UMass Stat697ML: Statistical Machine Learning December 12, 2016

Final project

Yubing Yao

1 Project Statement

We plan to introduce and develop the particle EM algorithm (Rockova V. 2016) under the framework of Latent Dirichlet Allocation (LDA) model (D. Blei et al. 2003). We will derive the adapted Evidence Lower Bound (ELBO) in Variational Bayesian inference based on a simpler distribution with particle EM approximation. In our project, we are planning to identify the multi-modal posterior in LDA model and to implement particle EM algorithm motivated by population-based optimization.
The Particle EM Algorithm was first introduced by Veronika Rockova in Bayesian variable selection to find the best multiple point approximation of posterior marginal distribution in 2016. The particle EM algorithm explores the whole search space by multiple repulsive particles and try to capture the multiple modes of posterior distribution. A lot of other variation inference approaches have already been implemented to LDA such as stochastic variational inference (Matthew D. Hoffman, et al. 2013) and nonparametric variational inference approach (Samuel J. Gershman, et al. 2012). So we are planning to compare those three approaches (Particle EM Algorithm, Stochastic variational inference and nonparametric variational inference) within LDA model.

2 Latent Dirichlet Allocation(LDA) model framework

The Latent Dirichlet Allocation(LDA) model is a probabilistic model for collections of discrete data such as text corpora introduced by (D. Blei et al. 2003). Following the same notation as D. Blei et al. 2003:

- Denote a collection of M documents-$\mathcal{D} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_M\}$.
- A document has a sequence of N words denoting $\boldsymbol{w} = \{w_1, w_2, \ldots, w_N\}$.
- A word defined as an item from a vocabulary indexed by $\{1, 2, \ldots, V\}$. Represent words using unit-basis vectors that have a single component equal to one and all other components equal to zero. Thus, using superscripts to denote components, the vth word in the vocabulary is represented by a V-vector w such that $w^v = 1$ and $w^u = 1$ for $u \neq v$.

Assume the documents are represented as random mixture over latent topics in LDA model, without slight modification, firstly we fix the number of latent topics-N is fixed,then the following generative process for each document-$\boldsymbol{w}$ in a corpus-$\mathcal{D}$:

1. Choose $\theta \sim Dir(\alpha)$

2. For each of the N words-w_n:
    (a) Choose a topic $z_n \sim Multinomial(\theta)$
    (b) Choose a word w_n from $p(w_n|z_n, \beta)$

Denote $k \times V$ matrix β where $\beta_{ij} = p(w^j = 1|z^i = 1), j = 1, \ldots, V, i = 1, 2, \ldots, k$, and $w^j_{n_d} = 1$ of jth component of the word $w_{n_d}, j = 1, \ldots, V, n_d = 1, \ldots, N_d, d = 1, 2, \ldots, D$.
Given the parameter-α and β, for a corpus-$\mathcal{D}$ with $d = 1, \ldots, D$ documents, then joint distribution of a topic mixture-θ_d, a set of N_d topics-$\boldsymbol{z}_d$ and a set of N_d words-$\boldsymbol{w}_d$ is given by:

$$p(\mathcal{D}|\alpha, \beta) = \prod_{d=1}^{D} \int p(\theta_d, \boldsymbol{z}_d, \boldsymbol{w}_d|\alpha, \beta) d\theta_d = \prod_{d=1}^{D} \int p(\theta_d|\alpha) \prod_{n_d=1}^{N} p(\boldsymbol{z}_{dn_d}|\theta_d) p(\boldsymbol{w}_{dn_d}|\boldsymbol{z}_{dn_d}, \beta) d\theta_d$$

where $p(\theta_d|\alpha)$ assuming K dimensional vector of θ_d:

$$p(\theta_d|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$$

And denote $\phi_{n_d i} = I(z^i_{w_{n_d}} = 1)$, that is, n_dth word is generated from latent topic-i, then $p(\boldsymbol{w}_{dn}|\boldsymbol{z}_{dn}, \beta)$ can be represented as:

$$p(\boldsymbol{w}_{dn}|\boldsymbol{z}_{dn}, \beta) = \prod_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w^j_{n_d}} = \prod_{i=1}^{k} \prod_{j=1}^{V} (\beta_{ij})^{w^j_{n_d} \phi_{n_d i}}$$

3 Variational Inference in Latent Dirichlet Allocation(LDA) model

Two free variational parameters-$\boldsymbol{\gamma}, \boldsymbol{\phi}$ was introduced in D. Blei et al. 2003 to break the coupling between model parameters-θ and β in LDA such that:

$$q(\theta_d, \boldsymbol{z}_d|\gamma, \phi) = q(\theta|\gamma) \prod_{n_d=1}^{N_d} q(\boldsymbol{z}_{n_d}|\phi_{n_d})$$

For any fixed d, we can achieve a evidence lower bound using the variational distribution-$q(\theta_d, \boldsymbol{z}_d|\gamma, \phi)$,

$$\begin{aligned} \log p(\boldsymbol{w}_d|\alpha, \beta) &= \log \int \sum_{\boldsymbol{z}_d} p(\theta_d, \boldsymbol{z}_d, \boldsymbol{w}_d|\alpha, \beta) d\theta_d \\ &= \log \int \sum_{\boldsymbol{z}_d} \frac{p(\theta_d, \boldsymbol{z}_d, \boldsymbol{w}_d|\alpha, \beta) q(\theta_d, \boldsymbol{z}_d|\gamma, \phi)}{q(\theta_d, \boldsymbol{z}_d|\gamma, \phi)} d\theta_d \\ &\geq \int \sum_{\boldsymbol{z}_d} q(\theta_d, \boldsymbol{z}_d|\gamma, \phi) \log p(\theta_d, \boldsymbol{z}_d, \boldsymbol{w}_d|\alpha, \beta) - \int \sum_{\boldsymbol{z}_d} q(\theta_d, \boldsymbol{z}_d|\gamma, \phi) \log q(\theta_d, \boldsymbol{z}_d|\gamma, \phi) d\theta_d \\ &\quad = \mathrm{E}_q[p(\theta, \boldsymbol{z}_d, \boldsymbol{w}_d|\alpha, \beta)] - \mathrm{E}_q[q(\theta_d, \boldsymbol{z}_d|\gamma, \phi)] \\ &\quad = \mathrm{E}_q[p(\theta, \boldsymbol{z}_d, \boldsymbol{w}_d|\alpha, \beta)] + H(\gamma, \phi) \end{aligned} \tag{1}$$

where the entropy-$H(\gamma, \phi) = -\mathrm{E}_q[q(\theta_d, \boldsymbol{z}_d|\gamma, \phi)]$.

Right-hand side of the inequality1 above is a lower bound on the log likelihood for an arbitrary variational distribution-$q(\theta, \boldsymbol{z}|\gamma, \phi)$.

This lower bound can be denoted as $\mathcal{L}(\gamma, \phi, \alpha, \beta)$,

$$\mathcal{L}(\gamma, \phi, \alpha, \beta) = \sum_{d=1}^{D} (\mathrm{E}_q[\log p(\theta_d|\alpha)] + \mathrm{E}_q[\log p(\boldsymbol{z}_d|\theta_d)] + \mathrm{E}_q[\log p(\boldsymbol{w}_d|\boldsymbol{z}_d, \beta)] - \mathrm{E}_q[\log q(\theta_d)]) \tag{2}$$

4 Particle approximation and Evidence Lower Bound

Motivated by the particle approximation in Rockova V. 2016, for $d = 1, 2, \ldots, D$, we use a weighted mixture of atoms to approximate $\pi(\boldsymbol{z}_{dn}|\mathcal{D}), n = 1, 2, \ldots, N_d, d = 1, 2, \ldots, D$, and we denote:

$$q_{PEM}(\boldsymbol{z}_{dn}|\boldsymbol{\Gamma}_{dn}, \boldsymbol{\omega}) = \sum_{p=1}^{P} \omega_p \mathbb{I}\{\boldsymbol{z}_{dn} = \boldsymbol{z}_{pdn}\}$$

where $\boldsymbol{\Gamma}_{dn} = [\boldsymbol{z}_{1dn}, \boldsymbol{z}_{2dn}, \ldots, \boldsymbol{z}_{Pdn}]$,corresponding importance weights-$\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_P)^T$, where $\sum_{p=1}^{P} \omega_p = 1, \forall d = 1, \ldots, D, n = 1, 2, \ldots, N_d; 0 \leq \omega_p \leq 1, \forall p = 1, 2, \ldots, P$.
For $\boldsymbol{z}_{pdn}, \forall p = 1, 2, \ldots, P, n = 1, 2, \ldots, N_d, d = 1, 2, \ldots, D$, z_{pdn} can only takes one of $1, 2, \ldots, k$ values, representing the possible k topics in LDA model.

And denote $\boldsymbol{z}^i_{pdn}$ will only can take the $i = 1, 2, \ldots, k$ values corresponding to k possible topics.

$$z^i_{pdn} = \begin{cases} 1 & \text{if word } z_{pdn} \text{ is from topic } i \\ 0 & otherwise \end{cases}$$

We follow the similar setup for the variational distribution setup for $q(\theta|\gamma)$ in D. Blei et al. 2003. Thus the new variational distribution of $q_{PEM}(\theta_d, \boldsymbol{z}_d|\boldsymbol{\gamma}_d, \boldsymbol{\Gamma}_d, \boldsymbol{\omega})$ is:

$$q_{PEM}(\theta, \boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{\Gamma}, \boldsymbol{\omega}) = \prod_{d=1}^{D} \left[q(\theta_d|\gamma_d) \prod_{n_d=1}^{N_d} q_{PEM}(\boldsymbol{z}_{dn_d}|\boldsymbol{\Gamma}_d, \boldsymbol{\omega}) \right]$$

where $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_D], \boldsymbol{\omega} = [\omega_1, \omega_2, \ldots, \omega_P]$.

Then the evidence lower bound-2 replacing $q(\theta, \boldsymbol{z}|\gamma, \phi)$ with $q_{PEM}(\theta_d, \boldsymbol{z}_d|\gamma_d, \boldsymbol{\Gamma}_d, \boldsymbol{\omega})$ can be written as :

$$\begin{aligned} \mathcal{L}_\lambda(\boldsymbol{\gamma}, \boldsymbol{\Gamma}, \boldsymbol{\omega}; \boldsymbol{\theta}, \alpha, \beta) &= \sum_{d=1}^{D} (\mathrm{E}_{q_{PEM}}[\log p(\theta_d|\alpha)] + \mathrm{E}_{q_{PEM}}[\log p(\boldsymbol{z}_d|\theta_d)] + \mathrm{E}_{q_{PEM}}[\log p(\boldsymbol{w}_d|\boldsymbol{z}_d, \beta)] \\ &\quad -\mathrm{E}_q[\log q(\theta_d|\boldsymbol{\gamma}_d)] - \lambda \mathrm{E}_{q_{PEM}}[\log q_{PEM}(\boldsymbol{z}_{n_d}|\boldsymbol{\Gamma}_d, \boldsymbol{\omega}_d)]) \\ &= \sum_{d=1}^{D} (\mathrm{E}_{q_{PEM}}[\log p(\theta_d|\alpha)] + \mathrm{E}_{q_{PEM}}[\log p(\boldsymbol{z}_d|\theta_d)] + \mathrm{E}_{q_{PEM}}[\log p(\boldsymbol{w}_d|\boldsymbol{z}_d, \beta)] \\ &\quad + \sum_{d=1}^{D} (-\mathrm{E}_q[\log q(\theta_d|\boldsymbol{\gamma}_d)] + H_\lambda(\boldsymbol{\Gamma}_d, \boldsymbol{\omega})) \end{aligned} \tag{3}$$

where $H_\lambda(\boldsymbol{\Gamma}_d, \boldsymbol{\omega}) = -\lambda \mathrm{E}_{q_{PEM}}[\log q_{PEM}(\boldsymbol{z}_d|\boldsymbol{\Gamma}_d, \boldsymbol{\omega})], \lambda \geq 0$.
When $\lambda = 1$, the above is regular ELBO from variational calculus. if $\lambda = 0$, it is equivalent to

parallel EM.

In the E step of particle EM algorithm, we need to maximize the variational parameters-$\boldsymbol{\gamma}, \boldsymbol{\Gamma}, \boldsymbol{\omega}$, Finding $\hat{\boldsymbol{\Gamma}}_d = [\hat{\boldsymbol{Z}}_{1d}, \hat{\boldsymbol{Z}}_{2d}, \ldots, \hat{\boldsymbol{Z}}_{Pd}]$ and $\hat{\boldsymbol{\omega}}$ such that

$$(\hat{\boldsymbol{\Gamma}}, \hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\gamma}}) = \underset{\boldsymbol{\Gamma},\boldsymbol{\omega},\boldsymbol{\gamma}}{\text{argmax}} \mathcal{L}_\lambda(\boldsymbol{\gamma}, \boldsymbol{\Gamma}, \boldsymbol{\omega}; \alpha, \beta) \quad \text{subject to} \sum_{p=1}^{P} \omega_p = 1, 0 \leq \omega_p \leq 1$$

where $\hat{\boldsymbol{\Gamma}} = [\hat{\boldsymbol{\Gamma}}_1, \hat{\boldsymbol{\Gamma}}_2, \ldots, \hat{\boldsymbol{\Gamma}}_D], \hat{\boldsymbol{\omega}} = [\hat{\boldsymbol{\omega}}_1, \hat{\boldsymbol{\omega}}_2, \ldots, \hat{\boldsymbol{\omega}}_D], \hat{\boldsymbol{\gamma}} = [\hat{\boldsymbol{\gamma}}_1, \hat{\boldsymbol{\gamma}}_2, \ldots, \hat{\boldsymbol{\gamma}}_D]$.

Denote the P^*_{dni} unique particles contained with each $\boldsymbol{\Gamma}_{dni}$ by: $\boldsymbol{\Gamma}^*_{ndi} = [\phi^*_{1dni}, \phi^*_{2dni}, \ldots, \phi^*_{P^*_{dni}dni}]$,denote $p^*_{l_{dni}}$ the cumulative importance weight associated with each unique particle-ϕ^*_{ldni},i.e.,

$$p^*_{ldni} = \sum_{p=1}^{P} \omega_p \mathbb{I}(\phi_{pdni} = \phi^*_{ldni})$$

Then the term related with $\boldsymbol{\Gamma}, \boldsymbol{\omega}$ in entropy of ELBO–3 can be expressed as:

$$H_\lambda(\boldsymbol{\Gamma}, \boldsymbol{\omega}) = -\lambda \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \sum_{l=1}^{P^*_{dni}} p^*_{ldni} \log(p^*_{ldni})$$

5 Particle EM

5.1 E step with single particle

Before entirely introduce the Particle EM algorithm in LDA model, firstly we assume single particle-$P = 1$ and fix some document-$\boldsymbol{w}_d$, the E step at m iteration, given the $\boldsymbol{Z}^{(m)}_d$ such that complete data surrogate objective function related with $\boldsymbol{Z}_d$:

$$\begin{aligned} Q(\boldsymbol{Z}_d | \boldsymbol{Z}^{(m)}_d, \boldsymbol{w}_d) &= \mathrm{E}_{\alpha,\beta,\theta_d|\boldsymbol{w}_d,\boldsymbol{Z}^{(m)}_d} \log \pi(\alpha, \beta, \theta_d | \boldsymbol{w}_d) \\ &= \sum_{n=1}^{N_d} \sum_{i=1}^{k} \phi_{dni}(\Psi(\gamma^{(m)}_i) - \Psi(\sum_{j=1}^{k} \gamma^{(m)}_j)) \\ &\quad + \sum_{n=1}^{N_d} \sum_{i=1}^{k} \sum_{j=1}^{V_d} \boldsymbol{z}^{i(m)}_{dn} w^j_n \log \beta_{ij} \end{aligned} \tag{4}$$

where $\phi_{dni} = \mathrm{E}(z^i_{dn}) = P(z^i_{dn} = 1)$.

Then focusing on the z_{dn}, then

$$z^{(m+1)}_{dn} = \underset{z_{dn} \in \{1,2,\ldots,k\}}{\text{argmax}} \left\{ \sum_{i=1}^{k} z^i_{dn}(\Psi(\gamma^{(m)}_i) - \Psi(\sum_{j=1}^{k} \gamma^{(m)}_j)) + \sum_{i=1}^{k} \boldsymbol{z}^{i(m)}_{dn} \log \beta_{iv_d} \right\} \tag{5}$$

The equation-5 can be treated as the log-likelihood function of a Categorical distribution with the probability-$\boldsymbol{\phi}_{nd} = [\phi_{dn1}, \phi_{dn2}, \ldots, \phi_{dnk}]$, $\sum_{i=1}^{k} \phi_{dni} = 1$ where $\phi_{dni} = \mathrm{E}(z^i_{dn} | \boldsymbol{Z}^{(m)}_d, \boldsymbol{\gamma}^{(m)}_d) = P(z^i_{dn} = 1 | \boldsymbol{Z}^{(m)}_d, \boldsymbol{\gamma}^{(m)}_d), \forall i = 1, 2, \ldots, k$.

The next update of $z_{dn}^{(m+1)} = i$ is obtained by choosing the $i, 1 \leq i \leq k$ such that $\phi_{ndi} = \max\{\phi_{dn1}, \phi_{dn2}, \ldots, \phi_{dnk}\}$.

And under $P = 1$ the entropy term:

$$\begin{aligned} -\mathrm{E}_q[\log q(\theta_d|\boldsymbol{\gamma}_d)] + H_\lambda(\boldsymbol{\Gamma}_d, \boldsymbol{\omega}_d) &= -\Psi(\textstyle\sum_{j=1}^k \gamma_j) + \displaystyle\sum_{i=1}^k \log\Gamma(\gamma_i) - \sum_{i=1}^k (\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^k \gamma_j)) \\ &\quad + H_\lambda(\boldsymbol{\Gamma}_d, \boldsymbol{\omega}_d) \end{aligned} \tag{6}$$

Then with single particle $P = 1$ and fix some $d, \boldsymbol{w}_d$ the evidence lower bound-3 becomes:

$$\begin{aligned} \mathcal{L}_{d\lambda}(\boldsymbol{\gamma}_d, \boldsymbol{\Gamma}, \boldsymbol{\omega}; \boldsymbol{\theta}, \alpha, \beta) &= \log\Gamma(\textstyle\sum_{j=1}^k \alpha_j) + \displaystyle\sum_{i=1}^k (\alpha_i - 1)(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^k \gamma_{dj})) \\ &\quad + \sum_{n=1}^{N_d} \sum_{i=1}^k \phi_{dni}(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^k \gamma_{dj})) \\ &\quad + \sum_{n=1}^{N_d} \sum_{i=1}^k \sum_{j=1}^V \phi_{dni} w_n^j \log \beta_{ij} \\ &\quad - \Psi(\textstyle\sum_{j=1}^k \gamma_{dj}) + \displaystyle\sum_{i=1}^k \log\Gamma(\gamma_{di}) - \sum_{i=1}^k (\gamma_{di} - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^k \gamma_{dj})) \\ &\quad + H_\lambda(\boldsymbol{\Gamma}_d, \boldsymbol{\omega}_d) \end{aligned} \tag{7}$$

Firstly maximize the ELBO-7 with respect to ϕ_{dni} with the constraint-$\sum_{j=1}^k \phi_{dnj} = 1$. Denote $\beta_{iv_d} = p(w_n^{v_d} = 1 | z_d^i = 1)$ for the appropriate v_d. Then add the Lagrange multiplier to the terms in ELBO-7 containing ϕ_{ni},

$$\mathcal{L}_{d[\phi_{dni}]} = \phi_{dni}(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^k \gamma_{dj})) + \phi_{dni} \log \beta_{iv_d} - \lambda \phi_{dni} \log \phi_{dni} + \lambda_\phi(\sum_{j=1}^k \phi_{dnj} - 1)$$

Take first derivative in terms of ϕ_{dni}, then:

$$\frac{\partial \mathcal{L}_d}{\partial \phi_{dni}} = \Psi(\gamma_{di}) - \Psi(\sum_{j=1}^k \gamma_{dj}) + \log \beta_{iv_d} - \lambda - \lambda \log \phi_{dni} + \lambda_\phi$$

Set the derivative above equal to zero, then the value of estimated $\hat{\phi}_{dni}$ maximizing the ELBO-7 is:

$$\hat{\phi}_{dni} = \frac{\beta_{iv_d}^{1/\lambda} \exp((\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^k \gamma_{dj}))/\lambda)}{\sum_{i=1}^k \beta_{iv_d}^{1/\lambda} \exp((\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^k \gamma_{dj}))/\lambda)}$$

Next, we maximize the equation-7 with respect to γ_{di}, the terms containing γ_i are:

$$\begin{aligned} \mathcal{L}_{d[\gamma_{di}]} &= (\alpha_i - 1)\Psi(\gamma_{di}) - \Psi(\textstyle\sum_{j=1}^k \gamma_{dj}) \displaystyle\sum_{j=1}^k (\alpha_j - 1) + \sum_{n=1}^{N_d} \phi_{dni}(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^k \gamma_{dj})) \\ &\quad - \Psi(\textstyle\sum_{j=1}^k \gamma_j) + \displaystyle\sum_{i=1}^k \log\Gamma(\gamma_{di}) - (\gamma_{di} - 1)\Psi(\gamma_{di}) + \Psi(\sum_{j=1}^k \gamma_{dj}) \sum_{j=1}^k (\gamma_{dj} - 1) \end{aligned} \tag{8}$$

Take first derivative in terms of γ_{di}:

$$\frac{\partial \mathcal{L}_{d[\gamma_{di}]}}{\partial \gamma_{di}} = (\alpha_i + \sum_{n=1}^{N_d} \phi_{dni} - \gamma_{di})\Psi'(\gamma_{di}) - \Psi'(\sum_{i=1}^{k} \gamma_{di}) \sum_{j=1}^{k} (\alpha_j + \sum_{n=1}^{N_d} \phi_{dni} - \gamma_{dj})$$

Setting the above first derivative equal to 0 then the value of estimated $\hat{\gamma}_{di}$ maximizing the ELBO-7 is:

$$\hat{\gamma}_{di} = \alpha_i + \sum_{n=1}^{N_d} \phi_{dni}$$

Another way to represent the updated Z_d given $Z_d^{(m)}$, then with single particle $P = 1$ and fix some $d, \boldsymbol{w}_d$ the evidence lower bound-3 becomes:

$$\begin{aligned} Q(\boldsymbol{Z}_d, \boldsymbol{\gamma}_d; \alpha, \beta, \theta_d | Z_d^{(m)}) &= \mathrm{E}_{\alpha,\beta,\theta_d|\boldsymbol{w}_d} \log \pi(\alpha, \beta, \theta_d | \boldsymbol{w}_d, Z_d^{(m)}) \\ &= \log \Gamma(\textstyle\sum_{j=1}^{k} \alpha_j) - \displaystyle\sum_{i=1}^{k} (\alpha_i - 1)(\Psi(\gamma_{di}) - \Psi(\sum_{i=1}^{k} \gamma_{di}) \\ &\quad + \sum_{n=1}^{N_d} \sum_{i=1}^{k} z_{dn}^{i(m)} (\Psi(\gamma_i) - \Psi(\sum_{i=1}^{k} \gamma_i) \\ &\quad + \sum_{n=1}^{N_d} \sum_{i=1}^{k} \sum_{j=1}^{V_d} z_{dn}^{i(m)} w_n^j \log \beta_{ij} \\ &\quad - \Psi(\textstyle\sum_{j=1}^{k} \gamma_j) + \displaystyle\sum_{i=1}^{k} \log \Gamma(\gamma_{di}) - \sum_{i=1}^{k} (\gamma_{di} - 1)(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_j)) \\ &\quad - \lambda \sum_{n=1}^{N_d} \sum_{i=1}^{k} z_{dn}^{i(m)} \log \phi_{dni} \end{aligned} \tag{9}$$

where $\phi_{dni} = p(Z_{dn}^i = 1), i = 1, 2, \ldots, k$. With similar method as above, we can get the estimate he value of estimated $\hat{\phi}_{dni}$ maximizing the ELBO-10 is:

5.2 M step with single particle

In M step with single particle-$P = 1$, with all the documents-$\boldsymbol{w}_d, d = 1, \ldots, D$, maximize the ELBO-7 with respect to model parameters-α, β.

Similarly to Blei et al. 2003, choose the terms related to β and add Lagrange multiplier,

$$\mathcal{L}_{[\beta]} = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \sum_{j=1}^{V} \phi_{dni} w_{dn}^j \log \beta_{ij} + \sum_{i=1}^{k} \lambda_{\beta i} (\sum_{j=1}^{V} \beta_{ij} - 1)$$

Take the first derivative in terms of β_{ij}, set it to zero, then:

$$\hat{\beta}_{ij} = \frac{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^j}{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{j'=1}^{V} \phi_{dni} w_{dn}^{j'}}$$

And also choose the terms containing α:

$$\mathcal{L}_{[\alpha]}=\sum_{d=1}^{D}\left(\log\Gamma(\sum_{j=1}^{k}\alpha_j)+\sum_{i=1}^{k}(\alpha_i-1)(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj}))\right)$$

Take the first derivative in terms of α_i, then:

$$\frac{\partial\mathcal{L}}{\partial\alpha_i}=D(\Psi(\sum_{j=1}^{k}\alpha_j)-\Psi(\alpha_i))+\sum_{d=1}^{D}(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj}))$$

Set it to zero, then $\forall i\neq j=1,2,\ldots,k,$:

$$\frac{\partial\mathcal{L}}{\partial\alpha_i\alpha_j}=\delta(i,j)D\Psi'(\alpha_i)-\Psi'(\sum_{j=1}^{k}\alpha_j)$$

5.3 E step with multiple particles

With $P>1$, we need alternatively collaborative updating the particle location-$[\boldsymbol{\phi}_{1dn},\boldsymbol{\phi}_{2dn},\ldots,\boldsymbol{\phi}_{Pdn}],\forall n=1,2,\ldots,N_d$ and their corresponding importance weights-$(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\ldots,\boldsymbol{\omega}_P)^T,\forall d=1,\ldots,D;n=1,2,\ldots,N_d$.

Denote $\boldsymbol{\Gamma}^{(m)}=[\boldsymbol{\Gamma}^{(m)}_{1dn},\boldsymbol{\Gamma}^{(m)}_{2dn},\ldots,\boldsymbol{\Gamma}^{(m)}_{Ddn}],\forall d=1,\ldots,D;n=1,2,\ldots,N_d$, the state of particle system at the mth iteration and denote $\boldsymbol{\omega}^{(m)}=(\boldsymbol{\omega}^{(m)}_{1dn},\boldsymbol{\omega}^{(m)}_{2dn},\ldots,\boldsymbol{\omega}^{(m)}_{Ddn})^T$.

Given $\boldsymbol{\omega}^{(m)}$ fix some d then the evidence lower bound-3 becomes:

$$\begin{aligned}\mathcal{L}_{d\lambda}(\boldsymbol{\gamma}_d,\boldsymbol{\Gamma},\boldsymbol{\omega};\boldsymbol{\theta},\alpha,\beta) \quad = \quad & \log\Gamma(\textstyle\sum_{j=1}^{k}\alpha_j)+\displaystyle\sum_{i=1}^{k}(\alpha_i-1)(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj}))\\ &+\sum_{p=1}^{P}\sum_{n=1}^{N_d}\omega_p^{(m)}\sum_{i=1}^{k}\phi_{pdni}(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj}))\\ &+\sum_{p=1}^{P}\sum_{n=1}^{N_d}\omega_p^{(m)}\sum_{i=1}^{k}\sum_{j=1}^{V}\phi_{pdni}w_n^j\log\beta_{ij}\\ &-\Psi(\textstyle\sum_{j=1}^{k}\gamma_{dj})+\displaystyle\sum_{i=1}^{k}\log\Gamma(\gamma_{di})-\sum_{i=1}^{k}(\gamma_{di}-1)(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj}))\\ &-\lambda\sum_{p=1}^{P}\sum_{n=1}^{N_d}\omega_p\sum_{i=1}^{k}\phi_{pdni}\log(\sum_{p=1}^{P}\omega_p\phi_{pdni})\end{aligned}\tag{10}$$

where $\phi_{pdni}=\mathrm{E}(z^i_{pdn})=P(z^i_{pdn}=1),\forall p=1,2,\ldots,P;d=1,2,\ldots,D,n=1,2,\ldots,N_d,i=1,2,\ldots,k$.
Similar to one particle system,firstly maximize the ELBO-10 with respect to ϕ_{pdni} with the constraint-$\sum_{j=1}^{k}\phi_{pdnj}=1$. Denote $\beta_{iv_d}=p(w_n^{v_d}=1|z^i_{pd}=1)$ for the appropriate v_d.
Then add the Lagrange multiplier to the terms in ELBO-10 containing ϕ_{pdni},

$$\begin{aligned}\mathcal{L}_{d[\phi_{pdni}]} &= \omega_p\phi_{pdni}(\Psi(\gamma_{di}) - \Psi(\textstyle\sum_{j=1}^{k}\gamma_{dj})) + \omega_p\phi_{pdni}\log\beta_{iv_d} \\ &\quad -\lambda\sum_{p'=1}^{P}\omega_{p'}\phi_{p'dni}\log(\sum_{p'=1}^{P}\omega_{p'}\phi_{p'dni}) + \lambda_{\phi_p}(\sum_{j=1}^{k}\phi_{pdnj} - 1)\end{aligned} \tag{11}$$

Take first derivative in terms of ϕ_{pdni}, then:

$$\frac{\partial\mathcal{L}_d}{\partial\phi_{pdni}} = \omega_p(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k}\gamma_{dj})) + \omega_p\log\beta_{iv_d} - \lambda\omega_p(\log(\sum_{p'=1}^{P}\omega_{p'}\phi_{p'dni}) + 1) + \lambda_{\phi_p}$$

Take second derivative in terms of ϕ_{pdni}, then:

$$\frac{\partial^2\mathcal{L}_d}{\partial\phi_{pdni}^2} = -\frac{\lambda\omega_p^2}{\sum_{p'=1}^{P}\omega_{p'}\phi_{p'dni}}$$

And the second order partial derivative in terms of $\phi_{pdni}, \phi_{p''dni}$ where $\forall p \neq p'' = 1, 2, \ldots, P$,

$$\frac{\partial^2\mathcal{L}_d}{\partial\phi_{pdni}\partial\phi_{p''dni}} = -\frac{\lambda\omega_p\omega_{p''}}{\sum_{p'=1}^{P}\omega_{p'}\phi_{p'dni}}$$

Apply the Newton-Raphson method, for each fixed $n = 1, 2, \ldots, N_d, d = 1, 2, \ldots, D$, denote $\boldsymbol{\phi}_{dni} = [\phi_{1dni}, \phi_{2dni}, \ldots, \phi_{Pdni}]^T$, then by iterating the equation below we can find the maximal-$\boldsymbol{\phi}_{dn}$:

$$\boldsymbol{\phi}_{dni(new)} = \boldsymbol{\phi}_{dni(old)} - H(\boldsymbol{\phi}_{dn(old)})^{-1}g(\boldsymbol{\phi}_{dni(old)})$$

where $H(\boldsymbol{\phi}_{dn}), g(\boldsymbol{\phi}_{dni})$ are the Hessian matrix and gradient respectively at the point-$\boldsymbol{\phi}_{dni}$ defined above.

A point need to note that the maximal $\boldsymbol{\phi}_{dni}$ achieve above need to satisfy $\sum_{i=1}^{k}\phi_{pdni} = 1$ for each fixed-$n = 1, 2, \ldots, N_d, d = 1, 2, \ldots, D, p = 1, 2, \ldots, P$.

In order to address the constraints-$\forall p = 1, 2, \ldots, P, n = 1, 2, \ldots, N_d, d = 1, 2, \ldots, D, i = 1, 2, \ldots, k, 0 \leq \phi_{pdni} \leq 1$, and for each fixed-$p, n, d, \sum_{i=1}^{k}\phi_{pdni} = 1$, we transform ϕ_{pdni} to some variable-$x_{pdni} \in (-\infty, \infty)$ with $x_{pdni} = logit(\phi_{pdni}), \phi_{pndi} = \frac{e^{x_{pdni}}}{1+e^{x_{pdni}}}, \forall p = 1, 2, \ldots, P, n = 1, 2, \ldots, N_d, d = 1, 2, \ldots, D, i = 1, 2, \ldots, k-1$ and based on the constraints-$\sum_{i=1}^{k}\phi_{pdni} = 1$, we have total $P(k-1)\sum_{d=1}^{D}N_d$ free parameters with $\phi_{pndk} = 1 - \sum_{i'=1}^{k-1}\frac{e^{x_{pdni'}}}{1+e^{x_{pdni'}}}$.

Apply the chain rule of differentiation, then the first derivative in terms of the new parameter-$x_{pdni'}, \forall p = 1, 2, \ldots, P, d = 1, 2, \ldots, D, n = 1, 2, \ldots, N_d, i' = 1, 2, \ldots, k-1$,

$$\begin{aligned}\frac{\partial\mathcal{L}_d}{\partial x_{pdni'}} = \frac{\partial\mathcal{L}_d}{\partial\phi_{pdni'}}\frac{\partial\phi_{pdni'}}{\partial x_{pdni'}} &= \Big(\omega_p(\Psi(\gamma_{di}) - \Psi(\textstyle\sum_{j=1}^{k}\gamma_{dj})) + \omega_p\log\beta_{iv_d} - \lambda\omega_p(\log(\sum_{p'=1}^{P}\omega_{p'}\frac{e^{x_{pdni'}}}{1+e^{x_{pdni'}}}) + 1)\Big) \\ &\quad \times\frac{e^{x_{pdni'}}}{(1+e^{x_{pdni'}})^2}\end{aligned} \tag{12}$$

Also the second derivative in terms of the parameter-$x_{pdni'}, \forall p = 1, 2, \ldots, P, d = 1, 2, \ldots, D, n = 1, 2, \ldots, N_d, i' = 1, 2, \ldots, k-1$,

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}_d}{\partial x^2_{pdni'}} &= \frac{\partial}{\partial x_{pdni'}}\left(\frac{\partial \mathcal{L}_d}{\partial x_{pdni'}}\right) = \frac{\partial}{\partial x_{pdni'}}\left(\frac{\partial \mathcal{L}_d}{\partial \phi_{pdni'}}\frac{\partial \phi_{pdni'}}{\partial x_{pdni'}}\right) \\
&= \frac{\partial \mathcal{L}_d}{\partial \phi_{pdni'}}\frac{\partial^2 \phi_{pdni'}}{\partial x^2_{pdni'}} + \frac{\partial^2 \mathcal{L}_d}{\partial \phi_{pdni'}\partial x_{pdni'}}\frac{\partial \phi_{pdni'}}{\partial x_{pdni'}} \\
&= \frac{\partial \mathcal{L}_d}{\partial \phi_{pdni'}}\frac{\partial^2 \phi_{pdni'}}{\partial x^2_{pdni'}} + \frac{\partial^2 \mathcal{L}_d}{\partial \phi^2_{pdni'}}(\frac{\partial \phi_{pdni'}}{\partial x_{pdni'}})^2 \\
&= \frac{\partial \mathcal{L}_d}{\partial \phi_{pdni'}}\frac{\partial \phi_{pdni'}}{\partial x_{pdni'}} \\
&= \Big(\omega_p(\Psi(\gamma_{di}) - \Psi(\textstyle\sum_{j=1}^k \gamma_{dj})) + \omega_p \log \beta_{iv_d} - \lambda\omega_p(\log(\sum_{p'=1}^P \omega_{p'} \frac{e^{x_{p'dni'}}}{1+e^{x_{p'dni'}}}) + 1)\Big) \\
&\quad \times \frac{e^{x_{pdni'}}}{(1+e^{x_{pdni'}})^2}(1 - \frac{e^{x_{pdni'}}}{1+e^{x_{pdni'}}}) - \frac{\lambda\omega_p^2}{\sum_{p'=1}^P \omega_{p'} \frac{e^{x_{pdni'}}}{1+e^{x_{pdni'}}}} * \frac{e^{2x_{pdni'}}}{(1+e^{x_{pdni'}})^4}
\end{aligned} \tag{13}$$

Also the second order partial derivative in terms of the parameter-$x_{pdni'}, x_{p''dni'}, \forall p \neq p'' = 1, 2, \ldots, P, d = 1, 2, \ldots, D, n = 1, 2, \ldots, N_d, i' = 1, 2, \ldots, k-1$,

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}_d}{\partial x_{pdni'}\partial x_{p''dni'}} &= \frac{\partial}{\partial x_{p''dni'}}\left(\frac{\partial \mathcal{L}_d}{\partial x_{pdni'}}\right) = \frac{\partial}{\partial x_{p''dni'}}\left(\frac{\partial \mathcal{L}_d}{\partial \phi_{pdni'}}\frac{\partial \phi_{pdni'}}{\partial x_{pdni'}}\right) \\
&= \frac{\partial \mathcal{L}_d}{\partial \phi_{pdni'}}\frac{\partial^2 \phi_{pdni'}}{\partial x_{pdni'}\partial x_{p''dni'}} + \frac{\partial^2 \mathcal{L}_d}{\partial \phi_{pdni'}\partial x_{p''dni'}}\frac{\partial \phi_{pdni'}}{\partial x_{pdni'}} \\
&= \frac{\partial \mathcal{L}_d}{\partial \phi_{pdni'}} \times 0 + \frac{\partial^2 \mathcal{L}_d}{\partial \phi_{pdni'}\partial \phi_{p''dni'}}\frac{\partial \phi_{pdni'}}{\partial x_{pdni'}}\frac{\partial \phi_{p''dni'}}{\partial x_{p''dni'}} \\
&= \frac{\partial^2 \mathcal{L}_d}{\partial \phi_{pdni'}\partial \phi_{p''dni'}}\frac{\partial \phi_{pdni'}}{\partial x_{pdni'}}\frac{\partial \phi_{p''dni'}}{\partial x_{p''dni'}} \\
&= -\frac{\lambda\omega_p\omega_{p''}}{\sum_{p'=1}^P \omega_{p'} \frac{e^{x_{p'dni'}}}{1+e^{x_{p'dni'}}}}\frac{e^{x_{pdni'}}}{(1+e^{x_{pdni'}})^2}\frac{e^{x_{p''dni'}}}{(1+e^{x_{p''dni'}})^2}
\end{aligned} \tag{14}$$

Thus the Newton-Raphson method in terms of the vector of P parameters-$\boldsymbol{x}_{dni'} = [x_{1dni'}, x_{2dni'}, \ldots, x_{Pdni'}]^T, \forall i' = 1, 2, \ldots, k-1, n = 1, 2, \ldots, N_d, d = 1, 2, \ldots, D$ without constraints, thus by iterating the equation below we can find the maximal-$\boldsymbol{x}_{dni'}$:

$$\boldsymbol{x}_{dni'(new)} = \boldsymbol{x}_{dni'(old)} - H(\boldsymbol{x}_{dni'(old)})^{-1} g(\boldsymbol{x}_{dni'(old)})$$

where $H(\boldsymbol{x}_{dni'}), g(\boldsymbol{x}_{dni'})$ are the Hessian matrix and gradient respectively at the vector point-$\boldsymbol{x}_{dni'}$ defined above respectively.

Apply the property of importance weights for any fixed $\displaystyle\sum_{p'=1}^{P} \omega_{p'} = 1$, and add the P equations-$\sum_{p'=1}^P \frac{\partial \mathcal{L}_d}{\partial \phi_{p'dni}}$ and set it equal to zero, then:

$$\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj}) + \log \beta_{iv_d} - \lambda(\log(\sum_{p'=1}^{P} \omega_{p'}\phi_{p'dni}) + 1) + \sum_{p'=1}^{P} \lambda_{\phi_{p'}} = 0$$

$$\log(\sum_{p'=1}^{P} \omega_{p'dn}\phi_{p'dni}) = (\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj}))/\lambda + \log \beta_{iv_d}/\lambda + \sum_{p'=1}^{P} \lambda_{\phi_{p'}}/\lambda - 1$$

Secondly, we need to update the particle importance weights-$\boldsymbol{\omega} = [\omega_1, \omega_2, \ldots, \omega_P]^T$ given $\hat{\phi}_{pdni}, \forall i = 1, 2, \ldots, k, n = 1, 2, \ldots, N_d, d = 1, 2, \ldots, D$, Then the ELBO-10 can be written as:

$$\begin{aligned}\mathcal{L}_\lambda(\boldsymbol{\gamma},\boldsymbol{\Gamma},\boldsymbol{\omega};\boldsymbol{\theta},\alpha,\beta) \quad = \quad & D\log\Gamma(\textstyle\sum_{j=1}^{k}\alpha_j) + \displaystyle\sum_{d=1}^{D}\sum_{i=1}^{k}(\alpha_i-1)(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj})) \\ & + \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\sum_{l=1}^{P}\omega_p\phi_{pdni}(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj})) \\ & + \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\sum_{j=1}^{V}\sum_{p=1}^{P}\omega_p\phi_{pdni}w_n^j\log\beta_{ij} \\ & - \sum_{d=1}^{D}\left(\Psi(\sum_{j=1}^{k}\gamma_{dj}) + \sum_{i=1}^{k}\log\Gamma(\gamma_{di}) - \sum_{i=1}^{k}(\gamma_{di}-1)(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj}))\right) \\ & - \lambda\sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\sum_{p=1}^{P}\omega_p\phi_{pdni}\log(\sum_{p=1}^{P}\omega_p\phi_{pdni})\end{aligned} \tag{15}$$

Maximize the ELBO-15 with respect to ω_p with the constraint-$\sum_{p'=1}^{P}\omega_{p'} = 1$. Denote $\beta_{iv_d} = p(w_n^{v_d} = 1|z_{pd}^i = 1)$ for the appropriate v_d.
Then add the Lagrange multiplier to the terms in ELBO-15 containing $\omega_p, \forall p = 1, 2, \ldots, P$,

$$\begin{aligned}\mathcal{L}_{[\omega_p]} \quad = \quad & \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\omega_p\phi_{pdni}(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj})) \\ & + \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\omega_p\phi_{pdni}\log\beta_{iv_d} \\ & - \lambda\sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\sum_{p'=1}^{P}\omega_{p'}\phi_{p'dni}\log(\sum_{p'=1}^{P}\omega_{p'}\phi_{p'dni}) + \lambda_{\omega_{p'}}(\sum_{p'=1}^{P}\omega_{p'}-1)\end{aligned} \tag{16}$$

Thus for any $p = 1, 2, \ldots, P$ given $\hat{\phi}_{pndi}, \forall p = 1, 2, \ldots, P, i = 1, 2, \ldots, k, n = 1, 2, \ldots, N_d, d = 1, 2, \ldots, D$

$$\hat{\omega}_p = \underset{\omega_p}{\text{argmax}}\mathcal{L}_{[\omega_p]}$$

where $0 \leq \hat{\omega}_p \leq 1, \displaystyle\sum_{p'=1}^{P}\omega_{p'} = 1, \forall p = 1, 2, \ldots, P$. Based on the expression of $\mathcal{L}_{[\omega_p]}$ above,

$$\begin{aligned}\hat{\omega}_p \quad \propto \quad & \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\left(\phi_{pdni}(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj})) + \phi_{pdni}\log\beta_{iv_d}\right) \\ = \quad & \frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\left(\phi_{pdni}(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj}))+\phi_{pdni}\log\beta_{iv_d}\right)}{\sum_{p'=1}^{P}\sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\left(\phi_{p'dni}(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj}))+\phi_{p'dni}\log\beta_{iv_d}\right)}\end{aligned} \tag{17}$$

Take first derivative in terms of ω_p, then:

$$\begin{aligned}\frac{\partial \mathcal{L}_{[\omega_p]}}{\partial \omega_p} &= \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{i=1}^{k}(\phi_{pdni}(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k}\gamma_{dj})) \\ &+\sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\phi_{pdni}\log\beta_{iv_d} \\ &-\lambda\sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\phi_{pdni}\log(\sum_{p'=1}^{P}\omega_{p'}\phi_{p'dni}) - \lambda\sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\phi_{pdni} + \lambda_{\omega_{p'}}\end{aligned} \tag{18}$$

Last, we maximize the equation-10 with respect to γ_{di}, the terms containing γ_{di} are:

$$\begin{aligned}\mathcal{L}_{d[\gamma_{di}]} &= (\alpha_i - 1)\Psi(\gamma_{di}) - \Psi(\textstyle\sum_{j=1}^{k}\gamma_{dj})\displaystyle\sum_{j=1}^{k}(\alpha_j - 1) + \sum_{n=1}^{N_d}\sum_{p=1}^{P}\omega_p\phi_{pdni}(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k}\gamma_{dj})) \\ &\quad -\Psi(\textstyle\sum_{j=1}^{k}\gamma_j) + \displaystyle\sum_{i=1}^{k}\log\Gamma(\gamma_{di}) - (\gamma_{di} - 1)\Psi(\gamma_{di}) + \Psi(\sum_{j=1}^{k}\gamma_{dj})\sum_{j=1}^{k}(\gamma_{dj} - 1)\end{aligned} \tag{19}$$

Take first derivative in terms of γ_{di}:

$$\frac{\partial \mathcal{L}_{d[\gamma_{di}]}}{\partial \gamma_{di}} = (\alpha_i + \sum_{n=1}^{N_d}\sum_{p=1}^{P}\omega_p\phi_{pdni} - \gamma_{di})\Psi'(\gamma_{di}) - \Psi'(\sum_{i=1}^{k}\gamma_{di})\sum_{j=1}^{k}(\alpha_j + \sum_{n=1}^{N_d}\sum_{p=1}^{P}\omega_p\phi_{pdni} - \gamma_{dj})$$

Setting the above first derivative equal to 0 then the value of estimated $\hat{\gamma}_{di}$ maximizing the ELBO-7 is:

$$\hat{\gamma}_{di} = \alpha_i + \sum_{n=1}^{N_d}\sum_{p=1}^{P}\omega_p\phi_{pdni}$$

5.4 M step with multiple particles

In M step with multiple particles-$P > 1$, with all the documents-$\boldsymbol{w}_d, d = 1, \ldots, D$, maximize the ELBO-15 with respect to model parameters-$\boldsymbol{\alpha}, \boldsymbol{\beta}$.

Similarly to Blei et al. 2003, choose the terms related to β and add Lagrange multiplier,

$$\mathcal{L}_{[\beta]} = \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\sum_{j=1}^{V}\sum_{p=1}^{P}\omega p\phi_{pdni}w_{dn}^{j}\log\beta_{ij} + \sum_{i=1}^{k}\lambda_{\beta i}(\sum_{j=1}^{V}\beta_{ij} - 1)$$

Take the first derivative in terms of β_{ij}, set it to zero, then:

$$\hat{\beta}_{ij} = \frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{p=1}^{P}\omega_p\phi_{pdni}w_{dn}^{j}}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{j'=1}^{V}\sum_{p=1}^{P}\omega_p\phi_{pdni}w_{dn}^{j'}}$$

And also choose the terms containing α:

$$\mathcal{L}_{[\alpha]} = \sum_{d=1}^{D}\left(\log\Gamma(\sum_{j=1}^{k}\alpha_j) + \sum_{i=1}^{k}(\alpha_i - 1)(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k}\gamma_{dj}))\right)$$

Take the first derivative in terms of $\alpha_i, \forall i = 1, 2, \ldots, k$, then:

$$\frac{\partial \mathcal{L}}{\partial \alpha_i} = D(\Psi(\sum_{j=1}^{k} \alpha_j) - \Psi(\alpha_i)) + \sum_{d=1}^{D} (\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj}))$$

Set it to zero, then $\forall i \neq j = 1, 2, \ldots, p,$:

$$\frac{\partial \mathcal{L}}{\partial \alpha_i \alpha_j} = \delta(i, j) D \Psi'(\alpha_i) - \Psi'(\sum_{j=1}^{k} \alpha_j)$$

Apply the Newton-Raphson method for a Hessian with special structure, denote $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_k]^T$, then by iterating the equation below we can find the maximal-$\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha}_{(new)} = \boldsymbol{\alpha}_{(old)} - H'(\boldsymbol{\alpha}_{(old)})^{-1} g'(\boldsymbol{\alpha}_{(old)})$$

where $H'(\boldsymbol{\alpha}), g'(\boldsymbol{\alpha})$ are the Hessian matrix and gradient respectively at the point-$\boldsymbol{\alpha}$ defined above.
And the Hessian matrix $H'(\boldsymbol{\alpha})$ with the special form:

$$H'(\boldsymbol{\alpha}) = diag(\boldsymbol{h}) - \Psi'(\sum_{j=1}^{k} \alpha_j) \mathbf{1}\mathbf{1}^T$$

where $\boldsymbol{h} = [D\Psi'(\alpha_1), D\Psi'(\alpha_2), \ldots, D\Psi'(\alpha_k)]^T$

And inverse of Hessian matrix $H'(\boldsymbol{\alpha})$ can be expressed as:

$$H'(\boldsymbol{\alpha})^{-1} = diag(\boldsymbol{h})^{-1} + \frac{D diag(\boldsymbol{h})^{-1} \mathbf{1}\mathbf{1}^T diag(\boldsymbol{h})^{-1}}{D(\Psi'(\sum_{j=1}^{k} \alpha_j))^{-1} - \sum_{j=1}^{k} (\Psi'(\alpha_j))^{-1}}$$