Supplementary material for "Scalable Nonparametric Bayesian Multilevel Clustering"

Viet Huynh [†] **Dinh Phung** [†] **Svetha Venkatesh** [†]
[†] Center for Pattern Recognition and Data Analytics (PRaDA)
Deakin University, Australia

XuanLong Nguyen
Department of Statistics
University of Michigan, Ann Arbor, USA

Matt Hoffman
Adobe Research
Adobe Systems, Inc.

Hung Hai Bui
Adobe Research
Adobe Systems, Inc.

Abstract

This document presents supplementary material to complement the manuscript entitled "*Scalable Nonparametric Bayesian Multilevel Clustering*". The two first sections provide details on generalized Dirichlet distributions and some properties of exponential family. The following section presents derivations for variational and stochastic variational Bayes updates for the models.

1 Generalized Dirichlet distribution

Generalized Dirichlet distribution were originally introduce by (Connor & Mosiman, 1969)) and later developed with Bayesian analysis by (Wong, 1998). In this section, we introduce formal definition for this distribution and conjugate property which is useful for our derivation in the following sections.

Definition 1. (Original definition (Connor & Mosiman, 1969)) The generalized Dirichlet distribution (GD) is a generalization of the Dirichlet distribution with a more general covariance structure and almost twice the number of parameters. The density probability function with random vector $\theta = (\theta_1, \ldots, \theta_K)$ and $\theta_K = 1 - \sum_{i=1}^{K-1} \theta_i$

$$p\left(\theta \mid a_1, \ldots, a_{K-1}, b_1, \ldots, b_{K-1}\right) = \frac{\theta_k^{b_{K-1}-1} \prod_{i=1}^{K-1} \left[\theta_i^{a_i-1} \left(\sum_{j=i}^{K} \theta_j\right)^{b_{i-1}-(a_i+b_i)}\right]}{\prod_{i=1}^{K-1} \mathrm{B}\left(a_i, b_i\right)}$$

The above distribution can be constructed as follows. If z_i's are i.i.d. random variables from Beta distributions, i.e. $z_i \sim \mathrm{Beta}\left(a_i, b_i\right)$ and $z_1 = x_1, \ldots, z_i = \theta_i / \left(1 - \sum_{j=1}^{i-1} \theta_j\right)$. We denote $\theta \sim \mathrm{GD}\left(a_1, \ldots, a_{k-1}, b_1, \ldots, b_{k-1}\right)$.

Lemma 2. *(Conjugate prior) Suppose $x \sim GD\left(a_1, \ldots, a_{k-1}, b_1, \ldots, b_{k-1}\right)$ and $y \mid x \sim Mult(x)$ then the posterior distribution is*

$$x \mid y \sim GD\left(\tilde{a}_1, \ldots, \tilde{a}_{k-1}, \tilde{b}_1, \ldots, \tilde{b}_{k-1}\right)$$

where $\tilde{a}_i = a_i + y_i$ and $\tilde{b}_i = b_i + \sum_{j=i+1}^{k} y_i$.

Proof. See Wong (1998, Lemma 1). □

2 Some properties of Exponential family

Proposition 3. *(Expectation of log likelihood) Let $p\left(\theta \mid \eta\right)$ be a distribution which we call likelihood function and $q\left(\theta \mid \lambda\right)$ be a variational distribution used to approximate $p\left(\theta \mid \eta\right)$. Both distributions are supposed belong to the same exponential family form (but different (hyper)parameters), i.e.*

$$p\left(\theta \mid \eta\right) \propto \exp\left(\langle \eta, T\left(\theta\right)\rangle - B\left(\eta\right)\right) \qquad q\left(\theta \mid \lambda\right) \propto \exp\left(\langle \lambda, T\left(\theta\right)\rangle - B\left(\lambda\right)\right)$$

Then

$$\mathbb{E}_{q(\theta|\lambda)}\left[\ln p\left(\theta \mid \eta\right)\right] = \left\langle \eta, \frac{\partial B\left(\lambda\right)}{\partial \lambda} \right\rangle - B\left(\eta\right)$$

Proof. We have

$$\mathbb{E}_{q(\theta|\lambda)}\left[\ln p\left(\theta \mid \eta\right)\right] = \mathbb{E}_{q(\theta|\lambda)}\left[\langle \eta, T\left(\theta\right)\rangle - B\left(\eta\right)\right]$$

However, recall the fact that $\mathbb{E}\left[T\left(\theta\right)\right]$ is the derivative of log partition function of $q\left(\theta \mid \lambda\right)$, i.e. $\mathbb{E}\left[T\left(\theta\right)\right] = \frac{\partial B(\lambda)}{\partial \lambda}$, we have

$$\mathbb{E}_{q(\theta|\lambda)}\left[\ln q\left(\theta \mid \lambda\right)\right] = \left\langle \eta, \frac{\partial B\left(\lambda\right)}{\partial \lambda} \right\rangle - B\left(\eta\right)$$

□

Proposition 4. *(Partial derivative of entropy) Let $q\left(\theta \mid \lambda\right)$ be a distribution belong to exponential family , i.e. $q\left(\theta \mid \lambda\right) \propto \exp\left(\langle \lambda, T\left(\theta\right)\rangle - B\left(\lambda\right)\right)$, then*

$$\frac{\partial}{\partial \lambda} \mathbb{E}_{q(\theta|\lambda)}\left[\ln q\left(\theta \mid \lambda\right)\right] = \frac{\partial^2 B\left(\lambda\right)}{\partial \lambda \partial \lambda^\top} . \lambda$$

Proof. We have

$$\frac{\partial}{\partial\lambda}\mathbb{E}_{q(\theta|\lambda)}\left[\ln q\left(\theta \mid \lambda\right)\right] = \frac{\partial}{\partial\lambda}\mathbb{E}_{q(\theta|\lambda)}\left[\langle\lambda, T\left(\theta\right)\rangle - B\left(\lambda\right)\right]$$

However, recall the fact that $\mathbb{E}\left[T\left(\theta\right)\right]$ is the derivative of log partition function of $q\left(\theta \mid \lambda\right)$, i.e. $\mathbb{E}\left[T\left(\theta\right)\right] = \frac{\partial B(\lambda)}{\partial\lambda}$, we have

$$\begin{aligned}\frac{\partial}{\partial\lambda}\mathbb{E}_{q(\theta|\lambda)}\left[\ln q\left(\theta \mid \lambda\right)\right] &= \frac{\partial}{\partial\lambda}\left[\lambda^{\mathsf{T}}\frac{\partial B\left(\lambda\right)}{\partial\lambda} - B\left(\lambda\right)\right] \\ &= \frac{\partial^2 B\left(\lambda\right)}{\partial\lambda\partial\lambda^{\top}}.\lambda + \frac{\partial B\left(\lambda\right)}{\partial\lambda} - \frac{\partial B\left(\lambda\right)}{\partial\lambda} \\ &= \frac{\partial^2 B\left(\lambda\right)}{\partial\lambda\partial\lambda^{\top}}.\lambda\end{aligned}$$

□

Proposition 5. *(Partial derivative of log likelihood) Let $p\left(\theta \mid \eta\right)$ be a distribution which we call likelihood function and $q\left(\theta \mid \lambda\right)$ be a variational distribution used to approximate $p\left(\theta \mid \eta\right)$. Both distributions are supposed belong to the same exponential family form (but different (hyper)parameters), i.e.*

$$p\left(\theta \mid \eta\right) \propto \exp\left(\langle\eta, T\left(\theta\right)\rangle - B\left(\eta\right)\right) \qquad q\left(\theta \mid \lambda\right) \propto \exp\left(\langle\lambda, T\left(\theta\right)\rangle - B\left(\lambda\right)\right)$$

Then

$$\frac{\partial}{\partial\lambda}\mathbb{E}_{q(\theta|\lambda)}\left[\ln p\left(\theta \mid \eta\right)\right] = \frac{\partial^2 B\left(\lambda\right)}{\partial\lambda\partial\lambda^{\top}}.\eta$$

Proof. We have

$$\frac{\partial}{\partial\lambda}\mathbb{E}_{q(\theta|\lambda)}\left[\ln p\left(\theta \mid \eta\right)\right] = \frac{\partial}{\partial\lambda}\mathbb{E}_{q(\theta|\lambda)}\left[\langle\eta, T\left(\theta\right)\rangle - B\left(\eta\right)\right]$$

However, recall the fact that $\mathbb{E}\left[T\left(\theta\right)\right]$ is the derivative of log partition function of $q\left(\theta \mid \lambda\right)$, i.e. $\mathbb{E}\left[T\left(\theta\right)\right] = \frac{\partial B(\lambda)}{\partial\lambda}$, we have

$$\begin{aligned}\frac{\partial}{\partial\lambda}\mathbb{E}_{q(\theta|\lambda)}\left[\ln q\left(\theta \mid \lambda\right)\right] &= \frac{\partial}{\partial\lambda}\left[\eta^{\mathsf{T}}\frac{\partial B\left(\lambda\right)}{\partial\lambda} - B\left(\eta\right)\right] \\ &= \frac{\partial^2 B\left(\lambda\right)}{\partial\lambda\partial\lambda^{\top}}.\eta\end{aligned}$$

□

3 Variational for MC2

The objective of inference problem with the model is to estimate the posterior distribution $p\left(\Theta \mid x, w\right)$ where Θ is the collection of parameter variable of the model, $\Theta \triangleq \{\beta, \epsilon, \tau, c, z, t, \psi, \phi\}$. However, since this posterior is intractable, in variational Bayes inference, this will be approximated with tractable distribution called variational distribution, $q\left(\Theta\right)$. In order to ensure that $q\left(\Theta\right)$, one usually use mean-field assumption which assumes all variational variables in Θ independent. However, because of the nature of the model, two group of variables z_i and $t_{j1}, \ldots, t_{jn_j}$ can not be totally factorized. We will maintain the joint distribution of these variables in variational inference.

The variational distribution is

$$q\left(\Theta\right) = q\left(\beta\right) q\left(\epsilon\right) q\left(\tau\right) q\left(c\right) q\left(z, t\right) q\left(\psi\right) q\left(\phi\right)$$

where (in truncation setting with K level in β and T level in τ_k and M level in ϵ). It is noticed that $q(z,t)$ is joint distribution in which t_{ji} conditional dependent on z_i.

$$q(\beta) = \mathrm{GD}\left(\beta \mid \lambda^{\beta}\right) \qquad q(\epsilon) = \mathrm{GD}\left(\epsilon \mid \lambda^{\epsilon}\right)$$

$$q(\tau) = \prod_{k=1}^{K} \mathrm{GD}\left(\tau_k \mid \lambda_k^{\tau}\right) \qquad q(c) = \prod_{k=1}^{K}\prod_{t=1}^{T} \mathrm{Mult}\left(c_{kt} \mid \mu_{kt}^{c}\right)$$

$$q(\psi) = \prod_{m=1}^{M} q\left(\psi_m \mid \lambda_m^{\psi}\right) \qquad q(\phi) = \prod_{k=1}^{K} q\left(\phi_k \mid \lambda_k^{\phi}\right)$$

$$q(z,t) = \prod_{j=1}^{J}\left[q(z_j)\prod_{i=1}^{n_j} q(t_{ji} \mid z_j)\right] = \prod_{j=1}^{J}\left[\mathrm{Mult}\left(z_j \mid \mu_j^{z}\right)\prod_{i=1}^{n_j}\mathrm{Mult}\left(t_{ji} \mid \mu_{jiz_j}^{t}\right)\right]$$

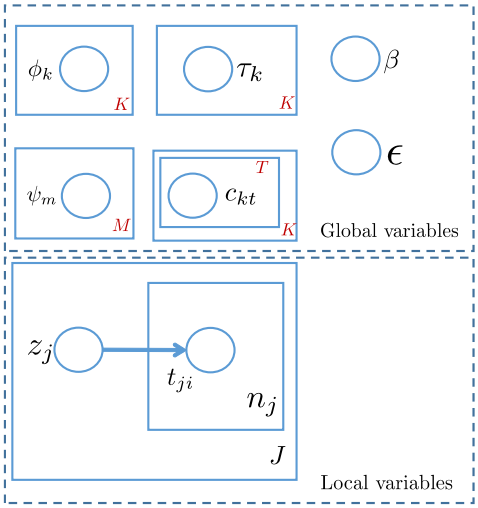

Figure 1: Variational distribution dependency.

The variational distributions is depicted in Fig. 1

- $\lambda^{\beta} = \left(\lambda_{11}^{\beta}, \ldots, \lambda_{1(K-1)}^{\beta}, \lambda_{21}^{\beta}, \ldots, \lambda_{2(K-1)}^{\beta}\right)$ is a $2K-2$ dimension vector;
- $\lambda^{\epsilon} = \left(\lambda_{11}^{\epsilon}, \ldots, \lambda_{1(M-1)}^{\epsilon}, \lambda_{21}^{\epsilon}, \ldots, \lambda_{2(M-1)}^{\epsilon}\right)$ is a $2M-2$ dimension vector;
- $\lambda^{\tau} = \left(\lambda_{11}^{\tau}, \ldots, \lambda_{1(T-1)}^{\tau}, \lambda_{21}^{\tau}, \ldots, \lambda_{2(T-1)}^{\tau}\right)$ is a $2T-2$ dimension vector;
- μ_j^z is a $K-$dimension vector; μ_{kt}^c is a $M-$dimension vector; μ_{jik}^t is a $T-$dimension vector.

3.1 Stick-breaking variable updates

Now we compute the updates for stick-breaking variables

Update equations for β, we have

$$q(\beta) \propto \exp\left(\mathbb{E}\left[\ln p(x,w,\Theta)\right]\right)$$

where

$$\begin{aligned}\mathbb{E}\left[\ln p(x,w,\Theta)\right] &= \ln p(\beta) + \mathbb{E}\left[\ln p(z_{1:J} \mid \beta)\right] + \text{const} \\ &= \ln p(\beta) + \sum_{k=1}^{K}\sum_{j=1}^{J} \mu_{jk}^{z} \ln \beta_k + \text{const}\end{aligned}$$

Hence,

$$q(\beta) \propto p(\beta) \exp\left(\sum_{k=1}^{K}\sum_{j=1}^{J} \mu_{jk}^{z} \ln \beta_k\right)$$

Since $p(\beta \mid \eta)$ has the form of generalized Dirichlet distribution with parameters ($K-2$ dimension)$[1,\ldots,1,\gamma,\ldots,\gamma]^{\mathsf{T}}$. Therefore, using result from Lemma 2 with the prior $p(\beta \mid \gamma)$ and observations $\left[\sum_{j=1}^{J}\mu_{j1}^{z},\ldots,\sum_{j=1}^{J}\mu_{jK}^{z}\right]^{\mathsf{T}}$, we obtain the updated distribution $q(\beta)$ with new hyperparameter λ^{β} where

$$\lambda_{k1}^{\beta} = 1 + \sum_{j=1}^{J}\mu_{jk}^{z} \qquad \lambda_{k2}^{\beta} = \eta + \sum_{j=1}^{J}\sum_{t=k+1}^{K}\mu_{jt}^{z}$$

Update equations for ϵ, The updates for $q(\epsilon)$ can be similarly computed as follows

$$\begin{aligned} q(\epsilon) &\propto \exp\left(\mathbb{E}\left[\ln p(x,w,\Theta)\right]\right) \\ &\propto \exp\left(\ln p(\epsilon) + \mathbb{E}\left[\ln p(c \mid \epsilon)\right]\right) \\ &\propto \exp\left(\ln p(\epsilon) + \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{t=1}^{T}\varepsilon_{ktm}\ln\epsilon_m\right) \\ &= p(\epsilon)\exp\left(\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{t=1}^{T}\varepsilon_{ktm}\ln\epsilon_m\right) \end{aligned}$$

Hence, we obtain the updates

$$\lambda_{m1}^{\epsilon} = 1 + \sum_{k=1}^{K}\sum_{t=1}^{T}\mu_{ktm}^{c} \qquad \lambda_{m2}^{\epsilon} = \gamma + \sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{l=m+1}^{M}\mu_{ktl}^{c}$$

Update equations for τ_k, we have

$$\begin{aligned} \mathbb{E}\left[\ln p(x,w,\Theta)\right] &= \mathbb{E}\left[\ln p(\tau_k) + \sum_{j=1}^{J}\left(\ln p(t_{j\cdot}, z_j = k \mid \tau, \beta)\right)\right] + \text{const} \\ &= \ln p(\tau_k) + \sum_{l=1}^{l}\left(\sum_{j=1}^{J}\mu_{jk}^{z}\sum_{i=1}^{n_j}\mu_{jikl}^{t}\right)\ln\tau_{kl} + \text{const} \end{aligned}$$

Hence,

$$q(\tau_{k,1:T}) \propto p(\tau_{k,1:T})\exp\left(\sum_{t=1}^{T}\left(\sum_{j=1}^{J}\mu_{jk}^{z}\sum_{i=1}^{n_j}\mu_{jikl}^{t}\right)\ln\tau_{kt}\right)$$

As a consequence, we have

$$\lambda_{kl1}^{\tau} = 1 + \sum_{j=1}^{J}\mu_{jk}^{z}\sum_{i=1}^{n_j}\mu_{jikl}^{t} \qquad \lambda_{kl2}^{\tau} = \upsilon + \sum_{j=1}^{J}\mu_{jk}^{z}\sum_{i=1}^{n_j}\sum_{r=l+1}^{T}\mu_{jikr}^{t}$$

3.2 Content and context atom updates

Update equations for context atomsϕ_k, using the standard update for VB, we have

$$
\begin{aligned}
q\left(\phi_k\right) &\propto p\left(\phi_k\right) \exp\left(\mathbb{E}\left[\ln p\left(x \mid z, \phi\right)\right]\right) \\
&= p\left(\phi_k\right) \exp\left(\sum_{j=1}^{J} \mu_{jk}^{z} \ln p\left(x_j \mid \phi_k\right)\right)
\end{aligned}
$$

using conjugacy, we obtain

$$
\lambda_k^{\phi} = \lambda_*^{\phi} + \sum_{j=1}^{J} \mu_{jk}^{z} \left[T\left(x_j\right); 1\right]
$$

Update equations for content atoms ψ_m, similarly, we compute $q\left(\psi_m\right)$ as follows

$$
\begin{aligned}
q\left(\psi_m\right) &\propto p\left(\psi_m\right) \exp\left(\mathbb{E}\left[\ln p\left(w \mid c, z, t, \psi\right)\right]\right) \\
&= p\left(\psi_m\right) \exp\left(\sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(\sum_{k=1}^{K} \vartheta_{jk} \sum_{t=1}^{T} \varepsilon_{ktm} \kappa_{jikt}\right) \ln p\left(w_{ji} \mid \psi_m\right)\right)
\end{aligned}
$$

Therefore

$$
\lambda_m^{\psi} = \lambda_*^{\psi} + \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(\sum_{k=1}^{K} \mu_{jk}^{z} \sum_{l=1}^{T} \mu_{ktm}^{c} \mu_{jikl}^{t}\right) \left[T\left(w_{ji}\right); 1\right]
$$

3.3 Indicator variable updates

Now we compute variational distribution for indicators variables, $q\left(z, t\right)$ and $q\left(c\right)$. The join distribution $q\left(z_j, t_{j.}\right)$ are factorized as follows $q\left(z_j, t_{j.}\right) = q\left(z_j\right) \prod_{i=1}^{n_j} q\left(t_{ji} \mid z_j\right)$ and compute

$$
\begin{aligned}
\mu_{jikl}^{t} &= q\left(t_{ji} = l \mid z_j = k\right) \propto q\left(t_{ji} = l, z_j = k\right) \\
&\propto \exp\left(\mathbb{E}\left[\ln p\left(w_{ji} \mid z_j = k, t_{ji} = l, c, \psi\right)\right] + \mathbb{E}\left[\ln p\left(t_{ji} = l \mid \tau_k\right)\right]\right). \\
&\propto \exp\left(\sum_{m=1}^{M} \mu_{klm}^{c} \mathbb{E}\left[\ln p\left(w_{ji} \mid \psi_m\right)\right] + \mathbb{E}\left[\ln \tau_{kl}\right]\right)
\end{aligned}
$$

This means

$$
\mu_{jikt}^{t} \propto \exp\left(\sum_{m=1}^{M} \mu_{klm}^{c} \mathbb{E}\left[\ln p\left(w_{ji} \mid \psi_m\right)\right] + \mathbb{E}\left[\ln \tau_{kl}\right]\right) \triangleq \tilde{\mu}_{jikl}^{t}
$$

In the above equation, $\tilde{\mu}_{jikl}^{t}$ is the unnormalized value for μ_{jikt}^{t}. This term will be used to compute μ_j^{z} as follows

$$
\begin{aligned}
\mu_{jk}^{z} &= \sum_{t_{j1}, \ldots, t_{jn_j}} q\left(z_j = k, t_{j.}\right) \\
&\propto \sum_{t_{j1}, \ldots, t_{jn_j}} \exp\left(\mathbb{E}\left[p\left(x_j \mid \phi_k\right)\right] + \mathbb{E}\left[\ln \beta_k\right] + \sum_{i=1}^{n_j} \left\{\mathbb{E}\left[\ln p\left(w_{ji} \mid z_j = k, t_{ji}, c, \psi\right)\right] + \mathbb{E}\left[\ln p\left(t_{ji} \mid \tau_k\right)\right]\right\}\right) \\
&= \exp\left(\mathbb{E}\left[p\left(x_j \mid \phi_k\right)\right] + \mathbb{E}\left[\ln \beta_k\right]\right) \left(\sum_{t_{j1}, \ldots, t_{jn_j}} \prod_{i} \exp\left(\left\{\mathbb{E}\left[\ln p\left(w_{ji} \mid z_j = k, t_{ji}, c, \psi\right)\right] + \mathbb{E}\left[\ln p\left(t_{ji} \mid \tau_k\right)\right]\right\}\right)\right) \\
&= \exp\left(\mathbb{E}\left[p\left(x_j \mid \phi_k\right)\right] + \mathbb{E}\left[\ln \beta_k\right]\right) \prod_{i} \left(\sum_{l=1}^{T} \tilde{\mu}_{jikl}^{t}\right) \\
&= \exp\left(\mathbb{E}\left[p\left(x_j \mid \phi_k\right)\right] + \mathbb{E}\left[\ln \beta_k\right] + \sum_{i} \ln\left(\sum_{l=1}^{T} \tilde{\mu}_{jikl}^{t}\right)\right)
\end{aligned}
$$

Therefore, we obtain the update equation for μ_j^z as follows

$$\mu_{jk}^z \propto \exp\left(\mathbb{E}\left[\ln p\left(x_j \mid \phi_k\right)\right] + \mathbb{E}\left[\ln \beta_k\right] + \sum_i \ln\left(\sum_{l=1}^{T} \tilde{\mu}_{jikl}^t\right)\right)$$

Finally, we need to compute update for μ_{kl}^c, we have

$$\begin{aligned}
\mu_{klm}^c &\propto \exp\left(\mathbb{E}\left[\ln p\left(x, w, \Theta\right)\right]\right) \\
&\propto \exp\left(\mathbb{E}\left[\ln p\left(w \mid c_{kl} = m, z, t, \psi\right)\right] + \mathbb{E}\left[\ln \epsilon_m\right]\right) \\
&= \exp\left(\sum_{j=1}^{J} \mu_{jk}^z \sum_{i=1}^{n_j} \mu_{jikl}^t \mathbb{E}\left[\ln p\left(x_{ji} \mid \psi_m\right)\right] + \mathbb{E}\left[\ln \epsilon_m\right]\right)
\end{aligned}$$

which is

$$\mu_{klm}^c \propto \exp\left(\sum_{j=1}^{J} \mu_{jk}^z \sum_{i=1}^{n_j} \mu_{jikl}^t \mathbb{E}\left[\ln p\left(x_{ji} \mid \psi_m\right)\right] + \mathbb{E}\left[\ln \epsilon_m\right]\right)$$

4 Stochastic Variational for MC2

The Evidence Lower BOund (ELBO) function for the model

$$\mathcal{F}\left(x, w, \Theta\right) \triangleq \mathbb{E}\left\{\ln p\left(x, w, \Theta\right)\right\} - \mathbb{E}\left\{\ln q\left(\Theta\right)\right\}$$

can be represented as $\mathcal{F} = \sum_{j=1}^{J} \mathcal{F}_j = \mathbb{E}\left[J\mathcal{F}_j\right]$ where J is the number of observation groups; $\mathcal{F}_j$ is lower bound function only related to document j-th and defined as

$$\begin{aligned}
\mathcal{L}_j = &\mathbb{E}\left[\ln p\left(x_j \mid z_j, \phi\right)\right] + \mathbb{E}\left[\ln p\left(w_{j\cdot} \mid z_j, t_{j\cdot}, c, \psi\right)\right] + \mathbb{E}\left[\ln p\left(z_j \mid \beta\right)\right] + \mathbb{E}\left[\ln p\left(t_{j\cdot} \mid z_j, \tau\right)\right] \\
&- \mathbb{E}\left[\ln q\left(z_j\right)\right] - \mathbb{E}\left[\ln q\left(t_{j\cdot} \mid z_j\right)\right] \\
&+ \frac{1}{J}\left(\mathbb{E}\left[\ln p\left(c \mid \epsilon\right)\right] + \mathbb{E}\left[\ln p\left(\beta\right)\right] + \mathbb{E}\left[\ln p\left(\tau\right)\right] + \mathbb{E}\left[\ln p\left(\epsilon\right)\right]\right) \\
&+ \frac{1}{J}\left(\mathbb{E}\left[\ln p\left(\psi \mid \lambda_*^{\psi}\right)\right] + \mathbb{E}\left[\ln p\left(\phi \mid \lambda_*^{\phi}\right)\right]\right) \\
&- \frac{1}{J}\left(\mathbb{E}\left[\ln q\left(c\right)\right] + \mathbb{E}\left[\ln q\left(\beta\right)\right] + \mathbb{E}\left[\ln q\left(\tau\right)\right] + \mathbb{E}\left[\ln q\left(\epsilon\right)\right]\right) \\
&- \frac{1}{J}\left(\mathbb{E}\left[\ln q\left(\psi\right)\right] + \mathbb{E}\left[\ln q\left(\phi\right)\right]\right)
\end{aligned}$$

4.1 Stochastic updates for stick-breaking variables

Update equations for β, using results from Proposition 4 and 5 with generalized Dirichlet distributions, we get

$$\frac{\partial}{\partial \lambda_k^{\beta}} \mathbb{E}\left[\ln q\left(\beta\right)\right] = \frac{\partial^2 B\left(\lambda_k^{\beta}\right)}{\partial \lambda_k^{\beta} \partial \left(\lambda_k^{\beta}\right)^{\top}} \left[\lambda_{k1}^{\beta} - 1; \lambda_{k2}^{\beta} - 1\right]$$

and

$$\frac{\partial}{\partial \lambda_k^{\beta}} \mathbb{E}\left[\ln p\left(\beta\right)\right] = \frac{\partial^2 B\left(\lambda_k^{\beta}\right)}{\partial \lambda_k^{\beta} \partial \left(\lambda_k^{\beta}\right)^{\top}} \left[0; \gamma - 1\right]$$

Therefore, using natural gradient, we have

$$\frac{\partial^{(\mathsf{ng})}\mathcal{L}_j}{\partial\lambda_{k1}^{\beta}} = \frac{-\lambda_{k1}^{\beta}+1+J\mu_{jk}^{z}}{J} \qquad \frac{\partial^{(\mathsf{ng})}\mathcal{L}_j}{\partial\lambda_{k2}^{\beta}} = \frac{-\lambda_{k2}^{\beta}+\eta+J\sum_{l=k+1}^{K}\mu_{jl}^{z}}{J}$$

Update equations for ϵ, we have

$$\frac{\partial\mathcal{L}_j}{\partial\lambda_m^{\epsilon}} = \frac{\partial}{\partial\lambda_m^{\epsilon}}\left(\frac{1}{J}\mathbb{E}\left[\ln p\left(c\mid\epsilon\right)\right]+\frac{1}{J}\mathbb{E}\left[\ln p\left(\epsilon\right)\right]-\frac{1}{J}\mathbb{E}\left[\ln q\left(\epsilon\right)\right]\right)$$

where

$$\begin{aligned}\frac{\partial}{\partial\lambda_m^{\epsilon}}\mathbb{E}\left[\ln p\left(c\mid\epsilon\right)\right] &= \frac{\partial}{\partial\lambda_m^{\epsilon}}\mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{r=1}^{M}\mu_{ktr}^{c}\ln p\left(c_{lt}=r\mid\epsilon\right)\right]\\ &= \frac{\partial}{\partial\lambda_m^{\epsilon}}\mathbb{E}\left[\left\langle\left[\ln\epsilon_m,\ln\left(1-\epsilon_m\right)\right],\left[\sum_{k=1}^{K}\sum_{t=1}^{T}\mu_{ktm}^{c},\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{r=m+1}^{M}\mu_{ktr}^{c}\right]\right\rangle\right]\\ &= \frac{\partial^2 B\left(\lambda_m^{\epsilon}\right)}{\partial\lambda_m^{\epsilon}\partial\left(\lambda_m^{\epsilon}\right)^{\top}}\left[\sum_{k=1}^{K}\sum_{t=1}^{T}\mu_{ktm}^{c},\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{r=m+1}^{M}\mu_{ktr}^{c}\right]\end{aligned}$$

Similarly

$$\frac{\partial}{\partial\lambda_m^{\epsilon}}\mathbb{E}\left[\ln q\left(\epsilon\right)\right] = \frac{\partial^2 B\left(\lambda_m^{\epsilon}\right)}{\partial\lambda_m^{\epsilon}\partial\left(\lambda_m^{\epsilon}\right)^{\top}}\left[\lambda_{m1}^{\epsilon}-1,;\lambda_{m2}^{\epsilon}-1\right]$$

and

$$\frac{\partial}{\partial\lambda_m^{\epsilon}}\mathbb{E}\left[\ln p\left(u\right)\right] = \left\langle\frac{\partial^2 B\left(\varphi_m\right)}{\partial\varphi_m\partial\varphi_m^{\top}},\left[0;\gamma-1\right]\right\rangle$$

Therefore, using natural gradient, we have

$$\frac{\partial^{(\mathsf{ng})}\mathcal{L}_j}{\partial\lambda_{m1}^{\epsilon}} = \frac{-\lambda_{m1}^{\epsilon}+1+\sum_{k=1}^{K}\sum_{t=1}^{T}\mu_{ktm}^{c}}{J} \qquad \frac{\partial^{(\mathsf{ng})}\mathcal{L}_j}{\partial\lambda_{m2}^{\epsilon}} = \frac{-\lambda_{m2}^{\epsilon}+\gamma+\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{r=m+1}^{M}\mu_{ktr}^{c}}{J}$$

Update equations for τ_k, we have

$$\frac{\partial\mathcal{L}_j}{\partial\lambda_{kt}^{\tau}} = \frac{\partial}{\partial\lambda_{kt}^{\tau}}\left(\mathbb{E}\left[\ln p\left(t_{j\cdot}\mid\tau,z_j\right)\right]+\frac{1}{J}\mathbb{E}\left[\ln p\left(\tau\right)\right]-\frac{1}{J}\mathbb{E}\left[\ln q\left(\tau\right)\right]\right)$$

where

$$\begin{aligned}\frac{\partial}{\partial\lambda_{kl}^{\tau}}\mathbb{E}\left[\ln p\left(t_{j\cdot}\mid\tau,z_j\right)\right] &= \mu_{jk}^{z}\sum_{i=1}^{n_j}\frac{\partial}{\partial\lambda_{kl}^{\tau}}\mathbb{E}\left[\mu_{jikl}^{t}\ln p\left(t_{ji}=l\mid\tau_k,z_j=k\right)\right]\\ &= \mu_{jk}^{z}\frac{\partial}{\partial\lambda_{kl}^{\tau}}\mathbb{E}\left[\sum_{i=1}^{n_j}\mu_{jikl}^{t}\ln\left(\tau_k\right)\right]\\ &= \mu_{jk}^{z}\frac{\partial}{\partial\lambda_{kl}^{\tau}}\mathbb{E}\left[\left\langle\left[\ln\tau_{kl},\ln\left(1-\tau_{kl}\right)\right],\left[\sum_{i=1}^{n_j}\mu_{jikl}^{t},\sum_{i=1}^{n_j}\sum_{r=l+1}^{T}\mu_{jikr}^{t}\right]\right\rangle\right]\\ &= \mu_{jk}^{z}\frac{\partial^2 B\left(\lambda_{kl}^{\tau}\right)}{\partial\lambda_{kl}^{\tau}\partial\left(\lambda_{kl}^{\tau}\right)^{\top}}\left[\sum_{i=1}^{n_j}\mu_{jikl}^{t},\sum_{i=1}^{n_j}\sum_{r=l+1}^{T}\mu_{jikr}^{t}\right]\end{aligned}$$

Similarly

$$\frac{\partial}{\partial \lambda_{kl}^{\tau}} \mathbb{E}\left[\ln q\left(\tau_{kl}\right)\right] = \frac{\partial^2 B\left(\lambda_{kl}^{\tau}\right)}{\partial \lambda_{kl}^{\tau} \partial\left(\lambda_{kl}^{\tau}\right)^{\top}}\left[\lambda_{kl1}^{\tau} - 1; \lambda_{kl2}^{\tau} - 1\right]$$

and

$$\frac{\partial}{\partial \lambda_{kl}^{\tau}} \mathbb{E}\left[\ln p\left(\tau_{kl}\right)\right] = \frac{\partial^2 B\left(\lambda_{kl}^{\tau}\right)}{\partial \lambda_{kl}^{\tau} \partial\left(\lambda_{kl}^{\tau}\right)^{\top}}\left[0; \upsilon - 1\right]$$

Finally, using natural gradient, we have

$$\frac{\partial^{(\mathsf{ng})} \mathcal{L}_j}{\partial \lambda_{kl1}^{\tau}} = \frac{-\lambda_{kl1}^{\tau} + 1 + J\mu_{jk}^{z} \sum_{i=1}^{n_j} \mu_{jikl}^{t}}{J} \qquad \frac{\partial^{(\mathsf{ng})} \mathcal{L}_j}{\partial \lambda_{kl2}^{\tau}} = \frac{-\lambda_{kl2}^{\tau} + \upsilon + J\mu_{jk}^{z} \sum_{i=1}^{n_j} \sum_{r=t+1}^{T} \mu_{jikr}^{t}}{J}$$

4.2 Stochastic updates for content and context atoms

Update equations for content atoms ϕ_k, compute gradient of $\mathcal{L}_j$ with respect to λ_k^{ϕ} (in $q\left(\phi_k \mid \lambda_k^{\phi}\right)$), we have

$$\begin{aligned}
\frac{\partial \mathcal{L}_j}{\partial \lambda_k^{\phi}} &= \frac{\partial}{\partial \lambda_k^{\phi}}\left(\mathbb{E}\left[\ln p\left(x_j \mid z, \phi\right)\right] + \frac{1}{J}\left(\mathbb{E}\left[\ln p\left(\phi \mid \lambda_*^{\phi}\right)\right] - \mathbb{E}\left[\ln q\left(\phi\right)\right]\right)\right) \\
&= \frac{\partial}{\partial \lambda_k^{\phi}}\left(\mathbb{E}\left[\ln p\left(x_j \mid z, \phi\right)\right] + \frac{1}{J}\left(\mathbb{E}\left[\ln p\left(\phi_k \mid \lambda_*^{\phi}\right)\right] - \mathbb{E}\left[\ln q\left(\phi_k \mid \lambda_k^{\phi}\right)\right]\right)\right)
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial}{\partial \lambda_k^{\phi}} \mathbb{E}\left[\ln p\left(x_j \mid z, \phi\right)\right] &= \frac{\partial}{\partial \lambda_k^{\phi}} \sum_{k=1}^{K} \mu_{jk}^{z} \mathbb{E}\left[\left\langle T\left(x_j\right), \phi_k\right\rangle - A\left(\phi_k\right)\right] \\
&= \frac{\partial}{\partial \lambda_k^{\phi}} \mu_{jk}^{z} \left\langle \left[T\left(x_j\right), 1\right], \frac{\partial B\left(\lambda_k^{\phi}\right)}{\partial \lambda_k^{\phi}} \right\rangle \\
&= \mu_{jk}^{z} \left\langle \left[T\left(x_j\right), 1\right], \frac{\partial^2 \lambda_k^{\phi}}{\partial \lambda_k^{\phi} \partial\left(\lambda_k^{\phi}\right)^{\top}} \right\rangle
\end{aligned}$$

here we use the fact that $\mathbb{E}\left[\phi_k, -A\left(\phi_k\right)\right]$ is the derivative of log partition function of $q\left(\phi_k \mid \lambda_k^{\phi}\right)$, i.e. $\frac{\partial B\left(\lambda_k^{\phi}\right)}{\partial \lambda_k^{\phi}}$.

Moreover, using results from Proposition 4 and 5, we get

$$\frac{\partial}{\partial \lambda_k^{\phi}} \mathbb{E}\left[\ln p\left(\phi \mid \lambda_*^{\phi}\right)\right] = \left\langle \lambda_*^{\phi}, \frac{\partial^2 \lambda_k^{\phi}}{\partial \lambda_k^{\phi} \partial\left(\lambda_k^{\phi}\right)^{\top}} \right\rangle$$

and

$$\frac{\partial}{\partial \lambda_k^{\phi}} \mathbb{E}\left[\ln q\left(\phi_k \mid \lambda_k^{\phi}\right)\right] = \left\langle \lambda_k^{\phi}, \frac{\partial^2 \lambda_k^{\phi}}{\partial \lambda_k^{\phi} \partial\left(\lambda_k^{\phi}\right)^{\top}} \right\rangle$$

Hence, using natural gradient, we have

$$\frac{\partial^{(\mathsf{ng})}\mathcal{L}_j}{\partial\lambda_k^{\phi}} = \frac{-\lambda_k^{\phi} + \lambda_*^{\phi} + J\mu_{jk}^{z}\left[T\left(x_j\right), 1\right]}{J}$$

Update equations for content atoms ψ_m, computing gradient of $\mathcal{L}_j$ with respect to λ_m^{ψ} (in $q\left(\psi_m \mid \lambda_m^{\psi}\right)$), we have

$$\begin{aligned}\frac{\partial\mathcal{L}_j}{\partial\lambda_m^{\psi}} &= \frac{\partial}{\partial\lambda_m^{\psi}}\left(\mathbb{E}\left[\ln p\left(w_{j\cdot} \mid z, t, c, \psi\right)\right] + \frac{1}{J}\left(\mathbb{E}\left[\ln p\left(\psi \mid \lambda_*^{\psi}\right)\right] - \mathbb{E}\left[\ln q\left(\psi\right)\right]\right)\right) \\ &= \frac{\partial}{\partial\lambda_m^{\psi}}\left(\mathbb{E}\left[\ln p\left(w_{j\cdot} \mid z, t, c, \psi\right)\right] + \frac{1}{J}\left(\mathbb{E}\left[\ln p\left(\psi_m \mid \lambda_*^{\psi}\right)\right] - \mathbb{E}\left[\ln q\left(\psi_m \mid \lambda_m^{\psi}\right)\right]\right)\right)\end{aligned}$$

where

$$\begin{aligned}\frac{\partial}{\partial\lambda_m^{\psi}}\mathbb{E}\left[\ln p\left(w_{j\cdot} \mid z, t, c, \psi\right)\right] &= \frac{\partial}{\partial\lambda_m^{\psi}}\sum_{i=1}^{n_j}\sum_{k=1}^{K}\mu_{jk}^{z}\sum_{l=1}^{T}\mu_{jikl}^{t}\sum_{m=1}^{M}\mu_{klm}^{c}\mathbb{E}\left[\left\langle T\left\langle w_{ji}\right\rangle, \psi_m\right\rangle - A\left(\psi_m\right)\right] \\ &= \frac{\partial}{\partial\lambda_m^{\psi}}\sum_{k=1}^{K}\mu_{jk}^{z}\sum_{l=1}^{T}\mu_{klm}^{c}\sum_{i=1}^{n_j}\mu_{jikl}^{t}\left\langle\left[T\left\langle w_{ji}\right\rangle; 1\right], \frac{\partial B\left(\lambda_m^{\psi}\right)}{\partial\lambda_m^{\psi}}\right\rangle \\ &= \sum_{k=1}^{K}\mu_{jk}^{z}\sum_{l=1}^{T}\mu_{klm}^{c}\sum_{i=1}^{n_j}\mu_{jikl}^{t}\frac{\partial^2 B\left(\lambda_m^{\psi}\right)}{\partial\lambda_m^{\psi}\partial\left(\lambda_m^{\psi}\right)^{\top}}\left[T\left\langle w_{ji}\right\rangle; 1\right]\end{aligned}$$

Here we use the fact that $\mathbb{E}\left[\psi_m, -A\left(\psi_m\right)\right]$ is the derivative of log partition function of $q\left(\psi_m \mid \lambda_m^{\psi}\right)$, i.e. $\frac{\partial B\left(\lambda_m^{\psi}\right)}{\partial\lambda_m^{\psi}}$. Moreover, using results from Proposition 4 and 5, we get

$$\frac{\partial}{\partial\lambda_m^{\psi}}\mathbb{E}\left[\ln p\left(\psi_m \mid \lambda_*^{\psi}\right)\right] = \frac{\partial^2 B\left(\lambda_m^{\psi}\right)}{\partial\lambda_m^{\psi}\partial\left(\lambda_m^{\psi}\right)^{\top}}\lambda_*^{\psi}$$

and

$$\frac{\partial}{\partial\lambda_m^{\psi}}\mathbb{E}\left[\ln q\left(\psi_m \mid \lambda_m\right)\right] = \frac{\partial^2 B\left(\lambda_m^{\psi}\right)}{\partial\lambda_m^{\psi}\partial\left(\lambda_m^{\psi}\right)^{\top}}\lambda_m^{\psi}$$

Hence, using natural gradient, we have

$$\frac{\partial^{(\mathsf{ng})}\mathcal{L}_j}{\partial\lambda_m^{\psi}} = \frac{-\lambda_m^{\psi} + \lambda_*^{\psi} + J\sum_{k=1}^{K}\mu_{jk}^{z}\sum_{l=1}^{T}\mu_{klm}^{c}\sum_{i=1}^{n_j}\mu_{jikl}^{t}\left[T\left\langle w_{ji}\right\rangle; 1\right]}{J}$$

4.3 Stochastic updates for global indicator variables

Since updating μ_{kt}^{c} involving computation form all data groups $j's$, we will use "lazy update" with stochastic gradient. The gradient of ELBO function over μ_{kt}^{c} is needed to compute. However, the mean-parameterization of $q\left(c_{kt} \mid \mu_{kt}^{c}\right)$ with constraints $\sum_m \mu_{ktm}^{c} = 1$ and $0 \leq \mu_{ktm}^{c} \leq 1$ is difficult to compute gradient. We therefore prefer to work with a minimal natural parameterization in exponential family form as follows

$$q\left(c_{kt} \mid \lambda_{kt}^{c}\right) = \exp\left(\left\langle\lambda_{kt}^{c}, T\left(c_{kt}\right)\right\rangle - B\left(\lambda_{kt}^{c}\right)\right)$$

where $\lambda_{kt}^{c} = \left[\lambda_{kt1}^{c}, \ldots, \lambda_{kt(M-1)}^{c}\right]^{\mathsf{T}}$; $T\left(c\right) = \left[\delta\left(c-1\right), \ldots, \delta\left(c-M+1\right)\right]^{\mathsf{T}}$and $B\left(\lambda_{kt}^{c}\right) = 1 + \sum_{m=1}^{M-1}\exp\left(\lambda_{ktm}^{c}\right)$. The relationship between these groups of parameters is

$$\begin{aligned}\mu_{ktm}^{c} &= \frac{\exp\left(\lambda_{ktm}^{c}\right)}{1 + \sum_{m=1}^{M-1}\exp\left(\lambda_{ktm}^{c}\right)}, m = 1, \ldots, M-1 \\ \text{and } \mu_{kM}^{c} &= \frac{1}{1 + \sum_{m=1}^{M-1}\exp\left(\lambda_{ktm}^{c}\right)}\end{aligned}$$

Now we can compute the gradient over new parameters λ_{kt}^c:

$$\frac{\partial^{(\text{ng})}\mathcal{L}_j}{\partial\lambda_{klm}^c} = \frac{-\lambda_{klm}^c + \mathbb{E}\left[\ln\frac{\epsilon_m}{\epsilon_M}\right]}{J} + (a_{klm} - a_{klM}) \tag{1}$$

where $a_{klm} = \mu_{jk}^z \sum_{i=1}^{n_j} \mu_{jikl}^t \mathbb{E}\left[\ln p\left(w_{ji} \mid \psi_m\right)\right]$, for all m.

References

Connor, R. J., & Mosiman, J. E. 1969. Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, **64**, 194–206.

Wong, T.-T. 1998. Generalized Dirichlet distribution in Bayesian analysis. *Applied Mathematics and Computation*, **97**, 165–181.