Posteriors, conjugacy, and exponential families for completely random measures

Posteriors, conjugacy, and exponential families for completely random measures

Tamara Broderick Ashia C. Wilson Michael I. Jordan

April 25, 2016

Abstract

We demonstrate how to calculate posteriors for general Bayesian nonparametric priors and likelihoods based on completely random measures (CRMs). We further show how to represent Bayesian nonparametric priors as a sequence of finite draws using a size-biasing approach—and how to represent full Bayesian nonparametric models via finite marginals. Motivated by conjugate priors based on exponential family representations of likelihoods, we introduce a notion of exponential families for CRMs, which we call exponential CRMs. This construction allows us to specify automatic Bayesian nonparametric conjugate priors for exponential CRM likelihoods. We demonstrate that our exponential CRMs allow particularly straightforward recipes for size-biased and marginal representations of Bayesian nonparametric models. Along the way, we prove that the gamma process is a conjugate prior for the Poisson likelihood process and the beta prime process is a conjugate prior for a process we call the odds Bernoulli process. We deliver a size-biased representation of the gamma process and a marginal representation of the gamma process coupled with a Poisson likelihood process.

1 Introduction

An important milestone in Bayesian analysis was the development of a general strategy for obtaining conjugate priors based on exponential family representations of likelihoods [DeGroot, 1970]. While slavish adherence to exponential-family conjugacy can be criticized, conjugacy continues to occupy an important place in Bayesian analysis, for its computational tractability in high-dimensional problems and for its role in inspiring investigations into broader classes of priors (e.g., via mixtures, limits, or augmentations). The exponential family is, however, a parametric class of models, and it is of interest to consider whether similar general notions of conjugacy can be developed for Bayesian nonparametric models. Indeed, the nonparametric literature is replete with nomenclature that suggests the exponential family, including familiar names such as "Dirichlet," "beta," "gamma," and "Poisson." These names refer to aspects of the

random measures underlying Bayesian nonparametrics, either the Lévy measure used in constructing certain classes of random measures or properties of marginals obtained from random measures. In some cases, conjugacy results have been established that parallel results from classical exponential families; in particular, the Dirichlet process is known to be conjugate to a multinomial process likelihood [Ferguson, 1973], the beta process is conjugate to a Bernoulli process [Kim, 1999, Thibaux and Jordan, 2007] and to a negative binomial process [Broderick et al., 2015]. Moreover, various useful representations for marginal distributions, including stick-breaking and size-biased representations, have been obtained by making use of properties that derive from exponential families. It is striking, however, that these results have been obtained separately, and with significant effort; a general formalism that encompasses these individual results has not yet emerged. In this paper, we provide the single, holistic framework so strongly suggested by the nomenclature. Within this single framework, we show that it is straightforward to calculate posteriors and establish conjugacy. Our framework includes the specification of a Bayesian nonparametric analog of the finite exponential family, which allows us to provide automatic and constructive nonparametric conjugate priors given a likelihood specification as well as general recipes for marginal and size-biased representations.

A broad class of Bayesian nonparametric priors—including those built on the Dirichlet process [Ferguson, 1973], the beta process [Hjort, 1990], the gamma process [Ferguson, 1973, Lo, 1982, Titsias, 2008], and the negative binomial process [Zhou et al., 2012, Broderick et al., 2015]—can be viewed as models for the allocation of data points to traits. These processes give us pairs of traits together with rates or frequencies with which the traits occur in some population. Corresponding likelihoods assign each data point in the population to some finite subset of traits conditioned on the trait frequencies. What makes these models nonparametric is that the number of traits in the prior is countably infinite. Then the (typically random) number of traits to which any individual data point is allocated is unbounded, but also there are always new traits to which as-yet-unseen data points may be allocated. That is, such a model allows the number of traits in any data set to grow with the size of that data set.

A principal challenge of working with such models arises in posterior inference. There is a countable infinity of trait frequencies in the prior which we must integrate over to calculate the posterior of trait frequencies given allocations of data points to traits. Bayesian nonparametric models sidestep the full infinite-dimensional integration in three principal ways: conjugacy, size-biased representations, and marginalization.

In its most general form, conjugacy simply asserts that the prior is in the same family of distributions as the posterior. When the prior and likelihood are in finite-dimensional conjugate exponential families, conjugacy can turn posterior calculation into, effectively, vector addition. As a simple example, consider a model with beta-distributed prior, $\theta \sim \mathrm{Beta}(\theta|\alpha, \beta)$, for some fixed hyperparameters α and β. For the likelihood, let each observation x_n with $n \in \{1, \ldots, N\}$ be iid Bernoulli-distributed conditional on parameter θ: $x_n \overset{iid}{\sim} \mathrm{Bern}(x|\theta)$. Then the

posterior is simply another beta distribution, $\text{Beta}(\theta|\alpha_{post}, \beta_{post})$, with parameters updated via addition: $\alpha_{post} := \alpha + \sum_{n=1}^{N} x_n$ and $\beta_{post} := \beta + N - \sum_{n=1}^{N} x_n$. While conjugacy is certainly useful and popular in the case of finite parameter cardinality, there is arguably a stronger computational imperative for its use in the infinite-parameter case. Indeed, the core prior-likelihood pairs of Bayesian nonparametrics are generally proven [Hjort, 1990, Kim, 1999, Lo, 1982, Thibaux and Jordan, 2007, Broderick et al., 2015], or assumed to be [Titsias, 2008, Thibaux, 2008], conjugate. When such proofs exist, though, thus far they have been specialized to specific pairs of processes. In what follows, we demonstrate a general way to calculate posteriors for a class of distributions that includes all of these classical Bayesian nonparametric models. We also define a notion of exponential family representation for the infinite-dimensional case and show that, given a Bayesian nonparametric exponential family likelihood, we can readily construct a Bayesian nonparametric conjugate prior.

Size-biased sampling provides a finite-dimensional distribution for each of the individual prior trait frequencies [Thibaux and Jordan, 2007, Paisley et al., 2010]. Such a representation has played an important role in Bayesian nonparametrics in recent years, allowing for either exact inference via slice sampling [Damien et al., 1999, Neal, 2003]—as demonstrated by Teh et al. [2007], Broderick et al. [2015]—or approximate inference via truncation [Doshi et al., 2009, Paisley et al., 2011]. This representation is particularly useful for building hierarchical models [Thibaux and Jordan, 2007]. We show that our framework yields such representations in general, and we show that our construction is especially straightforward to use in the exponential family framework that we develop.

Marginal processes avoid directly representing the infinite-dimensional prior and posterior altogether by integrating out the trait frequencies. Since the trait allocations are finite for each data point, the marginal processes are finite for any finite set of data points. Again, thus far, such processes have been shown to exist separately in special cases; for example, the Indian buffet process [Griffiths and Ghahramani, 2006] is the marginal process for the beta process prior paired with a Bernoulli process likelihood [Thibaux and Jordan, 2007]. We show that the integration that generates the marginal process from the full Bayesian model can be generally applied in Bayesian nonparametrics and takes a particularly straightforward form when using conjugate exponential family priors and likelihoods. We further demonstrate that, in this case, a basic, constructive recipe exists for the general marginal process in terms of only finite-dimensional distributions.

Our results are built on the general class of stochastic processes known as *completely random measures* (CRMs) [Kingman, 1967]. We review CRMs in Section 2.1 and we discuss what assumptions are needed to form a full Bayesian nonparametric model from CRMs in Section 2.3. Given a general Bayesian nonparametric prior and likelihood (Section 2.2), we demonstrate in Section 3 how to calculate the posterior. Although the development up to this point is more general, we next introduce a concept of exponential families for CRMs

(Section 4.1) and call such models *exponential CRMs*. We show that we can generate automatic conjugate priors given exponential CRM likelihoods in Section 4.2. Finally, we show how we can generate recipes for size-biased representations (Section 5) and marginal processes (Section 6), which are particularly straightforward in the exponential CRM case (Corollary 5.2 in Section 5 and Corollary 6.2 in Section 6). We illustrate our results on a number of examples and derive new conjugacy results, size-biased representations, and marginal processes along the way.

We note that some similar results have been obtained by Orbanz [2010] and James [2014]. In the present work, we focus on creating representations that allow tractable inference.

2 Bayesian models based on completely random measures

As we have discussed, we view Bayesian nonparametric models as being composed of two parts: (1) a collection of pairs of traits together with their frequencies or rates and (2) for each data point, an allocation to different traits. Both parts can be expressed as *random measures*. Recall that a random measure is a random element whose values are measures.

We represent each trait by a point ψ in some space Ψ of traits. Further, let θ_k be the frequency, or rate, of the trait represented by ψ_k, where k indexes the countably many traits. In particular, $\theta_k \in \mathbb{R}_+$. Then (θ_k, ψ_k) is a tuple consisting of the frequency of the kth trait together with its trait descriptor. We can represent the full collection of pairs of traits with their frequencies by the discrete measure on Ψ that places weight θ_k at location ψ_k:

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

where the cardinality K may be finite or infinity.

Next, we form data point X_n for the nth individual. The data point X_n is viewed as a discrete measure. Each atom of X_n represents a pair consisting of (1) a trait to which the nth individual is allocated and (2) a degree to which the nth individual is allocated to this particular trait. That is,

$$X_n = \sum_{k=1}^{K_n} x_{n,k} \delta_{\psi_{n,k}}, \tag{2}$$

where again $\psi_{n,k} \in \Psi$ represents a trait and now $x_{n,k} \in \mathbb{R}_+$ represents the degree to which the nth data point belongs to trait $\psi_{n,k}$. K_n is the total number of traits to which the nth data point belongs.

Here and in what follows, we treat $X_{1:N} = \{X_n : n \in [N]\}$ as our observed data points for $[N] := \{1, 2, 3, \ldots, N\}$. In practice $X_{1:N}$ is often incorporated

into a more complex Bayesian hierarchical model. For instance, in topic modeling, ψ_k represents a topic; that is, ψ_k is a distribution over words in a vocabulary [Blei et al., 2003, Teh et al., 2006]. θ_k might represent the frequency with which the topic ψ_k occurs in a corpus of documents. $x_{n,k}$ might be a positive integer and represent the number of words in topic $\psi_{n,k}$ that occur in the nth document. So the nth document has a total length of $\sum_{k=1}^{K_n} x_{n,k}$ words. In this case, the actual observation consists of the words in each document, and the topics are latent. Not only are the results concerning posteriors, conjugacy, and exponential family representations that we develop below useful for inference in such models, but in fact our results are especially useful in such models—where the traits and any ordering on the traits are not known in advance.

Next, we want to specify a full Bayesian model for our data points $X_{1:N}$. To do so, we must first define a prior distribution for the random measure Θ as well as a likelihood for each random measure X_n conditioned on Θ. We let Σ_Ψ be a σ-algebra of subsets of Ψ, where we assume all singletons are in Σ_Ψ. Then we consider random measures Θ and X_n whose values are measures on Ψ. Note that for any random measure Θ and any measurable set $A \in \Sigma_\Psi$, $\Theta(A)$ is a random variable.

2.1 Completely random measures

We can see from Eqs. (1) and (2) that we desire a distribution on random measures that yields discrete measures almost surely. A particularly simple form of random measure called a *completely random measure* can be used to generate a.s. discrete random measures [Kingman, 1967].

A completely random measure Θ is defined as a random measure that satisfies one additional property; for any disjoint, measurable sets $A_1, A_2, \ldots, A_K \in \Sigma_\Psi$, we require that $\Theta(A_1), \Theta(A_2), \ldots, \Theta(A_K)$ be independent random variables. Kingman [1967] showed that a completely random measure can always be decomposed into a sum of three independent parts:

$$\Theta = \Theta_{det} + \Theta_{fix} + \Theta_{ord}. \tag{3}$$

Here, Θ_{det} is the deterministic component, Θ_{fix} is the *fixed-location* component, and Θ_{ord} is the *ordinary* component. In particular, Θ_{det} is any deterministic measure. We define the remaining two parts next.

The fixed-location component is called the "fixed component" by Kingman [1967], but we expand the name slightly here to emphasize that Θ_{fix} is defined to be constructed from a set of random weights at fixed (i.e., deterministic) locations. That is,

$$\Theta_{fix} = \sum_{k=1}^{K_{fix}} \theta_{fix,k} \delta_{\psi_{fix,k}}, \tag{4}$$

where the number of fixed-location atoms, K_{fix}, may be either finite or infinity; $\psi_{fix,k}$ is deterministic, and $\theta_{fix,k}$ is a non-negative, real-valued random variable (since Φ is a measure). Without loss of generality, we assume that the locations

$\psi_{fix,k}$ are all distinct. Then, by the independence assumption of CRMs, we must have that $\theta_{fix,k}$ are independent random variables across k. Although the fixed-location atoms are often ignored in the Bayesian nonparametrics literature, we will see that the fixed-location component has a key role to play in establishing Bayesian nonparametric conjugacy and in the CRM representations we present.

The third and final component is the ordinary component. Let $\#(A)$ denote the cardinality of some countable set A. Let μ be any σ-finite, deterministic measure on $\mathbb{R}_+ \times \Psi$, where $\mathbb{R}_+$ is equipped with the Borel σ-algebra and $\Sigma_{\mathbb{R}_+ \times \Psi}$ is the resulting product σ-algebra given Σ_Ψ. Recall that a *Poisson point process* with rate measure μ on $\mathbb{R}_+ \times \Psi$ is a random countable subset Π of $\mathbb{R}_+ \times \Psi$ such that two properties hold [Kingman, 1992]:

1. For any $A \in \Sigma_{\mathbb{R}_+ \times \Psi}$, $\#(\Pi \cap A) \sim \text{Poisson}(\mu(A))$.
2. For any disjoint $A_1, A_2, \ldots, A_K \in \Sigma_{\mathbb{R}_+ \times \Psi}$, $\#(\Pi \cap A_1), \#(\Pi \cap A_2), \cdots, \#(\Pi \cap A_K)$ are independent random variables.

To generate an ordinary component, start with a Poisson point process on $\mathbb{R}_+ \times \Psi$, characterized by its rate measure $\mu(d\theta \times d\psi)$. This process yields Π, a random and countable set of points: $\Pi = \{(\theta_{ord,k}, \psi_{ord,k})\}_{k=1}^{K_{ord}}$, where K_{ord} may be finite or infinity. Form the ordinary component measure by letting $\theta_{ord,k}$ be the weight of the atom located at $\psi_{ord,k}$:

$$\Theta_{ord} = \sum_{k=1}^{K_{ord}} \theta_{ord,k} \delta_{\psi_{ord,k}}. \tag{5}$$

Recall that we stated at the start of Section 2.1 that CRMs may be used to produce a.s. discrete random measures. To check this assertion, note that Θ_{fix} is a.s. discrete by construction (Eq. (4)) and Θ_{ord} is a.s. discrete by construction (Eq. (5)). Θ_{det} is the one component that may not be a.s. atomic. Thus the prevailing norm in using models based on CRMs is to set $\Theta_{det} \equiv 0$; in what follows, we adopt this norm. If the reader is concerned about missing any atoms in Θ_{det}, note that it is straightforward to adapt the treatment of Θ_{fix} to include the case where the atom weights are deterministic. When we set $\Theta_{det} \equiv 0$, we are left with $\Theta = \Theta_{fix} + \Theta_{ord}$ by Eq. (3). So Θ is also discrete, as desired.

2.2 Prior and likelihood

The prior that we place on Θ will be a fully general CRM (minus any deterministic component) with one additional assumption on the rate measure of the ordinary component. Before incorporating the additional assumption, we say that Θ has a fixed-location component with K_{fix} atoms, where the kth atom has arbitrary distribution $F_{fix,k}$: $\theta_{fix,k} \overset{indep}{\sim} F_{fix,k}(d\theta)$. K_{fix} may be finite or infinity, and Θ has an ordinary component characterized by rate measure $\mu(d\theta \times d\psi)$. The additional assumption we make is that the distribution on

the weights in the ordinary component is assumed to be decoupled from the distribution on the locations. That is, the rate measure decomposes as

$$\mu(d\theta \times d\psi) = \nu(d\theta) \cdot G(d\psi), \tag{6}$$

where ν is any σ-finite, deterministic measure on $\mathbb{R}_+$ and G is any proper distribution on Ψ. While the distribution over locations has been discussed extensively elsewhere [Neal, 2000, Wang and Blei, 2013], it is the weights that affect the allocation of data points to traits.

Given the factorization of μ in Eq. (6), the ordinary component of Θ can be generated by letting $\{\theta_{fix,k}\}_{k=1}^{K_{ord}}$ be the points of a Poisson point process generated on $\mathbb{R}_+$ with rate ν.[1] We then draw the locations $\{\psi_{fix,k}\}_{k=1}^{K_{ord}}$ iid according to $G(d\psi)$: $\psi_{fix,k} \overset{iid}{\sim} G(d\psi)$. Finally, for each k, $\theta_{fix,k}\delta_{\psi_{fix,k}}$ is an atom in Θ_{ord}. This factorization will allow us to focus our attention on the trait frequencies, and not the trait locations, in what follows. Moreover, going forward, we will assume G is diffuse (i.e., G has no atoms) so that the ordinary component atoms are all at a.s. distinct locations, which are further a.s. distinct from the fixed locations.

Since we have seen that Θ is an a.s. discrete random measure, we can write it as

$$\Theta = \sum_{k=1}^{K} \theta_k \delta_{\psi_k}, \tag{7}$$

where $K := K_{fix} + K_{ord}$ may be finite or infinity, and every ψ_k is a.s. unique. That is, we will sometimes find it helpful notationally to use Eq. (7) instead of separating the fixed and ordinary components. At this point, we have specified the prior for Θ in our general model.

Next, we specify the likelihood; i.e., we specify how to generate the data points X_n given Θ. We will assume each X_n is generated iid given Θ across the data indices n. We will let X_n be a CRM with only a fixed-location component given Θ. In particular, the atoms of X_n will be located at the atom locations of Θ, which are fixed when we condition on Θ:

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

Here, $x_{n,k}$ is drawn according to some distribution H that may take θ_k, the weight of Θ at location ψ_k, as a parameter; i.e.,

$$x_{n,k} \overset{indep}{\sim} H(dx|\theta_k) \quad \text{independently across } n \text{ and } k. \tag{8}$$

Note that while every atom of X_n is located at an atom of Θ, it is not necessarily the case that every atom of Θ has a corresponding atom in X_n. In particular, if $x_{n,k}$ is zero for any k, there is no atom in X_n at ψ_k.

[1] Recall that K_{ord} may be finite or infinity depending on ν and is random when taking finite values.

We highlight that the model above stands in contrast to Bayesian nonparametric *partition* models, for which there is a large literature. In partition models (or clustering models), Θ is a random probability measure [Ferguson, 1974]; in this case, the probability constraint precludes Θ from being a completely random measure, but it is often chosen to be a normalized completely random measure [James et al., 2009, Lijoi and Prünster, 2010]. The choice of Dirichlet process (a normalized gamma process) for Θ is particularly popular due to a number of useful properties that coincide in this single choice [Doksum, 1974, Escobar, 1994, Escobar and West, 1995, 1998, Ferguson, 1973, Lo, 1984, MacEachern, 1994, Perman et al., 1992, Pitman, 1996a,b, Sethuraman, 1994, West and Escobar, 1994]. In partition models, X_n is a draw from the probability distribution described by Θ. If we think of such X_n as a random measure, it is a.s. a single unit mass at a point ψ with strictly positive probability in Θ.

One potential connection between these two types of models is provided by combinatorial clustering [Broderick et al., 2015]. In partition models, we might suppose that we have a number of data sets, all of which we would like to partition. For instance, in a document modeling scenario, each document might be a data set; in particular each data point is a word in the document. And we might wish to partition the words in each document. An alternative perspective is to suppose that there is a single data set, where each data point is a document. Then the document exhibits traits with multiplicities, where the multiplicities might be the number of words from each trait; typically a trait in this application would be a topic. In this case, there are a number of other names besides feature or trait model that may be applied to the overarching model—such as admixture model or mixed membership model [Airoldi et al., 2014].

2.3 Bayesian nonparametrics

So far we have described a prior and likelihood that may be used to form a Bayesian model. We have already stated above that forming a *Bayesian nonparametric* model imposes some restrictions on the prior and likelihood. We formalize these restrictions in Assumptions A0, A1, and A2 below.

Recall that the premise of Bayesian nonparametrics is that the number of traits represented in a collection of data can grow with the number of data points. More explicitly, we achieve the desideratum that the number of traits is unbounded, and may always grow as new data points are collected, by modeling a countable infinity of traits. This assumption requires that the prior have a countable infinity of atoms. These must either be fixed-location atoms or ordinary component atoms. Fixed-location atoms represent known traits in some sense since we must know the fixed locations of the atoms in advance. Conversely, ordinary component atoms represent unknown traits, as yet to be discovered, since both their locations and associated rates are unknown a priori. Since we cannot know (or represent) a countable infinity of traits a priori, we cannot start with a countable infinity of fixed-location atoms.

A0. The number of fixed-location atoms in Θ is finite.

Since we require a countable infinity of traits in total and they cannot come from the fixed-location atoms by Assumption A0, the ordinary component must contain a countable infinity of atoms. This assumption will be true if and only if the rate measure on the trait frequencies has infinite mass.

A1. $\nu(\mathbb{R}_+) = \infty$.

Finally, an implicit part of the starting premise is that each data point be allocated to only a finite number of traits; we do not expect to glean an infinite amount of information from finitely represented data. Thus, we require that the number of atoms in every X_n be finite. By Assumption A0, the number of atoms in X_n that correspond to fixed-location atoms in Θ is finite. But by Assumption A1, the number of atoms in Θ from the ordinary component is infinite. So there must be some restriction on the distribution of values of X at the atoms of Θ (that is, some restriction on H in Eq. (8)) such that only finitely many of these values are nonzero.

In particular, note that if $H(dx|\theta)$ does not contain an atom at zero for any θ, then a.s. every one of the countable infinity of atoms of X will be nonzero. Conversely, it follows that, for our desiderata to hold, we must have that $H(dx|\theta)$ exhibits an atom at zero. One consequence of this observation is that $H(dx|\theta)$ cannot be purely continuous for all θ. Though this line of reasoning does not necessarily preclude a mixed continuous and discrete H, we henceforth assume that $H(dx|\theta)$ is discrete, with support $\mathbb{Z}_* = \{0, 1, 2, \ldots\}$, for all θ.

In what follows, we write $h(x|\theta)$ for the probability mass function of x given θ. So our requirement that each data point be allocated to only a finite number of traits translates into a requirement that the number of atoms of X_n with values in $\mathbb{Z}_+ = \{1, 2, \ldots\}$ be finite. Note that, by construction, the pairs $\{(\theta_{ord,k}, x_{ord,k})\}_{k=1}^{K_{ord}}$ form a marked Poisson point process with rate measure $\mu_{mark}(d\theta \times dx) := \nu(d\theta)h(x|\theta)$. And the pairs with $x_{ord,k}$ equal to any particular value $x \in \mathbb{Z}_+$ further form a thinned Poisson point process with rate measure $\nu_x(d\theta) := \nu(d\theta)h(x|\theta)$. In particular, the number of atoms of X with weight x is Poisson-distributed with mean $\nu_x(\mathbb{R}_+)$. So the number of atoms of X is finite if and only if the following assumption holds.[2]

A2. $\sum_{x=1}^{\infty} \nu_x(\mathbb{R}_+) < \infty$ for $\nu_x := \nu(d\theta)h(x|\theta)$.

Thus Assumptions A0, A1, and A2 capture our Bayesian nonparametric desiderata. We illustrate the development so far with an example.

Example 2.1. The *beta process* [Hjort, 1990] provides an example distribution for Θ. In its most general form, sometimes called the *three-parameter beta*

[2] When we have the more general case of a mixed continuous and discrete H, Assumption A2 becomes

A2b. $\int_{x>0} \int_{\theta \in \mathbb{R}_+} \nu(d\theta)H(dx|\theta) < \infty$.

process [Teh and Görür, 2009, Broderick et al., 2012], the beta process has an ordinary component whose weight rate measure has a beta distribution kernel,

$$\nu(d\theta) = \gamma\theta^{-\alpha-1}(1-\theta)^{c+\alpha-1}d\theta, \tag{9}$$

with support on $(0, 1]$. Here, the three fixed hyperparameters are γ, the *mass parameter*; c, the *concentration parameter*; and α, the *discount parameter*.[3] Moreover, each of its K_{fix} fixed-location atoms, $\theta_k\delta_{\psi_k}$, has a beta-distributed weight [Broderick et al., 2015]:

$$\theta_{fix,k} \sim \text{Beta}(\theta|\rho_{fix,k}, \sigma_{fix,k}), \tag{10}$$

where $\rho_{fix,k}, \sigma_{fix,k} > 0$ are fixed hyperparameters of the model.

By Assumption A0, K_{fix} is finite. By Assumption A1, $\nu(\mathbb{R}_+) = \infty$. To achieve this infinite-mass restriction, the beta kernel in Eq. (9) must be improper; i.e., either $-\alpha \leq 0$ or $c + \alpha \leq 0$. Also, note that we must have $\gamma > 0$ since ν is a measure (and the case $\gamma = 0$ would be trivial).

Often the beta process is used as a prior paired with a *Bernoulli process* likelihood [Thibaux and Jordan, 2007]. The Bernoulli process specifies that, given $\Theta = \sum_{k=1}^{\infty} \theta_k\delta_{\psi_k}$, we draw

$$x_{n,k} \overset{indep}{\sim} \text{Bern}(x|\theta_k),$$

which is well-defined since every atom weight θ_k of Θ is in $(0, 1]$ by the beta process construction. Thus,

$$X_n = \sum_{k=1}^{\infty} x_{n,k}\delta_{\psi_k}.$$

The marginal distribution of the $X_{1:N}$ in this case is often called an *Indian buffet process* [Griffiths and Ghahramani, 2006, Thibaux and Jordan, 2007]. The locations of atoms in X_n are thought of as the dishes sampled by the nth customer.

We take a moment to highlight the fact that continuous distributions for $H(dx|\theta)$ are precluded based on the Bayesian nonparametric desiderata by considering an alternative likelihood. Consider instead if $H(dx|\theta)$ were continuous here. Then X_1 would have atoms at every atom of Θ. In the Indian buffet process analogy, any customer would sample an infinite number of dishes, which contradicts our assumption that our data are finite. Indeed, any customer would sample all of the dishes at once. It is quite often the case in practical applications, though, that the X_n are merely latent variables, with the observed variables chosen according to a (potentially continuous) distribution given X_n [Griffiths and Ghahramani, 2006, Thibaux and Jordan, 2007];

[3] In [Teh and Görür, 2009, Broderick et al., 2012], the ordinary component features the beta distribution kernel in Eq. (9) multiplied not only by γ but also by a more complex, positive, real-valued expression in c and α. Since all of γ, c, and α are fixed hyperparameters, and γ is an arbitrary positive real value, any other constant factors containing the hyperparameters can be absorbed into γ, as in the main text here.

consider, e.g., mixture and admixture models. These cases are not precluded by our development.

Finally, then, we may apply Assumption A2, which specifies that the number of atoms in each observation X_n is finite; in this case, the assumption means

$$\sum_{x=1}^{\infty} \int_{\theta \in \mathbb{R}_+} \nu(d\theta) \cdot h(x|\theta) = \int_{\theta \in (0,1]} \nu(d\theta) \cdot h(1|\theta)$$

since θ is supported on $(0, 1]$ and x is supported on $\{0, 1\}$

$$= \int_{\theta \in (0,1]} \gamma \theta^{-\alpha-1}(1-\theta)^{c+\alpha-1} d\theta \cdot \theta = \gamma \int_{\theta \in (0,1]} \theta^{1-\alpha-1}(1-\theta)^{c+\alpha-1} d\theta < \infty.$$

The integral here is finite if and only if $1 - \alpha$ and $c + \alpha$ are the parameters of a proper beta distribution: i.e., if and only if $\alpha < 1$ and $c > -\alpha$. Together with the restrictions above, these restrictions imply the following allowable parameter ranges for the beta process fixed hyperparameters:

$$\gamma > 0, \quad \alpha \in [0, 1), \quad c > -\alpha, \quad \rho_{fix,k}, \sigma_{fix,k} > 0 \quad \text{for all } k \in [K_{fix}]. \tag{11}$$

These correspond to the hyperparameter ranges previously found in [Teh and Görür, 2009, Broderick et al., 2012]. ■

3 Posteriors

In Section 2, we defined a full Bayesian model consisting of a CRM prior for Θ and a CRM likelihood for an observation X conditional on Θ. Now we would like to calculate the posterior distribution of $\Theta|X$.

Theorem 3.1 (Bayesian nonparametric posteriors)**.** *Let Θ be a completely random measure that satisfies Assumptions A0 and A1; that is, Θ is a CRM with K_{fix} fixed atoms such that $K_{fix} < \infty$ and such that the kth atom can be written $\theta_{fix,k}\delta_{\psi_{fix,k}}$ with*

$$\theta_{fix,k} \overset{indep}{\sim} F_{fix,k}(d\theta)$$

for proper distribution $F_{fix,k}$ and deterministic $\psi_{fix,k}$. Let the ordinary component of Θ have rate measure

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

where G is a proper distribution and $\nu(\mathbb{R}_+) = \infty$. Write $\Theta = \sum_{k=1}^{\infty} \theta_k \delta_{\psi_k}$, and let X be generated conditional on Θ according to $X = \sum_{k=1}^{\infty} x_k \delta_{\psi_k}$ with $x_k \overset{indep}{\sim} h(x|\theta_k)$ for proper, discrete probability mass function h. And suppose X and Θ jointly satisfy Assumption A2 so that

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

Then let Θ_{post} be a random measure with the distribution of $\Theta|X$. Θ_{post} is a completely random measure with three parts.

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

2. *Let $\{x_{new,k}\delta_{\psi_{new,k}} : k \in [K_{new}]\}$ be the atoms of X that are not at fixed locations in the prior of Θ. K_{new} is finite by Assumption A2. Then Θ_{post} has a fixed-location atom at $x_{new,k}$ with random weight $\theta_{post,new,k}$, whose distribution $F_{post,new,k}(d\theta)$ is proportional to*

$$\nu(d\theta)h(x_{new,k}|\theta).$$

3. *The ordinary component of Θ_{post} has rate measure*

$$\nu_{post}(d\theta) := \nu(d\theta)h(0|\theta).$$

Proof. To prove the theorem, we consider in turn each of the two parts of the prior: the fixed-location component and the ordinary component. First, consider any fixed-location atom, $\theta_{fix,k}\delta_{\psi_{fix,k}}$, in the prior. All of the other fixed-location atoms in the prior, as well as the prior ordinary component, are independent of the random weight $\theta_{fix,k}$. So it follows that all of X except $x_{fix,k} := X(\{\psi_{fix,k}\})$ is independent of $\theta_{fix,k}$. Thus the posterior has a fixed atom located at $\psi_{fix,k}$ whose weight, which we denote $\theta_{post,fix,k}$, has distribution

$$F_{post,fix,k}(d\theta) \propto F_{fix,k}(d\theta)h(x_{fix,k}|\theta),$$

which follows from the usual finite Bayes Theorem.

Next, consider the ordinary component in the prior. Let

$$\Psi_{fix} = \{\psi_{fix,1}, \ldots, \psi_{fix,K_{fix}}\}$$

be the set of fixed-location atoms in the prior. Recall that Ψ_{fix} is deterministic, and since G is continuous, all of the fixed-location atoms and ordinary component atoms of Θ are at a.s. distinct locations. So the measure X_{fix} defined by

$$X_{fix}(A) := X(A \cap \Psi_{fix})$$

can be derived purely from X, without knowledge of Θ. It follows that the measure X_{ord} defined by

$$X_{ord}(A) := X(A \cap (\Psi \backslash \Psi_{fix}))$$

can be derived purely from X without knowledge of Θ. X_{ord} is the same as the observed data measure X but with atoms only at atoms of the ordinary component of Θ and not at the fixed-location atoms of Θ.

Now for any value $x \in \mathbb{Z}_+$, let

$$\{\psi_{new,x,1}, \ldots, \psi_{new,x,K_{new,x}}\}$$

be all of the locations of atoms of size x in X_{ord}. By Assumption A2, the number of such atoms, $K_{new,x}$, is finite. Further let $\theta_{new,x,k} := \Theta(\{\psi_{new,x,k}\})$. Then

the values $\{\theta_{new,x,k}\}_{k=1}^{K_{new,x}}$ are generated from a thinned Poisson point process with rate measure

$$\nu_x(d\theta) := \nu(d\theta)h(x|\theta). \tag{12}$$

And since $\nu_x(\mathbb{R}_+) < \infty$ by assumption, each $\theta_{new,x,k}$ has distribution equal to the normalized rate measure in Eq. (12). Note that $\theta_{new,x,k}\delta_{\psi_{new,x,k}}$ is a fixed-location atom in the posterior now that its location is known from the observed X_{ord}.

By contrast, if a likelihood draw at an ordinary component atom in the prior returns a zero, that atom is not observed in X_{ord}. Such atom weights in Θ_{post} thus form a marked Poisson point process with rate measure

$$\nu(d\theta)h(0|\theta),$$

as was to be shown. □

In Theorem 3.1, we consider generating Θ and then a single data point X conditional on Θ. Now suppose we generate Θ and then N data points, $X_1, \ldots, X_N$, iid conditional on Θ. In this case, Theorem 3.1 may be iterated to find the posterior $\Theta|X_{1:N}$. In particular, Theorem 3.1 gives the ordinary component and fixed atoms of the random measure $\Theta_1 := \Theta|X_1$. Then, using Θ_1 as the prior measure and X_2 as the data point, another application of Theorem 3.1 gives $\Theta_2 := \Theta|X_{1:2}$. We continue recursively using $\Theta|X_{1:n}$ for n between 1 and $N-1$ as the prior measure until we find $\Theta|X_{1:N}$. The result is made explicit in the following corollary.

Corollary 3.2 (Bayesian nonparametric posteriors given multiple data points)**.** *Let Θ be a completely random measure that satisfies Assumptions A0 and A1; that is, Θ is a CRM with K_{fix} fixed atoms such that $K_{fix} < \infty$ and such that the kth atom can be written $\theta_{fix,k}\delta_{\psi_{fix,k}}$ with*

$$\theta_{fix,k} \overset{indep}{\sim} F_{fix,k}(d\theta)$$

for proper distribution $F_{fix,k}$ and deterministic $\psi_{fix,k}$. Let the ordinary component of Θ have rate measure

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

where G is a proper distribution and $\nu(\mathbb{R}_+) = \infty$. Write $\Theta = \sum_{k=1}^{\infty} \theta_k \delta_{\psi_k}$, and let $X_1, \ldots, X_n$ be generated conditional on Θ according to $X = \sum_{k=1}^{\infty} x_{n,k}\delta_{\psi_{n,k}}$ with $x_{n,k} \overset{indep}{\sim} h(x|\theta_k)$ for proper, discrete probability mass function h. And suppose X_1 and Θ jointly satisfy Assumption A2 so that

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

It is enough to make the assumption for X_1 since the X_n are iid conditional on Θ.

Then let Θ_{post} be a random measure with the distribution of $\Theta|X_{1:N}$. Θ_{post} is a completely random measure with three parts.

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

2. *Let $\{\psi_{new,k} : k \in [K_{new}]\}$ be the union of atom locations across $X_1, X_2, \ldots, X_N$ minus the fixed locations in the prior of Θ. K_{new} is finite. Let $x_{new,n,k}$ be the weight of the atom in X_n located at $\psi_{new,k}$. Note that at least one of $x_{new,n,k}$ across n must be non-zero, but in general $x_{new,n,k}$ may equal zero. Then Θ_{post} has a fixed-location atom at $x_{new,k}$ with random weight $\theta_{post,new,k}$, whose distribution $F_{post,new,k}(d\theta)$ is proportional to*

$$\nu(d\theta) \prod_{n=1}^{N} h(x_{new,n,k}|\theta).$$

3. *The ordinary component of Θ_{post} has rate measure*

$$\nu_{post,n}(d\theta) := \nu(d\theta) \left[h(0|\theta)\right]^n.$$

Proof. Corollary 3.2 follows from recursive application of Theorem 3.1. In order to recursively apply Theorem 3.1, we need to verify that Assumptions A0, A1, and A2 hold for the posterior $\Theta|X_{1:(n+1)}$ when they hold for the prior $\Theta|X_{1:n}$. Note that the number of fixed atoms in the posterior is the number of fixed atoms in the prior plus the number of new atoms in the posterior. By Theorem 3.1, these counts are both finite as long as $\Theta|X_{1:n}$ satisfies Assumptions A0 and A2, which both hold for $n = 0$ by assumption and $n > 0$ by the recursive assumption. So Assumption A0 holds for $\Theta|X_{1:(n+1)}$.

Next we notice that since Assumption A1 implies that there is an infinite number of ordinary component atoms in $\Theta|X_{1:n}$ and only finitely many become fixed atoms in the posterior by Assumption A2, it must be that $\Theta|X_{1:(n+1)}$ has infinitely many ordinary component atoms. So Assumption A1 holds for $\Theta|X_{1:(n+1)}$.

Finally, we note that

$$\begin{aligned}
&\sum_{x=1}^{\infty} \int_{\theta \in \mathbb{R}_+} \nu_{post,n}(d\theta) h(x|\theta) \\
&= \sum_{x=1}^{\infty} \int_{\theta \in \mathbb{R}_+} \nu(d\theta) \left[h(0|\theta)\right]^n h(x|\theta) \leq \sum_{x=1}^{\infty} \int_{\theta \in \mathbb{R}_+} \nu(d\theta) h(x|\theta) < \infty,
\end{aligned}$$

where the penultimate inequality follows since $h(0|\theta) \in [0, 1]$ and where the inequality follows by Assumption A2 on the original Θ (conditioned on no data). So Assumption A2 holds for $\Theta|X_{1:(n+1)}$. □

We now illustrate the results of the theorem with an example.

Example 3.3. Suppose we again start with a beta process prior for Θ as in Example 2.1. This time we consider a *negative binomial process likelihood*

[Zhou et al., 2012, Broderick et al., 2015]. The negative binomial process specifies that, given $\Theta = \sum_{k=1}^{\infty} \theta_k \delta_{\psi_k}$, we draw $X = \sum_{k=1}^{\infty} x_k \delta_{\psi_k}$ with

$$x_k \overset{indep}{\sim} \text{NegBin}(x|r, \theta_k),$$

for some fixed hyperparameter $r > 0$. So

$$X_n = \sum_{k=1}^{\infty} x_{n,k} \delta_{\psi_k}.$$

In this case, Assumption A2 translates into the following restriction.

$$\begin{aligned}
&\sum_{x=1}^{\infty} \int_{\theta \in \mathbb{R}_+} \nu(d\theta) \cdot h(x|\theta) \\
&= \int_{\theta \in \mathbb{R}_+} \nu(d\theta) \cdot [1 - h(0|\theta)] = \int_{\theta \in (0,1]} \gamma \theta^{-\alpha-1} (1-\theta)^{c+\alpha-1} d\theta \cdot [1 - (1-\theta)^r] < \infty,
\end{aligned}$$

where the penultimate equality follows since the support of $\nu(d\theta)$ is $(0, 1]$.

By a Taylor expansion, we have $1 - (1-\theta)^r = r\theta + o(\theta)$ as $\theta \to 0$, so we require

$$\int_{\theta \in (0,1]} \theta^{1-\alpha-1} (1-\theta)^{c+\alpha-1} d\theta < \infty,$$

which is satisfied if and only if $1 - \alpha$ and $c + \alpha$ are the parameters of a proper beta distribution. Thus, we have the same parameter restrictions as in Eq. (11).

Now we calculate the posterior given the beta process prior on Θ and the negative binomial process likelihood for X conditional on Θ. In particular, the posterior has the distribution of Θ_{post}, a CRM with three parts given by Theorem 3.1.

First, at each fixed atom $\psi_{fix,k}$ of the prior with weight $\theta_{fix,k}$ given by Eq. (10), there is a fixed atom in the posterior with weight $\theta_{post,fix,k}$. Let $x_{post,fix,k} := X(\{\psi_{fix,k}\})$. Then $\theta_{post,fix,k}$ has distribution

$$\begin{aligned}
F_{post,fix,k}(d\theta) &\propto F_{fix}(d\theta) \cdot h(x_{post,fix,k}|\theta) \\
&= \text{Beta}(\theta|\rho_{fix,k}, \sigma_{fix,k})\ d\theta \cdot \text{NegBin}(x_{post,fix,k}|r, \theta) \\
&\propto \theta^{\rho_{fix,k}-1} (1-\theta)^{\sigma_{fix,k}-1}\ d\theta \cdot \theta^{x_{post,fix,k}} (1-\theta)^r \\
&\propto \text{Beta}\left(\theta\,|\rho_{fix,k} + x_{post,fix,k}, \sigma_{fix,k} + r\right)\ d\theta.
\end{aligned} \tag{13}$$

Second, for any atom $x_{new,k}\delta_{\psi_{new,k}}$ in X that is not at a fixed location in the prior, Θ_{post} has a fixed atom at $\psi_{new,k}$ whose weight $\theta_{post,new,k}$ has distribution

$$\begin{aligned}
F_{post,new,k}(d\theta) &\propto \nu(d\theta) \cdot h(x_{new,k}|\theta) \\
&= \nu(d\theta) \cdot \text{NegBin}(x_{new,k}|r, \theta) \\
&\propto \theta^{-\alpha-1} (1-\theta)^{c+\alpha-1}\ d\theta \cdot \theta^{x_{new,k}} (1-\theta)^r \\
&\propto \text{Beta}\left(\theta\,|-\alpha + x_{new,k}, c + \alpha + r\right)\ d\theta,
\end{aligned} \tag{14}$$

which is a proper distribution since we have the following restrictions on its parameters. For one, by assumption, $x_{new,k} \geq 1$. And further, by Eq. (11), we have $\alpha \in [0, 1)$ as well as $c + \alpha > 0$ and $r > 0$.

Third, the ordinary component of Θ_{post} has rate measure

$$\nu(d\theta)h(0|\theta) = \gamma\theta^{-\alpha-1}(1-\theta)^{c+\alpha-1}\, d\theta \cdot (1-\theta)^r = \gamma\theta^{-\alpha-1}(1-\theta)^{c+r+\alpha-1}\, d\theta.$$

Not only have we found the posterior distribution Θ_{post} above, but now we can note that the posterior is in the same form as the prior with updated ordinary component hyperparameters:

$$\gamma_{post} = \gamma, \quad \alpha_{post} = \alpha, \quad c_{post} = c + r.$$

The posterior also has old and new beta-distributed fixed atoms with beta distribution hyperparameters given in Eq. (13) and Eq. (14), respectively. Thus, we have proven that the beta process is, in fact, conjugate to the negative binomial process. An alternative proof was first given by Broderick et al. [2015]. ■

As in Example 3.3, we can use Theorem 3.1 not only to calculate posteriors but also, once those posteriors are calculated, to check for conjugacy. This approach unifies existing disparate approaches to Bayesian nonparametric conjugacy. However, it still requires the practitioner to guess the right conjugate prior for a given likelihood. In the next section, we define a notion of exponential families for CRMs, and we show how to automatically construct a conjugate prior for any exponential family likelihood.

4 Exponential families

Exponential families are what typically make conjugacy so powerful in the finite case. For one, when a finite likelihood belongs to an exponential family, then existing results give an automatic conjugate, exponential family prior for that likelihood. In this section, we review finite exponential families, define *exponential CRMs*, and show that analogous automatic conjugacy results can be obtained for exponential CRMs. Our development of exponential CRMs will also allow particularly straightforward results for size-biased representations (Corollary 5.2 in Section 5) and marginal processes (Corollary 6.2 in Section 6).

In the finite-dimensional case, suppose we have some (random) parameter θ and some (random) observation x whose distribution is conditioned on θ. We say the distribution $H_{exp,like}$ of x conditional on θ is in an exponential family if

$$H_{exp,like}(dx|\theta) = h_{exp,like}(x|\theta)\, dx = \kappa(x) \exp\left\{\langle \eta(\theta), \phi(x)\rangle - A(\theta)\right\}\, \mu(dx), \tag{15}$$

where $\eta(\theta)$ is the *natural parameter*, $\phi(x)$ is the *sufficient statistic*, $\kappa(x)$ is the *base density*, and $A(\theta)$ is the *log partition function*. We denote the density

of $H_{exp,like}$ here, which exists by definition, by $h_{exp,like}$. The measure μ—with respect to which the density $h_{exp,like}$ exists—is typically Lebesgue measure when $H_{exp,like}$ is diffuse or counting measure when $H_{exp,like}$ is atomic. $A(\theta)$ is determined by the condition that $H_{exp,like}(dx|\theta)$ have unit total mass on its support.

It is a classic result [Diaconis and Ylvisaker, 1979] that the following distribution for $\theta \in \mathbb{R}^D$ constitutes a conjugate prior:

$$F_{exp,prior}(d\theta) = f_{exp,prior}(\theta)\, d\theta = \exp\left\{\langle \xi, \eta(\theta)\rangle + \lambda\left[-A(\theta)\right] - B(\xi,\lambda)\right\}\, d\theta. \tag{16}$$

$F_{exp,prior}$ is another exponential family distribution, now with natural parameter $(\xi', \lambda)'$, sufficient statistic $(\eta(\theta)', -A(\theta))'$, and log partition function $B(\xi,\lambda)$. Note that the logarithms of the densities in both Eq. (15) and Eq. (16) are linear in $\eta(\theta)$ and $-A(\theta)$. So, by Bayes Theorem, the posterior $F_{exp,post}$ also has these quantities as sufficient statistics in θ, and we can see $F_{exp,post}$ must have the following form.

$$\begin{aligned} &F_{exp,post}(d\theta|x) = f_{exp,post}(\theta|x)\, d\theta \\ &= \exp\left\{\langle \xi + \phi(x), \eta(\theta)\rangle + (\lambda+1)\left[-A(\theta)\right] - B(\xi+\phi(x), \lambda+1)\right\}\, d\theta. \end{aligned} \tag{17}$$

Thus we see that $F_{exp,post}$ belongs to the same exponential family as $F_{exp,prior}$ in Eq. (16), and hence $F_{exp,prior}$ is a conjugate prior for $H_{exp,like}$ in Eq. (15).

4.1 Exponential families for completely random measures

In the finite-dimensional case, we saw that for any exponential family likelihood, as in Eq. (15), we can always construct a conjugate exponential family prior, given by Eq. (16).

In order to prove a similar result for CRMs, we start by defining a notion of exponential families for CRMs.

Definition 4.1. We say that a CRM Θ is an *exponential CRM* if it has the following two parts. First, let Θ have K_{fix} fixed-location atoms, where K_{fix} may be finite or infinite. The kth fixed-location atom is located at any $\psi_{fix,k}$, unique from the other fixed locations, and has random weight $\theta_{fix,k}$, whose distribution has density $f_{fix,k}$:

$$f_{fix,k}(\theta) = \kappa(\theta)\exp\left\{\langle \eta(\zeta_k), \phi(\theta)\rangle - A(\zeta_k)\right\},$$

for some base density κ, natural parameter function η, sufficient statistic ϕ, and log partition function A shared across atoms. Here, ζ_k is an atom-specific parameter.

Second, let Θ have an ordinary component with rate measure $\mu(d\theta \times d\psi) = \nu(d\theta)\cdot G(d\psi)$ for some proper distribution G and weight rate measure ν of the form

$$\nu(d\theta) = \gamma\exp\left\{\langle \eta(\zeta), \phi(\theta)\rangle\right\}.$$

In particular, η and ϕ are shared with the fixed-location atoms, and fixed hyperparameters γ and ζ are unique to the ordinary component.

4.2 Automatic conjugacy for completely random measures

With Definition 4.1 in hand, we can specify an automatic Bayesian nonparametric conjugate prior for an exponential CRM likelihood.

Theorem 4.2 (Automatic conjugacy)**.** *Let $\Theta = \sum_{k=1}^{\infty} \theta_k \delta_{\psi_k}$, in accordance with Assumption A1. Let X be generated conditional on Θ according to an exponential CRM with fixed-location atoms at $\{\psi_k\}_{k=1}^{\infty}$ and no ordinary component. In particular, the distribution of the weight x_k at ψ_k of X has the following density conditional on the weight θ_k at ψ_k of Θ:*

$$h(x|\theta_k) = \kappa(x) \exp\left\{\langle \eta(\theta_k), \phi(x)\rangle - A(\theta_k)\right\}.$$

Then a conjugate prior for Θ is the following exponential CRM distribution. First, let Θ have $K_{prior,fix}$ fixed-location atoms, in accordance with Assumption A0. The kth such atom has random weight $\theta_{fix,k}$ with proper density

$$f_{prior,fix,k}(\theta) = \exp\left\{\langle \xi_{fix,k}, \eta(\theta)\rangle + \lambda_{fix,k}\left[-A(\theta)\right] - B(\xi_{fix,k}, \lambda_{fix,k})\right\},$$

where $(\eta', -A)'$ here is the sufficient statistic and B is the log partition function. $\xi_{fix,k}$ and $\lambda_{fix,k}$ are fixed hyperparameters for this atom weight.

Second, let Θ have ordinary component characterized by any proper distribution G and weight rate measure

$$\nu(d\theta) = \gamma \exp\left\{\langle \xi, \eta(\theta)\rangle + \lambda\left[-A(\theta)\right]\right\},$$

where γ, ξ, and λ are fixed hyperparameters of the weight rate measure chosen to satisfy Assumptions A1 and A2.

Proof. To prove the conjugacy of the prior for Θ with the likelihood for X, we calculate the posterior distribution of $\Theta|X$ using Theorem 3.1. Let Θ_{post} be a CRM with the distribution of $\Theta|X$. Then, by Theorem 3.1, Θ_{post} has the following three parts.

First, at any fixed location $\psi_{fix,k}$ in the prior, let $x_{fix,k}$ be the value of X at that location. Then Θ_{post} has a fixed-location atom at $\psi_{fix,k}$, and its weight $\theta_{post,fix,k}$ has distribution

$$\begin{aligned} F_{post,fix,k}(d\theta) &\propto f_{prior,fix,k}(\theta)\ d\theta \cdot h(x_{fix,k}|\theta) \\ &\propto \exp\left\{\langle \xi_{fix,k}, \eta(\theta)\rangle + \lambda_{fix,k}\left[-A(\theta)\right]\right\}\ d\theta \cdot \exp\left\{\langle \eta(\theta), \phi(x_{fix,k})\rangle - A(\theta)\right\}\ d\theta \\ &= \exp\left\{\langle \xi_{fix,k} + \phi(x_{fix,k}), \eta(\theta)\rangle + (\lambda_{fix,k}+1)\left[-A(\theta)\right]\right\}\ d\theta. \end{aligned}$$

It follows, from putting in the normalizing constant, that the distribution of $\theta_{post,fix,k}$ has density

$$\begin{aligned} f_{post,fix,k}(\theta) = \exp\{&\langle \xi_{fix,k} + \phi(x_{fix,k}), \eta(\theta)\rangle + (\lambda_{fix,k}+1)\left[-A(\theta)\right] \\ &- B(\xi_{fix,k} + \phi(x_{fix,k}), \lambda_{fix,k}+1)\}. \end{aligned}$$

Second, for any atom $x_{new,k}\delta_{\psi_{new,k}}$ in X that is not at a fixed location in the prior, Θ_{post} has a fixed atom at $\psi_{new,k}$ whose weight $\theta_{post,new,k}$ has distribution

$$F_{post,new,k}(\theta) \propto \nu(d\theta) \cdot h(x_{new,k}|\theta)$$

$$\propto \exp\left\{\langle \xi, \eta(\theta)\rangle + \lambda\left[-A(\theta)\right]\right\} \cdot \exp\left\{\langle \eta(\theta), \phi(x_{new,k})\rangle - A(\theta)\right\}\ d\theta$$
$$= \exp\left\{\langle \xi + \phi(x_{new,k}), \eta(\theta)\rangle + (\lambda+1)\left[-A(\theta)\right]\right\}\ d\theta$$

and hence density

$$f_{post,new,k}(\theta) = \exp\left\{\langle \xi + \phi(x_{new,k}), \eta(\theta)\rangle + (\lambda+1)\left[-A(\theta)\right] - B(\xi + \phi(x_{new,k}), \lambda+1)\right\}.$$

Third, the ordinary component of Θ_{post} has weight rate measure

$$\nu(d\theta) \cdot h(0|\theta)$$
$$= \gamma \exp\left\{\langle \xi, \eta(\theta)\rangle + \lambda\left[-A(\theta)\right]\right\} \cdot \kappa(0) \exp\left\{\langle \eta(\theta), \phi(0)\rangle - A(\theta)\right\}$$
$$= \gamma\kappa(0) \cdot \exp\left\{\langle \xi + \phi(0), \eta(\theta)\rangle + (\lambda+1)\left[-A(\theta)\right]\right\}.$$

Thus, the posterior rate measure is in the same exponential CRM form as the prior rate measure with updated hyperparameters:

$$\gamma_{post} = \gamma\kappa(0), \quad \xi_{post} = \xi + \phi(0), \quad \lambda_{post} = \lambda + 1.$$

Since we see that the posterior fixed-location atoms are likewise in the same exponential CRM form as the prior, we have shown that conjugacy holds, as desired. □

We next use Theorem 4.2 to give proofs of conjugacy in cases where conjugacy has not previously been established in the Bayesian nonparametrics literature.

Example 4.3. Let X be generated according to a *Poisson likelihood process*[4] conditional on Θ. That is, $X = \sum_{k=1}^{\infty} x_k \delta_{\psi_k}$ conditional on $\Theta = \sum_{k=1}^{\infty} \theta_k \delta_{\psi_k}$ has an exponential CRM distribution with only a fixed-location component. The weight x_k at location ψ_k has support on $\mathbb{Z}_*$ and has a Poisson density with parameter $\theta_k \in \mathbb{R}_+$:

$$h(x|\theta_k) = \frac{1}{x!}\theta_k^x e^{-\theta_k} = \frac{1}{x!}\exp\left\{x\log(\theta_k) - \theta_k\right\}. \tag{18}$$

The final line is rewritten to emphasize the exponential family form of this density, with

$$\kappa(x) = \frac{1}{x!}, \quad \phi(x) = x, \quad \eta(\theta) = \log(\theta), \quad A(\theta) = \theta.$$

By Theorem 4.2, this Poisson likelihood process has a Bayesian nonparametric conjugate prior for Θ with two parts.

First, Θ has a set of $K_{prior,fix}$ fixed-location atoms, where $K_{prior,fix} < \infty$ by Assumption A0. The kth such atom has random weight $\theta_{fix,k}$ with density

$$f_{prior,fix,k}(\theta) = \exp\left\{\langle \xi_{fix,k}, \eta(\theta)\rangle + \lambda_{fix,k}\left[-A(\theta)\right] - B(\xi_{fix,k}, \lambda_{fix,k})\right\}$$

[4]We use the term "Poisson likelihood process" to distinguish this specific Bayesian nonparametric likelihood from the Poisson point process.

$$= \theta^{\xi_{fix,k}} e^{-\lambda_{fix,k}\theta} \exp\{-B(\xi_{fix,k}, \lambda_{fix,k})\}$$
$$= \text{Gamma}(\theta\,|\xi_{fix,k}+1, \lambda_{fix,k}), \tag{19}$$

where $\text{Gamma}(\theta|a,b)$ denotes the gamma density with shape parameter $a>0$ and rate parameter $b>0$. So we must have fixed hyperparameters $\xi_{fix,k} > -1$ and $\lambda_{fix,k} > 0$. Further,

$$\exp\{-B(\xi_{fix,k}, \lambda_{fix,k})\} = \lambda_{fix,k}^{\xi_{fix,k}+1}/\Gamma(\xi_{fix,k}+1)$$

to ensure normalization.

Second, Θ has an ordinary component characterized by any proper distribution G and weight rate measure

$$\nu(d\theta) = \gamma \exp\{\langle \xi, \eta(\theta)\rangle + \lambda\,[-A(\theta)]\}\; d\theta = \gamma\theta^{\xi} e^{-\lambda\theta}\; d\theta. \tag{20}$$

Note that Theorem 4.2 guarantees that the weight rate measure will have the same distributional kernel in θ as the fixed-location atoms.

Finally, we need to choose the allowable hyperparameter ranges for γ, ξ, and λ. First, $\gamma > 0$ to ensure ν is a measure. By Assumption A1, we must have $\nu(\mathbb{R}_+) = \infty$, so ν must represent an improper gamma distribution. As such, we require either $\xi + 1 \leq 0$ or $\lambda \leq 0$. By Assumption A2, we must have

$$\sum_{x=1}^{\infty} \int_{\theta\in\mathbb{R}_+} \nu(d\theta)\cdot h(x|\theta) = \int_{\theta\in\mathbb{R}_+} \nu(d\theta)\cdot[1-h(0|\theta)] = \int_{\theta\in\mathbb{R}_+} \gamma\theta^{\xi}e^{-\lambda\theta}\; d\theta \cdot \left[1-e^{-\theta}\right] < \infty.$$

To ensure the integral over $[1,\infty)$ is finite, we must have $\lambda > 0$. To ensure the integral over $(0,1)$ is finite, we note that $1 - e^{-\theta} = \theta + o(\theta)$ as $\theta \to 0$. So we require

$$\int_{\theta\in(0,1)} \gamma\theta^{\xi+1} e^{-\lambda\theta}\; d\theta < \infty,$$

which is satisfied if and only if $\xi + 2 > 0$.

Finally, then the hyperparameter restrictions can be summarized as:

$$\gamma > 0, \quad \xi \in (-2,-1], \quad \lambda > 0; \quad \xi_{fix,k} > -1 \text{ and } \lambda_{fix,k} > 0 \quad \text{for all } k \in [K_{prior,fix}].$$

The ordinary component of the conjugate prior for Θ discovered in this example is typically called a *gamma process*. Here, we have for the first time specified the distribution of the fixed-location atoms of the gamma process and, also for the first time, proved that the gamma process is conjugate to the Poisson likelihood process. We highlight this result as a corollary to Theorem 4.2.

Corollary 4.4. *Let the Poisson likelihood process be a CRM with fixed-location atom weight distributions as in Eq. (18). Let the gamma process be a CRM with fixed-location atom weight distributions as in Eq. (19) and ordinary component weight measure as in Eq. (20). Then the gamma process is a conjugate Bayesian nonparametric prior for the Poisson likelihood process.*

■

Example 4.5. Next, let X be generated according to a new process we call an *odds Bernoulli process*. We have previously seen a typical Bernoulli process likelihood in Example 2.1. In the odds Bernoulli process, we say that X, conditional on Θ, has an exponential CRM distribution. In this case, the weight of the kth atom, x_k, conditional on θ_k has support on $\{0, 1\}$ and has a Bernoulli density with odds parameter $\theta_k \in \mathbb{R}_+$:

$$\begin{aligned} h(x|\theta_k) &= \theta_k^x (1+\theta_k)^{-1} \\ &= \exp\left\{x \log(\theta_k) - \log(1+\theta_k)\right\}. \end{aligned} \tag{21}$$

That is, if ρ is the probability of a successful Bernoulli draw, then $\theta = \rho/(1-\rho)$ represents the odds ratio of the probability of success over the probability of failure.

The final line of Eq. (21) is written to emphasize the exponential family form of this density, with

$$\kappa(x) = 1, \quad \phi(x) = x, \quad \eta(\theta) = \log(\theta), \quad A(\theta) = \log(1+\theta).$$

By Theorem 4.2, the likelihood for X has a Bayesian nonparametric conjugate prior for Θ. This conjugate prior has two parts.

First, Θ has a set of $K_{prior,fix}$ fixed-location atoms. The kth such atom has random weight $\theta_{fix,k}$ with density

$$\begin{aligned} f_{prior,fix,k}(\theta) &= \exp\left\{\langle \xi_{fix,k}, \eta(\theta)\rangle + \lambda_{fix,k}\left[-A(\theta)\right] - B(\xi_{fix,k}, \lambda_{fix,k})\right\} \\ &= \theta^{\xi_{fix,k}} (1+\theta)^{-\lambda_{fix,k}} \exp\left\{-B(\xi_{fix,k}, \lambda_{fix,k})\right\} \\ &= \text{BetaPrime}\left(\theta \,|\xi_{fix,k}+1, \lambda_{fix,k} - \xi_{fix,k} - 1\right), \end{aligned} \tag{22}$$

where $\text{BetaPrime}(\theta|a, b)$ denotes the beta prime density with shape parameters $a > 0$ and $b > 0$. Further,

$$\exp\left\{-B(\xi_{fix,k}, \lambda_{fix,k})\right\} = \frac{\Gamma(\lambda_{fix,k})}{\Gamma(\xi_{fix,k}+1)\Gamma(\lambda_{fix,k} - \xi_{fix,k} - 1)}$$

to ensure normalization.

Second, Θ has an ordinary component characterized by any proper distribution G and weight rate measure

$$\nu(d\theta) = \gamma \exp\left\{\langle \xi, \eta(\theta)\rangle + \lambda\left[-A(\theta)\right]\right\} \, d\theta = \gamma \theta^{\xi} (1+\theta)^{-\lambda} \, d\theta. \tag{23}$$

We need to choose the allowable hyperparameter ranges for γ, ξ, and λ. First, $\gamma > 0$ to ensure ν is a measure. By Assumption A1, we must have $\nu(\mathbb{R}_+) = \infty$, so ν must represent an improper beta prime distribution. As such, we require either $\xi + 1 \leq 0$ or $\lambda - \xi - 1 \leq 0$. By Assumption A2, we must have

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

since the support of x is $\{0, 1\}$

$$= \int_{\theta \in \mathbb{R}_+} \gamma \theta^{\xi} (1+\theta)^{-\lambda} \, d\theta \cdot \theta^{1} (1+\theta)^{-1} = \gamma \int_{\theta \in \mathbb{R}_+} \theta^{\xi+1} (1+\theta)^{-\lambda-1} \, d\theta < \infty.$$

Since the integrand is the kernel of a beta prime distribution, we simply require that this distribution be proper; i.e., $\xi + 2 > 0$ and $\lambda - \xi - 1 > 0$.

The hyperparameter restrictions can be summarized as:

$$\gamma > 0, \xi \in (-2, -1], \lambda > \xi + 1; \xi_{fix,k} > -1 \text{ and } \lambda_{fix,k} > \xi_{fix,k} + 1 \text{ for all } k \in [K_{prior,fix}].$$

We call the distribution for Θ described in this example the *beta prime process*. Its ordinary component has previously been defined by Broderick et al. [2015]. But this result represents the first time the beta prime process is described in full, including parameter restrictions and fixed-location atoms, as well as the first proof of its conjugacy with the odds Bernoulli process. We highlight the latter result as a corollary to Theorem 4.2 below.

Corollary 4.6. *Let the odds Bernoulli process be a CRM with fixed-location atom weight distributions as in Eq. (21). Let the beta process be a CRM with fixed-location atom weight distributions as in Eq. (22) and ordinary component weight measure as in Eq. (23). Then the beta process is a conjugate Bayesian nonparametric prior for the odds Bernoulli process.*

■

5 Size-biased representations

We have shown in Section 4.2 that our exponential CRM (Definition 4.1) is useful in that we can find an automatic Bayesian nonparametric conjugate prior given an exponential CRM likelihood. We will see in this section and the next that exponential CRMs allow us to build representations that allow tractable inference despite the infinite-dimensional nature of the models we are using.

The best-known size-biased representation of a random measure in Bayesian nonparametrics is the *stick-breaking* representation of the Dirichlet process Θ_{DP} [Sethuraman, 1994]:

$$\begin{aligned} \Theta_{DP} &= \sum_{k=1}^{\infty} \theta_{DP,k} \delta_{\psi_k}; \\ \text{For } k \in \mathbb{Z}_*, \, \theta_{DP,k} &= \beta_k \prod_{j=1}^{k-1} (1 - \beta_j), \quad \beta_k \overset{iid}{\sim} \text{Beta}(1, c), \quad \psi_k \overset{iid}{\sim} G, \end{aligned} \tag{24}$$

where c is a fixed hyperparameter satisfying $c > 0$.

The name "stick-breaking" originates from thinking of the unit interval as a stick of length one. At each round k, only some of the stick remains; β_k describes the proportion of the remaining stick that is broken off in round k, and $\theta_{DP,k}$ describes the total amount of remaining stick that is broken off in

round k. By construction, not only is each $\theta_{DP,k} \in (0,1)$ but in fact the $\theta_{DP,k}$ add to one (the total stick length) and thus describe a distribution.

Eq. (24) is called a *size-biased* representation for the following reason. Since the weights $\{\theta_{DP,k}\}_{k=1}^{\infty}$ describe a distribution, we can make draws from this distribution; each such draw is sometimes thought of as a multinomial draw with a single trial. In that vein, typically we imagine that our data points $X_{mult,n}$ are described as iid draws conditioned on Θ_{DP}, where $X_{mult,n}$ is a random measure with just a single atom:

$$X_{mult,n} = \delta_{\psi_{mult,n}}; \quad \psi_{mult,n} = \psi_k \text{ with probability } \theta_{DP,k}. \tag{25}$$

Then the limiting proportion of data points $X_{mult,n}$ with an atom at $\psi_{mult,1}$ (the first atom location chosen) is $\theta_{DP,1}$. The limiting proportion of data points with an atom at the next unique atom location chosen will have size $\theta_{DP,2}$, and so on [Broderick et al., 2013].

The representation in Eq. (24) is so useful because there is a familiar, finite-dimensional distribution for each of the atom weights $\theta_{DP,k}$ of the random measure Θ_{DP}. This representation allows approximate inference via truncation [Ishwaran and James, 2001] or exact inference via slice sampling [Walker, 2007, Kalli et al., 2011].

Since the weights $\{\theta_{DP,k}\}_{k=1}^{\infty}$ are constrained to sum to one, the Dirichlet process is not a CRM.[5] Indeed, there has been much work on size-biased representations for more general normalized random measures, which include the Dirichlet process as just one example [Perman et al., 1992, Pitman, 1996a,b, 2003].

By contrast, we here wish to explore size-biasing for non-normalized CRMs. In the normalized CRM case, we considered which atom of a random discrete probability measure was drawn first and what is the distribution of that atom's size. In the non-normalized CRM case considered in the present work, when drawing X conditional on Θ, there may be multiple atoms (or one atom or no atoms) of Θ that correspond to non-zero atoms in X. The number will always be finite though by Assumption A2. In this non-normalized CRM case, we wish to consider the sizes of all such atoms in Θ. Size-biased representations have been developed in the past for particular CRM examples, notably the beta process [Paisley et al., 2010, Broderick et al., 2012]. And even though there is typically no interpretation of these representations in terms of a single stick representing a unit probability mass, they are sometimes referred to as stick-breaking representations as a nod to the popularity of Dirichlet process stick-breaking.

In the beta process case, such size-biased representations have already been shown to allow approximate inference via truncation [Doshi et al., 2009, Paisley et al., 2011] or exact inference via slice sampling [Teh et al., 2007, Broderick et al., 2015]. Here we provide general recipes for the creation of these representations and illustrate our recipes by discovering previously unknown size-biased

[5] In fact, the Dirichlet process is a normalized gamma process (cf. Example 4.3) [Ferguson, 1973].

representations.

We have seen that a general CRM Θ takes the form of an a.s. discrete random measure:

$$\sum_{k=1}^{\infty} \theta_k \delta_{\psi_k}. \tag{26}$$

The fixed-location atoms are straightforward to simulate; there are finitely many by Assumption A0, their locations are fixed, and their weights are assumed to come from finite-dimensional distributions. The infinite-dimensionality of the Bayesian nonparametric CRM comes from the ordinary component (cf. Section 2.3 and Assumption A1). So far the only description we have of the ordinary component is its generation from the countable infinity of points in a Poisson point process. The next result constructively demonstrates that we can represent the distributions of the CRM weights $\{\theta_k\}_{k=1}^{\infty}$ in Eq. (26) as a sequence of finite-dimensional distributions, much as in the familiar Dirichlet process case.

Theorem 5.1 (Size-biased representations)**.** *Let Θ be a completely random measure that satisfies Assumptions A0 and A1; that is, Θ is a CRM with K_{fix} fixed atoms such that $K_{fix} < \infty$ and such that the kth atom can be written $\theta_{fix,k}\delta_{\psi_{fix,k}}$. The ordinary component of Θ has rate measure*

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

where G is a proper distribution and $\nu(\mathbb{R}_+) = \infty$. Write $\Theta = \sum_{k=1}^{\infty} \theta_k \delta_{\psi_k}$, and let X_n be generated iid given Θ according to $X_n = \sum_{k=1}^{\infty} x_{n,k}\delta_{\psi_k}$ with $x_{n,k} \overset{indep}{\sim} h(x|\theta_k)$ for proper, discrete probability mass function h. And suppose X_n and Θ jointly satisfy Assumption A2 so that

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

Then we can write

$$\begin{aligned}
\Theta &= \sum_{m=1}^{\infty} \sum_{x=1}^{\infty} \sum_{j=1}^{\rho_{m,x}} \theta_{m,x,j} \delta_{\psi_{m,x,j}} \\
\psi_{m,x,k} &\overset{iid}{\sim} G \text{ iid across } m, x, j \\
\rho_{m,x} &\overset{indep}{\sim} \text{Poisson}\left(\rho \left| \int_{\theta} \nu(d\theta) h(0|\theta)^{m-1} h(x|\theta) \right.\right) \text{ across } m, x \\
\theta_{m,x,j} &\overset{indep}{\sim} F_{size,m,x}(d\theta) \propto \nu(d\theta) h(0|\theta)^{m-1} h(x|\theta) \\
&\quad \text{iid across } j \text{ and independently across } m, x.
\end{aligned} \tag{27}$$

Proof. By construction, Θ is an a.s. discrete random measure with a countable infinity of atoms. Without loss of generality, suppose that for every (non-zero) value of an atom weight θ, there is a non-zero probability of generating an atom

with non-zero weight x in the likelihood. Now suppose we generate $X_1, X_2, \ldots$. Then, for every atom $\theta\delta_\psi$ of Θ, there exists some finite n with an atom at ψ. Therefore, we can enumerate all of the atoms of Θ by enumerating

- Each atom $\theta\delta_\psi$ such that there is an atom in X_1 at ψ.
- Each atom $\theta\delta_\psi$ such that there is an atom in X_2 at ψ but there is not an atom in X_1 at ψ.

  $\vdots$
- Each atom $\theta\delta_\psi$ such that there is an atom in X_m at ψ but there is not an atom in any of $X_1, X_2, \ldots, X_{m-1}$ at ψ.

  $\vdots$

Moreover, on the mth round of this enumeration, we can further break down the enumeration by the value of the observation X_m at the atom location:

- Each atom $\theta\delta_\psi$ such that there is an atom in X_m **of weight** 1 at ψ but there is not an atom in any of $X_1, X_2, \ldots, X_{m-1}$ at ψ.
- Each atom $\theta\delta_\psi$ such that there is an atom in X_m **of weight** 2 at ψ but there is not an atom in any of $X_1, X_2, \ldots, X_{m-1}$ at ψ.

  $\vdots$
- Each atom $\theta\delta_\psi$ such that there is an atom in X_m **of weight** x at ψ but there is not an atom in any of $X_1, X_2, \ldots, X_{m-1}$ at ψ.

  $\vdots$

Recall that the values θ_k that form the weights of Θ are generated according to a Poisson point process with rate measure $\nu(d\theta)$. So, on the first round, the values of θ_k such that $x_{1,k} = x$ also holds are generated according to a thinned Poisson point process with rate measure

$$\nu(d\theta)h(x|\theta).$$

In particular, since the rate measure has finite total mass by Assumption A2, we can define

$$M_{1,x} := \int_\theta \nu(d\theta)h(x|\theta),$$

which will be finite. Then the number of atoms θ_k for which $x_{1,k} = x$ is

$$\rho_{1,x} \sim \text{Poisson}(\rho|M_{1,x}).$$

And each such θ_k has weight with distribution

$$F_{size,1,x}(d\theta) \propto \nu(d\theta)h(x|\theta).$$

Finally, note from Theorem 3.1 that the posterior $\Theta|X_1$ has weight rate measure

$$\nu_1(d\theta) := \nu(d\theta)h(0|\theta).$$

Now take any $m > 1$. Suppose, inductively, that the ordinary component of the posterior $\Theta|X_1, \ldots, X_{m-1}$ has weight rate measure

$$\nu_{m-1}(d\theta) := \nu(d\theta)h(0|\theta)^{m-1}.$$

The atoms in this ordinary component have been selected precisely because they have not appeared in any of $X_1, \ldots, X_{m-1}$. As for $m = 1$, we have that the atoms θ_k in this ordinary component with corresponding weight in X_m equal to x are formed by a thinned Poisson point process, with rate measure

$$\nu_{m-1}(d\theta)h(x|\theta) = \nu(d\theta)h(0|\theta)^{m-1}h(x|\theta).$$

Since the rate measure has finite total mass by Assumption A2, we can define

$$M_{m,x} := \int_\theta \nu(d\theta)h(0|\theta)^{m-1}h(x|\theta),$$

which will be finite. Then the number of atoms θ_k for which $x_{1,k} = x$ is

$$\rho_{m,x} \sim \text{Poisson}(\rho|M_{m,x}).$$

And each such θ_k has weight

$$F_{size,m,x} \propto \nu(d\theta)h(0|\theta)^{m-1}h(x|\theta).$$

Finally, note from Theorem 3.1 that the posterior $\Theta|X_{1:m}$, which can be thought of as generated by prior $\Theta|X_{1:(m-1)}$ and likelihood $X_m|\Theta$, has weight rate measure

$$\nu(d\theta)h(0|\theta)^{m-1}h(0|\theta) = \nu_m(d\theta),$$

confirming the inductive hypothesis.

Recall that every atom of Θ is found in exactly one of these rounds and that $x \in \mathbb{Z}_+$. Also recall that the atom locations may be generated independently and identically across atoms, and independently from all the weights, according to proper distribution G (Section 2.2). To summarize, we have then

$$\Theta = \sum_{m=1}^{\infty}\sum_{x=1}^{\infty}\sum_{j=1}^{\rho_{m,x}} \theta_{m,x,j}\delta_{\psi_{m,x,j}},$$

where

$$\begin{aligned}\psi_{m,x,k} &\overset{iid}{\sim} G \text{ iid across } m, x, j\\ M_{m,x} &= \int_\theta \nu(d\theta)h(0|\theta)^{m-1}h(x|\theta) \text{ across } m, x\end{aligned}$$

$$\rho_{m,x} \overset{indep}{\sim} \text{Poisson}(\rho|M_{m,x}) \text{ across } m,x$$
$$F_{size,m,x}(d\theta) \propto \nu(d\theta)h(0|\theta)^{m-1}h(x|\theta) \text{ across } m,x$$
$$\theta_{m,x,j} \overset{indep}{\sim} F_{size,m,x}(d\theta) \text{ iid across } j \text{ and independently across } m,x,$$

as was to be shown. □

The following corollary gives a more detailed recipe for the calculations in Theorem 5.1 when the prior is in a conjugate exponential CRM to the likelihood.

Corollary 5.2 (Exponential CRM size-biased representations)**.** *Let Θ be an exponential CRM with no fixed-location atoms (thereby trivially satisfying Assumption A0) such that Assumption A1 holds.*

Let X be generated conditional on Θ according to an exponential CRM with fixed-location atoms at $\{\psi_k\}_{k=1}^{\infty}$ and no ordinary component. Let the distribution of the weight $x_{n,k}$ at ψ_k have probability mass function

$$h(x|\theta_k) = \kappa(x)\exp\left\{\langle\eta(\theta_k),\phi(x)\rangle - A(\theta_k)\right\}.$$

Suppose that Θ and X jointly satisfy Assumption A2. And let Θ be conjugate to X as in Theorem 4.2. Then we can write

$$\begin{aligned}
\Theta &= \sum_{m=1}^{\infty}\sum_{x=1}^{\infty}\sum_{j=1}^{\rho_{m,x}} \theta_{m,x,j}\delta_{\psi_{m,x,j}} \\
\psi_{m,x,j} &\overset{iid}{\sim} G \quad \textit{iid across } m,x,j \\
M_{m,x} &= \gamma\cdot\kappa(0)^{m-1}\kappa(x)\cdot\exp\left\{B(\xi+(m-1)\phi(0)+\phi(x),\lambda+m)\right\} \\
\rho_{m,x} &\overset{indep}{\sim} \text{Poisson}\left(\rho|M_{m,x}\right) \\
&\quad \textit{independently across } m,x \\
\theta_{m,x,j} &\overset{indep}{\sim} f_{size,m,x}(\theta)\,d\theta \\
&= \exp\left\{\langle\xi+(m-1)\phi(0)+\phi(x),\eta(\theta)\rangle + (\lambda+m)[-A(\theta)]\right. \\
&\qquad \left. - B(\xi+(m-1)\phi(0)+\phi(x),\lambda+m)\right\} \\
&\quad \textit{iid across } j \textit{ and independently across } m,x.
\end{aligned} \tag{28}$$

Proof. The corollary follows from Theorem 5.1 by plugging in the particular forms for $\nu(d\theta)$ and $h(x|\theta)$.

In particular,

$$\begin{aligned}
M_{m,x} &= \int_{\theta\in\mathbb{R}_+} \nu(d\theta)h(0|\theta)^{m-1}h(x|\theta) \\
&= \int_{\theta\in\mathbb{R}_+} \gamma\exp\left\{\langle\xi,\eta(\theta)\rangle + \lambda\left[-A(\theta)\right]\right\} \\
&\quad \cdot\left[\kappa(0)\exp\left\{\langle\eta(\theta),\phi(0)\rangle - A(\theta)\right\}\right]^{m-1} \\
&\quad \cdot\kappa(x)\exp\left\{\langle\eta(\theta),\phi(x)\rangle - A(\theta)\right\}\,d\theta
\end{aligned}$$

$$= \gamma\kappa(0)^{m-1}\kappa(x)\exp\left\{B\left(\xi + (m-1)\phi(0) + \phi(x), \lambda + m\right)\right\},$$

□

Corollary 5.2 can be used to find the known size-biased representation of the beta process [Thibaux and Jordan, 2007]; we demonstrate this derivation in detail in Example B.1 in Appendix B. Here we use Corollary 5.2 to discover a new size-biased representation of the gamma process.

Example 5.3. Let Θ be a gamma process, and let X_n be iid Poisson likelihood processes conditioned on Θ for each n as in Example 4.3. That is, we have

$$\nu(d\theta) = \gamma\theta^{\xi}e^{-\lambda\theta}\, d\theta.$$

And

$$h(x|\theta_k) = \frac{1}{x!}\theta_k^x e^{-\theta_k}$$

with

$$\gamma > 0, \quad \xi \in (-2, -1], \quad \lambda > 0; \quad \xi_{fix,k} > -1 \text{ and } \lambda_{fix,k} > 0 \quad \text{for all } k \in [K_{prior,fix}]$$

by Example 4.3.

We can pick out the following components of h:

$$\kappa(x) = \frac{1}{x!}, \quad \phi(x) = x, \quad \eta(\theta) = \log(\theta), \quad A(\theta) = \theta.$$

Thus, by Corollary 5.2, we have

$$f_{size,m,x}(\theta) \propto \theta^{\xi+x}e^{-(\lambda+m)\theta} \propto \text{Gamma}\left(\theta\,|\xi + x + 1, \lambda + m\right).$$

We summarize the representation that follows from Corollary 5.2 in the following result.

Corollary 5.4. *Let the gamma process be a CRM Θ with fixed-location atom weight distributions as in Eq. (19) and ordinary component weight measure as in Eq. (20). Then we may write*

$$\begin{aligned}
\Theta &= \sum_{m=1}^{\infty}\sum_{x=1}^{\infty}\sum_{j=1}^{\rho_{m,x}} \theta_{m,x,j}\delta_{\psi_{m,x,j}} \\
\psi_{m,x,j} &\overset{iid}{\sim} G \quad \textit{iid across } m, x, j \\
M_{m,x} &= \gamma \cdot \frac{1}{x!} \cdot \Gamma(\xi + x + 1) \cdot (\lambda + m)^{-(\xi+x+1)} \ \textit{across } m, x \\
\rho_{m,x} &\overset{indep}{\sim} \text{Poisson}\left(\rho|M_{m,x}\right) \ \textit{across } m, x \\
\theta_{m,x,j} &\overset{indep}{\sim} \text{Gamma}\left(\theta\,|\xi + x + 1, \lambda + m\right) \\
&\quad \textit{iid across } j \textit{ and independently across } m, x.
\end{aligned}$$

■

6 Marginal processes

In Section 5, although we conceptually made use of the observations $\{X_1, X_2, \ldots\}$, we focused on a representation of the prior Θ: cf. Eqs. (27) and (28). In this section, we provide a representation of the marginal of $X_{1:N}$, with Θ integrated out.

The canonical example of a marginal process again comes from the Dirichlet process (DP). In this case, the full model consists of the DP-distributed prior on Θ_{DP} (as in Eq. (24)) together with the likelihood for $X_{mult,n}$ conditional on Θ_{DP} (iid across n) described by Eq. (25). Then the marginal distribution of $X_{mult,1:N}$ is described by the *Chinese restaurant process*. This marginal takes the following form.

For each $n = 1, 2, \ldots, N$,

1. Let $\{\psi_k\}_{k=1}^{K_{n-1}}$ be the union of atom locations in $X_{mult,1}, \ldots, X_{mult,n-1}$. Then $X_{mult,n}|X_{mult,1}, \ldots, X_{mult,n-1}$ has a single atom at ψ, where

$$\psi = \begin{cases} \psi_k & \text{with probability} \propto \sum_{k=1}^{K_{n-1}} X_{mult,m}(\{\psi_k\}) \\ \psi_{new} & \text{with probability} \propto c \end{cases}$$
$$\psi_{new} \sim G$$

In the case of CRMs, the canonical example of a marginal process is the Indian buffet process [Griffiths and Ghahramani, 2006]. Both the Chinese restaurant process and Indian buffet process have proven popular for inference since the underlying infinite-dimensional prior is integrated out in these processes and only the finite-dimensional marginal remains. By Assumption A2, we know that the marginal will generally be finite-dimensional for our CRM Bayesian models. And thus we have the following general marginal representations for such models.

Theorem 6.1 (Marginal representations)**.** *Let Θ be a completely random measure that satisfies Assumptions A0 and A1; that is, Θ is a CRM with K_{fix} fixed atoms such that $K_{fix} < \infty$ and such that the kth atom can be written $\theta_{fix,k}\delta_{\psi_{fix,k}}$. The ordinary component of Θ has rate measure*

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

where G is a proper distribution and $\nu(\mathbb{R}_+) = \infty$. Write $\Theta = \sum_{k=1}^{\infty} \theta_k \delta_{\psi_k}$, and let X_n be generated iid given Θ according to $X_n = \sum_{k=1}^{\infty} x_{n,k}\delta_{\psi_k}$ with $x_{n,k} \overset{indep}{\sim} h(x|\theta_k)$ for proper, discrete probability mass function h. And suppose X_n and Θ jointly satisfy Assumption A2 so that

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

Then the marginal distribution of $X_{1:N}$ is the same as that provided by the following construction.

For each $n = 1, 2, \ldots, N$,

1. *Let* $\{\psi_k\}_{k=1}^{K_{n-1}}$ *be the union of atom locations in* $X_1, \ldots, X_{n-1}$. *Let* $x_{m,k} := X_m(\{\psi_k\})$. *Let* $x_{n,k}$ *denote the weight of* $X_n|X_1, \ldots, X_{n-1}$ *at* ψ_k. *Then* $x_{n,k}$ *has distribution described by the following probability mass function:*

$$h_{cond}\left(x_{n,k} = x \,\middle|\, x_{1:(n-1),k}\right) = \frac{\int_{\theta\in\mathbb{R}_+} \nu(d\theta)h(x|\theta)\prod_{m=1}^{n-1} h(x_{m,k}|\theta)}{\int_{\theta\in\mathbb{R}_+} \nu(d\theta)\prod_{m=1}^{n-1} h(x_{m,k}|\theta)}.$$

2. *For each* $x = 1, 2, \ldots$

   - X_n *has* $\rho_{n,x}$ *new atoms. That is,* X_n *has atoms at locations* $\{\psi_{n,x,j}\}_{j=1}^{\rho_{n,x}}$, *where*

   $$\{\psi_{n,x,j}\}_{j=1}^{\rho_{n,x}} \cap \{\psi_k\}_{k=1}^{K_{n-1}} = \emptyset \quad a.s.$$

   Moreover,

   $$\rho_{n,x} \overset{indep}{\sim} \text{Poisson}\left(\rho \left| \int_\theta \nu(d\theta)h(0|\theta)^{n-1}h(x|\theta)\right.\right) \text{ across } n, x$$

   $$\psi_{n,x,j} \overset{iid}{\sim} G(d\psi) \text{ across } n, x, j.$$

Proof. We saw in the proof of Theorem 5.1 that the marginal for X_1 can be expressed as follows. For each $x \in \mathbb{Z}_+$, there are $\rho_{1,x}$ atoms of X_1 with weight x, where

$$\rho_{1,x} \overset{indep}{\sim} \text{Poisson}\left(\int_\theta \nu(d\theta)h(x|\theta)\right) \text{ across } x.$$

These atoms have locations $\{\psi_{1,x,j}\}_{j=1}^{\rho_{1,x}}$, where

$$\psi_{1,x,j} \overset{iid}{\sim} G(d\psi) \text{ across } x, j.$$

For the upcoming induction, let $K_1 := \sum_{x=1}^{\infty} \rho_{1,x}$. And let $\{\psi_k\}_{k=1}^{K_1}$ be the (a.s. disjoint by assumption) union of the sets $\{\psi_{1,x,j}\}_{j=1}^{\rho_{1,x}}$ across x. Note that K_1 is finite by Assumption A2.

We will also find it useful in the upcoming induction to let $\Theta_{post,1}$ have the distribution of $\Theta|X_1$. Let $\theta_{post,1,x,j} = \Theta_{post,1}(\{\psi_{1,x,j}\})$. By Theorem 3.1 or the proof of Theorem 5.1, we have that

$$\theta_{post,1,x,j} \overset{indep}{\sim} F_{post,1,x,j}(d\theta) \propto \nu(d\theta)h(x|\theta)$$
$$\text{independently across } x \text{ and iid across } j.$$

Now take any $n > 1$. Inductively, we assume $\{\psi_{n-1,k}\}_{k=1}^{K_{n-1}}$ is the union of all the atom locations of $X_1, \ldots, X_{n-1}$. Further assume K_{n-1} is finite. Let $\Theta_{post,n-1}$ have the distribution of $\Theta|X_1, \ldots, X_{n-1}$. Let $\theta_{n-1,k}$ be the weight of $\Theta_{post,n-1}$ at $\psi_{n-1,k}$. And, for any $m \in [n-1]$, let $x_{m,k}$ be the weight of X_m at $\psi_{n-1,k}$. We inductively assume that

$$\theta_{n-1,k} \overset{indep}{\sim} F_{n-1,k}(d\theta) \propto \nu(d\theta)\prod_{m=1}^{n-1} h(x_{m,k}|\theta) \tag{29}$$
$$\text{independently across } k.$$

Now let $\psi_{n,k}$ equal $\psi_{n-1,k}$ for $k \in [K_{n-1}]$. Let $x_{n,k}$ denote the weight of X_n at $\psi_{n,k}$ for $k \in [K_{n-1}]$. Conditional on the atom weight of Θ at $\psi_{n,k}$, the atom weights of $X_1, \ldots, X_{n-1}, X_n$ are independent. Since the atom weights of Θ are independent as well, we have that $x_{n,k}|X_1, \ldots, X_{n-1}$ has the same distribution as $x_{n,k}|x_{1,k}, \ldots, x_{n-1,k}$. We can write the probability mass function of this distribution as follows.

$$\begin{aligned}
&h_{cond}\left(x_{n,k} = x \,|x_{1,k}, \ldots, x_{n-1,k}\right) \\
&= \int_{\theta \in \mathbb{R}_+} F_{n-1,k}(d\theta) h(x|\theta) \\
&= \frac{\int_{\theta \in \mathbb{R}_+} \left[\nu(d\theta) \prod_{m=1}^{n-1} h(x_{m,k}|\theta)\right] \cdot h(x|\theta)}{\int_{\theta \in \mathbb{R}_+} \nu(d\theta) \prod_{m=1}^{n-1} h(x_{m,k}|\theta)},
\end{aligned}$$

where the last line follows from Eq. (29).

We next show the inductive hypothesis in Eq. (29) holds for n and $k \in [K_{n-1}]$. Let $x_{n,k}$ denote the weight of X_n at $\psi_{n,k}$ for $k \in [K_{n-1}]$. Let $F_{n,k}(d\theta)$ denote the distribution of $x_{n,k}$ and note that

$$\begin{aligned}
F_{n,k}(d\theta) &\propto F_{n-1,k}(d\theta) \cdot h(x_{n,k}|\theta) \\
&= \nu(d\theta) \prod_{m=1}^{n} h(x_{m,k}|\theta),
\end{aligned}$$

which agrees with Eq. (29) for n when we assume the result for $n-1$.

The previous development covers atoms that are present in at least one of $X_1, \ldots, X_{n-1}$. Next we consider new atoms in X_n; that is, we consider atoms in X_n for which there are no atoms at the same location in any of $X_1, \ldots, X_{n-1}$.

We saw in the proof of Theorem 5.1 that, for each $x \in \mathbb{Z}_+$, there are $\rho_{n,x}$ new atoms of X_n with weight x such that

$$\rho_{n,x} \overset{indep}{\sim} \text{Poisson}\left(\rho \,\middle|\, \int_\theta \nu(d\theta) h(0|\theta)^{n-1} h(x|\theta)\right) \text{ across } x.$$

These new atoms have locations $\{\psi_{n,x,j}\}_{j=1}^{\rho_{n,x}}$ with

$$\psi_{n,x,j} \overset{iid}{\sim} G(d\psi) \text{ across } x, j.$$

By Assumption A2, $\sum_{x=1}^{\infty} \rho_{n,x} < \infty$. So

$$K_n := K_{n-1} + \sum_{x=1}^{\infty} \rho_{n,x}$$

remains finite. Let $\psi_{n,k}$ for $k \in \{K_{n-1}+1, \ldots, K_n\}$ index these new locations. Let $\theta_{n,k}$ be the weight of $\Theta_{post,n}$ at $\psi_{n,k}$ for $k \in \{K_{n-1}+1, \ldots, K_n\}$. And let $x_{n,k}$ be the value of X at $\psi_{n,k}$.

We check that the inductive hypothesis holds. By repeated application of Theorem 3.1, the ordinary component of $\Theta|X_1,\ldots,X_{n-1}$ has rate measure

$$\nu(d\theta)h(0|\theta)^{n-1}.$$

So, again by Theorem 3.1, we have that

$$\theta_{n,k} \overset{indep}{\sim} F_{n.k}(d\theta) \propto \nu(d\theta)h(0|\theta)^{n-1}h(x_{n,k}|\theta).$$

Since X_m has value 0 at $\psi_{n,k}$ for $m \in \{1,\ldots,n-1\}$ by construction, we have that the inductive hypothesis holds. □

As in the case of size-biased representations (Section 5 and Corollary 5.2), we can find a more detailed recipe when the prior is in a conjugate exponential CRM to the likelihood.

Corollary 6.2 (Exponential CRM marginal representations)**.** *Let Θ be an exponential CRM with no fixed-location atoms (thereby trivially satisfying Assumption A0) such that Assumption A1 holds.*

Let X be generated conditional on Θ according to an exponential CRM with fixed-location atoms at $\{\psi_k\}_{k=1}^{\infty}$ and no ordinary component. Let the distribution of the weight $x_{n,k}$ at ψ_k have probability mass function

$$h(x|\theta_k) = \kappa(x)\exp\left\{\langle\eta(\theta_k),\phi(x)\rangle - A(\theta_k)\right\}.$$

Suppose that Θ and X jointly satisfy Assumption A2. And let Θ be conjugate to X as in Theorem 4.2. Then the marginal distribution of $X_{1:N}$ is the same as that provided by the following construction.

For each $n = 1,2,\ldots,N$,

1. *Let $\{\psi_k\}_{k=1}^{K_{n-1}}$ be the union of atom locations in $X_1,\ldots,X_{n-1}$. Let $x_{m,k} := X_m(\{\psi_k\})$. Let $x_{n,k}$ denote the weight of $X_n|X_1,\ldots,X_{n-1}$ at ψ_k. Then $x_{n,k}$ has distribution described by the following probability mass function:*

$$\begin{aligned}&h_{cond}\left(x_{n,k} = x \,\middle|\, x_{1:(n-1),k}\right)\\ &= \kappa(x)\exp\left\{-B(\xi+\sum_{m=1}^{n-1}x_m,\lambda+n-1) + B(\xi+\sum_{m=1}^{n-1}x_m+x,\lambda+n)\right\}.\end{aligned}$$

2. *For each $x = 1,2,\ldots$*

   - *X_n has $\rho_{n,x}$ new atoms. That is, X_n has atoms at locations $\{\psi_{n,x,j}\}_{j=1}^{\rho_{n,x}}$, where*

$$\{\psi_{n,x,j}\}_{j=1}^{\rho_{n,x}} \cap \{\psi_k\}_{k=1}^{K_{n-1}} = \emptyset \quad a.s.$$

   Moreover,

$$\begin{aligned}M_{n,x} &:= \gamma\cdot\kappa(0)^{n-1}\kappa(x)\cdot\exp\left\{B(\xi+(n-1)\phi(0)+\phi(x),\lambda+n)\right\}\\ &\quad across\ n,x\\ \rho_{n,x} &\overset{indep}{\sim} \text{Poisson}\left(\rho\,|M_{n,x}\right)\ across\ n,x\\ \psi_{n,x,j} &\overset{iid}{\sim} G(d\psi)\ across\ n,x,j.\end{aligned}$$

Proof. The corollary follows from Theorem 6.1 by plugging in the forms for $\nu(d\theta)$ and $h(x|\theta)$.

In particular,

$$\begin{aligned}
&\int_{\theta\in\mathbb{R}_+} \nu(d\theta) \prod_{m=1}^{n} h(x_{m,k}|\theta) \\
&= \int_{\theta\in\mathbb{R}_+} \gamma \exp\left\{\langle \xi, \eta(\theta)\rangle + \lambda\left[-A(\theta)\right]\right\} \cdot \left[\prod_{m=1}^{n} \kappa(x_{m,k}) \exp\left\{\langle \eta(\theta), \phi(x_{m,k})\rangle - A(\theta)\right\}\right] \\
&= \gamma \left[\prod_{m=1}^{n} \kappa(x_{m,k})\right] B\left(\xi + \sum_{m=1}^{n} \phi(x_{m,k}), \lambda + n\right).
\end{aligned}$$

So

$$\begin{aligned}
&h_{cond}\left(x_{n,k} = x \,\middle|\, x_{1:(n-1),k}\right) \\
&= \frac{\int_{\theta\in\mathbb{R}_+} \nu(d\theta) h(x|\theta) \prod_{m=1}^{n-1} h(x_{m,k}|\theta)}{\int_{\theta\in\mathbb{R}_+} \nu(d\theta) \prod_{m=1}^{n-1} h(x_{m,k}|\theta)} \\
&= \kappa(x) \exp\left\{-B(\xi + \sum_{m=1}^{n-1} x_m, \lambda + n - 1) + B(\xi + \sum_{m=1}^{n-1} x_m + x, \lambda + n)\right\}.
\end{aligned}$$

□

In Example C.1 in Appendix C we show that Corollary 6.2 can be used to recover the Indian buffet process marginal from a beta process prior together with a Bernoulli process likelihood. In the following example, we discover a new marginal for the Poisson likelihood process with gamma process prior.

Example 6.3. Let Θ be a gamma process, and let X_n be iid Poisson likelihood processes conditioned on Θ for each n as in Example 4.3. That is, we have

$$\nu(d\theta) = \gamma\theta^{\xi} e^{-\lambda\theta}\, d\theta \quad \text{and} \quad h(x|\theta_k) = \frac{1}{x!}\theta_k^x e^{-\theta_k}$$

with

$$\gamma > 0, \quad \xi \in (-2, -1], \quad \lambda > 0; \quad \xi_{fix,k} > -1 \text{ and } \lambda_{fix,k} > 0 \quad \text{for all } k \in [K_{prior,fix}]$$

by Example 4.3.

We can pick out the following components of h:

$$\kappa(x) = \frac{1}{x!}, \quad \phi(x) = x, \quad \eta(\theta) = \log(\theta), \quad A(\theta) = \theta.$$

And we calculate

$$\exp\left\{B(\xi, \lambda)\right\} = \int_{\theta\in\mathbb{R}_+} \exp\left\{\langle \xi, \eta(\theta)\rangle + \lambda[-A(\theta)]\right\}\, d\theta = \int_{\theta\in\mathbb{R}_+} \theta^{\xi} e^{-\lambda\theta} = \Gamma(\xi + 1)\lambda^{-(\xi+1)}.$$

So, for $k \in \mathbb{Z}_*$, we have

$$\begin{aligned}
\mathbb{P}(x_n = x) &= \kappa(x) \exp\left\{ -B(\xi + \sum_{m=1}^{n-1} x_m, \lambda + n - 1) + B(\xi + \sum_{m=1}^{n-1} x_m + x, \lambda + n) \right\} \\
&= \frac{1}{x!} \cdot \frac{(\lambda + n - 1)^{\xi + \sum_{m=1}^{n-1} x_m + 1}}{\Gamma(\xi + \sum_{m=1}^{n-1} x_m + 1)} \cdot \frac{\Gamma(\xi + \sum_{m=1}^{n-1} x_m + x + 1)}{(\lambda + n)^{\xi + \sum_{m=1}^{n-1} x_m + x + 1}} \\
&= \frac{\Gamma(\xi + \sum_{m=1}^{n-1} x_m + x + 1)}{\Gamma(x+1)\Gamma(\xi + \sum_{m=1}^{n-1} x_m + 1)} \cdot \left(\frac{\lambda + n - 1}{\lambda + n} \right)^{\xi + \sum_{m=1}^{n} x_m + 1} \left(\frac{1}{\lambda + n} \right)^{x} \\
&= \text{NegBin}\left(x \,\middle|\, \xi + \sum_{m=1}^{n-1} x_m + 1, (\lambda + n)^{-1} \right).
\end{aligned}$$

And

$$\begin{aligned}
M_{n,x} &:= \gamma \cdot \kappa(0)^{n-1} \kappa(x) \cdot \exp\left\{ B(\xi + (n-1)\phi(0) + \phi(x), \lambda + n) \right\} \\
&= \gamma \cdot \frac{1}{x!} \cdot \Gamma(\xi + x + 1)(\lambda + n)^{-(\xi + x + 1)}.
\end{aligned}$$

We summarize the marginal distribution representation of $X_{1:N}$ that follows from Corollary 6.2 in the following result.

Corollary 6.4. *Let Θ be a gamma process with fixed-location atom weight distributions as in Eq. (19) and ordinary component weight measure as in Eq. (20). Let X_n be drawn, iid across n, conditional on Θ according to a Poisson likelihood process with fixed-location atom weight distributions as in Eq. (18). Then $X_{1:N}$ has the same distribution as the following construction.*

For each $n = 1, 2, \ldots, N$,

1. *Let $\{\psi_k\}_{k=1}^{K_{n-1}}$ be the union of atom locations in $X_1, \ldots, X_{n-1}$. Let $x_{m,k} := X_m(\{\psi_k\})$. Let $x_{n,k}$ denote the weight of $X_n | X_1, \ldots, X_{n-1}$ at ψ_k. Then $x_{n,k}$ has distribution described by the following probability mass function:*

$$h_{cond}\left(x_{n,k} = x \,\middle|\, x_{1:(n-1),k}\right) = \text{NegBin}\left(x \,\middle|\, \xi + \sum_{m=1}^{n-1} x_{m,k} + 1, (\lambda + n)^{-1} \right).$$

2. *For each $x = 1, 2, \ldots$*

   - *X_n has $\rho_{n,x}$ new atoms. That is, X_n has atoms at locations $\{\psi_{n,x,j}\}_{j=1}^{\rho_{n,x}}$, where*

$$\{\psi_{n,x,j}\}_{j=1}^{\rho_{n,x}} \cap \{\psi_k\}_{k=1}^{K_{n-1}} = \emptyset \quad a.s.$$

   Moreover,

$$M_{n,x} := \gamma \cdot \frac{1}{x!} \cdot \frac{\Gamma(\xi + x + 1)}{(\lambda + n)^{\xi + x + 1}}$$

   across n, x

$$\rho_{n,x} \overset{indep}{\sim} \text{Poisson}\left(\rho \,|M_{n,x}\right) \;\; \textit{independently across } n, x$$
$$\psi_{n,x,j} \overset{iid}{\sim} G(d\psi) \;\; \textit{independently across } n, x \textit{ and iid across } j.$$

■

7 Discussion

In the preceding sections, we have shown how to calculate posteriors for general CRM-based priors and likelihoods for Bayesian nonparametric models. We have also shown how to represent Bayesian nonparametric priors as a sequence of finite draws, and full Bayesian nonparametric models via finite marginals. We have introduced a notion of exponential families for CRMs, which we call exponential CRMs, that has allowed us to specify automatic Bayesian nonparametric conjugate priors for exponential CRM likelihoods. And we have demonstrated that our exponential CRMs allow particularly straightforward recipes for size-biased and marginal representations of Bayesian nonparametric models. Along the way, we have proved that the gamma process is a conjugate prior for the Poisson likelihood process and the beta prime process is a conjugate prior for the odds Bernoulli process. We have discovered a size-biased representation of the gamma process and a marginal representation of the gamma process coupled with a Poisson likelihood process.

All of this work has relied heavily on the description of Bayesian nonparametric models in terms of completely random measures. As such, we have worked very particularly with pairings of real values—the CRM atom weights, which we have interpreted as trait frequencies or rates—together with trait descriptors—the CRM atom locations. However, all of our proofs broke into essentially two parts: the fixed-location atom part and the ordinary component part. The fixed-location atom development essentially translated into the usual finite version of Bayes Theorem and could easily be extended to full Bayesian models where the prior describes a random element that need not be real-valued. Moreover, the ordinary component development relied entirely on its generation as a Poisson point process over a product space. It seems reasonable to expect that our development might carry through when the first element in this tuple need not be real-valued. And thus we believe our results are suggestive of broader results over more general spaces.

Acknowledgements

Support for this project was provided by ONR under the Multidisciplinary University Research Initiative (MURI) program (N00014-11-1-0688). T. Broderick was supported by a Berkeley Fellowship. A. C. Wilson was supported by an NSF Graduate Research Fellowship.

A Further automatic conjugate priors

We use Theorem 4.2 to calculate automatic conjugate priors for further exponential CRMs.

Example A.1. Let X be generated according to a Bernoulli process as in Example 2.1. That is, X has an exponential CRM distribution with $K_{like,fix}$ fixed-location atoms, where $K_{like,fix} < \infty$ in accordance with Assumption A0:

$$X = \sum_{k=1}^{K_{like,fix}} x_{like,k}\delta_{\psi_{like,k}}.$$

The weight of the kth atom, $x_{like,k}$, has support on $\{0,1\}$ and has a Bernoulli density with parameter $\theta_k \in (0,1]$:

$$\begin{aligned} h(x|\theta_k) &= \theta_k^x(1-\theta_k)^{1-x} \\ &= \exp\left\{x\log(\theta_k/(1-\theta_k)) + \log(1-\theta_k)\right\}. \end{aligned}$$

The final line is rewritten to emphasize the exponential family form of this density, with

$$\begin{aligned} \kappa(x) &= 1 \\ \phi(x) &= x \\ \eta(\theta) &= \log\left(\frac{\theta}{1-\theta}\right) \\ A(\theta) &= -\log(1-\theta). \end{aligned}$$

Then, by Theorem 4.2, X has a Bayesian nonparametric conjugate prior for

$$\Theta := \sum_{k=1}^{K_{like,fix}} \theta_k\delta_{\psi_k}.$$

This conjugate prior has two parts.

First, Θ has a set of $K_{prior,fix}$ fixed-location atoms at some subset of the $K_{like,fix}$ fixed locations of X. The kth such atom has random weight $\theta_{fix,k}$ with density

$$\begin{aligned} f_{prior,fix,k}(\theta) &= \exp\left\{\langle\xi_{fix,k},\eta(\theta)\rangle + \lambda_{fix,k}\left[-A(\theta)\right] - B(\xi_{fix,k},\lambda_{fix,k})\right\} \\ &= \theta^{\xi_{fix,k}}(1-\theta)^{\lambda_{fix,k}-\xi_{fix,k}}\exp\left\{-B(\xi_{fix,k},\lambda_{fix,k})\right\} \\ &= \text{Beta}\left(\theta\,|\xi_{fix,k}+1, \lambda_{fix,k}-\xi_{fix,k}+1\right), \end{aligned}$$

where $\text{Beta}(\theta|a,b)$ denotes the beta density with shape parameters $a > 0$ and $b > 0$. So we must have fixed hyperparameters $\xi_{fix,k} > -1$ and $\lambda_{fix,k} > \xi_{fix,k} - 1$. Further,

$$\exp\left\{-B(\xi_{fix,k},\lambda_{fix,k})\right\} = \frac{\Gamma(\lambda_{fix,k}+2)}{\Gamma(\xi_{fix,k}+1)\Gamma(\lambda_{fix,k}-\xi_{fix,k}+1)}$$

to ensure normalization.

Second, Θ has an ordinary component characterized by any proper distribution G and weight rate measure

$$\begin{aligned}\nu(d\theta) &= \gamma \exp\left\{\langle \xi, \eta(\theta)\rangle + \lambda\left[-A(\theta)\right]\right\}\, d\theta \\ &= \gamma\theta^{\xi}(1-\theta)^{\lambda-\xi}\, d\theta.\end{aligned}$$

Finally, we need to choose the allowable hyperparameter ranges for γ, ξ, and λ. $\gamma > 0$ ensures ν is a measure. By Assumption A1, we must have $\nu(\mathbb{R}_+) = \infty$, so ν must represent an improper beta distribution. As such, we require either $\xi + 1 \leq 0$ or $\lambda - \xi \leq 0$. By Assumption A2, we must have

$$\begin{aligned}&\sum_{x=1}^{\infty} \int_{\theta\in\mathbb{R}_+} \nu(d\theta)\cdot h(x|\theta) \\ &= \int_{\theta\in(0,1]} \nu(d\theta) h(1|\theta) \\ &\text{since the support of } x \text{ is } \{0,1\} \text{ and the support of } \theta \text{ is } (0,1] \\ &= \gamma \int_{\theta\in(0,1]} \theta^{\xi}(1-\theta)^{\lambda-\xi}\, d\theta\cdot\theta \\ &< \infty\end{aligned}$$

Since the integrand is the kernel of a beta distribution, the integral is finite if and only if $\xi + 2 > 0$ and $\lambda - \xi + 1 > 0$.

Finally, then the hyperparameter restrictions can be summarized as:

$$\begin{aligned}&\gamma > 0 \\ &\xi \in (-2, -1] \\ &\lambda > \xi - 1 \\ &\xi_{fix,k} > -1 \text{ and } \lambda_{fix,k} > \xi_{fix,k} - 1 \quad \text{for all } k \in [K_{prior,fix}]\end{aligned}$$

By setting $\alpha = \xi+1$, $c = \lambda+2$, $\rho_{fix,k} = \xi_{fix,k}+1$, and $\sigma_{fix,k} = \lambda_{fix,k}-\xi_{fix,k}+1$, we recover the hyperparameters of Eq. (11) in Example 2.1. Here, by contrast to Example 2.1, we found the conjugate prior and its hyperparameter settings given just the Bernoulli process likelihood. Henceforth, we use the parameterization of the beta process above. ■

B Further size-biased representations

Example B.1. Let Θ be a beta process, and let X_n be iid Bernoulli processes conditioned on Θ for each n as in Example A.1. That is, we have

$$\nu(d\theta) = \gamma\theta^{\xi}(1-\theta)^{\lambda-\xi}\, d\theta.$$

And

$$h(x|\theta_k) = \theta_k^x(1-\theta_k)^{1-x}$$

with

$$\begin{aligned}
&\gamma > 0 \\
&\xi \in (-2, -1] \\
&\lambda > \xi - 1 \\
&\xi_{fix,k} > -1 \text{ and } \lambda_{fix,k} > \xi_{fix,k} - 1 \quad \text{for all } k \in [K_{prior,fix}]
\end{aligned}$$

by Example A.1.

We can pick out the following components of h:

$$\begin{aligned}
\kappa(x) &= 1 \\
\phi(x) &= x \\
\eta(\theta) &= \log\left(\frac{\theta}{1-\theta}\right) \\
A(\theta) &= -\log(1-\theta).
\end{aligned}$$

Thus, by Corollary 5.2,

$$\begin{aligned}
\Theta &= \sum_{m=1}^{\infty} \sum_{x=1}^{\infty} \sum_{j=1}^{\rho_{m,x}} \theta_{m,x,j} \delta_{\psi_{m,x,j}} \\
\psi_{m,x,j} &\overset{iid}{\sim} G \quad \text{iid across } m, x, j \\
\theta_{m,x,j} &\overset{indep}{\sim} f_{size,m,x}(\theta)\, d\theta \\
&\propto \theta^{\xi+x}(1-\theta)^{\lambda+m-\xi-x}\, d\theta \\
&\propto \text{Beta}\left(\theta \,|\xi + x, \lambda - \xi + m - x\right)\, d\theta \\
&\quad \text{iid across } j \text{ and independently across } m, x \\
M_{m,x} &:= \gamma \cdot \frac{\Gamma(\xi + x + 1)\Gamma(\lambda - \xi + m - x + 1)}{\Gamma(\lambda + m + 2)} \\
\rho_{m,x} &\overset{indep}{\sim} \text{Poisson}\left(M_{m,x}\right) \\
&\quad \text{across } m, x
\end{aligned}$$

Broderick et al. [2012] and Paisley et al. [2012] have previously noted that this size-biased representation of the beta process arises from the Poisson point process. ■

C Further marginals

Example C.1. Let Θ be a beta process, and let X_n be iid Bernoulli processes conditioned on Θ for each n as in Examples A.1 and B.1.

We calculate the main components of Corollary 6.2 for this pair of processes. In particular, we have

$$\mathbb{P}(x_n = 1) = \kappa(k) \exp\left\{-B(\xi + \sum_{m=1}^{n-1} x_m, \lambda + n - 1) + B(\xi + \sum_{m=1}^{n-1} x_m + 1, \lambda + n)\right\}$$

$$
\begin{aligned}
&= \frac{\Gamma(\lambda + n - 1 + 2)}{\Gamma(\xi + \sum_{m=1}^{n-1} x_m + 1)\Gamma(\lambda + n - 1 - \xi - \sum_{m=1}^{n-1} x_m + 1)} \\
&\quad \cdot \frac{\Gamma(\xi + \sum_{m=1}^{n-1} x_m + 1 + 1)\Gamma(\lambda + n - \xi - \sum_{m=1}^{n-1} x_m - 1 + 1)}{\Gamma(\lambda + n + 2)} \\
&= \frac{\xi + \sum_{m=1}^{n-1} x_m + 1}{\lambda + n + 1}
\end{aligned}
$$

And

$$
\begin{aligned}
M_{n,1} &:= \gamma \cdot \kappa(0)^{n-1}\kappa(1) \cdot \exp\left\{B(\xi + (n-1)\phi(0) + \phi(1), \lambda + n)\right\} \\
&= \gamma \cdot \frac{\Gamma(\xi + 1 + 1)\Gamma(\lambda + n - \xi - 1 + 1)}{\Gamma(\lambda + n + 2)}
\end{aligned}
$$

Thus, the marginal distribution of $X_{1:N}$ is the same as that provided by the following construction.

For each $n = 1, 2, \ldots, N$,

1. At any location ψ for which there is some atom in $X_1, \ldots, X_{n-1}$, let x_m be the weight of X_m at ψ for $m \in [n-1]$. Then we have that $X_n | X_1, \ldots, X_{n-1}$ has weight x_n at ψ, where

$$
\mathbb{P}(dx_n) = \text{Bern}\left(x_n \,\middle|\, \frac{\xi + \sum_{m=1}^{n-1} x_m + 1}{\lambda + n + 1}\right)
$$

2. X_n has $\rho_{n,1}$ atoms at locations $\{\psi_{n,1,j}\}$ with $j \in [\rho_{n,1}]$ where there have not yet been atoms in any of $X_1, \ldots, X_{n-1}$. Moreover,

$$
\begin{aligned}
M_{n,1} &:= \gamma \cdot \frac{\Gamma(\xi + 1 + 1)\Gamma(\lambda + n - \xi - 1 + 1)}{\Gamma(\lambda + n + 2)} \\
&\quad \text{across } n \\
\rho_{n,1} &\overset{indep}{\sim} \text{Poisson}\left(M_{n,1}\right) \text{ across } n, x \\
\psi_{n,1,j} &\overset{iid}{\sim} G(d\psi) \text{ across } n, j
\end{aligned}
$$

Here, we have recovered the three-parameter extension of the Indian buffet process [Teh and Görür, 2009, Broderick et al., 2013]. ■

References

E. M. Airoldi, D. Blei, E. A. Erosheva, and S. E. Fienberg. *Handbook of Mixed Membership Models and Their Applications*. CRC Press, 2014.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

T. Broderick, M. I. Jordan, and J. Pitman. Beta processes, stick-breaking, and power laws. *Bayesian Analysis*, 7(2):439–476, 2012.

T. Broderick, M. I. Jordan, and J. Pitman. Cluster and feature modeling from combinatorial stochastic processes. *Statistical Science*, 2013.

T. Broderick, L. Mackey, J. Paisley, and M. I. Jordan. Combinatorial clustering and the beta negative binomial process. *IEEE TPAMI*, 2015.

P. Damien, J. Wakefield, and S. Walker. Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B*, 61(2):331–344, 1999.

M. H. DeGroot. *Optimal Statistical Decisions*. John Wiley & Sons, Inc, 1970.

P. Diaconis and D. Ylvisaker. Conjugate priors for exponential families. *The Annals of Statistics*, 7(2):269–281, 1979.

K. Doksum. Tailfree and neutral random probabilities and their posterior distributions. *The Annals of Probability*, pages 183–201, 1974.

F. Doshi, K. T. Miller, J. Van Gael, and Y. W. Teh. Variational inference for the Indian buffet process. In *AISTATS*, 2009.

M. D. Escobar. Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277, 1994.

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

M. D. Escobar and M. West. Computing nonparametric hierarchical models. In *Practical nonparametric and semiparametric Bayesian statistics*, pages 1–22. Springer, 1998.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

T. S. Ferguson. Prior distributions on spaces of probability measures. *The Annals of Statistics*, pages 615–629, 1974.

T. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *NIPS*, 2006.

N. L. Hjort. Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, pages 1259–1294, 1990.

H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453), 2001.

L. F. James. Poisson latent feature calculus for generalized Indian buffet processes. *arXiv preprint arXiv:1411.2936*, 2014.

L. F. James, A. Lijoi, and I. Prünster. Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36(1):76–97, 2009.

M. Kalli, J. E. Griffin, and S. G. Walker. Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105, 2011.

Y. Kim. Nonparametric Bayesian estimators for counting processes. *Annals of Statistics*, pages 562–588, 1999.

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

J. F. C. Kingman. *Poisson Processes*, volume 3. Oxford University Press, 1992.

A. Lijoi and I. Prünster. Models beyond the Dirichlet process. In N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors, *Bayesian Nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics, 2010.

A. Y. Lo. Bayesian nonparametric statistical inference for Poisson point processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 59 (1):55–66, 1982.

A. Y. Lo. On a class of Bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics*, 12(1):351–357, 1984.

S. N. MacEachern. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics-Simulation and Computation*, 23 (3):727–741, 1994.

R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

R. M. Neal. Slice sampling. *Annals of Statistics*, pages 705–741, 2003.

P. Orbanz. Conjugate projective limits. *arXiv preprint arXiv:1012.0363*, 2010.

J. W. Paisley, A. K. Zaas, C. W. Woods, G. S. Ginsburg, and L. Carin. A stick-breaking construction of the beta process. In *ICML*, pages 847–854, 2010.

J. W. Paisley, L. Carin, and D. M. Blei. Variational inference for stick-breaking beta process priors. In *ICML*, pages 889–896, 2011.

J. W. Paisley, D. M. Blei, and M. I. Jordan. Stick-breaking beta processes and the Poisson process. In *AISTATS*, pages 850–858, 2012.

M. Perman, J. Pitman, and M. Yor. Size-biased sampling of poisson point processes and excursions. *Probability Theory and Related Fields*, 92(1):21–39, 1992.

J. Pitman. Random discrete distributions invariant under size-biased permutation. *Advances in Applied Probability*, pages 525–539, 1996a.

J. Pitman. Some developments of the Blackwell-MacQueen urn scheme. *Lecture Notes-Monograph Series*, pages 245–267, 1996b.

J. Pitman. Poisson-Kingman partitions. *Lecture Notes-Monograph Series*, pages 1–34, 2003.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

Y. W. Teh and D. Görür. Indian buffet processes with power-law behavior. In *NIPS*, pages 1838–1846, 2009.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006.

Y. W. Teh, D. Görür, and Z. Ghahramani. Stick-breaking construction for the Indian buffet process. In *AISTATS*, pages 556–563, 2007.

R. Thibaux and M. I. Jordan. Hierarchical beta processes and the Indian buffet process. In *AISTATS*, pages 564–571, 2007.

R. J. Thibaux. *Nonparametric Bayesian Models for Machine Learning*. PhD thesis, UC Berkeley, 2008.

M. K. Titsias. The infinite gamma-Poisson feature model. In *NIPS*, pages 1513–1520, 2008.

S. G. Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics—Simulation and Computation*, 36(1):45–54, 2007.

C. Wang and D. M. Blei. Variational inference in nonconjugate models. *The Journal of Machine Learning Research*, 14(1):1005–1031, 2013.

M. West and M. D. Escobar. Hierarchical priors and mixture models, with application in regression and density estimation. In P. R. Freeman and A. F. M. Smith, editors, *Aspects of Uncertainty: A Tribute to D. V. Lindley*. Institute of Statistics and Decision Sciences, Duke University, 1994.

M. Zhou, L. Hannah, D. Dunson, and L. Carin. Beta-negative binomial process and Poisson factor analysis. *AISTATS*, 2012.