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

Appendices

This document contains supplementary material for the submission “Gamma Processes, Stick-Breaking, and Variational Inference”.

A Variational inference details

To effectively perform variational inference, we re-write 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). We re-state our construction of the gamma CRM that we use for the inference algorithms:

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

where $E_k \overset{iid}{\sim} \text{Exp}(c)$, $T_k \overset{ind}{\sim} \text{Gamma}(d_k, \alpha)$, $\sum_{k=1}^{\infty} \mathbb{1}_{(d_k=r)} \overset{iid}{\sim} \text{Poisson}(\gamma)$, $\omega_k \overset{iid}{\sim} \frac{1}{\gamma} H_0$. Here d_k denotes the round in which atom k appears, and may be defined as $d_k \triangleq 1 + \sum_{i=1}^{\infty} \mathbb{I}\left\{\sum_{j=1}^{i} C_j < k\right\}$. Conversely, given the round indicators $\mathbf{d} = \{d_k\}$, we can recover the round-specific atom counts as $C_i = \sum_{k=1}^{\infty} \mathbb{I}(d_k = i)$.

We 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 $P(\mathcal{D}, \Pi|\Lambda) =$

$$P(\mathcal{D}, \Pi_{-G}|\Pi_G, \Lambda) \cdot P(\alpha) \cdot P(\gamma) \cdot P(c) \cdot P(\mathbf{d}|\gamma) \cdot \prod_{k=1}^{K} P(E_k|c) \cdot P(T_k|d_k, \alpha) \cdot \prod_{n=1}^{N} P(z_{nk}|E_k, T_k),$$

with Π_{-G} denoting the set of the latent variables excluding those from the Poisson-Gamma prior. The distribution of $\mathbf{d}$ is given by $P(\mathbf{d}|\gamma) =$

$$\prod_{r=1}^{\infty} \frac{\gamma^{\sum_k \mathbb{1}_{(d_k=r)}}}{\left(\sum_k \mathbb{1}_{(d_k=r)}\right)!} \cdot \exp\left\{-\gamma \mathbb{I}\left(\sum_{r'=r}^{\infty} \sum_{k=1}^{\infty} \mathbb{1}_{(d_k=r')} > 0\right)\right\}.$$

See Paisley et al. (2011) for discussions on how to approximate some of these factors in the variational algorithm.

A.1 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 $q(E_k) \sim \text{Gamma}(\acute{\xi_k}, \acute{\epsilon_k})$, $q(T_k) \sim \text{Gamma}(\acute{u_k}, \acute{v_k})$, $q(\alpha) \sim \text{Gamma}(\kappa_1, \kappa_2)$, $q(\gamma) \sim \text{Gamma}(\tau_1, \tau_2)$, $q(c) \sim \text{Gamma}(\rho_1, \rho_2)$, $q(z_{nk}) \sim \text{Poisson}(\lambda_{nk})$, $q(d_k) \sim \text{Mult}(\varphi_k)$.

The *evidence lower bound* (ELBO) may therefore be written as $\mathcal{L} = \mathbb{E}_Q \log P(\mathcal{D}, \Pi|\Lambda) - \mathbb{E}_Q \log Q$, with the relevant distributions described above.

A.2 Variational parameter updates

We first re-state the closed form updates for the variational distributions on the prior variables. The 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:

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

Next we describe the gradient ascent updates on $q(T_k)$ and the updates on $q(\Pi_{-G})$ and $q(z_{nk})$.

The gradients for the two variational parameters in $q(T_k)$ are:

$$\frac{\partial\mathcal{L}}{\partial\acute{u_k}} = \sum_{r\geq 1}(r-1)\varphi_k(r)\psi^{'}(\acute{u_k}) - \frac{\mathbb{E}_Q(\alpha)}{\acute{\upsilon_k}} - \sum_{n=1}^{N}\mathbb{E}_Q(E_k)\left(\frac{\acute{\upsilon_k}}{\acute{\upsilon_k}+1}\right)^{\acute{u_k}}\cdot\log\left(\frac{\acute{\upsilon_k}}{\acute{\upsilon_k}+1}\right) - \sum_{n=1}^{N}\mathbb{E}_Q(z_{nk})\frac{1}{\acute{\upsilon_k}} - (\acute{u_k}-1)\psi^{'}(\acute{u_k}) - 1$$

$$\frac{\partial\mathcal{L}}{\partial\acute{\upsilon_k}} = -\sum_{r\geq 1}(r-1)\varphi_k(r)\frac{1}{\acute{\upsilon_k}} + \mathbb{E}_Q(\alpha)\frac{\acute{u_k}}{(\acute{\upsilon_k})^2} - \sum_{n=1}^{N}\mathbb{E}_Q(E_k)\acute{u_k}\frac{\acute{\upsilon_k}^{\acute{u_k}-1}}{(\acute{\upsilon_k}+1)^{\acute{u_k}+1}} + \sum_{n=1}^{N}\mathbb{E}_Q(z_{nk})\frac{\acute{u_k}}{(\acute{\upsilon_k})^2} - \frac{1}{\acute{\upsilon_k}}.$$

For the topic modeling problems, we model the observed vocabulary-vs-document corpus count matrix D as $D \sim \text{Poi}(\Phi Z)$, where the $V \times K$ matrix Φ models the factor loadings, and the $K \times N$ matrix Z models the actual factor counts in the documents. We put the $K-$truncated Poisson-Gamma prior on Z, and put a Dirichlet$(\beta_1, \ldots, \beta_V)$ prior on the columns of Φ.

The variational distribution Q consequently gets a Dirichlet$(\Phi|\{\mathbf{b}\}_k)$ distribution multiplied to it, where $\mathbf{b} = (b_1, \ldots, b_V)$ are the variational Dirichlet hyperparameters. This setup does not immediately lend itself to closed form updates for the b-s, so we resort to gradient ascent. The gradient of the ELBO with respect to each variational hyperparameter is

$$\frac{\partial\mathcal{L}}{\partial b_{vk}} = -\mathbb{E}_Q(z_{nk})\cdot\frac{\sum_v b_{vk} - b_{vk}}{\left(\sum_v b_{vk}\right)^2} + \psi^{'}(b_{vk}) \cdot\left(\beta_v - b_{vk} + \sum_n d_{vn}\right) + \psi^{'}(\sum_v b_{vk})\times \left(\sum_v b_{vk} - V - \beta_v - \sum_n d_{vn} + 1\right).$$

In practice however we found a closed-form update facilitated by a simple lower bound on the ELBO to converge faster. We describe the update here. First note that the part of the ELBO relevant to a potential closed form variational update of ϕ_{vk} can be written as

$$\mathcal{L} = -\phi_{vk}\cdot\sum_n\mathbb{E}_Q(z_{nk}) + \sum_n d_{vn}\cdot\log\phi_{vk} + \cdots,$$

which can then be lower bounded as

$$\mathcal{L} \geq \log\phi_{vk}\cdot\left(-\sum_n\mathbb{E}_Q(z_{nk}) + \sum_n d_{vn}\right) + \cdots.$$

This allows us to analytically update b_{vk} as $b_{vk} = -\sum_n\mathbb{E}_Q(z_{nk}) + \sum_n d_{vn} + \beta_v$. This frees us from having to choose appropriate corpus-specific initializations and learning rates for the Φs.

A similar lower bound on the ELBO allows us to update the variational parameters of $q(z_{nk})$ as $\lambda_{nk} = -1 - \sum_v d_{vn} + \mathbb{E}_Q(\log E_k) + \mathbb{E}_Q(T_k)$.

B Variational inference using denormalized DP construction

We describe our algorithm derived from the simpler construction of the Gamma process by multiplying 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 $G_0 \sim \text{Gamma}(\alpha, c)$, $\quad V_i \overset{iid}{\sim} \text{Beta}(1,\alpha)$, $\quad \omega_i \overset{iid}{\sim} H_0$.

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

As before, we place gamma priors on α and c : $\alpha \sim \text{Gamma}(a_1, a_2), c \sim \text{Gamma}(c_1, c_2)$.

Our variational distribution for this prior is as follows:

$$Q = q(G_0) \cdot q(\alpha) \cdot q(c) \cdot \prod_{k=1}^{K} q(V_k) \cdot q(T_k) \cdot \prod_{n=1}^{N} q(z_{nk}),$$

where $q(G_0) \sim \text{Gamma}(g_1, g_2)$, $q(V_k) \sim \text{Beta}(\nu_{k1}, \nu_{k2})$, $q(T_k) \sim \text{Gamma}(t_{k1}, t_{k2})$, $q(\alpha) \sim \text{Gamma}(\kappa_1, \kappa_2)$, $q(c) \sim \text{Gamma}(\rho_1, \rho_2)$, $q(z_{nk}) \sim \text{Poisson}(\lambda_{nk})$.

The closed form updates for the variational hyperparameters for $\alpha, G_0,$ and c are as follows:

$$\kappa_1 = a_1, \quad \kappa_2 = a_2 - \mathbb{E}_Q(\log G_0) - \sum_k \mathbb{E}_Q(\log(1 - V_k) + \sum_k \mathbb{E}_Q(T_k),$$

$$g_1 = \alpha + \sum_{n=1}^{N} \sum_k \mathbb{E}_Q(z_{nk}), \quad g_2 = N \cdot \sum_k \mathbb{E}_Q(V_k e^{-T_k}),$$

$$\rho_1 = c_1, \quad \rho_2 = c_2 + \mathbb{E}_Q(G_0).$$

The updates for $q(V_k)$ and $q(T_k)$ are not closed form, necessitating gradient ascent steps. The gradients for the variational parameters in $q(V_k)$ are:

$$\frac{\partial \mathcal{L}}{\partial \nu_{k1}} = \psi^{'}(\nu_{k1} + \nu_{k2}) \left[\nu_{k1} + \nu_{k2} - \alpha - \sum_{n=1}^{N} \mathbb{E}_Q(z_{nk}) - 1 \right] + \psi^{'}(\nu_{k1}) \cdot \left[\sum_{n=1}^{N} \mathbb{E}_Q(z_{nk}) - \nu_{k1} + 1 \right] - N \cdot \mathbb{E}_Q \left(G_0 e^{-T_k} \right) \frac{\nu_{k2}}{\nu_{k1} + \nu_{k2}}$$

$$\frac{\partial \mathcal{L}}{\partial \nu_{k2}} = \psi^{'}(\nu_{k1} + \nu_{k2}) \left[\nu_{k1} + \nu_{k2} - \alpha - \sum_{n=1}^{N} \mathbb{E}_Q(z_{nk}) - 1 \right] - N \cdot \mathbb{E}_Q \left(G_0 e^{-T_k} \right) \frac{\nu_{k1}}{\nu_{k1} + \nu_{k2}} + \psi^{'}(\nu_{k2}) \cdot [\alpha - \nu_{k2}].$$

The gradients for the variational parameters in $q(T_k)$ are:

$$\frac{\partial \mathcal{L}}{\partial t_{k1}} = 1 + \psi^{'}(t_{k1}) \cdot (k - t_{k1} - 1) - \log t_{k2} - \frac{1}{t_{k2}} \left(\alpha + \sum_{n=1}^{N} \mathbb{E}_Q(z_{nk}) \right) - N \cdot \mathbb{E}_Q(G_0 V_k) \cdot \frac{\partial}{\partial t_{k1}} \left(\frac{t_{k2}}{t_{k2} + 1} \right)^{t_{k1}}$$

$$\frac{\partial \mathcal{L}}{\partial t_{k2}} = \frac{t_{k1}}{t_{k2}^2} (\alpha + \sum_{n=1}^{N} \mathbb{E}_Q(z_{nk})) - \frac{1}{t_{k2}}(k-1) - N \cdot \mathbb{E}_Q(G_0 V_k) \cdot \frac{\partial}{\partial t_{k2}} \left(\frac{t_{k2}}{t_{k2} + 1} \right)^{t_{k1}}$$

.

C Markov chain Monte Carlo sampling details

We re-write the construction of the Poisson-Gamma prior:

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

$E_k \overset{iid}{\sim} \text{Exp}(c)$, $\quad T_k \overset{ind}{\sim} \text{Gamma}(d_k, \alpha)$, $\quad \sum_{k=1}^{\infty} \mathbb{1}_{(d_k = r)} \overset{iid}{\sim} \text{Pois}(\gamma)$, $\quad \omega_k \overset{iid}{\sim} \frac{1}{\gamma} H_0$. We put improper priors on α and c, and a noninformative Gamma prior on γ. The indicator counts are given by $Z_{nk} \sim \text{Pois}(g_k)$, where $g_k = E_k e^{-T_k}$. To avoid sampling the atom weights E_k and T_k, we integrate them out using Monte Carlo techniques in the sampling steps for the prior.

C.1 Sampling the round indicators

The conditional posterior for the round indicators $\mathbf{d} = \{d_k\}_{k=1}^{K}$ can be written as

$$p\left(d_k = i | \{d_l\}_{l=1}^{k-1}, \{Z_{nk}\}_{n=1}^{N}, \alpha, c, \gamma\right) \propto p\left(\{Z_{nk}\}_{n=1}^{N} | d_k = i, \alpha, c\right) p\left(d_k = i | \{d_l\}_{l=1}^{k-1}\right).$$

For the first factor, we collapse out the stick-breaking weights and approximate the resulting integral using Monte-Carlo techniques as follows:

$$p\left(\{Z_{nk}\}_{n=1}^{N} | d_k = i, \alpha, c\right) = \int_{[0,\infty]^i} \prod_{n=1}^{N} \text{Pois}(Z_{nk} | g_k) \mathrm{d}G \approx \frac{1}{S} \sum_{s=1}^{S} \prod_{n=1}^{N} \text{Pois}(Z_{nk} | g_k^{(s)}),$$

where $g_k^{(s)} = E_k^{(s)} e^{-T_k^{(s)}} \overset{d}{=} V_{k,d_k}^{(s)} \prod_{l=1}^{d_k} (1 - V_{kl}^{(s)})$. Here S is the number of simulated samples from the integral over the stick-breaking weights. We take $S = 1000$ in our experiments.

The second factor is the same as Paisley et al. (2010):

$$p(d_k = d | \gamma, \{d_l\}_{l=1}^{k-1}) = \begin{cases} 0 & \text{if } d < d_{k-1} \\ \frac{1 - \sum_{t=1}^{D_{k-1}} \text{Pois}(t|\gamma)}{1 - \sum_{t=1}^{D_{k-1}-1} \text{Pois}(t|\gamma)} & \text{if } d = d_{k-1} \\ \left(1 - \frac{1 - \sum_{t=1}^{D_{k-1}} \text{Pois}(t|\gamma)}{1 - \sum_{t=1}^{D_{k-1}-1} \text{Pois}(t|\gamma)}\right) (1 - \text{Pois}(0|\gamma)) \text{Pois}(0|\gamma)^{h-1} & \text{if } d = d_{k-1} + h \end{cases}$$

Here $D_k \triangleq \sum_{j=1}^{k} \mathbb{I}(d_j = d_k)$. Normalizing the product of these two factors over all i is infeasible, so we evaluate this product for increasing i till it drops below 10^{-2}, and normalize over the gathered values.

C.2 Sampling the factor variables

Here we consider the Poisson factor modeling scenario that we use to model vocabulary-document count matrices. Recall that a $V \times N$ count matrix D is modeled as $D = \text{Poi}(\Phi Z)$, where the $V \times K$ matrix Φ models the factor loadings, and the $K \times N$ matrix Z models the actual factor counts in the documents.. We put the Poisson-Gamma prior on Z and symmetric Dirichlet$(\beta_1, \ldots, \beta_V)$ priors on the columns of Φ. The sampling steps for Φ and Z are described next.

C.2.1 Sampling Φ

First note that the elements of the count matrix are modeled as $d_{vn} = \text{Poi}\left(\sum_{k=1}^{K} \phi_{vk} z_{kn}\right)$, which can be equivalently written as $d_{vn} = \sum_{k=1}^{K} d_{vkn}, \quad d_{vkn} = \text{Poi}(\phi_{vk} z_{kn})$. Standard manipulations then allow us to sample the d_{vkn}'s from Mult$(d_{vn}; p_{v1n}, \ldots, p_{vKn})$ where $p_{vkn} = \phi_{vk} z_{kn} / \sum_{k}^{K} \phi_{vk} z_{kn}$.

Now we have $\phi_k \sim \text{Dirichlet}(\beta_1, \ldots, \beta_V)$. Using standard relationships between Poisson and multinomial distributions, we can derive the posterior distribution of the ϕ_k's as Dirichlet$(\beta_1 + d_{1k}, \ldots, \beta_V + d_{Vk})$, where $d_{vk} = \sum_{n=1}^{N} d_{vkn}$.

C.2.2 Sampling Z

In our algorithm we sample each z_{nk} conditioned on all the other variables in the model; therefore the conditional posterior distribution can be written as

$$
\begin{aligned}
& p(z_{nk} | D, \Phi, Z_{n,-k}, \mathbf{d}, \alpha, c, \gamma) \\
& = p(D | Z_n, \Phi) p(z_{nk} | \mathbf{d}, \alpha, c, Z_{n,-k}) \\
& = \prod_{v=1}^{V} \text{Poi}\left(d_{vn} | \sum_{k=1}^{K} \phi_{vk} z_{kn}\right) \frac{p(Z_n | \mathbf{d}, \alpha, c)}{p(Z_{n,-k} | \mathbf{d}, \alpha, c)}.
\end{aligned}
$$

The distributions in both the numerator and denominator of the second factor can be sampled from using the Monte Carlo techniques described above, by integrating out the stick-breaking weights.

C.3 Sampling hyperparameters

As mentioned above, we put a noninformative Gamma prior on γ and improper (1) priors on α and c. The posterior sampling steps are described below:

C.3.1 Sampling γ

Given the round indicators $\mathbf{d} = \{d_k\}$, we can recover the round-specific atom counts as described above. Then the conjugacy between the Gamma prior on γ and the Poisson distribution of C_i gives us a closed form posterior distribution for γ: $p(\gamma | \mathbf{d}, Z, \alpha, c) = \text{Gamma}(\gamma | a + \sum_{i=1}^{K} C_i, b + d_K)$.

C.3.2 Sampling α

The conditional posterior distribution of α may be written as:

$$
p(\alpha | Z, \mathbf{d}, c) \propto p(\alpha) \prod_{n=1}^{N} \prod_{k=1}^{K} p(Z | \mathbf{d}, \alpha, c).
$$

We calculate the posterior distribution of Z using Monte Carlo techniques as described above. Then we discretize the search space for α around its current values as $(\alpha_{cur} + t\Delta\alpha)_{t=L}^{U}$, where the lower and upper bounds L and U are chosen so that the unnormalized posterior falls below 10^{-2}. The search space is also clipped below at 0. α is then drawn from a multinomial distribution on the search values after normalization.

C.3.3 Sampling c

We sample c in exactly the same way as α. We first write the conditional posterior as

$$
p(c | Z, \mathbf{d}, \alpha) \propto p(c) \prod_{n=1}^{N} \prod_{k=1}^{K} p(Z | \mathbf{d}, \alpha, c).
$$

The search space $(c > 0)$ is then discretized using appropriate upper and lower bounds as above, and Z is sampled using Monte Carlo techniques. c is then drawn from a multinomial distribution on the search values after normalization.

References

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*.