Categories and Functional Units: An Infinite Hierarchical Model for Brain Activations

Supplementary Materials

Danial Lashkari **Ramesh Sridharan** **Polina Golland**
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139
{danial, rameshvs, polina}@csail.mit.edu

In the Supplementary Materials, we derive the Gibbs free energy cost function for variational inference, and derive the update rules for inference using our variational approximation.

1 Joint Probability Distribution

Based on the generative model described in the paper, we can write the full joint distribution of all the observed and unobserved variables. For each variable, we use $\omega^{\cdot}$ to denote the natural parameters of the distribution for that variable. For example, the variable e_{jih} is associated with natural parameters $\omega^{e,1}_{jh}$ and $\omega^{e,2}_{jh}$.

1.1 fMRI model

The fMRI response across different sessions within the same subject can be highly variable. Therefore, we model the variables $\boldsymbol{a}$, the activation height, and $\boldsymbol{\lambda}$, the voxel-level noise parameter, individually for each of the sessions in the experiment (due to space constraints in the original paper, we omitted these implementation details there). We use the index m to denote session m, giving a_{jim} and λ_{jim} for $m \in \{1, \ldots, M\}$. We are also provided with binary indicators D_{jmt} for each subject j, which specifies whether time points t belonged to session m. For notational convenience, we let $\lambda_{jit} = \prod_m \lambda_{jim}^{D_{jmt}}$, and similarly let $a_{jit} = \prod_m a_{jim}^{D_{jmt}}$. Given the fMRI model parameters, we can then write the likelihood of the observed data y_{jit}:

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{\lambda}, \boldsymbol{e}) = \prod_{j,i,t} \sqrt{\frac{\lambda_{jit}}{2\pi}} \exp\left\{ -\frac{\lambda_{jit}}{2} \left(y_{jit} - \sum_h e_{jih} F_{ht} - a_{jit} \sum_s G_{jst} x_{jis} \right)^2 \right\} \quad (1)$$

For the nuisance parameters e, we have

$$p(e_{jih}) = \text{Normal}(\mu^e_{jh}, \sigma^e_{jh}) \quad (2)$$

$$\propto \exp\left\{ -\omega^{e,2}_{jh} e^2_{jih} + \omega^{e,1}_{jh} e_{jih} \right\}, \quad (3)$$

where $\omega^{e,2}_{jh} = \frac{1}{2}(\sigma^e_{jh})^{-1}$ and $\omega^{e,1}_{jh} = \frac{1}{2}\mu^e_h(\sigma^e_{jh})^{-1}$.

For the noise precision variables $\boldsymbol{\lambda}$, we have:

$$p(\lambda_{jim}) = \text{Gamma}(\kappa_{jm}, \theta_{jm}) \quad (4)$$

$$\propto \exp\left\{ \omega^{\lambda,1}_{jm} \log(\lambda_{jim}) - \omega^{\lambda,2}_{jm} \lambda_{jim} \right\}, \quad (5)$$

where $\omega_{jm}^{\lambda,1} = \kappa_{jm} - 1$ and $\omega_{jm}^{\lambda,2} = \theta_{jm}$.

The distribution over the activation heights $\boldsymbol{a}$ is given by

$$p(a_{jim}) = \text{Normal}_+(\mu_{jm}^a, \sigma_{jm}^a) \tag{6}$$

$$\propto \exp\left\{-\omega_{jm}^{a,2} a_{jim}^2 + \omega_{jm}^{a,1} a_{jim}\right\}, a_{jim} \geq 0 \tag{7}$$

We have that $\omega_{jm}^{a,2} = \frac{1}{2}(\sigma_{jm}^a)^{-1}$ and $\omega_{a,1}^{jm} = \frac{1}{2}\mu_{jm}^a \left(\sigma_{jm}^a\right)^{-1}$.

Note that the distribution $w \sim \text{Normal}_+(\rho, \eta^{-1})$ is a member of the exponential family of distributions and has the following properties:

$$p(w) = \sqrt{\tfrac{2\lambda}{\pi}}\left[1 + \text{erf}\left(\sqrt{\tfrac{\rho}{2}}\eta\right)\right]^{-1} e^{-\rho(w-\eta)^2/2}, \tag{8}$$

$$E[w] = \eta + \sqrt{\tfrac{2}{\pi\lambda}}\left[1 + \text{erf}\left(\sqrt{\tfrac{\rho}{2}}\eta\right)\right]^{-1} e^{-\rho\eta^2/2}, \tag{9}$$

$$E[w^2] = \eta^2 + \rho^{-1} + \eta\sqrt{\tfrac{2}{\pi\lambda}}\left[1 + \text{erf}\left(\sqrt{\tfrac{\rho}{2}}\eta\right)\right]^{-1} e^{-\rho\eta^2/2} \tag{10}$$

1.2 Nonparametric Hierarchical Co-clustering Model

The voxel activations x_{jis} are binary, with prior probability $\phi_{k,l}$ given according to cluster memberships. We have that $\phi \sim \text{Beta}(\omega^{\phi,1}, \omega^{\phi,2})$, giving us the joint density of $\boldsymbol{x}$ and $\boldsymbol{\phi}$ conditioned on the cluster memberships $\boldsymbol{z}$ and $\boldsymbol{c}$:

$$p(\boldsymbol{x}, \boldsymbol{\phi}|\boldsymbol{z}, \boldsymbol{c}) = \prod_{j,k,l}\left[\frac{\Gamma(\omega^{\phi,1} + \omega^{\phi,2})}{\Gamma(\omega^{\phi,1})\Gamma(\omega^{\phi,2})}\phi_{k,l}^{\omega^{\phi,1} - 1 + \sum_{i,s} x_{jis}\delta(z_{ji},k)\delta(c_s,l)} \cdot \right.$$

$$\left. (1-\phi_{k,l})^{\omega^{\phi,2} - 1 + \sum_{i,s}(1-x_{jis})\delta(z_{ji},k)\delta(c_s,l)}\right]$$

Our model assumes a Dirichlet Process prior over the stimulus category memberships, using the stick-breaking construction for the prior $\boldsymbol{\rho}$ [1]:

$$p(\boldsymbol{c}|\boldsymbol{\rho}) = \prod_s \prod_l \rho_l^{\delta(c_s,l)} \tag{11}$$

$$\rho_l = u_l \prod_{l'=1}^{l-1}(1 - u_{l'}) \tag{12}$$

$$p(\boldsymbol{u}) = \prod_l (\chi - 1)(1 - u_l)^{\chi - 1} \tag{13}$$

We assume a hierarchical Dirichlet process prior over the functional unit memberships, with subject-level weights $\boldsymbol{\beta}$. We will use a collapsed variational inference scheme [2], and therefore marginalize over these weights:

$$p(\boldsymbol{z}|\pi, \alpha) = \int_{\boldsymbol{\beta}} p(\boldsymbol{z}|\boldsymbol{\beta}) p(\boldsymbol{\beta}|\pi, \alpha), \tag{14}$$

$$= \prod_{j=1}^{J}\left[\tfrac{\Gamma(\alpha)}{\Gamma(\alpha + N_j)} \prod_{k=1}^{K} \tfrac{\Gamma(\alpha\pi_k + n_{jk})}{\Gamma(\alpha\pi_k)}\right], \tag{15}$$

where $n_{jk} = \sum_{i=1}^{N_j} \delta(z_{ji}, k)$ and K is the number of non-empty functional units in the configuration. To provide conjugacy with the Dirichlet prior over the group-level functional unit weights $\boldsymbol{\pi}$, we prefer the terms in Equation (15) involving this weight to appear as powers of π_k . However, the current form of the conditional distribution makes the computation of the posterior over π hard. Now, note that for $0 \leq r \leq n$, we have $\sum_{r=0}^{n} \left[{n \atop r}\right] \vartheta^r = \Gamma(\vartheta + n)/\Gamma(\vartheta)$, where $\left[{n \atop r}\right]$ are unsigned Stirling numbers of the first kind [3]. The collapsed variational approach uses this fact and the properties

of the Beta distribution to add an auxiliary variable $\boldsymbol{r} = \{r_{ji}\}$ to the model:

$$p(\boldsymbol{z}, \boldsymbol{r}, \mid \pi, \alpha) \propto \prod_{j=1}^{J} \prod_{k=1}^{K} \left[{n_{jk} \atop r_{jk}}\right] (\alpha\pi_k)^{r_{jk}}, \tag{16}$$

where $r_{jk} \in \{0, 1, \cdots, n_{ji}\}$. If we marginalize the distribution (16) over the auxiliary variable, we obtain the expression in (15).

2 Minimization of the Gibbs Free Energy

Let $\boldsymbol{h} = \{\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{r}, \boldsymbol{\phi}, \pi, \rho\}$ denote the set of all unobserved variables. In the framework of variational inference, we approximate the model posterior on $\boldsymbol{h}$ given the observed data $p(\boldsymbol{h}|\boldsymbol{y})$ by a distribution $q(\boldsymbol{h})$. The approximation is performed through minimization of the Gibbs free energy function

$$\mathcal{F}[q] = E[\log q(\boldsymbol{h})] - E[\log p(\boldsymbol{y}, \boldsymbol{h})].$$

Here, and in the remainder of the paper, $E[\cdot]$ and $V[\cdot]$ indicate expected value and variance with respect to the distribution q. We assume a distribution q of the form:

$$q(\boldsymbol{h}) = q(\boldsymbol{r}|\boldsymbol{z}) \left[\prod_k q(v_k)\right] \left[\prod_l q(u_l)\right] \left[\prod_{k,l} q(\phi_{k,l})\right] \left[\prod_s q(c_s)\right] \cdot \prod_{j,i} \left(q(a_{ji}) q(\lambda_{ji}) q(z_{ji}) \left[\prod_s q(x_{jis})\right] \left[\prod_h q(e_{jih})\right] \right). \tag{17}$$

where we explicitly account for the dependency of the auxiliary variables on the functional unit memberships. Including this structure maintains the quality of approximation despite the introduction of the auxiliary variables [4].

2.1 Auxiliary variables

Assuming that all the other components of the posterior q are constant, we have:

$$\mathcal{F}[q(\boldsymbol{r}, \mid \boldsymbol{z})] = E_{\boldsymbol{z}} \left[\sum_{\boldsymbol{r}} q(\boldsymbol{r}|\boldsymbol{z}) \Big(\log q(\boldsymbol{r}|\boldsymbol{z}) + - \sum_{j,k} \Big\{ \log \left[{n_{jk} \atop r_{jk}}\right] + r_{jk}\, E[\log(\alpha\pi_k)] \Big\} \Big) \right] + \text{const.} \tag{18}$$

From this, we find that the optimal posterior on the auxiliary variables takes the form

$$q^*(\boldsymbol{r}|\boldsymbol{z}) = \prod_j \prod_k q(r_{jk}|\boldsymbol{z}). \tag{19}$$

Under $q*$, we have for the auxiliary variable $\boldsymbol{r}$:

$$q(r_{jk}|\boldsymbol{z}) = \frac{\Gamma(\tilde{\omega}^r_{jk})}{\Gamma(\tilde{\omega}^r_{jk} + n_{jk})} \left[{n_{jk} \atop r_{jk}}\right] (\tilde{\omega}^r_{jk})^{r_{jk}}. \tag{20}$$

This distribution corresponds to the probability mass function for a random variable that describes the number of tables that n_{jk} customers occupy in a Chinese Restaurant Process with parameter $\tilde{\omega}^r_{jk}$ [3]. Here, the optimal value of the parameter $\tilde{\omega}^r_{jk}$ is given by

$$\begin{aligned} \log \tilde{\omega}^r_{jk} &= E[\log(\alpha\pi_k)] \qquad (21) \\ &= \log\alpha + E[\log v_k] + \textstyle\sum_{k'<k} E[\log(1 - v_{k'})]. \end{aligned}$$

Note that as a distribution parameterized by $\log \tilde{\omega}^r_{jk}$, Equation (20) defines a member of the exponential family of distributions. The expected value of the distribution is therefore computed by:

$$\begin{aligned} E[r_{jk}|\boldsymbol{z}] &= \frac{\partial}{\partial \log \tilde{\omega}^r_{jk}} \log \frac{\Gamma(\tilde{\omega}^r_{jk} + n_{jk})}{\Gamma(\tilde{\omega}^r_{jk})} \qquad (22) \\ &= \tilde{\omega}^r_{jk} \Psi(\tilde{\omega}^r_{jk} + n_{jk}) - \tilde{\omega}^r_{jk} \Psi(\tilde{\omega}^r_{jk}), \qquad (23) \end{aligned}$$

where $\Psi(\omega) = \frac{\partial}{\partial \omega} \log \Gamma(\omega)$. This expression is useful when we need the expected values of the auxiliary variable r_{jk} for updating the other components of the distribution:

$$E[r_{jk}] = E_{\boldsymbol{z}}[E_{\boldsymbol{r}}[r_{jk}|\boldsymbol{z}]] = \tilde{\omega}^r_{jk} E_{\boldsymbol{z}}[\Psi(\tilde{\omega}^r_{jk} + n_{jk}) - \Psi(\tilde{\omega}^r_{jk})]. \tag{24}$$

Under $q(\boldsymbol{z})$, each variable n_{jk} is the sum of N_j independent Bernoulli random variables $\delta(z_{ji}, k)$ for $1 \leq i \leq N_j$ with the probability of success $q(z_{ji} = k)$. Therefore, as suggested in [2], we can use the Central Limit Theorem and approximate this term using a Gaussian distribution for $n_{jk} > 0$. Due to the independence of these Bernoulli variables, we have

$$\Pr(n_{jk} > 0) = 1 - \prod_{i=1}^{N_j} q(z_{ji} \neq k), \tag{25}$$

$$E[n_{jk}] = E[n_{jk}|n_{jk} > 0] \Pr(n_{jk} > 0), \tag{26}$$

$$E[n^2_{jk}] = E[n^2_{jk}|n_{jk} > 0] \Pr(n_{jk} > 0), \tag{27}$$

which we can use to easily compute $E^+[n_{jk}] = E[n_{jk}|n_{jk} > 0]$ and $V^+[n_{jk}] = V[n_{jk}|n_{jk} > 0]$. Now, we can calculate $E[r_{jk}]$ using Equation (24) by noting that:

$$E_{\boldsymbol{z}}[\Psi(\tilde{\omega}^r_{jk} + n_{jk}) - \Psi(\tilde{\omega}^r_{jk})] \approx$$
$$\Pr(n_{jk} > 0) \left[\Psi(\tilde{\omega}^r_{jk} + E^+[n_{jk}]) - \Psi(\tilde{\omega}^r_{jk}) + \tfrac{V^+[n_{jk}]}{2} \Psi''(\tilde{\omega}^r_{jk} + E^+[n_{jk}])\right]. \tag{28}$$

Lastly, based on the auxiliary variable r, we find that the optimal posterior over the system weight stick-breaking parameters is given by $v_k \sim \text{Beta}(\tilde{\omega}^{v,1}_k, \tilde{\omega}^{v,2}_k)$, with parameters as follows:

$$\tilde{\omega}^{v,1}_k = 1 + \sum_j E[r_{jk}] \tag{29}$$

$$\tilde{\omega}^{v,2}_k = \gamma + \sum_{j,k'>k} E[r_{jk'}] \tag{30}$$

2.2 fMRI model variables

The free energy terms corresponding to $\boldsymbol{e}$ are

$$\mathcal{F}[q(\boldsymbol{e})] = \int_{\boldsymbol{e}} q(\boldsymbol{e}) \Bigg(\log q(\boldsymbol{e}) + e^2_{jih} \omega^{e,2}_{jh} - e_{jih} \omega^{e,1}_{jh} - \sum_{j,i,t,h} \frac{E[\lambda_{jit}]}{2} \Bigg[e^2_{jih} F^2_{ht} +$$
$$- e_{jih} F_{ht} \Bigg(y_{jit} - \sum_{h' \neq h} e_{jih'} F_{h't} - E[a_{jit}] \sum_s E[x_{jis}] G_{jst} \Bigg) \Bigg] \Bigg) + \text{const.} \tag{31}$$

Recall that we assume a factored form for $q(\boldsymbol{e}) = \prod_{j,i,h} q(e_{jih})$. Minimizing with respect to this distribution gives $q(e_{jih}) \propto \exp\left\{-e^2_{jih} \tilde{\omega}^{e,2}_{jih} + e_{jih} \tilde{\omega}^{e,1}_{jih}\right\}$, with the parameters given by

$$\tilde{\omega}^{e,2}_{jih} = \omega^{e,2}_{jh} + \sum_t \tfrac{\lambda_{jit}}{2} F^2_{ht} \tag{32}$$

$$\tilde{\omega}^{e,1}_{jih} = \omega^{e,1}_{jh} + \sum_t \lambda_{jit} F_{ht} \Bigg(y_{jit} - \sum_{h' \neq h} E[e_{jih'}] F_{h't} - E[a_{jit}] \sum_s E[x_{jis}] G_{jst} \Bigg) \tag{33}$$

For the activation heights $\boldsymbol{a}$, we find

$$\mathcal{F}[q(\boldsymbol{a})] = \int_{\boldsymbol{a}} q(\boldsymbol{a}) \Bigg(\log q(\boldsymbol{a}) + a^2_{jim} \omega^{a,2}_{jm} - a_{jim} \omega^{a,1}_{jm} +$$
$$\sum_{j,i,m} \frac{E[\lambda_{jim}]}{2} \sum_t D_{jmt} \Bigg[a^2_{jim} \sum_{s,s'} E[x_{jis} x_{jis'}] G_{jst} G_{js't} +$$
$$- a_{jim} \sum_s G_{jst} E[x_{jis}] \Bigg(y_{jit} - \sum_h E[e_{jih}] F_{ht} \Bigg) \Bigg] \Bigg) + \text{const.} \tag{34}$$

Assuming factorizability, minimizing gives $q(a_{jim}) \propto \exp\left\{-a_{jim}^2\tilde{\omega}_{jim}^{a,2} + a_{jim}\tilde{\omega}_{jim}^{a,1}\right\}, a \geq 0$, with parameters given by

$$\tilde{\omega}_{jim}^{a,2} = \omega_{jm}^{a,2} + \sum_t D_{jmt}\tfrac{\lambda_{jit}}{2}\sum_{s,s',t} G_{jst}G_{js't}E[x_{jis}x_{jis'}] \tag{35}$$

$$\tilde{\omega}_{jim}^{a,1} = \omega_{jm}^{a,1} + \sum_t D_{jmt}\sum_s G_{jst}E[x_{jis}]\left(y_{jit} - \sum_h E[e_{jih}]F_{ht}\right) \tag{36}$$

The terms relating to the noise precisions $\boldsymbol{\lambda}$ are computed as

$$\begin{aligned}\mathcal{F}[q(\boldsymbol{\lambda})] = \int_{\boldsymbol{\lambda}} q(\lambda)\Bigg[\log q(\boldsymbol{\lambda}) - \log(\lambda_{jim})\omega_{jm}^{\lambda,1} - \lambda_{jim}\omega_{jm}^{\lambda,2} + \\ - \sum_{j,i,m,t} D_{jmt}\Bigg(- \tfrac{\log(\lambda_{jim})}{2} + \tfrac{\lambda_{jim}}{2}\Bigg[y_{jit}^2 + \big(E[a_{jim}^2]\sum_{s,s'} E[x_{jis}x_{jis'}]G_{jst}G_{js't}\big) + \\ \sum_h \big(E[e_{jih}^2]F_{ht}^2 + \sum_{h'\neq h} E[e_{jih}]E[e_{jih'}]F_{ht}F_{h't}\big) + \\ - y_{jit}\big(\sum_h E[e_{jih}]F_{ht} + E[a_{jim}]\sum_s G_{jst}E[x_{jis}]\big) + \\ \left(\sum_h E[e_{jih}]F_{ht}\right)\left(E[a_{jim}]\sum_s G_{jst}E[x_{jis}]\right)\Bigg]\Bigg)\Bigg] + \text{const.}\end{aligned} \tag{37}$$

For a factorizable $q(\boldsymbol{\lambda}) = \prod_{jim} q(\lambda_{jim})$, minimization with respect to $q(\lambda_{jim})$ shows that $q(\lambda_{jim}) \propto \exp\left\{\log(\lambda_{jim})\tilde{\omega}_{jim}^{\lambda,1} - \lambda_{jim}\tilde{\omega}_{jim}^{\lambda,2}\right\}$, where the parameters are given by

$$\begin{aligned}\tilde{\omega}_{jim}^{\lambda,2} &= \omega_{jm}^{\lambda,2} + \textstyle\sum_t \frac{D_{jmt}}{2} \\ \tilde{\omega}_{jim}^{\lambda,1} &= \omega_{jm}^{\lambda,1} + \Bigg(y_{jit}^2 + \sum_h \big(E[e_{jih}^2]F_{ht}^2 + \sum_{h'\neq h} E[e_{jih}]E[e_{jih'}]F_{ht}F_{h't}\big) + \\ &\qquad \big(E[a_{jim}^2]\sum_{s,s'} E[x_{jis}x_{jis'}]G_{jst}G_{js't}\big) + \\ &\qquad - y_{jit}\big(\sum_h E[e_{jih}]F_{ht} + E[a_{jim}]\sum_s G_{jst}E[x_{jis}]\big) + \\ &\qquad \left(\sum_h E[e_{jih}]F_{ht}\right)\left(E[a_{jim}]\sum_s G_{jst}E[x_{jis}]\right)\Bigg)\end{aligned} \tag{38}$$

2.3 Voxel activation variables

The Gibbs free energy as a function only of the posterior over voxel activation variables $\boldsymbol{x}$ can be computed as below. For notational convenience, we let $\Phi_{jis} = \sum_{k,l} E[\log(\phi_{k,l})]\delta(z_{ji},k)\delta(c_s,l)$ and $\bar{\Phi}_{jis} = \sum_{k,l} E[\log(1-\phi_{k,l})]\delta(z_{ji},k)\delta(c_s,l)$.

$$\begin{aligned}\mathcal{F}[q(\boldsymbol{x})] = \sum_{\boldsymbol{x}} q(\boldsymbol{x})\Bigg[\log q(\boldsymbol{x}) - \sum_{jis}\Bigg[(1-x_{jis})\bar{\Phi}_{jis} + \\ x_{jis}\Bigg(\Phi_{jis} + \sum_m \tfrac{E[\lambda_{jim}]E[a_{jim}]}{2}\sum_t D_{jmt}G_{jst}\big(y_{jit} - \sum_h e_{jih}F_{ht}\big) + \\ \sum_m \tfrac{E[\lambda_{jim}]E[a_{jim}^2]}{2}\sum_t D_{jmt}\sum_{s'} G_{jst}G_{j'st}x_{jis'}\Bigg)\Bigg]\Bigg]\end{aligned} \tag{39}$$

Minimization of this function with respect to $q(\boldsymbol{x})$ yields

$$q(x_{jis}=1) \propto \exp\left\{\Phi_{jis} + \sum_m E[\lambda_{jim}] \sum_t D_{jmt}\Big[E[a_{jim}]G_{jst}\big(y_{jit} - \sum_h e_{jih}F_{ht}\big) + \right.$$
$$\left. - \tfrac{1}{2}E[a_{jim}^2]\big(G_{jst}^2 + \sum_{s'\neq s} 2G_{jst}G_{js't}x_{jis'}\big)\Big]\right\} \quad (40)$$

$$q(x_{jis}=0) \propto \exp\left\{\hat{\Phi}_{jis}\right\} \quad (41)$$

2.4 System membership and activation probabilities

First, note that with the optimal posterior over the auxiliary variables q^*, defined in Equation (19), we have:

$$E[\log q^*(\boldsymbol{r}|\boldsymbol{z}) - \log p(\boldsymbol{z}, \boldsymbol{r} \mid \pi, \alpha)] = \sum_j \Big(\log\Gamma(\alpha + N_j) - \log\Gamma(\alpha)\Big) + \sum_{jk} E_{\boldsymbol{z}}\Big[\log\Gamma(\tilde{\omega}_{jk}^r) - \log\Gamma(\tilde{\omega}_{jk}^r + n_{jk})\Big]. \quad (42)$$

The Gibbs free energy as a function of posterior over a single membership variable $q(z_{ji})$ then becomes:

$$\mathcal{F}[q(z_{ji})] = \sum_k q(z_{ji}=k)\log q(z_{ji}=k) - \sum_k E_{\boldsymbol{z}}\Big[\log\Gamma(\tilde{\omega}_{jk}^r + n_{jk})\Big] + \\ -\sum_k q(z_{ji}=k)\sum_{l,s}\Big[q(x_{jis}=1)q(c_s=l)E[\log\phi_{k,l}] + q(x_{jis}=0)q(c_s=l)E[\log(1-\phi_{k,l})]\Big] \quad (43)$$

We can simplify the second term on the right hand side of Equation (43) as:

$$E_{\boldsymbol{z}}\Big[\log\Gamma(\tilde{\omega}_{r_{ji}} + n_{jk})\Big] = E_{\boldsymbol{z}}\Big[\delta(z_{ji},k)\log(\tilde{\omega}_{jk}^r + n_{jk}^{\neg ji}) + \log\Gamma(\tilde{\omega}_{jk}^r + n_{jk}^{\neg ji})\Big], \quad (44)$$
$$= q(z_{ji}=k)E_{\boldsymbol{z}^{\neg ji}}[\log(\tilde{\omega}_{jk}^r + n_{jk}^{\neg ji})] + E_{\boldsymbol{z}^{\neg ji}}[\log\Gamma(\tilde{\omega}_{jk}^r + n_{jk}^{\neg ji})], \quad (45)$$

where $n_{jk}^{\neg ji}$ and $\boldsymbol{z}^{\neg ji}$ indicate the exclusion of voxel i in subject j and only the first term is a function of $q(z_{ji})$. Now, minimizing Equation (43) yields the following update for membership variables:

$$q(z_{ji}=k) \propto \exp\Big\{E_{\boldsymbol{z}^{\neg ji}}[\log(\tilde{\omega}_{jk}^r + n_{jk}^{\neg ji})] + \\ \sum_{l,s}\Big(q(x_{jis}=1)q(c_s=l)E[\log\phi_{k,l}] + q(x_{jis}=0)q(c_s=l)E[\log(1-\phi_{k,l})]\Big)\Big\},$$

In order to compute the first term on the right hand side, as with the Equation (28), we use a Gaussian approximation for the distribution of n_{jk} to find:

$$E_{\boldsymbol{z}^{\neg ji}}[\log(\tilde{\omega}_{jk}^r + n_{jk}^{\neg ji})] \approx \log(\tilde{\omega}_{jk}^r + E[n_{jk}^{\neg ji}]) - \tfrac{V[n_{jk}^{\neg ji}]}{2(\tilde{\omega}_{jk}^r + E[n_{jk}^{\neg ji}])^2}. \quad (46)$$

For the unit/category activation variables, we find

$$\mathcal{F}[q(\phi_{k,l})] = \int_v q(\phi_{k,l})\Bigg(\log q(\phi_{k,l}) - \sum_k\Big[\Big\{\omega^{\phi,1} + \sum_{j,i,s} q(z_{ji}=k)q(c_s=l)q(x_{jis}=1)\Big\}\log\phi_{k,l} + \\ \Big\{\omega^{\phi,2} + \sum_{j,i,s} q(z_{ji}=k)q(c_s=l)q(x_{jis}=0)\Big\}\log(1-\phi_{k,l})\Big]\Bigg) + \text{const.} \quad (47)$$

The posterior distribution is thus found to be $\phi_{k,l} \sim \mathrm{Beta}(\tilde{\omega}_{k,l}^{\phi,1}, \tilde{\omega}_{k,l}^{\phi,2})$, with parameters as described by:

$$\tilde{\omega}_{k,l}^{\phi,1} = \omega^{\phi,1} + \sum_{j,i,s} q(z_{ji} = k) q(c_s = l) q(x_{jis} = 1) \tag{48}$$

$$\tilde{\omega}_{k,l}^{\phi,2} = \omega^{\phi,2} + \sum_{j,i,s} q(z_{ji} = k) q(c_s = l) q(x_{jis} = 0) \tag{49}$$

The stimulus category memberships are given by

$$\log q(c_s = l) = \sum_{i,k} \Big[q(x_{jis} = 1) q(z_{ji} = k) E[\log \phi_{k,l}] + q(x_{jis} = 0) q(z_{ji} = k) E[\log(1 - \phi_{k,l})] \Big] + E[\log \rho] + \text{const.} \tag{50}$$

Where ρ is given as in Equation (12), and the posterior over the stick-breaking weights is $u_l \sim \mathrm{Beta}(\tilde{\omega}_l^{u,1}, \tilde{\omega}_l^{u,2})$:

$$\tilde{\omega}_l^{u,1} = 1 + \sum_s q(c_s = l) \tag{51}$$

$$\tilde{\omega}_l^{u,1} = \chi + \sum_{s, l' < l} q(c_s = l') \tag{52}$$

References

[1] D.M. Blei and M.I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144, 2006.

[2] Y.W. Teh, K. Kurihara, and M. Welling. Collapsed variational inference for HDP. *Advances in Neural Information Processing Systems*, 20:1481–1488, 2008.

[3] C.E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

[4] Y.W. Teh, D. Newman, and M. Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 19:1353–1360, 2007.