On collapsed representation of hierarchical Completely Random Measures

Gaurav Pandey GP88@CSA.IISC.ERNET.IN
Ambedkar Dukkipati AD@CSA.IISC.ERNET.IN
Department of Computer Science and Automation
Indian Institute of Science, Bangalore-560012, India

Abstract

The aim of the paper is to provide an exact approach for generating a Poisson process sampled from a hierarchical CRM, without having to instantiate the infinitely many atoms of the random measures. We use completely random measures (CRM) and hierarchical CRM to define a prior for Poisson processes. We derive the marginal distribution of the resultant point process, when the underlying CRM is marginalized out. Using well known properties unique to Poisson processes, we were able to derive an exact approach for instantiating a Poisson process with a hierarchical CRM prior. Furthermore, we derive Gibbs sampling strategies for hierarchical CRM models based on Chinese restaurant franchise sampling scheme. As an example, we present the sum of generalized gamma process (SGGP), and show its application in topic-modelling. We show that one can determine the power-law behaviour of the topics and words in a Bayesian fashion, by defining a prior on the parameters of SGGP.

1. Introduction

Mixed membership modelling is the problem of assigning an object to multiple latent classes/features simultaneously. Depending upon the problem, one can allow a single latent feature to be exhibited single or multiple times by the object. For instance, a document may comprise several topics, with each topic occurring in the document with variable multiplicity. The corresponding problem of mapping the words of a document to topics, is referred to as topic modelling.

While parametric solutions to mixed membership modelling have been available in literature since more than a decade (Landauer & Dumais, 1997; Hofmann, 1999; Blei et al., 2001), the first non-parametric approach, that allowed the number of latent classes to be determined as well, was the hierarchical Dirichlet process (HDP) (Teh et al., 2006). Both the approaches model the object as a set of repeated draws from an object-specific distribution, whereby the object specific distribution is itself sampled from a common distribution. On the other hand, recent approaches such as hierarchical beta-negative binomial process (Zhou et al., 2012; Broderick et al., 2015) and hierarchical gamma-Poisson process (Titsias, 2008; Zhou & Carin, 2015) model the object as a point process, sampled from an object specific random measure, which is itself sampled from a common random measure. In some sense, these approaches are more natural for mixed membership modelling, since they model the object as a single entity rather than as a sequence of draws from a distribution.

A straightforward implementation of any of the above non-parametric models would require sampling the atoms in the non-parametric distribution for the base as well as object-specific measure. However, since the number of atoms in these distributions are often infinite, a truncation step is required to ensure tractability. Alternatively, for the HDP, a Chinese restaurant franchise scheme (Teh et al., 2006) can be used for collapsed inference in the model (that is, without explicitly instantiating the atoms). Fully collapsed inference scheme has also been proposed for beta-negative binomial process (BNBP) (Heaukulani & Roy, 2013; Zhou, 2014) and Gamma-Gamma-Poisson process (Zhou et al., 2015). Of particular relevance is the work by Roy (2014), whereby a Chinese restaurant fanchise scheme has been proposed for hierarchies of beta proceses (and its generalizations), when coupled with Bernoulli process.

In this paper, it is our aim to extend fully collapsed sampling so as to allow any completely random measure (CRM) for the choice of base and object-specific measure. As proposed in Roy (2014) for hierarchies of generalized beta processes, we propose Chinese restaurant franchise schemes for hierarchies of CRMs, when coupled with Pois-

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

son process. We hope that this will encourage the use of hierarchical random measures, other than HDP and BNBP, for mixed-membership modelling and will lead to further research into an understanding of the applicability of the various random measures. To give an idea about the flexibility that can be obtained by using other measures, we propose the sum of generalized gamma process (SGGP), which allows one to determine the power term in the power-law distribution of topics with documents, by defining a prior on the parameters of SGGP. Alternatively, one can also define a prior directly on the discount parameter.

The main contributions in this paper are as follows:

- We derive marginal distributions of Poisson process, when coupled with CRMs,
- We provide an exact approach for generating a Poisson process sampled from a hierarchical CRM, without having to instantiate the infinitely many atoms of the random measure.
- We provide a Gibbs sampling approach for sampling a Poisson process from a hierarchical CRM.
- In the experiments section, we propose the sum of generalized gamma process (SGGP), and show its applicability for topic-modelling. By defining a prior on the parameters of SGGP, one can determine the power-law distribution of the topics and words in a Bayesian fashion.

2. Preliminaries and background

In this section, we fix the notation and recall a few well known results from the theory of point processes.

2.1. Poisson process

Let $(S, \mathcal{S})$ be a measurable space and Π be a random countable collection of points on S. Let $N(A) = |\Pi \cap A|$, for any measurable set A. N is also known as the counting process of Π. Π is called a Poisson process if $N(A)$ is independent of $N(B)$, whenever A and B are disjoint measurable sets, and $N(A)$ is Poisson distributed with mean $\mu(A)$ for a fixed σ-finite measure μ. In sequel, we refer to both the random collection Π and its counting process N as Poisson process.

Let $(T, \mathcal{T})$ be another measurable space and $f : S \to T$ be a measurable function. If the push forward measure of μ via f, that is, $\mu \circ f^{-1}$ is non-atomic, then $f(\Pi) = \{f(x) : x \in \Pi\}$ is also a Poisson process with mean measure $\mu \circ f^{-1}$. This is also known as the mapping proposition for Poisson processes (Kingman, 1992). Moreover, if $\Pi_1, \Pi_2, \ldots$ is a countable collection of independent Poisson processes with mean measures $\mu_1, \mu_2, \ldots$ respectively, then the union $\Pi = \cup_{i=1}^{\infty} \Pi_i$ is also a Poisson process with mean measure $\mu = \sum_{i=1}^{\infty} \mu_i$. This is known as the superposition proposition. Equivalently, if N_i is the counting process of Π_i, then $N = \sum_{i=1}^{\infty} N_i$ is the counting process of a Poisson process with mean measure $\mu = \sum_{i=1}^{\infty} \mu_i$.

Finally, let g be a measurable function from S to $\mathbb{R}$, and $\Sigma = \sum_{x \in \Pi} g(x)$. By Campbell's proposition (Kingman, 1992), Σ is absolutely convergent with probability, if and only if

$$\int_S \min(|g(x)|, 1)\mu(\mathrm{d}x) < \infty. \tag{1}$$

If this condition holds, then for any $t > 0$,

$$\mathbb{E}[e^{-t\Sigma}] = \exp\left\{-\int_S (1 - e^{-tg(x)})\mu(\mathrm{d}x)\right\}. \tag{2}$$

2.2. Completely random measures

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability space. Let $(M(S), \mathcal{B})$ be the space of all σ-finite measures on $(S, \mathcal{S})$ supplied with an appropriate σ-algebra. A completely random measure (CRM) Λ on $(S, \mathcal{S})$, is a measurable mapping from Ω to $M(S)$ such that

1. $\mathbb{P}\{\Lambda(\emptyset) = 0\} = 1$,
2. For any disjoint countable collection of sets $A_1, A_2, \ldots$, the random variables $\Lambda(A_i), i = 1, 2, \ldots$ are independent, and $\Lambda(\cup A_i) = \sum_i \Lambda(A_i)$, holds almost surely. (the independent increments property)

An important characterization of CRMs in terms of Poisson processes is as follows (Kingman, 1967). For any CRM Λ on $(S, \mathcal{S})$ without any fixed atoms or deterministic component, there exists a Poisson process N on $(\mathbb{R}^+ \times S, \mathcal{B}_{\mathbb{R}^+} \otimes \mathcal{S})$, such that $\Lambda(\mathrm{d}x) = \int_{\mathbb{R}^+} zN(\mathrm{d}z, \mathrm{d}x)$. Using Campbell's proposition, the Laplace transform of $\Lambda(A)$ for a measurable set A, is given by the following formula:

$$\mathbb{E}[e^{-t\Lambda(A)}] = \exp\left(-\int_{\mathbb{R}^+ \times A} (1 - e^{-tz})\nu(\mathrm{d}z, \mathrm{d}x)\right), \, t \geq 0, \tag{3}$$

where ν denotes the mean measure of the underlying Poisson process N. ν is also referred to as the Poisson intensity measure of Λ. If $\nu(\mathrm{d}z, \mathrm{d}x) = \rho(\mathrm{d}z)\mu(\mathrm{d}x)$, for a σ-finite measure μ on S, and a σ-finite measure ρ on $\mathbb{R}^+$ that satisfies $\int_{\mathbb{R}^+}(1 - e^{-tz})\rho(\mathrm{d}z) < \infty$, then $\Lambda(.)$ is known as homogenous CRM. In sequel, we assume $\mu(.)$ to be finite. Moreover, unless specified, whenever we refer to CRM, it means a homogeneous completely random measure without any fixed atoms or deterministic component.

Let N be the Poisson process of the CRM Λ, that is, $\Lambda(\mathrm{d}x) = \int_{\mathbb{R}^+} sN(\mathrm{d}z, \mathrm{d}x)$. If Π is the random collection of points corresponding to N, then Λ can equivalently be written as $\Lambda = \sum_{(z,x)\in\Pi} z\delta_x$. $\{z : (z,x) \in \Pi\}$ constitute the weights of the CRM Λ. By the mapping proposition for Poisson processes, they form a Poisson process with mean measure $\mu^*(\mathrm{d}z) = \mu \circ f^{-1}(\mathrm{d}z)$, where $f(x,y) = x$ is the projection map on $\mathbb{R}^+$. Hence, the weights of Λ form a Poisson process on $\mathbb{R}^+$ with mean measure $\mu^*(\mathrm{d}z) = \nu(\mathrm{d}z, S) = \rho(\mathrm{d}z)\mu(S)$. We formally state this result below.

Lemma 2.1. *The weights of a homogenous CRM with no atoms or deterministic component, whose Poisson intensity measure $\nu(\mathrm{d}z, \mathrm{d}x) = \rho(\mathrm{d}z)\mu(\mathrm{d}x)$ form a Poisson process with mean measure $\rho(\mathrm{d}z)\mu(S)$.*

Note 1: A completely random measure without any fixed atoms or deterministic component is a purely-atomic random measure.

Note 2: Every such homogeneous CRM Λ on $(S, \mathcal{S})$ has an underlying Poisson process N on $(\mathbb{R}^+ \times S, \mathcal{B}_{\mathbb{R}^+} \otimes \mathcal{S})$, such that

$$\Lambda(\mathrm{d}x) = \int_{\mathbb{R}^+} zN(\mathrm{d}z, \mathrm{d}x) \tag{4}$$

almost surely.

3. The proposed model

Let $X_1, \ldots, X_n$ be n observed samples, for instance, n documents. We assume that each sample X_i is generated as follows:

- The base measure Φ is CRM(ρ, μ), where ρ and μ are σ-finite and finite (non-atomic) measures on $(S, \mathcal{S})$ respectively.
- Object specific measures $\Lambda_i, 1 \leq i \leq n$ are CRM$(\bar{\rho}, \Phi)$, where $\bar{\rho}$ is another σ-finite non-atomic measure on $(S, \mathcal{S})$.
- The latent feature set N_i for each object X_i is a Poisson process with mean measure Λ_i.
- Finally, the visible features X_i are sampled from N_i.

Note: For topic modelling, S corresponds to the space of all probability measures on the words in the dictionary, also known as topics. Hence, when we sample Φ, we sample a subset of topics, along with the weights for those topics. This follows from the discreteness of Φ. Sampling object-specific random measures Λ_i corresponds to sampling the document specific weights for all the topics in Φ. Sampling the latent features N_i then corresponds to selecting a subset of topics from Λ_i based on the corresponding document-specific weights. Since, all the $\Lambda_i's$ have access to the same set of topics, this leads to sharing of topics among N_is. Finally, the words in X_i is sampled from the corresponding topic in N_i using categorical distribution.

Our aim is to infer the latent features $N_i, 1 \leq i \leq n$ from $X_i, 1 \leq i \leq n$. By Bayes' rule

$$\begin{aligned} &P(N_1, \ldots, N_n | X_1, \ldots, X_n) \propto \\ &\quad P(X_1, \ldots, X_n | N_1, \ldots, N_n) P(N_1, \ldots, N_n) \\ &\quad = \Pi_{i=1}^n P(X_i|N_i) P(N_1, \ldots, N_n) \end{aligned}$$

The conditional distribution of X_i given N_i are often very simple to compute, for instance, in the case of topic modelling, it is simply the product of categorical distributions. Hence, all we need to compute is the prior distribution of the latent features $N_1, \ldots, N_n$. This can be obtained by marginalizing out the base and object-specific random measures Φ and $\Lambda_i, 1 \leq i \leq n$. This is what we wish to achieve in the next few sections.

We will address the problem of marginalizing out the base and object-specific random measures in two steps. Firstly, in section 3.1, we will derive results for the case when the base measure is held fixed and the object-specific random measure is marginalized out. Next, in section 3.2, we will derive results for the case, when the base random measure Φ is also marginalized out. All the proofs are provided in the appendix.

3.1. Marginalizing out the object specific measure

Let ϕ be a realization of the base random measure Φ. Let $\Lambda_i, 1 \leq i \leq n$, be independent CRM$(\bar{\rho}, \phi)$. It is straightforward to see that if ϕ is a finite measure Λ_is will almost-surely be finite. Because of the independence among Λ_is, we can focus on marginalizing out a single object-specific random measure, say Λ. Although, in our original formulation, only 1 object is sampled from its object-specific random measure, we will present results for the case when n objects, $N_1, \ldots, N_n$ are sampled from the object specific random measure. This extended result will be needed in the next section when marginalizing the base measure.

There are several ways to instantiate the random measure Λ. For instance, one can use the fact that since the underlying base measure ϕ is purely-atomic, the support of CRM$(\bar{\rho}, \phi)$ will be restricted to only those measures whose support is a subset of the support of ϕ. In particular, if $\phi = \sum_{j=1}^{\infty} \beta_j \delta_{x_j}$, then Λ will be of the form $\sum_{j=1}^{\infty} L_j \delta_{x_j}$, where L_j are independent random variables. The independence of L_js follows from the complete randomness of the measure.

However, we found that this approach doesn't lead us far. Hence, we derive the marginal distribution of the Poisson processes $N_1, \ldots, N_n$ in proposition 3.1 and 3.2, by first assuming ϕ to be a continuous measure and then generalizing it to the case where ϕ is any finite measure.

In the sequel, $\psi(t) = \int_{\mathbb{R}^+} (1 - e^{-tz}) \bar{\rho}(\mathrm{d}z)$, and $\psi^{(k)}$ is the k^{th} derivative of ψ.

Proposition 3.1. *Let Λ be a CRM on $(S, \mathcal{S})$ with Poisson intensity measure $\rho(\mathrm{d}z)\mu(\mathrm{d}x)$, where both $\mu(.)$ and $\rho(.)$ are non-atomic. Let $N_1, \ldots, N_n$ be n independent Poisson process with random mean measure Λ, and M be the distinct points of N_i, $1 \leq i \leq n$. Then, M is a Poisson process with mean measure $\mathbb{E}[M(\mathrm{d}x)] = \mu(\mathrm{d}x)\int_{\mathbb{R}^+}(1 - e^{-nz})\rho(\mathrm{d}z)$.*

The above proposition provides the distribution of distinct points of the n point processes, $N_1, \ldots, N_n$. In order to complete the description of the distribution of $N_1, \ldots, N_n$, we also need to specify the joint distribution of the counts of each distinct feature in each N_i. This distribution is referred to as CRM-Poisson distribution in the rest of the paper. Let $M(S) = k$ and m_{ij} be the count of the j^{th} distinct feature in the i^{th} object. Furthermore, let $[m_{\boldsymbol{\cdot}j}]$ be the count of the j^{th} distinct feature for each object and $[m_{ij}]_{1\leq i\leq n, 1\leq j\leq k}$ be the set of count vectors for the each latent feature.

Proposition 3.2. *The joint distribution of the set of count vectors for the each latent feature $[m_{ij}]_{(n,k)}$ is given by*

$$P([m_{ij}]_{(n,k)}) = (-1)^{m_{\boldsymbol{\cdot\cdot}}-k} \frac{\theta^k e^{-\theta\psi(n)}}{\prod_{i=1}^{n}(m_{i\boldsymbol{\cdot}})!} \prod_{j=1}^{k} \psi^{(m_{\boldsymbol{\cdot}j})}(n)\,, \tag{5}$$

*where $m_{i\boldsymbol{\cdot}} = \sum_{j=1}^{k} m_{ij}$, $m_{\boldsymbol{\cdot}j} = \sum_{i=1}^{n} m_{ij}$, $m_{\boldsymbol{\cdot\cdot}} = \sum_{i=1}^{n}\sum_{j=1}^{k} m_{ij}$, $\theta = \mu(S)$ and $\psi(t) = \int_{\mathbb{R}^+}(1 - e^{-tz})\rho(\mathrm{d}z)$ is the Laplace exponent of Λ, and $\psi^{(l)}(t)$ is the l^{th} derivative of $\psi(t)$. This distribution will be referred to as **CRM-Poisson**$(\mu(S), \rho, n)$.*

Corollary 3.3. *Conditioned on $M(S) = k$, the set of count vectors for the each latent feature $[m_{ij}]_{(n,k)}$ is distributed as*

$$\begin{aligned} &P([m_{ij}]_{(n,k)} | M(S) = k) \\ &= \frac{\theta^k (-1)^{m_{\boldsymbol{\cdot\cdot}}-k} k!}{\prod_{i=1}^{n}(m_{i\boldsymbol{\cdot}})!} \prod_{j=1}^{k} \frac{\psi^{(m_{i\boldsymbol{\cdot}})}(n)}{\psi(n)} \end{aligned} \tag{6}$$

Note that both $\psi^{(k)}$ and ψ contain a multiple involving $\mu(S)$, which cancels out when they are divided in (6). Hence, conditioned on the number of points in the Poisson process M, the distribution of the set of counts for each latent feature $[m_{ij}]_{(n,k)}$ does not depend on the measure μ. In sequel, this distribution will be referred to as **conditional CRM-Poisson**(ρ, n, k) or **CCRM-Poisson**(ρ, n, k).

Example 1: The Gamma-Poisson process
The Poisson-intensity measure of gamma process is given by $\rho(\mathrm{d}z) = e^{-z}z^{-1}\,\mathrm{d}z$. The corresponding Laplace exponent is $\psi(t) = \ln(1 + t)$. Replacing it in equation (5), we get

$$P([m_{ij}]_{(n,k)}) = \frac{\theta^k \prod_{j=1}^{k}\Gamma(m_{\boldsymbol{\cdot}j})}{\prod_{i=1}^{n} m_{i\boldsymbol{\cdot}}!(1+n)^{m_{\boldsymbol{\cdot\cdot}}+\theta}} \tag{7}$$

Next, we generalize these results for the case when ϕ is an atomic measure.

Proposition 3.4. *Let Λ be a completely random measure with Poisson intensity measure $\nu(\mathrm{d}z, \mathrm{d}x) = \phi(\mathrm{d}x)\bar{\rho}(\mathrm{d}z)$, where $\bar{\rho}$ is non-atomic. Let N be a Poisson process with mean measure Λ. Then, N can be obtained by sampling a Poisson process with mean measure $\phi(\mathrm{d}x)\psi(1)$, say M, and then sampling the count of each feature in M using the conditional CRM-Poisson distribution.*

Note: The points in M won't be distinct anymore, since the underlying mean measure is non-atomic.

3.2. Marginalizing out the base measure

The previous section derived the marginal distribution of the Poisson processes, for a fixed realization ϕ of the base random measure Φ. In this section, we want to marginalize the CRM Φ as well. Marginalizing Φ does away with the independence among the latent features N_is, hence, we need to model the joint distribution of $N_1, \ldots, N_n$.

The model under study is

$$\begin{aligned} \Phi &\sim \text{CRM}(\rho, \mu)\,, \\ \Lambda_i | \Phi &\sim \text{CRM}(\rho', \Phi),\ 1 \leq i \leq n\,, \\ N_i | \Lambda_i &\sim \text{Poisson Process}(\Lambda_i),\ 1 \leq i \leq n\,. \end{aligned} \tag{8}$$

We use Proposition 3.4 to marginalize out Λ_i from the above description. Thus N_i can equivalently be obtained by sampling a Poisson processes with mean measure $\Phi(\mathrm{d}x)\int_{\mathbb{R}^+}(1 - e^{-z})\rho(\mathrm{d}z)$, and then sampling the count of each feature in M_i for each point process N_i using Corollary 3.3. In particular, let M_i be the corresponding Poisson process, and m_{ij} be the count of the j^{th} feature in M_i for the point process N_i and $r_{i\boldsymbol{\cdot}} = M_i(S)$. The reason for the symbol $r_{i\boldsymbol{\cdot}}$ will become clear, when we have a picture of the entire generative model. Let $[m_{ij}]_{\boldsymbol{\cdot}, r_{i\boldsymbol{\cdot}}}$ be the set of counts of the latent features for the i^{th} individual. The distribution of the set of counts $[m_{ij}]_{\boldsymbol{\cdot}, r_{i\boldsymbol{\cdot}}}$ conditioned on $M_i(S)$ does not depend on Φ. Hence, an alternative description of the N_i via M_i and m_{ij}, $1 \leq j \leq r_{i\boldsymbol{\cdot}}$ is as follows:

$$
\begin{aligned}
&M_i|\Phi \sim \text{Poisson Process}\left(\Phi(.)\int_{\mathbb{R}+}(1-e^{-z})\bar{\rho}(\mathrm{d}z)\right)\,,\\
&[m_{ij}]_{(\bullet,r_{i\bullet})}|\{M_i(S)=r_{i\bullet}\} \sim \text{CCRM-Poisson}(\bar{\rho},1,r_{i\bullet})\\
&N_i=\sum_{i=1}^{r_{i\bullet}} m_{ij}\delta_{M_{ij}}\,,
\end{aligned}
\tag{9}
$$

where M_{ij} are the points in the point process M_i.

M_i, $1 \le i \le n$ are independent Poisson processes, whose mean measure is a scaled CRM, and hence, also a CRM. Hence, we are again in the domain of CRM-Poisson models. Let $\bar{\psi}(1) = \int_{\mathbb{R}+}(1-e^{-z})\bar{\rho}(\mathrm{d}z)$. If we define $\Phi'(\mathrm{d}x) = \bar{\psi}(1)\Phi(\mathrm{d}x)$, then

$$
\begin{aligned}
&\mathbb{E}[e^{-t\Phi'(A)}] = \mathbb{E}[e^{-t\bar{\psi}(1)\Phi(A)}]\\
&= \exp\left\{-\mu(A)\int_{\mathbb{R}+}(1-e^{-t\bar{\psi}(1)z})\rho(\mathrm{d}z)\right\}\\
&= \exp\left\{-\mu(A)\int_{\mathbb{R}+}(1-e^{-tz'})\rho(\mathrm{d}(z'/\bar{\psi}(1)))\right\}
\end{aligned}
$$

Hence, the Poisson intensity measure of the scaled CRM Φ' is given by $\rho(\mathrm{d}(z/\bar{\psi}(1)))\mu(\mathrm{d}x)$. Applying Proposition 3.4 to marginalize out Φ, we get that M_i's can be obtained by sampling a Poisson process R with mean measure

$$
\begin{aligned}
\mathbb{E}[R(\mathrm{d}x)] &= \mu(\mathrm{d}x)\int_{\mathbb{R}+}(1-e^{-nz'})\rho(\mathrm{d}(z'/\bar{\psi}(1)))\\
&= \mu(\mathrm{d}x)\int_{\mathbb{R}+}(1-e^{-\bar{\psi}(1)nz})\rho(\mathrm{d}z)\,.
\end{aligned}
$$

The count of each feature in R for each point process M_i can then be obtained by using Corollary 3.3. In particular, let r_{ik} be the count of the k^{th} point in R for the point process M_i and $p = R(S)$.

A complete generative model for generating the point processes N_i, $1 \le i \le n$ is as follows:

$$
\begin{aligned}
&R \sim \text{Poisson Process}\left(\mu(.)\int_{\mathbb{R}+}(1-e^{-\bar{\psi}(1)nz})\rho(\mathrm{d}z)\right),\\
&[r_{ik}]_{(n,p)}|\,\{R(S)=p\} \sim \text{CCRM-Poisson}(\rho,\bar{\psi}(1)n,p)
\end{aligned}
\tag{10}
$$

$$
\begin{aligned}
&M_i = \sum_{k=1}^{p} r_{ik}\delta_{R_k}\\
&[m_{ij}]_{(\bullet,r_{i\bullet})}|\{M_i(S)=r_{i\bullet}\} \sim \text{CCRM-Poisson}(\bar{\rho},1,r_{i\bullet})\\
&N_i = \sum_{j=1}^{r_{i\bullet}} m_{ij}\delta_{M_{ij}}\,,
\end{aligned}
$$

Since R is again a Poisson process, it is straightforward to extend this hierarchy further by sampling $\mu(.)$ again from a CRM.

4. Implementation via Gibbs sampling

Section 3 provided an approach for sampling a Poisson process, when sampled from a hierarchical CRM, without having to instantiate the infinitely many atoms of the base or object-specific CRM. However, it is not clear how the above derivations can be used for determining the latent features $N_1, \ldots, N_n$ for the objects $X_1, \ldots, X_n$, which is the aim of this work.

In this section, we provide a Gibbs sampling approach for sampling the latent features from its prior distribution that is $P(N_1, \ldots, N_n)$. In order to sample from the posterior, one simply needs to multiply the equations in this section with the likelihood of the latent feature. In order to be able to perform MCMC sampling in hierarchical CRM-Poisson models, we need to marginalize out $R(S)$ and $M_i(S)$ from distributions of $[r_{ik}]_{(n,p)}$ and $[m_{ij}]_{(\bullet,r_{i\bullet})}$ respectively. By marginalizing out the Poisson distributed random variable $R(S)$ from (10), we get that

$$
[r_{ik}]_{(n,p)} \sim \text{CRM-Poisson}(\mu(S),\rho,\bar{\psi}(1)n)\,.
$$

The marginal distribution of the set of counts of each latent feature for the i^{th} individual $[m_{ij}]_{(\bullet,r_{i\bullet})}$ (where $r_{i\bullet}$ is also random) is given by the following lemma.

Lemma 4.1. *Let*

$$
h(u) = \mathbb{E}[e^{-u\psi(S)}] = \exp\left\{-\mu(S)\int_{\mathbb{R}+}(1-e^{-uz})\rho(\mathrm{d}z)\right\}.
$$

Furthermore, if we let

$$
\begin{aligned}
\psi(u) &= \int_{\mathbb{R}+}(1-e^{-uz})\rho(\mathrm{d}z)\\
\bar{\psi}(u) &= \int_{\mathbb{R}+}(1-e^{-uz})\bar{\rho}(\mathrm{d}z)\,,
\end{aligned}
$$

then, $[m_{ij}]_{(\bullet,r_{i\bullet})}$ is marginally distributed as

$$
P([m_{ij}]_{(\bullet,r_{i\bullet})}) = (-1)^{m_{i\bullet}}h^{(r_{i\bullet})}\left(\bar{\psi}(1)\right)\frac{\prod_{j=1}^{r_{i\bullet}}\bar{\psi}^{(m_{ij})}(1)}{m_{i\bullet}!}\,,
\tag{11}
$$

In the case of topic-modelling, the number of latent features, $\#N_i$ is equal to the number of observed features $\#X_i$. Hence, let X_{il} be the l^{th} observed feature associated with the i^{th} object and N_{il} be the corresponding latent feature. Here, we discuss the MCMC approach for sampling from the prior distribution of N_{il}, $1 \le l \le m_{i\bullet}$.

As discussed in (Neal, 2000), it is more efficient to sample the index of the latent feature, rather than the latent feature itself. Hence, let T_{il} be the index of the point in M_i associated with N_{il}, and D_{ij} be the index of the point in R associated with M_{ij}. In an analny with the Chinese restaurant franchise model (Teh et al., 2006), one can think of T_{il}

to be the index of the table assigned to the l^{th} customer in the i^{th} restaurant, and D_{ij} to be the index of the dish associated with the j^{th} table in i^{th} restaurant. Moreover, m_{ij} refers to the number of customers sitting on the j^{th} table in i^{th} restaurant, and r_{ik} refers to the number of tables in the i^{th} restaurant with the k^{th} dish. Hence $r_{i\bullet} = \sum_{k=1}^{p} r_{ik}$ is the number of tables in the i^{th} restaurant.

The distribution of the number of customers per table in the i^{th} restaurant, $[m_{ij}]_{(\bullet,r_{i\bullet})}$ follows from Lemma 4.1. Hence, in order to sample the table of l^{th} customer, T_{il}, given the indices of the tables of all the other customers in i^{th} restaurant, we treat it as the table corresponding to the last customer of the i^{th} restaurant. Let $m_{i'j}^{-(il)}$ be the number of customers sitting on the j^{th} table in the i'^{th} restaurant, excluding the l^{th} customer. The probability that the l^{th} customer in the i^{th} restaurant occupies the j^{th} table is proportional to $P(m_{ij'}^{-(il)} + 1_{j'=j}, 1 \leq j' \leq r_{i\bullet})$ as given in (11). We divide the expression by $P(m_{ij'}^{-(il)}, 1 \leq j' \leq r_{i\bullet})$ to get a simpler form for the unnormalized probability distribution. Hence, the probability of assigning an existing table with index j is given by

$$P(T_{il} = j|\mathbf{T}^{-(il)}) \propto -\frac{\bar{\psi}^{(m_{ij}^{-(il)}+1)}(1)}{\bar{\psi}^{(m_{ij}^{-(il)})}(1)}\,, \tag{12}$$

and the probability of sampling a new table for the customer is given by

$$P(T_{il} = r_{i\bullet} + 1|\mathbf{T}^{-(il)}) \propto= -\frac{h^{(r_{i\bullet}+1)}(\bar{\psi}(1))}{h^{(r_{i\bullet})}(\bar{\psi}(1))}\bar{\psi}^{(1)}(1)\,, \tag{13}$$

where $\bar{\psi}(t) = \int_{\mathbb{R}+}(1 - e^{-tz})\bar{\rho}(\mathrm{d}z)$ and $h^{(k)}$ is the k^{th} derivative of h.

Moreover, whenever a new table is sampled for a customer, a dish is sampled for the table from the distribution on tables per dish. By the discussion in the beginning of this section, the number of tables per dish $[r_{ik}]_{(n,p)}$ follow a CRM-Poisson$(\mu(S), \rho, \bar{\psi}(1)n)$ distribution. Hence, in order to sample the dish at j^{th} table, D_{ij}, given the indices of the dishes at all the other tables, we treat it as the dish corresponding to the last table of the last restaurant. Let $r_{\bullet k}^{-(ij)}$ be the total number of tables served with the k^{th} dish, excluding the j^{th} table of i^{th} restaurant. The probability that the k^{th} dish is served at the j^{th} table in the i^{th} restaurant is proportional to $P(r_{i'k'}^{-(ij)} + 1_{i'=i',k'=k}, 1 \leq i' \leq n, 1 \leq k' \leq p)$ as given in (11). We divide the expression by $P(r_{i'k'}^{-(ij)}, 1 \leq i' \leq n, 1 \leq k' \leq p)$ to get a simpler form for the unnormalized probability distribution. Hence, the probability of serving an existing dish with index k is given by

$$P(D_{ij} = k|\mathbf{D}^{-(ij)}) \propto -\frac{\psi^{(r_{\bullet k}^{-(ij)}+1)}(\bar{\psi}(1)n)}{\psi^{(r_{\bullet k}^{-(ij)})}(\bar{\psi}(1)n)}\,, \tag{14}$$

and the probability of sampling a new dish for the table is given by

$$P(D_{ij} = p + 1|\mathbf{D}^{-(ij)}) \propto \theta\psi^{(1)}(\bar{\psi}(1)n)\,, \tag{15}$$

where $\psi(t) = \int_{\mathbb{R}+}(1 - e^{-tz})\rho(\mathrm{d}z)$ and $\theta = \mu(S)$.

Hence, a complete description of one iteration of MCMC sampling, from the prior distribution, in hierarchical CRM-Poisson models is as follows:

1. For each customer in each restaurant, sample his table index conditioned on the indices of table of other customers, according to equations (12) and (13).
2. If the table selected is a new table, sample the index of dish corresponding to that table from equations (14) and (15).
3. Sample the index of dish for each table, conditioned on the indices of dishes at the other tables, according to equations (14) and (15).

Example 2: The Gamma-Gamma-Poisson process
We compute the dish and table sampling probabilities for the Gamma-Gamma-Poisson process using the above equations. The Poisson intensity measure for both the base and object specific measures Φ and $\Lambda_i, 1 \leq i \leq n$ is $z^{-1}e^{-z}\,\mathrm{d}z$. The corresponding Laplace exponent is given by $\psi(t) = \bar{\psi}(t) = \ln(1 + t)$. Moreover, let the mean measure for the base measure Φ be $\mu(.)$ and $\mu(S) = \theta$. Then, $h(u) = \mathbb{E}e^{-u\Phi(S)} = \frac{1}{(1+u)^{\theta}}$. The corresponding derivatives are given by

$$\psi^{(k)} = \bar{\psi}^{(k)}(t) = \frac{(-1)^{k-1}\Gamma(k)}{(1+t)^k} \tag{16}$$

$$h^{(k)}(u) = \frac{(-1)^k\Gamma(k+\theta)}{(1+u)^{k+\theta}\Gamma(\theta)} \tag{17}$$

The corresponding dish sampling probabilities are given by

$$P(D_{ij} = k|\mathbf{D}^{-(ij)}) \propto \frac{r_{\bullet k}^{-(ij)}}{1 + n\ln 2} \tag{18}$$

$$P(D_{ij} = p + 1|\mathbf{D}^{-(ij)}) \propto \frac{\theta}{1 + n\ln 2} \tag{19}$$

for an existing and new dish respectively. Normalizing these probabilities, we get

$$P(D_{ij} = k|\mathbf{D}^{-(ij)}) = \frac{r_{\bullet k}^{-(ij)}}{\sum_{k=1}^{p} r_{\bullet k} + \theta} \tag{20}$$

$$P(D_{ij} = p + 1|\mathbf{D}^{-(ij)}) = \frac{\theta}{\sum_{k=1}^{p} r_{\bullet k} + \theta} \tag{21}$$

The table sampling probabilities are given by

$$P(T_{il} = j|\mathbf{T}^{-(il)}) \propto \frac{m_{ij}^{-(il)}}{1+\ln(2)} \quad (22)$$

$$P(T_{il} = r_{i\boldsymbol{\cdot}} + 1|\mathbf{T}^{-(il)}) \propto \frac{\theta + r_{i\boldsymbol{\cdot}}}{(1+\ln(2))^2} \quad (23)$$

for an existing and new table respectively. Normalizing these probabilities, we get

$$P(T_{il} = j|\mathbf{T}^{-(il)}) = \frac{m_{ij}^{-(il)}}{\sum_{j=1}^{r_{i\boldsymbol{\cdot}}} m_{ij}^{-(il)} + \frac{\theta + r_{i\boldsymbol{\cdot}}}{1+\ln(2)}} \quad (24)$$

$$P(T_{il} = r_{i\boldsymbol{\cdot}} + 1|\mathbf{T}^{-(il)}) = \frac{(\theta + r_{i\boldsymbol{\cdot}})/(1+\ln(2))}{\sum_{j=1}^{r_{i\boldsymbol{\cdot}}} m_{ij}^{-(il)} + \frac{\theta + r_{i\boldsymbol{\cdot}}}{1+\ln(2)}} \quad (25)$$

Example 3: The Gamma-Generalized Gamma-Poisson process

In this scenario, the base random measure has $\rho(\mathrm{d}z) = e^{-z}z^{-1}\,\mathrm{d}z$, whereas the object specific measure has $\bar{\rho}(\mathrm{d}z) = e^{-z}z^{-d-1}\,\mathrm{d}z$, where $0 < d < 1$ is known as the discount parameter. The corresponding Laplace exponents are given by $\psi(t) = \ln(1+t)$ and $\bar{\psi}(t) = \frac{(1+t)^d - 1}{d}$ respectively. The derivative of $\bar{\psi}$ is given by

$$\bar{\psi}^{(k)}(t) = \frac{(-1)^{k-1}\Gamma(k-d)}{(1+t)^{k-d}\Gamma(1-d)} \quad (26)$$

$$(27)$$

Other derivatives remain same as in the previous example. Moreover, the dish sampling probabilities remain same. The table sampling probabilities are given by

$$P(T_{il} = j|\mathbf{T}^{-(il)}) = \frac{m_{ij}^{-(il)} - d}{\sum_{j=1}^{r_{i\boldsymbol{\cdot}}} (m_{ij}^{-(il)} - d) + \frac{\theta + r_{i\boldsymbol{\cdot}}}{1+\ln_d(2)}} \quad (28)$$

$$P(T_{il} = r_{i\boldsymbol{\cdot}} + 1|\mathbf{T}^{-(il)}) = \frac{(\theta + r_{i\boldsymbol{\cdot}})/(1+\ln_d(2))}{\sum_{j=1}^{r_{i\boldsymbol{\cdot}}} (m_{ij}^{-(il)} - d) + \frac{\theta + r_{i\boldsymbol{\cdot}}}{1+\ln_d(2)}} \quad (29)$$

where $\ln_d(2) = \frac{2^d - 1}{d}$.

5. Experimental results

We use hierarchical CRM-Poisson models for learning topics from the NIPS corpus [1].

[1]The dataset can be downloaded from http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

5.1. Evaluation

For evaluating the different models, we divide each document into a training section and a test section by independently sampling a boolean random variable for each word. The probability of sending the word to the training section is varied from 0.3 to 0.7. We run 2000 iterations of Gibbs sampling. The first 500 iterations are discarded, and every sample in every 5 iterations afterwards is used to update the document-specific distribution on topics and the topic specific distribution on words. In particular, let W be the number of words, K be the number of topics, $(\beta_{dk})_{1\le k\le K}$ be the document specific distribution on topics for the document d, and $(\tau_{kw})_{1\le w\le W}$ be the topic specific distribution on words for the k^{th} topic. Then, the probability of observing a word w in document d is given by $\sum_{k=1}^{K} \beta_{dk}\tau_{kw}$. For the evaluation metric, we use perplexity, which is simply the inverse of the geometric mean of the probability of all the words in the test set.

5.2. Varying the Common CRM

In our experiments, we fix the object specific random measure Λ_i in (8) to be the gamma process, with $\bar{\rho}(\mathrm{d}z) = e^{-z}z^{-1}\,\mathrm{d}z$. For the base CRM Φ, we consider two specific choices of random measures.

- **Generalized gamma process (GGP):** The Poisson intensity measure of Φ is given by $\nu(\mathrm{d}z, \mathrm{d}x) = \rho(\mathrm{d}z)\mu(\mathrm{d}x)$, where $\rho(\mathrm{d}z) = \frac{\theta}{\Gamma(1-d)} e^{-z}z^{-d-1}\,\mathrm{d}z$, $0 \le d < 1, \theta > 0$ and $\mu(S) = 1$. The corresponding Laplace exponent is given by $\theta((1+t)^d - 1)/d$.
- **Sum of Generalized gamma processes (SGGP):** The Poisson intensity measure of the CRM is given by $\nu(\mathrm{d}z, \mathrm{d}x) = \rho(\mathrm{d}z)\mu(\mathrm{d}x)$, where

$$\rho(\mathrm{d}z) = \sum_{q=1}^{m} \frac{\theta_q}{\Gamma(1-d_q)} e^{-z}z^{-d_q-1}\,\mathrm{d}z \quad (30)$$

  and $\mu(S) = 1$. The corresponding Laplace exponent is given by

$$\psi(t) = \left(\sum_{q=1}^{m} \theta_q \frac{(1+t)^{d_q} - 1}{d_q}\right). \quad (31)$$

For the case of GGP, the value of the discount parameter d is chosen from the set $\{0, .1, .2, .3, .4\}$. Furthermore, a gamma prior with rate parameter 2 and shape parameter 4 is defined on θ.

Note: The generalized gamma process with discount parameter 0 corresponds to the Gamma process. Using a gamma process prior for the base and object-specific CRM corresponds exactly to the hierarchical Dirichlet process with a gamma prior on the concentration parameter of the object specific Dirichlet process. We did not add comparison results with HDP separately, because the same perplexity is obtained in both the models.

For the case of SGGP, we consider $m = 5$, and $d_1 = 0$, $d_2 = .1 \ldots, d_5 = .4$. Furthermore, independent gamma priors with rate parameter 2 and shape parameter 4 are defined for each θ_q, $1 \leq q \leq 5$. The posterior of each parameter θ_q is sampled via uniform sampling. We use equations (12)-(15) to compute the dish sampling and table sampling probabilities. The probability of sampling an existing dish is given by

$$P(D_{ij} = k|\mathbf{D}^{-(ij)}) \propto \frac{\sum_{q=1}^{m} \theta_q \frac{\Gamma(r_{\bullet k}^{-(ij)}+1-d_q)}{\Gamma(1-d_q)}(1+\bar{\psi}(1)n)^{d_q}}{\sum_{q=1}^{m} \theta_q \frac{\Gamma(r_{\bullet k}^{-(ij)}-d_q)}{\Gamma(1-d_q)}(1+\bar{\psi}(1)n)^{d_q}},$$

where $\bar{\psi}(1) = \int_{\mathbb{R}+}(1-e^{-z})\bar{\rho}(\mathrm{d}z) = \ln(2)$. Similarly, the probability of a new dish is given by

$$P(D_{ij} = p+1|\mathbf{D}^{-(ij)}) \propto \sum_{q=1}^{m} \theta_q (1+\bar{\psi}(1)n)^{d_q}.$$

The table-sampling probabilities can be computed similarly. We approximated the Laplace transform of $\Phi(S)$ (h in (13)), by a weighted sum of exponential functions to simplify the computation of its derivatives. The perplexity for the hierarchical CRM-Poisson models as a function of training percentage is plotted in Figure 1. Note that Figure 1 doesn't necessarily imply that SGGM-based models will always outperform GGM based models as the results have been obtained by defining a specific gamma prior for each hyperparameter, as mentioned above.

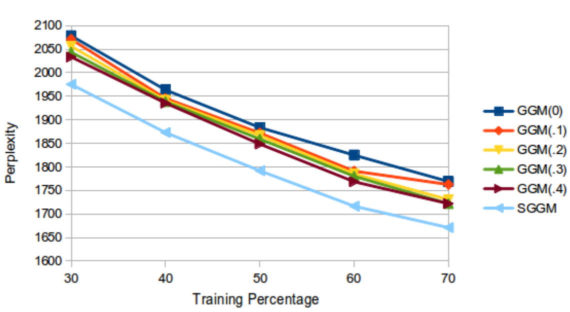

Figure 1. Variation of perplexity with training percentage for various hierarchical CRM-Poisson models

6. Conclusion

For years, hierarchical Dirichlet processes have been the standard tool for nonparametric topic modelling, since collapsed inference in HDP can be performed using the Chinese restaurant franchise scheme. In this paper, our aim was to show that collapsed Gibbs sampling can be extended to a much larger set of hierarchical random measures using the same Chinese restaurant franchise scheme, thereby opening doors for further research into the efficacy of various hierarchical priors. We hope that this will encourage a better understanding of applicability of various hierarchical CRM priors. Furthermore, the results of the paper can be used to prove results for hierarchical CRMs in other contexts, for instance, nonparametric hidden Markov models.

Acknowledgement

Gaurav Pandey is supported by IBM PhD Fellowship for the academic year 2015-2016.

References

Blei, David M, Ng, Andrew Y, and Jordan, Michael I. Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems*, pp. 601–608, 2001.

Broderick, Tamara, Mackey, Lester, Paisley, John, and Jordan, Michael I. Combinatorial Clustering and the Beta Negative Binomial Process. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37:290–306, 2015.

Heaukulani, Creighton and Roy, Daniel M. The combinatorial structure of beta negative binomial processes. *arXiv preprint arXiv:1401.0062*, 2013.

Hofmann, Thomas. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in Artificial Intelligence*, pp. 289–296. Morgan Kaufmann Publishers Inc., 1999.

James, Lancelot F. Bayesian Poisson process Partition Calculus with an application to Bayesian Lévy Moving Averages. *Annals of Statistics*, pp. 1771–1799, 2005.

Kingman, John. Completely Random Measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967.

Kingman, John Frank Charles. *Poisson Processes*, volume 3. Oxford University Press, 1992.

Landauer, Thomas K and Dumais, Susan T. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

Neal, Radford M. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

Roy, Daniel M. The continuum-of-urns scheme, generalized beta and indian buffet processes, and hierarchies thereof. *arXiv preprint arXiv:1501.00208*, 2014.

Teh, Yee Whye, Jordan, Michael I, Beal, Matthew J, and Blei, David M. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476), 2006.

Titsias, Michalis K. The Infinite Gamma-Poisson Feature Model. In *Advances in Neural Information Processing Systems*, pp. 1513–1520, 2008.

Zhou, Mingyuan. Beta-negative binomial process and exchangeable random partitions for mixed-membership modeling. In *Advances in Neural Information Processing Systems*, pp. 3455–3463, 2014.

Zhou, Mingyuan and Carin, Lawrence. Negative Binomial Process Count and Mixture Modelling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37: 307–320, 2015.

Zhou, Mingyuan, Hannah, Lauren A., Dunson, David B., and Carin, Lawrence. Beta-Negative Binomial Process and Poisson Factor Analysis. In *AISTATS*, 2012.

Zhou, Mingyuan, Padilla, Oscar Hernan Madrid, and Scott, James G. Priors for random count matrices derived from a family of negative binomial processes. *Journal of the American Statistical Association*, (just-accepted):00–00, 2015.