Bayesian Nonparametrics Rough Notes

Yee Whye Teh
Gatsby Computational Neuroscience Unit
University College London
17 Queen Square
London WC1N 3AR, United Kingdom
ywteh@gatsby.ucl.ac.uk

March 2, 2012

Contents

1 Random Partitions

For a positive integer n, we will denote by $[n]$ the set $\{1, \ldots, n\}$. We will consider partitions of $[n]$, that is, a set of disjoint non-empty subsets of $[n]$ whose union is $[n]$. Let $\mathbf{\Pi}_{[n]}$ denote the set of partititons of $[n]$. In general, for a countable set S, $\mathbf{\Pi}_S$ denotes the partitions of S. Let Π_n be a random partition (random variable taking values in $\mathbf{\Pi}_{[n]}$).

Let $\pi \in \mathbf{\Pi}_{[n]}$ by a partition. Given a permutation ϕ of $[n]$, let $\phi(\pi)$ be the partition obtained by permuting the members of clusters in π by ϕ. Given a subset $S \subset [n]$, let $J_S(\pi)$ be the partition of S obtained by removing all the members of clusters in π which are not in S.

$$\pi = \{\{1,3,5\},\{2,4\},\{6,7,8\},\{9\}\} \tag{1.1}$$

$$\phi = (123)(489)(567) \qquad \phi(\pi) = \{\{2,1,6\},\{3,8\},\{7,5,9\},\{4\}\} \tag{1.2}$$

$$S = \{1,3,4,5,7,8\} \qquad J_S(\pi) = \{\{1,3,5\},\{4\},\{7,8\}\} \tag{1.3}$$

Definition 1.4 (Exchangeability)**.** *Π_n is called exchangeable if its distribution is invariant to permutations ϕ of $[n]$:*

$$\mathbb{P}(\Pi_n = \pi_n) = \mathbb{P}(\Pi_n = \phi(\pi_n)) \tag{1.5}$$

Exchangeability implies that $\mathbb{P}(\Pi_n = \pi_n)$ only depends on the number of clusters and the sizes of the clusters in π_n. We will denote the number of clusters of π_n by k and the sizes of the clusters by $n_1, \ldots, n_k$. Then the probability function can be written as

$$\mathbb{P}(\Pi_n = \pi_n) = f_n(n_1, \ldots, n_k) \tag{1.6}$$

where f_n is some symmetry function of its arguments, which have to satisfy

$$n_j \geq 1 \qquad \sum_{j=1}^{k} n_j = n \tag{1.7}$$

Definition 1.8 (Projectivity)**.** *A sequence of random partitions $\Pi_1, \Pi_2, \ldots$ of $[1], [2], \ldots$ respectively is called projective if*

$$\mathbb{P}(J_{[m]}(\Pi_n) = \pi_m) = \mathbb{P}(\Pi_m = \pi_m) \tag{1.9}$$

for all $\pi_m \in \mathbf{\Pi}_{[m]}$.

Projectivity implies that the probability function has to satisfy the addition rule:

$$f_n(n_1, \ldots, n_k) = f_{n+1}(n_1, \ldots, n_k, 1) + \sum_{j=1}^{k} f_{n+1}(n_1, \ldots, n_j + 1, \ldots, n_k) \tag{1.10}$$

The formula means that a projective sequence of partitions can be constructed by iteratively adding members $1, 2, \ldots,$ into it. Suppose that after iteration n the partition constructed has k clusters of sizes $n_1, \ldots, n_k$. At iteration $n+1$, a new member is added into a new cluster with probability proportional to $f_{n+1}(n_1, \ldots, n_k, 1)$, while it is added to the jth cluster with probability proportional to $f_{n+1}(n_1, \ldots, n_j + 1, \ldots, n_k)$.

Theorem 1.11 (Kingman (1978))**.** *Given a projective sequence of exchangeable random partitions $\Pi_1, \Pi_2, \ldots$, there is a unique exchangeable random partition Π_∞ of $\mathbb{N}$, such that $\Pi_n \sim J_{[n]}(\Pi_\infty)$. The converse is true as well.*

We will simply refer to a projective sequence of exchangeable random partitions simply as exchangeable random partitions (on $\mathbb{N}$) from now on.

Definition 1.12 (Exchangeable Partition Probability Function)**.** *A function f with arguments finite sequences of positive integers is called an exchangeable partition probability function (EPPF) if:*

- *f is a symmetric function of its arguments;*

- *f satisfies the addition rule:*

$$f_n(n_1, \ldots, n_k) = f_{n+1}(n_1, \ldots, n_k, 1) + \sum_{j=1}^{k} f_{n+1}(n_1, \ldots, n_j + 1, \ldots, n_k) \tag{1.13}$$

An EPPF governs the distribution for an exchangeable random partition, with

$$\mathbb{P}(\Pi_n = \{c_1, \ldots, c_k\}) = f(|c_1|, \ldots, |c_k|) \tag{1.14}$$

Definition 1.15 (Chinese Restaurant Process)**.** *A Chinese restaurant process (CRP) is a random partition on $[n]$ constructed iteratively, where at each iteration, a new member is added to a new cluster with probability proportional to α, and to a cluster of size m with probability proportional to m. We denote the distribution by $\mathrm{CRP}(\alpha, [n])$.*

A CRP is obviously projective. It is also exchangeable, since its probability function, obtained by multiplying the probabilities associated with each stage of the construction, is:

$$\mathbb{P}(\Pi_n = \pi_n) = \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \alpha^{|\pi_n|} \prod_{c \in \pi_n} \Gamma(|c|) \tag{1.16}$$

Definition 1.17 (Two-parameter Chinese Restaurant Process)**.** *A two-parameter CRP is a random partition on $[n]$ constructed iteratively, where at each iteration, a new member is added to a new cluster with probability proportional to $\alpha + dk$ where k is the current number of clusters, and to a cluster of size m with probability proportional to $m - d$. We denote the distribution by $\mathrm{CRP}(\alpha, d, [n])$.*

A $\mathrm{CRP}(\alpha, d, [n])$ is projective and exchangeable as well, since its EPPF is given by:

$$\mathbb{P}(\Pi_n = \pi_n) = \frac{[\alpha + d]_d^{|\pi_n|-1}}{[\alpha + 1]_1^{n-1}} \prod_{c \in \pi_n} [1 - d]_1^{|c|-1} \tag{1.18}$$

where $[x]_b^a = x(x+b)\cdots(x+(a-1)b)$.

It is in general hard to come up with exchangeable random partitions by fiddlying with its EPPF or an iterative construction. It is easier to think about exchangeable random partitions as induced by a probability measure μ. Consider

$$X_i | \mu \sim \mu \tag{1.19}$$

for each $i = 1, 2, \ldots$. Construct a partition on $\mathbb{N}$ such that i and j belong to the same cluster if and only if $X_i = X_j$. This partition is obviously exchangeable since X_i's are iid. Further, if μ is random, the induced random partition is still exchangeable.

Theorem 1.20 (Kingman (1978))**.** *Every exchangeable random partition on $\mathbb{N}$ can be constructed by inducing it from a random probability measure.*

This theorem motivates study of exchangeable random partitions via random probability measures. It also motivates the question of what random probability measure underlies the Chinese restaurant processes (the Dirichlet process and the Pitman-Yor process respectively).

The structure of the random partition can be read off from the atomic structure of μ. Note that an atom x of μ of mass w will correspond to a cluster c_x, with asymptotic size w (as $n \to \infty$, $c_x/n \to w$.) If μ has a smooth component, say of total mass w_0, then every draw X_i coming from this component will be unique with probability 1, so that it corresponds to a cluster of size 1. The total proportion of integers belonging to such tiny clusters will asymptote to w_0. These clusters are called dust.

2 Poisson Processes

Definition 2.1 (Poisson Process)**.** *A Poisson process is a random measure N on space (S, Ω) such that for each measurable set A, $N(A)$ is Poisson distributed with mean $\lambda(A)$, where λ is the mean measure of the process. We write:*

$$N \sim \text{Poisson}(\lambda) \tag{2.2}$$

A Poisson process is composed of atoms of unit mass:

$$N = \sum_{k=1}^{N(S)} \delta_{x_k} \qquad\qquad N(A) = \sum_{k=1}^{N(S)} \delta_{x_k}(A) \tag{2.3}$$

If $\lambda(A)$ is finite, then the point masses in N can be generated in the following fashion:

- First generate Poisson number of points $n \sim \text{Poisson}(\lambda(A))$.
- For each $k \in [n]$, generate $x_k \sim \lambda(A \cap \cdot)/\lambda(A)$

Proposition 2.4 (Completely Randomness)**.** *A Poisson process is completely random: if A and B are disjoint, then $N(A) \perp\!\!\!\perp N(B)$.*

Proposition 2.5 (Transformations)**.**

- *Superposition: if N and N' are independent Poisson processes with means λ and λ', then $N + N'$ is a Poisson process with mean $\lambda + \lambda'$.*
- *Mapping: if N is a Poisson process with mean λ and $f : S \to S'$ is measurable, then $N \circ f^{-1}$ is Poisson with mean $\lambda \circ f^{-1}$.*

Theorem 2.6 (Campbell)**.** *If $N \sim \text{Poisson}(\lambda)$, then*

$$\mathbb{E}\left[\int f(x)N(dx)\right] = \mathbb{E}\left[\sum_{k=1}^{N(S)} f(x_k)\right] = \int f(x)\lambda(dx) \tag{2.7}$$

A useful characterization of point processes is via the characteristic functional.

$$\Psi[f] = \mathbb{E}[e^{-\int f(x)N(dx)}] \tag{2.8}$$

This is a generalization of the characteristic function of a random variable, and completely determines the point process.

Theorem 2.9 (Characteristic Functional)**.** *For a Poisson process $N \sim \text{Poisson}(\lambda)$, the characteristic functional is:*

$$\Psi[f] = e^{-\int 1 - e^{-f(x)}\lambda(dx)} \tag{2.10}$$

Proposition 2.11 (Change of Probability)**.** *Let f be such that $-\int 1 - e^{-f(x)}\lambda(dx) < \infty$. We can define a point process N absolutely continuous wrt $\text{Poisson}(\lambda)$, with density:*

$$p(N) = \frac{e^{-\int f(x)N(dx)}}{e^{-\int 1 - e^{-f(x)}\lambda(dx)}} \tag{2.12}$$

Then N is a Poisson process with exponentially tilted mean measure $e^{-f(x)}\lambda(dx)$.

Theorem 2.13 (Palm Formula)**.** *If $N \sim \text{Poisson}(\lambda)$ and f and G are functions of $x \in S$ and of (x, N) respectively, then*

$$\mathbb{E}\left[\int f(x)G(x, N)N(dx)\right] = \int \mathbb{E}[G(x, \delta_x + N)]f(x)\lambda(dx) \tag{2.14}$$

The LHS can be interpreted as approximately the following: the expectation of $f(x)G(x, N)N(S)$ where we first draw $N \sim \text{Poisson}(\lambda)$, then draw $x \sim N/N(S)$. The RHS is approximately the same expectation but where we first draw $x \sim \lambda/\lambda(S)$, then $\delta_x + N$ is a Poisson process conditioned on there being an atom at x.

3 Completely Random Measures

Definition 3.1 (Completely Random Measure (CRM)). *A completely random measure μ over a measure space (S, Ω) is a random measure such that if A and B are disjoint then $\mu(A) \perp\!\!\!\perp \mu(B)$.*

CRMs cannot be random and non-atomic, since a random smooth density function must have some local correlations. Examples of CRMs include (non-random) measures that can contain both atoms and a smooth component, and atomic random measures $\mu = \sum_j v_j \delta_{x_j}$ where v_j are random and independent, but x_j are fixed (non-random). A more non-trivial example can be constructed using Poisson processes. Let $N = \sum_k \delta_{(w_k, y_k)}$ be a Poisson process over the product space $\mathbb{R}_+ \times S$, and define

$$\mu(A) = \sum_k w_k \chi_A(y_k) \tag{3.2}$$

where χ_A is the characteristic function for A. Then the completely randomness of N will be inherited by μ.

Theorem 3.3 (Kingman (1967)). *A CRM μ can be decomposed into a sum of 3 independent components:*

$$\mu = \mu_0 + \sum_{j=1}^{J} v_j \delta_{x_j} + \sum_{k=1}^{K} w_k \delta_{y_k} \tag{3.4}$$

where μ_0 is a non-random measure, J is constant, $\{x_j\}$ are fixed members of S, v_j are mutually independent random variables on $\mathbb{R}_+$, and $N = \sum_{k=1}^{K} \delta_{(w_k, y_k)}$ is a Poisson process over $\mathbb{R}_+ \times S$.

The atoms $\{x_j\}$ are called the fixed atoms of μ, while $\{y_k\}$ are the random atoms. Let $q_j(dv_j)$ be the distribution of v_j. The mean measure λ is called the Lévy measure. To ensure that μ is σ-finite, it has to satisfy the property that

$$\int_{\mathbb{R}_+} \int_A \min(w, 1) \lambda(dw, dy) < \infty \tag{3.5}$$

for each $A \in \Omega$. Further, the number of random atoms in A is infinite iff $\lambda(\mathbb{R}_+ \times A) = \infty$. When λ decomposes into a product of two measures, $\lambda(dw, dy) = \rho(dw)H(dy)$, the CRM is called homogeneous. It implies that the masses of the random atoms are independent of their locations.

The characteristic functional of the underlying Poisson process N governing the distribution of the random atoms translates to a representation of CRMs called the Lévy-Khintchine representation.

Theorem 3.6 (Lévy-Khintchine Representation). *The characteristic functional of μ is*

$$\Psi[f] = \mathbb{E}\left[e^{-\int f(x)\mu(dx)}\right] = \exp\left(-\int f(x)\mu_0(dx) - \sum_{j=1}^{J} \psi_j(f(x_j)) - \int_{\mathbb{R}_+ \times S} 1 - e^{-wf(y)} \lambda(dw, dy)\right) \tag{3.7}$$

where $\psi_j(z) = -\log \mathbb{E}[e^{-zv_j}]$ is the Laplace transform of q_j.

4 Normalized Random Measures

Define a normalized random measure (NRM) as a random probability measure obtained by normalizing a homogeneous completely random measure (CRM):

$$\mu \sim \mathrm{CRM}(\rho H) \tag{4.1}$$

$$T = \mu(\Theta) \tag{4.2}$$

$$\nu = \mu / T \tag{4.3}$$

T is the total mass of the unnormalized μ. ρ is the Lévy measure of the CRM, and H the base distribution. Generalizations to other CRMs are possible; homogeneity is assumed for simplicity here.

4.1 Posterior Distribution

Consider a sample of size n drawn iid from ν:

$$X_i|\nu \sim \nu \tag{4.4}$$

We are interested in the posterior distribution of μ and ν given observations $X_i = x_i$ for $i = 1, \ldots, n$. Noting that μ is discrete, different observations can take on the same values. Let x_j^* for $j = 1, \ldots, k$ be the k unique values, with x_j^* appearing n_j times. Since ν is determined given μ, the probability of the observations is

$$\mathbb{P}(X_i \in dx_i \forall i \in [n]|\mu) = T^{-n} \prod_{j=1}^{k} \mu(dx_j^*)^{n_j} \tag{4.5}$$

The difficulty here is in the T^{-n} term, which we can address by introducing an auxiliary variable U:

$$U|\mu \sim \text{Gamma}(n, T) \tag{4.6}$$

$$\mathbb{P}(U \in du|\mu) = \frac{T^n}{\Gamma(n)} u^{n-1} e^{-Tu} du \tag{4.7}$$

$$\mathbb{P}(U \in du) = \frac{u^{n-1}}{\Gamma(n)} \mathbb{E}[T^n e^{-Tu}] du \tag{4.8}$$

Multiplying the probabilities, we get:

$$\mathbb{P}(X_i \in dx_i \forall i \in [n], U \in du|\mu) = \Gamma(n)^{-1} \prod_{j=1}^{k} \mu(dx_j^*)^{n_j} u^{n-1} e^{-Tu} du \tag{4.9}$$

Consider the characteristic functional of the posterior process. This is given by:

$$\mathbb{E}\left[e^{-\int f(x)\mu(dx)}|X_i \in dx_i \forall i \in [n], U \in du\right] = \frac{\mathbb{E}[e^{-\int f(x)\mu(dx)}\mathbb{P}(X_i \in dx_i \forall i \in [n], U \in du|\mu)]}{\mathbb{E}\left[\mathbb{P}(X_i \in dx_i \forall i \in [n], U \in du|\mu)\right]} \tag{4.10}$$

To obtain the numerator,

$$\mathbb{E}\left[e^{-\int f(x)\mu(dx)}\Gamma(n)^{-1} \prod_{j=1}^{k} \mu(dx_j^*)^{n_j} u^{n-1} e^{-Tu} du\right] \tag{4.11}$$

$$= \Gamma(n)^{-1} u^{n-1} du \mathbb{E}\left[e^{-\int (f(x)+u)\mu(dx)} \prod_{j=1}^{k} \mu(dx_j^*)^{n_j}\right] \tag{4.12}$$

Applying the Palm formula,

$$= \Gamma(n)^{-1} u^{n-1} du \int \mathbb{E}\left[e^{-\int (f(x)+u)(\mu + w_k \delta_{x_k'})(dx)} \prod_{j=1}^{k-1} \mu(dx_j^*)^{n_j}\right] (w_k \delta_{x_k'}(dx_k^*))^{n_j} \rho(dw_k) h(dx_k') \tag{4.13}$$

$$= \Gamma(n)^{-1} u^{n-1} du \mathbb{E}\left[e^{-\int (f(x)+u)\mu(dx)} \prod_{j=1}^{k-1} \mu(dx_j^*)^{n_j}\right] h(dx_k^*) \int e^{-(f(x_k^*)+u)w_k} w_k^{n_j} \rho(dw_k) \tag{4.14}$$

$$= \Gamma(n)^{-1} u^{n-1} du \mathbb{E}\left[e^{-\int (f(x)+u)\mu(dx)}\right] \prod_{j=1}^{k} h(dx_j^*) \int e^{-(f(x_j^*)+u)w_j} w_j^{n_j} \rho(dw_j) \tag{4.15}$$

By the Lévy-Khintchine representation,

$$=\Gamma(n)^{-1}u^{n-1}du\exp\left(-\int(1-e^{-w(f(x)+u)})\rho(dw)h(dx)\right)\prod_{j=1}^{k}h(dx_j^*)\int e^{-(f(x_j^*)+u)w_j}w_j^{n_j}\rho(dw_j) \tag{4.16}$$

Setting $f(x)=0$ gives us the denominator, which is the marginal probability of the observation and auxiliary variable:

$$\mathbb{E}\left[\mathbb{P}(X_i\in dx_i\forall i\in[n],U\in du|\mu)\right] \tag{4.17}$$

$$=\mathbb{P}(X_i\in dx_i\forall i\in[n],U\in du) \tag{4.18}$$

$$=\Gamma(n)^{-1}u^{n-1}du\mathbb{E}\left[e^{-\int u\mu(dx)}\right]\prod_{j=1}^{k}h(dx_j^*)\int e^{-uw_j}w_j^{n_j}\rho(dw_j) \tag{4.19}$$

$$=\Gamma(n)^{-1}u^{n-1}e^{-\psi(u)}du\prod_{j=1}^{k}h(dx_j^*)\kappa(u,n_j) \tag{4.20}$$

where $\psi(u)=-\log\mathbb{E}[e^{-\int u\mu(dx)}]=\int(1-e^{-uw})\rho(dw)$ is the Laplace transform of ρ, and $\kappa(u,n)=\int e^{-uw}w^n\rho(dw)$ is the nth moment of the u-exponentially tilted Lévy measure. Now dividing the numerator by the denominator, the characteristic functional of the posterior μ is:

$$\mathbb{E}\left[e^{-\int f(x)\mu(dx)}|X_i\in dx_i\forall i\in[n],U\in du\right] \tag{4.21}$$

$$=\exp\left(-\int(e^{-wu}-e^{-w(f(x)+u)})\rho(dw)h(dx)\right)\prod_{j=1}^{k}\int e^{-f(x_j^*)w_j}\frac{e^{-uw_j}w_j^{n_j}\rho(dw_j)}{\kappa(u,n_j)} \tag{4.22}$$

$$=\exp\left(-\int(1-e^{-wf(x)})e^{-wu}\rho(dw)h(dx)\right)\prod_{j=1}^{k}\int e^{-f(x_j^*)w_j}\frac{e^{-uw_j}w_j^{n_j}\rho(dw_j)}{\kappa(u,n_j)} \tag{4.23}$$

$$\tag{4.24}$$

Thus the posterior μ can be expressed as:

$$\mu=\mu'+\sum_{j=1}^{k}W_j\delta_{x_j^*} \tag{4.25}$$

where the fixed atoms at the k observed values $\{x_j^*\}$ has random masses distributed as

$$\mathbb{P}(W_j\in dw_j|X_i\in dx_i\forall i\in[n],U\in du)=\frac{e^{-uw_j}w_j^{n_j}\rho(dw_j)}{\kappa(u,n_j)} \tag{4.26}$$

and the random atoms are described by a CRM μ' with an exponentially tilted Lévy measure

$$e^{-wu}\rho(dw)h(dx) \tag{4.27}$$

4.2 Exchangeable Partition Probability Function

The exchangeable partition probability function (EPPF) can be easily ready off from the marginal probability (4.16). Denoting by Π_n the partition on $[n]$ induced by the observations $\{x_j\}$, we get:

$$\mathbb{P}(\Pi_n=\pi_n,U\in du)=\Gamma(n)^{-1}u^{n-1}e^{-\psi(u)}du\prod_{c\in\pi_n}\kappa(u,|c|) \tag{4.28}$$

The conditional distribution given $U \in du$ is obtained by dividing by (4.8):

$$\mathbb{P}(\Pi_n = \pi_n | U \in du) = \frac{e^{-\psi(u)}}{\mathbb{E}[T^n e^{-Tu}]} \prod_{c \in \pi_n} \kappa(u, |c|) \tag{4.29}$$

And the EPPF itself is:

$$\mathbb{P}(\Pi_n = \pi_n) = \Gamma(n)^{-1} \int u^{n-1} e^{-\psi(u)} \prod_{c \in \pi_n} \kappa(u, |c|) du \tag{4.30}$$

4.3 Stick-breaking Construction

To derive the stick-breaking construction for NRMs, consider a size-biased sample from ν, i.e. if

$$\mu = \sum_{k=1}^{\infty} w_k \delta_{x_k^*} \tag{4.31}$$

then let 1^* be a variable taking on value k with probability w_k/T, and let $W_1^* = w_{1^*}$. Consider the joint distribution of W_1^* and T:

$$\mathbb{P}(W_1^* \in dw, T \in dt | \mu) = \sum_{k=1}^{\infty} \frac{w_k}{T} \delta_{w_k}(dw) \delta_T(dt) \tag{4.32}$$

$$\mathbb{P}(W_1^* \in dw, T \in dt) = \mathbb{E}\left[\sum_{k=1}^{\infty} \frac{w_k}{T} \delta_{w_k}(dw) \delta_T(dt) \right] \tag{4.33}$$

$$= \int \mathbb{E}\left[\frac{1}{T + w'} \delta_{T+w'}(dt) \right] w' \delta_{w'}(dw) \rho(dw') \tag{4.34}$$

$$= \frac{w}{t} g(t - w) dt \rho(dw) \tag{4.35}$$

where g is the density of T under Lévy measure ρ.[1] Intuitively, the probability that the first size-biased mass W_1^* takes on value w and that the total mass T is t is the "probability" of w under ρ, times the probability of the rest of the mass being $t - w$ under g, times the probability w/t that mass w is chosen.

$$\mathbb{P}(W_1^* \in dw | T \in dt) = \frac{w}{t} \frac{g(t - w)}{g(t)} \rho(dw) \tag{4.36}$$

The above gives the first size-biased mass W_1^*. Now we can recurse on the rest of the atoms in μ, with total mass left $t - w$. This gives:

$$\mathbb{P}(W_k^* \in dw_k | T \in dt, W_j^* \in dw_j \forall j \in [k-1]) = \frac{w_k}{t - \sum_{j=1}^{k-1} w_j} \frac{g(t - \sum_{j=1}^{k} w_k)}{g(t - \sum_{j=1}^{k-1} w_k)} \rho(w_k) \tag{4.37}$$

Note that the conditionals are Markovian in the sense that each iteration depends only on the total mass left $t - \sum_{j \in [k-1]} w_j$, and not on the original total mass, nor the sizes of the previous size-biased masses.

To complete the stick-breaking construction, we have to first sample the total mass T before starting to break the sticks. Note that this stick-breaking construction generalizes to Poisson-Kingman processes, the only difference now being that the first total mass T has a different distribution.

[1] I have not found any statements regarding absolute continuity of the distribution of T. Bertoin (2006), Pitman (2003) both assumed it without proof. For the NRMs in use, they do exist.

4.4 Examples

We will give two examples of how the above results pan out: the Dirichlet process and the normalized generalized gamma process. The normalized generalized gamma process (NGGP) Brix (1999), Favaro and Teh (2012), Lijoi et al. (2007) is the current most general tractable family of NRMs. It encompasses the DP (a normalized gamma process) Ferguson (1973), the normalized stable Kingman (1975), Perman (1990), and the normalized inverse Gaussian Lijoi et al. (2005).

4.4.1 Dirichlet Process

The Dirichlet process is a normalized gamma process. The Lévy measure for the gamma process is given by

$$\rho(dw) = \alpha w^{-1} e^{-w} dw \tag{4.38}$$

The exponentially tilted moments are:

$$\kappa(u, n) = \int e^{-uw} w^n \rho(dw) = \alpha \int_0^\infty w^{n-1} e^{(u+1)w} dw = \alpha \frac{\Gamma(n)}{(u+1)^n} \tag{4.39}$$

The Laplace transform for ρ is:

$$\psi(u) = \int (1 - e^{-uw}) \rho(dw) = \alpha \int_0^\infty (1 - e^{-uw}) w^{-1} e^{-w} dw = \alpha \log(u+1) \tag{4.40}$$

And the total mass T is gamma distributed with shape α and scale 1, so

$$g(t) = \frac{1}{\Gamma(\alpha)} t^{\alpha-1} e^{-t} \tag{4.41}$$

Plugging these into (4.28), we get:

$$\begin{aligned}
\mathbb{P}(\Pi_n = \pi_n, U \in du) =& \Gamma(n)^{-1} u^{n-1} e^{-\psi(u)} du \prod_{c \in \pi_n} \kappa(u, |c|) && (4.42)\\
=& \Gamma(n)^{-1} \frac{u^{n-1}}{(u+1)^\alpha} du \prod_{c \in \pi_n} \alpha \frac{\Gamma(|c|)}{(u+1)^{|c|}} && (4.43)\\
=& \Gamma(n)^{-1} \frac{u^{n-1}}{(u+1)^{n+\alpha}} du \; \alpha^{|\pi_n|} \prod_{c \in \pi_n} \Gamma(|c|) && (4.44)
\end{aligned}$$

Thus we note that the random partition Π_n is independent of the auxiliary variable U. Integrating out U,

$$\begin{aligned}
\mathbb{P}(\Pi_n = \pi_n) =& \frac{1}{\Gamma(n)} \int_0^\infty \frac{u^{n-1}}{(u+1)^{n+\alpha}} du \; \alpha^{|\pi_n|} \prod_{c \in \pi_n} \Gamma(|c|) && (4.45)\\
=& \frac{1}{\Gamma(n)} \frac{\Gamma(n)\Gamma(\alpha)}{\Gamma(n+\alpha)} \alpha^{|\pi_n|} \prod_{c \in \pi_n} \Gamma(|c|) && (4.46)\\
=& \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \alpha^{|\pi_n|} \prod_{c \in \pi_n} \Gamma(|c|) && (4.47)
\end{aligned}$$

For the stick-breaking construction, (4.37) is

$$\begin{aligned}
& \mathbb{P}(W_k^* \in dw_k | T \in dt, W_j^* \in dw_j \forall j \in [k-1]) && (4.48)\\
=& \frac{w_k}{t - \sum_{j=1}^{k-1} w_j} \frac{g(t - \sum_{j=1}^k w_k)}{g(t - \sum_{j=1}^{k-1} w_k)} \rho(w_k) && (4.49)\\
=& \frac{w_k}{t - \sum_{j=1}^{k-1} w_j} \frac{(t - \sum_{j=1}^k w_k)^{\alpha-1} e^{-(t - \sum_{j=1}^k w_k)}}{(t - \sum_{j=1}^{k-1} w_k)^{\alpha-1} e^{-(t - \sum_{j=1}^{k-1} w_k)}} \alpha w_k^{-1} e^{-w_k} && (4.50)
\end{aligned}$$

Transforming the variables to $V_k = W_k^*/(T - \sum_{j=1}^{k-1} W_k^*)$, we get:

$$\mathbb{P}(V_k \in dv_k | T \in dt, V_j \in dv_j \forall j \in [k-1]) = \alpha(1-v_k)^{\alpha-1} \tag{4.51}$$

That is, each $v_k \sim \text{Beta}(1, \alpha)$ independently, and independent from the total mass T.

4.4.2 Normalized Generalized Gamma Process

The NGGP hsa Lévy measure

$$\rho(dw) = \frac{\alpha}{\Gamma(1-\sigma)} w^{-\sigma-1} e^{-\tau w} dw \tag{4.52}$$

When $\sigma = 0, \tau = 1$ we get the DP. When $\tau = 0$ we get the normalized stable, and when $\sigma = .5$ we get the normalized inverse Gaussian.

The exponentially tilted moments are:

$$\kappa(u, n) = \int e^{-uw} w^n \rho(dw) = \frac{\alpha}{\Gamma(1-\sigma)} \int_0^\infty w^{n-\sigma-1} e^{(u+\tau)w} dw = \frac{\alpha\Gamma(n-\sigma)}{(u+\tau)^{n-\sigma}\Gamma(1-\sigma)} \tag{4.53}$$

The Laplace transform for ρ is (via integration by parts):

$$\psi(u) = \int (1 - e^{-uw}) \rho(dw) = \frac{\alpha}{\Gamma(1-\sigma)} \int_0^\infty (1 - e^{-uw}) w^{-\sigma-1} e^{-\tau w} dw = \alpha \frac{(u+\tau)^\sigma - \tau^\sigma}{\sigma} \tag{4.54}$$

The joint distribution of Π_n and U is:

$$\mathbb{P}(\Pi_n = \pi_n, U \in du) = \Gamma(n)^{-1} u^{n-1} e^{-\psi(u)} du \prod_{c \in \pi_n} \kappa(u, |c|) \tag{4.55}$$

$$= \Gamma(n)^{-1} u^{n-1} e^{-\alpha \frac{(u+\tau)^\sigma - \tau^\sigma}{\sigma}} du \prod_{c \in \pi_n} \frac{\alpha\Gamma(|c|-\sigma)}{(u+\tau)^{|c|-\sigma}\Gamma(1-\sigma)} \tag{4.56}$$

$$= \frac{\alpha^{|\pi_n|}}{\Gamma(n)} \frac{u^{n-1}}{(u+\tau)^{n-|\pi_n|}} e^{-\alpha \frac{(u+\tau)^\sigma - \tau^\sigma}{\sigma}} du \prod_{c \in \pi_n} \frac{\Gamma(|c|-\sigma)}{\Gamma(1-\sigma)} \tag{4.57}$$

Marginalizing out U gives the EPPF. Note that the EPPF is of Gibbs type. The stick-breaking construction can also be obtained in closed (but ugly) form.

5 Poisson-Kingman Processes

NRMs form a wide class of exchangeable random partitions/random probability measures, but it does not include an important example, which is the Pitman-Yor process. A generalization of NRMs called Poisson-Kingman processes encompasses this (Pitman, 2003). These form the currently largest class of exchangeable random partitions/random probability measures. They generalize NRMs by allowing the total mass to come from a distribution different from the one under the homogeneous CRM.

Let ρ be the Lévy measure and H the base distribution of the CRM. Denote by $\text{CRM}(\rho H | t)$ the random measure obtained by conditioning the CRM on having total mass t. Let γ be some distribution on $\mathbb{R}_+$. A Poisson-Kingman process $\text{PKP}(\rho H, \gamma)$ is a random probability measure μ constructed as follows:

$$T \sim \gamma \tag{5.1}$$

$$\nu | T \sim \text{CRM}(\rho H | T) \tag{5.2}$$

$$\mu = \mu / T \tag{5.3}$$

If $\gamma = \delta_t$ is a point mass, then $\mathrm{PKP}(\rho H, \delta_t)$ is a random probability measure constructed from an underlying CRM, with the total mass T fixed at t. We can understand $\mathrm{PKP}(\rho H, \gamma)$ as a mixture over $\mathrm{PKP}(\rho H, \delta_t)$, where t has mixing distribution γ:

$$\mathrm{PKP}(\rho H, \gamma) = \int \mathrm{PKP}(\rho H, \delta_t)\gamma(dt) \tag{5.4}$$

5.1 Pitman-Yor Process

The Pitman-Yor process is a Poisson-Kingman process with the following parameters: The underlying CRM is a stable process, with

$$\rho(dx) = \frac{\sigma}{\Gamma(1-\sigma)} w^{-\sigma-1} \tag{5.5}$$

The total mass Tunder the CRM is a positive stable distribution with index σ, whose density we denote $g_\sigma(T)$.

The total mass distribution we use to get the Pitman-Yor is a polynomial-tilting of $g_\sigma(T)$:

$$\gamma_{\sigma,\alpha}(T) = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha/\sigma+1)} T^{-\alpha} g_\sigma(T) \tag{5.6}$$

Deriving the useful properties of the Pitman-Yor process (its EPPF, stick-breaking construction) is an interesting mathematical exercise (left to the user :p). For practical applications these derivations are not needed, but it is good to know where the various constructions came from.

6 Gibbs Type Exchangeable Random Partitions

Definition 6.1 (Gibbs Type Exchangeable Random Partition)**.** *An exchangeable random partition is said to be of Gibbs type if its EPPF can be written as:*

$$f(n_1, \ldots, n_k) = V_{nk} \prod_{j=1}^{k} W_{n_j} \tag{6.2}$$

where $n = \sum_{j=1}^{k} n_j$.

These are also related to product partition models Barry and Hartigan (1992), Müller and Quintana (ress). Product partition models are used in a larger range of problems, e.g. in regression problems where the partition structure is dependent on covariates, and in change point analysis. They do not have the nice theoretical results of Gibbs type partitions.

The Chinese restaurant processes are of Gibbs type. In the one-parameter ($\sigma = 0$) case, we have:

$$f_0(n_1, \ldots, n_k|\alpha) = \frac{1}{[\alpha]_1^n} \alpha^k \prod_{j=1}^{k} [1]_1^{n_j - 1} \tag{6.3}$$

If we mix over α, say with prior $\gamma(\alpha)$, we still get a Gibbs type partition with EPPF:

$$f_0(n_1, \ldots, n_k) = \int \frac{1}{[\alpha]_1^n} \alpha^k \gamma(\alpha) d\alpha \prod_{j=1}^{k} [1]_1^{n_j - 1} \tag{6.4}$$

Another class of Gibbs type partitions are the random partitions induced by a finite symmetric Dirichlet of length K and pseudocount $\beta > 0$:

$$\mathbb{P}(p_1, \ldots, p_K|\beta, K) = \frac{\Gamma(K\beta)}{\Gamma(\beta)^K} \prod_{j=1}^{K} p_j^{\beta-1} \tag{6.5}$$

The EPPF can be obtained by marginalizing out the Dirichlet, giving:

$$f_{-\beta}(n_1, \ldots, n_k|K) = \frac{[K]^k_{-1}}{[K\beta]^n_1}\beta^k \prod_{j=1}^{k}[1+\beta]^{n_j-1}_1 \tag{6.6}$$

If we mix over K (but not β), say with prior $\gamma(K)$, we still get a Gibbs type partition with EPPF:

$$f_{-\beta}(n_1, \ldots, n_k) = \sum_{K=1}^{\infty} \frac{[K]^k_{-1}}{[K\beta]^n_1}\beta^k \gamma(K) \prod_{j=1}^{k}[1+\beta]^{n_j-1}_1 \tag{6.7}$$

These Gibbs type partitions have finite but unbounded number of clusters: a partition drawn from the distribution will have $K < \infty$ number of clusters as $n \to \infty$, but K is not bounded. E.g. a Poisson prior on K.

A final class of Gibbs type priors are the Poisson-Kingman processes $\mathrm{PKP}(\rho_\sigma, \gamma)$ constructed from an underlying stable process. The EPPF for $\mathrm{PKP}(\rho_\sigma, \delta_t)$ can be derived as:

$$f_\sigma(n_1, \ldots, n_k|t) = \frac{\sigma^k \int_0^t g_\sigma(t-v)v^{n-k\sigma-1}dv}{t^n \Gamma(n-k\sigma)g_\sigma(t)} \prod_{j=1}^{k}[1-\sigma]^{n_j-1}_1 \tag{6.8}$$

Mixing over γ, we still get a Gibbs type partition with EPPF:

$$f_\sigma(n_1, \ldots, n_k) = \int_0^\infty \frac{\sigma^k \int_0^t g_\sigma(t-v)v^{n-k\sigma-1}dv}{t^n \Gamma(n-k\sigma)g_\sigma(t)}\gamma(dt) \prod_{j=1}^{k}[1-\sigma]^{n_j-1}_1 \tag{6.9}$$

There is a characterization of all Gibbs type exchangeable random partitions, which states that these are the only possible Gibbs type priors:

Theorem 6.10 (Gnedin and Pitman (2006))**.** *A Gibbs type random partition is exchangeable if and only if there is an index $\sigma \in [-\infty, 1]$ such that*

$$W_m = [1-\sigma]^{m-1}_1 = \frac{\Gamma(m-\sigma)}{\Gamma(1-\sigma)} \tag{6.11}$$

The index σ determines the type of random partition it is:

- $\sigma = 1$*: trivial partitions with all singleton clusters.*
- $\sigma < 0$*: These are mixtures of finite symmetric Dirichlets with pseudocounts $\beta = -\sigma$, with mixing distribution $\gamma(K)$ over the number of components of the Dirichlet.*
- $\sigma = 0$*: These are mixtures of DPs with mixing distribution $\gamma(\alpha)$ over the concentration parameter.*
- $0 < \sigma < 1$*: These are the Poisson-Kingman processes $\mathrm{PKP}(\rho_\sigma, \gamma)$, which are mixtures of $\mathrm{PKP}(\rho_\sigma, \delta_t)$ with fixed total mass t, with mixing distribution $\gamma(t)$ over the total mass.*

Further the basic processes over which we mix with γ are the extremal points of the space of Gibbs type exchangeable random partitions of index σ.

A simplification of the Gibbs type exchangeable random partition leads to the two-parameter CRP:

Theorem 6.12. *If an exchangeable random partition is of Gibbs type with $V_{nk} = V'_n V''_k$, then it is a two-parameter CRP.*

References

Barry, D. and Hartigan, J. A. (1992). Product partition models for change point problems. *Annals of Statistics*, 20(1):260–279.

Bertoin, J. (2006). *Random Fragmentation and Coagulation Processes*. Cambridge University Press.

Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability*, 31:929–953.

Favaro, S. and Teh, Y. W. (2012). MCMC for normalized random measure mixture models. In preparation.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230.

Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, 138(3):5674–5684.

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

Kingman, J. F. C. (1975). Random discrete distributions. *Journal of the Royal Statistical Society*, 37:1–22.

Kingman, J. F. C. (1978). The representation of partition structures. *Journal of the London Mathematical Society*, 18:374–380.

Lijoi, A., Mena, R. H., and Prüenster, I. (2005). Hierarchical mixture modelling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association*, 100:1278–1291.

Lijoi, A., Mena, R. H., and Prüenster, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture model. *Journal of the Royal Statistical Society B*, 69:715–740.

Müller, P. and Quintana, F. (In Press). Random partition models with regression on covariates. *Journal of Statistical Inference and Planning*.

Perman, M. (1990). *Random Discrete Distributions Derived from Subordinators*. PhD thesis, Department of Statistics, University of California at Berkeley.

Pitman, J. (2003). Poisson-Kingman partitions. In Goldstein, D. R., editor, *Statistics and Science: a Festschrift for Terry Speed*, pages 1–34. Institute of Mathematical Statistics.