Theory of Dependent Hierarchical Normalized Random Measures

Changyou Chen[1,2]
cchangyou@gmail.com
[1]RSISE, The Australian National University,
Canberra, ACT, Australia

Wray Buntine[2,1]
Wray.Buntine@nicta.com.au
[2]NICTA, ACT, Australia

Nan Ding[3]
ding10@purdue.edu
[3]Department of Computer Science, Purdue University, USA

June 7, 2013

Abstract

This paper presents theory for Normalized Random Measures (NRMs), Normalized Generalized Gammas (NGGs), a particular kind of NRM, and Dependent Hierarchical NRMs which allow networks of dependent NRMs to be analysed. These have been used, for instance, for time-dependent topic modelling. In this paper, we first introduce some mathematical background of completely random measures (CRMs) and their construction from Poisson processes, and then introduce NRMs and NGGs. Slice sampling is also introduced for posterior inference. The dependency operators in Poisson processes and for the corresponding CRMs and NRMs is then introduced and Posterior inference for the NGG presented. Finally, we give dependency and composition results when applying these operators to NRMs so they can be used in a network with hierarchical and dependent relations.

Keywords

completely random measures; normalized randomized measures; normalized Generalized gamma process; dependent hierarchical normalized randomized measures; hierarchical models;

Contents

1 Introduction

This paper presents theory for Normalized Random Measures (NRMs), Normalized Generalized Gammas (NGGs), a particular kind of NRM, and Dependent Hierarchical NRMs which allow networks of dependent NRMs to be analysed. These have been used, for instance, for time-dependent topic modelling [CDB12].

Dependency models are getting more and more popular in machine learning recently due to the fact of correlated data we are facing at, *e.g.*, real data is always correlated with each other rather than independent. The pioneer work on this topic goes to MachEachern [Mac99, Mac00], where he treated the jumps and atoms to be stochastic between dependent models. While there are many ways of constructing dependent nonparametric models, *e.g.*, from a stick-breaking construction [GS09], or from a hierarchical construction [TJBB06], in this paper, following the idea of [LGF10], we construct dependency normalized random measures from the underlying Poisson processes of the corresponding completely random measures [Kin67]. This construction is intuitive and allow flexibly controlling of the dependencies. A related construction in the statistical literature is made by Lijoi *et al.* [A. 12] that deals with modeling two groups of data.

In this paper, we first introduce in Section 2 some mathematical background of completely random measures (CRMs) and their construction from Poisson processes, and then introduce NRMs and NGGs. Slice sampling is also introduced to do the posterior sampling of NRMs using techniques from [GW11]. The dependency operators in Poisson processes and for the corresponding CRMs and NRMs is then introduced in Section 3 following the work of [Kin93, LGF10]. Posterior inference for the NGG are then developed in Section 4 based on the results of [JLP09]. Then we give the dependency and composition results when applying these operators to NRMs in Section 5. Proofs are given in the Appendix, Section A.

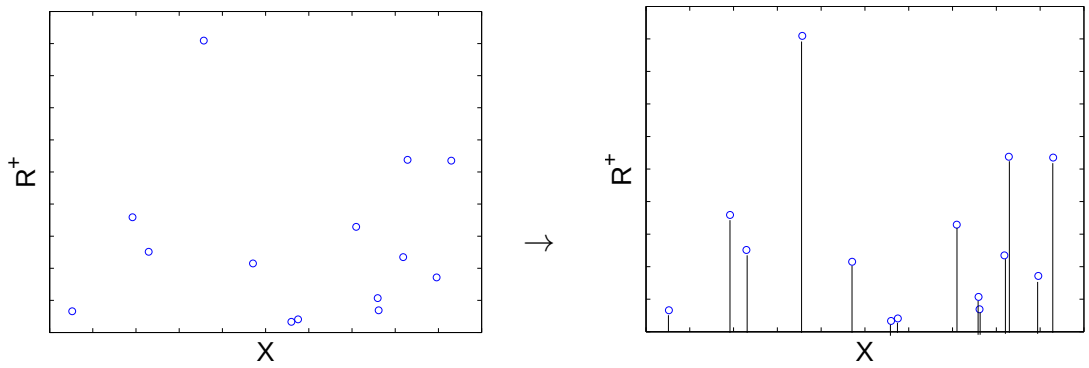
2 Background

In this section we briefly introduce background of Poisson processes, the corresponding completely random measures, dependency operations on these random measures, and normalized random measures.

Section 2.1 explains how to construct completely random measures from Poisson processes. Section 3.1 introduces operations on Poisson processes to construct dependent Poisson processes. Section 3.2 adapts these operations to the corresponding completely random measures (CRMs). Constructing normalized random measures (NRMs) from CRMs is discussed in Section 2.2 along with details of the Normalized Generalized Gamma (NGG), a particular kind of NRM for which the details have been worked out. A slice sampler for sampling an NRM is described in Section 2.3.

We first give an illustration of the basic construction for an NRM. for a target domain $\mathbb{X}$. The Poisson process is used to create a countable (and usually) infinite set of points in a product space of $\mathbb{R}^+$ with the target domain $\mathbb{X}$, as shown in the left

of Figure 1. The distribution is then a discrete one on these points. The distribution can be pictured by dropping lines from each point (t, x) down to $(0, x)$, and then normalizing all these lines so their sum is one. The resulting picture shows the set of weighted impulses that make up the constructed NRM on the target domain.



Counting process:

$N(\cdot) = \sum_k \delta_{(J_k, x_k)}(\cdot)$

Completely random measure:

$\tilde{\mu}(\cdot) = \sum_k J_k \delta_{x_k}(\cdot)$

Figure 1: Constructing a completely random measure from a counting process $N(\cdot)$ with points at (J_k, x_k).

2.1 Constructing Completely Random Measures from Poisson processes

In contrast to the general class of completely random measure (CRM) [Kin67], which admits a unique decomposition as the summation over there parts: a deterministic measure, a purely atomic measure with fixed atom locations and a discrete measure with random jumps and atoms, in this paper, we restrict it to the class of pure jump processes [FK72], which has the following form

$$\tilde{\mu} = \sum_{k=1}^{\infty} J_k \delta_{x_k}, \tag{1}$$

where $J_1, J_2, \cdots > 0$ are called the jumps of the process, and $x_1, x_2, \cdots$ are a sequence of independent random variables drawn from a base measurable space $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$[1].

It is shown that these kinds of CRMs can be constructed from Poisson processes with specific mean measures $\nu(\cdot)$. We will start from some definitions.

Poisson Distributions: A random variable X taking values in $\mathbb{N} = \{0, 1, \cdots, \infty\}$ is said to have the Poisson distribution with mean c in $(0, \infty)$ if

$$p(X = k|c) = \frac{e^{-c} c^k}{k!}, k \in \mathbb{N}, \tag{2}$$

[1] $\mathcal{B}(\mathbb{X})$ means the σ-algebra of $\mathbb{X}$, we sometimes omit this and use $\mathbb{X}$ to denote the measurable space.

then $X < \infty$ almost surely and $\mathbb{E}[X] = \text{Var}[X] = c$.

Poisson Processes: Let $(\mathbb{S}, \mathcal{S})$ be a measure space where $\mathcal{S}$ is the σ-algebra of $\mathbb{S}$. Let $\nu(\cdot)$ be a measure on it. A Poisson process on $\mathbb{S}$ is defined to be a random subset $\Pi \in \mathbb{S}$ such that if $N(A)$ is the number of points of Π in the measurable subset $A \subseteq \mathbb{S}$, then

a) $N(A)$ is a random variable having the Poisson distribution with mean $\nu(A)$, and

b) whenever $A_1, \cdots, A_n$ are in $\mathcal{S}$ and disjoint, the random variables $N(A_1), \cdots, N(A_n)$ are independent.

The integer-value random measure $N(\cdot)$ is called a *Poisson random measure* and the Poisson process is denoted as $\Pi \sim \text{PoissonP}(\nu)$, where ν is called the mean measure of the Poisson process.

Completely Random Measure: In this paper, we define a random measure on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ to be a linear functional of the Poisson random measure $N(\cdot)$, whose mean measure $\nu(\mathrm{d}t, \mathrm{d}x)$ defined on a product space $\mathbb{S} = R^+ \times \mathbb{X}$:

$$\tilde{\mu}(B) = \int_{\mathbb{R}^+ \times B} tN(\mathrm{d}t, \mathrm{d}x), \forall B \in \mathcal{B}(\mathbb{X}). \tag{3}$$

The mean measure $\nu(\mathrm{d}t, \mathrm{d}x)$ is called the *Lévy measure* of $\tilde{\mu}$.

The general treatment of constructing random measures from Poisson random measures can be found in [Jam05]. Note that the random measure $\tilde{\mu}$ in construction (3) has the same form as Equation (1) because $N(\cdot)$ is composed of a countable number of points. It can be proven to be a completely random measure [Kin67] on $\mathbb{X}$, meaning that for arbitrary disjoint subsets $\{A_i \in \mathbb{X}\}$ of the measurable space, the random variables $\{\tilde{\mu}(A_i)\}$ are independent.

For the completely random measure defined above to always be finite, it is necessary that $\int_{\mathbb{R}^+ \times \mathbb{X}} t\, \nu(\mathrm{d}t, \mathrm{d}x)$ be finite, and therefore for every $z > 0$, $\nu([z, \infty) \times \mathbb{X}) = \int_z^\infty \int_{\mathbb{X}} \nu(\mathrm{d}t, \mathrm{d}x)$ is finite [Kin93]. It follows that there will always be a finite number of points with jumps $J_k > z$ for that $z > 0$. Therefore in the bounded product space $[z, \infty) \times \mathbb{X}$ the measure $\nu(\mathrm{d}t, \mathrm{d}x)$ is finite. So it is meaningful to sample those points (J_k, x_k) with $J_k > z$ by first getting the count of points K sampled from a Poisson with (finite) mean $\nu([z, \infty) \times \mathbb{X})$, and then to sample the K points according to the distribution of $\frac{\nu(\mathrm{d}t, \mathrm{d}x)}{\nu([z,\infty)\times\mathbb{X})}$.

Without loss of generality, we assume the Lévy measure of Equation (3) can be decomposed as $\nu(\mathrm{d}t, \mathrm{d}x) = M\rho_\eta(\mathrm{d}t|x)H(\mathrm{d}x)$, where η denotes the hyper-parameters if any, $H(\mathrm{d}x)$ is a probability measure on $\mathbb{X}$ so $H(\mathbb{X}) = 1$, and M is called the *mass* of the Lévy measure. Note the total measure of $\rho_\eta(\mathrm{d}t|x)$ is not standardized in any way so in principle some mass could also appear in $\rho_\eta(\mathrm{d}t|x)$. The mass is used as a concentration parameter for the random measure.

A realization of $\tilde{\mu}$ on $\mathbb{X}$ can be constructed by sampling from the underlying Poisson process in a number of ways, either in rounds for decreasing bounds z using the logic just given, or by explicitly sampling the jumps in order. The later goes as follows [FK72]:

Lemma 1 (Sampling a CRM) *Sample a CRM $\tilde{\mu}$ with Lévy measure $\nu(\mathrm{d}t, \mathrm{d}x) = M\rho_\eta(\mathrm{d}t|x)H(\mathrm{d}x)$ as follows.*

- *Draw i.i.d. samples x_i from the base measure $H(\mathrm{d}x)$.*
- *Draw the corresponding weights J_i for these i.i.d. samples in decreasing order, which goes as:*
  - *Draw the largest jump J_1 from the cumulative distribution function $P(J_1 \leq j_1) = \exp\left\{-M\int_{j_1}^{\infty} \rho_\nu(\mathrm{d}t|x_i)\right\}$.*
  - *Draw the second largest jump J_2 from the cumulative distribution function $P(J_2 \leq j_2) = \exp\left\{-M\int_{j_2}^{j_1} \rho_\nu(\mathrm{d}t|x_2)\right\}$.*
  - $\ldots$
- *The random measure $\tilde{\mu}$ then can now be realized as $\tilde{\mu} = \sum_i J_i\delta_{x_i}$.*

As a random variable is uniquely determined by its Laplace transformation, the random measure $\tilde{\mu}$ is uniquely characterized by its *Laplace functional* through the Lévy-Khintchine representation of a Lévy process [Ç10]. That is, for any measurable function $f : \mathbb{X} \to \mathbb{R}^+$, we have

$$\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} f(x)\tilde{\mu}(\mathrm{d}x)\right\}\right] = \exp\left\{-\int_{\mathbb{R}^+\times\mathbb{X}} \left[1-\exp\left\{-tf(x)\right\}\right]\nu(\mathrm{d}t,\mathrm{d}x)\right\} \quad (4)$$

Now instead of dealing with $\tilde{\mu}$ itself, we deal with $\nu(\mathrm{d}t, \mathrm{d}x)$, which is called the Lévy measure of $\tilde{\mu}$, whose role in generating the measure via a Poisson process was explained above.

In the case where the measure on the jumps is not dependent on the data x, so $\rho_\eta(\mathrm{d}t|x) = \rho_\eta(\mathrm{d}t)$, then $\tilde{\mu}$ is called homogeneous, which is the case considered in this paper. When f does not depend on x, (4) simplifies to

$$\mathbb{E}\left[\exp\left\{-f\,\tilde{\mu}(B)\right\}\right] = \exp\left\{-M\,p(B)\int_{\mathbb{R}^+}\left[1-\exp\left\{-tf\right\}\right]\rho_\eta(\mathrm{d}t)\right\}\ . \quad (5)$$

Note the term inside the exponential plays an important role in subsequent theory, so it is given a name.

Laplace exponent: The *Laplace exponent*, denoted as $\psi_\eta(f)$ for a CRM with parameters η is given by

$$\begin{aligned}\psi_\eta(f) &= \int_{\mathbb{R}^+\times\mathbb{X}}\left[1-\exp\left\{-tf\right\}\right]\nu(\mathrm{d}t,\mathrm{d}x)\\ &= M\int_{\mathbb{R}^+}\left[1-\exp\left\{-tf\right\}\right]\rho_\eta(\mathrm{d}t) \qquad \text{(homogeneous case)}\ . \quad (6)\end{aligned}$$

Note that to guarantee the positiveness of jumps in the random measure, $\rho(\mathrm{d}t)$ in the Lévy measure should satisfy $\int_0^\infty \rho_\eta(\mathrm{d}t) = +\infty$ [ELP03], which leads to the following equations:

$$\psi_\eta(0) = 0, \qquad \psi_\eta(+\infty) = +\infty. \tag{7}$$

That $\psi_\eta(f)$ is finite for finite positive f implies (or is a consequence of) $\int_0^\infty t\rho_\eta(\mathrm{d}t)$ being finite.

Remark There are thus four different ways to define or interpret a CRM:

1. via the linear functional of Equation (3),
2. through the Lévy-Khintchine representation of Equation (4) using the Laplace exponent,
3. sampling in order of decreasing jumps using Lemma 1, and
4. sampling in blocks of decreasing jump values as discussed before Lemma 1.

2.2 Normalized random measures

Normalized Random Measures (NRM) Based on (3), a normalized random measure on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ is defined as[2]

$$\mu = \frac{\tilde{\mu}}{\tilde{\mu}(\mathbb{X})}. \tag{8}$$

The original idea of constructing random probabilities by normalizing completely random measures on $\mathbb{R}$, namely increasing additive processes, can be found in [ELP03], where it is termed normalized random measures with independent increment (NRMI) and the existence of such random measures is proved. This idea can be easily generalized from $\mathbb{R}$ to any parameter space $\mathbb{X}$, *e.g.*, $\mathbb{X}$ being the Dirichlet distribution space in topic modeling. Also note that the idea of normalized random measures can be taken as doing a transformation $Tr(\cdot)$ on completely random measures, that is $\mu = Tr(\tilde{\mu})$. In the normalized random measure case, $Tr(\cdot)$ is a transformation such that $Tr(\tilde{\mu}(\mathbb{X})) = 1$. A concise survey of other kinds of transformations can be found in [LP10].

Taking different Lévy measures $\nu(\mathrm{d}t, \mathrm{d}x)$ of (4), we can obtain different NRMs. We use $\mathrm{NRM}(\eta, M, H(\cdot))$ to denote the normalized random measure, where M is the total mass, which usually needs to be sampled in the model, and $H(\cdot)$ is the base probability measure, η is the set of other hyper-parameters to the measure on the jumps, depending on the specific NRMs. In this paper, we are interested in a class of NRMs called *normalized generalized Gamma processes*:

[2]In this paper, we use μ to denote a normalized random measure, while use $\tilde{\mu}$ to denote its unnormalized counterpart.

Normalized Generalized Gamma Processes: Generalized Gamma processes are random measures proposed by Brix [Bri99] for constructing shot noise Cox processes. They have the Lévy measures as

$$\nu(\mathrm{d}t, \mathrm{d}x) = \frac{e^{-bt}}{t^{1+a}} H(\mathrm{d}x), b > 0, 0 < a < 1. \tag{9}$$

By normalizing the generalized Gamma process as in (8), we obtain the normalized generalized Gamma process (NGG).

Sometimes we also need the Gamma distribution. Because there are several parameterisations of this in use, we define it here.

Gamma distribution: The Gamma distribution has two parameters, shape a and scale b, and is denoted $\mathrm{Ga}(a, b)$ with density function

$$p(x|\mathrm{Ga}(a,b)) \;=\; \frac{1}{\Gamma(a) b^a} x^{a-1} e^{-bx} \;.$$

For ease of representation and sampling, we convert the NGG into a different form using the following lemma.

Lemma 2 *Let a normalised random measure be defined using Lévy density $\nu(\mathrm{d}x, \mathrm{d}t)$. Then scaling t by $\lambda > 0$ yields an equivalent NRM up to a factor. That is, the normalised measure obtained using $\nu(\mathrm{d}x, \mathrm{d}t/\lambda)$ is equivalent to the normalised measure obtained using $\lambda\,\nu(\mathrm{d}x, \mathrm{d}t)$.*

By this lemma, without loss of generality, we can instead represent the NGG by eliminating the parameter b above.

Normalized Generalized Gamma: The NGG with shape parameter a, total mass (or concentration) parameter M and base distribution $H(\cdot)$, denoted $\mathrm{NGG}(a, M, H(\cdot))$, has Lévy density $M\,\rho_a(\mathrm{d}t) H(\mathrm{d}x)$ where

$$\rho_a(t) \;=\; \frac{a}{\Gamma(1-a)} \frac{e^{-t}}{t^{1+a}} \;.$$

Note that similar to the two parameter Poisson-Dirichlet process [PY97], the normalized generalized Gamma process with $a \neq 0$ can also produce power-law phenomenon, making it different from the Dirichlet process and suitable to model real data.

Proposition 1 ([LMP07]) *Let K_n be the number of components induced by the NGG with parameter a and mass M or the Dirichlet process with total mass M. Then for the NGG, $K_n/n^a \to S_{a,M}$ almost surely, where $S_{a,M}$ is a strictly positive random variable parameterized by a and M. For the DP, $K_n/\log(n) \to M$.*

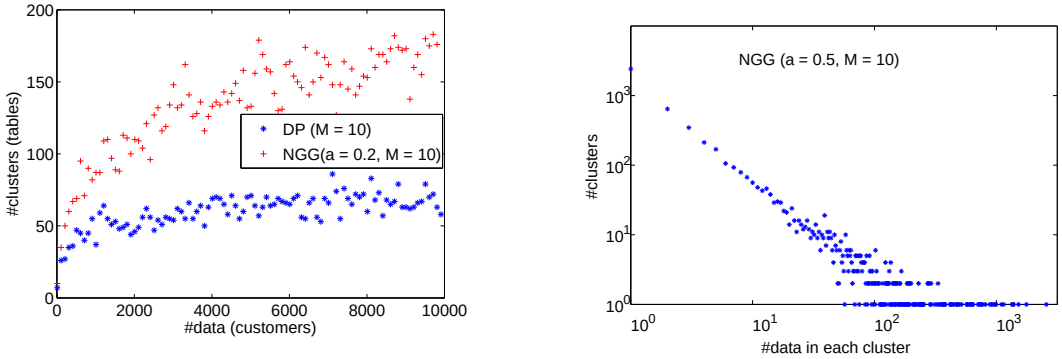


Figure 2: Power-law phenomena in NGG. The first plot shows the #data versus #clusters compared with DP, the second plot shows the size s of each cluster versus total number of clusters with size s.

Figure 2 demonstrates the power law phenomena in the NGG compared to the Dirichlet process (DP). We sample it using the generalized Blackwell-MacQueen sampling scheme [JLP09]. Each data to be sampled can choose an existing cluster or create a new cluster, resulting in K clusters with N data points in total.

Many familiar stochastic processes are special/limiting cases of normalized generalized Gamma processes, *e.g.*, *Dirichlet processes* arise when $a \to 0$. *Normalized inverse-Gaussian processes* (N-IG) arise when $a = \frac{1}{2}$ and $b = \frac{1}{2}$. If $b \to 0$, we get the *σ-stable process*, and if $a \to 0$ and b depends on x, we get the *extended Gamma process*.

Remark For the NGG, key formula used subsequently are as follows:

$$\begin{aligned}
\psi_a(v) &= M\left((1+v)^a - 1\right) \\
\int_L^\infty \rho_a(\mathrm{d}t) &= |Q(-a, L)| \\
\int_L^\infty e^{-vt}\rho_a(\mathrm{d}t) &= (1+v)^a |Q(-a, L(1+v))| \\
\int_0^L \left(1 - e^{-vt}\right)\rho_a(t)\mathrm{d}t &= ((1+v)^a - 1) + (1+v)^a |Q(-a, L(1+v))| - |Q(-a, L)|
\end{aligned}$$

where $Q(x, y) = \Gamma(x, y)/\Gamma(x)$ is the regularized upper incomplete Gamma function. Some mathematical libraries provide it for a negative first argument, or it can be evaluated using

$$Q(-a, z) = Q(1-a, z) - \frac{1}{\Gamma(1-a)} z^{-a} e^{-z},$$

using an upper incomplete Gamma function defined only for positive arguments.

Finally, because probabilities for a NRM necessarily have the divisor $\tilde{\mu}(\mathcal{X}) = \sum_{k=1}^{\infty} J_k$, and thus likelihoods of the NRM should involve powers of $\tilde{\mu}(\mathcal{X})$, a trick widely used to eliminate these terms is to do data augmentation via the Gamma identity.

Latent relative mass: Consider the case where N data are observed. By introducing the auxiliary variable, called *latent relative mass*, $U_N = \Gamma_N/\tilde{\mu}(\mathcal{X})$ where $\Gamma_N \sim \text{Gamma}(N, 1)$, then it follows that

$$\frac{1}{\tilde{\mu}(\mathcal{X})^N} p(\Gamma_N) \mathrm{d}\Gamma_N = \frac{U_N^{N-1}}{\Gamma(N)} e^{-U_N \tilde{\mu}(\mathcal{X})} \mathrm{d}U_N$$

Thus the N-th power of the normaliser can be replaced by an exponential term in the jumps which factorizes, at the expense of introducing the new latent variable U_N. To the best of our knowledge, the idea of this latent variable originals from [Jam05] and is future explicitly studied in [JLP06, JLP09, GW11], *etc.*.

2.3 Slice sampling normalized random measure mixtures

Slice sampling an NRM has been discussed in several papers, here we follow the method in [GW11], to briefly introduce the ideas behind it. It deals with the normalized random measure mixture of the type

$$\begin{aligned} \theta_{s_i} &\sim \mu \qquad\qquad \text{where } \mu = \sum_{k=1}^{\infty} \omega_k \delta_{\theta_k}, \\ X_i &\sim g_0(\cdot|\theta_{s_i}) \end{aligned} \tag{10}$$

where $\omega_k = J_k / \sum_{l=1}^{\infty} J_l$, $J_1, J_2, \cdots$ are the jumps of the corresponding CRM defined in (3), θ_k's are the components of the mixture model drawn *i.i.d.* from a parameter space $H(\cdot)$, s_i denotes the component that X_i belongs to, and $g_0(\cdot|\theta_k)$ is the density function to generate data on component k.

Given the observations $\vec{X}$, we introduce a slice latent variable u_i for each x_i so that we only consider those components whose jump sizes J_k's are larger than the corresponding u_i's. Moreover, only jumps with size greater than L are considered, and this is maintained my setting $L \leq \min_i u_i$. Sampling of the NRM can then be done by only considering jumps greater than L. The count of such jumps, K has a Poisson distribution, $K \sim \text{Poisson}(M \int_L^{\infty} \rho_\eta(\mathrm{d}t))$, while each jump has density $\frac{\rho_\eta(J_k)}{\int_L^{\infty} \rho_\eta(s)\mathrm{d}t}$.

Furthermore, the auxiliary variable U_N (latent relative mass) is introduced to decouple each individual jump J_k and their infinite sum of the jumps $\sum_{l=1}^{\infty} J_l$ appeared in the denominators of ω_k's. For clarification, we list the notation and its description in Table 1. Based on [GW11], we have the following posterior Lemma.

Table 1: List of notation.

Notation	Description
K	#components with jump sizes larger than a threshold L
$\theta_k, k = 1, \cdots, K$	components in the mixture model
M	total mass of the random measure
$J_k, k = 1, \cdots, K$	jump sizes of the random measure with all $J_k > L$
$X_i, i = 1, \cdots, N$	observed data
$n_k, k = 1, \cdots, K$	#data attached to each component
N	total number of data points
$s_i, i = 1, \cdots, N$	Variables indicating which component y_i belongs to
$u_i, i = 1, \cdots, N$	slice variable uniformly distributed in $(0, J_{s_i}]$ for y_i
L	$L = \min\{\vec{u}\}$
U_N	an auxiliary variable introduced to make the sampling feasible, the latent relative mass
$g_0(\cdot\|\theta_k)$	density function to generate data on component θ_k
$h(\theta_k)$	density of $H(\theta_k)$
$p_M(M)$	prior for M
$\nu(\mathrm{d}t, \mathrm{d}x)$	Lévy measure of the random measure with decomposition $\nu(\mathrm{d}t, \mathrm{d}x) = M\rho_\eta(\mathrm{d}t)H(\mathrm{d}x)$ considered in this paper

Lemma 3 *The posterior of the infinite mixture model (10) with the above auxiliary variables is proportional to*

$$
\begin{aligned}
&p(\vec{\theta}, J_1, \cdots, J_K, K, \vec{u}, \vec{s}, U_N, \vec{X}|L, NRM(\eta, M, H(\cdot))) \propto \\
&U_N^{N-1} \exp\left\{-U_N \sum_{k=1}^{K} J_k\right\} \exp\left\{-M \int_0^L \left(1 - e^{-U_N t}\right) \rho_\eta(t)\mathrm{d}t\right\} \\
&M^K \exp\left\{-M \int_L^\infty \rho_\eta(t)\mathrm{d}t\right\} \prod_{k=1}^{K} \rho_\eta(J_k) h(\theta_k) \prod_{i=1}^{N} 1(J_{s_i} > u_i) g_0(x_i|\theta_{s_i}), \qquad (11)
\end{aligned}
$$

where $1(a)$ *is a indicator function returning 1 if* a *is true and 0 otherwise,* $h(\cdot)$ *is the density of* $H(\cdot)$*,* $L \leq \min\{\vec{u}\}$*.*

The expressions for the NGG needed to work with this lemma were given in the remark at the end of Section 2.2. Thus the integral term in Equation (11) can be turned into an expression involving incomplete Gamma functions.

2.3.1 Sampling:

First, we denote the parameter set as $C = \left\{\vec{\theta}, J_1, \cdots, J_K, K, \vec{u}, L, \vec{s}, U_N, M\right\}$, then the sampling goes as

- **Sampling $\vec{s}$:** From (11) we get

$$p(s_i = k|C\backslash\{s_i\}) \propto 1(J_k > u_i)g_0(y_i|\theta_k) \tag{12}$$

- **Sampling U_N:** Similarly

$$\begin{aligned} p(U_N|C\backslash\{U_N\}) \quad \propto \quad & U_N^{N-1} \exp\left\{-U_N \sum_{k=1}^{K} J_k\right\} \\ & \exp\left\{-M \int_0^L \left[1 - \exp\left\{-U_N t\right\}\right] \rho_\eta(\mathrm{d}t)\right\}, \end{aligned} \tag{13}$$

  which can be sampled using rejection sampling from a proposal distribution $Ga\left(n, \sum_{k=1}^{K} J_k\right)$.

- **Sampling θ:** The posterior of θ_k with prior density $h(\theta_k)$ is

$$p(\theta_k|C\backslash\{\theta_k\}) \propto h(\theta) \prod_{i|s_i=k} g_0(y_i|\theta_k). \tag{14}$$

- **Sampling $K, \{J_1, \cdots, J_K\}$:** Sampling for J_k can be done separately for those associated with data points (fixed points) and for those that are not. Based on [JLP09], when integrating out $\vec{u}$ in (11), the posterior of the jump J_k with data attached ($n_k > 0$) is proportional to

$$J_k^{n_k} \exp\left\{-U_N J_k\right\} \rho_\eta(J_k), \tag{15}$$

  While for those without data attached ($n_k = 0$), based on [GW11], conditional on U_N, the number of these jumps follows a Poisson distribution with mean

$$M \int_L^\infty \exp\{-U_N t\}\rho_\eta(\mathrm{d}t),$$

  while their lengths t have densities proportional to

$$\exp\{-U_N t\}\rho_\eta(\mathrm{d}t)1(t > L).$$

- **Sampling $\vec{u}$:** $\vec{u}$ are uniformly distributed in the interval $(0, J_{s_i}]$ for each i. After sampling the $\vec{u}$, L is set to $L = \min\{\vec{u}\}$.

- **Sampling M:** The posterior of M with prior $p_M(M)$ is

$$p(M|C\backslash\{M\}) \propto p_M(M)M^K \exp\left\{-M\left[\int_L^\infty \rho_\eta(\mathrm{d}t) + \int_0^L \left[1 - \exp\left\{-U_N t\right\}\right] \rho_\eta(\mathrm{d}t)\right]\right\} \tag{16}$$

  $p_M(M)$ is usually taken to be Gamma distributed, so the posterior of M can be sampled conveniently.

3 Operations

This section introduces the dependency operations used. These are developed for Poisson processes, CRMs and NRMs.

3.1 Operations on Poisson processes

We review three operations that transform Poisson processes in order to construct dependent completely random measures. For details, refer to [Kin93, LGF10].

Superposition of Poisson processes Given a set of Poisson processes $\Pi_1, \Pi_2, \cdots, \Pi_n$, the superposition of these Poisson processes is defined as the union of the points in these Poisson processes:

$$\Pi := \bigcup_{i=1}^{n} \Pi_i. \tag{17}$$

Lemma 4 (Superposition Theorem) *Let $\Pi_1, \cdots, \Pi_n$ be n independent Poisson processes on $\mathbb{S}$ with $\Pi_k \sim PoissonP(\nu_k)$, then the superposition of these n Poisson processes is still a Poisson process with $\Pi \sim PoissonP(\sum_i \nu_i)$.*

Subsampling of Poisson processes Subsampling of a Poisson process with sampling rate $q(\theta)$ is defined to be selecting the points of the Poisson process via independent Bernoulli trials with acceptance rate $q(\theta)$.

Lemma 5 (Subsampling Theorem) *Let $\Pi \sim PoissonP(\nu)$ be a Poisson process on the space $\mathbb{S}$ and $q : \mathbb{S} \to [0, 1]$ be a measurable function. If we independently draw $z_\theta \in \{0, 1\}$ for each $\theta \in \Pi$ with $P(z_\theta = 1) = q(\theta)$, and let $\Pi_k = \{\theta \in \Pi : z_\theta = k\}$ for $k = 0, 1$, then Π_0, Π_1 are independent Poisson processes on $\mathbb{S}$ with $S^{1-q}(\Pi) := \Pi_0 \sim PoissonP((1 - q)\nu)$ and $S^q(\Pi) := \Pi_1 \sim PoissonP(q\nu)$.*

Point transition of Poisson processes Point transition of a Poisson process Π on space $(\mathbb{S}, \mathcal{S})$, denoted as $T(\Pi)$, is defined as moving each point of the Poisson process independently to other locations following a probabilistic transition kernel T, which is defined to be a function $\mathcal{T} : \mathbb{S} \times \mathcal{S} \to [0, 1]$ [3] such that for each $\theta \in \mathbb{S}$, $\mathcal{T}(\theta, \cdot)$ is a probability measure on E that describes the distribution of where the point θ moves, and for each $A \in \mathcal{S}$, $\mathcal{T}(\cdot, A)$ is integrable. Thus, $T(\Pi) := \{\theta' : \theta' \sim \mathcal{T}(\theta, \cdot)|\theta \in \Pi\}$. With a little abuse of notation, we use $\mathcal{T}(\theta)$ to denote a sample from $\mathcal{T}(\theta, \cdot)$ in this paper. Thus $\mathcal{T}(\theta)$ is a stochastic function.

[3] In the following we will use $T(\cdot)$ to denote the point transition operation, while use $\mathcal{T}(\cdot, \cdot)$ to denote the corresponding transition kernel.

Lemma 6 (Transition Theorem) *Let $\Pi \sim PoissonP(\nu)$ be a Poisson process on space $(\mathbb{S}, \mathcal{S})$, $\mathcal{T}$ a probability transition kernel, then*

$$T(\Pi) \sim PoissonP(\mathcal{T}\nu). \tag{18}$$

where $\mathcal{T}\nu$ can be considered as a transformation of measures over $\mathbb{S}$ defined as $(\mathcal{T}\nu)(A) := \int_{\mathbb{S}} \mathcal{T}(\theta, A)\nu(\mathrm{d}\theta)$ for $A \in \mathbb{S}$.

3.2 Operations on random measures

3.2.1 Operations on CRMs

The dependency operations defined on Poisson processes in Section 3.1 can be naturally generalized to the completely random measures given the construction in (3). Formally, we have

Superposition of CRMs Given n independent CRMs $\tilde{\mu}_1, \cdots, \tilde{\mu}_n$ on $\mathbb{X}$, the superposition ($\tilde{\oplus}$) is defined as:

$$\tilde{\mu}_1 \tilde{\oplus} \tilde{\mu}_2 \tilde{\oplus} \cdots \tilde{\oplus} \tilde{\mu}_n := \mu_1 + \mu_2 + \cdots + \mu_n \ .$$

Subsampling of CRMs Given a CRM $\tilde{\mu} = \sum_{k=1}^{\infty} J_k \delta_{\theta_k}$ on $\mathbb{X}$, and a measurable function $q : \mathbb{X} \to [0, 1]$. If we independently draw $z(\theta) \in \{0, 1\}$ for each $\theta \in \mathbb{X}$ with $p(z(\theta) = 1) = q(\theta)$, the subsampling of $\tilde{\mu}$, is defined as

$$\tilde{S}^q(\tilde{\mu}) := \sum_k z(\theta_k) J_k \delta_{\theta_k}, \tag{19}$$

Point transition of CRMs Given a CRM $\tilde{\mu} = \sum_{k=1}^{\infty} J_k \delta_{\theta_k}$ on $\mathbb{X}$, the point transition of $\tilde{\mu}$, is to draw atoms θ'_k from a transformed base measure to yield a new random measure as

$$\tilde{T}(\tilde{\mu}) := \sum_{k=1}^{\infty} J_k \delta_{\theta'_k}.$$

3.2.2 Operations on NRMs

The operations on NRMs can be naturally generalized from those on CRMs:

Superposition of NRMs Given n independent NRMs $\mu_1, \cdots, \mu_n$ on $\mathbb{X}$, the superposition ($\oplus$) is:

$$\mu_1 \oplus \mu_2 \oplus \cdots \oplus \mu_n := c_1\mu_1 + c_2\mu_2 + \cdots + c_n\mu_n \ .$$

where the weights $c_m = \frac{\tilde{\mu}_m(\mathbb{X})}{\sum_j \tilde{\mu}_j(\mathbb{X})}$ and $\tilde{\mu}_m$ is the unnormalized random measures corresponding to μ_m.

Subsampling of NRMs Given a NRM $\mu = \sum_{k=1}^{\infty} r_k \delta_{\theta_k}$ on $\mathbb{X}$, and a measurable function $q : \mathbb{X} \to [0,1]$. If we independently draw $z(\theta) \in \{0,1\}$ for each $\theta \in \mathbb{X}$ with $p(z(\theta) = 1) = q(\theta)$, the subsampling of μ, is defined as

$$S^q(\mu) := \sum_{k:z(\theta_k)=1} \frac{r_k}{\sum_j z(\theta_j) r_j} \delta_{\theta_k}, \tag{20}$$

Point transition of NRMs Given a NRM $\mu = \sum_{k=1}^{\infty} r_k \delta_{\theta_k}$ on $\mathbb{X}$, the point transition of μ, is to draw atoms θ'_k from a transformed base measure to yield a new NRM as

$$T(\mu) := \sum_{k=1}^{\infty} r_k \delta_{\theta'_k} \ .$$

The definitions are constructed so the following simple lemma holds.

Lemma 7 *Superposition, subsampling or point transition of NRMs is equivalent to superposition, subsampling or point transition of their underlying CRMs.*

Thus one does not need to distinguish between whether these operations are on CRMs or NRMs.

4 Posteriors for the NGG

This section develops posteriors for the single NGG, for a standard version $p\left(\vec{X}|\text{NGG}(a, M, H(\cdot)\right)$ and a version conditioned on the latent relative mass U_N, $p\left(\vec{X}|U_N, \text{NGG}(a, M, H(\cdot)\right)$. The second version is done because, as shown, the first version requires computing a complex recursive function.

4.1 Simple Posterior

James *et al.* [JLP09] develop posterior analysis as follows. This theorem simplifies their results and specialises them to the NGG.

Theorem 2 (Posterior Analysis for the NGG) *Consider the $NGG(a, M, H(\cdot))$. For a data vector $\vec{X}$ of length N there are K distinct values $X_1^*, ..., X_K^*$ with counts $n_1, ..., n_K$ respectively (where each $n_k > 0$). The posterior marginal is given by*

$$p\left(\vec{X}, K|NGG(a, M, H(\cdot)\right) \ = \ \frac{e^M T_{a,M}^{N,K}}{a^{N-K+1}} \prod_{k=1}^{K} (1-a)_{n_k-1} h(X_k^*) \ . \tag{21}$$

where

$$T_{a,M}^{N,K} \ = \ \frac{a^{N-1}}{\Gamma(N)} \int_M^{\infty} \left(1 - \left(\frac{M}{t}\right)^{1/a}\right)^{N-1} t^{K-1} e^{-t} \mathrm{d}t \tag{22}$$

is defined for $N \in \mathcal{N}^+$, $K \in \mathcal{R}$ and $M \in \mathcal{R}^+$ so $M > 0$. Moreover, the predictive posterior is given by:

$$p(X_{N+1} \in \mathrm{d}x | \vec{X}, NGG(a, M, H(\cdot)) \;=\; \omega_0 H(\mathrm{d}x) + \sum_{k=1}^{K} \omega_k \delta_{X_k^*}(\mathrm{d}x)$$

where the weights sum to 1 ($\sum_{k=0}^{K} \omega_k = 1$) are derived as

$$\begin{aligned} \omega_0 &\propto a \frac{T_{a,M}^{N+1,K+1}}{T_{a,M}^{N+1,K}} \\ \omega_k &\propto (n_k - a) \end{aligned} \tag{23}$$

Note, $T_{a,M}^{N,K}$ is a strictly decreasing function of N and M, but an increasing function of K and a. Moreover, an alternative definition of $T_{a,M}^{N,K}$ derived using the transformation $t = M(1+u)^a$ is

$$T_{a,M}^{N,K} \;=\; \frac{a^N M^K}{\Gamma(N) e^M} \int_{\mathbb{R}^+} \frac{u^{N-1}}{(1+u)^{N-Ka}} e^{M - M(1+u)^a} \mathrm{d}u \;,$$

and various scaled versions of this integral are presented in the literature. Introducing a $\Gamma(b/a, 1)$ prior on M and then marginalising out M makes the term in $e^{M-M(1+u)^a}$ disappear since the integral over M can be carried inside the integral over u.

Corollary 1 *Let $\vec{\mu} \sim NGG(a, M, H(\cdot))$ and suppose $M \sim \Gamma(b/a, 1)$ then it follows that $\vec{\mu} \sim PDP(a, b, H(\cdot))$*

For computation, the issue here will be computing the terms $T_{a,M}^{N,K}$. Therefore we present some results for this. These use $\Gamma(x, y)$, the upper incomplete Gamma function, defined for $y > 0$ and all real x.

Lemma 8 (Evaluating $T_{a,M}^{N,K}$) *Have $T_{a,M}^{N,K}$ defined as in Theorem 2. Then the following formula hold:*

$$T_{a,M}^{1,K} \;=\; \Gamma(K, M) \;, \tag{24}$$

$$T_{a,M}^{N,K} \;\leq\; \frac{a^{N-1}}{\Gamma(N)} T_{a,M}^{1,K} \;, \tag{25}$$

$$T_{a,M}^{N,K} \;=\; \frac{a^{N-1}}{\Gamma(N)} \sum_{n=0}^{N-1} (-1)^n \binom{N-1}{n} \Gamma\left(K - \frac{n}{a}, M\right) M^{n/a} \;, \tag{26}$$

$$T_{a,M}^{N-1,K-1} \;=\; T_{a,M}^{N,K} + \left(\frac{N-1}{a} - (K-1)\right) T_{a,M}^{N,K-1} \qquad \forall N \geq 2 \;. \tag{27}$$

A variant of Equation (26) only applies for $K \in \mathbb{N}^+$,

$$T_{a,M}^{N,K} \;=\; \frac{a^{N-1}}{\Gamma(N)} \sum_{n=0}^{N-1} (-1)^n \binom{N-1}{n} \left(1 - \frac{n}{a}\right)_{K-1} \Gamma\left(1 - \frac{n}{a}, M\right) M^{n/a} \;. \tag{28}$$

Other recursions involve factors of $1/a$ *and can be used when* $a = 1/R$ *for some* $R \in \mathbb{N}^+, R > 1$. *Note the function* $T_{a,M}^{N,K}$ *is well defined for non-integral* K. *Then*

$$T_{a,M}^{N,K} = \frac{a}{N-1}\left(T_{a,M}^{N-1,K} - M^{1/a}T_{a,M}^{N-1,K-1/a}\right) \quad \forall N \geq 2 \ , \quad (29)$$

$$T_{a,M}^{N,K} = (K-1)T_{a,M}^{N,K-1} + M^{1/a}\, T_{a,M}^{N-1,K-1-1/a} \quad \forall N \geq 2 \ . \quad (30)$$

Note the upper incomplete Gamma function becomes infinitesimal quickly for large y and negative x because $\Gamma(x,y) \to y^{x-1}e^{-y}$ as $y \to \infty$, and for positive y and $x \leq 1$, $\Gamma(x,y) \leq y^{x-1}e^{-y}$. As $y \to 0$ and $x < 0$, $\Gamma(x,y) \to -y^x/x$. Moreover, for $x < -1$ a good approximation is given by $\Gamma(x,y) \approx y^x e^{-y}/(y-x+1)$. This implies the series summation of Equation (26) will be unstable for large N since to a first approximation it is a binomial expansion of $(1-1)^N$. Experiments show this can happen for $N > 20$, so the summation is not practically useful but good for checking small values.

The recursion of Equation (27) recurses down on K. The inverted version, recursing up with $T_{a,M}^{N,K}$ on the left-hand side is unstable because it involves the subtraction of two terms, $T_{a,M}^{N-1,K-1}$ and $\left(\frac{N-1}{a} - (K-1)\right) T_{a,M}^{N,K-1}$. Thus errors magnify and it is not practically useful for $N > 20$. However, the inverted version shows that $T_{a,M}^{N,K}$ is related to a generalised Stirling number of the second kind.

Computing $T_{a,M}^{N,K}$ would go as follows. Fix an upper bound on K to be used, denote in as K_{max}. Values of $T_{a,M}^{N,K}$ need to be initialised for $K = K_{max} \& N \geq K_{max}$ and for $K < K_{max} \& N = K$. This can be done using either numerical integration or a saddle point approximation using Equation (22). The saddle point approximation requires an initial maximisation step, which can be done using Newton-Raphson convergence, and typically has 6-decimal place accuracy for $N > 50$. Thereafter the recursion of Equation (27) can be applied to recurse down on K.

Remark The Poisson-Dirichlet Process and Dirichlet Process are well known for their ease of use in a hierarchical context [TJBB06, CDB11, BH12]. The NGG has the same general form, which comes from it being a species sampling model.

The major issue with this posterior theory is that one needs to precompute the terms $T_{a,M}^{N,K}$. While the Poisson-Dirichlet Process has a similar style, it has a generalised Stirling number dependent only on the discount a [BH12]. The difference is that for the PDP we can tabulate these terms for a given discount parameter a and still vary the concentration parameter (b above, but corresponding to M) easily. For the NGG, any tables of $T_{a,M}^{N,K}$ would need to be recomputed with every change in mass parameter M. This might represent a significant computational burden.

4.2 Conditional Posterior

James *et al.* [JLP09] also develop conditional posterior analysis as follows. This theorem simplifies their results and specialises them to the NGG.

Theorem 3 (Conditional Posterior Analysis for the NGG) *Consider the $NGG_{a,M}$ and the situation of Theorem 2. The conditional posterior marginal, conditioned on the auxiliary variable U_N, is given by*

$$p\left(\vec{X}, K|U_N = u, NGG\left(a, M, H(\cdot)\right), N\right) \;=\; \frac{\left(Ma\left(1+u\right)^a\right)^K}{\sum_{k=1}^N S_{k,a}^N \left(Ma\left(1+u\right)^a\right)^k} \prod_{k=1}^K (1-a)_{n_k-1} h(X_k^*) \; . \tag{31}$$

Moreover, the predictive posterior is given by:

$$p\left(X_{N+1} \in \mathrm{d}x|\vec{X}, U_N = u, NGG\left(a, M, H(\cdot)\right), N\right) \;=\; \omega_0 H(\mathrm{d}x) + \sum_{k=1}^K \omega_k \delta_{X_k^*}(\mathrm{d}x)$$

where the weights sum to 1 ($\sum_{k=0}^K \omega_k = 1$) are derived as

$$\begin{aligned} \omega_0 &\propto Ma\left(1+u\right)^a \\ \omega_k &\propto n_k - a \; . \end{aligned} \tag{32}$$

The posterior for U_N is given by:

$$p\left(U_N = u|\vec{X}, NGG\left(a, M, H(\cdot)\right), N\right) \;=\; \frac{aM^K}{T_{a,M}^{N,K}} \frac{u^{N-1}}{\left(1+u\right)^{N-Ka}} e^{-M(1+u)^a} \; . \tag{33}$$

A posterior distribution is also presented by James *et al.* as their major result of Theorem 1 [JLP09]. We adapt it here to the NGG.

Theorem 4 *In the context of Theorem 3 the conditional posterior of the normalised random measure $\tilde{\mu}$ given data $\vec{X}$ of length N and latent relative mass $U_N = u$ is given by*

$$\vec{\mu} \;=\; \frac{T}{T+J_+}\vec{\mu}' + \frac{J^+}{T+J_+} \sum_{k=1}^K p_k \delta_{X_k^*}$$

where

$$\begin{aligned} \vec{\mu}' &\sim NGG\left(a, \frac{M}{1+u}, H(\cdot)\right) \; , \\ T &\sim f_T(t) \quad \textit{where Lévy measure of } f_T(t) = \frac{Ma}{\Gamma(1-a)} t^{-a-1} e^{-(1+u)t} \; , \\ J^+ &\sim \Gamma(N - Ka, 1+u) \; , \\ \vec{p} &\sim \textit{Dirichlet}_K\left(\vec{n} - a\right) \; . \end{aligned}$$

Here, $\vec{\mu}'$, J_+ and $\vec{p}$ are jointly independent and T, J_+ and $\vec{p}$ are jointly independent.

Note in particular the densities given for $\vec{\mu}'$ and T are not independent from each other. While an explicit density is not given for T, its expected value is easily computed via the Laplace transform as $Ma(1+u)^{a-1}$.

A joint form of the conditional posteriors presented in Theorem 3 can be developed, and can be derived from the general sampling form in Lemma 3. by marginalising out jumps J_k and then taking the limit as $L \to 0$. This matches the conditionals of Theorem 3 so is seen to be correct.

Corollary 2 (Collapsed Sampling Posterior) *In the context of Theorem 3, assume there are K jumps with attached data (J_k such that $n_k > 0$). The resultant posterior is as follows:*

$$p\left(\vec{X}, U_N = u, K \mid N, NGG(a, M, H(\cdot))\right) = \frac{u^{N-1}}{(1+u)^{N-Ka}} (Ma)^K e^{M-M(1+u)^a} \prod_{k=1}^{K} (1-a)_{n_k-1} h(X_k^*) \ . \tag{34}$$

Moreover, the posterior for jumps J_k with data count n_k given $\vec{X}, U_N = u, K, N$ is $J_k \sim Ga(n_k - a, 1 + u)$.

Remark With the use of the latent relative mass U_N, the NGG lends itself to hierarchical reasoning without a need to compute the recursive series $T_{a,M}^{N,K}$. This can be done with either the jumps integrated out, or the jumps retained.

5 Dependencies and Properties of Operations

This section presents a number of results to do with the operations applied to the NRMs. First dependencies such as covariances are presented. Then some further properties are developed for when the operations are used in a network.

5.1 Dependencies between NRMs via Operations

Properties of the NRMs here are given in terms of the Laplace exponent and its derivatives. In the Dirichlet process case, we have $\psi(v) = M \log(1 + v)$, while in the normalized generalized Gamma process case, we have $\psi_a(v) = M\left((1 + v)^a - 1\right)$. Because the dependencies involve the total masses significantly, we use a modified version of the Laplace exponent in all these results. Define $\tilde{\psi}_\eta(v) = \frac{1}{M}\psi_\eta(v)$, which has the mass removed.

Different from the Dirichlet process, the total masses M are no longer independent from their normalized jumps in general normalized random measures. However, we can still derive the correlations between different NRMs. The following Theorems summarize these results.

Lemma 9 (Mean and Variance of an NRM) *Given a normalized random measure μ on $\mathbb{X}$ with the underlying Lévy measure $\nu(\mathrm{d}t, \mathrm{d}x) = M\rho_\eta(\mathrm{d}t)P(\mathrm{d}x)$, for $\forall B \in \mathcal{B}(\mathbb{X})$. The mean of this NRM is given by*

$$\mathbb{E}[\mu(B)] = P(B) \ . \tag{35}$$

The variance of this NRM is given by

$$\begin{aligned} Var(\mu(B)) &= P(B)(P(B)-1)M \\ &\int_0^\infty v\tilde{\psi}''_\eta(v)\exp\left\{-M\tilde{\psi}_\eta(v)\right\}\mathrm{d}v \ . \end{aligned} \tag{36}$$

Remark For DP, the corresponding variances are:

$$\mathrm{Var}_{DP}(\mu(B)) = \frac{P(B)(1-P(B))}{M+1}.$$

For NGG, it is

$$\mathrm{Var}_{NGG}(\mu(B)) = P(B)(1-P(B))\frac{1-a}{a}e^M M^{\frac{1}{a}}|\Gamma(-\frac{1}{a},M)|.$$

For large M the upper incomplete Gamma function used here has the property that $e^M M^{1+\frac{1}{a}}|\Gamma(-\frac{1}{a},M)| \to 1$ and so we get for large M

$$\mathrm{Var}_{NGG}(\mu(B)) \to P(B)(1-P(B))\frac{1-a}{Ma} \ .$$

Theorem 5 (Dependency via superposition) *Suppose $\mu_i, i = 1,\cdots,n$ are n independent normalized random measures on $\mathbb{X}$ with the underlying Lévy measures $\nu_i(\mathrm{d}t,\mathrm{d}x) = M_i\rho_\eta(\mathrm{d}t)P(\mathrm{d}x)$, let $\mu = \mu_1 \oplus \cdots \oplus \mu_n$, $B \in \mathcal{B}(\mathbb{X})$, then the covariance between $\mu_k(k<n)$ and μ is*

$$\begin{aligned} &Cov(\mu_k(B),\mu(B)) = \\ &P(B)M_k\int_0^\infty \gamma(M_k,P(B),v)\exp\left\{-(\sum_{j\neq k}M_j)\tilde{\psi}_\eta(v)\right\}\mathrm{d}v \\ &+P(B)^2\left(\frac{2M_k}{\sum_j M_j}-1\right) \ . \end{aligned} \tag{37}$$

where

$$\begin{aligned} &\gamma(M_k,P(B),v) = \\ &\int_0^v \left(P(B)M_k\tilde{\psi}'_\eta(v_1)^2 - \tilde{\psi}''_\eta(v_1)\right)\exp\left\{-M_k\tilde{\psi}_\eta(v_1)\right\}\mathrm{d}v_1 \end{aligned} \tag{38}$$

Theorem 6 (Dependency via subsampling) *Let $\tilde{\mu}$ be a completely random measure on $\mathbb{X}$ with Lévy measure $\nu(\mathrm{d}t,\mathrm{d}x) = M\rho_\eta(\mathrm{d}t)P(\mathrm{d}x)$, $\mu = \frac{\tilde{\mu}}{\tilde{\mu}(\mathbb{X})}$. The covariance between μ and its subsampling version $S^q(\mu)$, denoted as μ^q, with sampling rate $q(\cdot)$ on $B \in \mathcal{B}(\mathbb{X})$ is*

$$\begin{aligned} &Cov(\mu^q(B),\mu(B)) = \\ &P(B)M_q\int_0^\infty \gamma(M_q,P(B),v)\exp\left\{-(M-M_q)\tilde{\psi}_\eta(v)\right\}\mathrm{d}v \\ &+ \ P(B)^2\left(\frac{2M_q-M}{M}\right) \ , \end{aligned} \tag{39}$$

where $M_q := (q\tilde{\mu})(\mathbb{X}) = \int_{\mathbb{X}} q(x)\tilde{\mu}(x)\mathrm{d}x$.

Theorem 7 (Dependency via point transition) *Let $\tilde{\mu}$ be a random measure on $\mathbb{X}$ with Lévy measure $\nu(\mathrm{d}t, \mathrm{d}x) = M\rho_\eta(\mathrm{d}t)P(\mathrm{d}x)$, $\mu = \frac{\tilde{\mu}}{\tilde{\mu}(\mathbb{X})}$. Let $B \in \mathcal{B}(\mathbb{X})$, $A = \mathcal{T}(B) := \{x : x \sim \mathcal{T}(y, \cdot), y \in B\}$ be the set of points obtained after the point transition on B, thus $P(A) = \int_B P(\mathcal{T}(x))\mathrm{d}x$. Suppose A and B are disjoint (which is usually the case when the transition operator T is appropriately defined), the covariance between μ and its point transition version $T(\mu)$ on $B \in \mathcal{B}(\mathbb{X})$ is*

$$Cov(\mu(B), (T\mu)(B)) = P(A)P(B) \left(M^2 \int_0^\infty \int_0^{v_1} \tilde{\psi}'_\eta(v_2)^2 \exp\left\{-M\tilde{\psi}_\eta(v_2)\right\} \mathrm{d}v_2 \mathrm{d}v_1 - 1 \right) \tag{40}$$

5.2 Properties of the three dependency operations

We first prove the following two Lemmas about superposition and subsampling of CRMs.

A straightforward extension of [Theorem 1 JLP09] leads to the following Lemma about the posterior of CRMs under superposition.

Lemma 10 (Posterior of CRMs under superposition) *Let $\tilde{\mu}_1, \tilde{\mu}_2, \cdots, \tilde{\mu}_n$ be n independent CRMs defined on space $\mathbb{X}$, with Lévy measures $\nu_i(\mathrm{d}t, \mathrm{d}x)$ for $i = 1, \cdots, n$. Let*

$$\tilde{\mu} = \oplus_{i=1}^n \tilde{\mu}_i. \tag{41}$$

Then given observed data $X = \{X_i\}$ (we use X_k^* to denote the distinct values among X) and a latent relative mass U_n, the posterior of $\tilde{\mu}$ is given by (we use $x|(y)$ to denote the variable x conditioned on y)

$$\tilde{\mu}|(U_n, X) = \tilde{\mu}|(U_n) + \sum_{k=1}^K J_k \delta_{X_k^*}, \tag{42}$$

where

1. *$\tilde{\mu}|(U_n)$ is a CRM with Lévy measure*

$$\nu(\mathrm{d}t, \mathrm{d}x) = e^{-ut} \left(\sum_{i=1}^n \nu_i(\mathrm{d}t, \mathrm{d}x) \right),$$

2. *X_k^* $(k = 1, \cdots, K)$ are the fixed points of discontinuity and J_k's are the corresponding jumps with densities proportional to*

$$t^{n_k} e^{-ut} \left(\sum_{i=1}^n \nu_i(\mathrm{d}t, \mathrm{d}x) \right),$$

where n_k is the number of data attached at jump J_k.

3. $\tilde{\mu}|(U_n)$ *and* J_k*'s are independent.*

We can prove the following formula of the Lévy measure under different dependency operations.

Lemma 11 (Lévy measure under dependency operations) *Let* $\tilde{\mu} = \sum_{k=1}^{\infty} J_k \delta_{X_k^*}$ *be a CRM with Lévy measure* $\nu(\mathrm{d}t, \mathrm{d}x)$.

- *Let* $S^q(\tilde{\mu})$ *be its subsampling version with acceptance rate* $q(\cdot)$*, then* $S^q(\tilde{\mu})$ *has the Lévy measure of* $q(\mathrm{d}x)\nu(\mathrm{d}t, \mathrm{d}x)$.
- *Let* $T(\tilde{\mu})$ *be its point transition version, where* $\nu(\mathrm{d}t, \mathrm{d}x) = M\rho_\eta(\mathrm{d}t)H(\mathrm{d}x)$. *Then its Lévy measure is* $M\rho_\eta(\mathrm{d}t)T(H)(\mathrm{d}x)$ *where* $T(H)$ *is the transformed base measure.*
- *Let* $\tilde{\mu}_1 \oplus \tilde{\mu}_2$ *be the superposition, then its Lévy measure is* $\nu_1(\mathrm{d}t, \mathrm{d}x) + \nu_2(\mathrm{d}t, \mathrm{d}x)$.

Now we give some properties about compositions of of the dependency operations which follow simply.

Lemma 12 (Composition of dependency operators) *Given CRMs* $\tilde{\mu}$*,* $\tilde{\mu}'$ *and* $\tilde{\mu}''$*, the following hold:*

- *Two subsampling operations are commutative. So with acceptance rates* $q(\cdot)$ *and* $q'(\cdot)$*, then* $S^{q'}(S^q(\tilde{\mu})) = S^q(S^{q'}(\tilde{\mu}))$. *Both are equal to* $S^{q'q}(\tilde{\mu}))$.
- *A constant subsampling operation commutes with a point transition operation. Thus* $S^q(T(\tilde{\mu})) = T(S^q(\tilde{\mu}))$ *where the acceptance rate* q *is indepenent of the data space.*
- *Subsampling and point transition operations distribute over superposition. Thus for acceptance rate* $q(\cdot)$ *and point transition* $T(\cdot)$*,*

$$S^q(\tilde{\mu} \oplus \tilde{\mu}') \;=\; S^q(\tilde{\mu}) \oplus S^q(\tilde{\mu}') \;, \qquad T(\tilde{\mu} \oplus \tilde{\mu}') \;=\; T(\tilde{\mu}) \oplus T(\tilde{\mu}') \;.$$

- *Superposition is commutative and associative. Thus* $\tilde{\mu} \oplus \tilde{\mu}' = \tilde{\mu}' \oplus \tilde{\mu}$ *and* $(\tilde{\mu} \oplus \tilde{\mu}') \oplus \tilde{\mu}'' = \tilde{\mu} \oplus (\tilde{\mu}' \oplus \tilde{\mu}'')$.

Thus when subsampling operations are all constant, a composition of subsampling, point transition and superposition operations admits a normal form where all the subsampling operations are applied first, then the transition operations and lastly the superposition operations.

Lemma 13 (Normal form for compositions) *Assume subsampling operations all have a constant acceptance rate. A normal form for a composition of subsampling,*

point transition and superposition operations is obtained by applying the following rules until no further can apply.

$$\begin{aligned} S^q(S^{q'}(\tilde{\mu})) &\rightarrow S^{qq'}(\tilde{\mu}) \ , \\ S^q(T(\tilde{\mu})) &\rightarrow T(S^q(\tilde{\mu})) \ , \\ S^q(\tilde{\mu} \oplus \tilde{\mu}') &\rightarrow S^q(\tilde{\mu}) \oplus S^q(\tilde{\mu}') \ , \\ T(\tilde{\mu} \oplus \tilde{\mu}') &\rightarrow T(\tilde{\mu}) \oplus T(\tilde{\mu}') \ . \end{aligned}$$

The remaining top level set of superpositions are then flattened out by removing any precedence ordering.

Note that Lemmas 10, 11, 12 and 13 all apply to NRMs as well due to Lemma 7. Now it is ready to state the main theorem about the relation between the CRM and the corresponding NRM under the three dependency operations.

Theorem 8 (Equivalence Theorem) *Assume the subsampling rates $q(\cdot)$ are independent (constant)[4] for each point of the corresponding Poisson process, the following dependent random measures (43) and (44) are equivalent:*

- *Manipulate the normalized random measures:*

$$\mu'_m \sim T(S^q(\mu'_{m-1})) \oplus \mu_m, \qquad \text{for } m > 1. \tag{43}$$

- *Manipulate the completely random measures:*

$$\begin{aligned} &\tilde{\mu}'_m \sim \tilde{T}(\tilde{S}^q(\tilde{\mu}'_{m-1})) \oplus \tilde{\mu}_m, \qquad \text{for } m > 1. \\ &\mu'_m = \frac{\tilde{\mu}'_m}{\tilde{\mu}'_m(\mathbb{X})}, \end{aligned} \tag{44}$$

Furthermore, both resulting NRMs μ'_m's correspond to:

$$\mu'_m = \sum_{j=1}^{m} \frac{(q^{m-j}\tilde{\mu}_j)(\mathbb{X})}{\sum_{j'=1}^{m}(q^{m-j'}\tilde{\mu}_{j'})(\mathbb{X})} T^{m-j}(\mu_j), \qquad \text{for } m > 1$$

where $q^{m-j}\tilde{\mu}$ is the random measure with Lévy measure $q^{m-j}(\mathrm{d}x)\nu(\mathrm{d}t, \mathrm{d}x)$, and $\nu(\mathrm{d}t, \mathrm{d}x)$ is the Lévy measure of $\tilde{\mu}$. $T^{m-j}(\mu)$ denotes point transition on μ for $(m-j)$ times .

Finally, in the posterior sampling for subsampling operation, we can prove the following posterior of the Bernoulli variables.

[4]This assumption is to deal with the case when considering point transition, meaning we can drop this assumption if no point transition operation is considered.

Theorem 9 (Posterior acceptance rates for subsampling) *Let $\tilde{\mu}' = \sum_k J_k \theta_k$ be a completely random measure on $\mathbb{X}$, $\tilde{\mu} = S^q(\tilde{\mu}') := \sum_k z_k J_k \delta_k$ be its subsampling version, where z_k's are independent Bernoulli random variables with acceptance rate q. Further define $\mu = \frac{\tilde{\mu}}{\tilde{\mu}(\mathbb{X})}$. Given $n = \sum_k n_k$ observed data in μ, the posterior of z_k is:*

$$p(z_k = 1|\tilde{\mu}, n) = \begin{cases} 1 & \text{if } n_k > 0, \\ \frac{q/J}{q/J+(1-q)/J^{-k}} & \text{if } n_k = 0. \end{cases} \tag{45}$$

where $J = (\sum_{k'} z_{k'} J_{k'})^n$, $J^{-k} = \left(\sum_{k' \neq k} z_{k'} J_{k'}\right)^n$.

Based on the above result, we give the posterior formula of the acceptance indicator z_{mk} used in our paper *dependent hierarchical normalized random measures for dynamic topic modeling* [CDB12]. Note this is only an approximated posterior and is lacked of theoretical guarantee. Exact posterior will be given in our future work.

Corollary 3 (Posterior acceptance rates in sampling J'_{mk} in Section 4 [CDB12]) *Using the terminology as in Section 4 [CDB12], the posterior $p(z_{mk} = 1|\tilde{\mu}_m, \{\tilde{n}'_{mk}\})$ is computed as:*

- *If $\tilde{n}'_{mk} > 0$, then $p(z_{mk} = 1|\tilde{\mu}_m, \{\tilde{n}'_{mk}\}) = 1$.*
- *Otherwise,*

$$p(z_{mk} = 1|\tilde{\mu}_m, \{\tilde{n}'_{mk}\}) = \frac{q^{m-m'}/J_m}{q^{m-m'}/J_m + (1-q^{m-m'})/J_m^{-k}},$$

where $J_m = \left(\sum_{m' \leq m} \sum_{k'} z_{m'k'} J_{m'k'}\right)^{\tilde{n}'_{m\cdot}}$, $J_m^{-k} = \left(\sum_{m' \leq m} \sum_{k' \neq k} z_{m'k'} J_{m'k'}\right)^{\tilde{n}'_{m\cdot}}$, and $\tilde{n}'_{m\cdot} = \sum_{k'} \tilde{n}'_{mk'}$.

Acknowledgements. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

[A. 12] A. Lijoi and B. Nipoti and I. Prünster. Bayesian inference with dependent normalized completely random measures. *working paper*, 2012.

[BH12] W. Buntine and M. Hutter. A Bayesian view of the Poisson-Dirichlet process. Technical Report arXiv:1007.0296, NICTA and ANU, Australia, February 2012.

[Bri99] A. Brix. Generalized Gamma measures and shot-noise Cox processes. *Advances in Applied Probability*, 31:929–953, 1999.

[Ç10] E. Çinlar. *Probability and stochastics*. Springer, 2010.

[CDB11] C. Chen, L. Du, and W. Buntine. Sampling table configurations for the hierarchical Poisson-Dirichlet process. In *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, pages 296–311. 2011.

[CDB12] C. Chen, N. Ding, and W. Buntine. Dependent hierarchical normalized random measures for dynamic topic modeling. In *International conference on machine learning (ICML)*. 2012.

[ELP03] E. R. Egazzini, A. Lijoi, and I. Prünster. Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics*, 31(2):560–585, 2003.

[FK72] T.S. Ferguson and M.J. Klass. A representation of independent increment processes without Gaussian component. *The Annals of Mathematical Statistics*, 43(5):1634–1643, 1972.

[GKS11] J. E. Griffin, M. Kolossiatis, and M. F. J. Steel. Comparing distributions using dependent normalized random measure mixtures. *working paper*, 2011.

[GS09] J. E. Griffin and M. F. J Steel. Time-dependent stick-breaking processes. *working paper*, 2009.

[GW11] J.E. Griffin and S.G. Walker. Posterior simulation of normalized random measure mixtures. *Journal of Computational and Graphical Statistics*, 20(1):241–259, 2011.

[Jam05] L. F. James. Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. *The Annals of Statistics*, 33(4):1771–1799, 2005.

[JLP06] L.F. James, A. Lijoi, and I. Prünster. Conjugacy as a distinctive feature of the Dirichlet process. *Scandinavian Journal of Statistics*, 33:105–120, 2006.

[JLP09] L.F. James, A. Lijoi, and I. Prünster. Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36:76–97, 2009.

[Kin67] J.F.C. Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967.

[Kin93] J.F.C. Kingman. *Poisson Processes*. Oxford University Press, 1993.

[LGF10] D. Lin, E. Grimson, and J. Fisher. Construction of dependent Dirichlet processes based on Poisson processes. In *Neural Information Processing Systems (NIPS)*. 2010.

[LMP07] A. Lijoi, R.H. Mena, and I. Prünster. Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of Royal Statistical Society B*, 69(4):715–740, 2007.

[LP10] A. Lijoi and I. Prünster. *Models beyond the Dirichlet process*. Cambridge University Press, 2010.

[Mac99] S. N. MacEachern. Dependent nonparametric processes. In *Proceedings of the Section on Bayesian Statistical Science*. 1999.

[Mac00] S. N. MacEachern. Dependent dirichlet processes. Technical report, Ohio State University, 2000.

[PY97] J. Pitman and M. Yor. The two-parameter Poisson-Diriclet distribution derived from a stable subordinator. *Annals of Probability*, 25(2):855–900, 1997.

[TJBB06] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

A Proofs

Proof of Lemma 2 We have $\nu(\mathrm{d}x, \mathrm{d}t/\lambda)$. Doing a change of variables $t' = t/\lambda$ and some rearranging of the Lévy-Khintchine formula yields the following:

$$\mathbb{E}\left[e^{-\int_{\mathbb{X}}(\lambda f(x))(\tilde{\mu}(\mathrm{d}x)/\lambda)}\right] = e^{-\int_{\mathbb{R}^+\times\mathbb{X}}\left(1-e^{-t'(\lambda f(x))}\right)\lambda\nu(\mathrm{d}x,\mathrm{d}t')}$$

Since $\tilde{\mu}(\mathrm{d}x)/\lambda$ normalises to the same measure as $\tilde{\mu}(\mathrm{d}x)$, and saying something holds for any $f(x)$ is the same as saying something holds for any $\lambda f(x)$ (when $\lambda > 0$), the result follows.

Proof of Lemma 3 First, for the infinite mixture model, we have infinite number of components, thus given the observed data $(x_1, \cdots, x_N)$ and their allocation indicators $\vec{s}$, the model likelihood is

$$f_\mu(\vec{x}, \vec{s}|\vec{\theta}, \vec{J}) = \prod_{i=1}^{N} \frac{J_{s_i}}{J_+} g_0(x_i|\theta_{s_i}),$$

where $J_+ = \sum_{k=1}^{\infty} J_k$. Now introduce the slice auxiliary variables $\vec{u}$ for each data, such that we only consider the components whose jumps are larger than a threshold u_i for data x_i, in this auxiliary space we have

$$f_\mu(\vec{x}, \vec{u}, \vec{s}|\vec{\theta}, \vec{J}) = \frac{1}{J_+^N} \prod_{i=1}^{N} 1(u_i < J_{s_i}) g_0(x_i|\theta_{s_i}).$$

Now using the fact that

$$\frac{1}{J_+^N} = \frac{\int_0^\infty U_N^{N-1} \exp\{-U_N J_+\} \,\mathrm{d}U_N}{\Gamma(N)},$$

after introducing the auxiliary variable U_N, we have

$$f_\mu(\vec{x}, \vec{u}, \vec{s}, U_N|\vec{\theta}, \vec{J}) \propto U_N^{N-1} \exp\{-U_N J_+\} \prod_{i=1}^{N} 1(u_i < J_{s_i}) g_0(x_i|\theta_{s_i}).$$

Further decomposing J_+ as

$$J_+ = J^* + \sum_{k=1}^{K} J_k,$$

where K is the number of jumps which are large than a threshold L, $J^* = \sum_{k=K+1}^{\infty} J_k$, then we get

$$\begin{aligned} & f_\mu(\vec{x}, \vec{u}, \vec{s}, U_N|\vec{\theta}, J_1, \cdots, J_K, K) \\ \propto\ & U_N^{n-1} \exp\left\{-U_N \sum_{k=1}^{K} J_k\right\} \mathbb{E}\left[\exp\{-U_N J^*\}\right] \prod_{i=1}^{N} 1(u_i < J_{s_i}) g_0(x_i|\theta_{s_i}). \quad (46) \end{aligned}$$

Now use the Lévy-Khintchine representation of a Lévy process (4) to evaluate $\mathbb{E}\left[\exp\left\{-U_N J^*\right\}\right]$, we get

$$f_\mu(\vec{x},\vec{u},\vec{s},U_N|\vec{\theta},J_1,\cdots,J_K,K) \propto U_N^{N-1}\exp\left\{-U_N\sum_{k=1}^{K}J_k\right\}$$
$$\exp\left\{-M\int_0^L\left(1-\exp\left\{-U_N t\right\}\right)\rho_\eta(t)\mathrm{d}t\right\}\prod_{i=1}^{N}1(u_i<J_{s_i})g_0(x_i|\theta_{s_i}). \quad (47)$$

Now combining with the priors

$$p(J_1,\cdots,J_K)=\prod_{k=1}^{K}\frac{\rho_\eta(J_k)}{\int_L^\infty\rho_\eta(t)\mathrm{d}t},$$

$$K\sim\text{Poisson}(M\int_L^\infty\rho_\eta(\mathrm{d}t)),\qquad \theta_k\sim h(\theta_k),$$

the result follows.

Proof of Theorem 2 The definition for $\tau_n(u)$ comes from [Proposition 1][JLP09]. The posterior marginal of Equation (21) comes from [Proposition 3][JLP09] and is simplified using the change of variables $t=M\left(1+u\right)^a$. For the predictive posterior, the weights in Equation (23) are derived directly from the posterior. The posterior proportionality for $p(U_N=u|\vec{X},\text{NGG}(a,M,H(\cdot)))$ discards terms not containing u.

Proof of Corollary 1 Marginalise out M from the posterior of Equation (21) using the alternative definition of $T_{a,M}^{N,K}$. It can be seen this yields the posterior of a Poisson-Dirichlet distribution with discount parameter a and concentration parameter b. Since the posteriors are equivalent for all data, the distributions are equivalent almost surely.

Proof of Lemma 8 Equation (25) holds by noticing $T_{a.M}^{N,K}$ is decreasing in N and then using the definition of the upper incomplete Gamma function. To prove Equation (26), expand the term $\left(1-\left(\frac{M}{t}\right)^{1/a}\right)^{N-1}$ using the binomial expansion and absorbing the powers $t^{-n/a}$ into t^{K-1} as an incomplete Gamma integral.

Now manipulate Equation (26). Expand $\Gamma\left(K-\frac{n}{a},M\right)$ using the recursion for the upper incomplete Gamma function, which can be applied for all first arguments when $M>0$.

$$\begin{aligned}
&= \sum_{n=0}^{N-1}\binom{N-1}{n}\left(-M^{1/a}\right)^n\left(\left(K-1-\frac{n}{a}\right)\Gamma\left(K-1-\frac{n}{a},M\right)+M^{K-1-\frac{n}{a}}e^{-M}\right)\\
&= \sum_{n=0}^{N-1}\binom{N-1}{n}\left(-M^{1/a}\right)^n\left(K-1-\frac{n}{a}\right)\Gamma\left(K-1-\frac{n}{a},M\right)+M^{K-1}e^{-M}\sum_{n=0}^{N-1}\binom{N-1}{n}(-1)^n
\end{aligned}$$

The second sum is a binomial expansion of $(1-1)^{N-1}$ and therefore disappears. Apply this step repeatedly. For $K \in \mathcal{N}+$, this terminates after $K-1$ steps to get Equation (28).

Equation (29) holds by expanding

$$\left(1-\left(\frac{M}{t}\right)^{1/a}\right)^{N-1} = \left(1-\left(\frac{M}{t}\right)^{1/a}\right)^{N-2} - \left(1-\left(\frac{M}{t}\right)^{1/a}\right)^{N-2}\left(\frac{M}{t}\right)^{1/a}$$

inside the integral definition of $T_{a,M}^{N+1,K}$ and then recognising the terms.

Equation (27) and Equation (30) hold by applying the integration by parts formula on the terms $A(t) = \left(1-\left(\frac{M}{t}\right)^{1/a}\right)^{N-1}$ and $B(t) = t^{K-1}e^{-t}$. Rearranging the resultant integrals and recognising the terms yields

$$0 = M^{1/a}T_{a,M}^{N-1,K-1-1/a} + (K-1)T_{a,M}^{N,K-1} - T_{a,M}^{N,K} .$$

This proves Equation (30). Equation (27) follows by then applying Equation (29).

Proof of Theorem 3 The posterior marginal of Equation (31) comes from [Proposition 4][JLP09]. Although the denominator is difficult to evaluate, and it can be derived through a recursion, the easiest way is simply to normalise the renumerator. Sum over $(Ma\,(1+u)^a)^K \prod_{k=1}^{K}(1-a)_{n_k-1}$ for all length K partitions $(n_1, n_2, ..., n_K)$ yields $(Ma\,(1+u)^a)^K\, S_{K,a}^N$ and the result follows by again summing over K. The predictive posterior, as before, follows directly from the posterior marginal. The posterior proportionality for U_N, $p(U_N = u|\vec{X}, \mathrm{NGG}(a, M, H(\cdot)))$, comes from [Proposition 4][JLP09] after discarding terms not containing u. The normalising constant is obtained using the methods of Theorem 2.

Proof of Corollary 2 Equation (34) can be seen to hold true since conditioning it on $U_N = u$ and $\vec{X}$ yields respectively Equation (31) and Equation (33). The posterior on J_k comes from [GW11].

This can also be proven from [GW11] at the end of Section 3, and includes the prior on $K_L, J_1, ..., J_K$ described in Section 4. The mixture model component $k(y_i|\theta_{s_i})$ has also been stripped and the slice sampling variables marginalised out. One then takes the limit as $L \to 0$.

Proof of Lemma 9 This uses a similar technique to that of Theorem 1 in [GKS11]. Using the identity $1/b = \int_0^\infty e^{-v\,b}\mathrm{d}v$ we get

$$\begin{aligned}\mathbb{E}\left[\mu(B)\right] &= \mathbb{E}\left[\frac{\tilde{\mu}(B)}{\tilde{\mu}(\mathbb{X})}\right] \\ &= \int_0^\infty \mathbb{E}\left[\tilde{\mu}(B)\exp\left\{-v\tilde{\mu}(B)\right\}\right]\mathbb{E}\left[\exp\left\{-v\tilde{\mu}(\mathbb{X}\setminus B)\right\}\right]\mathrm{d}v\ . \qquad (48)\end{aligned}$$

According to the Lévy-Khintchine representation of $\tilde{\mu}$ and definition (6), we have

$$\mathbb{E}\left[\exp\left\{-v\tilde{\mu}(B)\right\}\right] = \exp\left\{-P(B)M\tilde{\psi}_\eta(v)\right\} \tag{49}$$

$$\mathbb{E}\left[\tilde{\mu}(B)\exp\left\{-v\tilde{\mu}(B)\right\}\right] = -\mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}v}\exp\left\{-v\tilde{\mu}(B)\right\}\right]$$

$$= P(B)M\tilde{\psi}'_\eta(v)\exp\left\{-P(B)M\tilde{\psi}_\eta(v)\right\} \tag{50}$$

$$\mathbb{E}\left[\tilde{\mu}(B)^2\exp\left\{-v\tilde{\mu}(B)\right\}\right] = \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}v^2}\exp\left\{-v\tilde{\mu}(B)\right\}\right]$$

$$= \left(P(B)^2M^2\left(\tilde{\psi}'_\eta(v)\right)^2 - P(B)M\tilde{\psi}''_\eta(v)\right)\exp\left\{-P(B)M\tilde{\psi}_\eta(v)\right\} \tag{51}$$

Substituting (49) and (50) into (48) and using the fact in (7), after simplifying we have

$$\mathbb{E}\left[\mu(B)\right] = P(B).$$

Since $\mathrm{Var}\left(\mu(B)\right) = \mathbb{E}\left[\mu(B)^2\right] - \left(\mathbb{E}\left[\mu(B)\right]\right)^2$, and the last term is equal to $\left(P(B)\right)^2$, we now deal with the first term.

$$\mathbb{E}\left[\mu(B)^2\right] = \mathbb{E}\left[\frac{\tilde{\mu}(B)^2}{\tilde{\mu}(\mathbb{X})^2}\right]$$

$$= \int_0^\infty\int_0^\infty \mathbb{E}\left[\tilde{\mu}(B)^2 \times \exp\left\{-v_1\tilde{\mu}(\mathbb{X}) - v_2\tilde{\mu}(\mathbb{X})\right\}\right]\mathrm{d}v_1\mathrm{d}v_2 \tag{52}$$

$$= \int_0^\infty\int_0^\infty \mathbb{E}\left[\tilde{\mu}(B)^2\exp\left\{-(v_1+v_2)\tilde{\mu}(B)\right\}\right]\mathbb{E}\left[\exp\left\{-(v_1+v_2)\tilde{\mu}(\mathbb{X}\setminus B)\right\}\right]\mathrm{d}v_1\mathrm{d}v_2$$

Substituting (49)(51) into (52) we have

$$(52) = \int_0^\infty\int_0^\infty \left[P(B)^2M^2\left(\tilde{\psi}'_\eta(v_1+v_2)\right)^2 - P(B)M\tilde{\psi}''_\eta(v_1+v_2)\right] \exp\left\{-M\tilde{\psi}_\eta(v_1+v_2)\right\}\mathrm{d}v_1\mathrm{d}v_2\ . \tag{53}$$

Furthermore, let $v = v_1 + v_2, B = \mathbb{X}$ in (51), after integrating out v_1, v_2 in $[0,\infty]$, we have

$$\int_0^\infty\int_0^\infty M^2\left(\tilde{\psi}'_\eta(v_1+v_2)\right)^2\exp\left\{-M\tilde{\psi}_\eta(v_1+v_2)\right\}\mathrm{d}v_1\mathrm{d}v_2 \tag{54}$$

$$= 1 + \int_0^\infty\int_0^\infty M\tilde{\psi}''_\eta(v_1+v_2)\exp\left\{-M\tilde{\psi}_\eta(v_1+v_2)\right\}\mathrm{d}v_1\mathrm{d}v_2$$

Substitute (54) into (53) and simplify we get

$$\mathrm{Var}(\mu(B)) =$$

$$P(B)(1-P(B))M\int_0^\infty\int_0^\infty -\tilde{\psi}''_\eta(v_1+v_2)\exp\left\{-M\tilde{\psi}_\eta(v_1+v_2)\right\}\mathrm{d}v_1\mathrm{d}v_2\ . \tag{55}$$

Now use a change of variables, let $v'_1 = v_1, v'_2 = v_1 + v_2$ and simplify we get the result of (36).

Proof of Theorem 5 From the definition we have

$$\begin{aligned}
\mathrm{Cov}\left(\mu_k(B), \mu(B)\right) &= \sum_{i=1}^{n} \mathrm{Cov}\left(\frac{M_i}{\sum_j M_j}\mu_i(B), \mu_k(B)\right) \\
&= \mathrm{Cov}\left(\frac{M_k}{\sum_j M_j}\mu_k(B), \mu_k(B)\right) + \sum_{i\neq k} \mathrm{Cov}\left(\frac{M_i}{\sum_j M_j}\mu_i(B), \mu_k(B)\right) \qquad (56) \\
&= \mathbb{E}\left[\frac{\tilde{\mu}_k(B)^2}{\left(\sum_j \tilde{\mu}_j(\mathbb{X})\right)\tilde{\mu}_k(\mathbb{X})}\right] - \mathbb{E}\left[\frac{\tilde{\mu}_k(B)}{\sum_j \tilde{\mu}_j(\mathbb{X})}\right]\mathbb{E}\left[\frac{\tilde{\mu}_k(B)}{\tilde{\mu}_k(\mathbb{X})}\right] \\
&+ \sum_{i\neq k}\left\{\mathbb{E}\left[\frac{\tilde{\mu}_i(B)\tilde{\mu}_k(B)}{\left(\sum_j \tilde{\mu}_j(\mathbb{X})\right)\tilde{\mu}_k(\mathbb{X})}\right] - \mathbb{E}\left[\frac{\tilde{\mu}_i(B)}{\sum_j \tilde{\mu}_j(\mathbb{X})}\right]\mathbb{E}\left[\frac{\tilde{\mu}_k(B)}{\tilde{\mu}_k(\mathbb{X})}\right]\right\}
\end{aligned}$$

Note that for the Dirichlet process, the last $n-1$ terms of (56) vanish because μ_i's are independent from their total mass M_i's, but this is not the case for general NRMs. Now we calculate these term by term.

For the first term, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{\tilde{\mu}_k(B)^2}{\left(\sum_j \tilde{\mu}_j(\mathbb{X})\right)\tilde{\mu}_k(\mathbb{X})}\right] \\
&= \int_0^\infty\int_0^\infty \mathbb{E}\left[\tilde{\mu}_k(B)^2 \exp\left\{-v_1(\sum_j \tilde{\mu}_j)(\mathbb{X}) - v_2\tilde{\mu}_k(\mathbb{X})\right\}\right]\mathrm{d}v_1\mathrm{d}v_2 \\
&= \int_0^\infty\int_0^\infty \mathbb{E}\left[\tilde{\mu}_k(B)^2 \exp\left\{-(v_1+v_2)\tilde{\mu}_k(B)\right\}\right]\mathbb{E}\left[\exp\left\{-(v_1+v_2)\tilde{\mu}_k(\mathbb{X}\setminus B)\right\}\right] \\
&\quad \mathbb{E}\left[\exp\left\{-v_1(\sum_{j\neq k}\tilde{\mu}_j(\mathbb{X}))\right\}\right]\mathrm{d}v_1\mathrm{d}v_2 \\
&= \int_0^\infty\int_0^{v_2} \left(P(B)^2 M_k^2 \tilde{\psi}_\eta'(v_1)^2 - P(B)M_k\tilde{\psi}_\eta''(v_1)\right)\exp\left\{-M_k\tilde{\psi}_\eta(v_1)\right\} \\
&\quad \exp\left\{-(\sum_{j\neq k} M_j)\tilde{\psi}_\eta(v_2)\right\}\mathrm{d}v_1\mathrm{d}v_2 \\
&= P(B)M_k\int_0^\infty \gamma(M_k, P(B), v)\exp\left\{-(\sum_{j\neq k} M_j)\tilde{\psi}_\eta(v)\right\}\mathrm{d}v \qquad (57)
\end{aligned}$$

For the second term, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{\tilde{\mu}_k(B)}{\sum_j \tilde{\mu}_j(\mathbb{X})}\right]\mathbb{E}\left[\frac{\tilde{\mu}_k(B)}{\tilde{\mu}_k(\mathbb{X})}\right] &= P(B)\int_0^\infty \mathbb{E}\left[\tilde{\mu}_k(B)\exp\left\{-v\sum_j \tilde{\mu}_j(\mathbb{X})\right\}\right]\mathrm{d}v \\
&= P(B)^2 M_k \int_0^\infty \tilde{\psi}'_\eta(v)\exp\left\{-(\sum_j M_j)\tilde{\psi}_\eta(v)\right\}\mathrm{d}v \\
&= \frac{P(B)^2 M_k \exp\left\{-\left(\sum_j M_j\right)\tilde{\psi}_\eta(0)\right\}}{\sum_j M_j} \\
&= \frac{P(B)^2 M_k}{\sum_j M_j} \qquad (58)
\end{aligned}$$

For the third term, similarly

$$\begin{aligned}
&\mathbb{E}\left[\frac{\tilde{\mu}_i(B)\tilde{\mu}_k(B)}{\left(\sum_j \tilde{\mu}_j(\mathbb{X})\right)\tilde{\mu}_k(\mathbb{X})}\right] \\
=& \int_0^\infty\int_0^\infty \mathbb{E}\left[\tilde{\mu}_i(B)\tilde{\mu}_k(B)\exp\left\{-v_1(\sum_j \tilde{\mu}_j)(\mathbb{X}) - v_2\tilde{\mu}_k(\mathbb{X})\right\}\right]\mathrm{d}v_1\mathrm{d}v_2 \\
=& \int_0^\infty\int_0^\infty \mathbb{E}\left[\tilde{\mu}_k(B)\exp\left\{-(v_1+v_2)\tilde{\mu}_k(B)\right\}\right]\mathbb{E}\left[\exp\left\{-(v_1+v_2)\tilde{\mu}_k(\mathbb{X}\setminus B)\right\}\right] \\
&\mathbb{E}\left[\tilde{\mu}_i(B)\exp\left\{-v_1\tilde{\mu}_i(B)\right\}\right]\mathbb{E}\left[\exp\left\{-v_1\tilde{\mu}_i(\mathbb{X}\setminus B)\right\}\right]\mathbb{E}\left[\exp\left\{-v_1(\sum_{j\neq\{i,k\}}\tilde{\mu}_j(\mathbb{X}))\right\}\right]\mathrm{d}v_1\mathrm{d}v_2 \\
=& \int_0^\infty\int_0^\infty P(B)M_k\tilde{\psi}'_\eta(v_1+v_2)\exp\left\{-M_k\tilde{\psi}_\eta(v_1+v_2)\right\} \\
&P(B)M_i\tilde{\psi}'_\eta(v_1)\exp\left\{-M_i\tilde{\psi}_\eta(v_1)\right\} \\
&\exp\left\{-(\sum_{j\neq\{i,k\}}M_j)\tilde{\psi}_\eta(v_1)\right\}\mathrm{d}v_1\mathrm{d}v_2 \\
=& P(B)^2 M_i M_k \int_0^\infty \tilde{\psi}'_\eta(v_1)\exp\left\{-(\sum_{j\neq k}M_j)\tilde{\psi}_\eta(v_1)\right\}\int_0^{v_1}\tilde{\psi}'_\eta(v_2)\exp\left\{-M_k\tilde{\psi}_\eta(v_2)\right\}\mathrm{d}v_2\mathrm{d}v_1 \\
=& P(B)^2 M_i\left(\frac{1}{\sum_{j\neq k}M_j} - \frac{1}{\sum_j M_j}\right)\exp\left\{-(\sum_j M_j)\tilde{\psi}_\eta(0)\right\} \\
=& P(B)^2 M_i\left(\frac{1}{\sum_{j\neq k}M_j} - \frac{1}{\sum_j M_j}\right) \qquad (59)
\end{aligned}$$

The fourth term is similar to the second term, and is equal to

$$\begin{aligned}\mathbb{E}\left[\frac{\tilde{\mu}_i(B)}{\sum_j \tilde{\mu}_j(\mathbb{X})}\right]\mathbb{E}\left[\frac{\tilde{\mu}_k(B)}{\tilde{\mu}_k(\mathbb{X})}\right] &= \frac{P(B)^2 M_i \exp\left\{-\left(\sum_j M_j\right)\tilde{\psi}_\eta(0)\right\}}{\sum_j M_j} \\ &= \frac{P(B)^2 M_i}{\sum_j M_j} \end{aligned} \tag{60}$$

The result follows.

Proof of Theorem 6 By subsampling, we obtain two independent NRMs μ^q and μ_0^q, corresponding to those points selected and those rejected by the independent Bernoulli trials, respectively.

We denote the total mass of the corresponding unnormalized μ^q as M_q, and M_q^0 for μ_0^q. From the definition of subsampling, we have

$$M_q := (q\tilde{\mu})(\mathbb{X}) = \int_{\mathbb{X}} q(x)\tilde{\mu}(x)\mathrm{d}x,$$

$$M_q^0 = M - M_q.$$

Furthermore, notice that the original NRM μ is the superposition of μ^q and μ_0^q. Thus according to Theorem 5, the covariance between μ and μ^q is

$$P(B)M_q \int_0^\infty \gamma(M_q, P(B), v)\exp\left\{-(M-M_q)\tilde{\psi}_\eta(v)\right\}\mathrm{d}v + P(B)^2\left(\frac{2M_q - M}{M}\right),$$

Proof of Theorem 7 Note that $\tilde{\mu}$ and $\tilde{\mu}'$ are not independent, thus they can not be separated when taking the expectation. Now let A and B are defined as in the theorem, then:

$$\begin{aligned}&\mathbb{E}\left[\mu(B)\left((T\mu)(B)\right)\right] = \mathbb{E}\left[\frac{\tilde{\mu}(B)}{\tilde{\mu}(\mathbb{X})}\frac{\tilde{\mu}'(B)}{\tilde{\mu}'(\mathbb{X})}\right] = \mathbb{E}\left[\frac{\tilde{\mu}(B)}{\tilde{\mu}(\mathbb{X})}\frac{\tilde{\mu}(A)}{\tilde{\mu}(\mathbb{X})}\right] \\ =& \int_0^\infty\int_0^\infty \mathbb{E}\left[\tilde{\mu}(B)\tilde{\mu}(A)\times\exp\left\{-(v_1+v_2)\tilde{\mu}(\mathbb{X})\right\}\right]\mathrm{d}v_1\mathrm{d}v_2 \\ =& \int_0^\infty\int_0^\infty \mathbb{E}\left[\tilde{\mu}(B)\exp\left\{-(v_1+v_2)\tilde{\mu}(B)\right\}\right] \\ &\qquad \mathbb{E}\left[\tilde{\mu}(A)\exp\left\{-(v_1+v_2)\tilde{\mu}(A)\right\}\right] \\ &\qquad \mathbb{E}\left[\exp\left\{-(v_1+v_2)\tilde{\mu}(X/\{A\cup B\})\right\}\right]\mathrm{d}v_1\mathrm{d}v_2 \\ =& \int_0^\infty\int_0^\infty P(B)M\tilde{\psi}'_\eta(v_1+v_2)\exp\left\{-P(B)M\tilde{\psi}_\eta(v_1+v_2)\right\} \\ &\qquad P(A)M\tilde{\psi}'_\eta(v_1+v_2)\exp\left\{-P(A)M\tilde{\psi}_\eta(v_1+v_2)\right\} \\ &\qquad P(\mathbb{X}/\{A\cup B\})M\tilde{\psi}'_\eta(v_1+v_2)\exp\left\{-P(\mathbb{X}/\{A\cup B\})M\tilde{\psi}_\eta(v_1+v_2)\right\}\mathrm{d}v_1\mathrm{d}v_2 \\ =& P(A)P(B)M^2\int_0^\infty\int_0^{v_1}\tilde{\psi}'_\eta(v_2)^2\exp\left\{-M\tilde{\psi}_\eta(v_2)\right\}\mathrm{d}v_2\mathrm{d}v_1\end{aligned}$$

Then the covariance is:

$$\begin{aligned} & \mathrm{Cov}\left(\mu(B), (T\mu)(B)\right) \\ = \; & \mathbb{E}\left[\mu(B)\left((T\mu)(B)\right)\right] - \mathbb{E}\left[\mu(B)\right]\mathbb{E}\left[(T\mu)(B)\right] \\ = \; & P(A)P(B) \\ & \left(M^2 \int_0^\infty \int_0^{v_1} \tilde{\psi}'_\eta(v_2)^2 \exp\left\{-M\tilde{\psi}_\eta(v_2)\right\} \mathrm{d}v_2 \mathrm{d}v_1 - 1\right) \end{aligned} \tag{61}$$

Proof of Lemma 10 From the existing of Poisson processes, each Lévy measure $\nu_i(\mathrm{d}t, \mathrm{d}x)$ corresponds to a Poisson random measure $N_i(\mathrm{d}t, \mathrm{d}x)$ with

$$\mathbb{E}\left[N_i(\mathrm{d}t, \mathrm{d}x)\right] = \nu_i(\mathrm{d}t, \mathrm{d}x).$$

Also we have $\forall i$,

$$\tilde{\mu}_i(\mathrm{d}x) = \int_0^\infty t N_i(\mathrm{d}t, \mathrm{d}x).$$

Thus from (41) we have

$$\tilde{\mu}(\mathrm{d}x) = \int_0^\infty t \left(\sum_{i=1}^n N_i(\mathrm{d}t, \mathrm{d}x)\right) = \int_0^\infty t N(\mathrm{d}t, \mathrm{d}x),$$

where $N(\cdot) = \sum_{i=1}^n N_i(\cdot)$ is again a Poisson random measure. Thus the Lévy intensity for $\tilde{\mu}(\cdot)$ is

$$\nu(\mathrm{d}t, \mathrm{d}x) \;\; = \;\; \sum_{i=1}^n \nu_i(\mathrm{d}t, \mathrm{d}x). \tag{62}$$

Because Theorem 1 in [JLP09] applies for any CRMs with Lévy measure $\nu(\mathrm{d}t, \mathrm{d}x)$, the Lemma is proved.

Proof of Lemma 11 The case for point transition and superposition are developed similarly to the case for subsampling, so we only consider the later here.

The case for subsampling follows by merging the impact of the subsampling operation with the sampling step in Lemma 1. Suppose the Lévy measure is in the form $M\rho(\mathrm{d}t|x)H(\mathrm{d}x)$. The infinitesimal rate at data point x_i when sampling the jump is now $q(x_i)M\rho(\mathrm{d}t|x)$. Thus the Lévy measure for the subsampled measure must be $M\rho(\mathrm{d}t|x)q(x)H(\mathrm{d}x)$.

This argument can also be seen from the detailed derivation below. First note that $S^q(\tilde{\mu})$ is equivalent to

$$S^q(\tilde{\mu}) = \int_{R^+ \times \mathbb{X}} z(\mathrm{d}x) s N(\mathrm{d}s, \mathrm{d}x). \tag{63}$$

Let $B \in \mathbb{X}$, we divide B into n non-overlap patches and use A_{nm} to denote the m-th patch of them. So we have

$$\begin{aligned}
\mathbb{E}_{N(\cdot),z}\left[e^{-uS^q(\tilde{\mu})(B)}\right] &\overset{n\to\infty}{=} \mathbb{E}_{N(\cdot),z}\left[e^{-\sum_{A_{nm}\in B} uz(A_{nm})s_{nm}N(A_{nm},s_{nm})}\right] \\
&= \mathbb{E}_{N(\cdot),z}\left[\prod_{A_{nm}\in B} e^{-uz(A_{nm})s_{nm}N(A_{nm},s_{nm})}\right] \\
&= \prod_{A_{nm}\in B} \mathbb{E}_{N(\cdot),z}\left[e^{-uz(A_{nm})s_{nm}N(A_{nm},s_{nm})}\right] \\
&= e^{\sum_{A_{nm}\in B}\log\left\{\mathbb{E}_{N(\cdot),z}\left[e^{-uz(A_{nm})s_{nm}N(A_{nm},s_{nm})}-1\right]+1\right\}} \\
&\overset{(a)}{=} e^{\sum_{A_{nm}\in B}\mathbb{E}_{N(\cdot),z}\left[e^{-uz(A_{nm})s_{nm}N(A_{nm},s_{nm})}-1\right]} \\
&\overset{(b)}{=} e^{q\sum_{A_{nm}\in B}\mathbb{E}_{N(\cdot)}\left[e^{-us_{nm}N(A_{nm},s_{nm})}-1\right]} \\
&\overset{n\to\infty}{=} e^{-\int_{R^+\times B}\left(1-e^{-us}\right)(q\nu(\mathrm{d}s,\mathrm{d}x))}
\end{aligned} \tag{64}$$

Here (a) above follows because $\mathbb{E}_{N(\cdot)}\left[\left(e^{-uz(A_{nm})s_{nm}N(A_{nm},s_{nm})}-1\right)\right]$ is infinitesimal thus $\log(1+x) \overset{x\to 0}{\sim} x$ applies. (b) is obtained by integrating out $z(A_{nm})$ with Bernoulli distribution. Thus it can be seen from (64) that $S^q(\tilde{\mu})$ has the Lévy measure of $q(\mathrm{d}x)\nu(\mathrm{d}t,\mathrm{d}x)$.

Proof of Theorem 8 We show that starting from (44) and (43), we can both end up the random measures defined in (45).

First, for the operations in (44), adapting from Theorem 2.17 of [Ç10], a Poisson random measure with mean measure ν on the space $\mathbb{R}^+ \times \mathbb{X}$ has the form

$$N = \sum_{n=1}^{\infty}\sum_{i<K_n}\delta_{(s,x)}, \tag{65}$$

where K_n is a Poisson distributed random variable with mean ν, and $(s \in \mathbb{R}^+, x \in \mathbb{X})$ are points in the corresponding Poisson processes. Then a realization of N composes of points in a Poisson process Π_1, and the corresponding Poisson random measure can be written as $N_1 = \sum_{(s,x)\in\Pi_1}\delta_{(s,x)}$.

Now consider doing a subsampling S^q and a point transition T on Π_1, by the definitions and (65) we get a new random measure

$$\begin{aligned}
\tilde{N} &= T(S_q(N_1)) = T(S_q(\sum\delta_{(s,x)})) \\
&\overset{(*)}{=} \sum z(q(T(x)))\delta_{(s,T(x))} \overset{(**)}{=} \sum z(q(x))\delta_{(s,T(x))},
\end{aligned} \tag{66}$$

where $z(q(\cdot))$ means a Bernoulli random variable with acceptance rate $q(\cdot)$, $(*)$ follows from definitions, $(**)$ follows from the assumption of constant subsampling rate.

It is easy to show by induction that by subsampling and point transitioning i times of the Poisson process Π_1, we get a random measure as

$$\tilde{N}' = \sum z(q^i(x))\delta_{(s,T^i(x))}. \tag{67}$$

By the definition, when superpositioning the this Poisson process $T^i(S_i^q(\Pi_1))$ with another Poisson process Π_2 with mean measure ν_2, we get another random measure as

$$N'' = \sum_{(s,x)\in\Pi_1} z(q^i(x))\delta_{(s,T^i(x))} + \sum_{(s,x)\in\Pi_2} \delta_{(s,T(x))}. \tag{68}$$

This Poisson random measure is then used to construct a completely random measure $\tilde{\mu}$ using (3) as:

$$\begin{aligned} \tilde{\mu}(A) &= \int_{\mathbb{R}^+\times\mathbb{X}} sN''(\mathrm{d}s,\mathrm{d}x) \\ &= \sum_{(s,x)\in\Pi_1} z(q^i(x))s\delta_{(s,T^i(x))} + \sum_{(s,x)\in\Pi_2} s\delta_{(s,x)}. \end{aligned} \tag{69}$$

By marginalize over r's and normalizing this random measure, we get

$$\begin{aligned} \mu(A) &= \frac{\tilde{\mu}(A)}{\tilde{\mu}(\mathbb{X})} \\ &= \frac{M_1'}{M_1+M_2'}\frac{\sum_{(s,x)\in\Pi_1\cap A} s\delta_{(s,T^i(x))}}{\sum_{(s,x)\in\Pi_1\cap\mathbb{X}} s\delta_{(s,T^i(x))}} \\ &+ \frac{M_2'}{M_1'+M_2'}\frac{\sum_{(s,x)\in\Pi_2\cap A} s\delta_{(s,T^i(x))}}{\sum_{(s,x)\in\Pi_2\cap\mathbb{X}} s\delta_{(s,T^i(x))}} \\ &= \frac{M_1'}{M_1'+M_2'}(T^i\mu_1)(A) + \frac{M_2'}{M_1'+M_2'}(T^i\mu_2)(A), \end{aligned} \tag{70}$$

where by apply Lemma 11 we conclude that $M_1' = (q^i\tilde{\mu}_1)(\mathbb{X})$ is the total mass of the random measure with Lévy measure $q^j(\mathrm{d}x)\nu(\mathrm{d}t,\mathrm{d}x)$ and $M_2' = \tilde{\mu}_2(\mathbb{X})$. We use the fact that $(T^k\tilde{\mu}_i)(\mathbb{X}) = \tilde{\mu}_i(\mathbb{X})$ in the derivation of (70), because the point transition operation only moves the points (s,x) of the Poisson process to other locations $(s,x+\mathrm{d}x)$, thus does not affect the total mass of the corresponding random measure.

This means by superpositioning after subsampling, the mass of the normalized random measure decades exponentially fast with respect to the distance i. Based on Eq. (70), when taking i from 1 to n, and taking superposition for all these random measure induced, the resulting normalized random measure is:

$$\mu_n' = \sum_{i=1}^{n} \frac{(q^{n-i}\tilde{\mu}_i)(\mathbb{X})}{\sum_{j=1}^{n}(q^{n-j}\tilde{\mu}_j)(\mathbb{X})} T^{n-i}(\mu_i). \tag{71}$$

Next, for the operations in (43), from the definition we have

$$\begin{aligned}\mu_2' &= T\left(S^q\left(\mu_1'\right)\right) \oplus \mu_2 \\ &= \frac{\left(q\tilde{\mu}_1\right)(\mathbb{X})}{\left(q\tilde{\mu}_1+\tilde{\mu}_2\right)(\mathbb{X})} T\left(\mu_1\right) + \frac{\left(\tilde{\mu}_1\right)(\mathbb{X})}{\left(q\tilde{\mu}_1+\tilde{\mu}_2\right)(\mathbb{X})}\mu_2 \end{aligned} \tag{72}$$

Now μ_2' has a total mass of $(q\tilde{\mu}_1+\tilde{\mu}_2)(\mathbb{X})$, by induction on i, we get the formula in (45) for $i=n$.

This completes the proof.

Proof of Theorem 9 Given the current data configuration $\{n_k, k=1,2,\cdots\}$, for a particular k,

- If $n_k>0$, this means this jump J_k must exist in μ, otherwise it is impossible to have $n_k>0$, thus $p(z_k=1|\tilde{\mu},n)=1$.
- Otherwise,since $\mu=\sum_{k:z_k=1}\frac{J_k\delta_k}{\sum_{k'}z_{k'}J_{k'}}$, we have the likelihood as:

$$\prod_{k'':n_{k''}>0}\frac{J_{k''}^{n_{k''}}}{\left(\sum_{k'\neq k}z_{k'}J_{k'}+z_kJ_k\right)^{n_k}} = \frac{\prod_{k'':n_{k''}>0}J_{k''}^{n_{k''}}}{\left(\sum_{k'\neq k}z_{k'}J_{k'}+z_kJ_k\right)^{n}}.$$

  Furthermore, we know that the prior for z_k is $p(z_k=1)=q$, thus the posterior is:

$$p(z_k=1|\tilde{\mu},n)\propto\frac{q}{\left(\sum_{k'\neq k}z_{k'}J_{k'}+J_k\right)^n}.$$

$$p(z_k=0|\tilde{\mu},n)\propto\frac{1-q}{\left(\sum_{k'\neq k}z_{k'}J_{k'}\right)^n}.$$

  After normalizing, we get the posterior for the case $n_k=0$ in (45).

Proof of Corollary 3 Note that J_{mk}' is obtained by subsampling of $\{J_{m'k}, m'\leq m\}$, the number of data points in $\tilde{\mu}_m'$ is denoted as $\tilde{n}_{m\cdot}'=\sum_{k'}\tilde{n}_{mk'}'$.

Following the same arguments as in the proof of Theorem 9, when $\tilde{n}_{mk}'>0$, $p(z_{mk}=1|\tilde{\mu}_m,\tilde{n}_{m\cdot}')=1$. Otherwise, by subsampling, μ_m' can be written as:

$$\mu_m' = \sum_{m'\leq m}\sum_{k':z_{m'k'}=1}\frac{z_{m'k'}J_{m'k'}\delta_{\theta_{m'k'}}}{\sum_{m''\leq m}\sum_{k''}z_{m''k''}J_{m''k''}}.$$

Now following the same proof of Theorem 9, if we define

$$J_m=\left(\sum_{m'\leq m}\sum_{k'}z_{m'k'}J_{m'k'}\right)^{\tilde{n}_{m\cdot}'}, J_m^{-k}=\left(\sum_{m'\leq m}\sum_{k'\neq k}z_{m'k'}J_{m'k'}\right)^{\tilde{n}_{m\cdot}'},$$

then we get the likelihood as

$$\frac{\prod_{k'':\tilde{n}_{mk''}'>0}{J_{mk''}'}^{n_{k''}}}{J_m}.$$

Furthermore, from subsampling, we know that the Bernoulli prior for z_{mk} is $q^{m-m'}$, and the posterior can then be derived using the Bayes rule as in the proof of Theorem 9.