Particle EM algorithm in Latent Dirichlet Allocation model

Yubing Yao
Department of Biostatistics and Epidemiology,
University of Massachussetts, Amherst
yyao@umass.edu

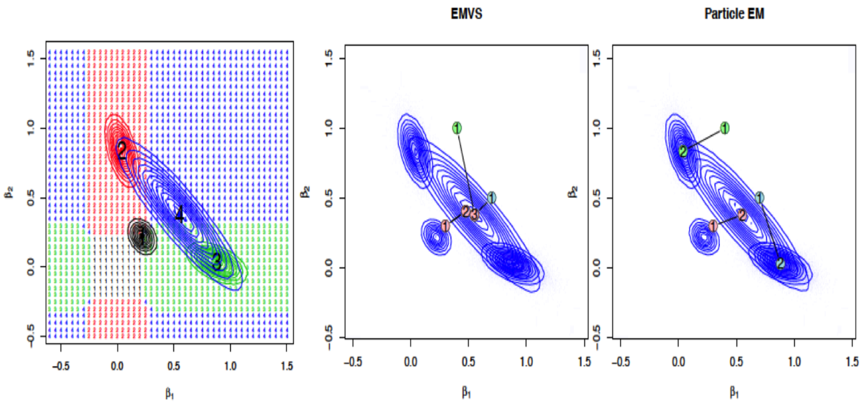
Abstract

We develop a particle EM algorithm under the framework of Latent Dirichlet Allocation (LDA) model for text documents, and explore the global mode of evidence lower bound (ELBO) in Bayesian variational inference with penalized entropy for topic probabilities through generating multiple repulsive particles to explore the search space to identify the comprehensive set of modes.

1 Introduction

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 is a population-based optimization method, and aims to overcome the vulnerability of local entrapment of the traditional EM algorithm when dealing with the multi-modal posterior/likelihood. The particle EM algorithm explores the whole search space by multiple repulsive particles and tries to capture the multiple modes of posterior distribution.

The difference between EM and Particle EM algorithms can be exemplified in a simple illustrative the example below, assume we have two predictors-$\{X_1, X_2\}$ with the corresponding parameters-$\{\beta_1, \beta_2\}$ with high correlation-0.9, then assume posterior marginal distribution-$\pi(\boldsymbol{\beta}|\boldsymbol{Y})$ with four modes with mode 3 global mode-$\{\beta_1 = 1, \beta_2 = 0\}$, we can observe in EM algorithm with multiple particles they tend to converge to one strong local mode 4 while in Particle EM algorithm we force multiple particles to separate by penalizing the entropy term and thus are possible to detect global mode-mode 3.



1.1 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-$\mathcal{D} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_M\}$.
- A document has a sequence of N words denoting $\boldsymbol{w} = \{w_1, w_2, \dots, w_N\}$.
- A word defined as an item from a vocabulary indexed by $\{1, 2, \dots, V\}$. 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 $\beta_{ij} = p(w^j = 1|z^i = 1), j = 1, \dots, V, i = 1, 2, \dots, k$, and $w^j_{n_d} = 1$ of jth component of the word $w_{n_d}, j = 1, \dots, V, n_d = 1, \dots, N_d, d = 1, 2, \dots, D$.
Given the parameter-α and β, for a corpus-$\mathcal{D}$ with $d = 1, \dots, D$ 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:

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

1.2 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:

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

Right-hand side of the inequality4 defined in Appendix 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 $\mathcal{L}(\gamma, \phi, \alpha, \beta)$,

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

2 Particle EM algorithm in LDA model

2.1 Particle approximation and Evidence Lower Bound

Motivated by the particle approximation in Rockova V. 2016, for $d = 1, 2, \dots, D$, we use a weighted mixture of atoms to approximate $\pi(\boldsymbol{z}_{dn}|\mathcal{D}), n = 1, 2, \dots, N_d, d = 1, 2, \dots, D$, 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 $\boldsymbol{\Gamma}_{dn} = [\boldsymbol{z}_{1dn}, \boldsymbol{z}_{2dn}, \dots, \boldsymbol{z}_{Pdn}]$,corresponding importance weights-$\boldsymbol{\omega} = (\omega_1, \omega_2, \dots, \omega_P)^T$, where $\sum_{p=1}^{P} \omega_p = 1, \forall d = 1, \dots, D, n = 1, 2, \dots, N_d; 0 \leq \omega_p \leq 1, \forall p = 1, 2, \dots, P$.
For $\boldsymbol{z}_{pdn}, \forall p = 1, 2, \dots, P, n = 1, 2, \dots, N_d, d = 1, 2, \dots, D$, z_{pdn} can only takes one of $1, 2, \dots, k$ values, representing the possible k topics in LDA model.

And denote $\boldsymbol{z}^i_{pdn}$ will only can take the $i = 1, 2, \dots, k$ 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 $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \dots, \boldsymbol{\gamma}_D], \boldsymbol{\omega} = [\omega_1, \omega_2, \dots, \omega_P]$.

Then the evidence lower bound-1 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 :

$$\begin{aligned} \mathcal{L}_\lambda(\boldsymbol{\gamma}, \boldsymbol{\Gamma}, \boldsymbol{\omega}; \boldsymbol{\theta}, \alpha, \beta) \quad = \quad & \sum_{d=1}^{D} (\mathrm{E}_{q_{PEM}}[\log p(\theta_d|\alpha)] + \mathrm{E}_{q_{PEM}}[\log p(\boldsymbol{z}_d|\theta_d)] + \mathrm{E}_{q_{PEM}}[\log p(\boldsymbol{w}_d|\boldsymbol{z}_d, \beta)] \\ & -\mathrm{E}_q[\log q(\theta_d|\gamma_d)] - \lambda \mathrm{E}_{q_{PEM}}[\log q_{PEM}(\boldsymbol{z}_{n_d}|\boldsymbol{\Gamma}_d, \boldsymbol{\omega}_d)]) \end{aligned} \tag{2}$$

where $\lambda \geq 0$.
When $\lambda = 1$, the above is regular ELBO from variational inference. if $\lambda = 0$, it is equivalent to parallel EM.

2.2 Particle EM-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-$\boldsymbol{w}_d$.

Denote $\beta_{iv_d} = p(w_n^{v_d} = 1|z_d^i = 1)$ for the appropriate v_d, the value of estimated $\hat{\phi}_{dni}$ maximizing the ELBO 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, the value of estimated $\hat{\gamma}_{di}$ maximizing the ELBO is:

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

2.3 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 with respect to model parameters-$\boldsymbol{\alpha}, \boldsymbol{\beta}$.

Similarly to Blei et al. 2003, then estimated value of model parameter-$\boldsymbol{\beta}$ is:

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

The estimation of $\alpha_i, i = 1, 2, \ldots, k$ by maximizing the ELBO can be solved numerically by Newton-Raphson method.

2.4 E step with multiple particles

With $P > 1$, we need alternatively update 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-$(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_P)^T, \forall d = 1, \ldots, D; n = 1, 2, \ldots, N_d$.

Apply 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.

The value of estimated importance weight-$\hat{\omega}_p, \forall p = 1, 2, \ldots, P$ is:

$$\hat{\omega}_p \quad = \quad \frac{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \big(\phi_{pdni}(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})) + \phi_{pdni} \log \beta_{iv_d}\big)}{\sum_{p'=1}^{P} \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \big(\phi_{p'dni}(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})) + \phi_{p'dni} \log \beta_{iv_d}\big)} \tag{3}$$

The value of estimated $\hat{\gamma}_{di}$ maximizing the ELBO is:

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

2.5 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 with respect to model parameters-$\boldsymbol{\alpha}, \boldsymbol{\beta}$.

Similarly to Blei et al. 2003, then estimated value of model parameter-$\boldsymbol{\beta}$ is:

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

Similarly to the estimator at M step with single particle, estimation of model parameter-$\boldsymbol{\alpha}$ is from Newton-Raphson method for a Hessian with special structure.

After randomly initialize the values of model parameters- $\boldsymbol{\beta}^{(0)}, \boldsymbol{\alpha}^{(0)}$, we recursively iterate the estimators of variational parameters - $\boldsymbol{\phi}_{pnd}, \boldsymbol{\omega}, \boldsymbol{\gamma}_d$ until convergence at E step and also recursively iterate model parameters $\boldsymbol{\beta}, \boldsymbol{\alpha}$ until convergence at M step.

3 Appendix

3.1 Variational Inference in LDA model

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

$$\begin{aligned}
\log p(\boldsymbol{w}_d|\alpha, \beta) &= \log \int \sum_{\boldsymbol{z}_d} p(\theta_d, \boldsymbol{z}_d, \boldsymbol{w}_d|\alpha, \beta) d\theta_d \\
&= \log \int \sum_{\boldsymbol{z}_d} \frac{p(\theta_d, \boldsymbol{z}_d, \boldsymbol{w}_d|\alpha, \beta) q(\theta_d, \boldsymbol{z}_d|\gamma, \phi)}{q(\theta_d, \boldsymbol{z}_d|\gamma, \phi)} d\theta_d \\
&\geq \int \sum_{\boldsymbol{z}_d} q(\theta_d, \boldsymbol{z}_d|\gamma, \phi) \log p(\theta_d, \boldsymbol{z}_d, \boldsymbol{w}_d|\alpha, \beta) - \int \sum_{\boldsymbol{z}_d} q(\theta_d, \boldsymbol{z}_d|\gamma, \phi) \log q(\theta_d, \boldsymbol{z}_d|\gamma, \phi) d\theta_d \\
& \quad = \mathrm{E}_q[p(\theta, \boldsymbol{z}_d, \boldsymbol{w}_d|\alpha, \beta)] - \mathrm{E}_q[q(\theta_d, \boldsymbol{z}_d|\gamma, \phi)] \\
& \quad = \mathrm{E}_q[p(\theta, \boldsymbol{z}_d, \boldsymbol{w}_d|\alpha, \beta)] + H(\gamma, \phi)
\end{aligned} \tag{4}$$

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

3.2 Parameter estimation in Particle EM on LDA model

3.2.1 Variational parameter estimation at E step of Particle EM with single particle

Before entirely introduce the Particle EM algorithm in LDA model, firstly we assume single particle-$P = 1$ and fix some document-$\boldsymbol{w}_d$, and denote $\phi_{dni} = \mathrm{E}(z^i_{dn}) = P(z^i_{dn} = 1), \forall n = 1, 2, \ldots, N_d, i = 1, 2, \ldots, k$.

Under $P = 1$ the entropy term becomes:

$$\begin{aligned}
-\mathrm{E}_q[\log q(\theta_d|\boldsymbol{\gamma}_d)] + H_\lambda(\boldsymbol{\Gamma}_d, \boldsymbol{\omega}_d) &= -\Psi(\textstyle\sum_{j=1}^k \gamma_j) + \displaystyle\sum_{i=1}^k \log \Gamma(\gamma_i) - \sum_{i=1}^k (\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^k \gamma_j)) \\
&\quad + H_\lambda(\boldsymbol{\Gamma}_d, \boldsymbol{\omega}_d)
\end{aligned} \tag{5}$$

Then with single particle $P = 1$ and fix some $d, \boldsymbol{w}_d$ the evidence lower bound-2 becomes:

$$\begin{aligned}
\mathcal{L}_{d\lambda}(\boldsymbol{\gamma}_d, \boldsymbol{\Gamma}, \boldsymbol{\omega}; \boldsymbol{\theta}, \alpha, \beta) &= \log \Gamma(\textstyle\sum_{j=1}^k \alpha_j) + \displaystyle\sum_{i=1}^k (\alpha_i - 1)(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^k \gamma_{dj})) \\
&\quad + \sum_{n=1}^{N_d} \sum_{i=1}^k \phi_{dni}(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^k \gamma_{dj})) \\
&\quad + \sum_{n=1}^{N_d} \sum_{i=1}^k \sum_{j=1}^V \phi_{dni} w^j_n \log \beta_{ij} \\
&\quad - \Psi(\textstyle\sum_{j=1}^k \gamma_{dj}) + \displaystyle\sum_{i=1}^k \log \Gamma(\gamma_{di}) - \sum_{i=1}^k (\gamma_{di} - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^k \gamma_{dj})) \\
&\quad + H_\lambda(\boldsymbol{\Gamma}_d, \boldsymbol{\omega}_d)
\end{aligned} \tag{6}$$

Firstly maximize the ELBO-6 with respect to ϕ_{dni} with the constraint-$\sum_{j=1}^k \phi_{dnj} = 1$. Denote $\beta_{iv_d} = p(w^{v_d}_n = 1|z^i_d = 1)$ for the appropriate v_d. Then add the Lagrange multiplier to the terms in ELBO-6 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-6 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-6 with respect to γ_{di}, the terms containing γ_i are:

$$\begin{aligned} \mathcal{L}_{d[\gamma_{di}]} \quad = \quad & (\alpha_i - 1)\Psi(\gamma_{di}) - \Psi(\textstyle\sum_{j=1}^{k} \gamma_{dj}) \displaystyle\sum_{j=1}^{k} (\alpha_j - 1) + \sum_{n=1}^{N_d} \phi_{dni}(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})) \\ & -\Psi(\textstyle\sum_{j=1}^{k} \gamma_j) + \displaystyle\sum_{i=1}^{k} \log \Gamma(\gamma_{di}) - (\gamma_{di} - 1)\Psi(\gamma_{di}) + \Psi(\sum_{j=1}^{k} \gamma_{dj}) \sum_{j=1}^{k} (\gamma_{dj} - 1) \end{aligned} \tag{7}$$

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-6 is:

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

3.2.2 Model parameter estimation at M step of Particle EM 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-6 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 $\forall i \neq j = 1, 2, \ldots, k,$:

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

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-$\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha}_{(new)} = \boldsymbol{\alpha}_{(old)} - H_1(\boldsymbol{\alpha}_{(old)})^{-1} g_1(\boldsymbol{\alpha}_{(old)})$$

where $H_1(\boldsymbol{\alpha}), g_1(\boldsymbol{\alpha})$ are the Hessian matrix and gradient respectively at the point-$\boldsymbol{\alpha}$ defined above. And the Hessian matrix $H_1(\boldsymbol{\alpha})$ with the special form:

$$H_1(\boldsymbol{\alpha}) = diag(\boldsymbol{h}) - \Psi'(\sum_{j=1}^{k} \alpha_j)\mathbf{1}\mathbf{1}^T$$

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

And inverse of Hessian matrix $H_1(\boldsymbol{\alpha})$ can be expressed as:

$$H_1(\boldsymbol{\alpha})^{-1} = diag(\boldsymbol{h})^{-1} + \frac{D diag(\boldsymbol{h})^{-1}\mathbf{1}\mathbf{1}^T diag(\boldsymbol{h})^{-1}}{D(\Psi'(\sum_{j=1}^{k}\alpha_j))^{-1} - \sum_{j=1}^{k}(\Psi'(\alpha_j))^{-1}}$$

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

Given $\boldsymbol{\omega}^{(m)}$ fix some d then the evidence lower bound-2 becomes:

$$\begin{aligned}
\mathcal{L}_{d\lambda}(\boldsymbol{\gamma}_d, \boldsymbol{\Gamma}, \boldsymbol{\omega}; \boldsymbol{\theta}, \alpha, \beta) &= \log\Gamma(\textstyle\sum_{j=1}^{k}\alpha_j) + \displaystyle\sum_{i=1}^{k}(\alpha_i - 1)(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k}\gamma_{dj})) \\
&+ \sum_{p=1}^{P}\sum_{n=1}^{N_d}\omega_p^{(m)}\sum_{i=1}^{k}\phi_{pdni}(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k}\gamma_{dj})) \\
&+ \sum_{p=1}^{P}\sum_{n=1}^{N_d}\omega_p^{(m)}\sum_{i=1}^{k}\sum_{j=1}^{V}\phi_{pdni}w_n^j\log\beta_{ij} \\
&- \Psi(\textstyle\sum_{j=1}^{k}\gamma_{dj}) + \displaystyle\sum_{i=1}^{k}\log\Gamma(\gamma_{di}) - \sum_{i=1}^{k}(\gamma_{di} - 1)(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k}\gamma_{dj})) \\
&- \lambda\sum_{p=1}^{P}\sum_{n=1}^{N_d}\omega_p\sum_{i=1}^{k}\phi_{pdni}\log(\sum_{p=1}^{P}\omega_p\phi_{pdni})
\end{aligned} \tag{8}$$

where $\phi_{pdni} = \mathrm{E}(z_{pdn}^i) = P(z_{pdn}^i = 1), \forall p = 1, 2, \ldots, P; d = 1, 2, \ldots, D, n = 1, 2, \ldots, N_d, i = 1, 2, \ldots, k$.
Similar to one particle system,firstly maximize the ELBO-8 with respect to ϕ_{pdni} with the constraint-$\sum_{j=1}^{k}\phi_{pdnj} = 1$. Denote $\beta_{iv_d} = p(w_n^{v_d} = 1|z_{pd}^i = 1)$ for the appropriate v_d.
Then add the Lagrange multiplier to the terms in ELBO-8 containing ϕ_{pdni},

$$\begin{aligned}
\mathcal{L}_{d[\phi_{pdni}]} &= \omega_p\phi_{pdni}(\Psi(\gamma_{di}) - \Psi(\textstyle\sum_{j=1}^{k}\gamma_{dj})) + \omega_p\phi_{pdni}\log\beta_{iv_d} \\
&- \lambda\displaystyle\sum_{p'=1}^{P}\omega_{p'}\phi_{p'dni}\log(\sum_{p'=1}^{P}\omega_{p'}\phi_{p'dni}) + \lambda_{\phi_p}(\sum_{j=1}^{k}\phi_{pdnj} - 1)
\end{aligned} \tag{9}$$

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 $\displaystyle\sum_{i=1}^{k} \phi_{pdni} = 1$ 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-$p, n, d$, $\displaystyle\sum_{i=1}^{k} \phi_{pdni} = 1$, 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-$\displaystyle\sum_{i=1}^{k} \phi_{pdni} = 1$, we have total $P(k-1)\sum_{d=1}^{D} N_d$ free parameters with $\phi_{pndk} = 1 - \displaystyle\sum_{i'=1}^{k-1} \frac{e^{x_{pdni'}}}{1+e^{x_{pdni'}}}$.

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

$$\begin{aligned} \frac{\partial \mathcal{L}_d}{\partial x_{pdni'}} = \frac{\partial \mathcal{L}_d}{\partial \phi_{pdni'}} \frac{\partial \phi_{pdni'}}{\partial x_{pdni'}} \quad = \quad & \Big(\omega_p(\Psi(\gamma_{di}) - \Psi(\textstyle\sum_{j=1}^{k} \gamma_{dj})) + \omega_p \log \beta_{iv_d} - \lambda\omega_p(\log(\textstyle\sum_{p'=1}^{P} \omega_{p'} \frac{e^{x_{pdni'}}}{1+e^{x_{pdni'}}}) + 1)\Big) \\ & \times \tfrac{e^{x_{pdni'}}}{(1+e^{x_{pdni'}})^2} \end{aligned} \tag{10}$$

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

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}_d}{\partial x_{pdni'}^2} &= \frac{\partial}{\partial x_{pdni'}}\left(\frac{\partial \mathcal{L}_d}{\partial x_{pdni'}}\right) = \frac{\partial}{\partial x_{pdni'}}\left(\frac{\partial \mathcal{L}_d}{\partial \phi_{pdni'}}\frac{\partial \phi_{pdni'}}{\partial x_{pdni'}}\right) \\
&= \frac{\partial \mathcal{L}_d}{\partial \phi_{pdni'}}\frac{\partial^2 \phi_{pdni'}}{\partial x_{pdni'}^2} + \frac{\partial^2 \mathcal{L}_d}{\partial \phi_{pdni'} \partial x_{pdni'}}\frac{\partial \phi_{pdni'}}{\partial x_{pdni'}} \\
&= \frac{\partial \mathcal{L}_d}{\partial \phi_{pdni'}}\frac{\partial^2 \phi_{pdni'}}{\partial x_{pdni'}^2} + \frac{\partial^2 \mathcal{L}_d}{\partial \phi_{pdni'}^2}(\frac{\partial \phi_{pdni'}}{\partial x_{pdni'}})^2 \\
&= \frac{\partial \mathcal{L}_d}{\partial \phi_{pdni'}}\frac{\partial \phi_{pdni'}}{\partial x_{pdni'}} \\
&= \Big(\omega_p(\Psi(\gamma_{di}) - \Psi(\textstyle\sum_{j=1}^{k}\gamma_{dj})) + \omega_p \log \beta_{iv_d} - \lambda\omega_p(\log(\textstyle\sum_{p'=1}^{P}\omega_{p'}\frac{e^{x_{p'dni'}}}{1+e^{x_{p'dni'}}}) + 1)\Big) \\
&\quad \times \tfrac{e^{x_{pdni'}}}{(1+e^{x_{pdni'}})^2}(1 - \tfrac{e^{x_{pdni'}}}{1+e^{x_{pdni'}}}) - \frac{\lambda\omega_p^2}{\sum_{p'=1}^{P}\omega_{p'}\frac{e^{x_{pdni'}}}{1+e^{x_{pdni'}}}} * \tfrac{e^{2x_{pdni'}}}{(1+e^{x_{pdni'}})^4}
\end{aligned} \tag{11}$$

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

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}_d}{\partial x_{pdni'}\partial x_{p''dni'}} &= \frac{\partial}{\partial x_{p''dni'}}\left(\frac{\partial \mathcal{L}_d}{\partial x_{pdni'}}\right) = \frac{\partial}{\partial x_{p''dni'}}\left(\frac{\partial \mathcal{L}_d}{\partial \phi_{pdni'}}\frac{\partial \phi_{pdni'}}{\partial x_{pdni'}}\right) \\
&= \frac{\partial \mathcal{L}_d}{\partial \phi_{pdni'}}\frac{\partial^2 \phi_{pdni'}}{\partial x_{pdni'}\partial x_{p''dni'}} + \frac{\partial^2 \mathcal{L}_d}{\partial \phi_{pdni'}\partial x_{p''dni'}}\frac{\partial \phi_{pdni'}}{\partial x_{pdni'}} \\
&= \frac{\partial \mathcal{L}_d}{\partial \phi_{pdni'}} \times 0 + \frac{\partial^2 \mathcal{L}_d}{\partial \phi_{pdni'}\partial \phi_{p''dni'}}\frac{\partial \phi_{pdni'}}{\partial x_{pdni'}}\frac{\partial \phi_{p''dni'}}{\partial x_{p''dni'}} \\
&= \frac{\partial^2 \mathcal{L}_d}{\partial \phi_{pdni'}\partial \phi_{p''dni'}}\frac{\partial \phi_{pdni'}}{\partial x_{pdni'}}\frac{\partial \phi_{p''dni'}}{\partial x_{p''dni'}} \\
&= -\frac{\lambda\omega_p\omega_{p''}}{\sum_{p'=1}^{P}\omega_{p'}\frac{e^{x_{p'dni'}}}{1+e^{x_{p'dni'}}}}\frac{e^{x_{pdni'}}}{(1+e^{x_{pdni'}})^2}\frac{e^{x_{p''dni'}}}{(1+e^{x_{p''dni'}})^2}
\end{aligned} \tag{12}$$

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 $\sum_{p'=1}^{P}\omega_{p'} = 1$, 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-8 can be written as:

$$
\begin{aligned}
\mathcal{L}_\lambda(\boldsymbol{\gamma}, \boldsymbol{\Gamma}, \boldsymbol{\omega}; \boldsymbol{\theta}, \alpha, \beta) \quad = \quad & D \log \Gamma(\textstyle\sum_{j=1}^{k} \alpha_j) + \displaystyle\sum_{d=1}^{D} \sum_{i=1}^{k} (\alpha_i - 1)(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})) \\
& + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \sum_{l=1}^{P} \omega_p \phi_{pdni} (\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})) \\
& + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \sum_{j=1}^{V} \sum_{p=1}^{P} \omega_p \phi_{pdni} w_n^j \log \beta_{ij} \\
& - \sum_{d=1}^{D} \left(\Psi(\sum_{j=1}^{k} \gamma_{dj}) + \sum_{i=1}^{k} \log \Gamma(\gamma_{di}) - \sum_{i=1}^{k} (\gamma_{di} - 1)(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})) \right) \\
& - \lambda \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \sum_{p=1}^{P} \omega_p \phi_{pdni} \log(\sum_{p=1}^{P} \omega_p \phi_{pdni})
\end{aligned} \tag{13}
$$

Maximize the ELBO-13 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-13 containing $\omega_p, \forall p = 1, 2, \ldots, P$,

$$
\begin{aligned}
\mathcal{L}_{[\omega_p]} \quad = \quad & \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \omega_p \phi_{pdni} (\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})) \\
& + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \omega_p \phi_{pdni} \log \beta_{iv_d} \\
& - \lambda \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \sum_{p'=1}^{P} \omega_{p'} \phi_{p'dni} \log(\sum_{p'=1}^{P} \omega_{p'} \phi_{p'dni}) + \lambda_{\omega_{p'}} (\sum_{p'=1}^{P} \omega_{p'} - 1)
\end{aligned} \tag{14}
$$

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

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

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,

$$
\begin{aligned}
\hat{\omega}_p \quad & \propto \quad \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \left(\phi_{pdni} (\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})) + \phi_{pdni} \log \beta_{iv_d} \right) \\
& = \quad \frac{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \left(\phi_{pdni} (\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})) + \phi_{pdni} \log \beta_{iv_d} \right)}{\sum_{p'=1}^{P} \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \left(\phi_{p'dni} (\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})) + \phi_{p'dni} \log \beta_{iv_d} \right)}
\end{aligned} \tag{15}
$$

Last, we maximize the equation-8 with respect to γ_{di}, the terms containing γ_{di} are:

$$
\begin{aligned}
\mathcal{L}_{d[\gamma_{di}]} \quad = \quad & (\alpha_i - 1)\Psi(\gamma_{di}) - \Psi(\textstyle\sum_{j=1}^{k} \gamma_{dj}) \displaystyle\sum_{j=1}^{k} (\alpha_j - 1) + \sum_{n=1}^{N_d} \sum_{p=1}^{P} \omega_p \phi_{pdni} (\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})) \\
& - \Psi(\textstyle\sum_{j=1}^{k} \gamma_j) + \displaystyle\sum_{i=1}^{k} \log \Gamma(\gamma_{di}) - (\gamma_{di} - 1)\Psi(\gamma_{di}) + \Psi(\sum_{j=1}^{k} \gamma_{dj}) \sum_{j=1}^{k} (\gamma_{dj} - 1)
\end{aligned} \tag{16}
$$

Take first derivative in terms of γ_{di}:

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

Set the above first derivative equal to 0 then the value of estimated $\hat{\gamma}_{di}$ maximizing the ELBO-13 is:

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

3.2.4 Variational parameter estimation at M step of Particle EM with multiple particles

n M step with multiple particles-$P > 1$, with all the documents-$\boldsymbol{w}_d, d = 1, \ldots, D$, maximize the ELBO-13 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 $\forall i \neq j = 1, 2, \ldots, p,$:

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

Similar as estimation of model parameter-$\boldsymbol{\alpha}$ at M step with single particle, apply the Newton-Raphson method for a Hessian with special structure with exact same expression.

4 Future work

4.1 Comparison with other variational inference methods in twitter-LDA data

We are now developing R and C++ code for particle EM algorithm in LDA model, and analyze some real dataset with our algorithm.

The dataset we are going to use is the Twitter-LDA data, which is available at minghui's Github, https://github.com/minghui/Twitter-LDA. The Twitter-LDA datasets (W. X. Zhao et al. 2011) contain 12 files, each file is all tweets of a single user, roughly each file contains 400-700 lines of records. The Twitter-LDA data are short and noisy, compared with common corpora. They also provided the result of their own Twitter-LDA(T-LDA) results, which they claimed work well with twitter data. So we can compare our particle EM algorithm with their T-LDA results and also with variational inference method in Blei 2003 et. al, nonparametric variational inference (Gershman et.al. 2012), stochastic variational inference method (Hoffman et. al. 2013) in LDA model.

4.1.1 Stochastic Variational Inference in LDA

Stochastic variational inference optimizes object function $f(\lambda)$ by introducing noisy natural gradients. Specifically in LDA, the ρ size of steps can be regarded as a proportional weight between the updated estimate of λ at step $t+1$ and the previous estimate of λ at step t, denoted by $\hat{\lambda}_t = (1-\rho_t)\hat{\lambda}_{t-1} + \rho_t\hat{\lambda}$. In the stochastic variational inference frame work, the parameter β is the global parameter and governed by a variational parameter λ, which is also a global parameter for each topic in vocabulary; the document d, word $w_{d,1:N}$, topic parameter $\theta - d$ and topic assignment $z_{d,1:N}$ are all local. As the result of the parameters' setting, one of the differences between the classical variational inference in LDA and stochastic variational inference in LDA is the additional V -dimensional variational distribution for each topic in vocabulary, which can be given by:

$$q(\beta_k) = Dir(\lambda_k)$$

By the good properties of Dirichlet distribution, we can have following complete conditionals:

$$\begin{aligned}
p(z_{dn} &= k|\theta_d, \beta_{1:K}, w_{dn}) \propto \exp\{\log\theta_{dk} + \log\beta_{k,w_{dn}}\} \\
p(\theta_d|z_d) &= Dir(\alpha + \sum_{m=1}^{N} z_{dn}) \\
p(\beta_k|z,w) &= Dir(\eta + \sum_{d=1}^{D}\sum_{n_d=1}^{N} z_{dn_d}^k w_{dn_d}).
\end{aligned}$$

We can easily get following updates equations:

$$\begin{aligned}
\phi_{dn}^k &\propto exp\{\Psi(\gamma_{dk} + \Psi(\lambda_{k,w_{dn}}) - \Psi(\sum_v \lambda_{kv}))\} \\
\gamma_d &= \alpha + \sum_n = 1^N \phi_{dn} \\
\lambda_k &= \eta + \sum_{n_d=1}^{N} \phi_{dn}^k w_{dn_d}
\end{aligned}$$

To give a better understanding of how the stochastic variational inference works in LDA, Hoffman (2013) provided pesudo-codes:

```
Initialize $\lambda^{(0)}$ randomly.
Set the step-size schedule $\rho_t$ appropriately
repeat
  Sample a document $w_d$ uniformly from the data set.
  Initialize $\gamma_{dk} = 1$, for $k \in \{1, \dots, K\}$.
  repeat
    For $n \in \{1, \dots, N\}$ set
        $\phi_{dn}^k \propto \exp\{\mathbb{E}[\log\theta_{dk}] + [\log\beta_{k,w_{dn}}]\}, k \in \{1, \dots, K\}$.
    Set $\gamma_d = \alpha + \sum_n \phi_{dn}$.
  until local parameters $\phi_{dn}$ and $\gamma_d$ converge.
  $k \in \{1, \dots, K\}$ set intermediate topics
        $\hat{\lambda}_k = \eta + D\sum_{n=1}^{N} \phi_{dn_d}^k w_{dn_d}$.
  $\lambda^{(t)} = (1-\rho_t)\lambda^{(t-1)} + \rho_t\hat{\lambda}$
  until forever
```

It is worth to mention that, because there is only one document w_d being sampled in each repeat, stochastic variational inference will be much more efficient in computation as compared with other variational inference methods.

4.1.2 Nonparametric Variational Inference in LDA

In 2012, Samuel J. Gershman introduced nonparametric variational inference for the model with continuous-valued hidden random variables and non-conjugacy between pairs of variables. Benefited from the less requirement for the conjugacy requirement and advantages from kernel approximation, the nonparametric variational inference can handle models with multimodal posterior and non-conjugate distribution. Inspired by such good properties of continuous version nonparametric variational inference, we were trying to explore its implementation in discrete case—LDA.
We proposed to take the $logit$ transformation of variational parameter ϕ,

$$\eta_{nd} = logit(\phi_{nd}),$$

such that a discrete $[0, 1]$ variable can be transformed to a continuous $[-\infty, \infty]$ variable and apply similar approximation in continuous nonparametric variational inference. and use second Gaussian kernel to approximate the variational distribution of hidden variable z_{nd},

$$\begin{aligned} q(z_{nd}|\phi_{nd}) &= \prod_{i=1}^{k}[K(\eta_{ndi})]^{z^i_{nd}} \\ q(z_{nd}|\phi_{nd}) &= \prod_{i=1}^{k}[\frac{1}{\sqrt{2\pi}}\exp(-\frac{\eta^2_{ndi}}{2})]^{z^i_{nd}}, \end{aligned}$$

where z^i_{nd} is indicator variable for ith topic in document nd.
The evidence lower bound (ELBO) can be written as:

$$\begin{aligned} L(\boldsymbol{\gamma}, \boldsymbol{\eta}; \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= log\Gamma(\sum_{j=1}^{k}\alpha_j) - \sum_{i=1}^{k}\log\Gamma(\alpha_i) + \sum_{i=1}^{k}(\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)) \\ &+ \sum_{nd=1}^{Nd}\sum_{i=1}^{k}\frac{\exp(\eta_{ndi})}{1+\exp(\eta_{ndi})}(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)) \\ &+ \sum_{nd=1}^{Nd}\sum_{i=1}^{k}\sum_{j=1}^{V_d}\frac{\exp(\eta_{ndi})}{1+\exp(\eta_{ndi})}w^j_n\log\beta_{ij} \\ &- \log\Gamma(\sum_{j=1}^{k}\gamma_j) + \sum_{i=1}^{k}\log\Gamma(\gamma_i) - \sum_{i=1}^{k}(\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)) \\ &- \sum_{nd=1}^{Nd}\sum_{i=1}^{k} - \frac{1}{2\sqrt{2\pi}}\exp(-\frac{\eta^2_{nd^i}}{2})(\log(2\pi) + \eta^2_{nd^i}). \end{aligned}$$

Further derivation of the update equations are recommended in future work.

4.2 Selection of optimal λ in ELBO of Particle EM in LDA model

The penalized parameter-λ is added in entropy term of Evidence Lower Bound (ELBO)-2 to create repulsive power among the multiple particles, and the strength of this repulsive power will be determined by the value of $\lambda > 1$.

An intuitive way to select optimal value of the penalized parameter-λ is through cross-validation. With K folder cross validation we divide training and test datasets to examine a grid of possible values of $\lambda > 1$ with lowest prediction test errors in terms of posterior marginal probabilities of topics.

4.3 Stochastic Gradient Particle EM algorithm in LDA model

When dealing with massive documents, estimation of model parameters in particle EM algorithm combining all the documents will be computational inhibitive, and thus stochastic gradient descent method will replace the coordinate descent method for parameter estimation of particle EM algorithm. The detailed procedure will be very similar to section-3.2-`Stochastic Variational Inference` in Hoffman et. al. 2013 in LDA model.

4.4 Nonparametric Bayesian Topic model with hierarchical Dirichlet process (HDP)

For very large collection of documents pre-fixed number of topics is not practical. Thus random even infinite number of topics can be addressed by a Bayesian nonparametric topic model, where the document data itself decides the number of topics in topic model.

Thus Particle EM algorithm incorporating stochastic gradient descent method will be derived and implemented under the framework of Bayesian nonparametric variant of LDA where the number of topics is not fixed.

5 Reference

Particle EM for Variable Selection Rockova V. (2016) Journal of the American Statistical Association, Theory and Methods (Invited revision)

Latent Dirichlet allocation. D. Blei, A. Ng, and M. Jordan. Journal of Machine Learning Research, 3:993–1022, January 2003.

Nonparametric variational inference. S. Gershman, M. Hoffman, and D. Blei. In International Conference on Machine Learning, 2012.

Stochastic variational Inference M. Hoffman, D. Blei, C. Wang, and J. Paisley Journal of Machine Learning Research, vol. 14, pp.1303–1347, 2013.

Comparing twitter and traditional media using topic models. W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. In Proceedings of the 33rd European Conference on Advances in Information Retrieval, pages 338–349, 2011