribution of U given G is:

$$P(U \in du \,|\, G) = \frac{T^N}{\Gamma(N)} u^{N-1} e^{-Tu} du, \tag{178}$$

and so the joint conditional distribution becomes,

$$P(\boldsymbol{\pi} = \pi, \phi_c \in d\varphi_c \text{ for } c \in \pi, U \in du \,|\, G) = \frac{1}{\Gamma(N)} u^{N-1} e^{-Tu} du \prod_{c\in\pi} G(d\varphi_c)^{|c|} \tag{179}$$

Given $\boldsymbol{\pi} = \pi$, $\phi_c \in d\varphi_c$ for $c \in \pi$, and $U \in du$, we will now attempt to determine the posterior distribution of G. For smooth real-valued random variables we can often determine the distribution by calculating the density. G is a random measure and it seems less clear whether we can proceed by calculating its density. For real-valued random variables, there are other quantities that we can calculate to determine the distribution instead. For example, in Sec. (9.3) we used the Laplace transform to determine the marginal distribution of $G(A)$ when G is the gamma process.

For random measures, a generalization of the Laplace transform which we can use to determine the distribution is called the Laplace functional. Instead of being a function of a positive real number s, the Laplace functional is a function of a function $f(\phi)$:

$$f \mapsto \mathbb{E}\left[e^{-\int f(\phi) G(d\phi)}\right] \tag{180}$$

If two random measures G and G' have the same Laplace functional, that is, Eq. (180) takes the same value for both G and G', for all functions f in a large enough class, then G and G' will have the same distribution.

For a completely random measure G, we can compute its Laplace functional by using the fact that its atoms are constructed from the atoms of a Poisson random measure N, and using the Lévy-Khinchin formula Eq. (144) in Sec. (9.3). Using Eq. (161), we have,

$$\int f(\phi) G(d\phi) = \sum_k W_k f(\phi_k) = \int w f(\phi) \mathsf{N}(d\phi, dw). \tag{181}$$

Evaluating the Lévy-Khinchin formula at the function $(\phi, w) \mapsto wf(\phi)$ and at $t = 1$, we can write Eq. (180) as:

$$\mathbb{E}\left[e^{-\int f(\phi)G(d\phi)}\right] = \exp\left(-\int_\Theta \int_0^\infty (1 - e^{-wf(\phi)})\nu(dw)G_0(d\phi)\right). \tag{182}$$

Returning to the computation of the posterior random measure given observations, we would like to compute the Laplace functional of the posterior G. This can be achieved by calculating the posterior expectation,

$$\begin{aligned}
&\mathbb{E}\left[e^{-\int f(\phi)G(d\phi)} \,\middle|\, \boldsymbol{\pi} = \pi, \phi_c \in d\varphi_c \text{ for } c \in \pi, U \in du\right] \\
=&\frac{\mathbb{E}\left[e^{-\int f(\phi)G(d\phi)} P(\boldsymbol{\pi} = \pi, \phi_c \in d\varphi_c \text{ for } c \in \pi, U \in du \,|\, G)\right]}{\mathbb{E}\left[P(\boldsymbol{\pi} = \pi, \phi_c \in d\varphi_c \text{ for } c \in \pi, U \in du \,|\, G)\right]}
\end{aligned} \tag{183}$$

If we can compute the numerator, the denominator can be obtained by just setting $f(\phi) = 0$. Plugging in Eq. (179), the numerator becomes,

$$\begin{aligned}
&\mathbb{E}\left[e^{-\int f(\phi)G(d\phi)} \frac{1}{\Gamma(N)} u^{N-1} e^{-Tu} du \prod_{c\in\pi} G(d\varphi_c)^{|c|}\right] \\
=&\frac{1}{\Gamma(N)} u^{N-1} du \mathbb{E}\left[e^{-\int (f(\phi)+u)G(d\phi)} \prod_{c\in\pi} G(d\varphi_c)^{|c|}\right],
\end{aligned} \tag{184}$$

where the expectation is with respect to the completely random measure G. We can rewrite it in terms of the underlying Poisson random measure N as follows,

$$=\frac{1}{\Gamma(N)} u^{N-1} du \mathbb{E}\left[e^{-\int (f(\phi)+u)w\mathsf{N}(d\xi)} \prod_{c\in\pi} \int \delta_\phi(d\varphi_c) w^{|c|} \mathsf{N}(d\xi)\right], \tag{185}$$

which is the expectation of a somewhat complicated functional of a Poisson random measure. We can attempt to evaluate this expectation using the two useful formulas derived in Sec. (9.3). The first, Eq. (140), allows us to compute the mean of a functional of a Poisson random measure which can be written in the form,

$$\int q(\xi)\mathsf{N}(d\xi), \tag{186}$$

while Eq. (144) allows us to compute the Laplace transform of Eq. (186) instead, which is the mean of a functional of the form,

$$e^{\int -g(\xi)\mathsf{N}(d\xi)}. \tag{187}$$

Unfortunately, the functional in Eq. (185) is a product of terms of both forms, and neither formula is applicable. To evaluate this expectation we will have to derive a third formula called the Palm formula.

13.1 Palm Formula

Suppose $g(\xi)$ and $q(\xi)$ are two functions over a measure space and N is a Poisson random measure defined over the same space with mean measure μ. We would like to compute an expectation of the form,

$$\mathbb{E}\left[e^{-\int g(\xi)\mathsf{N}(d\xi)}\int q(\xi)\mathsf{N}(d\xi)\right]. \tag{188}$$

We can derive this expectation using the Lévy-Khinchin formula Eq. (144) as follows. Suppose $s \geq 0$. Then,

$$\mathbb{E}\left[e^{-\int -sq(\xi)+g(\xi)\mathsf{N}(d\xi)}\right] = \exp\left(-\int 1 - e^{sq(\xi)-g(\xi)}\mu(d\xi)\right). \tag{189}$$

Differentiating both sides with respect to s,

$$\mathbb{E}\left[e^{-\int -sq(\xi)+g(\xi)\mathsf{N}(d\xi)}\int q(\xi)\mathsf{N}(d\xi)\right] = \exp\left(-\int 1 - e^{sq(\xi)-g(\xi)}\mu(d\xi)\right)\int q(\xi)e^{sq(\xi)-g(\xi)}\mu(d\xi).$$

Finally, setting $s = 0$ and rearranging the integrals,

$$\begin{aligned}\mathbb{E}\left[e^{-\int g(\xi)\mathsf{N}(d\xi)}\int q(\xi)\mathsf{N}(d\xi)\right] &= \exp\left(-\int 1 - e^{-g(\xi)}\mu(d\xi)\right)\int q(\xi)e^{-g(\xi)}\mu(d\xi) \\ &= \int q(\xi)\exp\left(-g(\xi) - \int 1 - e^{-g(\xi')}\mu(d\xi')\right)\mu(d\xi),\end{aligned} \tag{190}$$

which is the desired formula.

It is worthwhile pausing here to give an interpretation to the above result. Using the Lévy-Khinchin formula, we can identify the terms involving $g(x)$ on the RHS as the expectation of a functional of $\mathsf{N} + \delta_\xi$, a random measure constructed by adding a single fixed atom δ_ξ to the Poisson random measure N,

$$\mathbb{E}\left[e^{-\int g(\xi)\mathsf{N}(d\xi)}\int q(\xi)\mathsf{N}(d\xi)\right] = \int \mathbb{E}\left[e^{-\int g(\xi')(\mathsf{N}+\delta_\xi)(d\xi')}\right]q(\xi)\mu(d\xi). \tag{191}$$

Writing the exponential term on the LHS as simply a functional $g' : \mathsf{N} \mapsto e^{-\int g(\xi)\mathsf{N}(d\xi)}$ and denoting by $\mathcal{P}$ the distribution of N, we have,

$$\int\int g'(\mathsf{N})q(\xi)\mathsf{N}(d\xi)\mathcal{P}(d\mathsf{N}) = \int\int g'(\mathsf{N}+\delta_\xi)q(\xi)\mathcal{P}(d\mathsf{N})\mu(d\xi). \tag{192}$$

The LHS is an integral of $g'(\mathsf{N})q(\xi)$ with respect to a measure associated with first drawing N from $\mathcal{P}$, followed by integrating ξ with respect to N. On the other hand, the RHS is an integral with respect to a measure whereby we first integrate ξ with respect to the mean measure μ, followed by a conditional measure given by a Poisson random measure with an additional fixed atom at ξ. Loosely speaking, both sides describe the same expectation over a pair of random elements N and ξ, one a Poisson random measure, and the other a random

atom chosen from the measure. The RHS states loosely that conditioned on there being an atom at ξ, the conditional structure of N on the rest of the space is precisely the same as its prior, a Poisson random measure with the same mean measure μ. This structure follows from the fact that N is completely random, so knowing that there is an atom at ξ tells us precisely no information about the measure on the rest of the space.

Generally, the theory of Palm distributions studies the distribution of a point process (not necessarily Poisson) conditioned on there being an atom at some location ξ, and Eq. (192) is sometimes known as the Palm formula. As before, we can construct progressively more complex functions from simpler ones, and show that the Palm formula for Poisson processes holds for general functions $g(\xi, \mathsf{N})$ of both ξ and N:

$$\int \int g(\xi, \mathsf{N}) q(\xi) \mathsf{N}(d\xi) \mathcal{P}(d\mathsf{N}) = \int \int g(\xi, \mathsf{N} + \delta_\xi) q(\xi) \mathcal{P}(d\mathsf{N}) \mu(d\xi). \tag{193}$$

Further, this formula characterizes the Poisson random measure, that is, it holds precisely when N is a Poisson random measure.

13.2 Deriving the Posterior of a Normalized Random Measure

Equipped with knowledge of the Palm formula, we can now tackle the expectation in Eq. (185) with a slight generalization. Specifically, let $g(\xi) = (f(\phi) + u)w$ and for each $c \in \pi$ let $q_c(\xi) = \delta_\phi(d\varphi_c) w^{|c|}$ and $s_c \geq 0$. Then the Lévy-Khinchin formula gives,

$$\mathbb{E}\left[e^{-\int g(\xi) - \sum_{c\in\pi} s_c q_c(\xi) \mathsf{N}(d\xi)}\right] = \exp\left(-\int 1 - e^{-(g(\xi) - \sum_{c\in\pi} s_c q_c(\xi))} \mu(d\xi)\right) \tag{194}$$

Differentiating with respect to each s_c, and evaluating the derivative at $s_c = 0$ for all $c \in \pi$, we get,

$$\begin{aligned}
&\mathbb{E}\left[e^{-\int (f(\phi)+u) w \mathsf{N}(d\xi)} \prod_{c\in\pi} \int \delta_\phi(d\varphi_c) w^{|c|} \mathsf{N}(d\xi)\right] \\
=&\int \cdots \int \mathbb{E}\left[e^{-\int (f(\phi)+u) w (\mathsf{N} + \sum_{c\in\pi} \delta_{\phi_c, w_c})(d\phi, dw)}\right] \prod_{c\in\pi} \delta_{\phi_c}(d\varphi_c) w_c^{|c|} \mu(d\phi_c, dw_c) \\
=&\mathbb{E}\left[e^{-\int\int (f(\phi)+u) w \mathsf{N}(d\phi, dw)}\right] \prod_{c\in\pi} G_0(d\varphi_c) \int e^{-(f(\varphi_c)+u)w_c} w_c^{|c|} \nu(dw_c) \\
=&\exp\left(-\int\int 1 - e^{-(f(\phi)+u)w} G_0(d\phi)\nu(dw)\right) \prod_{c\in\pi} G_0(d\varphi_c) \int e^{-(f(\varphi_c)+u)w_c} w_c^{|c|} \nu(dw_c).
\end{aligned} \tag{195}$$

Setting $f(\phi) = 0$ in the above gives the denominator in Eq. (183), which is the marginal probability of the observations:

$$\begin{aligned}
&P(\boldsymbol{\pi} = \pi, \phi_c \in d\varphi_c \text{ for } c \in \pi, U \in du) \\
=&\mathbb{E}\left[P(\boldsymbol{\pi} = \pi, \phi_c \in d\varphi_c \text{ for } c \in \pi, U \in du \,|\, G)\right] \\
=&\frac{1}{\Gamma(N)} u^{N-1} du \exp\left(-\int 1 - e^{-uw} \nu(dw)\right) \prod_{c\in\pi} G_0(d\varphi_c) \int e^{-uw_c} w_c^{|c|} \nu(dw_c).
\end{aligned} \tag{196}$$

Finally, substituting both results for numerator and denominator back into Eq. (183), we get,

$$\begin{aligned}
&\mathbb{E}\left[e^{-\int f(\phi)G(d\phi)} \,\middle|\, \boldsymbol{\pi} = \pi, \phi_c \in d\varphi_c \text{ for } c \in \pi, U \in du\right] \\
=&\frac{\exp\left(-\int\int 1 - e^{-(f(\phi)+u)w} G_0(d\phi)\nu(dw)\right) \prod_{c\in\pi} G_0(d\varphi_c) \int e^{-(f(\varphi_c)+u)w_c} w_c^{|c|} \nu(dw_c)}{\exp\left(-\int 1 - e^{-uw}\nu(dw)\right) \prod_{c\in\pi} G_0(d\varphi_c) \int e^{-uw_c} w_c^{|c|}\nu(dw_c)} \\
=&\exp\left(-\int\int 1 - e^{-f(\phi)w} G_0(d\phi) e^{-uw}\nu(dw)\right) \prod_{c\in\pi} \frac{\int e^{-f(\varphi_c)w_c} e^{-uw_c} w_c^{|c|}\nu(dw_c)}{\int e^{-uw_c} w_c^{|c|}\nu(dw_c)}
\end{aligned} \tag{197}$$

The Laplace functional above is a product of terms, so the posterior G can be expressed as a sum of independent random elements, one corresponding to each term. The first term is in the form of the Laplace functional for a completely random measure with an exponentially tilted Lévy measure. Specifically, the mean measure of the underlying Poisson random measure is given by $G_0 \otimes \nu'$ where $\nu'(dw) = e^{-uw}\nu(dw)$. For each $c \in \pi$, the cth term above corresponds to the Laplace functional of a random measure consisting of a single fixed atom at φ_c, with random mass distribution given by,

$$P(W_c \in dw_c | \boldsymbol{\pi} = \pi, \phi_c \in d\varphi_c \text{ for } c \in \pi, U \in du) = \frac{e^{-uw_c} w_c^{|c|}\nu(dw_c)}{\int e^{-uv_c} v_c^{|c|}\nu(dv_c)}. \tag{198}$$

In summary, the posterior unnormalized random measure G can be written as follows:

$$\begin{aligned}
G|\{\boldsymbol{\pi} = \pi, \phi_c \in d\varphi_c \text{ for } c \in \pi, U \in du\} &= G' + \sum_{c\in\pi} W_c \delta_{\varphi_c} \\
G' \sim \mathrm{CRM}(\nu', G_0) \qquad \nu'(dw) &= e^{-uw}\nu(dw) \\
W_c \sim p_c \qquad\qquad\quad p_c(dw_c) &= \frac{e^{-uw_c} w_c^{|c|}\nu(dw_c)}{\int e^{-uv_c} v_c^{|c|}\nu(dv_c)}
\end{aligned} \tag{199}$$

The posterior normalized random measure is then obtained by normalizing the posterior G. Note that the posterior G is still a completely random measure, albeit one with fixed atoms. The fixed atoms correspond to locations where we have observed draws from $\tilde{G}$, while the non-fixed atoms in the smooth component G' correspond to other atoms in the discrete random measure which do not correspond to observed draws. The exponential tilting to the Lévy measure of G' encourages atoms in G' to have lower masses than those in the prior distribution, and is due to the observation that there are no drawn values elsewhere besides the observed values $\{\varphi_c\}_{c\in\pi}$.

The actual posterior of G given only $\theta = (\theta_1, \ldots, \theta_N)$ can now be obtained by integrating over the auxiliary variable U. We can achieve this by noting that the posterior distribution of U given θ can be obtained analytically by normalizing Eq. (196) with respect to u, though this posterior is not of a standard form. However it is straightforward, within a Markov chain Monte Carlo framework, to derive an algorithm for sampling from the joint posterior of both U and G given θ, which we shall see in a later section.

Returning to the normalized generalized gamma process running example, the posterior G has Lévy measure given in Eq. (164), so that the posterior smooth component G' has Lévy measure,

$$\nu'(dw) = \frac{\alpha}{\Gamma(1-\sigma)} w^{-1-\sigma} e^{-(\beta+u)w} dw = \nu_{\alpha,\beta+u,\sigma}(dw), \tag{200}$$

that is, G' is also a generalized gamma process, with an updated rate parameter $\beta + u$. On the other hand, the mass of the fixed atom at φ_c has density,

$$p_c(W_c \in dw_c) \propto e^{-uw_c} w_c^{|c|} \frac{\alpha}{\Gamma(1-\sigma)} w_c^{-1-\sigma} e^{-\beta w_c} dw_c \propto w_c^{|c|-\sigma-1} e^{-(\beta+u)w_c} dw_c, \tag{201}$$

that is, W_c is gamma distributed with shape $|c| - \sigma$ and rate $\beta + u$.

When $\sigma = 0$, we see that the posterior smooth component G' is a gamma process while the fixed atoms have gamma distributed masses. Both the gamma process and the gamma distributions of the masses of the fixed atoms have the same rate parameter $\beta + u$, so that the posterior normalized random measure $\tilde{G}$ is still a DP, recovering the posterior DP result we derived in Eq. (60). Note that in normalizing this random measure, the effect of the rate parameter is normalized away so that the posterior DP does not depend on β nor on U. In particular, $\tilde{G}$ and U are conditionally independent in the posterior given θ. This is one of the simplifying properties which is unique to the DP.

13.3 Random Partitions Induced by a Normalized Random Measure

As noted previously, the discrete nature of the normalized random measure $\tilde{G}$ naturally induces an exchangeable random partition $\boldsymbol{\pi}$ of $[N]$, where the clusters in $\boldsymbol{\pi}$ correspond to the unique values in the iid draws $\theta_1, \ldots, \theta_N \overset{iid}{\sim} \tilde{G}$. The probability distribution of the random partition can be obtained by integrating out the unique values and U from Eq. (196),

$$P(\boldsymbol{\pi} = \pi) = \frac{1}{\Gamma(N)} \int_0^\infty u^{N-1} \exp\left(- \int 1 - e^{-uw} \nu(dw) \right) \prod_{c \in \pi} \int_0^\infty e^{-uw_c} w_c^{|c|} \nu(dw_c) du. \tag{202}$$

For the normalized generalized gamma process, plugging in the Lévy measure and the Laplace transform of the total mass Eq. (175),

$$\begin{aligned} P(\boldsymbol{\pi}_{\alpha,\beta,\sigma} = \pi) =& \frac{1}{\Gamma(N)} \int_0^\infty u^{N-1} e^{-\frac{\alpha}{\sigma}((u+\beta)^\sigma - \beta^\sigma)} \prod_{c \in \pi} \int_0^\infty e^{-uw_c} w_c^{|c|} \frac{\alpha}{\Gamma(1-\sigma)} w_c^{-1-\sigma} e^{-\beta w_c} dw_c du \\ =& \frac{1}{\Gamma(N)} \int_0^\infty u^{N-1} e^{-\frac{\alpha}{\sigma}((u+\beta)^\sigma - \beta^\sigma)} \prod_{c \in \pi} \alpha \frac{\Gamma(|c| - \sigma)}{\Gamma(1-\sigma)} (u+\beta)^{\sigma - |c|} du \\ =& \frac{\alpha^{|\pi|}}{\Gamma(N)} \int_0^\infty u^{N-1} (u+\beta)^{\sigma|\pi| - N} e^{-\frac{\alpha}{\sigma}((u+\beta)^\sigma - \beta^\sigma)} du \prod_{c \in \pi} (1-\sigma)^{(|c|-1)}, \end{aligned} \tag{203}$$

where $(1-\sigma)^{(|c|-1)} = (1-\sigma)(2-\sigma)\cdots(|c|-1-\sigma)$.

Comparing Eq. (203) to Eq. (77) for the Pitman-Yor process, the terms involving the clusters of $\boldsymbol{\pi}_{\alpha,\beta,\sigma}$ are exactly the same, with the main difference between the two being the first term which involves the number of clusters in $\boldsymbol{\pi}_{\alpha,\beta,\sigma}$. This similarity begets an underlying relationship between the two processes, which we will explore in Sec. (16).

Recall also the power law properties of the Pitman-Yor process, the origins of which can be traced to the effect a positive value for σ has on the probabilities of data items (customers) joining clusters (tables) of varying sizes. Specifically, it reduces the relative probabilities of joining small clusters much more than large clusters, and as a result there is a predilection for a large number of small clusters and a small number of large ones. An effect of this is in the form of the cluster terms in Eq. (77). Since Eq. (203) has the same cluster terms, we expect similar power law properties are in effect in the random partition induced by the normalized generalized gamma process, which is indeed the case.

14 Sampling Algorithms for NRM Mixture Model

Equipped with an understanding of the distribution of the exchangeable random partition induced by a normalized random measure, and of the posterior distribution of the normalized random measure, we can now specify a mixture model using a normalised random measure as its mixing measure, and derive sampling algorithms for posterior simulation. The full specification of the model is as follows,

$$\begin{aligned}
&G \sim \mathrm{CRM}(\nu, G_0) && T = G(\Theta) && \tilde{G} = \frac{G}{T} \\
&U|G \sim \mathrm{Gamma}(N, T) \\
&\theta_i|G \overset{iid}{\sim} \tilde{G} && \text{for } i = 1, \ldots, N, \\
&x_i|\theta_i \overset{ind}{\sim} F(\theta_i) && \text{for } i = 1, \ldots, N.
\end{aligned} \tag{204}$$

For completeness, we have included the auxiliary variable U as well. In the following we will describe both a marginal sampler and a conditional sampler.

14.1 Marginal Gibbs Sampling for NRM Mixture Model

Integrating out the random measure G and introducing the induced random partition $\boldsymbol{\pi}$ of $[N]$, (196) leads to the joint probability over $\boldsymbol{\pi}$, U, the component parameters $\phi = (\phi_c)_{c\in\boldsymbol{\pi}}$, and the observations $x = (x_i)_{i\in[N]}$,

$$p(\boldsymbol{\pi}, \phi, U, x) = U^{N-1} \exp\left(-\int 1 - e^{-Uw}\nu(dw)\right) \prod_{c\in\boldsymbol{\pi}} \kappa_\nu(|c|, U) g_0(\phi_c) \prod_{i\in c} f(x_i|\phi_c) \tag{205}$$

where g_0 and f are the densities for the base measure G_0 and component distribution F respectively, and $\kappa_\nu(m, u)$ is the gamma integral,

$$\kappa_\nu(m, u) = \int_0^\infty e^{-uw} w^m \nu(dw) \tag{206}$$

For simplicity, we assume we can further marginalise out the component parameters, leading to,

$$p(\boldsymbol{\pi}, U, x) = U^{N-1} \exp\left(-\int 1 - e^{-Uw}\nu(dw)\right) \prod_{c\in\boldsymbol{\pi}} \kappa_\nu(|c|, U) f(x_c), \tag{207}$$

where

$$f(x_c) = \int g_0(\varphi_c) \prod_{i\in c} f(x_i|\varphi_c) d\varphi_c \tag{208}$$

At this point it is now straightforward to derive a Gibbs sampler which alternately updates U and $\boldsymbol{\pi}$. The conditional distribution of U given $\boldsymbol{\pi}$ is not of standard form, but a variety of techniques can be straightforwardly used as U is a one-dimensional random variable, for example slice sampling or Metropolis-within-Gibbs. For updating $\boldsymbol{\pi}$, consider updating the cluster assignment of data point i. Let $\boldsymbol{\pi}_{-i}$ denote the partition of $[N]\backslash\{i\}$ with i removed from $\boldsymbol{\pi}$, and $\boldsymbol{\pi}^+$ be the partition resulting once i is assigned to either a cluster $c \in \boldsymbol{\pi}^+$ or a new cluster. From Eq. (207), the conditional probabilities of the two cases are,

$$p(\boldsymbol{\pi}^+ \,|\, x, U) \propto \begin{cases} \frac{\kappa_\nu(|c|+1,U)}{\kappa_\nu(|c|,U)} f(x_i \,|\, x_c) & \text{for } \boldsymbol{\pi}^+ = \boldsymbol{\pi}_{-i} - c + (c \cup \{i\}), c \in \boldsymbol{\pi}_{-i}, \\ \kappa_\nu(1, U) f(x_i) & \text{for } \boldsymbol{\pi}^+ = \boldsymbol{\pi}_{-i} + \{i\}, \end{cases} \tag{209}$$

where $f(x_i|x_c) = f(x_{c\cup\{i\}})/f(x_c)$ is the conditional probability of x_i under cluster c which currently contains data points x_c. This sampler is a direct generalisation of the Gibbs sampler for the CRP mixture model in Sec. (3.3)

Returning to our running example, for the normalised generalised gamma process, the gamma integral is,

$$\begin{aligned} \kappa_{\alpha,\beta,\sigma}(m, u) &= \int_0^\infty \frac{\alpha}{\Gamma(1-\sigma)} w^{-1-\sigma} e^{-\beta w} w^m e^{-uw} dw \\ &= \frac{\alpha}{(\beta+u)^{m-\sigma}} \frac{\Gamma(m-\sigma)}{\Gamma(1-\sigma)} \end{aligned} \tag{210}$$

and the conditional probabilities for assigning data point x_i to the various clusters simplify to:

$$p(\boldsymbol{\pi}^+ \,|\, x, U) \propto \begin{cases} (|c| - \sigma) f(x_i \,|\, x_c) & \text{for } \boldsymbol{\pi}^+ = \boldsymbol{\pi}_{-i} - c + (c \cup \{i\}), c \in \boldsymbol{\pi}_{-i}, \\ \alpha(\beta + u)^\sigma f(x_i) & \text{for } \boldsymbol{\pi}^+ = \boldsymbol{\pi}_{-i} + \{i\}. \end{cases} \tag{211}$$

When $\sigma = 0$, we see that the above reduces to the Gibbs sampling update for the DP mixture model in Sec. (3.3).

For the DP mixture model, besides the marginal Gibbs sampler described here, other samplers based on alternative representations of the Dirichlet process can also be derived. These can be extended to the normalised random measures as well. In the next section, we will derive a stick-breaking representation for normalized random measures which is a direct generalisation of that for Dirichlet processes. Before that, we derive a conditional slice sampler for NRM mixture models that uses the represention of the posterior NRM given in Eq. (199).

14.2 Conditional Slice Sampler for NRM Mixture Model

The structure of the sampler follows closely the analogous sampler in Sec. (6.4) for the DP mixture model. Recall that the NRM mixture model Eq. (204) consists of a completely random measure G and parameters $\theta = (\theta_1, \ldots, \theta_N)$ corresponding to observations $x = (x_1, \ldots, x_N)$. In addition, there is an auxiliary variable U used in the characterization of the posterior NRM. We also introduce slice variables $s = (s_1, \ldots, s_N)$ with

$$s_i \,|\, G, \theta_i \sim \mathcal{U}[0, G(\{\theta_i\})]. \tag{212}$$

The conditional sampler consists mainly of two phases: sampling G, U and s given θ, and Gibbs sampling θ given G, U and s. Let $\boldsymbol{\pi}$ be the partition of $[N]$ induced by θ, with $(\phi_c)_{c\in\boldsymbol{\pi}}$ being the unique values, corresponding to the parameters of the clusters.

In the first phase, we update G, U and s given θ. We first update U, with G and s marginalized out, then sample both G and s from their joint conditional distribution given U and θ. As in Sec. (14.1), U can be updated using either slice sampling or Metropolis-within-Gibbs. The conditional distribution of G given θ is given by Eq. (199). In particular,

$$G \,|\, \theta = G' + \sum_{c\in\boldsymbol{\pi}} W_c \delta_{\phi_c}, \tag{213}$$

where G' consists of the atoms in G besides those corresponding to the values in θ. We first simulate the masses W_c of each atom ϕ_c, which are mutually independent, with distribution $P_c(dw_c) \propto e^{-Uw_c} w_c^{|c|} \nu(dw_c)$. Given these masses, the slice variables can be simulated independently, with s_i having distribution $\mathcal{U}[0, W_c]$ where $\in \boldsymbol{\pi}$ is the cluster index containing i. Finally, the smooth component G' can be simulated, with atoms drawn iid from G_0 and masses distributed according to a Poisson process with mean measure $v'(dw) = e^{-Uw}\nu(dw)$. The smooth component has infinitely many atoms, however only the finitely many with masses above $s_{\min} = \min_i s_i$ are needed in the second phase. We denote these $L \geq 0$ atoms in G' by $\phi'_1, \ldots, \phi'_L$, with masses $W'_1, \ldots, W'_L > s_{\min}$. A direct approach is to first simulate the number L of such atoms from a Poisson with rate

$$\int_{s_{\min}}^{\infty} \nu'(dw) \tag{214}$$

then to draw the masses $W'_1, \ldots, W'_L$ by sampling iid from the distribution supported on $[s_{\min}, \infty)$ with density proportional to ν' (Griffin and Walker, 2011). This approach requires computation of the integral Eq. (214) as well drawing samples from the non-standard distribution. Another approach which avoids both uses an adaptive thinning procedure (Favaro and Teh, 2013).

In the second phase, θ is updated given G and s. For each $i = 1, \ldots, N$, the values that θ_i can take on are those corresponding to the (finitely many) atoms in G with mass above s_i. As in Sec. (6.4), the probability that θ_i takes on the value of one of these atoms in G is simply proportional to the probability of observation x_i given θ_i:

$$p(\theta_i \,|\, G, s_i, x_i) \propto \begin{cases} f(x_i \,|\, \phi_c) & \text{for } \theta_i = \phi_c \text{ for some } c \in \boldsymbol{\pi} \text{ with } W_c > s_i, \\ f(x_i \,|\, \phi'_\ell) & \text{for } \theta_i = \phi'_\ell \text{ for some } \ell = 1, \ldots, L \text{ with } W'_\ell > s_i, \\ 0 & \text{otherwise.} \end{cases} \tag{215}$$

Finally, the cluster parameters ϕ_c for each cluster $c \in \boldsymbol{\pi}$ can be updated using a MCMC kernel with invariant distribution

$$p(\phi_c \,|\, x, \boldsymbol{\pi}) \propto g_0(\phi_c) \prod_{i \in c} f(x_i \,|\, \phi_c) \tag{216}$$

15 Stick-breaking Representation for Normalized Random Measure

In this section we will derive a stick-breaking representation for normalized random measures. As for the Dirichlet and Pitman-Yor processes, this gives a generative procedure for an enumeration of the atoms of the random measure. We will start with a discussion of the order in which the atoms of the random measure are enumerated, before using the Palm formula to derive the required generative procedure.

15.1 Size-biased Sampling

We start by recalling the derivation of the stick-breaking representation for Dirichlet processes in Sec. (5). The construction was derived starting with the Chinese restaurant process, with the mass of each atom being derived as the limiting proportion of customers sitting at the corresponding table. The first table is distinguished as being the table at which the first customer sat, and the limiting proportion of customers sitting at the table is derived to be $\beta_1 \sim \text{Beta}(1, \alpha)$, by identifying the sequence of whether subsequent customers sit at the table with a Pólya urn scheme with initial parameters 1 and α. For the second table, its proportion of customers is derived similarly, but now with the argument limited to the subsequence of customers who did not sit at the first table (which has asymptotic proportion $1 - \beta_1$). Among these, the second table is distinguished as the one that the first customer sits at, and the sequence of whether subsequent customers sit at the table is again described by a Pólya urn scheme with initial parameters 1 and α, so that the proportion of customers (among those who did not sit at the first table) who sat at the second table is $\beta_2 \sim \text{Beta}(1, \alpha)$. Among all customers, the proportion who sat at the second table is thus $\beta_2(1 - \beta_1)$. This process is repeated for each subsequent table to form the stick-breaking representation for the Dirichlet process.

There is a natural permutation of the atoms of the DP associated with the stick-breaking representation, known as a *size-biased permutation*. Suppose that instead of starting with the Chinese restaurant process, we start with a discrete probability measure $\tilde{G}$ distributed according to $\text{DP}(\alpha, G_0)$, with the collection of atoms and their associated masses in $\tilde{G}$ being $\{(\phi_k, \tilde{W}_k) : k = 1, 2, \ldots\}$, where the indexing by k is arbitrary. For example, we may simulate the atoms and their masses according to a gamma process (Sec. (9.2)), normalizing, and index the atoms by decreasing mass: $\tilde{W}_1 \geq \tilde{W}_2 \geq \cdots$. The Chinese restaurant process corresponds to the partition induced by a sequence $\theta_1, \theta_2, \ldots$ of iid draws from $\tilde{G}$, with the first table simply corresponding to the atom of $\tilde{G}$ picked by θ_1. This atom is ϕ_k with probability $\tilde{W}_k$, for each $k \in \mathbb{N}$. In other words it is a size-biased draw from among the atoms of $\tilde{G}$, with atoms with larger masses having higher chances of being drawn. Let 1^* be the index of the atom corresponding to the first table. Continuing the process, the second

table corresponds to the next distinct atom of $\tilde{G}$ encountered along the sequence $\theta_1, \theta_2, \ldots$, and it equals ϕ_k with probability $\tilde{W}_k / \sum_{\ell \neq 1^*} \tilde{W}_\ell$ for each $k \in \mathbb{N} \backslash \{1^*\}$. In other words, it is a size-biased draw from the atoms of $\tilde{G}$ other than ϕ_{1^*}. Let the index of the second atom be 2^*. In a similar fashion, we get $3^*, 4^*, \ldots$, with each index k^* being drawn without replacement, and equalling k with probability proportional to $\tilde{W}_k$. The sequence $1^*, 2^*, \ldots$ is a permutation of $\{1, 2, \ldots\}$, as each $k = 1, 2, \ldots$ will appear exactly once in the sequence. It is size-biased in the sense that indices with larger masses tend to appear earlier in the sequence.

15.2 The Stick-breaking Construction

Returning now to the normalized random measure, we can write the underlying unnormalized completely random measure as

$$G = \sum_{k=1}^{\infty} W_k \delta_{\phi_k} \tag{217}$$

where $W_k > 0$ are the masses of the atoms ϕ_k. Let the Lévy measure be ν and base distribution be G_0, and let the total mass be denoted $T = \sum_{k=1}^{\infty} W_k$, which we assume is positive and finite almost surely, and has a density f_T (which depends on ν).

Let $\theta_1, \theta_2, \ldots$ be a sequence of iid draws from the normalized random measure G/T. As above, this sequence induces a size-biased permutation $1^*, 2^*, \ldots$ of the atoms of G, where for each $k = 1, 2, \ldots$, we define k^* to be the index of the kth unique atom among those in G appearing in the sequence $\theta_1, \theta_2, \ldots$. In the case of the DP, this corresponds exactly to the kth table in the Chinese restaurant process. We shall use this size-biased permutation to construct a stick-breaking representation. In particular, let $W_k^* = W_{k^*}$ be the mass of the kth size-biased atom of G. Our aim is to derive the joint distribution of $W_1^*, W_2^*, \ldots$ as a sequential process, in which we can first draw W_1^*, followed by W_2^*, W_3^* and so forth. This enumerates all the atoms of G in a different order than that in Eq. (217):

$$G = \sum_{k=1}^{\infty} W_k^* \delta_{\phi_{k^*}}. \tag{218}$$

We will derive the joint distribution of the size-biased masses along with the total mass T, starting with the joint distribution of T and the first size-biased mass W_1^*. For each $k = 1, 2, \ldots$, the index 1^* is equal to k with probability proportional to W_k, so that the joint distribution of W_1^* and T is,

$$\begin{aligned} P(W_1^* \in dw_1, T \in dt | G) &= \sum_{k=1}^{\infty} \frac{W_k}{T} \delta_{W_k}(dw_1) \delta_T(dt) \qquad (219) \\ P(W_1^* \in dw_1, T \in dt) &= \mathbb{E}\left[\sum_{k=1}^{\infty} \frac{W_k}{T} \delta_{W_k}(dw_1) \delta_T(dt) \right] \\ &= \mathbb{E}\left[\int \frac{w}{T} \delta_w(dw_1) \delta_T(dt) \mathsf{N}(d\xi) \right] \end{aligned}$$

where $\mathsf{N} = \sum_{k=1}^{\infty} \delta_{\phi_k, W_k}$ denotes the Poisson random measure underlying the CRM G, and $\xi = (\phi, w)$. Using the Palm Formula and the fact that the mean measure of N is $G_0 \otimes \nu$,

$$\begin{aligned} &= \int \mathbb{E}\left[\frac{w}{T+w}\delta_w(dw_1)\delta_{T+w}(dt)\right]\nu(dw) \\ &= \frac{w_1}{t} f_T(t-w_1)dt\,\nu(dw_1) \end{aligned} \tag{220}$$

Normalizing the joint distribution, we get the conditional distribution of W_1^* given T,

$$P(W_1^* \in dw_1 \,|\, T \in dt) = \frac{w_1}{t}\frac{f_T(t-w_1)}{f_T(t)}\nu(dw_1) \tag{221}$$

The joint distribution above can roughly be interpreted as follows: $\nu(dw_1)$ is the "rate" of observing that an atom of mass w_1 exists in G, $f_T(t-w_1)$ is the density that the total mass of all other atoms is $t - w_1$ (so that the total mass is t), and w_1/t is the probability that the atom of mass w_1 was picked during size-biased sampling. Notice that the distribution of the total mass of all other atoms, given that there is an atom of mass w_1, is precisely the same as the distribution of the total mass of all atoms without knowledge of the atom with mass w_1. This is a consequence of the theory of the Palm distributions of the Poisson random measure described in Sec. (13.1), which implies that the distribution of the other atoms in G given the atom of mass w_1 is still as before and given by $\mathrm{CRM}(\nu, G_0)$.

We can apply the same argument to determine the distribution of the subsequent size-biased masses $W_2^*, W_3^*, \ldots$. One approach is to derive, for each $k \geq 1$, the joint probability

$$P(W_1^* \in dw_1, \ldots, W_k^* \in dw_k, T \in dt) \tag{222}$$

using the same technique as above, by first writing down the probability conditioned on G, then integrating out G using repeated applications of the Palm formula. Another, simpler but more indirect, way is to make use of the theory of Palm distributions. In particular, given $W_1^* \in dw_1, \ldots, W_{k-1}^* \in dw_{k-1}$, the distribution of the masses of all other atoms in G is precisely given by $\mathrm{CRM}(\nu, G_0)$. The difference is that while originally we condition on the total mass being t, now the total mass is reduced to $t - \sum_{j=1}^{k-1} w_j$ necessarily. Thus the conditional distribution of W_k^* is analogous to Eq. (221) with a reduced total mass,

$$P(W_k^* \in dw_k \,|\, T \in dt, W_1^* \in dw_1, \ldots, W_{k-1}^* \in dw_{k-1}) = \frac{w_k}{t - \sum_{j=1}^{k-1} w_j}\frac{f_T(t - \sum_{j=1}^{k} w_j)}{f_T(t - \sum_{j=1}^{k-1} w_j)}\nu(dw_k) \tag{223}$$

This completes our derivation for the distribution of the atom masses of G in size-biased order. It has a stick-breaking metaphor as follows: We start by drawing the total mass T from its distribution given ν (with density f_T). This presents a stick of length T. We now iteratively break pieces of the stick off, with the first piece being of length W_1^*, with conditional distribution Eq. (221), and for each $k > 1$, the kth subsequent stick is obtained by breaking a length of W_k^* off the stick, which has reduced length $T - \sum_{j=1}^{k-1} W_j^*$, with W_k^* having conditional distribution Eq. (223).

Returning again to our running example of the normalized generalized gamma process, the only unknown quantity here is the form of the distribution of its total mass $T_{\alpha,\beta,\sigma}$. When

$\beta = 0$, the NGGP reduces to a normalized stable process, and $T_{\alpha,0,\sigma}$ has a positive stable distribution, a well-studied distribution with known characteristic function, efficient simulation algorithms, but no closed analytically tractable density. When $\beta > 0$, the distribution becomes an exponentially-tilted (or Esscher-transformed) positive stable distribution. This was studied by Devroye (2009) who gave an algorithm for efficiently simulating from such distributions. While methods for numerically computing the density of stable distributions exist, they tend to be computationally expensive, so that the stick-breaking construction does not lead directly to practically relevant inference algorithms. The only special cases for which the density is known are when $\sigma = 0$, in which case the NGGP reduces to a DP, and when $\sigma = 1/2$, in which case it reduces to a normalized inverse Gaussian process.

15.3 Back to the Induced Random Partition

We started this section by relating a sequence of iid draws from the normalized random measure G/T, via size-biased sampling, to the stick-breaking construction. Using the results for the stick-breaking construction derived above, we can now return to the setting of iid draws, in particular to giving an explicit generative procedure for simulating the sequence $\theta_1, \theta_2, \ldots$ which does not require simulating the whole random measure G.

We start with simulating the total mass T of G, which has distribution given by f_T. For the first draw θ_1, its marginal distribution is simply given by the base distribution G_0. The value of θ_1 thus generated corresponds to the first size-biased sample ϕ_1^* from G. Further, the distribution of the corresponding mass W_1^* of the atom in G, conditioned on the simulated value of T, is given by Eq. (221).

For the second draw θ_2, this can either take on the value ϕ_1^*, with probability W_1^*/T, or the value of the second size-biased atom ϕ_2^* of G, with probability $(T - W_1^*)/T$. In case of the latter, the atom ϕ_2^* can be simulated from the base distribution G_0, while the corresponding mass can be simulated using Eq. (223) with $k = 2$. For subsequent draws, say θ_n for $n \geq 3$, suppose at this point that there were $k-1$ unique values among $\theta_1, \ldots, \theta_{n-1}$, corresponding to the first $k-1$ size-biased samples $\phi_1^*, \ldots, \phi_{k-1}^*$ from the atoms of G, with corresponding masses $W_1^*, \ldots, W_{k-1}^*$. Then θ_n will take on value ϕ_ℓ^*, for $\ell = 1, \ldots, k-1$, with probability W_ℓ^*/T, or the value of the kth size-biased atom ϕ_k^* of G, with probability $(T - \sum_{\ell=1}^{k-1} W_\ell^*)/T$. In the latter case the atom ϕ_k^* is again simulated from the base distribution G_0 and the corresponding mass simulated using the stick-breaking formula Eq. (223). The above gives an explicit procedure to simulate the sequence $\theta_1, \theta_2, \ldots$ along with the corresponding atoms $\phi_1^*, \phi_2^*, \ldots$ and their masses $W_1^*, W_2^*, \ldots$ in G in size-biased order.

Suppose that once the total mass T and $\theta_1, \ldots, \theta_N$ are all simulated, there are K unique values $\phi_1^*, \ldots, \phi_K^*$ with corresponding masses $W_1^*, \ldots, W_K^*$ in G. Further, let $\boldsymbol{\pi}$ be the random partition of $[N]$ induced by $\theta_1, \ldots, \theta_N$, and $n_1, \ldots, n_K$ be the number of occurrences of each unique value. The joint distribution of all the simulated random variables is then given by multiplying all the probabilities and densities involves in each step of the above

procedure, giving,

$$
\begin{aligned}
&P(T \in dt, \boldsymbol{\pi} = \pi, \phi_k^* \in d\psi_k, W_k^* \in dw_k \text{ for } k = 1, \ldots, K) \\
=&f_T(t) \prod_{k=1}^{K} \frac{t - \sum_{j=1}^{k-1} w_j}{t} \frac{w_k}{t - \sum_{j=1}^{k-1} w - j} \frac{f_T(t - \sum_{j=1}^{k} w_j)}{f_T(t - \sum_{j=1}^{k-1} w_j)} \nu(dw_k) G_0(d\psi_k) \left(\frac{w_k}{t}\right)^{n_k - 1} \\
=&f_T(t - \textstyle\sum_{k=1}^{K} w_k) t^{-N} dt \displaystyle\prod_{k=1}^{K} w_k^{n_k} \nu(dw_k) G_0(d\psi_k)
\end{aligned} \tag{224}
$$

Notice although the generative procedure introduces a particular size-biased order among the unique values (and particularly that the masses $W_1^*, \ldots, W_K^*$ are size-biased), the joint distribution above is in fact invariant to this order. This is down to the fact that the sequence $\theta_1, \ldots, \theta_N$ is exchangeable.

We conclude this section by noting that the above joint distribution, which describes both the induced random partition $\boldsymbol{\pi}$ as well as the masses of the associated atoms and the total mass, is consistent with the EPPF derived in Eq. (202). To see this, introduce an auxiliary variable $U \sim \text{Gamma}(n, T)$ and a change of variable $T' = T - \sum_{k=1}^{K} W_k^*$, which can be interpreted as the total mass among all other atoms except the K associated with $\theta_1, \ldots, \theta_N$. The resulting joint distribution becomes,

$$
\begin{aligned}
&P(T' \in dt', U \in du, \boldsymbol{\pi} = \pi, \phi_k^* \in d\psi_k, W_k^* \in dw_k \text{ for } k = 1, \ldots, K) \\
=&f_T(t') \frac{1}{\Gamma(N)} u^{N-1} e^{-u(t' + \sum_{k=1}^{K} w_k)} \prod_{k=1}^{K} w_k^{n_k} \nu(dw_k) G_0(d\psi_k) \\
=&\frac{1}{\Gamma(N)} u^{N-1} f_T(t') e^{-ut'} \prod_{k=1}^{K} e^{-uw_k} w_k^{n_k} \nu(dw_k) G_0(d\psi_k)
\end{aligned} \tag{225}
$$

Marginalizing out T' and $W_1^*, \ldots, W_K^*$, and using the fact that f_T is the distribution of the total mass of the CRM with Lévy measure ν,

$$
\begin{aligned}
&P(U \in du, \boldsymbol{\pi} = \pi, \phi_k^* \in d\psi_k \text{ for } k = 1, \ldots, K) \\
=&\frac{1}{\Gamma(N)} u^{N-1} \exp\left(-\int 1 - e^{-uw} \nu(dw)\right) \prod_{k=1}^{K} G_0(d\psi_k) \int e^{-uw_k} w_k^{n_k} \nu(dw_k)
\end{aligned} \tag{226}
$$

which exactly agrees with Eq. (196), modulo a few notational differences. Specifically, the number of clusters is $K = |\boldsymbol{\pi}|$, and here the sizes of the clusters are given by $n_1, \ldots, n_K$, with the clusters ordered according to the stick-breaking construction. More precisely, the clusters are ordered in increasing order of their least elements. Note that the first exponential term is the Laplace transform of the total mass of a CRM with Lévy measure ν, and follows from the Lévy-Khinchin formula Eq. (144).

16 Poisson-Kingman Processes

Over the last few sections, we have developed the various useful techniques and representations for working with normalized completely random measures. These form a large

class of random probability measures including the Dirichlet process, the normalized stable process, the normalized inverse Gaussian process and the normalized generalized gamma process. Unfortunately, the Pitman-Yor process, which is the second random probability measure studied in Part I, does not fall into this class. In this section we will elucidate the relationship between these two classes, by introducing a small generalization of normalized random measures. We refer to this class as the Poisson-Kingman processes (PKPs), in view of the study by Pitman (2003) on the exchangeable random partitions induced by random probability measures in this class, which he called the Poisson-Kingman partitions.

Let G be a completely random measure with Lévy measure ν and base distribution G_0, and let $T = G(\Theta)$ be its total mass. For simplicity, suppose that T is positive and finite almost surely, and has a density f_T. The Poisson-Kingman processes are obtained by imposing a different distribution, say γ, over the total mass T. For each positive value $t > 0$ such that $f_T(t) > 0$, denote the conditional distribution of the random measure G, given that its total mass is $T \in dt$, by $\mathrm{PK}(\nu, G_0, \delta_t)$. The general construction of a Poisson-Kingman process $\mathrm{PK}(\nu, G_0, \gamma)$ is then obtained by mixing the total mass over the alternative distribution γ:

$$\begin{aligned} T &\sim \gamma \\ G|T &\sim \mathrm{PK}(\nu, G_0, \delta_T). \end{aligned} \tag{227}$$

In other words, we require that the total mass T has distribution γ, and otherwise G is completely random with Lévy measure ν and base distribution G_0.

Based on Sec. (15), both the stick-breaking construction and the induced random partition of normalized random measures can be extended to the Poisson-Kingman process directly. Consider first the stick-breaking representation, which was derived by first drawing the total mass T from its distribution f_T, then by breaking off pieces of a stick with initial length T. In the case of the Poisson-Kingman process, we simply replace the initial distribution of T by γ. The resulting joint distribution over the total mass T, random partition $\boldsymbol{\pi}$, and associated size-biased draws is, following from Eq. (224),

$$\begin{aligned} &P(T \in dt, \boldsymbol{\pi} = \pi, \phi_k^* \in d\psi_k, W_k^* \in dw_k \text{ for } k = 1, \ldots, K) \\ =&\gamma(dt) \frac{f_T(t - \sum_{k=1}^K w_k)}{f_T(t)} t^{-N} dt \prod_{k=1}^K w_k^{n_k} \nu(dw_k) G_0(d\psi_k), \end{aligned} \tag{228}$$

where all the terms except $\gamma(dt)$ constitute the conditional distribution of the other random variables given $T \in dt$.

A subclass consists of the σ-stable Poisson-Kingman processes, which is obtained when the underlying completely random measure is a σ-stable process with Lévy measure

$$\nu_{\sigma,0,\sigma}(dw) = \frac{\sigma}{\Gamma(1-\sigma)} w^{-\sigma-1} dw. \tag{229}$$

This subclass is particularly mathematical tractable, as the form of the Lévy measure allows for the atom masses W_k^*'s in the above joint distribution to be marginalized out. Define a change of variables with $S = \sum_{k=1}^K W_k^*$ and $V_k = W_k^*/S$ for each $k = 1, \ldots, K-1$. Then,

noting that the Jacobian is S^{K-1} (see Appendix B), we have,

$$\begin{aligned}
&P(T \in dt, S \in ds, \boldsymbol{\pi} = \pi, V_k \in dv_k \text{ for } k = 1, \ldots, K-1) \\
=&\gamma(dt)\frac{f_{\sigma,0,\sigma}(t-s)}{f_{\sigma,0,\sigma}(t)} t^{-N} s^{K-1} dt\, ds \prod_{k=1}^{K} \frac{\sigma}{\Gamma(1-\sigma)} (v_k s)^{n_k - \sigma - 1} dv_k \\
=&\gamma(dt)\frac{f_{\sigma,0,\sigma}(t-s)}{f_{\sigma,0,\sigma}(t)} t^{-N} s^{N-K\sigma-1} dt\, ds \frac{\sigma^K}{\Gamma(1-\sigma)^K} \prod_{k=1}^{K} v_k^{n_k - \sigma - 1} dv_k,
\end{aligned}$$

where we defined $v_K = 1 - \sum_{k=1}^{K-1} v_k$, and $f_{\sigma,0,\sigma}$ is the density of the total mass $T_{\sigma,0,\sigma}$ of $G_{\sigma,0,\sigma}$, a positive σ-stable random variable. The terms containing the v_k's are proportional to the density for a Dirichlet distribution with parameters $(n_1 - \sigma, \ldots, n_K - \sigma)$, so that marginalizing out $V_1, \ldots, V_{K-1}$ gives,

$$P(T \in dt, S \in ds, \boldsymbol{\pi} = \pi) = \frac{\sigma^K}{\Gamma(N - K\sigma)} \gamma(dt) \frac{f_{\sigma,0,\sigma}(t-s)}{f_{\sigma,0,\sigma}(t)} t^{-N} s^{N-K\sigma-1} dt\, ds \prod_{k=1}^{K} \frac{\Gamma(n_k - \sigma)}{\Gamma(1-\sigma)}.$$

Integrating out T and S,

$$P(\boldsymbol{\pi} = \pi) = \frac{\sigma^K}{\Gamma(N - K\sigma)} \int_0^\infty \int_0^t \gamma(dt) \frac{f_{\sigma,0,\sigma}(t-s)}{f_{\sigma,0,\sigma}(t)} t^{-N} s^{N-K\sigma-1} ds\, dt \prod_{k=1}^{K} (1-\sigma)^{(n_k - 1)}, \tag{230}$$

and we see that the EPPF factorizes into a term which depends only upon the number of clusters K, and terms of the form $(1-\sigma)^{(n_k-1)}$, which are the same as for the Pitman-Yor process Eq. (77) and the normalized generalized gamma process Eq. (203). In fact, both are special cases of the σ-stable Poisson-Kingman process for specific choices of the distribution γ over the total mass.

16.1 Back to the Normalized Generalized Gamma Process

For the normalized generalized gamma process, the total mass distribution γ is an exponentially-tilted stable distribution, with density

$$\gamma^{\text{NGGP}}_{\beta,\sigma}(dt) = e^{\beta^\sigma - \beta t} f_{\sigma,0,\sigma}(t) dt, \tag{231}$$

where e^{β^σ} is the normalization constant, as can be observed from the Laplace transform Eq. (175) for a stable random variable: $\int_0^\infty e^{-\beta t} f_{\sigma,0,\sigma}(t) dt = \mathbb{E}[e^{-\beta T_{\sigma,0,\sigma}}] = e^{-\beta^\sigma}$. Some algebra allows us to verify that Eq. (230) reduces to the EPPF of a NGGP Eq. (203),

$$\begin{aligned}
&P(\boldsymbol{\pi} = \pi) \\
=&\frac{\sigma^K}{\Gamma(N - K\sigma)} \int_0^\infty \int_0^t e^{\beta^\sigma - \beta t} f_{\sigma,0,\sigma}(t-s) t^{-N} s^{N-K\sigma-1} ds\, dt \prod_{k=1}^{K} (1-\sigma)^{(n_k - 1)}.
\end{aligned}$$

With a change of variable from t to $t' = t - s$,

$$
\begin{aligned}
=&\frac{\sigma^K e^{\beta^\sigma}}{\Gamma(N-K\sigma)} \int_0^\infty \int_0^\infty e^{-\beta(t'+s)} f_{\sigma,0,\sigma}(t')(t'+s)^{-N} s^{N-K\sigma-1} ds\, dt' \prod_{k=1}^K (1-\sigma)^{(n_k-1)} \\
=&\frac{\sigma^K e^{\beta^\sigma}}{\Gamma(N-K\sigma)} \int_0^\infty \int_0^\infty e^{-\beta(t'+s)} f_{\sigma,0,\sigma}(t') \frac{1}{\Gamma(N)} \int_0^\infty u^{N-1} e^{-(t'+s)u} du\, s^{N-K\sigma-1} ds\, dt' \prod_{k=1}^K (1-\sigma)^{(n_k-1)} \\
=&\frac{\sigma^K e^{\beta^\sigma}}{\Gamma(N-K\sigma)\Gamma(N)} \int_0^\infty u^{N-1} \left(\int_0^\infty e^{-(u+\beta)t'} f_{\sigma,0,\sigma}(t') dt' \right) \left(\int_0^\infty e^{-(u+\beta)s} s^{N-K\sigma-1} ds \right) du \prod_{k=1}^K (1-\sigma)^{(n_k-1)} \\
=&\frac{\sigma^K e^{\beta^\sigma}}{\Gamma(N-K\sigma)\Gamma(N)} \int_0^\infty u^{N-1} e^{-(u+\beta)^\sigma} \Gamma(N-K\sigma)(u+\beta)^{K\sigma-N} du \prod_{k=1}^K (1-\sigma)^{(n_k-1)} \\
=&\frac{\sigma^K}{\Gamma(N)} \int_0^\infty u^{N-1} (u+\beta)^{K\sigma-N} e^{-((u+\beta)^\sigma-\beta^\sigma)} du \prod_{k=1}^K (1-\sigma)^{(n_k-1)}. \qquad (232)
\end{aligned}
$$

We see that the EPPF agrees with that for a NGGP with parameters (σ, β, σ). Noting that an NGGP with parameters (α, β, σ) is equivalent to one with parameters $(\alpha c^\sigma, \beta/c, \sigma)$ obtained by rescaling the underlying CRM by c, for any $c > 0$, we see that any NGGP is equivalent to one with parameters (σ, β, σ), for some σ and β, and in turn to the σ-stable Poisson-Kingman process with $\gamma^{\text{NGGP}}_{\beta,\sigma}$ as its total mass distribution.

16.2 Back to the Pitman-Yor Process

For the Pitman-Yor process, the total mass distribution is given by,

$$\gamma^{\text{PYP}}_{\alpha,\sigma}(dt) = \frac{\sigma\Gamma(\alpha)}{\Gamma(\alpha/\sigma)} t^{-\alpha} f_{\sigma,0,\sigma}(t) dt, \qquad (233)$$

where $\alpha > -\sigma$ is the concentration parameter of the Pitman-Yor process. The distribution $\gamma^{\text{PYP}}_{\alpha,\sigma}$ is a polynomially-tilted positive stable distribution (Devroye, 2009), as it is obtained by tilting the positive stable distribution by a polynomial factor $t^{-\alpha}$. The leading fraction is the normalization constant. This can be seen most easily when the concentration parameter $\alpha > 0$, using the Gamma identity $t^{-\alpha} = \frac{1}{\Gamma(\alpha)} \int_0^\infty u^{\alpha-1} e^{-ut} du$. To also include the case when $-\sigma < \alpha \leq 0$, we make use of the Laplace transform for the stable distribution, which is obtained as a special case of Eq. (175),

$$\mathbb{E}[e^{-uT_{\sigma,0,\sigma}}] = \int_0^\infty e^{-ut} f_{\sigma,0,\sigma}(t) dt = e^{-u^\sigma} \qquad (234)$$

Differentiating both sides with respect to u, we see that,

$$\int_0^\infty t e^{-ut} f_{\sigma,0,\sigma}(t) dt = \sigma u^{\sigma-1} e^{-u^\sigma}. \qquad (235)$$

Now,

$$\begin{aligned}
&\int_0^\infty t^{-\alpha} f_{\sigma,0,\sigma}(t)dt = \int_0^\infty t^{-(\alpha+1)} t f_{\sigma,0,\sigma}(t)dt \\
=&\int_0^\infty \frac{1}{\Gamma(\alpha+1)} \int_0^\infty u^\alpha e^{-ut} du\, t f_{\sigma,0,\sigma}(t)dt && \text{using the Gamma identity,} \\
=&\frac{1}{\Gamma(\alpha+1)} \int_0^\infty u^\alpha \int_0^\infty t e^{-ut} f_{\sigma,0,\sigma}(t)dt\, du \\
=&\frac{1}{\Gamma(\alpha+1)} \int_0^\infty u^\alpha \sigma u^{\sigma-1} e^{-u^\sigma} du && \text{using Eq. (235),} \\
=&\frac{1}{\Gamma(\alpha+1)} \int_0^\infty v^{\alpha/\sigma} e^{-v} dv && \text{using a change of variable } v = u^\sigma, \\
=&\frac{\Gamma(\alpha/\sigma+1)}{\Gamma(\alpha+1)} = \frac{\Gamma(\alpha/\sigma)}{\sigma\Gamma(\alpha)}. && (236)
\end{aligned}$$

To see that this choice of $\gamma^{\text{PYP}}_{\alpha,\sigma}$ leads to the Pitman-Yor process, we again check that the σ-stable Poisson-Kingman EPPF Eq. (230) reduces to that for the Pitman-Yor process:

$$\begin{aligned}
&P(\boldsymbol{\pi} = \pi) \\
=&\frac{\sigma^K}{\Gamma(N-K\sigma)} \int_0^\infty \int_0^t \frac{\sigma\Gamma(\alpha)}{\Gamma(\alpha/\sigma)} t^{-\alpha} f_{\sigma,0,\sigma}(t-s) t^{-N} s^{N-K\sigma-1} ds\, dt \prod_{k=1}^K (1-\sigma)^{(n_k-1)} \\
=&\frac{\sigma^{K+1}\Gamma(\alpha)}{\Gamma(\alpha/\sigma)\Gamma(N-K\sigma)} \int_0^\infty \int_0^t t^{-\alpha-N} f_{\sigma,0,\sigma}(t-s) s^{N-K\sigma-1} ds\, dt \prod_{k=1}^K (1-\sigma)^{(n_k-1)}.
\end{aligned}$$

Again introducing a change of variable from t to $t' = t - s$, and using the Gamma identity,

$$\begin{aligned}
=&\frac{\sigma^{K+1}\Gamma(\alpha)}{\Gamma(\alpha/\sigma)\Gamma(N-K\sigma)} \int_0^\infty \int_0^\infty \int_0^\infty \frac{1}{\Gamma(\alpha+N)} u^{\alpha+N-1} e^{-u(t'+s)} du f_{\sigma,0,\sigma}(t') s^{N-K\sigma-1} ds\, dt' \prod_{k=1}^K (1-\sigma)^{(n_k-1)} \\
=&\frac{\sigma^{K+1}\Gamma(\alpha)}{\Gamma(\alpha/\sigma)\Gamma(N-K\sigma)\Gamma(\alpha+N)} \int_0^\infty u^{\alpha+N-1} \int_0^\infty e^{-ut'} f_{\sigma,0,\sigma}(t') dt' \int_0^\infty e^{-us} s^{N-K\sigma-1} ds du \prod_{k=1}^K (1-\sigma)^{(n_k-1)} \\
=&\frac{\sigma^{K+1}\Gamma(\alpha)}{\Gamma(\alpha/\sigma)\Gamma(N-K\sigma)\Gamma(\alpha+N)} \int_0^\infty u^{\alpha+N-1} e^{-u^\sigma} \Gamma(N-K\sigma) u^{K\sigma-N} du \prod_{k=1}^K (1-\sigma)^{(n_k-1)} \\
=&\frac{\sigma^{K+1}\Gamma(\alpha)}{\Gamma(\alpha/\sigma)\Gamma(\alpha+N)} \int_0^\infty u^{\alpha+K\sigma-1} e^{-u^\sigma} du \prod_{k=1}^K (1-\sigma)^{(n_k-1)}.
\end{aligned}$$

Introducing a change of variable $v = u^\sigma$,

$$
\begin{aligned}
&= \frac{\sigma^{K+1}\Gamma(\alpha)}{\Gamma(\alpha/\sigma)\Gamma(\alpha+N)} \int_0^\infty \frac{1}{\sigma} v^{\alpha/\sigma+K-1} e^{-v} dv \prod_{k=1}^{K} (1-\sigma)^{(n_k-1)} \\
&= \frac{\sigma^{K}\Gamma(\alpha)}{\Gamma(\alpha/\sigma)\Gamma(\alpha+N)} \Gamma(\alpha/\sigma+K) \prod_{k=1}^{K} (1-\sigma)^{(n_k-1)} \\
&= \frac{\alpha(\alpha+\sigma)\cdots(\alpha+\sigma(K-1))}{\alpha^{(N)}} \prod_{k=1}^{K} (1-\sigma)^{(n_k-1)}
\end{aligned} \tag{237}
$$

which is precisely the Pitman-Yor EPPF derived in Eq. (77).

The normalized generalized gamma process and the Pitman-Yor process are special cases of Poisson-Kingman processes where all the integrals above just work out analytically, simplifying the derivations significantly. Generally, they do not simplify in such a fashion. However, armed with the joint distribution Eq. (228) or the EPPF Eq. (230), it is possible to derive tractable inference algorithms and work with the general class of Poisson-Kingman processes.

17 Gibbs-Type Exchangeable Random Partitions

Poisson-Kingman processes can be thought of as extensions of normalized random measures where the total mass is allowed to have any arbitrary marginal distribution γ. One may question why this particular form of extension makes mathematical sense, besides the observation that the Pitman-Yor process is a particular example. In this section, we will discuss a class of exchangeable random partitions that are in a sense natural, and how these lead to a number of natural classes of random probability measures, including the σ-stable Poisson-Kingman processes.

Recall that an exchangeable random partition is a distribution over partitions that is invariant to the ordering of its elements. That is, for each N, the probability of a partition $\pi_{[N]}$ of $[N]$ depends only on the number of clusters and the sizes of the clusters. Mathematically, there is a function p such that the exchangeable partition probability function (EPPF) is,

$$
P(\boldsymbol{\pi}_{[N]} = \pi_{[N]}) = p(K, n_1, \ldots, n_K) \tag{238}
$$

where $K = |\pi_{[N]}|$ is the number of clusters and $n_1, \ldots, n_K$ are the sizes of the clusters in $\pi_{[N]}$. The function p is symmetric in all but the first argument, that is, it is invariant to the ordering of the clusters.

We have seen examples of EPPFs for the random partitions induced by the Dirichlet process Eq. (6), by the Pitman-Yor process Eq. (77), by the normalized generalized gamma process Eq. (203) and most generally for the σ-stable Poisson-Kingman processes Eq. (230). All of them take the form of a product of factors,

$$
p(K, n_1, \ldots, n_K) = V_{NK} \prod_{k=1}^{K} W_{n_k} \tag{239}
$$

where V_{nk} and W_m are both positive functions for positive integers n, k and m. This form for the EPPF is sensible, and reminiscent of Gibbs distributions or exponential families, which define distributions over a collection of random variables as products of factors, each factor being a function of a sufficient statistic of the random variables. We can think of the above as an exponential family distribution over exchangeable partitions, with the sufficient statistics being the number of clusters K and the sizes of the clusters $n_1, \ldots, n_K$ (N is a fixed constant here and it is not necessary to include it as a sufficient statistic).

Given de Finetti's Theorem which gives a direct correspondence between random partitions and random probability measures, a natural question to ask is: What classes of random probability measures will induce exchangeable random partitions whose EPPFs are of the form Eq. (239)? Gnedin and Pitman (2006) gave a complete characterization of the classes of random probability measures inducing Gibbs-type EPPFs. Specifically, they showed that the W_m factors can always be written in the form

$$W_m = (1 - \sigma)^{(m-1)} \tag{240}$$

where $\sigma \in [-\infty, 1]$. When $\sigma = -\infty$, the random partition is degenerate and always consist of a single cluster, while in the other extreme when $\sigma = 1$, the random partition is also degenerate, with all items belonging to their own singleton clusters. In between, when $\sigma \in (-\infty, 1)$, the random partition can be non-trivial, and the corresponding random probability measures break into three distinct classes, each corresponding to a disjoint range of σ:

- $\sigma = 0$. This includes the one-parameter CRP Eq. (6) induced by the DP. It was shown that the class of Gibbs-type exchangeable random partitions corresponds to these, but with a randomized concentration parameter:

$$p(K, n_1, \ldots, n_K) = \int \frac{\alpha^K}{\alpha(\alpha+1)\cdots(\alpha+N-1)} \gamma(d\alpha) \prod_{k=1}^{K} 1^{(n_k-1)}, \tag{241}$$

  where γ is some distribution over $\mathbb{R}^+$ and can be interpreted as a prior for α.

- $0 < \sigma < 1$. This corresponds precisely to the σ-stable Poisson-Kingman processes. From Eq. (230), we already see that these are of Gibbs-type. The converse is true as well: any Gibbs-type exchangeable random partition with index $\sigma \in (0, 1)$ is the induced random partition of some σ-stable Poisson-Kingman process.

- $\sigma < 0$. This corresponds precisely to Bayesian finite mixtures. Let K^* be the number of components and $W_1, \ldots, W_{K^*}$ be the mixing proportions. The finite mixture can be represented succinctly as a random probability measure,

$$G = \sum_{k=1}^{K^*} W_k \delta_{\phi_k^*} \tag{242}$$

  where $\phi_1^*, \ldots, \phi_{K^*}^*$ are the component parameters. Suppose that out of N observations, n_k^* are assigned to component k, for each $k = 1, \ldots, K^*$, and denote these assignments

by z^*. A typical prior for the mixing proportions is a symmetric Dirichlet distribution. If this has parameters $(|\sigma|, \ldots, |\sigma|)$, then the marginal distribution of z^* is,

$$P(z^*|K^*) = \frac{\Gamma(K^*|\sigma|)}{\Gamma(N - K^*\sigma|)} \prod_{k=1}^{K^*} \frac{\Gamma(n_k^* - \sigma)}{\Gamma(|\sigma|)} = \frac{|\sigma|^{K^*}\Gamma(K^*|\sigma|)}{\Gamma(N - K^*\sigma|)} \prod_{k=1}^{K^*} (1-\sigma)^{(n_k^*-1)} \quad (243)$$

For a partition $\pi_{[N]}$ of $[N]$ consisting of K clusters, with sizes $n_1, \ldots, n_K$, the probability of $\pi_{[N]}$ under the finite mixture model is then,

$$P(\boldsymbol{\pi}_{[N]} = \pi_{[N]}|K^*) = \frac{|\sigma|^{K^*}\Gamma(K^*|\sigma|)}{\Gamma(N - K^*\sigma|)} K^*(K^*-1)\cdots(K^*-K+1) \prod_{k=1}^{K^*} (1-\sigma)^{(n_k-1)}. \quad (244)$$

The additional factors are due to the fact that the clusters in $\pi_{[N]}$ are unlabelled, while those in z^* are labelled by numbers in $[K^*]$. As a result, for the partition $\pi_{[N]}$, there are $K^*(K^*-1)\cdots(K^*-K+1)$ corresponding assignments, all the same probability. Note that when $K > K^*$ the resulting probability is 0, which is due to the fact that the finite mixture cannot represent more than K^* clusters. Now if we use a prior γ over the number of components of the finite mixture, we will have a marginal distribution over partitions,

$$P(\boldsymbol{\pi}_{[N]} = \pi_{[N]}) = \sum_{K^*=1}^{\infty} \gamma(K^*) \frac{|\sigma|^{K^*}\Gamma(K^*|\sigma|)}{\Gamma(N - K^*\sigma|)} K^*(K^*-1)\cdots(K^*-K+1) \prod_{k=1}^{K^*} (1-\sigma)^{(n_k-1)}, \quad (245)$$

which is of Gibbs-type with a negative index σ. The converse is true as well: any Gibbs-type exchangeable random partition with index $\sigma < 0$ is the induced random partition of a Bayesian finite mixture.

A discussion of the framework of Gibbs-type exchangeable random partitions and their use in Bayesian nonparametrics is provided by (De Blasi et al., 2015).

Part IV: A Few Final Words

For a more in-depth review of various constructions for random measures based on completely random measures, see Lijoi and Pruenster (2010).

Again mention DDPs and cite some literature.

Cite something by Lancelot.

Theory.

Variational approaches.

Hjort et al book.

Appendix A Some Background on Measure Theory

The purpose of this section is to provide some comfort to readers who may find the following notation mysterious:

$$\int f(x)\mu(dx), \tag{246}$$

where μ is a measure, not to mention the notation:

$$\int f(\omega, p)N(d\omega, dp), \tag{247}$$

where N is a discrete random measure.

Probability theory reposes on measure theory, and any serious user of probability theory will eventually need to master some measure theory; this is particularly true for a user working in an area such as Bayesian nonparametrics in which the principal objects are random measures. That said, in this monograph we have assumed little prior exposure to measure theory on the part of the reader (and we have accordingly skirted several opportunities to make rigorous statements that would require some measure theory). The main exception to this statement is that we have made use of the basic idea that a measure can be used to define an integral.

To give an interpretation to the object $\int f(x)\mu(dx)$, which is known as a *Lebesgue integral*, let us recall the interpretation given to the classical Riemann integral,

$$I = \int f(x)dx, \tag{248}$$

where x ranges over the real line $\mathbb{R}$. Note first that such integrals are generally specified with lower and upper limits:

$$I = \int_a^b f(x)dx, \tag{249}$$

but it can be useful to lighten our notation by writing

$$I = \int_a^b f(x)dx = \int \mathbb{1}_A(x)f(x)dx, \tag{250}$$

where the indicator function $\mathbb{1}_A(x)$ is equal to one if x lies in the interval A, and zero otherwise, and where $A = [a, b]$. We then redefine $f(x)$ to include the factor $\mathbb{1}_A(x)$; this allows us to avoid having to write out lower and upper limits explicitly.

The integral $\int f(x)dx$ is defined in elementary calculus as a limit of sums over partitions of the interval $[a, b]$. For example, consider the partition $(a = x_0, x_1, \ldots, x_{m-1}, x_m = b)$, where the length of each subinterval, $|x_i - x_{i-1}|$, for $i = 1, \ldots, m$, is equal to $1/m$. Evaluating the function $f(x)$ somewhere in each subinterval, for example at the right boundary of the subinterval, we form the finite sum:

$$I_m = \sum_{i=1}^{m} f(x_i)|x_i - x_{i-1}|, \tag{251}$$

and define I as the limit of I_m as m goes to infinity. While such a limit won't exist for arbitrary f, it does exist for certain classes of f, and with attention paid to ensuring that the limit is independent of the choice of evaluation point in the subinterval, as well as the choice of partition, one can obtain a basic theory of integration. (The full development of such theoretical results can consume several months of a young student's life.)

Note that the symbol "dx" doesn't have an interpretation all by itself (in particular, it does not mean a "small interval"). Rather, "dx" has a meaning in the context of the notation $\int f(x)dx$, simply as a reminder that the size of subintervals in the approximating sums I_m is obtained by using the classical notion of length.

Turning to the Lebesgue integral $\int f(x)\mu(dx)$, the change to the notation is that "dx" is replaced with "$\mu(dx)$." Again, the latter symbols do not have an interpretation all by themselves; rather, they are meant to suggest that a different notion of the length is being used in the approximating sums. In particular, let μ be a measure on $\mathbb{R}$; this is a function that assigns a real number to intervals and other subsets of the real line, doing so in a coherent way (e.g., such that the measure of a union of disjoint subsets is the sum of the measures of the subsets). Let $\mu(C)$ denote the measure assigned to the subset C; in particular, $\mu((x_{i-1}, x_i))$ is the measure assigned by μ to the subinterval (x_{i-1}, x_i). This suggests that we define the following approximating sequence,

$$I_m = \sum_{i=1}^{m} f(x_i)\mu((x_{i-1}, x_i)), \tag{252}$$

and take the limit as m goes to infinity as the definition of $\int f(x)\mu(dx)$.

This is indeed the correct intuition. The advanced student of real analysis will know that the actual definition of a Lebesgue integral is more sophisticated than this, relying on a better way of defining partitions and requiring a notion of "measurable function," but such considerations are better left to a full-blown study of real analysis. For our purposes, interpreting $\int f(x)\mu(dx)$ as a limit of sums in which μ is used in place of the classical notion of length will serve just fine. Note, moreover, that the same intuition serves when x is allowed to range over $\mathbb{R}^k$ or over more general spaces. As long as we can measure the size of subsets associated with increasingly fine partitions of sets via a measure μ, we can define integrals over those sets.

One aspect of the generalization from "dx" to "$\mu(dx)$" that is worthy of particular mention in our context is that in Lebesgue integration μ is allowed to contain atoms. Recall that if an atom is located at a point z then we have $\mu(\{z\}) > 0$. The classical measure of length assigns zero length to single points and thus does not allow atoms. Let us consider what happens when we integrate with respect to a measure that is precisely a single atom; i.e., let $\mu = \delta_z$. We have $\mu((x_{i-1}, x_i)) = 0$ for all intervals that do not contain the atom, and $\mu((x_{i^*-1}, x_{i^*})) = 1$ for the single interval that contains the atom. Thus,

$$I_m = \sum_{i=1}^{m} f(x_i)\mu((x_{i-1}, x_i)) = f(x_{i^*}) \approx f(z), \tag{253}$$

where the final approximation is based on an assumption about f (its continuity). Thus as m goes to infinity we expect I_m to approach $f(z)$, yielding:

$$\int f(x)\mu(dx) = f(z). \tag{254}$$

Similarly, if $\mu = \sum_{j=1}^{J} \delta_{z_j}$ is a sum of atoms, we obtain

$$\int f(x)\mu(dx) = \sum_{j=1}^{J} f(z_j). \tag{255}$$

Noting that a sum of atoms is known as a "counting measure," we see that we can express sums as Lebesgue integrals with respect to counting measure. Finally, we can also consider weighted sums of atoms where $\mu = \sum_{j=1}^{J} w_j \delta_{z_j}$, in which case we obtain a weighted sum:

$$\int f(x)\mu(dx) = \sum_{j=1}^{J} w_j f(z_j). \tag{256}$$

A weighted sum of atoms is also known as a "discrete measure."

Putting these ideas together, we can now provide an interpretation of the integral

$$\int f(\omega, p)N(d\omega, dp). \tag{257}$$

This is simply a Lebesgue integral on a space in which the coordinates are (ω, p) and where N is a measure on that space. Moreover, for the case of interest to us in which N is a counting measure, let $\{(\omega_i, p_i)\}$ denote the set of atoms forming N. We have:

$$\int f(\omega, p)N(d\omega, dp) = \sum_{i} f(\omega_i, p_i). \tag{258}$$

Finally, if N is a *random* measure, then the integral with respect to N is a real-valued random quantity; i.e., a random variable. Thus it makes sense to talk about quantities such as

$$\mathbb{E} \int f(\omega, p)N(d\omega, dp), \tag{259}$$

and

$$\mathbb{E}e^{-t \int f(\omega,p)N(d\omega,dp)}, \tag{260}$$

as we have done in Sec. (9).

Appendix B The Dirichlet Distribution

In this appendix, we provide some of the basic definitions and properties associated with the Dirichlet distribution. We intend for the material to be reasonably complete, so that it can be used as a resource not only by the beginning reader, but also by a more advanced reader who is attempting to go beyond the material presented in the main text. The material here does not need to be mastered before reading the main text, but it eventually should be mastered.

The Dirichlet distribution is a distribution on a simplex—a finite-dimensional set of nonnegative numbers that sum to one. In particular, define the simplex S_{K-1} as the set of vectors $v = (v_1, v_2, \ldots, v_K)$ that satisfy $0 < v_i < 1$ and $\sum_{i=1}^{K} v_i = 1$. Note that S_{K-1}

has measure zero when viewed as a subset of $\mathbb{R}^K$; thus, if we wish to define a probability density we will need to view S_{K-1} as embedded in $\mathbb{R}^{K-1}$. Another way to put this is to note that v is a redundant parameterization; only $K-1$ numbers are needed to specify a point in S_{K-1}. Let us in particular parameterize a point in S_{K-1} as $(v_1, v_2, \ldots, v_{K-1})$ and define $v_K := \sum_{i=1}^{K-1} v_i$.

The Dirichlet distribution is an exponential family distribution on the simplex. Its density takes the following form:

$$p(v_1, v_2, \ldots, v_{K-1}) = \frac{\Gamma(\alpha.)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K v_i^{\alpha_i - 1}, \tag{261}$$

where $\alpha_i > 0$ are parameters and where $\alpha. := \sum_{i=1}^K \alpha_i$. This density has its support on the set of points $(v_1, v_2, \ldots, v_{K-1})$ such that $0 < v_i < 1$ and $\sum_{i=1}^{K-1} v_i < 1$. Let us use the symbol Ω_{K-1} to denote this set. Taking the exponential of the logarithm on both sides of Eq. (261), we see that the natural parameters of the Dirichlet distribution are α_i and the sufficient statistics are $\log v_i$.

One often sees the Dirichlet distribution written using the simplified representation:

$$p(v_1, v_2, \ldots, v_K) = \frac{\Gamma(\alpha.)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K v_i^{\alpha_i - 1}. \tag{262}$$

Although this representation is convenient in calculations, it must be kept in mind that Eq. (262) does not define a density; it is merely a shorthand for the density defined in Eq. (261).

To show that Eq. (261) in fact defines a density (i.e., that it integrates to one), we establish a link between the Dirichlet distribution and the gamma distribution.

Representation in terms of normalized gamma random variables

A Dirichlet distribution has a natural representation in terms of a normalized set of independent gamma random variables; indeed, this representation is often used as the definition of the Dirichlet distribution. Many of the properties of the Dirichlet distribution are most easily obtained starting from the gamma representation.

Recall that the gamma density has the following form:

$$p(x \,|\, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \tag{263}$$

where $\Gamma(\alpha) = \int x^{\alpha-1} e^{-x} dx$ defines the gamma function for $\alpha > 0$.

Proposition 2. *Let $Z_i \sim \text{Gamma}(\alpha_i, \beta)$ denote independent gamma random variables for $i = 1, \ldots, K$, where $\alpha_i > 0$ and $\beta > 0$. Let $S = \sum_{i=1}^K Z_i$ and define $V_i = Z_i / S$. Then*

$$(V_1, \ldots, V_K) \perp\!\!\!\perp S \tag{264}$$

and

$$(V_1, \ldots, V_K) \sim \text{Dir}(\alpha_1, \ldots, \alpha_K). \tag{265}$$

Proof. This is an exercise in transformation of variables. Consider the transformation $(Z_1, \ldots, Z_K) \to (V_1, \ldots, V_{K-1}, S)$. The inverse has the form

$$Z_i = SV_i, \qquad i = 1, \ldots, K-1 \tag{266}$$

$$Z_K = S(1 - \sum_{i=1}^{K-1} V_i). \tag{267}$$

This yields the following Jacobian matrix:

$$\frac{\partial(z_1, \ldots, z_K)}{\partial(v_1, \ldots, v_{K-1}, s)} = \begin{bmatrix} s & 0 & \ldots & v_1 \\ 0 & s & \ldots & v_2 \\ 0 & \vdots & \ddots & \vdots \\ -s & -s & \ldots & (1 - \sum_{i=1}^{K-1} v_i) \end{bmatrix}. \tag{268}$$

Recalling that adding multiples of a given row to other rows does not change the determinant of a matrix, add each of the first $K-1$ rows to the final row; this transforms that row into $(0, 0, \ldots, 1)$. The resulting matrix clearly has a determinant of s^{K-1}.

We now use the change of variables theorem to obtain the joint density of $(V_1, \ldots, V_{K-1}, S)$:

$$\begin{aligned} p(v_1, \ldots, v_{K-1}, s) &= p(z_1, \ldots, z_K) s^{K-1} \\ &= \left(\prod_{i=1}^{K-1} \frac{\beta^{\alpha_i}}{\Gamma(\alpha_i)} (sv_i)^{\alpha_i - 1} e^{-s\beta v_i} \right) \left(\frac{\beta^{\alpha_K}}{\Gamma(\alpha_K)} \left(s(1 - \sum_{i=1}^{K-1} v_i) \right)^{\alpha_K - 1} e^{-s\beta(1 - \sum_{i=1}^{K-1} v_i)} \right) s^{K-1} \\ &= \frac{\beta^{\alpha_\cdot}}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \left(\prod_{i=1}^{K-1} v_i^{\alpha_i - 1} \right) \left(1 - \sum_{i=1}^{K-1} v_i \right)^{\alpha_K - 1} s^{\alpha_\cdot - 1} e^{-s\beta} \\ &= \left(\frac{\Gamma(\alpha_\cdot)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \left(\prod_{i=1}^{K-1} v_i^{\alpha_i - 1} \right) \left(1 - \sum_{i=1}^{K-1} v_i \right)^{\alpha_K - 1} \right) \left(\frac{\beta^{\alpha_\cdot}}{\Gamma(\alpha_\cdot)} s^{\alpha_\cdot - 1} e^{-s\beta} \right). \end{aligned} \tag{269}$$

The result is the product of a Dirichlet density and a gamma density, which proves both of the statements in the proposition. □

Moments

The derivation in the previous section establishes that Eq. (261) in fact defines a density; in particular, it shows that the ratio of gamma functions appearing in Eq. (261) is the correct normalization for the Dirichlet density. Let us record this fact:

$$\frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\alpha_\cdot)} = \int_{\Omega_{K-1}} \left(\prod_{i=1}^{K} v_i^{\alpha_i - 1} \right) dv_1 \cdots dv_{K-1}. \tag{270}$$

We can compute moments of the Dirichlet distribution as a straightforward application of Eq. (270). For example, the mean is obtained as follows:

$$\begin{aligned}
\mathbb{E}(V_i) &= \int_{\Omega_{K-1}} v_i \frac{\Gamma(\alpha_.)}{\prod_{i=1}^K \Gamma(\alpha_i)} \left(\prod_{i=1}^K v_i^{\alpha_i - 1} \right) dv_1 \cdots dv_{K-1} \\
&= \frac{\Gamma(\alpha_.)}{\prod_{i=1}^K \Gamma(\alpha_i)} \int_{\Omega_{K-1}} v_i^{\alpha_i} \left(\prod_{k \neq i} v_k^{\alpha_k - 1} \right) dv_1 \cdots dv_{K-1} \\
&= \frac{\Gamma(\alpha_.)}{\prod_{i=1}^K \Gamma(\alpha_i)} \frac{\Gamma(\alpha_i + 1) \prod_{k \neq i} \Gamma(\alpha_i)}{\Gamma(\alpha_. + 1)} \\
&= \frac{\Gamma(\alpha_.)}{\Gamma(\alpha_. + 1)} \frac{\Gamma(\alpha_i + 1)}{\Gamma(\alpha_i)} \\
&= \frac{\alpha_i}{\alpha_.}, \qquad (271)
\end{aligned}$$

where we have used $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$. This result shows that the parameters α_i specify the mean of the Dirichlet up to scaling.

In general, the moments of the Dirichlet distribution take the form of ratios of gamma functions. For example, plugging in V_i^m in place of V_i in the previous derivation, we obtain:

$$\begin{aligned}
\mathbb{E}(V_i^m) &= \frac{\Gamma(\alpha_.)}{\Gamma(\alpha_. + m)} \frac{\Gamma(\alpha_i + m)}{\Gamma(\alpha_i)} \qquad (272) \\
&= \frac{\alpha_i(\alpha_i + 1) \cdots (\alpha_i + m - 1)}{\alpha_.(\alpha_. + 1) \cdots (\alpha_. + m - 1)} \qquad (273) \\
&= \frac{\alpha_i^{(m)}}{\alpha_.^{(m)}}, \qquad (274)
\end{aligned}$$

where $a^{(m)} := a(a + 1) \cdots (a + m - 1)$. In the general case we have:

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=1}^K V_i^{m_i} \right] &= \frac{\Gamma(\alpha_.)}{\Gamma(\alpha_. + \sum_{i=1}^K m_i)} \prod_{i=1}^K \frac{\Gamma(\alpha_i + m_i)}{\Gamma(\alpha_i)} \qquad (275) \\
&= \frac{\prod_{i=1}^K \alpha_i^{(m_i)}}{\alpha_.^{(\sum_{i=1}^K m_i)}}. \qquad (276)
\end{aligned}$$

Finally, we use these formulas to compute the variance of the Dirichlet distribution:

$$\mathsf{Var}(V_i) = \frac{\alpha_i^{[2]}}{\alpha_.^{[2]}} - \left(\frac{\alpha_i}{\alpha_.} \right)^2 = \frac{\mathbb{E}(V_i)(1 - \mathbb{E}(V_i))}{\alpha_. + 1} \qquad (277)$$

as well as the covariance:

$$\mathsf{Cov}(V_i, V_j) = \frac{\alpha_i^{[1]} \alpha_j^{[1]}}{\alpha_.^{[2]}} - \frac{\alpha_i \alpha_j}{\alpha_.^2} = -\frac{\mathbb{E}(V_i)\mathbb{E}(V_j)}{\alpha_. + 1}. \qquad (278)$$

These formulas show that $\alpha_.$ plays the role of a concentration parameter; an increase in $\alpha_.$ leads to a decrease in variance. It is also worth noting that the covariances are negative. It is not the case in all problems involving proportions that pairs of proportions should negatively co-vary, and if they do not, a Dirichlet assumption should be questioned.

Further properties

In this section we collect together several additional properties of the Dirichlet distribution.

Let α and γ denote nonnegative vectors with K components: $\alpha = (\alpha_1, \ldots, \alpha_K)$ and $\gamma = (\gamma_1, \ldots, \gamma_K)$. Note also the following notational convention: when "Dir(α)" appears in an equation, it should be viewed as a shorthand for a random variable having the Dir(α) distribution.

The Dirichlet distribution satisfies an important aggregation property: Summing across subsets of Dirichlet variables yields a Dirichlet distribution in which the parameters are sums across the corresponding subsets of parameters. The simplest case is the following:

Proposition 3. *Let $V \sim \mathrm{Dir}(\alpha)$. Then*

$$(V_1, \ldots, V_i + V_{i+1}, \ldots, V_K) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_i + \alpha_{i+1}, \ldots, \alpha_K). \tag{279}$$

Proof. Let $V_i = Z_i / \sum_{j=1}^K Z_j$ where Z_i are independent Gamma(α_i, 1) variables. Now $Z_i + Z_{i+1}$ is distributed as Gamma($\alpha_i + \alpha_{i+1}$, 1). Normalizing by $\sum_{j=1}^K Z_j$ yields the result. □

A straightforward induction yields the following general result:

Proposition 4. *Let $V \sim \mathrm{Dir}(\alpha)$. Let $(B_1, B_2, \ldots, B_r)$ be a partition of the indices $(1, 2, \ldots, K)$. Then*

$$\left(\sum_{i \in B_1} V_i, \sum_{i \in B_2} V_i, \ldots \sum_{i \in B_r} V_i \right) \sim \mathrm{Dir}\left(\sum_{i \in B_1} \alpha_i, \sum_{i \in B_2} \alpha_i, \ldots \sum_{i \in B_r} \alpha_i \right). \tag{280}$$

The next proposition shows that the Dirichlet distribution can be decomposed into a random convex combination of Dirichlet distributions.

Proposition 5. *Define independent random variables $U \sim \mathrm{Dir}(\alpha)$ and $V \sim \mathrm{Dir}(\gamma)$. Let $W \sim \mathrm{Beta}(\sum_i \alpha_i, \sum_i \gamma_i)$, independently of V and W. Then*

$$WU + (1 - W)V \sim \mathrm{Dir}(\alpha + \gamma). \tag{281}$$

Proof. Let $\{Z_i\}_{i=1}^K$ be independent Gamma(α_i, 1) variables and let $\{Z_{K+i}\}_{i=1}^K$ be independent Gamma(γ_i, 1) variables. Then

$$\frac{\sum_{i=1}^K Z_i}{\sum_{i=1}^{2K} Z_i} \left(\frac{Z_1}{\sum_{i=1}^K Z_i}, \ldots, \frac{Z_K}{\sum_{i=1}^K Z_i} \right) + \frac{\sum_{i=K+1}^{2K} Z_i}{\sum_{i=1}^{2K} Z_i} \left(\frac{Z_{K+1}}{\sum_{i=K+1}^{2K} Z_i}, \ldots, \frac{Z_{2K}}{\sum_{i=K+1}^{2K} Z_i} \right) \tag{282}$$

has the same distribution as $WU + (1 - W)V$. But this expression is equal to

$$\left(\frac{Z_1 + Z_{K+1}}{\sum_{i=1}^{2K} Z_i}, \ldots, \frac{Z_K + Z_{2K}}{\sum_{i=1}^{2K} Z_i} \right), \tag{283}$$

which has a Dir($\alpha + \gamma$) distribution. □

In the following proposition we see that a Dirichlet distribution can be expressed as a mixture of Dirichlet distributions.

Proposition 6. *Let e_j denote a unit basis vector. Let $\beta_j = \gamma_j / \sum_i \gamma_i$. Then*

$$\sum_j \beta_j \operatorname{Dir}(\gamma + e_j) = \operatorname{Dir}(\gamma). \tag{284}$$

Proof. The equality in this proposition is to be understood as equality of distributions of random variables. The proof is a simple consequence of the general theorem $\mathbb{E}[P(U \mid X)] = P(U)$. In particular, define random variables U and X such that

$$U \sim \operatorname{Dir}(\gamma) \tag{285}$$
$$X \mid U = \operatorname{Discrete}(U). \tag{286}$$

From Dirichlet-multinomial conjugacy (see Sec. (**??**)) we have

$$U \mid X = x \sim \operatorname{Dir}(\gamma + x). \tag{287}$$

Moreover, from Eq. (271) the marginal of X is given by:

$$\mathbb{E}(X^j) = \mathbb{E}\mathbb{E}(X^j \mid U) = \mathbb{E}(U_j) = \frac{\gamma_j}{\sum_i \gamma_i} = \beta_j. \tag{288}$$

Thus $\mathbb{E}[P(U \mid X)] = P(U)$ implies

$$\sum_j \beta_j \operatorname{Dir}(\gamma + e_j) = \operatorname{Dir}(\gamma). \tag{289}$$

□

Finally, the next proposition shows that subsets of entries of Dirichlet-distributed random vectors, once normalized, are still Dirichlet.

Proposition 7. *Let $(V_1, \ldots, V_K) \sim \operatorname{Dir}(\alpha_1, \ldots, \alpha_K)$ and $1 \leq L \leq K$. Let $U = \sum_{\ell=L}^K V_\ell$ and $W = (V_L, \ldots, V_K)/U$. Then $V' = (V_1, \ldots, V_{L-1}, U)$ and W are independent with distributions*

$$V' \sim \operatorname{Dir}(\alpha_1, \ldots, \alpha_{L-1}, \textstyle\sum_{\ell=L}^K \alpha_\ell) \tag{290}$$
$$W \sim \operatorname{Dir}(\alpha_L, \ldots, \alpha_K) \tag{291}$$

Proof. Let $V_i = Z_i / \sum_{j=1}^K Z_j$ where $Z_i \sim \operatorname{Gamma}(\alpha_i, 1)$ are independent. Then $U = \sum_{\ell=L}^K Z_\ell / \sum_{j=1}^K Z_j$ where $\sum_{\ell=L}^K Z_\ell \sim \operatorname{Gamma}(\sum_{\ell=L}^K \alpha_\ell, 1)$, so both V' and W are expressed as a normalized vectors of independent Gamma random variables and Eq. (290) and Eq. (291) hold. To see that V' and W are independent, note from Proposition 2 that W and U are independent. Since W and $(Z_1, \ldots, Z_{L-1})$ are independent as well, and V' is a function of $(Z_1, \ldots, Z_{L-1})$ and U, W and V' are independent. □

Appendix C The Sethuraman Proof

Recall the definition of a stick-breaking random measure:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}, \tag{292}$$

where the atoms ϕ_k are independent draws from G_0 and the weights π_k are drawn according to a stick-breaking process:

$$\begin{aligned} \pi_1 &= \beta_1 \\ \pi_k &= \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) \qquad k = 2, 3, \ldots, \end{aligned} \tag{293}$$

where $\beta_k \sim \text{Beta}(1, \alpha_0)$.

Consider a partition $(B_1, \ldots, B_r)$ of Θ. Let V denote the random vector obtained when G is evaluated on this partition:

$$V = \left(\sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}(B_1) \ldots, \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}(B_r) \right) \tag{294}$$

We would like to show that V has a finite-dimensional Dirichlet distribution.

We make use of the particular form of the stick-breaking representation. Breaking off the first piece of the stick, we have:

$$G = \pi_1 \delta_{\phi_1} + \sum_{k=2}^{\infty} \pi_k \delta_{\phi_k} \tag{295}$$

$$G = \pi_1 \delta_{\phi_1} + (1 - \pi_1) \sum_{k=2}^{\infty} \pi'_k \delta_{\phi_k}, \tag{296}$$

where

$$\begin{aligned} \pi'_2 &= \beta_2 \\ \pi'_k &= \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) \qquad k = 3, 4, \ldots. \end{aligned} \tag{297}$$

But we can simply reindex the sum on the right-hand side of Eq. (296) to run from 1 to infinity. We see from Eq. (297), and the fact that the $\{\phi_k\}$ are iid draws from G_0, that this sum has the same distribution as G itself. That is, we have:

$$G \sim \pi_1 \delta_{\phi_1} + (1 - \pi_1) G. \tag{298}$$

In words, a stick-breaking random measure can be represented as a random convex sum of two random measures: a delta function at a random location and a stick-breaking random measure.

To turn this into a finite-dimensional assertion, we evaluate G on the partition $(B_1, \ldots, B_r)$. This yields:

$$V \sim \pi_1 X + (1 - \pi_1) V; \tag{299}$$

where $X = (\delta_{\phi_1}(B_1), \ldots, \delta_{\phi_1}(B_r))$ is a random vector which ranges over the unit basis vectors $\{e_j\}$. That is, we obtain a distributional equation for the finite-dimensional random vector V that parallels Eq. (298). We claim that the Dirichlet distribution satisfies this distributional equation.

To see this, assume that V has the Dirichlet distribution $\mathrm{Dir}(\alpha)$, where we define $\alpha := (\alpha_0 G_0(B_1), \ldots, \alpha_0 G_0(B_r))$. Condition on the event $\{X = e_j\}$. For fixed X, the equation $\pi_1 X + (1 - \pi_1) V$ reduces to a random convex combination of a constant and a Dirichlet random variable. But the constant e_j can be viewed as a draw from a Dirichlet distribution that has a parameter equal to e_j (and thus places all of its mass at e_j). We are thus in the setting of Prop. (6), where a random variable is a random convex combination of a pair of Dirichlet random variables. In particular, as required by the proposition, π_1 has a $\mathrm{Beta}(1, \alpha_0)$ distribution. The proposition therefore yields:

$$V \,|\, \{X = e_j\} \sim \mathrm{Dir}(\alpha + e_j). \tag{300}$$

Finally, take an expectation with respect to X. The event $\{X = e_j\}$ has probability $G_0(B_j)$, which is equal to $\alpha_j / \sum_{k=1}^{r} \alpha_k$. We can thus appeal to Prop. (5), which implies

$$\pi_1 X + (1 - \pi_1) V \sim \mathrm{Dir}(\alpha). \tag{301}$$

This shows that the assumption $V \sim \mathrm{Dir}(\alpha)$ is consistent with Eq. (299).

It turns out that it is also possible to show that the distributional equation Eq. (299) has a unique solution (Sethuraman, 1994). Thus, by showing that the equation is satisfied by the Dirichlet distribution, we have in fact obtained that unique solution.

Appendix D Tail-free Property of the Dirichlet Process

In this appendix we show that the Dirichlet process has the tail-free property. The setup is that of Sec. (**??**), in which we draw $G \sim \mathrm{DP}(\alpha, G_0)$ and then draw $\theta_i \overset{iid}{\sim} G$ for $i = 1, \ldots, N$. We will only consider the case in which we have one observation θ_1; the general case follows straightforwardly given that $\{\theta_i\}$ are mutually independent given G. To lighten the notation, define $\tilde{G} = \alpha G_0$.

Let $(A_1, A_2, \ldots, A_K)$ be a partition of Θ. Fix $j \in \{1, \ldots, K\}$ and consider a subset B such that $B \subset A_j$. We are interested in the distribution of the random vector $(G(A_1), \ldots, G(A_K))$ conditioned on $\theta_1 \in B$ or on $\theta_1 \in A_j$. The tail-free property of the DP states that $(G(A_1), \ldots, G(A_K))$ depends only on whether θ_1 is in A_j, and not on whether it is in B. Since $B \subset A_j$ is arbitrary, this shows that the actual location of θ_1 in A_j does not affect $(G(A_1), \ldots, G(A_K))$.

To show the tail-free property, we consider the joint distribution of $\theta_1 \in B$ and of $(G(A_1), \ldots, G(A_K))$ in terms of the finer partition $(B, A_1, \ldots, A_j \backslash B, \ldots, A_K)$ where A_j is split into B and $A_j \backslash B$. Let $(y_1, \ldots, y_{K-1})$ be a vector of non-negative numbers with $\sum_{k=1}^{K-1} y_k \leq 1$, and define $y_K = 1 - \sum_{k=1}^{K-1} y_k$. Let x be such that $0 \leq x \leq y_j$. Using the

fact that $(G(B), G(A_1), \ldots, G(A_j \backslash B), \ldots, G(A_K))$ is Dirichlet distributed with parameters $(\tilde{G}(B), \tilde{G}(A_1), \ldots, \tilde{G}(A_j \backslash B), \ldots, \tilde{G}(A_K))$, the joint distribution can be:

$$\begin{aligned}
&P(\theta_1 \in B, G(B) \in dx, G(A_1) \in dy_1, \ldots, G(A_{K-1}) \in dy_{K-1}) \\
=&P(\theta_1 \in B, G(B) \in dx, G(A_1) \in dy_1, \ldots, G(A_j \backslash B) \in d(y_j - x), \ldots, G(A_{K-1}) \in dy_{K-1}) \\
=&x \frac{\Gamma(\alpha)}{\Gamma(\tilde{G}(B))\Gamma(\tilde{G}(A_j \backslash B)) \prod_{k \neq j} \Gamma(\tilde{G}(A_k))} x^{\tilde{G}(B)-1} (y_j - x)^{\tilde{G}(A_1 \backslash B)-1} \prod_{k:1 \leq k \leq K, k \neq j} y_k^{\tilde{G}(A_k)-1} dx \prod_{k=1}^{K-1} dy_k
\end{aligned} \tag{302}$$

Integrating over x,

$$\begin{aligned}
&P(\theta_1 \in B, G(A_1) \in dy_1, \ldots, G(A_{K-1}) \in dy_{K-1}) \\
=&\int_0^{y_j} dx \left(\frac{\Gamma(\alpha)}{\Gamma(\tilde{G}(B))\Gamma(\tilde{G}(A_j \backslash B)) \prod_{k \neq j} \Gamma(\tilde{G}(A_k))} x^{\tilde{G}(B)} (y_j - x)^{\tilde{G}(A_1 \backslash B)-1} \prod_{k \neq j} y_k^{\tilde{G}(A_k)-1} \prod_{k=1}^{K-1} dy_k \right) \\
=&\frac{\Gamma(\alpha)}{\Gamma(\tilde{G}(B))\Gamma(\tilde{G}(A_j \backslash B)) \prod_{k \neq j} \Gamma(\tilde{G}(A_k))} \prod_{k \neq j} y_k^{\tilde{G}(A_j)-1} \prod_{k=1}^{K-1} dy_k \int_0^1 (z y_j)^{\tilde{G}(B)} (y_j (1-z))^{\tilde{G}(A_j \backslash B)-1} y_j dz \\
=&\frac{\Gamma(\alpha)}{\Gamma(\tilde{G}(B))\Gamma(\tilde{G}(A_j \backslash B)) \prod_{k \neq j} \Gamma(\tilde{G}(A_k))} \prod_{k:1 \leq k \leq K, k \neq j} y_k^{\tilde{G}(A_k)-1} y_j^{\tilde{G}(A_j)} \prod_{k=1}^{K-1} dy_k \frac{\Gamma(\tilde{G}(B)+1)\Gamma(\tilde{G}(A_j \backslash B))}{\Gamma(\tilde{G}(B)+1+\tilde{G}(A_j \backslash B))} \\
=&\frac{\Gamma(\alpha)}{\prod_{k=1}^{K} \Gamma(\tilde{G}(A_k))} \prod_{k=1}^{K} y_k^{\tilde{G}(A_k)-1} \prod_{k=1}^{K-1} dy_k \times y_j \frac{H(B)}{H(A_j)}
\end{aligned} \tag{303}$$

The first term in the resultant formula can be read as the density at $(y_1, \ldots, y_K)$ for the Dirichlet distribution of $(G(A_1), \ldots, G(A_K))$, while the second term is the probability of θ_1 being in subset A_j, $G(A_j) = y_j$, followed by the probability $\frac{H(B)}{H(A_j)}$ of θ_1 being in B conditioned on it being in A_j already. In particular, this last probability does not depend on $(G(A_1), \ldots, G(A_K))$. Dividing by $P(\theta_1 \in B) = H(B)$, a further conclusion is that

$$\begin{aligned}
&P(G(A_1) \in dy_1, \ldots, G(A_{K-1}) \in dy_{K-1} | \theta_1 \in B) \\
=&P(G(A_1) \in dy_1, \ldots, G(A_{K-1}) \in dy_{K-1} | \theta_1 \in A_j)
\end{aligned} \tag{304}$$

This demonstrates that the Dirichlet process is tail-free.

Appendix E Some Laplace Transforms

The Laplace transform provides a useful tool for establishing distributional results. By showing that the Laplace transform of a certain distribution has a known form, one identifies that distribution.

In this appendix we derive the Laplace transforms of the Poisson and gamma distributions. These are elementary results that can be found in introductory books on probability theory, but we place them here to encourage the reader to be thoroughly familiar with the Poisson and gamma distributions in approaching the main text.

The Poisson distribution

The probability mass function of a Poisson random variable X with parameter λ is:

$$p(X = k \,|\, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{305}$$

for $k = 0, 1, \ldots$. Thus the Laplace transform can be computed as follows:

$$\begin{aligned} \mathbb{E}[e^{-tX}] &= \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} e^{-tk} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^{-t})^k}{k!} \\ &= e^{-\lambda} e^{(\lambda e^{-t})} \\ &= \exp\{-\lambda(1 - e^{-t})\}. \end{aligned} \tag{306}$$

As a simple application of this result, the reader can verify that $\mathbb{E}[X] = \lambda$ for the Poisson distribution by taking the negative of the first derivative of the Laplace transform and setting $t = 0$.

The gamma distribution

The density of a gamma random variable X with shape parameter α and scale parameter β is:

$$p(x \,|\, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \tag{307}$$

for $x > 0$. We compute the Laplace transform as follows:

$$\begin{aligned} \mathbb{E}[e^{-tX}] &= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} e^{-tx} dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\beta+t)x} dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha)}{(\beta + t)^\alpha} \\ &= \left(\frac{\beta}{\beta + t}\right)^\alpha. \end{aligned} \tag{308}$$

Taking first and second derivatives and setting $t = 0$ yields the mean and variance of a gamma random variable:

$$\mathbb{E}[X] = \frac{\alpha}{\beta} \qquad \mathsf{Var}[X] = \frac{\alpha}{\beta^2}. \tag{309}$$

Appendix F Zipf's Law and the PY Process

As we have discussed in Sec. (**??**), when $0 < \sigma < 1$ the Pitman-Yor process yields power-law behavior. In this appendix we consider the power law known as Zipf's law, and give

a first-order derivation of this law. The claim is that for large N, the proportion of tables with m customers scales as $\mathcal{O}(m^{-1-\sigma})$.

To derive this law, consider a PYCRP with parameters α and σ, with $N+1$ customers. We will tag one of the customers, say the one indexed by $N+1$. Denote the partition among the other N customers by $\boldsymbol{\pi}$, and the table in $\boldsymbol{\pi}$ that customer $N+1$ sits at by Z (if the tagged customer sits by himself we let $Z = \emptyset$). We call Z the tagged table. Together, $\boldsymbol{\pi}$ and Z fully describe the partition of $[N+1]$. Let $K_{N,m}$ denote the number of tables in $\boldsymbol{\pi}$ with size m and let K_N denote the total number of tables.

Define a Markov chain over partitions of the $N+1$ customers, represented by $\boldsymbol{\pi}$ and Z, given by a random-scan Gibbs sampler: At each iteration one of the customers is chosen uniformly at random, and this customer is reseated in the restaurant according to the PYCRP predictive probabilities. This Markov chain is reversible, particularly it satisfies detailed balance. Further, it has stationary distribution given by a PYCRP over $\boldsymbol{\pi}$ and $Z \,|\, \boldsymbol{\pi}$ given by the PYCRP predictive probabilities, by making use of exchangeability and considering the tagged customer as the last one to enter the restaurant.

Consider a pair of states $(\boldsymbol{\pi}, Z)$ and $(\boldsymbol{\pi}', Z')$, with $|Z| = m+1$, $m \geq 1$, and where $(\boldsymbol{\pi}', Z')$ is obtained by a customer sitting at the tagged table other than the tagged customer being reseated elsewhere. Note that $|Z'| = m$. The detailed balance equation gives:

$$P(\boldsymbol{\pi}, Z)P((\boldsymbol{\pi}, Z) \to (\boldsymbol{\pi}', Z')) = P(\boldsymbol{\pi}', Z')P((\boldsymbol{\pi}', Z') \to (\boldsymbol{\pi}, Z)) \tag{310}$$

$$P(\boldsymbol{\pi})\frac{m+1-\sigma}{\alpha+N}\frac{1}{N+1}\left(1-\frac{m+1-\sigma}{\alpha+N}\right) = P(\boldsymbol{\pi}')\frac{m-\sigma}{\alpha+N}\frac{1}{N+1}\frac{m+1-\sigma}{\alpha+N}. \tag{311}$$

Summing over all $(\boldsymbol{\pi}, Z)$ and $(\boldsymbol{\pi}', Z')$ satisfying the description above, we get:

$$\mathbb{E}_{\boldsymbol{\pi}}\left[\frac{m+1-\sigma}{\alpha+N}\frac{m+1}{N+1}\left(1-\frac{m+1-\sigma}{\alpha+N}\right)K_{N,m+1}\right] = \mathbb{E}_{\boldsymbol{\pi}'}\left[\frac{m-\sigma}{\alpha+N}\frac{N-m}{N+1}\frac{m+1-\sigma}{\alpha+N}K'_{N,m}\right] \tag{312}$$

The additional term of $(m+1)K_{N,m+1}$ on the left-hand side arises because there are $K_{N,m+1}$ tables at which the tagged customer could have sat, and there are $m+1$ other customers at the table which could have been reseated elsewhere. Similarly, on the right-hand sid3 we have $K'_{N,m}$ choices for the tagged table, and $N-m$ customers from other tables that could be reseated at the tagged table. Simplifying, we get:

$$\mathbb{E}[K_{N,m+1}] = \frac{N-m}{\alpha+N-m-1+\sigma}\frac{m-\sigma}{m+1}\mathbb{E}[K_{N,m}] \asymp \frac{m-\sigma}{m+1}\mathbb{E}[K_{N,m}], \tag{313}$$

where the asymptotic expression is obtained for $N \gg m \geq 1$.

To solve the recurrence in Eq. (313), we have to solve for the boundary condition $\mathbb{E}[K_{N,1}]$. Consider again a pair of states $(\boldsymbol{\pi}, Z)$ and $(\boldsymbol{\pi}', Z')$, but now with $|Z| = 1$ and $|Z'| = 0$. Now there are two possibilities which cannot be distinguished based on $(\boldsymbol{\pi}, Z)$ and $(\boldsymbol{\pi}', Z')$: either the tagged customer was chosen to be reseated at his own table, or the only other customer at the tagged table was chosen to be reseated elsewhere. The detailed balance equation

now becomes:

$$P(\boldsymbol{\pi})\frac{1-\sigma}{\alpha+N}\left(\frac{1}{N+1}\frac{\alpha+\sigma K_N}{\alpha+N}+\frac{1}{N+1}\left(1-\frac{1-\sigma}{\alpha+N}\right)\right)$$
$$=P(\boldsymbol{\pi}')\frac{\alpha+\sigma K'_N}{\alpha+N}\left(\frac{1}{N+1}\frac{1-\sigma}{\alpha+N}+\frac{1}{N+1}\frac{1-\sigma}{\alpha+N}\right). \tag{314}$$

Again, summing over all $(\boldsymbol{\pi}, Z)$ and $(\boldsymbol{\pi}', Z')$ satisfying the description,

$$\mathbb{E}_{\boldsymbol{\pi}}\left[\frac{1-\sigma}{\alpha+N}K_{N,1}\left(\frac{1}{N+1}\frac{\alpha+\sigma K_N}{\alpha+N}+\frac{1}{N+1}\left(1-\frac{1-\sigma}{\alpha+N}\right)\right)\right] \tag{315}$$
$$=\mathbb{E}_{\boldsymbol{\pi}'}\left[\frac{\alpha+\sigma K'_N}{\alpha+N}\left(\frac{1}{N+1}\frac{1-\sigma}{\alpha+N}K'_{N,1}+\frac{N}{N+1}\frac{1-\sigma}{\alpha+N}\right)\right]. \tag{316}$$

Cancelling terms and simplifying, we get:

$$\mathbb{E}[K_{N,1}]=\frac{N}{\alpha+\sigma+N-1}(\alpha+\sigma\mathbb{E}[K_N]). \tag{317}$$

For large N, we see that

$$\mathbb{E}[K_{N,1}]\asymp\sigma\mathbb{E}[K_N]. \tag{318}$$

Now expanding the recurrence in Eq. (313), we have:

$$\mathbb{E}[K_{N,m}]\asymp\frac{\sigma}{m!}\frac{\Gamma(m-\sigma)}{\Gamma(1-\sigma)}\mathbb{E}[K_N] \qquad \text{for } m\geq 1. \tag{319}$$

We can also verify that $\sum_{m\geq 1}\mathbb{E}[K_{N,m}]$ is indeed equal to $\mathbb{E}[K_N]$ by noting that

$$\sum_{m\geq 1}\frac{\sigma}{m!}\frac{\Gamma(m-\sigma)}{\Gamma(1-\sigma)}=1 \tag{320}$$

using the Taylor expansion of the function $-(1-x)^{\sigma}$ about $x_0=0$ evaluated at $x=1$. Finally, applying Stirling's formula, assuming $N\gg m\gg 1$,

$$\mathbb{E}[K_{N,m}]\asymp\frac{\sigma}{\Gamma(1-\sigma)}m^{-1-\sigma}\mathbb{E}[K_N]. \tag{321}$$

References

Airoldi, E. M., Blei, D. M., Xing, E. P., and Fienberg, S. E. (2006). Mixed membership stochastic block models for relational data with application to protein-protein interactions. In *Proceedings of the International Biometrics Society Annual Meeting*.

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2002). The infinite hidden Markov model. In *Advances in Neural Information Processing Systems*, volume 14.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, 1:353–355.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bush, C. A. and MacEachern, S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika*, 83:275–285.

Chu, W., Ghahramani, Z., Krause, R., and Wild, D. L. (2006). Identifying protein complexes in high-throughput protein interaction screens using an infinite latent feature model. In *BIOCOMPUTING: Proceedings of the Pacific Symposium*.

Cohn, T., Blunsom, P., and Goldwater, S. (2010). Inducing tree-substitution grammars. *The Journal of Machine Learning Research*, 9999:3053–3096.

De Blasi, P., Favaro, S., Lijoi, A., Prüenster, I., and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence Special Issue on Bayesian Nonparametrics*.

de Finetti, B. (1931). Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Academia Nazionale dei Lincei, Serie 6. Memorie, Classe di Scienze Fisiche, Mathematice e Naturale*, 4.

Devroye, L. (2009). Random variate generation for exponentially and polynomially tilted stable distributions. *ACM Transactions Modeling and Computer Simulation*, 19:18:1–18:20.

Erosheva, E. (2003). Bayesian estimation of the grade of membership model. In *Bayesian Statistics*, volume 7, pages 501–510.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89:268–277.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588.

Fall, M. D., Barat, É., Mohammad-Djafari, A., and Comtat, C. (2009). Spatial emission tomography reconstruction using Pitman-Yor process. In *BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS IN SCIENCE AND ENGINEERING: The 29th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*. AIP Publishing.

Favaro, S. and Teh, Y. W. (2013). MCMC for normalized random measure mixture models. *Statistical Science*, 28(3):335–359.

Fei-Fei, L. and Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230.

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2011). A Sticky HDP-HMM with Application to Speaker Diarization. *Annals of Applied Statistics*, 5(2A):1020–1056.

Getoor, L. and Taskar, B., editors (2007). *Introduction to Statistical Relational Learning*. MIT Press.

Ghahramani, Z., Griffiths, T. L., and Sollich, P. (2007). Bayesian nonparametric latent feature models (with discussion and rejoinder). In *Bayesian Statistics*, volume 8.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall.

Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, 138(3):5674–5684.

Goldwater, S., Griffiths, T., and Johnson, M. (2006a). Interpolating between types and tokens by estimating power-law generators. In *Advances in Neural Information Processing Systems*, volume 18.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2006b). Contextual dependencies in unsupervised word segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*.

Görür, D., Jäkel, F., and Rasmussen, C. E. (2006). A choice model with infinitely many latent features. In *Proceedings of the International Conference on Machine Learning*, volume 23.

Griffin, J. E. and Walker, S. G. (2011). Posterior simulation of normalized random measure mixtures. *Journal of Computational and Graphical Statistics*, 20(1):241–259.

Griffiths, T. L. and Ghahramani, Z. (2006). Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems*, volume 18.

Heaps, H. S. (1978). *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, New York.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Annals of Statistics*, 18(3):1259–1294.

Ho, M. W., James, L. F., and Lau, J. W. (2006). Coagulation fragmentation laws induced by general coagulations of two-parameter Poisson-Dirichlet processes. http://arxiv.org/abs/math.PR/0601608.

Hoppe, F. M. (1984). Pólya-like urns and the Ewing sampling formula. *Journal of Mathematical Biology*, 20:91–94.

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

Jeffreys, W. and Berger, J. (1992). Ockham's razor and bayesian analysis. *American Statistician*, 80:64–72.

Johnson, N. L. and Kotz, S. (1977). *Urn Models and Their Application*. John Wiley, New York.

Kallenberg, O. (2002). *Foundations of Modern Probability*. Springer, New York.

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

Knowles, D. and Ghahramani, Z. (2007). Infinite sparse factor analysis and infinite independent components analysis. In *International Conference on Independent Component Analysis and Signal Separation*, volume 7 of *Lecture Notes in Computer Science*. Springer.

Lijoi, A. and Pruenster, I. (2010). Models beyond the Dirichlet process. In Hjort, N., Holmes, C., Müller, P., and Walker, S., editors, *Bayesian Nonparametrics*, pages 80–136. Cambridge University Press.

MacEachern, S., Kottas, A., and Gelfand, A. (2001). Spatial nonparametric Bayesian models. Technical Report 01-10, Institute of Statistics and Decision Sciences, Duke University. http://ftp.isds.duke.edu/WorkingPapers/01-10.html.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation*, 23:727–741.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. University of California Press.

Meeds, E., Ghahramani, Z., Neal, R. M., and Roweis, S. T. (2007). Modeling dyadic data with binary latent factors. In *Advances in Neural Information Processing Systems*, volume 19.

Navarro, D. J. and Griffiths, T. L. (2007). A nonparametric Bayesian method for inferring features from similarity judgements. In *Advances in Neural Information Processing Systems*, volume 19.

Neal, R. M. (1992). Bayesian mixture modeling. In *Proceedings of the Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, volume 11, pages 197–211.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265.

Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351.

Perman, M., Pitman, J., and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, 92(1):21–39.

Pitman, J. (2002). Combinatorial stochastic processes. Technical Report 621, Department of Statistics, University of California at Berkeley. Lecture notes for St. Flour Summer School.

Pitman, J. (2003). Poisson-Kingman partitions. In Goldstein, D. R., editor, *Statistics and Science: a Festschrift for Terry Speed*, pages 1–34. Institute of Mathematical Statistics.

Pitman, J. (2006). *Combinatorial Stochastic Processes*. Lecture Notes in Mathematics. Springer-Verlag, Berlin.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900.

Pritchard, J., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–285.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Stein, M. L. (1999). *Statistical Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York.

Sudderth, E. and Jordan, M. I. (2009). Shared segmentation of natural scenes using dependent Pitman-Yor processes. In *Advances in Neural Information Processing Systems*, volume 21.

Teh, Y. W. (2006). A Bayesian interpretation of interpolated Kneser-Ney. Technical Report TRA2/06, School of Computing, National University of Singapore.

Teh, Y. W., Görür, D., and Ghahramani, Z. (2007). Stick-breaking construction for the Indian buffet process. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 11.

Teh, Y. W. and Jordan, M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In Hjort, N., Holmes, C., Müller, P., and Walker, S., editors, *Bayesian Nonparametrics*. Cambridge University Press.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Thibaux, R. and Jordan, M. I. (2007). Hierarchical beta processes and the Indian buffet process. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, volume 11, pages 564–571.

Ting, D., Wang, G., Shapovalov, M., Mitra, R., Jordan, M. I., and Dunbrack Jr, R. L. (2010). Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model. *PLoS Computational Biology*, 6(4):e1000763.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36:45.

Wolpert, R. L. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika*, 85(2):251–267.

Wood, F., Griffiths, T. L., and Ghahramani, Z. (2006). A non-parametric Bayesian method for inferring hidden causes. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, volume 22.

Zipf, G. (1932). *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press, Cambridge, MA.