Gamma Processes, Stick-Breaking, and Variational Inference

Anirban Roychowdhury
Dept. of Computer Science and Engineering
Ohio State University
roychowdhury.7@osu.edu

Brian Kulis
Dept. of Computer Science and Engineering
Ohio State University
kulis@cse.ohio-state.edu

Abstract

While most Bayesian nonparametric models in machine learning have focused on the Dirichlet process, the beta process, or their variants, the gamma process has recently emerged as a useful nonparametric prior in its own right. Current inference schemes for models involving the gamma process are restricted to MCMC-based methods, which limits their scalability. In this paper, we present a variational inference framework for models involving gamma process priors. Our approach is based on a novel stick-breaking constructive definition of the gamma process. We prove correctness of this stick-breaking process by using the characterization of the gamma process as a completely random measure (CRM), and we explicitly derive the rate measure of our construction using Poisson process machinery. We also derive error bounds on the truncation of the infinite process required for variational inference, similar to the truncation analyses for other nonparametric models based on the Dirichlet and beta processes. Our representation is then used to derive a variational inference algorithm for a particular Bayesian nonparametric latent structure formulation known as the infinite Gamma-Poisson model, where the latent variables are drawn from a gamma process prior with Poisson likelihoods. Finally, we present results for our algorithm on nonnegative matrix factorization tasks on document corpora, and show that we compare favorably to both sampling-based techniques and variational approaches based on beta-Bernoulli priors, as well as a direct DP-based construction of the gamma process.

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

1 Introduction

The gamma process is a versatile pure-jump Lévy process with widespread applications in various fields of science. Of late it is emerging as an increasingly popular prior in the Bayesian nonparametric literature within the machine learning community; it has recently been applied to exchangeable models of sparse graphs Caron and Fox (2013) as well as for nonparametric ranking models Caron et al. (2013). It also has been used as a prior for infinite-dimensional latent indicator matrices Titsias (2008). This latter application is one of the earliest Bayesian nonparametric approaches to count modeling, and as such can be thought of as an extension of the venerable Indian Buffet Process to modeling latent structures where each feature can occur multiple times for a datapoint, instead of being simply binary.

The flexibility of gamma process models allows them to be applied in a wide variety of Bayesian nonparametric settings, but their relative complexity makes principled inference nontrivial. In particular, most direct applications of the gamma process in the Bayesian nonparametric literature use Markov chain Monte Carlo samplers (typically Gibbs sampling) for posterior inference, which often suffers from poor scalability. For other Bayesian nonparametric models—in particular those involving the Dirichlet process or beta process—a successful thread of research has considered variational alternatives to standard sampling methods Blei and Jordan (2003); Teh et al. (2007a); Wang et al. (2011). One first derives an explicit construction of the underlying "weights" of the atomic measure component of the random measures underlying the infinite priors; so-called "stick-breaking" processes for the Dirichlet and beta processes yield such a construction. Then these weights are truncated and integrated into a mean-field variational inference algorithm. For instance, stick-breaking was derived for the Dirichlet process in the seminal paper by Sethuraman Sethuraman (1994), which was in turn used for variational inference in Dirichlet process models Blei and Jordan (2003). Similar stick-breaking representations for a special case of the Indian Buffet Pro-

cess Teh et al. (2007b) and the beta process Paisley et al. (2010) have been constructed, and have naturally led to mean-field variational inference algorithms for nonparametric models involving these priors Doshi-Velez et al. (2009); Paisley et al. (2011). Such variational inference algorithms have been shown to be more scalable than the sampling-based inference techniques normally used; moreover they work with the full model posterior without marginalizing out any variables.

In this paper we propose a variational inference framework for gamma process priors using a novel stick-breaking construction of the process. We use the characterization of the gamma process as a *completely random measure* (CRM), which allows us to leverage Poisson process properties to arrive at a simple derivation of the rate measure of our stick-breaking construction, and show that it is indeed equal to the Lévy measure of the gamma process. We also use the Poisson process formulation to derive a bound on the error of the truncated version compared to the full process, analogous to the bounds derived for the Dirichlet process Ishwaran and James (2001), the Indian Buffet Process Doshi-Velez et al. (2009) and the beta process Paisley et al. (2011). We then, as a particular example, focus on the infinite Gamma-Poisson model of Titsias (2008) (note that variational inference need not be limited to this model). This model is a prior on infinitely wide latent indicator matrices with non-negative integer-valued entries; each column has an associated parameter independently drawn from a gamma distribution, and the matrix values are independently drawn from Poisson distributions with these parameters as means. We develop a mean-field variational technique using a truncated version of our stick-breaking construction, and a sampling algorithm that uses Monte Carlo integration for parameter marginalization, similar to Paisley et al. (2010), as a baseline inference algorithm for comparison. We also derive a variational algorithm based on the naïve construction of the gamma process. Finally we compare these with variational algorithms based on Beta-Bernoulli priors on a non-negative matrix factorization task involving the Psychological Review, NIPS, KOS and New York Times document corpora, and show that the variational algorithm based on our construction performs and scales better than all the others.

Related Work. To our knowledge this is the first explicit "stick-breaking"-like construction of the gamma CRM, apart from the naïve approach of denormalizing the construction of the DP with a suitable gamma random variable Miller (2011), Gopalan et al. (2014); moreover, as mentioned above, we develop a variational inference algorithm using the naïve construction (see 5.1) and show that it performs worse than our main algorithm on both synthetic and real datasets. The very general inverse Lévy measure algorithm of Wolpert and Ickstadt (1998) requires inversion of the exponential integral, as does the generalized CRM construction technique of Orbanz and Williamson (2012) when applied to the gamma process; since the closed form solution of the inverse of an exponential integral is not known, these techniques do not give us an analytic construction of the weights, and hence cannot be adapted to variational techniques in a straightforward manner. Other constructive definitions of the gamma process include Thibaux (2008), who discusses a sampling-based scheme for the weights of a gamma process by sampling from a Poisson process. As an alternative to gamma process-based models for count modeling, recent research has examined the negative binomial-beta process and its variants Zhou and Carin (2012); Zhou et al. (2012); Broderick et al. (2014); the stick-breaking construction of Paisley et al. (2010) readily extends to such models since they have beta process priors. The beta stick-breaking construction has also been used for variational inference in beta-Bernoulli process priors Paisley et al. (2011), though they have scalability issues when applied to the count modeling problems addressed in this work, as we show in the experimental section.

2 Background

2.1 Completely random measures

A completely random measure Kingman (1967); Jordan (2010) $\mathbb{G}$ on a space $(\Omega, \mathcal{F})$ is defined as a stochastic process on $\mathcal{F}$ such that for any two disjoint Borel subsets $\mathcal{A}_1$ and $\mathcal{A}_2$ in $\mathcal{F}$, the random variables $\mathbb{G}(\mathcal{A}_1)$ and $\mathbb{G}(\mathcal{A}_2)$ are independent. The canonical way of constructing a completely random measure $\mathbb{G}$ is to first take a σ-finite product measure H on $\Omega \otimes \mathbb{R}^+$, then draw a countable set of points $\{(\omega_k, p_k)\}$ from a Poisson process on a Borel σ-algebra on $\Omega \otimes \mathbb{R}^+$ with H as the rate measure. Then the CRM is constructed as $\mathbb{G} = \sum_{k=0}^{\infty} p_k \delta_{\omega_k}$, where the measure given to a measurable Borel set $B \subset \Omega$ is $\mathbb{G}(B) = \sum_{k:\omega_k \in B} p_k$. In this notation p_k are referred to as weights and the ω_k as atoms.

If the rate measure is defined on $\Omega \otimes [0,1]$ as $H(d\omega, dp) = cp^{-1}(1-p)^{c-1}B_0(d\omega)dp$, where B_0 is an arbitrary finite continuous measure on Ω and c is some constant (or function of ω), then the corresponding CRM constructed as above is known as a beta process. If the rate measure is defined as $H(d\omega, dp) = cp^{-1}e^{-cp}G_0(d\omega)dp$, with the same restrictions on c and G_0, then the corresponding CRM constructed as above is known as the gamma process.

The total mass of the gamma process $G, G(\Omega)$, is distributed as Gamma$(cG_0(\Omega), c)$. The improper distributions in these rate measures integrate to infinity over their respective domains, ensuring a countably infinite set of points in a draw from the Poisson process. For the beta process, the weights p_k are in [0,1], whereas for the gamma process they are in $[0, \infty)$. In both cases however the sum of the weights is finite, as can be seen from Campbell's theorem Kingman (1967), and is governed by c and the total mass of the base measure on Ω. For completeness we note that completely random measures as defined in Kingman (1967) have three components: a set of fixed atoms, a deterministic measure (usually assumed absent), and a random discrete measure. It is this third component that is explicitly generated using a Poisson process, though the fixed component can be readily incorporated into this construction Kingman (1993).

If we create an atomic measure by normalizing the weights $\{p_k\}$ from the gamma process, i.e. $D = \sum_{k=0}^{\infty} \pi_k \delta_{\omega_k}$ where $\pi_k = p_k / \sum_{i=0}^{\infty} p_i$, then D is known as a *Dirichlet process* Ferguson (1973), denoted as $D \sim \text{DP}(\alpha_0, H_0)$ where $\alpha_0 = G_0(\Omega)$ and $H_0 = G_0/\alpha_0$. It is not a CRM as the random variables induced on disjoint sets lack independence because of the normalization; it belongs to the class of normalized random measures with independent increments (NRMIs).

2.2 Stick-breaking for the Dirichlet and Beta Processes

A recursive way to generate the weights of random measures is given by stick-breaking, where a unit interval is subdivided into fragments based on draws from suitably chosen distributions. For example, the sick-breaking construction of the Dirichlet process Sethuraman (1994) is given by

$$D = \sum_{i=1}^{\infty} V_i \prod_{j=1}^{i-1} (1 - V_j) \delta_{\omega_i},$$

where $V_i \overset{iid}{\sim} \text{Beta}(1, \alpha), \quad \omega_i \overset{iid}{\sim} H_0$. Here the length of the first break from a unit-length stick is given by V_1. In the next round, a fraction V_2 of the remaining stick of length $1 - V_1$ is broken off, and we are left with a piece of length $(1-V_2)(1-V_1)$. The length of the piece in the next round is therefore given by $V_3(1 - V_2)(1 - V_1)$, and so on. Note that the weights belong to (0,1), and since this is a normalized measure, the weights sum to 1 almost surely. This is consistent with the use of the Dirichlet process as a prior on probability distributions.

This construction was generalized in Paisley et al. (2010) to yield stick-breaking for the beta process:

$$B = \sum_{i=1}^{\infty} \sum_{j=1}^{C_i} V_{ij}^{(i)} \prod_{l=1}^{i-1} (1 - V_{ij}^{(l)}) \delta_{\omega_{ij}}, \tag{1}$$

where $V_{ij}^{(i)} \overset{iid}{\sim} \text{Beta}(1, \alpha), \quad C_i \overset{iid}{\sim} \text{Poisson}(\gamma), \quad \omega_{ij} \overset{iid}{\sim} \frac{1}{\gamma} B_0$. We use this representation as the basis for our stick breaking-like construction of the Gamma CRM, and use Poisson process-based proof techniques similar to Paisley et al. (2012) to derive the rate measure.

3 The Stick-breaking Construction of the Gamma Process

3.1 Constructions and proof of correctness

We propose a simple recursive construction of the gamma process CRM, based on the stick-breaking construction for the beta process proposed in Paisley et al. (2010, 2012). In particular, we augment (or 'mark') a slightly modified stick-breaking beta process with an independent gamma-distributed random measure and show that the resultant Poisson process has the rate measure $H(d\omega, dp) = cp^{-1}e^{-cp} G_0(d\omega) dp$ as defined above. We show this by directly deriving the rate measure of the marked Poisson process using product distribution formulae. Our proposed stick-breaking construction is as follows:

$$G = \sum_{i=1}^{\infty} \sum_{j=1}^{C_i} G_{ij}^{(i)} V_{ij}^{(i)} \prod_{l=1}^{i} (1 - V_{ij}^{(l)}) \delta_{\omega_{ij}}, \tag{2}$$

where $G_{ij}^{(i)} \overset{iid}{\sim} \text{Gamma}(\alpha + 1, c), \quad V_{ij}^{(i)} \overset{iid}{\sim} \text{Beta}(1, \alpha), \quad C_i \overset{iid}{\sim} \text{Poisson}(\gamma), \quad \omega_{ij} \overset{iid}{\sim} \frac{1}{\gamma} H_0$. As with the beta process stick-breaking construction, the product of beta random variables allows us to interpret each j as corresponding to a stick that is being broken into an infinite number of pieces. Note that the expected weight on an atom in round i is $\alpha^i / c(1 + \alpha)^i$. The parameter c can therefore be used to control the weight decay cadence along with α.

The above representation provides the clearest view of the construction, but is somewhat cumbersome to deal with in practice, mostly due to the introduction of the additional gamma random variable. We reduce the number of random variables by noting that the product of a Beta$(1, \alpha)$ and a Gamma$(\alpha + 1, c)$ random variable has an Exp(c) distribution; we also perform a change of variables on the product of the $(1 - V_{ij})$s to arrive at the following equivalent construction, for which we now prove its correctness:

Theorem 1. *A gamma CRM with positive concentration parameters α and c and finite base measure H_0*

may be constructed as

$$G = \sum_{i=1}^{\infty} \sum_{j=1}^{C_i} E_{ij} e^{-T_{ij}} \delta_{\omega_{ij}}, \tag{3}$$

where $E_{ij} \overset{iid}{\sim} Exp(c)$, $T_{ij} \overset{ind}{\sim} Gamma(i, \alpha)$, $C_i \overset{iid}{\sim} Poisson(\gamma)$, $\omega_{ij} \overset{iid}{\sim} \frac{1}{\gamma} H_0$.

Proof. Note that, by construction, in each round i in (3), each set of weighted atoms $\{(\omega_{ij}, E_{ij}e^{-T_{ij}})\}_{j=1}^{C_i}$ forms a Poisson point process since the C_i are drawn from a Poisson(γ) distribution. In particular, each of these sets is a *marked* Poisson process Kingman (1993), where the atoms ω_{ij} of the Poisson process on Ω are marked with the random variables $E_{ij}e^{-T_{ij}}$ that have a probability measure on $(0, \infty)$. The superposition theorem of Kingman (1993) tells us that the countable union of Poisson process is itself a Poisson process on the same measure space; therefore denoting $G_i = \sum_{j=1}^{C_i} E_{ij} e^{-T_{ij}} \delta_{\omega_{ij}}$, we can say $G = \bigcup_{i=1}^{\infty} G_i$ is a Poisson process on $\Omega \times [0, \infty)$. We show below that the rate measure of this process equals that of the Gamma CRM.

Now, we note that the random variable $E_{ij}e^{-T_{ij}}$ has a probability measure on $[0, \infty)$; denote this by q_{ij}. We are going to mark the underlying Poisson process with this measure. The density corresponding to this measure can be readily derived using product distribution formulae. To that end, ignoring indices, if we denote $W = \exp(-T)$, then we can derive its distribution by a change of variable. Then, denoting $Q = E \times W$ where $E \sim \text{Exp}(c)$, we can use the product distribution formula to write the density of Q as

$$f_Q(q) = \int_0^1 \frac{\alpha^i}{\Gamma(i)} (-\log w)^{i-1} w^{\alpha-2} c e^{-c\frac{q}{w}} \mathrm{d}w,$$

where $T \sim \text{Gamma}(i, \alpha)$. Formally speaking, this is the Radon-Nikodym density corresponding to the measure q, since it is absolutely continuous with respect to the Lebesgue measure on $[0, \infty)$ and σ-finite by virtue of being a probability measure. Furthermore, these conditions hold for all the measures that we have in our union of marked Poisson processes; this allows us to write the density of the combined measure as

$$\begin{aligned} f(p) &= \sum_{i=1}^{\infty} \int_0^1 \frac{\alpha^i}{\Gamma(i)} (-\log w)^{i-1} w^{\alpha-2} c e^{-c\frac{p}{w}} \mathrm{d}w \\ &= \int_0^1 \sum_{i=1}^{\infty} \frac{\alpha^i}{\Gamma(i)} (-\log w)^{i-1} w^{\alpha-2} c e^{-c\frac{p}{w}} \mathrm{d}w \\ &= \int_0^1 \alpha w^{-2} c e^{-c\frac{p}{w}} \mathrm{d}w \\ &= \alpha p^{-1} e^{-cp} \\ &= c p^{-1} e^{-cp} \frac{\alpha}{c}, \end{aligned}$$

where we have used monotone convergence to get the Taylor expansion of $\exp(-\alpha \log w)$ inside the integral. Note that the measure defined on $\mathcal{B}([0, \infty))$ by the "improper" gamma distribution $p^{-1}e^{-cp}$ is σ-finite, in the sense that we can decompose $[0, \infty)$ into the countable union of disjoint intervals $[1/k, 1/(k-1))$, $k = 1, 2, \ldots \infty$, each of which has finite measure. In particular, the measure of the interval $[1, \infty)$ is given by the exponential integral.

Therefore the rate measure of the process G as constructed here is $G(d\omega, dp) = cp^{-1}e^{-cp}G_0(d\omega)dp$ where G_0 is the same as H_0 up to the multiplicative constant $\frac{\alpha}{c}$, and therefore satisfies the finiteness assumption imposed on H_0. □

We use the form specified in the theorem above in our variational inference algorithm since the variational distributions on almost all the parameters and variables in this construction lend themselves to simple closed-form exponential family updates. As an aside, we note that the random variables $(1 - V_{ij})$ have a Beta$(\alpha, 1)$ distribution; therefore if we denote $U_{ij} = 1 - V_{ij}$ then the construction in (2) is equivalent to

$$G = \sum_{i=1}^{\infty} \sum_{j=1}^{C_i} E_{ij}^{(i)} \prod_{l=1}^{i} U_{ij}^{(l)} \delta_{\omega_{ij}},$$

where $E_{ij}^{(i)} \overset{iid}{\sim} \text{Exp}(c)$, $U_{ij}^{(i)} \overset{iid}{\sim} \text{Beta}(\alpha, 1)$, $C_i \overset{iid}{\sim} \text{Poisson}(\gamma)$, $\omega_{ij} \overset{iid}{\sim} \frac{1}{\gamma} H_0$. This notation therefore relates our construction to the stick-breaking construction of the Indian Buffet Process Teh et al. (2007b), where the Bernoulli probabilities π_k are generated as products of iid Beta$(\alpha, 1)$ random variables : $\pi_1 = \nu_1$, $\pi_k = \prod_{i=1}^{k} \nu_i$ where $\nu_i \overset{iid}{\sim} \text{Beta}(\alpha, 1)$. In particular, we can view our construction as a generalization of the IBP stick-breaking, where the stick-breaking weights are multiplied with independent Exp(c) random variables, with the summation over j providing an explicit Poissonization.

3.2 Truncation analysis

The variational algorithm requires a truncation level for the number of atoms for tractability. Therefore we need to analyze the closeness between the marginal distributions of the data drawn from the full prior and the truncated prior, with the stick-breaking prior

weights integrated out. Our construction leads to a simpler truncation analysis if we truncate the number of rounds (indexed by i in the outer sum), which automatically truncates the atoms to a finite number. For this analysis, we will use the stick-breaking gamma process as the base measure of a Poisson likelihood process, which we denote by PP; this is precisely the model for which we develop variational inference in the next section. If we denote the gamma process as $G = \sum_{k=0}^{\infty} g_k \delta_{\omega_k}$, with g_k as the recursively constructed weights, then PP can be written as $PP = \sum_{k=0}^{\infty} p_k \delta_{\omega_k}$ where $p_k = \text{Poisson}(g_k)$. Under this model, we can obtain the following result, which is analogous to error bounds derived for other nonparametric models Ishwaran and James (2001); Doshi-Velez et al. (2009); Paisley et al. (2011) in the literature.

Theorem 2. *Let N samples $\boldsymbol{X} = (X_1, .., X_N)$ be drawn from $PP(G)$. If $G \sim \Gamma P(c, G_0)$, the full gamma process, then denote the marginal density of $\boldsymbol{X}$ as $\boldsymbol{m}_\infty(\boldsymbol{X})$. If G is a gamma process truncated after R rounds, denote the marginal density of $\boldsymbol{X}$ as $\boldsymbol{m}_R(\boldsymbol{X})$. Then*

$$\frac{1}{4}\int |\boldsymbol{m}_\infty(\boldsymbol{X}) - \boldsymbol{m}_R(\boldsymbol{X})| d\boldsymbol{X} \leq 1 - \exp\left\{-N\gamma\frac{\alpha}{c}\left(\frac{\alpha}{1+\alpha}\right)^R\right\}.$$

Proof. The starting intuition is that if we truncate the process after R rounds, then the error in the marginal distribution of the data will depend on the probability of positive indicator values appearing for atoms after the R^{th} round in the infinite version. Combining this with ideas analogous to those in Ishwaran and James (2000) and Ishwaran and James (2001), we get the following bound for the difference between the marginal distributions:

$$\frac{1}{4}\int |\mathbf{m}_\infty(\mathbf{X}) - \mathbf{m}_R(\mathbf{X})| d\mathbf{X}$$

$$\leq \mathbb{P}\left\{\exists (k,j), k > \sum_{r=1}^{R} C_r, 1 \leq n \leq N \text{ s.t. } X_n(\omega_{kj}) > 0\right\}.$$

Since we have a Poisson likelihood on the underlying gamma process, this probability can be written as

$$\mathbb{P}(\cdot) = 1 - \mathbb{E}\left[\mathbb{E}\left\{\left(\prod_{r=R+1}^{\infty}\prod_{j=1}^{C_r} e^{-\pi_{rj}}\right)^N \middle| C_r\right\}\right],$$

where $\pi_{rj} = G_{rj}^{(r)} V_{rj}^{(r)} \prod_{l=1}^{r}(1 - V_{rj}^{(l)})$. We may then use Jensen's inequality to bound it as follows:

$$\mathbb{P}(\cdot) \leq 1 - \exp\left[N \sum_{r=R+1}^{\infty} \mathbb{E}\left\{\sum_{j=1}^{C_r} \log(e^{-\pi_{rj}})\right\}\right]$$

$$= 1 - \exp\left[N\gamma\frac{1}{c}\sum_{r=R+1}^{\infty}\left(\frac{\alpha}{1+\alpha}\right)^r\right]$$

$$= 1 - \exp\left\{-N\gamma\frac{\alpha}{c}\left(\frac{\alpha}{1+\alpha}\right)^R\right\}.$$

□

4 Variational Inference

As discussed in Section 3.2, we will focus on the infinite Gamma-Poisson model, where a gamma process prior is used in conjunction with a Poisson likelihood function. When integrating out the weights of the gamma process, this process is known to yield a nonparametric prior for sparse, infinite count matrices Titsias (2008). We note that our approach should easily be applicable to other models involving gamma process priors.

4.1 The Model

To effectively perform variational inference, we rewrite G as a single sum of weighted atoms, using indicator variables $\{d_k\}$ for the rounds in which the atoms occur, similar to Paisley et al. (2010):

$$G = \sum_{k=1}^{\infty} E_k e^{-T_k} \delta_{\omega_k}, \tag{4}$$

where $E_k \overset{iid}{\sim} \text{Exp}(c)$, $T_k \overset{ind}{\sim} \text{Gamma}(d_k, \alpha)$, $\sum_{k=1}^{\infty} \mathbb{1}_{(d_k=r)} \overset{iid}{\sim} \text{Poisson}(\gamma)$, $\omega_k \overset{iid}{\sim} \frac{1}{\gamma} H_0$. We also place gamma priors on α, γ and $c : \alpha \sim \text{Gamma}(a_1, a_2), \gamma \sim \text{Gamma}(b_1, b_2), c \sim \text{Gamma}(c_1, c_2)$. Denoting the data, the latent prior variables and the model hyperparameters by $\mathcal{D}, \Pi$ and Λ respectively, the full likelihood may be written as $P(\mathcal{D}, \Pi|\Lambda) = P(\mathcal{D}, \Pi_{-G}|\Pi_G, \Lambda) \cdot P(\Pi_G|\Lambda)$ where $P(\Pi_G|\Lambda) = P(\alpha) \cdot P(\gamma) \cdot P(c) \cdot P(\mathbf{d}|\gamma) \cdot \prod_{k=1}^{K} P(E_k|c) \cdot P(T_k|d_k, \alpha) \cdot \prod_{n=1}^{N} P(z_{nk}|E_k, T_k)$. We truncate the infinite gamma process to K atoms, and take N to be the total number of datapoints. Π_{-G} denotes the set of the latent variables excluding those from the Poisson-Gamma prior; for instance, in factor analysis for topic models, this contains the Dirichlet-distributed factor variables (or topics).

From the Poisson likelihood, we have $z_{nk}|E_k, T_k \sim \text{Poisson}(E_k e^{-T_k})$, independently for each n. The distributions of T_k and $\mathbf{d}$ involve the indicator functions on the round indicator variables d_k:

$$P(T_k|d_k, \alpha) = \frac{\alpha^{\nu_k(0)}}{\prod_{r\geq 1}\Gamma(r)^{\mathbb{1}_{(d_k=r)}}} T_k^{\nu_k(1)} e^{-\alpha T_k},$$

where $\nu_k(s) = \sum_{r \geq 1} (r-s)\mathbb{1}_{(d_k=r)}$. We use the same weighting factors in our distribution on $\mathbf{d}$ as Paisley et al. (2011). See Paisley et al. (2011) for a discussions on how to approximate these factors in the variational algorithm.

4.2 The Variational Prior Distribution

Mean-field variational inference involves minimizing the KL divergence between the model posterior, and a suitably constructed *variational* distribution which is used as a more tractable alternative to the actual posterior distribution. To that end, we propose a fully-factorized variational distribution on the Poisson-Gamma prior as follows:

$$Q = q(\alpha)\cdot q(\gamma)\cdot q(c)\cdot \prod_{k=1}^{K} q(E_k)\cdot q(T_k)\cdot q(d_k)\cdot \prod_{n=1}^{N} q(z_{nk}),$$

where $q(E_k) \sim \text{Gamma}(\acute{\xi_k}, \acute{\epsilon_k})$, $q(T_k) \sim \text{Gamma}(\acute{u_k}, \acute{v_k})$, $q(\alpha) \sim \text{Gamma}(\kappa_1, \kappa_2)$, $q(\gamma) \sim \text{Gamma}(\tau_1, \tau_2)$, $q(c) \sim \text{Gamma}(\rho_1, \rho_2)$, $q(z_{nk}) \sim \text{Poisson}(\lambda_{nk})$, $q(d_k) \sim \text{Mult}(\varphi_k)$.

Instead of working with the actual KL divergence between the full posterior and the factorized proxy distribution, variational inference maximizes what is canonically known as the *evidence lower bound* (ELBO), a function that is the same as the KL divergence up to a constant. In our case it may be written as $\mathcal{L} = \mathbb{E}_Q \log P(\mathcal{D}, \Pi|\Lambda) - \mathbb{E}_Q \log Q$. We omit the full representation here for brevity.

4.3 The Variational Parameter Updates

Since we are using exponential family variational distributions, we leverage the closed form variational updates for exponential families wherever we can, and perform gradient ascent on the ELBO for the parameters of those distributions which do not have closed form updates. We list the updates on the distributions of the prior below. The closed-form updates for the hyperparameters in $q(E_k), q(\alpha), q(c)$ and $q(\gamma)$ are as follows:

$$\acute{\xi_k} = \sum_{n=1}^{N} \mathbb{E}_Q(z_{nk}) + 1, \quad \acute{\epsilon_k} = \mathbb{E}(c) + N \times \mathbb{E}_Q\left[e^{-T_k}\right],$$

$$\kappa_1 = \sum_{k=1}^{K}\sum_{r\geq 1} r\varphi_k(r) + a_1, \kappa_2 = \sum_{k=1}^{K} \mathbb{E}_Q(T_k) + a_2,$$

$$\rho_1 = c_1 + K, \quad \rho_2 = \sum_{k=1}^{K} \mathbb{E}_Q(E_k) + c_2,$$

$$\tau_1 = b_1 + K, \quad \tau_2 = \sum_{r\geq 1}\left\{1 - \prod_{k=1}^{K}\sum_{\acute{r}=1}^{r-1}\varphi_k(\acute{r})\right\} + b_2.$$

The updates for the multinomial probabilities in $q(d_k)$ are given by:

$$\varphi_k(r) \propto \exp\{r\mathbb{E}_Q(\log\alpha) - \log\Gamma(r) + (r-1)\mathbb{E}_Q(\log T_k) - \zeta\cdot\sum_{i\neq k}\varphi_i(r) - \mathbb{E}_Q(\gamma)\sum_{j=2}^{r}\prod_{k'\neq k}\sum_{r'=1}^{j-1}\varphi_{k'}(r')\}.$$

The variational distribution $q(T_k)$ does not lend itself to closed-form analytical updates, so we perform gradient ascent on the evidence lower bound. The variational updates for $q(z_{nk})$ and for the variational distributions on the latent variables in Π_{-G} are model dependent, and require some approximations for the factor analysis case. See Roychowdhury and Kulis (2014) for details.

5 Other Algorithms

Here we briefly describe the two primary competing algorithms we developed based on constructions of the Gamma process: a variational inference algorithm from the naïve construction, and a Markov chain Monte Carlo sampler based on our construction.

5.1 Naïve Variational Inference

We derive a variational inference algorithm from a simpler construction of the Gamma process, where we multiply the stick-breaking construction of the Dirichlet process by a Gamma random variable. The construction can be written as:

$$G = G_0 \sum_{i=1}^{\infty} V_i \prod_{j=1}^{i-1}(1-V_j)\delta_{\omega_i},$$

where $G_0 \sim \text{Gamma}(\alpha, c)$, $V_i \overset{iid}{\sim} \text{Beta}(1, \alpha)$, $\omega_i \overset{iid}{\sim} H_0$.

We use an equivalent form of the construction that is similar to the one used above :

$$G = G_0 \sum_{k=1}^{\infty} V_k e^{-T_k}\delta_{\omega_k},$$

where $G_0 \sim \text{Gamma}(\alpha, c)$, $V_k \overset{iid}{\sim} \text{Beta}(1, \alpha)$, $T_k \overset{ind}{\sim} \text{Gamma}(k-1, \alpha)$, $\omega_i \overset{iid}{\sim} H_0$.

As before, we place gamma priors on α and c : $\alpha \sim \text{Gamma}(a_1, a_2), c \sim \text{Gamma}(c_1, c_2)$. The closed-form coordinate ascent updates for G_0, α and c and the gradient ascent updates for $\{V_k, T_k\}$ are detailed in the supplementary.

5.2 The MCMC Sampler

As a baseline, we also derive and compare the variational algorithm with a standard MCMC sampler for this model. We use the construction in (4) for sampling

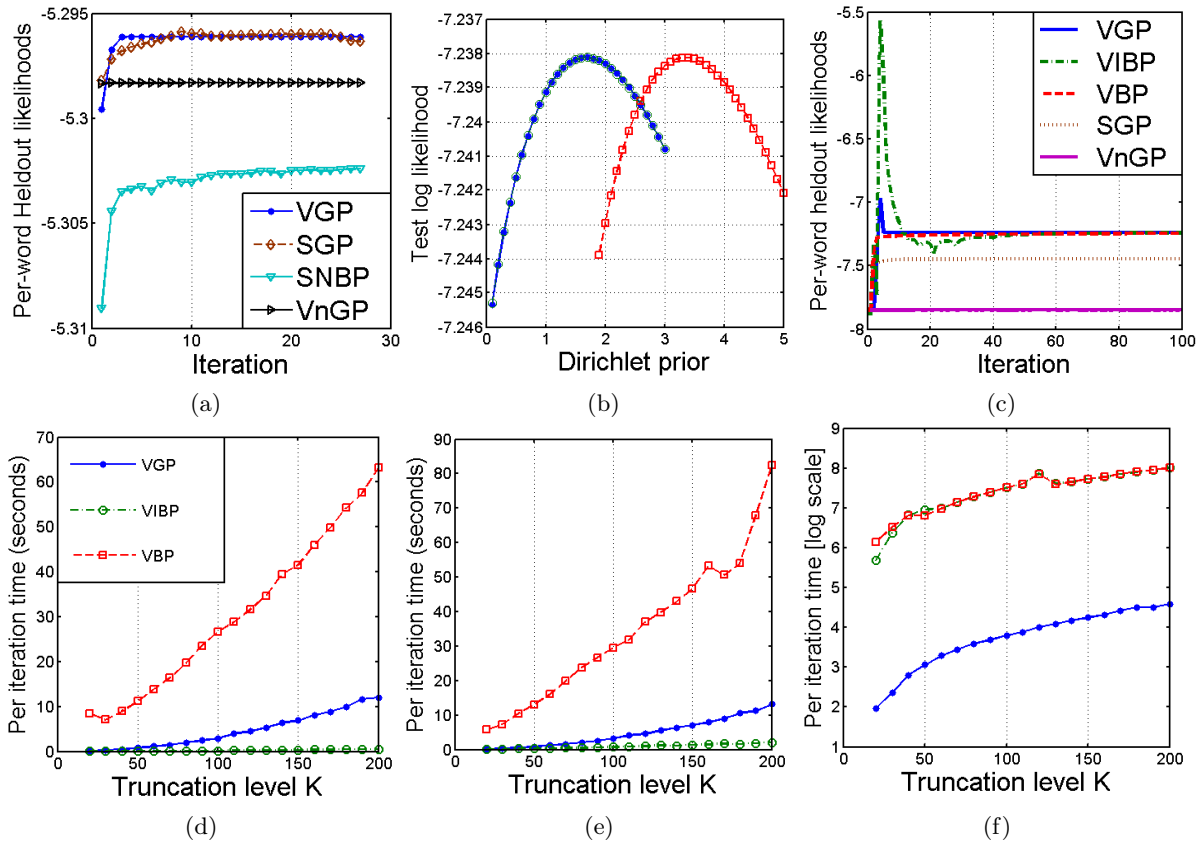

Figure 1: Plots of held-out test likelihoods and per-iteration running times (best viewed in color). Plots (d), (e) and (f) are for PsyRev, KOS, and NYT respectively. Plots (b) and (c) are for the PsyRev dataset. Algorithm trace colors are common to plots. See text for full details.

from the model. To avoid inferring the latent variables in all the atom weights of the Poisson-Gamma prior, we use Monte Carlo techniques to integrate them out, as in Paisley et al. (2010). This affects posterior inference for the indicators z_{nk}, the round indicators $\mathbf{d}$ and the hyperparameters c and α. The posterior distribution for γ is closed form, as are those for the likelihood latent variables in Π_{-G}. The complete updates are described in the supplementary.

6 Experiments

We consider the problem of learning latent topics in document corpora. Given an observed set of counts of vocabulary words in a set of documents, represented by say a $V \times N$ count matrix, where V is the vocabulary size and N the number of documents, we aim to learn K latent factors and their vocabulary realizations using Poisson factor analysis. In particular, we model the observed corpus count matrix D as $D \sim \text{Poi}(\Phi\mathbf{I})$, where the $V \times K$ matrix Φ models the factor loadings, and the $K \times N$ matrix $\mathbf{I}$ models the actual factor counts in the documents.

We implemented and analyzed the performance of three variational algorithms corresponding to four different priors on $\mathbf{I}$: the Poisson-gamma process prior from this paper (abbreviated hereafter as VGP), a Poisson-gamma prior using the naïve construction of the gamma process (VnGP), the Bernoulli-beta prior from Paisley et al. (2011) (VBP) and the IBP prior from Doshi-Velez et al. (2009) (VIBP), along with the MCMC sampler mentioned above (SGP). For the Bernoulli-beta priors we modeled $\mathbf{I}$ as $\mathbf{I} = W \circ Z$ as in Paisley et al. (2011), where the nonparametric priors are put on Z and a vague Gamma prior is put on W. For the VGP and SGP models we set $\mathbf{I} = Z$. In addition, for all four algorithms, we put a symmetric Dirichlet$(\beta_1, \ldots, \beta_V)$ prior on the columns of Φ. We added corresponding variational distributions for the variables in the collection denoted as Π_{-G} above. We use held-out per-word test log-likelihoods and times required to update all variables in Π in each iteration as our comparison metrics, with 80% of the data used for training. We used the same likelihood metric as Zhou and Carin (2012), with the samples replaced by the expectations of the variational distributions.

Synthetic Data. As a warm-up, we consider the performances of VGP and SGP on some synthetic data generated from this model. We generate 200 weighted atoms from the gamma prior using the stick-breaking construction, and use the Poisson likelihood to gen-

erate 3000 values for each atom to yield the indicator matrix Z. We simulated a vocabulary of 200 terms, generated a 200×200 factor-loading matrix Φ using symmetric Dirichlet priors, and then generated $D = \text{Poi}(\Phi Z)$. For the VGP and VnGP, we measure the test likelihood after every iteration and average the results across 10 random restarts. These measurements are plotted in fig.1a. As shown, VGP's measured heldout likelihood converges within 10 iterations. The SGP traceplot shows the first thirty heldout likelihoods measured after burn-in. Per-iteration times were 15 seconds and 2.36 minutes for VGP (with K=125) and SGP respectively. The SGP learned K online, with values oscillating around 50. SNBP refers to the Poisson-Gamma mixture ("NB process") sampler from Zhou and Carin (2012). Its traceplot shows the first 30 likelihoods measured after 1000 burn-in iterations. We see that it performed similarly to our algorithms, though slightly worse.

Real data. We used a similar framework to model the count data from the KOS[1], NIPS[2], Psychological Review (PsyRev)[3], and New York Times[1] corpora. The vocabulary sizes are 2566, 13649, 6906 and 100872 respectively, while the document counts are 1281, 1740, 3430 and 300000 respectively. For each dataset, we ran all three variational algorithms with 10 random restarts each, measuring the held-out log-likelihoods and per-iteration runtimes for different values of the truncation factor K. The learning rates for gradient ascent updates were kept on the order of 10^{-4} for both VGP and VBP, with 5 gradient steps per iteration. A representative subset of results is shown in figs.1b through 1f.

We used vague gamma priors on the hyperparameters α, γ and c in the variational algorithms, and improper (1) priors for the sampler. We found the test likelihoods to be independent of these initializations. The results for the variational algorithms were dependent on the Dirichlet prior β on Φ, as noted in fig.1b. We therefore used the learned test likelihood after 100 iterations as a heuristic to select β. We found the three variational algorithms to attain very similar test likelihoods across all four datasets after a few hours of CPU time, with the VGP and VBP having a slight edge over the VIBP. The sampler somewhat unexpectedly did not attain a competitive score for any dataset, unlike the synthetic case. For instance, as shown in fig.1c, it oscillated around -7.45 for the PsyRev dataset, whereas the variational algorithms attained -7.23. For comparison, the NB process sampler from Zhou and Carin (2012) attains -7.25 each iteration after 1000 iterations of burn-in. VnGP was the worst performer, with stable log-likelihood of -7.85. Also as seen in fig.1c, VGP was faster to convergence (in less than 10 iterations in ∼5 seconds) than VIBP and VBP (∼50 iterations each). The test log-likelihoods after a few hours of runtime were largely independent of the truncation K for the three variational algorithms. Behavior for the other datasets was similar.

Among the three variational algorithms, the VIBP scaled best for small to medium datasets as a function of the truncation factor due to all updates being closed-form, in spite of having to learn the additional weight matrix W. The VGP running times were competitive for small values of K for these datasets. However, in the large NYT dataset, VGP was orders of magnitude faster than the Bernoulli-beta algorithms (note the log-scale in fig.1f). For example, with a truncation of 100 atoms, VGP took around 45 seconds per iteration, whereas both VIBP and VBP took more than 3 minutes. The VBP scaled poorly for all datasets, as seen in figs.1d through 1f. The reason for this is three-fold: learning the parameters for the additional matrix W which is directly affected by dimensionality (also the reason for VIBP being slow for NYT dataset), gradient updates for two variables (as opposed to one for VGP) and a Taylor approximation required for these gradient updates (see Paisley et al. (2011)). The sampler SGP required around 7 minutes per iteration for the small datasets and an hour and 40 minutes on average for NYT.

To summarize, we found the VGP to post running times that are competitive with the fastest algorithm (VIBP) in small to medium datasets, and outperform the other methods completely in the large NYT dataset, all the while providing similar accuracy compared to the other variational algorithms, as measured by held-out likelihood. It was also the fastest to converge, typically taking less than 15 iterations. Compared with SGP, our variational method is substantially faster (particularly on large-scale data) and produces higher likelihood scores on real data.

7 Conclusion

We have described a novel stick-breaking representation for gamma processes and used it to derive a variational inference algorithm. This algorithm has been shown to be far more scalable for large datasets than related variational algorithms, while attaining similar accuracy and outperforming sampling-based methods. We expect that recent improvements to variational techniques can also be applied to our algorithm, potentially yielding even further scalability.

Acknowledgements

This work was supported by NSF award IIS-1217433.

[1] https://archive.ics.uci.edu/ml/datasets/Bag+of+Words
[2] http://www.stats.ox.ac.uk/ teh/data.html
[3] http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

References

Blei, D. and Jordan, M. (2003). Variational methods for Dirichlet process mixtures. *Bayesian Analysis*, 1:121–144.

Broderick, T., Mackey, L., Paisley, J., and Jordan, M. I. (2014). Combinatorial clustering and the beta negative binomial process. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Caron, F. and Fox, E. B. (2013). Bayesian Nonparametric Models of Sparse and Exchangeable Random Graphs. arXiv:1401.1137.

Caron, F., Teh, Y. W., and Murphy, B. T. (2013). Bayesian Nonparametric Plackett-Luce models for the Analysis of Clustered Ranked Data. arXiv:1211.5037.

Doshi-Velez, F., Miller, K., Gael, J. V., and Teh, Y. W. (2009). Variational Inference for the Indian Buffet Process. In *AISTATS*.

Ferguson, T. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230.

Gopalan, P., Ruiz, F., Ranganath, R., and Blei, D. (2014). Bayesian nonparametric Poisson factorization. In *AISTATS*.

Ishwaran, H. and James, L. F. (2000). Approximate Dirichlet Process Computing in Finite Normal Mixtures: Smoothing and Prior Information. *Journal of Computational and Graphical Statistics*, 11:508–532.

Ishwaran, H. and James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96:161–173.

Jordan, M. I. (2010). Hierarchical Models, Nested Models and Completely Random Measures. In Chen, M.-H., Dey, D., Mueller, P., Sun, D., and Ye, K., editors, *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger*. New York: Springer.

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

Kingman, J. F. C. (1993). *Poisson Processes*, volume 3 of *Oxford Studies in Probability*. Oxford University Press, New York.

Miller, K. (2011). *Bayesian Nonparametric Latent Feature Models*. PhD thesis, University of California at Berkeley.

Orbanz, P. and Williamson, S. (2012). Unit-rate Poisson representations of completely random measures.

Paisley, J., Blei, D. M., and Jordan, M. I. (2012). Stick-Breaking Beta Processes and the Poisson Process. In *Artificial Intelligence and Statistics*.

Paisley, J., Carin, L., and Blei, D. M. (2011). Variational Inference for Stick-Breaking Beta Process Priors. In *International Conference on Machine Learning*.

Paisley, J., Zaas, A., Woods, C. W., Ginsburg, G. S., and Carin, L. (2010). A Stick-Breaking Construction of the Beta Process. In *International Conference on Machine Learning*.

Roychowdhury, A. and Kulis, B. (2014). Gamma Processes, Stick-Breaking, and Variational Inference. arXiv:1410.1068.

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

Teh, Y., Kurihara, K., and Welling, M. (2007a). Collapsed variational inference for HDP. In *NIPS*.

Teh, Y. W., Görür, D., and Ghahramani, Z. (2007b). Stick-breaking construction for the Indian buffet process. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 11.

Thibaux, R. (2008). *Nonparametric Bayesian Models for Machine Learning*. PhD thesis, University of California at Berkeley.

Titsias, M. (2008). The Infinite Gamma-Poisson Model. In *Advances in Neural Information Processing Systems*.

Wang, C., Paisley, J., and Blei, D. (2011). Online variational inference for the hierarchical Dirichlet process. In *AISTATS*.

Wolpert, R. and Ickstadt, K. (1998). Simulation of Lévy Random Fields. In *Practical Nonparametric and Semiparametric Bayesian Statistics*. Springer-Verlag.

Zhou, M. and Carin, L. (2012). Augment-and-conquer negative binomial processes. In *NIPS*.

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