

Topic Embedding for Documents

July 21, 2015

1 Introduction

In the previous chapter, a generative word embedding model is presented, along with a learning algorithm to find a set of word embeddings. In this chapter, we extend this model by incorporating topics of a document into this generative model, and develop a continuous counterpart of Latent Dirichlet Allocation (LDA). Through learning the latent topics, the semantics of a document will be summarized as a few topic vectors, which could be used in different applications.

2 Notations

We assume each word in a document is semantically similar to a *topic embedding* in the embedding space. We often refer to topic embeddings simply as *topics*. Specifically, each document has K candidate topics, arranged in the matrix form $\mathbf{T}_i = (\mathbf{t}_{i1} \cdots \mathbf{t}_{iK})$, referred to as the *topic matrix*. Particularly, we fix $\mathbf{t}_{i1} = \mathbf{0}$, referred to as the *null topic*. As there are many words which have no obvious semantics, these words can be assigned to this null topic. Similar to words, each topic \mathbf{t}_{ik} accompanies a residual $r_{i,k}$. In addition, there is a topic weight β , a hyperparameter controlling their degree of impact to the distribution of words.

The above assumption that each word is semantically similar to a topic, is formulated as follows. In a document d_i , each word w_{ij} is assigned to a topic indexed by $z_{ij} \in \{1, \cdots, K\}$. Geometrically this means the embedding $\mathbf{v}_{w_{ij}}$ tends to align with the direction of $\mathbf{t}_{i,z_{ij}}$. Each topic \mathbf{t}_{ik} has a document-specific prior probability to be assigned to a word, denoted as $\phi_{ik} = P(k|d_i)$. The vector $\boldsymbol{\phi}_i = (\phi_{i1}, \cdots, \phi_{iK})$ is referred to as the *mixing proportions* of these topics in document d_i . As in LDA, $\boldsymbol{\phi}_i$ is governed by a Dirichlet prior $\text{Dir}(\boldsymbol{\alpha})$.

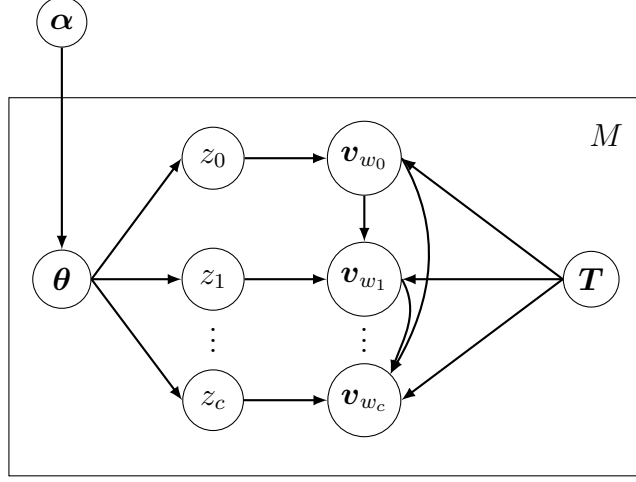


Figure 1: The Graphical Model of Topic Embedding

3 Distribution of a Text Window Parameterized by Word and Topic Embeddings

3.1 Conditional Distribution of a Word Given Context and Topic

Using the similar idea, we extend eq.(7) in [1] to incorporate the impact of the topic:

$$P(w_c \mid w_0:w_{c-1}, z_c, d_i) = P(w_c) \exp \left\{ \mathbf{v}_{w_c}^\top \left(\sum_{i=0}^{c-1} \mathbf{v}_{w_i} + \beta \mathbf{t}_{i,z_c} \right) + \sum_{i=0}^{c-1} a_{w_i w_c} + r_{i,z_c} \right\}, \quad (1)$$

where d_i is the current document, and $\beta > 0$ is a hyperparameter, named the *topic weight*, controlling their degree of impact to the distribution of w_c . The topic residual r_{i,z_c} only depends on the topic assignment z_c , but not on the value of w_c .

The topic weight β determines the “polarity” of the topics: a bigger β means that if a word is assigned to topic k , then its embedding is more strongly driven towards the direction of \mathbf{t}_{ik} . In particular, when $\beta = 0$, our model reduces to a model without topics.

This equation is equivalent to

$$\log \frac{P(w_c \mid w_0:w_{c-1}, z_c, d_i)}{P(w_c)} = \mathbf{v}_{w_c}^\top \left(\sum_{i=0}^{c-1} \mathbf{v}_{w_i} + \beta \mathbf{t}_{i,z_c} \right) + \sum_{i=0}^{c-1} a_{w_i w_c} + r_{i,z_c}. \quad (2)$$

In order to estimate r_{ik} , we let the context size $c = 0$ and $z_c = k$, and then (1) becomes:

$$P(s_j | k, d_i) = P(s_j) \exp \left\{ \beta \mathbf{v}_{s_j}^\top \mathbf{t}_{ik} + r_{ik} \right\}. \quad (3)$$

It is required that $\sum_{s_j \in \mathcal{S}} P(s_j | k, d_i) = 1$ to make (3) a distribution. It follows that

$$r_{ik} = -\log \left(\sum_{s_j \in \mathcal{S}} P(s_j) \exp \{ \beta \mathbf{v}_{s_j}^\top \mathbf{t}_{ik} \} \right). \quad (4)$$

That is, r_{ik} is uniquely determined by β and \mathbf{t}_{ik} . Specifically, when $\beta = 0$, $r_{ik} = 0$. Remind that when $\forall i, \mathbf{t}_{i1} = 0$, and thus $r_{i1} = 0$.

Our decision of making r_{ik} invariant to different values of w_c is a trade-off between computational efficiency and modeling accuracy. Intuitively, the distribution of w_c is primarily determined by its context $w_0:w_{c-1}$, and less influenced by the topic \mathbf{t}_{ik} . Then the magnitude of $\beta \mathbf{v}_{w_c}^\top \mathbf{t}_{ik} + r_{ik}$ should usually be smaller than the that of the context vectors. Within this expression, the magnitude of r_{ik} should also be smaller than the residuals between two words. As such, approximating it by a constant value will not result in big errors of the distribution of w_c .

4 The Generative Process

Now we have proposed the basic distributions of the words. Before the generative process begins, a few hyperparameters need to be specified:

1. The parameter $\boldsymbol{\alpha}$ of the Dirichlet prior of the mixing proportions $\boldsymbol{\phi}_i$, $\text{Dir}(\boldsymbol{\alpha})$;
2. The topic weight β ;

The generative process is as follows:

1. Draw the residual matrix \mathbf{A} from the Truncated Gaussian prior $\mathcal{N}_{\text{Fea}(\mathbf{G}, N)}(\mathbf{A}; 0, \mathbf{H})$;
2. Draw the embeddings \mathbf{V} uniformly from the solution set $\text{Sol}(\mathbf{V}; \mathbf{G}, \mathbf{A})$, of $\mathbf{V}^\top \mathbf{V} = \mathbf{G} - \mathbf{A}$;
3. For each document d_i :
 - (a) Draw the mixing proportions $\boldsymbol{\phi}_i$ from the Dirichlet prior $\text{Dir}(\boldsymbol{\alpha})$;
 - (b) For the j -th word, do the following:
 - i. Draw topic assignment z_{ij} from the categorical distribution $\text{Cat}(\boldsymbol{\phi}_i)$;
 - ii. Draw word w_{ij} with probability $P(w_{ij} | w_{i,j-c}:w_{i,j-1}, z_{ij}, d_i)$.

5 Likelihood Function

Given the embeddings \mathbf{V} and the bigram residuals \mathbf{A} , the topics \mathbf{T} and the hyperparamters α, β , the complete-data likelihood of a document d_i is:

$$\begin{aligned}
& p(d_i, \mathbf{Z}_i, \phi_i | \alpha, \beta, \mathbf{V}, \mathbf{A}, \mathbf{T}_i) \\
&= p(\phi_i | \alpha) p(\mathbf{Z}_i | \phi_i) p(d_i | \beta, \mathbf{V}, \mathbf{A}, \mathbf{T}_i, \mathbf{Z}_i) \\
&= \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{j=1}^K \phi_{ij}^{\alpha_j - 1} \cdot \prod_{j=1}^{L_i} \left(\phi_{i,z_{ij}} P(w_{ij}) \right. \\
&\quad \left. \cdot \exp \left\{ \mathbf{v}_{w_{ij}}^\top \left(\sum_{k=j-c}^{j-1} \mathbf{v}_{w_{ik}} + \beta \mathbf{t}_{z_{ij}} \right) + \sum_{k=j-c}^{j-1} a_{w_{ik} w_{ij}} + r_{i,z_{ij}} \right\} \right), \quad (5)
\end{aligned}$$

where $\mathbf{Z}_i = (z_{i1}, \dots, z_{iL_i})$, and $\Gamma(\cdot)$ is the Gamma function. The topic residuals $\mathbf{r}_i = \{r_{ik}\}_k$ are uniquely determined by \mathbf{T}_i and β , and thus are implicit in the likelihood functions.

We denote the latent variables of all documents $\{\mathbf{Z}_i\}_{i=1}^M$ collectively by \mathbf{Z} , and all the document-specific $\{\phi_i\}_{i=1}^M$ by ϕ . Then the complete-data likelihood of the whole corpus is:

$$\begin{aligned}
& p(\mathbf{D}, \mathbf{B}, \mathbf{A}, \mathbf{V}, \mathbf{Z}, \phi | \alpha, \beta, \mathbf{T}) \\
&= \mathcal{N}_{\text{Fea}(\mathbf{G}, N)}(\mathbf{A}; 0, \mathbf{H}) \cdot U(\text{Sol}(\mathbf{V}; \mathbf{G}, \mathbf{A})) \\
&\quad \cdot \prod_{i=1}^M \{p(\phi_i | \alpha) p(\mathbf{Z}_i | \phi_i) p(d_i | \beta, \mathbf{V}, \mathbf{A}, \mathbf{T}_i, \mathbf{Z}_i)\} \\
&= \frac{1}{\mathcal{Z}(\mathbf{A}, \mathbf{V}; \mathbf{B})} \exp \left\{ - \sum_{i,j=1}^{W,W} f(h_{i,j}) a_{s_i s_j}^2 \right\} \prod_{i=1}^M \left\{ \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{j=1}^K \phi_{ij}^{\alpha_j - 1} \right. \\
&\quad \left. \cdot \prod_{j=1}^{L_i} \left(\phi_{i,z_{ij}} P(w_{ij}) \cdot \exp \left\{ \mathbf{v}_{w_{ij}}^\top \left(\sum_{k=j-c}^{j-1} \mathbf{v}_{w_{ik}} + \beta \mathbf{t}_{z_{ij}} \right) + \sum_{k=j-c}^{j-1} a_{w_{ik} w_{ij}} + r_{i,z_{ij}} \right\} \right) \right\}, \quad (6)
\end{aligned}$$

where $U(\text{Sol}(\mathbf{V}; \mathbf{G}, \mathbf{A}))$ is a uniform distribution over $\text{Sol}(\mathbf{V}; \mathbf{G}, \mathbf{A})$, and $\mathcal{Z}(\mathbf{A}, \mathbf{V}; \mathbf{B})$ is the normalizing function of $\mathcal{N}_{\text{Fea}(\mathbf{G}, N)}(\mathbf{A}; 0, \mathbf{H}) \cdot U(\text{Sol}(\mathbf{V}; \mathbf{G}, \mathbf{A}))$:

$$\mathcal{Z}(\mathbf{A}, \mathbf{V}; \mathbf{B}) = \int_{\text{Fea}(\mathbf{G}, N)} \exp \{ - \|\mathbf{A}\|_{f(\mathbf{H})}^2 \} \cdot \lambda(\text{Sol}(\mathbf{V}; \mathbf{G}, \mathbf{A})) d\mathbf{A}, \quad (7)$$

where $\lambda(\text{Sol}(\mathbf{V}; \mathbf{G}, \mathbf{A}))$ is the Lebesgue measure of $\text{Sol}(\mathbf{V}; \mathbf{G}, \mathbf{A})$.

Taking the logarithm of both sides, we obtain

$$\begin{aligned}
& \log p(\mathbf{D}, \mathbf{B}, \mathbf{A}, \mathbf{V}, \mathbf{Z}, \phi | \alpha, \beta, \mathbf{T}) \\
&= C_0 - \log \mathcal{Z}(\mathbf{A}, \mathbf{V}; \mathbf{B}) - \|\mathbf{A}\|_{f(\mathbf{H})}^2 + \sum_{i=1}^M \left\{ \log \phi_{ik} \cdot \sum_{k=1}^K (m_{ik} + \alpha_{0k} - 1) \right. \\
&\quad \left. + \sum_{j=1}^{L_i} \left(\mathbf{v}_{w_{ij}}^\top \left(\sum_{k=j-c}^{j-1} \mathbf{v}_{w_{ik}} + \beta \mathbf{t}_{z_{ij}} \right) + \sum_{k=j-c}^{j-1} a_{w_{ik}w_{ij}} + r_{i,z_{ij}} \right) \right\}, \tag{8}
\end{aligned}$$

where $m_{ik} = \sum_{j=1}^{L_i} \delta(z_{ij} = k)$ counts the number of words assigned with the k -th topic in d_i , $C_0 = M \log \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} + \sum_{i,j=1}^{M,L_i} \log P(w_{ij})$ is constant given α .

6 Two Stage Learning Algorithm

6.1 Learning Objective and Process

Given the hyperparameters α, β , the learning objective is to find the estimates of the bigram probabilities \mathbf{B} , the embeddings and residuals \mathbf{V}, \mathbf{A} , the topics \mathbf{T} , and the word-topic and document-topic distributions $p(\mathbf{Z}_i, \phi_i | d_i, \mathbf{B}, \mathbf{A}, \mathbf{V}, \mathbf{T})$. Here the hyperparameters α, β are fixed after specified manually and effectively constants, and hence we hide them in the distribution notations.

We denote $\{\mathbf{Z}_i, \phi_i\}_{i=1}^M$ collectively as \mathbf{Z}, ϕ . Then the above objective is to find the optimal $\mathbf{B}^*, \mathbf{A}^*, \mathbf{V}^*, \mathbf{T}^*$ and the posterior $p(\mathbf{Z}, \phi | \mathbf{D}, \mathbf{B}^*, \mathbf{A}^*, \mathbf{V}^*, \mathbf{T}^*)$. This posterior is analytically intractable, and we use a simpler variational distribution $q(\mathbf{Z}, \phi)$ to approximate it.

The coupling between \mathbf{A}, \mathbf{V} and $\mathbf{T}, \mathbf{Z}, \phi$ in (8) makes it very difficult to find the optimal $\mathbf{A}^*, \mathbf{V}^*, \mathbf{T}^*$ and the corresponding posterior of \mathbf{Z}, ϕ . To get around this difficulty, we divide the learning into *two stages*.

1. In the first stage, considering that the topics have relatively small impact to word distributions, we simplify the model by disabling topics temporarily, and obtain the optimal solution $\mathbf{B}^*, \mathbf{A}^*, \mathbf{V}^*$ of this reduced model. The optimal solution could be calculated in closed-form;
2. In the second stage, we use $\mathbf{B}^*, \mathbf{A}^*, \mathbf{V}^*$ as an approximate solution, and then enable the topics, and find the corresponding optimal \mathbf{T}^* , $p(\mathbf{Z}, \phi | \mathbf{D}, \mathbf{B}^*, \mathbf{A}^*, \mathbf{V}^*, \mathbf{T}^*)$ of the full model. In the presence of a lot of hidden variables, a variational EM algorithm is pertinent. During the VEM iterations, we fix $\mathbf{B} = \mathbf{B}^*, \mathbf{A} = \mathbf{A}^*, \mathbf{V} = \mathbf{V}^*$.

6.2 Estimating $\mathbf{B}, \mathbf{A}, \mathbf{V}$ on the Reduced Model with Topics Disabled

As the first step, we disable topics by setting the topic weight β temporarily to 0. In this reduced model, different choices of the topic embeddings \mathbf{T} , document-topic distributions ϕ and topic assignments \mathbf{Z} only bring a constant offset to the log-likelihood of the corpus, so they are chosen arbitrarily as $\mathbf{T}_0, \phi_0, \mathbf{Z}_0$.

The matrix \mathbf{B} is estimated using the Maximum Likelihood Estimation, and \mathbf{A}, \mathbf{V} are estimated using the Low Rank Positive Semidefinite Approximation algorithm in Section 5, [1].

6.3 Estimating $\mathbf{T}, \mathbf{Z}, \phi$ using Variational EM Algorithm on the Full Model

In this stage, we use $\mathbf{B}^*, \mathbf{A}^*, \mathbf{V}^*$ obtained in the previous subsection as their approximate solutions, and then enable the topics by setting β to the prespecified value. Then we proceed to find the corresponding optimal $\mathbf{T}^*, p(\mathbf{Z}, \phi | \mathbf{D}, \mathbf{B}^*, \mathbf{A}^*, \mathbf{V}^*, \mathbf{T}^*)$ of this full model. In the presence of a lot of hidden variables, a variational EM algorithm is pertinent. During the VEM iterations, we fix $\mathbf{B} = \mathbf{B}^*, \mathbf{A} = \mathbf{A}^*, \mathbf{V} = \mathbf{V}^*$.

To simplify notation, in the following, we make the hyperparameters α, β , and the fixed parameters $\mathbf{B}^*, \mathbf{A}^*, \mathbf{V}^*$ implicit in the probabilistic functions. As the topic residuals $\mathbf{r} = \{r_{ik}\}_{i,k}$ are uniquely determined by \mathbf{T} and β , they are also kept implicit whenever they are irrelevant to the discussion.

We use p to denote the posterior $p(\mathbf{Z}, \phi | \mathbf{D}, \mathbf{T})$ when it is clear from context. Then for an arbitrary variational distribution $q(\mathbf{Z}, \phi)$, the following equalities hold

$$\begin{aligned} & E_q \log \left[\frac{p(\mathbf{D}, \mathbf{Z}, \phi | \mathbf{T})}{q(\mathbf{Z}, \phi)} \right] \\ &= E_q [\log p(\mathbf{D}, \mathbf{Z}, \phi | \mathbf{T})] + \mathcal{H}(q) \\ &= \log p(\mathbf{D} | \mathbf{T}) - \text{KL}(q || p), \end{aligned} \tag{9}$$

which implies

$$\text{KL}(q || p) = \log p(\mathbf{D} | \mathbf{T}) - \left(E_q [\log p(\mathbf{D}, \mathbf{Z}, \phi | \mathbf{T})] + \mathcal{H}(q) \right). \tag{10}$$

In (10), $E_q [\log p(\mathbf{D}, \mathbf{Z}, \phi | \mathbf{T})] + \mathcal{H}(q)$ is usually referred to as the *variational free energy* $\mathcal{L}(q, \mathbf{T})$, which is a lower bound of $\log p(\mathbf{D} | \mathbf{T})$. Directly

maximizing $\log p(\mathbf{D}|\mathbf{T})$ w.r.t. \mathbf{T} is intractable due to the hidden variables \mathbf{Z}, ϕ , so we maximize its lower bound $\mathcal{L}(q, \mathbf{T})$ instead. We adopt a mean-field approximation of the true posterior as the variational distribution, and use a Variational Expectation Maximization (VEM) algorithm to find q^*, \mathbf{T}^* maximizing $\mathcal{L}(q, \mathbf{T})$.

6.3.1 Mean-Field Approximation and VEM Algorithm

We assume that the mean-field approximation of the true posterior factorizes as follows:

$$q(\mathbf{Z}, \phi; \boldsymbol{\pi}, \boldsymbol{\theta}) = q(\phi; \boldsymbol{\theta})q(\mathbf{Z}; \boldsymbol{\pi}) = \prod_{i=1}^M \left\{ \text{Dir}(\phi_i; \boldsymbol{\theta}_i) \prod_{j=1}^{L_i} \text{Cat}(z_{ij}; \boldsymbol{\pi}_{ij}) \right\}.$$

Taking the logarithm of both sides, we obtain

$$\begin{aligned} \log q(\mathbf{Z}, \phi; \boldsymbol{\pi}, \boldsymbol{\theta}) &= \sum_{i=1}^M \left\{ \log \Gamma(\theta_{i0}) - \sum_{k=1}^K \log \Gamma(\theta_{ik}) \right. \\ &\quad \left. + \sum_{k=1}^K (\theta_{ik} - 1) \log \phi_{ik} + \sum_{j,k=1}^{L_i, K} \delta(z_{ij} = k) \log \pi_{ij}^k \right\}, \quad (11) \end{aligned}$$

where $\theta_{i0} = \sum_{k=1}^K \theta_{ik}$, π_{ij}^k is the k -th component of $\boldsymbol{\pi}_{ij}$.

It follows that

$$\begin{aligned} &\mathcal{H}(q) \\ &= -E_q[\log q(\mathbf{Z}, \phi; \boldsymbol{\pi}, \boldsymbol{\theta})] \\ &= \sum_{i=1}^M \left\{ \sum_{k=1}^K \log \Gamma(\theta_{ik}) - \log \Gamma(\theta_{i0}) - \sum_{k=1}^K (\theta_{ik} - 1) \psi(\theta_{ik}) + (\theta_{i0} - K) \psi(\theta_{i0}) - \sum_{j,k=1}^{L_i, K} \pi_{ij}^k \log \pi_{ij}^k \right\}. \quad (12) \end{aligned}$$

Plugging q into $\mathcal{L}(q, \mathbf{T})$, we have

$$\begin{aligned}
& \mathcal{L}(q, \mathbf{T}) \\
&= \mathcal{H}(q) + E_q [\log p(\mathbf{Z}, \phi | \mathbf{T})] \\
&= \mathcal{H}(q) + C_0 - \log \mathcal{Z}(\mathbf{A}^*, \mathbf{V}^* | \mathbf{B}^*) - \|\mathbf{A}\|_{f(\mathbf{H})}^2 \\
& \quad + \sum_{i=1}^M \left\{ \sum_{k=1}^K (E_{q(\mathbf{Z}_i | \boldsymbol{\pi}_i)}[m_{ik}] + \alpha_{0k} - 1) \cdot E_{q(\phi_{ik} | \boldsymbol{\theta}_i)}[\log \phi_{ik}] \right. \\
& \quad \left. + \sum_{j=1}^{L_i} \left(\mathbf{v}_{w_{ij}}^\top \left(\sum_{k=j-c}^{j-1} \mathbf{v}_{w_{ik}} + \beta E_{q(z_{ij} | \boldsymbol{\pi}_{ij})}[\mathbf{t}_{z_{ij}}] \right) + \sum_{k=j-c}^{j-1} a_{w_{ik} w_{ij}} + E_{q(z_{ij} | \boldsymbol{\pi}_{ij})}[r_{i, z_{ij}}] \right) \right\} \\
&= C_1 + \mathcal{H}(q) + \sum_{i=1}^M \left\{ \sum_{k=1}^K \left(\sum_{j=1}^{L_i} \pi_{ij}^k + \alpha_{0k} - 1 \right) \left(\psi(\theta_{ik}) - \psi(\theta_{i0}) \right) + \sum_{j=1}^{L_i} \left(\beta \mathbf{v}_{w_{ij}}^\top \mathbf{T}_i \boldsymbol{\pi}_{ij} + \mathbf{r}_i^\top \boldsymbol{\pi}_{ij} \right) \right\}, \tag{13}
\end{aligned}$$

where \mathbf{T}_i is the topic matrix of the i -th document, and \mathbf{r}_i is the vector constructed by concatenating all the topic residuals r_{ik} . $C_1 = C_0 - \log \mathcal{Z}(\mathbf{A}^*, \mathbf{V}^* | \mathbf{B}^*) - \|\mathbf{A}\|_{f(\mathbf{H})}^2 + \sum_{i,j=1}^{M, L_i} \left(\mathbf{v}_{w_{ij}}^\top \sum_{k=j-c}^{j-1} \mathbf{v}_{w_{ik}} + \sum_{k=j-c}^{j-1} a_{w_{ik} w_{ij}} \right)$ is constant. $\psi(\cdot)$ is the digamma function.

Then the Variational EM algorithm alternately optimize w.r.t. q and \mathbf{T}, \mathbf{r} as follows:

1. Initialize all the topics $\mathbf{T}_i = \mathbf{0}$, and correspondingly their residuals $\mathbf{r}_i = \mathbf{0}$;
2. Iterate over the following two steps until convergence. In the l -th step:
 - (a) Let the topics and residuals be $\mathbf{T} = \mathbf{T}^{(l-1)}, \mathbf{r} = \mathbf{r}^{(l-1)}$, find $q^{(l)}(\mathbf{Z}, \phi)$ that maximizes $\mathcal{L}(q, \mathbf{T}^{(l-1)})$. This is the Expectation step (E-step). In this step, $\log p(\mathbf{D} | \mathbf{T})$ is constant. Then the q that maximizes $\mathcal{L}(q, \mathbf{T}^{(l)})$ will minimize $\text{KL}(q || p)$, i.e. such a q is the closest variational distribution to p measured by KL-divergence;
 - (b) Given the variational distribution $q^{(l)}(\mathbf{Z}, \phi)$, find $\mathbf{T}^{(l)}, \mathbf{r}^{(l)}$ that maximizes $\mathcal{L}(q^{(l)}, \mathbf{T})$. This is the Maximization step (M-step). In this step, $\boldsymbol{\pi}, \boldsymbol{\theta}, \mathcal{H}(q)$ are constant;

6.3.2 Update Equations of $\boldsymbol{\pi}, \boldsymbol{\theta}$ in E-Step

In the E-step, $\mathbf{T} = \mathbf{T}^{(l-1)}, \mathbf{r} = \mathbf{r}^{(l-1)}$ are constant. For notational simplicity, we drop their superscripts (l) and denote them as \mathbf{T}, \mathbf{r} .

Plugging (12) into (13), we obtain

$$\begin{aligned}
& \mathcal{L}(q, \mathbf{T}^{(l-1)}) \\
&= \sum_{i=1}^M \left\{ \sum_{k=1}^K \log \Gamma(\theta_{ik}) - \log \Gamma(\theta_{i0}) - \sum_{k=1}^K (\theta_{ik} - 1) \psi(\theta_{ik}) + (\theta_{i0} - K) \psi(\theta_{i0}) - \sum_{j,k=1}^{L_i, K} \pi_{ij}^k \log \pi_{ij}^k \right. \\
&\quad \left. + \sum_{k=1}^K \left(\sum_{j=1}^{L_i} \pi_{ij}^k + \alpha_{0k} - 1 \right) \left(\psi(\theta_{ik}) - \psi(\theta_{i0}) \right) + \sum_{j=1}^{L_i} \left(\beta \mathbf{v}_{w_{ij}}^\top \mathbf{T}_i \boldsymbol{\pi}_{ij} + \mathbf{r}_i^\top \boldsymbol{\pi}_{ij} \right) \right\} + C_5.
\end{aligned} \tag{14}$$

We first maximize (14) w.r.t. π_{ij}^k , the probability that the j -th word in the i -th document takes the k -th latent topic. Note that this optimization is subject to the normalization constraint that $\sum_{k=1}^K \pi_{ij}^k = 1$.

We isolate terms containing $\boldsymbol{\pi}_{ij}$, and form a Lagrange function by incorporating the normalization constraint:

$$\Lambda(\boldsymbol{\pi}_{ij}) = - \sum_{k=1}^K \pi_{ij}^k \log \pi_{ij}^k + \sum_{k=1}^K \left(\psi(\theta_{ik}) - \psi(\theta_{i0}) \right) \pi_{ij}^k + \beta \mathbf{v}_{w_{ij}}^\top \mathbf{T}_i \boldsymbol{\pi}_{ij} + \mathbf{r}_i^\top \boldsymbol{\pi}_{ij} + \lambda_{ij} \left(\sum_{k=1}^K \pi_{ij}^k - 1 \right). \tag{15}$$

Taking the derivative w.r.t. π_{ij}^k , we obtain

$$\frac{\partial \Lambda(\boldsymbol{\pi}_{ij})}{\partial \pi_{ij}^k} = -1 - \log \pi_{ij}^k + \psi(\theta_{ik}) - \psi(\theta_{i0}) + \beta \mathbf{v}_{w_{ij}}^\top \mathbf{t}_{ik} + r_{ik} + \lambda_{ij}. \tag{16}$$

Setting this derivative to 0 yields the maximizing value of π_{ij}^k :

$$\pi_{ij}^k \propto \exp\{\psi(\theta_{ik}) + \beta \mathbf{v}_{w_{ij}}^\top \mathbf{t}_{ik} + r_{ik}\}. \tag{17}$$

Next, we maximize (14) w.r.t. θ_{ik} , the k -th component of the posterior Dirichlet parameter:

$$\begin{aligned}
& \frac{\partial \mathcal{L}(q, \mathbf{T}^{(l-1)})}{\partial \theta_{ik}} \\
&= \frac{\partial}{\partial \theta_{ik}} \left\{ \log \Gamma(\theta_{ik}) - \log \Gamma(\theta_{i0}) + \left(\sum_{j=1}^{L_i} \pi_{ij}^k + \alpha_{0k} - \theta_{ik} \right) \psi(\theta_{ik}) - \left(L_i + \sum_k \alpha_{0k} - \theta_{i0} \right) \psi(\theta_{i0}) \right\} \\
&= \left(\sum_{j=1}^{L_i} \pi_{ij}^k + \alpha_{0k} - \theta_{ik} \right) \psi'(\theta_{ik}) - \left(L_i + \sum_k \alpha_{0k} - \theta_{i0} \right) \psi'(\theta_{i0}),
\end{aligned} \tag{18}$$

where $\psi'(\cdot)$ is the derivative of the digamma function $\psi(\cdot)$, commonly referred to as the *trigamma function*.

Setting (18) to 0 yields a maximum at

$$\theta_{ik} = \sum_{j=1}^{L_i} \pi_{ij}^k + \alpha_{0k}. \quad (19)$$

Note this solution depends on the values of π_{ij}^k , which in turn depends on θ_{ik} in (17). Then we have to alternate between (17) and (19) until convergence.

6.3.3 Update Equations of $\mathbf{T}_i, \mathbf{r}_i$ in M-Step

In the M-step, $\boldsymbol{\pi} = \boldsymbol{\pi}^{(l)}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(l)}$ are constant. For notational simplicity, we drop their superscripts (l) and denote them as $\boldsymbol{\pi}, \boldsymbol{\theta}$.

Given these parameter values, (13) is a constant plus the sum of many $\beta \mathbf{v}_{w_{ij}}^\top \mathbf{T}_i \boldsymbol{\pi}_{ij} + \mathbf{r}_i^\top \boldsymbol{\pi}_{ij}$, each of which in turn is a linear transformation of the vector $\beta \mathbf{v}_{w_{ij}}^\top \mathbf{T}_i + \mathbf{r}_i^\top$. The k -th component of this vector is $\log \frac{\exp\{\beta \mathbf{v}_s^\top \mathbf{t}_{ik}\}}{E_{P(s)}[\exp\{\beta \mathbf{v}_s^\top \mathbf{t}_{ik}\}]}$, the logarithm of a softmax function of \mathbf{t}_{ik} . As a softmax function is concave w.r.t. the weight \mathbf{t}_{ik} , this component is concave, and so is $\beta \mathbf{v}_{w_{ij}}^\top \mathbf{T}_i + \mathbf{r}_i^\top$. Therefore $\mathcal{L}(q^{(l)}, \mathbf{T})$ is a concave function of \mathbf{T} , and its maximum is achieved when its derivative w.r.t. \mathbf{T} is 0.

The topic residuals \mathbf{r}_i are uniquely determined by \mathbf{T}_i and β . Thus we first solve \mathbf{T}_i , and then \mathbf{r}_i is readily determined.

As the first column of \mathbf{T}_i is fixed to 0, we only need to find the maximum w.r.t. other columns. We denote the submatrix of all columns of \mathbf{T}_i except the first column as $\mathbf{T}_{-1,i}$. To find this maximum, we take the derivative of (13) w.r.t. $\mathbf{T}_{-1,i}$:

$$\begin{aligned} & \frac{\partial \mathcal{L}(q^{(l)}, \mathbf{T})}{\partial \mathbf{T}_{-1,i}} \\ &= \frac{\partial \sum_{j=1}^{L_i} \left(\beta \mathbf{v}_{w_{ij}}^\top \mathbf{T}_i \boldsymbol{\pi}_{ij} + \boldsymbol{\pi}_{ij}^\top \mathbf{r}_i \right)}{\partial \mathbf{T}_{-1,i}} \\ &= \beta \frac{\partial}{\partial \mathbf{T}_{-1,i}} \text{Tr}(\mathbf{T}_i \sum_{j=1}^{L_i} \boldsymbol{\pi}_{ij} \mathbf{v}_{w_{ij}}^\top) + \left(\sum_{j=1}^{L_i} \boldsymbol{\pi}_{ij} \right)^\top \frac{\partial \mathbf{r}_i}{\partial \mathbf{T}_{-1,i}} \\ &= \beta \sum_{j=1}^{L_i} \mathbf{v}_{w_{ij}} \boldsymbol{\pi}_{-1,ij}^\top + \left(\sum_{j=1}^{L_i} \boldsymbol{\pi}_{ij} \right)^\top \frac{\partial \mathbf{r}_i}{\partial \mathbf{T}_{-1,i}} \\ &= \beta \sum_{j=1}^{L_i} \mathbf{v}_{w_{ij}} \boldsymbol{\pi}_{-1,ij}^\top + \sum_{k=2}^K \bar{\pi}_i^k \frac{\partial r_{ik}}{\partial \mathbf{T}_{-1,i}}, \end{aligned} \quad (20)$$

where $\bar{\pi}_i^k = \sum_{j=1}^{L_i} \pi_{ij}^k$, the sum of the variational probabilities of each word being assigned to the k -th topic in the i -th document. $\boldsymbol{\pi}_{-1,ij}^\top$ is the subvector of all elements of $\boldsymbol{\pi}_{ij}$ except the first: $(\pi_{ij}^2, \dots, \pi_{ij}^K)^\top$. The index of k in the second term in (20) starts from 2 because r_{i1} is fixed to be 0.

Solving the critical point $\mathbf{T}_{-1,i}$ of (20) requires the computation of $\frac{\partial r_{ik}}{\partial \mathbf{T}_i}$. (4) states that $r_{ik} = -\log(E_{P(s)}[\exp\{\beta \mathbf{v}_s^\top \mathbf{t}_{ik}\}])$. Then the derivative of r_{ik} w.r.t. \mathbf{T}_i is difficult to compute. Alternatively we use a second-order approximation to ease the computation.

As discussed above, $\|\beta \mathbf{t}_{ik}\|$ is small, and thus $\|\beta \mathbf{v}_s^\top \mathbf{t}_{ik}\|$ is usually small too (the Gaussian prior over \mathbf{v}_s strongly discourage big $\|\mathbf{v}_s\|$). Then a second-order approximation to $\exp\{\beta \mathbf{v}_s^\top \mathbf{t}_{ik}\}$ is appropriate: $\exp\{\beta \mathbf{v}_s^\top \mathbf{t}_{ik}\} \approx 1 + \beta \mathbf{v}_s^\top \mathbf{t}_{ik} + \frac{1}{2} \beta^2 (\mathbf{v}_s^\top \mathbf{t}_{ik})^2$. It follows that

$$\begin{aligned} & E_{P(s)}[\exp\{\beta \mathbf{v}_s^\top \mathbf{t}_{ik}\}] \\ & \approx 1 + \beta \mathbf{t}_{ik}^\top E_{P(s)}[\mathbf{v}_s] + \frac{1}{2} \beta^2 \mathbf{t}_{ik}^\top E_{P(s)}[\mathbf{v}_s \mathbf{v}_s^\top] \mathbf{t}_{ik}. \\ & = 1 + \beta \mathbf{t}_{ik}^\top \bar{\mathbf{v}} + \frac{1}{2} \beta^2 \mathbf{t}_{ik}^\top \mathbf{X} \mathbf{t}_{ik}, \end{aligned} \quad (21)$$

where $\bar{\mathbf{v}} = E_{P(s)}[\mathbf{v}_s]$ and $\mathbf{X} = E_{P(s)}[\mathbf{v}_s \mathbf{v}_s^\top]$. As \mathbf{V} is fixed, $\bar{\mathbf{v}}$ and \mathbf{X} can be precomputed. The dimensionality of \mathbf{X} is $N \times N$, and N is usually chosen as hundreds. Thus \mathbf{X} can easily fit into the memory.

It follows that

$$\begin{aligned} \frac{\partial r_{ik}}{\partial \mathbf{t}_{ik}} &= -\frac{1}{E_{P(s)}[\exp\{\beta \mathbf{v}_s^\top \mathbf{t}_{ik}\}]} \frac{\partial}{\partial \mathbf{t}_{ik}} E_{P(s)}[\exp\{\beta \mathbf{v}_s^\top \mathbf{t}_{ik}\}] \\ &\approx -e^{r_{ik}} \cdot \beta (\bar{\mathbf{v}} + \beta \mathbf{X} \mathbf{t}_{ik}). \end{aligned} \quad (22)$$

To summarize, $\frac{\partial r_{ik}}{\partial \mathbf{t}_{ij}}$ are divided into two cases:

$$\begin{cases} \frac{\partial r_{ik}}{\partial \mathbf{t}_{ik}} \approx -e^{r_{ik}} \cdot \beta (\bar{\mathbf{v}} + \beta \mathbf{X} \mathbf{t}_{ik}), & k \neq 1 \\ \frac{\partial r_{ik}}{\partial \mathbf{t}_{ij}} = 0, & k = 1 \text{ or } j \neq k. \end{cases} \quad (23)$$

Plugging (23) into (20), we obtain

$$\frac{\partial \mathcal{L}(q^{(l)}, \mathbf{T})}{\partial \mathbf{T}_{-1,i}} \approx \beta \sum_{j=1}^{L_i} \mathbf{v}_{w_{ij}} \boldsymbol{\pi}_{-1,ij}^\top - \beta (\bar{\mathbf{V}} + \beta \mathbf{X} \mathbf{T}_{-1,i}) \Pi_i, \quad (24)$$

where $\bar{\mathbf{V}} = (\bar{\mathbf{v}} \cdots \bar{\mathbf{v}})_{N \times (K-1)}$, whose first column is 0 and other columns are all $\bar{\mathbf{v}}$, and $\Pi_i = \begin{pmatrix} \bar{\pi}_i^2 e^{r_{i2}} & & 0 \\ & \ddots & \\ 0 & & \bar{\pi}_i^K e^{r_{iK}} \end{pmatrix} = \text{diag}(\bar{\boldsymbol{\pi}}_{-1,i}) \text{diag}(\exp\{\mathbf{r}_{-1,i}\})$. Here $\mathbf{r}_{-1,i}$ is the subvector of all elements of \mathbf{r}_i except the first.

Setting the RHS of (24) to 0 leads to an equation whose solution is near $\arg \max_{\mathbf{T}_{-1,i}} \mathcal{L}(q^{(l)}, \mathbf{T})$:

$$(\bar{\mathbf{V}} + \beta \mathbf{X} \mathbf{T}_{-1,i}) \Pi_i = \sum_{j=1}^{L_i} \mathbf{v}_{w_{ij}} \boldsymbol{\pi}_{-1,ij}^\top. \quad (25)$$

However, (25) cannot be solved directly, because the terms $e^{r_{ik}}$ in Π_i are complicated functions of \mathbf{t}_{ik} . To circumvent this complexity, we adopt an iterative algorithm. In the m -th iteration, $\mathbf{r}_{-1,i}$ take the values $\mathbf{r}_{-1,i}^{(m-1)}$ found in the $(m-1)$ -th iteration (if $m = 1$, then $\mathbf{r}_{-1,i}$ take the values computed in the last E-step), yielding a solution

$$\mathbf{T}_{-1,i}^{(m)} = \frac{1}{\beta} \mathbf{X}^{-1} \left\{ \left(\sum_{j=1}^{L_i} \mathbf{v}_{w_{ij}} \boldsymbol{\pi}_{-1,ij}^\top \right) \text{diag}(\bar{\boldsymbol{\pi}}_{-1,i})^{-1} \text{diag}(\exp\{-\mathbf{r}_{-1,i}^{(m-1)}\}) - \bar{\mathbf{V}} \right\}. \quad (26)$$

In the next iteration, $\mathbf{r}_{-1,i}^{(m)}$ is computed using (4). This iterative process continues until convergence.

References

- [1] Anonymous. A generative word embedding model and its low rank positive semidefinite solution. Submitted to EMNLP'2015.