Notes on Latent Direchlet Allocation with Full Bayesian Treatment

Shoichiro Yamanishi

April 27, 2020

1 Overview

This is an expository document for latent Direchlet allocation in the full Bayesian setting for β matrix. This is an outcome of my own self study into the original article, and it has turned out to be a very good streamlined study material for EM-algorithm and variational Bayes worth documenting by myself for my own better understanding.

The main purpose of this document is to quickly and effortlessly refresh my memory in the future as my own memory aid.

Another purpose is to fill the gap between the original article[1] and the blog articles available on Internet. The original article is too terse, and it takes me a lot of paper-and-pencil work to comprehend the contents. It also briefly touches on the full Bayesian treatment. On the other hand, the blog articles focus on a quick grasp of the concept with rich illustrations, and rigorous mathematical treatment is usually ommitted.

This document is characterized as follows.

- Comprehensive math treatment from the modeling down to just before implementation for both training and inference.
- Full-bayesian β with prior η.
- Small gaps between two adjacent equations through the course of deductions at the cost of lengthiness.
- Discussion on possibility of treatment of using the columns of β as word embeddings with non-informative priors.

This document is organized as follows. First, the generative model is defined, and then the problems of the training and the inference are defined. Then the training problem is solved by a version of EM-algorithm. First the expectation part of finding the posterior distribution is approximated by variational Bayes, specifically by the mean field approximation. And then the training parameters $\boldsymbol{\alpha}$ and η are updated in the maximization step. The maximization of those parameters are solved by Newton-Raphson method and it is then explained. Solving Newton-Raphson method involves evaluating Digamma and Trigamma functions, and numerical approximation techniques are explained. Finally, a possibility of using the columns of β as word embeddings in relation to the topic inference of a new document with non-informative priors is explained.

2 Generative Model Formation

Latent Direchlet allocation is a generative probabilistic model proposed by [1] for documents utilizing something called exchangeability of words in a document (de Finetti's Theorem). It is also used to generate word embeddings, notably lda2vec[2].

Definition 1 $\mathbf{w}_n$ *represents a word for the n-th position of a document. It is a one-of-V vector such that* $\mathbf{w}_n \in \{0,1\}^V$ *where always only one element is 1, and the rest is 0. We use the notation* $w_n^v = 1$ *if* w_n *represents the v-th word, or* $w_n^v = 0$ *if* w_n *does not represent the v-th word,*

Definition 2 $W_m = (\mathbf{w}_{m1}, \mathbf{w}_{m2}, \cdots, \mathbf{w}_{mN_m})$ *represents a document. It is an ordered list of* N_m *words.*

Definition 3 $\mathcal{D} = \{W_1, W_2, \cdots, W_M\}$ *represents a corpus. It is a set of M documents.*

The aim of this model is to represent a document W_m by **a probability distribution** $\boldsymbol{\theta}_m$ of K topics. Naturally it is a multinomial distribution. It is drawn from a Dirichlet prior $\boldsymbol{\alpha} \in \mathcal{R}^K$, which is learned empirically.

$$p(\boldsymbol{\theta}_m|\boldsymbol{\alpha}) = Dir(\boldsymbol{\theta}_m|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} (\theta_m^k)^{\alpha_k - 1} \tag{1}$$

where $0 \leq \theta_m^k \leq 1$ denotes the k-th element of the vector $\boldsymbol{\theta}_m$. Please note that $\boldsymbol{\theta}_m$ itself represents a multinomial probability distribution.

In LDA, each word $\mathbf{w}_{mn}$ in a document W_m is drawn from a distribution in β, which is dictated by a latent variable $\mathbf{z}_{mn}$, which is a one-of-K vector drawn from $\boldsymbol{\theta}_m$. This means the model assigns a topic for each word $\mathbf{w}_{mn}$, and it can be different for each word in a document.

Once the topic for $\mathbf{w}_n$ is determined by $\mathbf{z}_{mn}$ as k, then the word assignment (which v of V is assigned) is drawn from a probability distribution $\beta_{[k:*]}$. β is a $K \times V$ matrix, and $\beta_{[k:*]}$ represents a multinomial distribution for topic k. Each $\beta_{[k:*]}$ is drawn from a common uniform Direchlet prior η. Here we denote the row $\beta_{[k:*]}$ by a vector $\mathbf{b}_k$ and $0 \leq b_k^v \leq 1$ denotes the v-th element of the vector $\mathbf{b}_k$ for notational convenience.

$$p(\mathbf{b}_k|\eta) = Dir(\mathbf{b}_k|\eta) = \frac{\Gamma(V\eta)}{\Gamma(\eta)^V} \prod_{v=1}^{V} (b_k^v)^{\eta - 1} \tag{2}$$

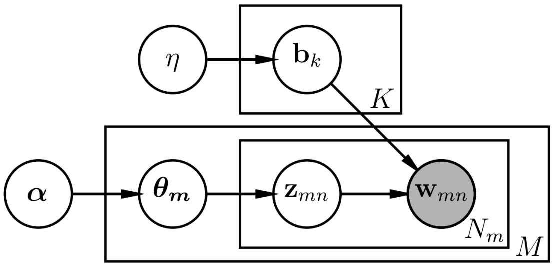

Figure 1: Generative Model

$\boldsymbol{\alpha}$ and η are learned empirically using a version of EM. The model for a document W_m is expressed as a marginalization of the joint distribution $p(W_m, \boldsymbol{\theta}_m, Z_m, \beta|\boldsymbol{\alpha}, \eta)$ over $\boldsymbol{\theta}_m$, Z_m, β where $Z_m = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_{N_m})$ as follows. The joint distribution is:

$$\begin{aligned} p(W_m, \boldsymbol{\theta}_m, Z_m, \beta|\boldsymbol{\alpha}, \eta) &= p(\boldsymbol{\theta}_m|\boldsymbol{\alpha}) \prod_{k=1}^{K} (p(\mathbf{b}_k|\eta)) \prod_{n=1}^{N_m} (p(\mathbf{z}_{mn}|\boldsymbol{\theta}_m) p(\mathbf{w}_{mn}|\mathbf{z}_{mn}, \beta)) \\ &= Dir(\boldsymbol{\theta}_m|\boldsymbol{\alpha}) \prod_{k=1}^{K} (Dir(\mathbf{b}_k|\eta)) \prod_{n=1}^{N_m} \left(\prod_{k=1}^{K} (\theta_m^k)^{z_{mn}^k} (b_k^{v_{mn}})^{z_{mn}^k} \right) \end{aligned} \tag{3}$$

where $z_{mn}^k \in \{0, 1\}$ denotes the k-th element of the vector $\mathbf{z}_{mn}$, and v_{mn} denotes the column in β (v_{mn}-th element of the vector $\mathbf{b}_k$) such that $(\mathbf{w}_{mn})^{v_{mn}} = 1$. And the marginalized ditribution for W_m is:

$$\begin{aligned} p(W_m|\boldsymbol{\alpha}, \eta) = &\int_{|\mathbf{b}_1|=1,\, 0 \le \mathbf{b}_1 \le 1} \int_{|\mathbf{b}_2|=1,\, 0 \le \mathbf{b}_2 \le 1} \cdots \int_{|\mathbf{b}_K|=1,\, 0 \le \mathbf{b}_K \le 1} \\ &\int_{|\boldsymbol{\theta}_m|=1,\, 0 \le \boldsymbol{\theta}_m \le 1} \sum_{Z_m} p(W_m, \boldsymbol{\theta}_m, Z_m, \beta|\boldsymbol{\alpha}, \eta) d\boldsymbol{\theta}_m d\mathbf{b}_K \cdots d\mathbf{b}_2 d\mathbf{b}_1 \end{aligned} \tag{4}$$

where the summing over Z_m indicates all the possible K^{N_m} combinations of $Z_m = (\mathbf{z}_{m1}, \mathbf{z}_{m1}, \cdots, \mathbf{z}_{N_m})$. Clearly this summation itself is already intractable. The corpus is modeled as follows.

$$p(\mathcal{D}|\boldsymbol{\alpha}, \eta) = \prod_{m} p(W_m|\boldsymbol{\alpha}, \eta) \tag{5}$$

The equation 5 is the basis for the EM algorithm assuming each document $\mathbf{w}_m$ is i.i.d.

3 Training and Inference

Basically we employ empirical Bayes in which we use MLE to learn $\boldsymbol{\alpha}$ and η from the data. We define **training** as the point estimate of $\boldsymbol{\alpha}$ and η such that the posterior is maximized in MLE for the observed $\mathcal{D}$.

$$\begin{aligned}\boldsymbol{\alpha}_{max}, \eta_{max} &= \underset{\boldsymbol{\alpha},\eta}{\operatorname{argmax}} \left(\ln p(\mathcal{D}|\boldsymbol{\alpha}, \eta)\right) \\ &= \underset{\boldsymbol{\alpha},\eta}{\operatorname{argmax}} \left(\sum_{m=1}^{M} \ln p(W_m, \boldsymbol{\theta}_m, Z_m, \beta|\boldsymbol{\alpha}, \eta) \right) \\ &= \underset{\boldsymbol{\alpha},\eta}{\operatorname{argmax}} \Big(\sum_{m=1}^{M} \Big(\int \cdots \int \int \sum_{Z_m} \ln p(W_m, \boldsymbol{\theta}_m, Z_m, \beta|\boldsymbol{\alpha}, \eta) d\boldsymbol{\theta}_m d\mathbf{b}_K \cdots d\mathbf{b}_1 \Big) \Big)\end{aligned} \tag{6}$$

This is not a tractable problem and we use an EM algorithm in which $\boldsymbol{\theta}_m$, Z_m, and β are treated as latent variables. For the EM-algorithm, we define the evidence lower bound as follows.

$$ELBO_m = \mathbb{E}_{\boldsymbol{\theta}_m, Z_m, \beta \sim q_m(\boldsymbol{\theta}_m, Z_m, \beta)} [\ln p(W_m, \boldsymbol{\theta}_m, Z_m, \beta|\boldsymbol{\alpha}, \eta) - \ln q_m(\boldsymbol{\theta}_m, Z_m, \beta)] \tag{7}$$

where $q_m(\boldsymbol{\theta}_m, Z_m, \beta)$ is a certain joint probability distribution for $\boldsymbol{\theta}_m$, Z_m, and β.

3.1 E-step

In the E-step of the normal EM algorithm, we identify $q_m(\boldsymbol{\theta}_m, Z_m, \beta)$ with the posterior with certain fixed $\boldsymbol{\alpha}^*$ and η^*.

$$\begin{aligned}q_m^*(\boldsymbol{\theta}_m, Z_m, \beta) &= \underset{q_m(\boldsymbol{\theta}_m, Z_m, \beta)}{\operatorname{argmax}} (ELBO_m) \\ &= p(\boldsymbol{\theta}_m, Z_m, \beta|W_m, \boldsymbol{\alpha}^*, \eta^*)\end{aligned} \tag{8}$$

However as we later explain, $p(\boldsymbol{\theta}_m, Z_m, \beta|W_m, \boldsymbol{\alpha}^*, \eta^*)$ is still intractable, and we need an approximation of it. One way of approximation is to use variational mean field approximation by breaking conditional dependences among $\boldsymbol{\theta}_m$, Z_m, and β. This is further explained in the following section.

3.2 M-step

In the M-step we raise $ELBO_m$ w.r.t. $\boldsymbol{\alpha}$ and η with fixed $q_m^*(\boldsymbol{\theta}_m, Z_m, \beta)$.

$$
\begin{aligned}
\boldsymbol{\alpha}^*, \eta^* &= \operatorname*{argmax}_{\boldsymbol{\alpha},\eta} \left(\sum_m ELBO_m \right) \\
&= \sum_m \mathbb{E}_{\boldsymbol{\theta}_m, Z_m, \beta \sim q^*_m(\boldsymbol{\theta}_m, Z_m, \beta)} [\ln p(W_m, \boldsymbol{\theta}_m, Z_m, \beta | \boldsymbol{\alpha}, \eta) - \ln q^*_m(\boldsymbol{\theta}_m, Z_m, \beta)] \\
&= \sum_m \mathbb{E}_{\boldsymbol{\theta}_m, Z_m, \beta \sim q^*_m(\boldsymbol{\theta}_m, Z_m, \beta)} [\ln p(W_m, \boldsymbol{\theta}_m, Z_m, \beta | \boldsymbol{\alpha}, \eta)]
\end{aligned} \tag{9}
$$

The RHS of equation 9 is concave in terms of $\boldsymbol{\alpha}$, η, and this is solved by taking the derivative of it.

$$
\nabla_{\boldsymbol{\alpha},\eta} \left(\sum_m \mathbb{E}_{\boldsymbol{\theta}_m, Z_m, \beta \sim q^*_m(\boldsymbol{\theta}_m, Z_m, \beta)} [\ln p(W_m, \boldsymbol{\theta}_m, Z_m, \beta | \boldsymbol{\alpha}, \eta)] \right) = 0 \tag{10}
$$

3.3 Topic Vector and Per-topic Word Probability Distribution

After the training has converged, we have the following, which together maximize $\sum ELBO_m$: $\boldsymbol{\alpha}^*$, η^*, $q^*(\boldsymbol{\theta}_m)$ & $q^*(Z_m)$ for each W_m, and $q^*(\beta)$. $q^*(\boldsymbol{\theta}_m)$ can be considered a topic vector for the document W_m, $q^*(Z_m)$ is further factorized into $\prod_{n=1}^{N_m} q^*(\mathbf{z}_{mn})$ and $q^*(\mathbf{z}_{mn})$ is considered a word-level topic vector, which can also be considered a document-dependent word embedding. $q^*(\beta)$ is further factorized into $\prod_{k=1}^{K} q^*(\mathbf{b}_k)$ and each $q^*(\mathbf{b}_k)$ can be considered the probability distribution over all the words for that topic.

3.4 Topic Inference for a New Document

We define **inference** of a new Document W_{new} as finding $q^{max}(\boldsymbol{\theta}_{new}, Z_{new})$ such that $q^{max}(\boldsymbol{\theta}_{new}, Z_{new})$ maximizes $ELBO_{new}$ for the fixed $\boldsymbol{\alpha}$ and η as well as $q^*(\beta)$. This is solved by the mean value approximation, and it is further explained in the following section. $q^{max}(\boldsymbol{\theta}_{new})$ can be considered the infered topic.

As a single point estimate, we can apply MAP on $q(\boldsymbol{\theta})$. In fact, as explained in the subsection 4.1, $q(\boldsymbol{\theta})$ is formed as $Dir(\boldsymbol{\theta}|\boldsymbol{\gamma})$ and $\boldsymbol{\gamma}$ is numerically computerd. Then we can find the mode as follows.

$$
\theta^k_{MAP} = \frac{\gamma^k - 1}{\sum_{j=1}^{K} \gamma^j - K} \tag{11}
$$

However, the original article [1] simply suggests to use $\boldsymbol{\gamma}$ as the topic vector.

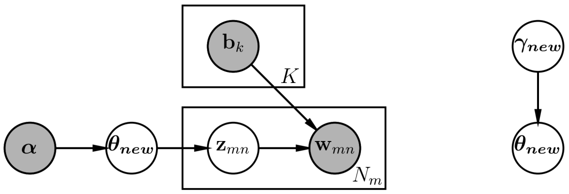

Figure 2: Infering θ_{new}. The left side is the modeling and the right side is the approximation of θ_{new}

4 Variational E-Step

This is the inner loop of the algorithm for approximating the posterior with the variational mean field approximation EM requires to represent the terms with $\boldsymbol{\theta}_m$, Z_m, and β in equation 21 in the expectation of the following conditional distribution.

$$p(\boldsymbol{\theta}_m, Z_m, \beta | W_m, \boldsymbol{\alpha}, \eta) = \frac{p(W_m, \boldsymbol{\theta}_m, Z_m, \beta | \boldsymbol{\alpha}, \beta)}{\sum_{W_m} p(W_m, \boldsymbol{\theta}_m, Z_m, \beta | \boldsymbol{\alpha}, \eta)} \tag{12}$$

However, both LHS and the denominator of RHS are intractable. Especially the denominator of RHS sums over all the elements of W_m, whose number is V^{N_m}. It is not factorized per word $\mathbf{w}_{mn}$ due to the explaining-away effect between $\mathbf{z}_{mn}$ for all n and $\mathbf{b}_k$ for all k conditioned on $\mathbf{w}_{mn}$. Therefore, an approximation of this is required. For that we use mean field approximation as follows.

$$\begin{aligned} p(\boldsymbol{\theta}_m, Z_m, \beta | \mathbf{w}_m, \boldsymbol{\alpha}, \eta) &\approx q(\boldsymbol{\theta}_m, Z_m, \beta) \\ &= q(\boldsymbol{\theta}_m) \prod_{n=1}^{N_m} q(\mathbf{z}_{mn}) \prod_{k=1}^{K} q(\mathbf{b}_k) \end{aligned} \tag{13}$$

Please note that $q(\mathbf{b}_k)$ are not per-document, but corpus-wise approximators.

4.1 Approximator $q(\boldsymbol{\theta}_m)$

Since $0 \leq \boldsymbol{\theta}_m \leq 1$ is dictated by a Direchlet distribution, we introduce a variational parameter $\boldsymbol{\gamma}_m$ such that $q(\boldsymbol{\theta}_m) = Dir(\boldsymbol{\theta}_m | \boldsymbol{\gamma}_m)$. The following holds from the properties of the Direchlet distribution.

$$\mathbb{E}_{\boldsymbol{\theta}_m \sim Dir(\boldsymbol{\theta}_m | \boldsymbol{\gamma}_m)}[\theta_m^i] = \frac{\gamma_m^i}{\sum_{k=1}^{K} \gamma_m^k} \tag{14}$$

$$\mathbb{E}_{\boldsymbol{\theta}_m \sim Dir(\boldsymbol{\theta}_m | \boldsymbol{\gamma}_m)}[\ln(\theta_m^i)] = \Psi(\gamma_m^i) - \Psi(\sum_{k=1}^{K} \gamma_m^i) \tag{15}$$

where $\Psi(x) = \frac{d\Gamma(x)}{dx}$ is a Digamma function (the first derivative of the gamma function). Digamma and Trigamma functions can be numerically computerd by approximation.

4.2 Approximator $q(\mathbf{b}_k)$

$0 \leq \mathbf{b}_k \leq 1$ is dictated by a Direchlet distribution, and we introduce a variational parameter $\boldsymbol{\omega}_k$ such that $q(\mathbf{b}_k) = Dir(\mathbf{b}_k|\boldsymbol{\omega}_k)$.

$$\mathbb{E}_{\mathbf{b}_k \sim Dir(\mathbf{b}_k|\boldsymbol{\omega}_k)}[b_k^v] = \frac{\omega_k^v}{\sum_{v=1}^V \omega_k^v} \tag{16}$$

$$\mathbb{E}_{\mathbf{b}_k \sim Dir(\mathbf{b}_k|\boldsymbol{\omega}_k)}[\ln(b_k^v)] = \Psi(\omega_k^v) - \Psi(\sum_{v=1}^V \omega_k^v) \tag{17}$$

4.3 Approximator $q(\mathbf{z}_{mn})$

$\mathbf{z}_{mn}$ is a multinomial distribution, we introduce a variational parameter $\boldsymbol{\phi}_{mn} \in \mathcal{R}^K, |\boldsymbol{\phi}_{mn}| = 1$, $0 \leq \boldsymbol{\phi}_{mn} \leq 1$ such that $q(\mathbf{z}_{mn}) = Mult(\mathbf{z}_{mn}|\ \boldsymbol{\phi}_{mn}) = \prod_{k=1}^K (\boldsymbol{\phi}_{mn}^k)^{\mathbf{z}_{mn}^k}$. The following trivially holds.

$$\mathbb{E}_{\mathbf{z}_{mn} \sim Multi(\mathbf{z}_{mn}|\boldsymbol{\phi}_{mn})}[\mathbf{z}_{mn}] = \boldsymbol{\phi}_{mn} \tag{18}$$

4.4 Putting all the approximators together

The equation 48 is reformed as follows.

$$\begin{aligned} q(\boldsymbol{\theta}_m, Z_m, \beta) &= q(\boldsymbol{\theta}_m|\boldsymbol{\gamma}) \prod_{k=1}^K q(\mathbf{b}_k|\boldsymbol{\omega}_k) \prod_{n=1}^{N_m} q(\mathbf{z}_{mn}|\boldsymbol{\phi}_{mn}) \\ &= Dir(\boldsymbol{\theta}_m|\boldsymbol{\gamma}) \prod_{k=1}^K Dir(\mathbf{b}_k|\boldsymbol{\omega}_k) \prod_{n=1}^{N_m} Mult(\mathbf{z}_{mn}|\boldsymbol{\phi}_{mn}) \\ &= Dir(\boldsymbol{\theta}_m|\boldsymbol{\gamma}) \prod_{k=1}^K Dir(\mathbf{b}_k|\boldsymbol{\omega}_k) \prod_{n=1}^{N_m} \prod_{v=1}^V (\phi_{mn}^v)^{z_{mn}^v} \end{aligned} \tag{19}$$

4.5 Evidence Lower Bound with the Approximators

We form a variational ELBO for the observed document W_m as follows.

$$ELBO_m = \mathbb{E}_{\boldsymbol{\theta}_m, Z_m, \beta \sim q(\boldsymbol{\theta}_m, Z_m, \beta)}[\ln p(W_m, \boldsymbol{\theta}_m, Z_m, \beta|\boldsymbol{\alpha}, \eta) - \ln q(\boldsymbol{\theta}_m, Z_m, \beta)] \tag{20}$$

The first term on the RHS of the equation 20 will be:

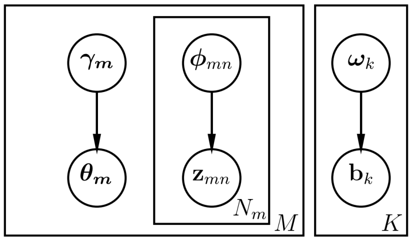

Figure 3: Approximator $q(\boldsymbol{\theta}_m, Z_m, \beta)$

$$\begin{aligned}
ELBO_{pm} &= \mathbb{E}[\ln p(W_m, \boldsymbol{\theta}_m, Z_m, \beta|\boldsymbol{\alpha}, \eta)] \\
&= \mathbb{E}\left[\ln Dir(\boldsymbol{\theta}_m|\boldsymbol{\alpha}) + \sum_{k=1}^{K} (\ln Dir(\mathbf{b}_k|\eta)) + \sum_{n=1}^{N_m}\sum_{k=1}^{K} z_{mn}^k \left(\ln \theta_m^k + \ln b_k^{v_{mn}}\right) \right] \\
&= \ln\left(\Gamma(\sum_k \alpha_k)\right) - \sum_{k=1}^{K} \ln\Gamma(\alpha_k) + \sum_{k=1}^{K} (\alpha_k - 1)\mathbb{E}_{\boldsymbol{\theta_m}\sim q(\boldsymbol{\theta_m})}[\ln(\theta_m^k)] \\
&\quad + \sum_{k=1}^{K}\left(\ln\left(\Gamma(K\eta)\right) - K\ln\Gamma(\eta) + \sum_{v=1}^{V}(\eta - 1)\mathbb{E}_{\mathbf{b}_k\sim q(\mathbf{b}_k)}[\ln(b_k^v)] \right) \\
&\quad + \sum_{n=1}^{N_m}\sum_{k=1}^{K} \mathbb{E}_{\mathbf{z}_{mn}\sim q(\mathbf{z_{mn}})}[z_{mn}^k] \left(\mathbb{E}_{\boldsymbol{\theta_m}\sim q(\boldsymbol{\theta_m})}[\ln \theta_m^k] + \mathbb{E}_{\mathbf{b}_k\sim q(\mathbf{b}_k)}[\ln b_k^{v_{mn}}] \right)
\end{aligned} \tag{21}$$

where we used the assumed independence of the approximator as follows.

$$\mathbb{E}_{\mathbf{z}_{mn},\boldsymbol{\theta_m}\sim q(\mathbf{z}_{mn})q(\boldsymbol{\theta_m})}[z_{mn}^k \ln \theta_m^k] = \left(\mathbb{E}_{\mathbf{z}_{mn}\sim q(\mathbf{z}_{mn})}[z_{mn}^k]\right)\left(\mathbb{E}_{\boldsymbol{\theta}_m\sim q(\boldsymbol{\theta}_m)}[\ln \theta_m^k]\right) \tag{22}$$

$$\mathbb{E}_{\mathbf{z}_{mn},\mathbf{b}_{\boldsymbol{k}}\sim q(\mathbf{z}_{mn})q(\mathbf{b}_k)}[z_{mn}^k \ln b_m^k] = \left(\mathbb{E}_{\mathbf{z}_{mn}\sim q(\mathbf{z}_{mn})}[z_{mn}^k]\right)\left(\mathbb{E}_{\mathbf{b}_k\sim q(\mathbf{b}_k)}[\ln b_k^v]\right) \tag{23}$$

The second term on the RHS of the equation 20 will be:

$$
\begin{aligned}
ELBO_{qm} &= -\mathbb{E}\left[\ln q(\boldsymbol{\theta}_m, Z_m, \beta)\right] \\
&= -\mathbb{E}\left[\ln Dir(\boldsymbol{\theta}_m|\boldsymbol{\gamma}) + \sum_{k=1}^{K} \ln Dir(\mathbf{b}_k|\boldsymbol{\omega}_k) + \sum_{n=1}^{N_m}\sum_{v=1}^{V} z_{mn}^{v} \ln \phi_{mn}^{v}\right] \\
&= -\ln\left(\Gamma(\sum_k \gamma_m^k)\right) + \sum_{k=1}^{K} \ln\Gamma(\gamma_m^k) - \sum_{k=1}^{K}(\gamma_m^k - 1)\mathbb{E}_{\boldsymbol{\theta_m}\sim q(\boldsymbol{\theta_m})}[\ln(\theta_m^k)] \\
&\quad + \sum_{k=1}^{K}\left(-\ln\left(\Gamma(\sum_v \omega_k^v)\right) + \sum_{v=1}^{V}\ln\Gamma(\omega_k^v) - \sum_{v=1}^{V}(\omega_k^v - 1)\mathbb{E}_{\mathbf{b}_k\sim q(\mathbf{b}_k)}[\ln(b_k^v)]\right) \\
&\quad - \sum_{n=1}^{N_m}\sum_{k=1}^{K}\mathbb{E}_{z_{mn}\sim q(z_{mn})}[z_{mn}^k]\ln\phi_{mn}^k
\end{aligned}
\tag{24}
$$

4.6 Raising ELBO by Finding a Better $q(\boldsymbol{\theta}_m, Z_m, \beta)$

Now we have formed $ELBO = ELBO_{pm} + ELBO_{qm}$ in terms of the expectations. We are ready to raise ELBO in turn for each of $q(\boldsymbol{\theta}_m)$, $q(\mathbf{z}_{mn})$, and $q(\mathbf{b}_k)$.

4.6.1 Finding a Better $q(\boldsymbol{\theta}_m)$

First we raise ELBO in terms of $q(\boldsymbol{\theta}_m)$. For that we gather the relevant terms in ELBO into the following.

$$
\begin{aligned}
\mathcal{L}_{q(\boldsymbol{\theta}_m)} &= \sum_{k=1}^{K}(\alpha_k - 1)\mathbb{E}_{\boldsymbol{\theta_m}\sim q(\boldsymbol{\theta_m})}[\ln(\theta_m^k)] \\
&\quad + \sum_{n=1}^{N_m}\sum_{k=1}^{K}\mathbb{E}_{\mathbf{z}_{mn}\sim q(\mathbf{z_{mn}})}[z_{mn}^k]\left(\mathbb{E}_{\boldsymbol{\theta_m}\sim q(\boldsymbol{\theta_m})}[\theta_m^k]\right) \\
&\quad - \ln\left(\Gamma(\sum_k \gamma_m^k)\right) + \sum_{k=1}^{K}\ln\Gamma(\gamma_m^k) - \sum_{k=1}^{K}(\gamma_m^k - 1)\mathbb{E}_{\boldsymbol{\theta_m}\sim q(\boldsymbol{\theta_m})}[\ln(\theta_m^k)]
\end{aligned}
\tag{25}
$$

In the following we gradually reorganize the RHS of equation 25 in terms of γ_m^k, and remove the unnecessary terms step by step.

$$
\begin{aligned}
\mathcal{L}'_{q(\boldsymbol{\theta}_m)} &= \sum_{k=1}^{K} (\alpha_k - 1) \left(\Psi(\gamma_m^k) - \Psi(\sum_{j=1}^{K} \gamma_m^j) \right) \\
&+ \sum_{n=1}^{N_m} \sum_{k=1}^{K} \phi_{mn}^k \left(\Psi(\gamma_m^k) - \Psi(\sum_{j=1}^{K} \gamma_m^j) \right) \\
&- \ln \left(\Gamma(\sum_k \gamma_m^k) \right) + \sum_{k=1}^{K} \ln \Gamma(\gamma_m^k) - \sum_{k=1}^{K} (\gamma_m^k - 1) \left(\Psi(\gamma_m^k) - \Psi(\sum_{j=1}^{K} \gamma_m^j) \right)
\end{aligned} \tag{26}
$$

$$
\begin{aligned}
\mathcal{L}''_{q(\boldsymbol{\theta}_m)} = &\left(\Psi(\gamma_m^k) - \Psi(\sum_{j=1}^{K} \gamma_m^j) \right) \left(\sum_{k=1}^{K} (\alpha_k - 1) + \sum_{n=1}^{N_m} \sum_{k=1}^{K} \phi_{mn}^k - \sum_{k=1}^{K} (\gamma_m^k - 1) \right) \\
&- \ln \left(\Gamma(\sum_k \gamma_m^k) \right) + \sum_{k=1}^{K} \ln \Gamma(\gamma_m^k)
\end{aligned} \tag{27}
$$

$$
\begin{aligned}
\mathcal{L}'''_{q(\boldsymbol{\theta}_m)} = &\left(\Psi(\gamma_m^k) - \Psi(\sum_{j=1}^{K} \gamma_m^j) \right) \left(\sum_{k=1}^{K} \left(\alpha_k - \gamma_m^k + \sum_{n=1}^{N_m} \phi_{mn}^k \right) \right) \\
&- \ln \left(\Gamma(\sum_k \gamma_m^k) \right) + \sum_{k=1}^{K} \ln \Gamma(\gamma_m^k)
\end{aligned} \tag{28}
$$

Then take a derivative of $\mathcal{L}'''_{q(\boldsymbol{\theta}_m)}$ w.r.t.γ_m^k and equate it with zero.

$$
\begin{aligned}
\frac{\partial \mathcal{L}_{q(\boldsymbol{\theta}_m)}}{\partial \gamma_m^k} = &\left(\Psi'(\gamma_m^k) - \Psi'(\sum_{j=1}^{K} \gamma_m^j) \right) \left(\sum_{k=1}^{K} \left(\alpha_k - \gamma_m^k + \sum_{n=1}^{N_m} \phi_{mn}^k \right) \right) \\
&- \left(\Psi(\gamma_m^k) - \Psi(\sum_{j=1}^{K} \gamma_m^j) \right) \\
&- \Psi(\sum_k \gamma_m^k) + \sum_{k=1}^{K} \Psi(\gamma_m^k) \\
= &\left(\Psi'(\gamma_m^k) - \Psi'(\sum_{j=1}^{K} \gamma_m^j) \right) \left(\sum_{k=1}^{K} \left(\alpha_k - \gamma_m^k + \sum_{n=1}^{N_m} \phi_{mn}^k \right) \right) = 0
\end{aligned} \tag{29}
$$

In order for this to hold for all $1 \leq k \leq K$, the following must hold.

$$\sum_{k=1}^{K} \left(\alpha_k - \gamma_m^k + \sum_{n=1}^{N_m} \phi_{mn}^k \right) = 0 \tag{30}$$

and eventually $\forall\, k, (1 \leq k \leq K)$,

$$\alpha_k - \gamma_m^k + \sum_{n=1}^{N_m} \phi_{mn}^k = 0 \tag{31}$$

and reorganized into

$$\gamma_m^k = \alpha_k + \sum_{n=1}^{N_m} \phi_{mn}^k \tag{32}$$

Equation 32 is the update step for γ_m^k with which $q(\boldsymbol{\theta}_m)$ raises $ELBO_m$.

4.6.2 Finding a Better $q(\mathbf{z}_{mn})$

Second we raise ELBO in terms of $q(\mathbf{z}_{mn})$. For that we gather the relevant terms in ELBO into the following.

$$\begin{aligned} \mathcal{L}_{q(\mathbf{z}_{mn})} = &\sum_{n=1}^{N_m} \sum_{k=1}^{K} \phi_{mn}^k \left(\mathbb{E}_{\boldsymbol{\theta_m} \sim q(\boldsymbol{\theta_m})}[\ln \theta_m^k] + \mathbb{E}_{\mathbf{b}_k \sim q(\mathbf{b}_k)}[\ln b_k^{v_{mn}}] \right) \\ &- \sum_{n=1}^{N_m} \sum_{k=1}^{K} \phi_{mn}^k \ln \phi_{mn}^k \end{aligned} \tag{33}$$

Removing the unnecessary terms from $\mathcal{L}_{q(\mathbf{z}_{mn})}$ in terms of ϕ_{mn}^k,

$$\begin{aligned} \mathcal{L}'_{q(\mathbf{z}_{mn})} = &\sum_{k=1}^{K} \phi_{mn}^k \left(\mathbb{E}_{\boldsymbol{\theta_m} \sim q(\boldsymbol{\theta_m})}[\ln \theta_m^k] + \mathbb{E}_{\mathbf{b}_k \sim q(\mathbf{b}_k)}[\ln b_k^{v_{mn}}] \right) \\ &- \sum_{k=1}^{K} \phi_{mn}^k \ln \phi_{mn}^k \end{aligned} \tag{34}$$

We need an equality constraint $\sum_{k=1}^{K}(\phi_{mn}^k) = 1$, and need to form a Lagrangian multiplier as follows.

$$\begin{aligned} \mathcal{L}(\boldsymbol{\phi}_{mn}, \lambda) = &\sum_{k=1}^{K} \phi_{mn}^k \left(\mathbb{E}_{\boldsymbol{\theta_m} \sim q(\boldsymbol{\theta_m})}[\ln \theta_m^k] + \mathbb{E}_{\mathbf{b}_k \sim q(\mathbf{b}_k)}[\ln b_k^{v_{mn}}] \right) \\ &- \sum_{k=1}^{K} \phi_{mn}^k \ln \phi_{mn}^k - \lambda \left(\sum_{k=1}^{K}(\phi_{mn}^k) - 1 \right) \end{aligned} \tag{35}$$

Then take a derivative of $\mathcal{L}(\boldsymbol{\phi}_{mn}, \lambda)$ w.r.t. ϕ^k_{mn} and equate it with zero.

$$\begin{aligned}\frac{\partial \mathcal{L}(\boldsymbol{\phi}_{mn}, \lambda)}{\partial \phi^k_{mn}} &= \frac{\partial}{\partial \phi^k_{mn}} \Big(\phi^k_{mn} \left(\mathbb{E}_{\boldsymbol{\theta_m} \sim q(\boldsymbol{\theta_m})}[\ln \theta^k_m] + \mathbb{E}_{\mathbf{b}_k \sim q(\mathbf{b}_k)}[\ln b_k^{v_{mn}}] \right) \\ &\qquad - \phi^k_{mn} \ln \phi^k_{mn} - \lambda \phi^k_{mn}) \Big) \qquad (36) \\ &= \left(\mathbb{E}_{\boldsymbol{\theta_m} \sim q(\boldsymbol{\theta_m})}[\ln \theta^k_m] + \mathbb{E}_{\mathbf{b}_k \sim q(\mathbf{b}_k)}[\ln b_k^{v_{mn}}] \right) - \ln \phi^k_{mn} - 1 - \lambda \\ &= 0\end{aligned}$$

This implies

$$\begin{aligned}\phi^k_{mn} &\approx \exp\left(\mathbb{E}_{\boldsymbol{\theta_m} \sim q(\boldsymbol{\theta_m})}[\ln \theta^k_m] + \mathbb{E}_{\mathbf{b}_k \sim q(\mathbf{b}_k)}[\ln b_k^{v_{mn}}] \right) \\ &\approx \exp\left(\Psi(\gamma^k_m) - \Psi(\sum_{j=1}^{K} \gamma^j_m) + \Psi(\omega_k^{v_{mn}}) - \Psi(\sum_{j=1}^{V} \omega^j_k) \right) \qquad (37) \\ &\approx \exp\left(\Psi(\gamma^k_m) + \Psi(\omega_k^{v_{mn}}) - \Psi(\sum_{j=1}^{V} \omega^j_k) \right).\end{aligned}$$

and we let

$$\hat{\phi}^k_{mn} = \exp\left(\Psi(\gamma^k_m) + \Psi(\omega_k^{v_{mn}} - \Psi(\sum_{j=1}^{V} \omega^j_k)) \right) \qquad (38)$$

and

$$\phi^k_{mn} = \frac{\hat{\phi}^k_{mn}}{\sum_{j=1}^{K} \hat{\phi}^k_{mn}} \qquad (39)$$

Equation 38 and 39 are the update step for ϕ^k_{mn} with which $q(\boldsymbol{\phi}_{mn})$ raises $ELBO_m$.

4.6.3 Finding a Better $q(\mathbf{b}_k)$

Third we raise ELBO in terms of $q(\mathbf{b}_k)$. Please note that $q(\mathbf{b}_k)$ is a corus-wide approximator that does not depend on particular document W_m. For that we gather the relevant terms in ELBO into the following.

$$\begin{aligned}\mathcal{L}_{q(\mathbf{b}_k)} = &\sum_{k=1}^{K} \left(\ln\left(\Gamma(K\eta)\right) - K \ln \Gamma(\eta) + \sum_{v=1}^{V} (\eta - 1) \mathbb{E}_{\mathbf{b}_k \sim q(\mathbf{b}_k)}[\ln(b^v_k)] \right) \\ &+ \sum_{n=1}^{N_m} \sum_{k=1}^{K} \mathbb{E}_{\mathbf{z}_{mn} \sim q(\mathbf{z_{mn}})}[z^k_{mn}] \left(\mathbb{E}_{\boldsymbol{\theta_m} \sim q(\boldsymbol{\theta_m})}[\ln \theta^k_m] + \mathbb{E}_{\mathbf{b}_k \sim q(\mathbf{b}_k)}[\ln b_k^{v_{mn}}] \right) \\ &+ \sum_{k=1}^{K} \left(- \ln\left(\Gamma(\sum_{v} \omega^v_k) \right) + \sum_{v=1}^{V} \ln \Gamma(\omega^v_k) - \sum_{v=1}^{V} (\omega^v_k - 1) \mathbb{E}_{\mathbf{b}_k \sim q(\mathbf{b}_k)}[\ln(b^v_k)] \right) \\ &\qquad (40)\end{aligned}$$

In the following we gradually reorganize the RHS of equation 40 in terms of ω_k, and remove the unnecessary terms step by step.

$$
\begin{aligned}
\mathcal{L}'_{q(\mathbf{b}_k)} = &+ \sum_{v=1}^{V} (\eta - 1) \mathbb{E}_{\mathbf{b}_k \sim q(\mathbf{b}_k)}[\ln(b_k^v)] \\
&+ \sum_{n=1}^{N_m} \mathbb{E}_{\mathbf{z}_{mn} \sim q(\mathbf{z_{mn}})}[z_{mn}^k] \left(\mathbb{E}_{\mathbf{b}_k \sim q(\mathbf{b}_k)}[\ln b_k^{v_{mn}}] \right) \\
&- \ln \left(\Gamma(\sum_v \omega_k^v) \right) + \sum_{v=1}^{V} \ln \Gamma(\omega_k^v) - \sum_{v=1}^{V} (\omega_k^v - 1) \mathbb{E}_{\mathbf{b}_k \sim q(\mathbf{b}_k)}[\ln(b_k^v)]
\end{aligned} \tag{41}
$$

$$
\begin{aligned}
\mathcal{L}''_{q(\mathbf{b}_k)} = &+ \sum_{v=1}^{V} (\eta - 1) \left(\Psi(\omega_k^v) - \Psi(\sum_{j=1}^{V} \omega_k^j) \right) \\
&+ \sum_{n=1}^{N_m} \phi_{mn}^k \left(\Psi(\omega_k^{v_{mn}}) - \Psi(\sum_{j=1}^{V} \omega_k^j) \right) \\
&- \ln \left(\Gamma(\sum_v \omega_k^v) \right) + \sum_{v=1}^{V} \ln \Gamma(\omega_k^v) - \sum_{v=1}^{V} (\omega_k^v - 1) \left(\Psi(\omega_k^v) - \Psi(\sum_{j=1}^{V} \omega_k^j) \right)
\end{aligned} \tag{42}
$$

$$
\begin{aligned}
\mathcal{L}'''_{q(\mathbf{b}_k)} = &+ \sum_{v=1}^{V} (\eta - 1) \left(\Psi(\omega_k^v) - \Psi(\sum_{j=1}^{V} \omega_k^j) \right) \\
&+ \sum_{n=1}^{N_m} \phi_{mn}^k \Psi(\omega_k^{v_{mn}}) - \sum_{n=1}^{N_m} \phi_{mn}^k \Psi(\sum_{j=1}^{V} \omega_k^j) \\
&- \ln \left(\Gamma(\sum_v \omega_k^v) \right) + \sum_{v=1}^{V} \ln \Gamma(\omega_k^v) - \sum_{v=1}^{V} (\omega_k^v - 1) \left(\Psi(\omega_k^v) - \Psi(\sum_{j=1}^{V} \omega_k^j) \right)
\end{aligned} \tag{43}
$$

Then take a derivative of $\mathcal{L}_{q(\mathbf{b}_k)}$ w.r.t. ω_k^v and equate it with zero.

$$\begin{aligned}
\frac{\partial \mathcal{L}_{q(\mathbf{b}_k)}}{\partial \omega_k^v} = \frac{\partial}{\partial \omega_k^v}\Big(& \sum_{v'=1}^{V} (\eta - 1) \left(\Psi(\omega_k^{v'}) - \Psi(\sum_{j=1}^{V} \omega_k^j) \right) \\
& + \sum_{n:v_{mn}=v} \phi_{mn}^k \Psi(\omega_k^{v_{mn}}) - \sum_{n=1}^{N_m} \phi_{mn}^k \Psi(\sum_{j=1}^{V} \omega_k^j) \\
& - \ln \left(\Gamma(\sum_{v'} \omega_k^{v'}) \right) + \sum_{v'=1}^{V} \ln \Gamma(\omega_k^{v'}) \\
& - \sum_{v'=1}^{V} (\omega_k^{v'} - 1) \left(\Psi(\omega_k^{v'}) - \Psi(\sum_{j=1}^{V} \omega_k^j) \right) \Big) \\
= \frac{\partial}{\partial \omega_k^v}\Big(& (\eta - 1)\Psi(\omega_k^v) - V(\eta - 1)\Psi(\sum_{j=1}^{V} \omega_k^j) \\
& + \sum_{n:v_{mn}=v} \phi_{mn}^k \Psi(\omega_k^v) - \sum_{n=1}^{N_m} \phi_{mn}^k \Psi(\sum_{j=1}^{V} \omega_k^j) \\
& - \ln \left(\Gamma(\sum_{v} \omega_k^{v}) \right) + \sum_{v=1}^{V} \ln \Gamma(\omega_k^{v}) \\
& - (\omega_k^v - 1)\Psi(\omega_k^v) + \sum_{v=1}^{V} (\omega_k^v - 1)\Psi(\sum_{j=1}^{V} \omega_k^j) \Big) \\
= \frac{\partial}{\partial \omega_k^v}\Big(& \Psi(\omega_k^v) \left((\eta - 1) + \sum_{n:v_{mn}=v} \phi_{mn}^k - (\omega_k^v - 1) \right) \\
& - \Psi(\sum_{j=1}^{V} \omega_k^j) \left(V(\eta - 1) + \sum_{n=1}^{N_m} \phi_{mn}^k - \sum_{v=1}^{V} (\omega_k^v - 1) \right) \\
& - \ln \left(\Gamma(\sum_{v} \omega_k^{v}) \right) + \sum_{v=1}^{V} \ln \Gamma(\omega_k^{v}) \Big)
\end{aligned} \tag{44}$$

Please note that $\sum_{n:v_{mn}=v} \phi_{mn}^k \Psi(\omega_k^{v_{mn}}) = \sum_{n:v_{mn}=v} \phi_{mn}^k \Psi(\omega_k^v)$ if k and n are both fixed. We continue the derivation as follows.

$$\begin{aligned}
\frac{\partial \mathcal{L}_{q(\mathbf{b}_k)}}{\partial \omega_k^v} &= \frac{\partial}{\partial \omega_k^v}\Big(\ \Psi(\omega_k^v)\left(\eta - \omega_k^v + \sum_{n:v_{mn}=v} \phi_{mn}^k\right) \\
&\qquad - \Psi(\sum_{j=1}^{V} \omega_k^j)\left(V\eta + \sum_{n=1}^{N_m} \phi_{mn}^k - \sum_{v=1}^{V} \omega_k^v\right) \ \Big) \\
&= \frac{\partial}{\partial \omega_k^v}\Big(\ \Psi(\omega_k^v)\left(-(\omega_k^v - \eta) + \sum_{n:v_{mn}=v} \phi_{mn}^k\right) \\
&\qquad - \Psi(\sum_{j=1}^{V} \omega_k^j)\left(-\sum_{v=1}^{V}(\omega_k^v - \eta) + \sum_{n=1}^{N_m} \phi_{mn}^k\right) \ \Big) = 0
\end{aligned} \tag{45}$$

In order for this to hold for all $1 \leq v \leq V$, the following must hold.

$$\begin{aligned}
-(\omega_k^v - \eta) + \sum_{n:v_{mn}=v} \phi_{mn}^k = 0 \\
\omega_k^v = \eta + \sum_{n:v_{mn}=v} \phi_{mn}^k
\end{aligned} \tag{46}$$

In fact, it must hold for the entire corpus, then

$$\omega_k^v = \eta + \sum_{\mathbf{w}_m, m=1}^{M}\left(\sum_{n:v_{mn}=v} \phi_{mn}^k\right) \tag{47}$$

Equation 47 is the update step for ω_k^v with which $q(\boldsymbol{\omega_k})$ raises $ELBO_m$.

4.7 The Variational E-step with $q(\boldsymbol{\theta}_m, Z_m, \beta)$

We can iterate this process of finding $q(\boldsymbol{\theta}_m)$, $q(\mathbf{z}_{mn})$, and $q(\mathbf{b}_k)$, in terms of finding better $\boldsymbol{\gamma}_m$, $\boldsymbol{\phi}_{mn}$, and $\boldsymbol{\omega}_k$ until $\sum_m ELBO_m$ convergeances.

5 M-step

This is the outer Loop of the training algorithm with the fixed approximated posterior. In section 4, we have approximated $p(\boldsymbol{\theta}_m, \mathbf{z}_m, \beta | \mathbf{w}_m, \boldsymbol{\alpha}, \eta)$ with

$$q(\boldsymbol{\theta}_m) \prod_{k=1}^{K} q(\mathbf{b}_k) \prod_{n=1}^{N_m} q(\mathbf{z}_{mn}) \tag{48}$$

This corresponds to the E-step of the EM-altorighm. In this section we develop the M-step where we raise ELBO by finding better $\boldsymbol{\alpha}$ and η. Please note this is just a simple maximization problem on $\boldsymbol{\alpha}$ and η in MLE setting without any probabilistic treatment for them.

First we raise ELBO w.r.t. $\boldsymbol{\alpha}$. Gathering the relevant terms of the equations 23 and 24 into the following for each document $\mathbf{w}_m$.

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\alpha}}(m) &= \ln\left(\Gamma(\sum_k \alpha_k)\right) - \sum_{k=1}^{K}\ln\Gamma(\alpha_k) + \sum_{k=1}^{K}(\alpha_k - 1)\mathbb{E}_{\boldsymbol{\theta_m}\sim q(\boldsymbol{\theta_m})}[\ln(\theta_m^k)] \\ &= \ln\left(\Gamma(\sum_k \alpha_k)\right) - \sum_{k=1}^{K}\ln\Gamma(\alpha_k) + \sum_{k=1}^{K}(\alpha_k - 1)\left(\Psi(\gamma_m^k) - \Psi(\sum_{j=1}^{K}\gamma_m^j)\right)\end{aligned} \tag{49}$$

Taking the derivative of α_k over the entier corpus,

$$\frac{\partial\left(\sum_{m=1}^{M}\mathcal{L}_{\boldsymbol{\alpha}}(m)\right)}{\partial\alpha_k} = M\left(\Psi(\sum_{j=1}^{K}\alpha_j) - \Psi(\alpha_k)\right) + \sum_{m=1}^{M}\left(\Psi(\gamma_m^k) - \Psi(\sum_{j=1}^{K}\gamma_m^j)\right) = 0 \tag{50}$$

This can not be analytically solved, and we need to numerically compute it. The original article suggests Newton-Raphson algorithm that runs in $O(K)$. This is explained in section 6.

Second we raise ELBO w.r.t. η. Gathering the relevant terms of the equations 23 and 24 into the following for each document $\mathbf{w}_m$.

$$\begin{aligned}\mathcal{L}_{\eta}(m) &= \sum_{k=1}^{K}\left(\ln\left(\Gamma(K\eta)\right) - K\ln\Gamma(\eta) + \sum_{v=1}^{V}(\eta-1)(\Psi(\omega_k^v) - \Psi(\sum_{j=1}^{V}\omega_k^j))\right) \\ &= K\Gamma(K\eta) - K^2\ln\Gamma(\eta) + (\eta-1)\sum_{k=1}^{K}\left(\left(\sum_{v=1}^{V}\Psi(\omega_k^v)\right) - V\Psi(\sum_{j=1}^{V}\omega_k^j)\right) \\ &= K\Gamma(K\eta) - K^2\ln\Gamma(\eta) + C(\eta-1)\end{aligned} \tag{51}$$

Taking the derivative of η over the entier corpus,

$$\frac{d\sum_{m=1}^{M}\mathcal{L}_{\eta}(m)}{d\eta} = M(K^2\Psi(K\eta) - K^2\Psi(\eta) + C) = 0 \tag{52}$$

We solve the following but it is not analytically closed.

$$\Psi(K\eta) - \Psi(\eta) + \frac{C}{K^2} = 0 \tag{53}$$

This can also be numerically solved by Newton-Raphson as follows.

$$\eta_{i+1} = \eta_i - \frac{\Psi(K\eta_i) - \Psi(\eta_i) + \frac{C}{K^2}}{K\Psi'(K\eta_i) - \Psi'(\eta_i)} \tag{54}$$

The evaluation of Digamma function $\Psi(x)$ and Trigamma function $\Psi'(x)$ are explained in section 7.

6 Solving Newton-Raphson Methods for $\boldsymbol{\alpha}$ and η

This section solves the equation 50 with Newton-Raphson Methods. Let $\mathbf{g}(\boldsymbol{\alpha})$ denote the LHS of the equation 50. Note that,

$$\frac{\partial^2 \left(\sum_{m=1}^{M} \mathcal{L}_{\boldsymbol{\alpha}}(m)\right)}{\partial \alpha_k^2} = M \left(\Psi'(\sum_{j=1}^{K} \alpha_j) - \Psi'(\alpha_k) \right) \tag{55}$$

$$\frac{\partial^2 \left(\sum_{m=1}^{M} \mathcal{L}_{\boldsymbol{\alpha}}(m)\right)}{\partial \alpha_k \partial \alpha_i} = M \Psi'(\sum_{j=1}^{K} \alpha_j) \quad (k \neq i) \tag{56}$$

then the Hessian matrix $H(\boldsymbol{\alpha})$ has the special form such that $H(\boldsymbol{\alpha}) = D(\boldsymbol{\alpha}) + m\mathbf{e}\mathbf{e}^T$ where $\mathbf{e} = (1, 1, \cdots, 1)^T$, $m = M\Psi'(\sum_{j=1}^{K} \alpha_j)$, and $D(\boldsymbol{\alpha})$ is a diagonal matrix whose elements are $-M\Psi'(\alpha_k)$. Then we can apply the following Sherman-Morrison formula.

$$\left(A + \mathbf{u}\mathbf{v}^T\right)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^T A^{-1}}{1 + \mathbf{v}^T A^{-1}\mathbf{u}} \tag{57}$$

Identifying $A = D(\boldsymbol{\alpha})$, $\mathbf{u} = m^{\frac{1}{2}}\mathbf{1}$, and $\mathbf{v} = m^{\frac{1}{2}}\mathbf{1}$, we obtain

$$H^{-1}(\boldsymbol{\alpha}) = D(\boldsymbol{\alpha})^{-1} - \frac{D(\boldsymbol{\alpha})^{-1}\mathbf{e}\mathbf{e}^T D(\boldsymbol{\alpha})^{-1}}{m^{-1} - KM\sum_{k=1}^{K}(\Psi'(\alpha_k))} \tag{58}$$

then

$$\begin{aligned} \left[H^{-1}(\boldsymbol{\alpha})\mathbf{g}(\boldsymbol{\alpha})\right]_k &= -\frac{[\mathbf{g}(\boldsymbol{\alpha})]_k}{M\Psi'(\alpha_k)} - \frac{\frac{1}{M\Psi'(\alpha_k)}}{m^{-1} - KM\sum_{j=1}^{K}(\Psi'(\alpha_j))} \sum_{j=1}^{K} \frac{[\mathbf{g}(\boldsymbol{\alpha})]_j}{-M\Psi'(\alpha_j)} \\ &= -\frac{1}{M\Psi'(\alpha_k)} \left([\mathbf{g}(\boldsymbol{\alpha})]_k - \frac{\sum_{j=1}^{K} \frac{[\mathbf{g}(\boldsymbol{\alpha})]_j}{-M\Psi'(\alpha_j)}}{m^{-1} - KM\sum_{j=1}^{K}(\Psi'(\alpha_j))} \right) \end{aligned} \tag{59}$$

and the update equation is:

$$\boldsymbol{\alpha}_{i+1} = \boldsymbol{\alpha}_i - H^{-1}(\boldsymbol{\alpha_i})\mathbf{g}(\boldsymbol{\alpha_i}) \tag{60}$$

7 Evaluating Digamma and Trigamma Functions

.

7.1 Evaluating Digamma Function

- Find k such that $x + k \geq 6$. Let $x' = x + k$.

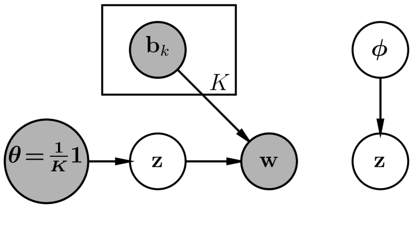

Figure 4: Inferring $\mathbf{z}$ with non-informative prior $\boldsymbol{\theta}$. The left side is the modeling and the right side is the approximation of $\mathbf{z}$

- Approximate $\Psi(x') \approx \ln(x') - \frac{1}{2x'} - \frac{1}{12x'^2} + \frac{1}{120x'^4} - \frac{1}{252x'^6} + \frac{1}{240x'^8} - \frac{5}{660x'^{10}} + \frac{691}{32760x'^{12}} - \frac{1}{12x'^{14}}$
- $\Psi(x) = \Psi(x') + \sum_{i=1}^{k} \frac{1}{x+k-1}$

7.2 Evaluating Trigamma Function

Basically use the derivative of the approximation of the digamma function.

$$\Psi'(x) \approx \frac{1}{x} + \frac{1}{2x^2} + \frac{1}{6x^3} - \frac{1}{30x^5} + \frac{1}{42x^7} - \frac{1}{30x^9} + \frac{5}{66x^{11}} - \frac{691}{2730x^{13}} + \frac{7}{6x^{15}} \tag{61}$$

8 Discussion : Columns of β as Word Embeddings

Here we discuss the theoretical interpretation of using the columns of β as word embeddings. It is formed in the inference of a new document $W_{m'}$, which consists of only one word $\boldsymbol{w'}$. And we assume $\boldsymbol{\alpha} = \mathbf{1}$, i.e., non-informative uniform prior, and the topic vector is also fixed as the non-informative uniform distribution such that $\boldsymbol{\theta_{m'}} = \frac{1}{K}\mathbf{1}$. This means the inference process has no prior information about which topics the document belongs to, and it makes no attempt to infere the information about the topic. Also, β is fixed at $\mathbb{E}[\beta_{[k,v]}] = \omega_k^v$.

Then the inference for $W_{m'}$ becomes merely a problem of finding $q(\boldsymbol{z'})$, and then $\mathbb{E}[\boldsymbol{z'}] = \phi'$. And after the inference, $\phi'_k = \frac{\beta_{[k,v]}}{\sum_{j=1}^{V} \beta_{[k,j]}}$.

So the normalized column $\beta_{[*,v]}$of a fixed β that corresponds to the given word $\boldsymbol{w'}$ such that $w'^v = 1$, is the expectation of the latent variable $\boldsymbol{z'}$ when no prior information about the topics is given.

In reality, the number of topics is less than 100, while a typical word embedding is of dimention of 300. It could be a competitive word embedding if we se the topic dimension to 300.

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 601–608. MIT Press, 2002.

[2] Christopher E. Moody. Mixing dirichlet topic models and word embeddings to make lda2vec. *CoRR*, abs/1605.02019, 2016.