LDA Notes

Jeff Lee

2014/3/5

1 Content

This notes is about the LDA (Latent Dirichlet Allocation ,Blei,2003). Here is mainly about the mathematics technology applied in the LDA and LDA practise in R language.

- Introduction
- LDA model
- Variational Distribution
- Jensen's Inequality
- Variational Inference
- Parameter Estimation
- LDA In R

2 Introduction

We are facing a world full of informations. And we need algorithmic tools to organize,analysis and understand these informations automatically. LDA, based on the topic model, is a simple but smart algorithm to provide us the topics of documents in the corpus. In the same words:it allows us to find the themes quickly in the documents.

3 LDA Model

LDA is a popular generative model based on topic model in information processing fields. In the LDA context,we have a corpus, a datasets of documents as D. When we consider a document,we regard it as a bag of words(words sequence in the document is ignored). So the LDA is applied to describe how a document contains words. The basic idea is that documents are represented as random mixtures over latent topics,where each topic is characterized by a distribution over words[1] . To obtain our LDA model, we need define the following terms at the beginning:

- The basic element in the LDA model is a word,discrete and defined be an item from a vocabulary indexed by $\{1,....V\}$. So we can regard a word indexed by v in the vocabulary as a unit-basis vector which has the length of V and the vth component equals one and other components in the vector are zeros.
- As we mentioned before, a document is a bag of words, so we can define a document with N words by $\boldsymbol{w} = \{w_1, w_2, ..., w_N\}$, where w_n presents the n th word in the document.
- A corpus or a dataset of M documents is defined by $D = \{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_M\}$.

With the terms denoted before, LDA can be assumed as the following process for each document $\mathbf{w}$ in a corpus D:

1. Choose $N \sim Poisson(\xi)$
2. Choose $\theta \sim Dir(\alpha)$
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim Multinomial(\theta)$.
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n

[1]Blei 2003

Some assumptions must be listed here:

- The dimension of the Dirichlet distribution(topic variable) is known and fixed.
- The word probabilities are parameterized by a $k \times V$ matrix β for each topic (row) and each term (column) where $\beta_{ij} = p(w^j = 1|z^i = 1)$.

A k-dimensional Dirichlet distribution has the following probability density:

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

The Dirichlet distribution is in the exponential family, has finite dimensional sufficient statistics, and is conjugate to the multinomial distribution. All the properties provide us a convenient development of the inference and parameter estimation algorithms for LDA.

Described in the LDA Processing, given the parameter α and β, the joint distribution of a topic mixture θ, a set of topics z, and a set of N words $\mathbf{w}$, we have :

$$p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{2}$$

With the Eq.(2), by integrating over θ and summing over z, we can get the marginal distribution of a document:

$$p(\boldsymbol{w}|\alpha, \beta) = \int p(\theta|\alpha) (\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \theta)) d\theta \tag{3}$$

Followed by producting the marginal distribution of a single document, we can obtain the probability of a corpus:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d, \alpha) (\prod_{n=1}^{N} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta)) d\theta_d \tag{4}$$

The central inferential problem for LDA is determining the posterior distribution of the latent variables given the document[2]:

$$p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta) = \frac{p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)}{p(\boldsymbol{w}|\alpha, \beta)} \tag{5}$$

This distribution is the crux of LDA, we can decompose it into a hierarchy with the graphical model technology:

$$p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) = p(\boldsymbol{w}|\boldsymbol{z}, \beta)p(\boldsymbol{z}|\theta)p(\theta|\alpha) \tag{6}$$

Here the $p(\boldsymbol{w}|\boldsymbol{z}, \beta)$ represents the prbability of observing a document with N words given a topic vector of length N that assigns a topic each word from the $k \times V$ probability β.So we can multiply them together to obtain the observing document:

$$p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) = \prod_{n=1}^{N} \beta_{z_n, w_n} \tag{7}$$

The $p(\boldsymbol{z}|\theta)$ is simple θ_i for the unique i such that $z_n^i = 1$,With all these and Eq.(1),we can break the Eq(6) into :

$$p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) = (\frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} ... \theta_k^{\alpha_k - 1}) \prod_{n=1}^{N} \beta_{z_n, w_n} \theta_{z_n} \tag{8}$$

Where θ_{z_n} represents the component of θ chosen for z_n.If we use the entire vocabulary of size V to replaced the nntation mentioned above:

$$p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) = (\frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} ... \theta_k^{\alpha_k - 1}) \prod_{n=1}^{N} \prod_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{i,j})^{w_n^j z_n^i} \tag{9}$$

As we marginalize over θ and $\boldsymbol{z}$, we get the denominator in Eq.(5):

$$p(\boldsymbol{w}|\alpha, \beta) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \int (\prod_{i=1}^{k} \theta_i^{\alpha_i - 1}) (\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{i,j})^{w_n^j}) d\theta \tag{10}$$

[2]Latent Dirichlet Allocation:Towards a Deeper Understanding,Colorado Reed

As we marginalize over θ and $\boldsymbol{z}$ From Eq.(10), we can not compute the distribution directly with the problem that $\theta_i, \beta_{i,j}$ twist together. As described in Blei et al.(2003), By dropping some of the edges and nodes in the original graphical model, we can obtain a simplified graphical model in thre form:

$$q(\theta, \boldsymbol{z}|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n) \tag{11}$$

Where the Dirichlet parameter γ and the multinomial parameters$(\phi_1, ...\phi_N)$ are the freee variational parameters.

4 Variational Distribution

Having specified a simplified family of probability distribution, the next step is to set up an optimization problem that determines the value of variational parameters γ, ϕ. Here we use the KL(Kullback Lerbler) divergence technology, which is a measure in statistics (Cover and Thomas, 1991) that quantifies in bits how close a probability distribution $p = \{p_i\}$ is to a model (or candidate) distribution $q = \{q_i\}$, as defined like:

$$D_{KL}(P||Q) = \int_{-\infty}^{+\infty} P(x) \log \frac{P(x)}{Q(x)} dx$$

Or

$$D_{KL}(P||Q) = \sum_{i=1}^{N} p_i \log \frac{p_i}{q_i}$$

D_{KL} is non-negative($\geqslant 0$), not symmetric in p and q, zero if the distributions match exactly and can potentially equal infinity. Here P is the real distribution and Q is a distribution we obtain by our model or experiments,If we use the real distribution P to obtain the bytes $E(X) = \sum_i (p_i(x) \times \log(\frac{1}{p_i(x)}))$, or we can also use the model distribution Q to obtain the bytes $E(X) = \sum_i (q_i(x) \times \log(\frac{1}{p_i(x)}))$. It is known that the the no model distributions can be more exactly than the real distribution. So the definition of D_KL can be

writen like this:

$$
\begin{aligned}
&D_KL(Q||P) \\
&= \sum_{x\in X} Q(x)\log(1/P(x)) - \sum_{x\in X} Q(x)\log(1/Q(x)) \\
&= \sum_{x\in X} Q(x)\log(\frac{Q(x)}{P(x)}) \\
&= E_Q[(\log\frac{Q(x)}{P(x)})] \\
&= E_Q[\log Q(x)] - E_Q[\log P(x)]
\end{aligned}
$$

The problem of LDA can now be transformed to the form:

$$(\gamma^*, \phi^*) = arg\min_{(\gamma,\phi)} D_{KL}(q(\theta, \boldsymbol{z}|\gamma, \phi)||p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta)) \tag{12}$$

Let q represent $q(\theta, \boldsymbol{z}|\gamma, \phi)$, So we can break The D_{KL} part down step by step :

$$
\begin{aligned}
&D_{KL}(q||p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta)) \\
&= E_q[\log q] - E_q[\log p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta)] \\
&= E_q[\log q] - E_q[\log\frac{p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)}{p(\boldsymbol{w}|\alpha, \beta)}] \\
&= E_q[\log q] - E_q[\log p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)] + \log p(\boldsymbol{w}|\alpha, \beta)
\end{aligned} \tag{13}
$$

In the Eq.(13),we use the Eq.(5) to obtain the right hand expression.

5 Jensen's Inequality

5.1 Convex Functions

Definition 1 *Lef f be a real valued function defined on an interval $I = [a, b]$. f is said to be convex onI if $\forall x_1, x_2 \in I, \lambda \in [0, 1]$,*

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$$

Throrem 1 (Jensen's inequality) *Lef f be a convex function defined on an interval I. If $x_1, x_2, ..., x_n \in I and \lambda_1, \lambda_2, ..., \lambda_n \geq 0 with \sum_{i=1}^{n} \lambda_i = 1$,*

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$$

5.2 Jensen's inequality in LDA

Using the Jensen's inequality ,we bound $p(\boldsymbol{w}|\alpha, \beta)$ by:

$$\begin{aligned}
&\log p(\boldsymbol{w}|\alpha, \beta) \\
&= \log \int \sum_{\boldsymbol{z}} p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) d\theta \\
&= \log \int \sum_{\boldsymbol{z}} \frac{p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) q(\theta, \boldsymbol{z})}{q(\theta, \boldsymbol{z})} d\theta \\
&\geq \int \sum_{\boldsymbol{z}} q(\theta, \boldsymbol{z}) \log \frac{p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)}{q(\theta, \boldsymbol{z})} d\theta \qquad (14) \\
&= -\int \sum_{\boldsymbol{z}} q(\theta, \boldsymbol{z}) \log \frac{q(\theta, \boldsymbol{z})}{p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)} d\theta \\
&= -E_q[\log \frac{q(\theta, \boldsymbol{z})}{p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)}] \\
&= E_q[\log p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)] - E_q[\log q(\theta, \boldsymbol{z})]
\end{aligned}$$

Here we regard $\lambda_z = \int q(\theta, \boldsymbol{z}) d\theta; x_{\boldsymbol{z}} = \frac{p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)}{q(\theta, \boldsymbol{z})}$ so we have $\sum_{\boldsymbol{z}} \lambda_z = 1$ and can apply the Jensen's inequality on the Eq.(14), and we obtain the form of KL divergence to make our process forward. If we denote the right side of Eq.(14) by $L(\gamma, \phi, \alpha, \beta)$,with the $q(\theta, \boldsymbol{z}) = q(\theta, \boldsymbol{z}|\gamma, \phi)$we can reach to the Eq.(15):

$$\log p(\boldsymbol{w}|\alpha, \beta) = L(\gamma, \phi, \alpha, \beta) + D_{KL}(q(\theta, \boldsymbol{z}|\gamma, \phi)||p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta)) \qquad (15)$$

So when we minimize the $D_{KL}(q(\theta, \boldsymbol{z}|\gamma, \phi)||p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta))$,we are expecting the $L(\gamma, \phi, \alpha, \beta)$ to get as close as $\log p(\boldsymbol{w}|\alpha, \beta)$.

6 Variational Inference

Using the factorizations of p and q,To expand the $L(\gamma, \phi, \alpha, \beta)$,we obtain:

$$\begin{aligned} & L(\gamma, \phi, \alpha, \beta) \\ & = E_q[\log p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)] - E_q[\log q(\theta, \boldsymbol{z})] \\ & = E_q[\log p(\theta|\alpha)] + E_q[\log p(\boldsymbol{z}|\theta)] + E_q[\log p(\boldsymbol{w}|\boldsymbol{z}, \beta)] \\ & - E_q[\log q(\theta)] - E_q[\log q(\boldsymbol{z})] \end{aligned} \tag{16}$$

Here when we mention $q(\theta, \boldsymbol{z})$,we mean $q(\theta, \boldsymbol{z}|\gamma, \phi)$,so do with the $q(\theta), q(\boldsymbol{z})$ short for $q(\theta|\gamma), q(\boldsymbol{z}|\phi)$ respectively.

The next step is to break the right side of Eq.(16) to five terms which are respectively expended by the entropy technology:

$$E_{\boldsymbol{\alpha}}[\log \theta_i] = E[\log \theta_i|\boldsymbol{\alpha}] = \Psi(\alpha_i) - \Psi(\sum_{j=1}^{k} \alpha_j) \tag{17}$$

In the next three sections we will prove the Eq.(17) with the properties of the exponential family of distribution.

6.1 The Exponential Family of Distributions

$$p(x|\theta) = h(x)e^{\theta^T T(x) - A(\theta)} \tag{18}$$

To get a normalized distribution ,for any θ

$$\int p(x)dx = e^{-A(\theta)} \int h(x)e^{\theta^T T(x)}dx = 1$$

So we obtain

$$A(x) = \log \int h(x)e^{\theta^T T(x)}dx$$

If we denote the $\int h(x)e^{\theta^T T(x)}dx$ as $Q(\theta)$,then

$$\begin{aligned}
&\frac{dA(\theta)}{d\theta} \\
&= \frac{1}{Q(\theta)} \times \frac{dQ(\theta)}{d\theta} \\
&= \frac{\int h(x)e^{\theta^T T(x)}T(x)dx}{\int h(x)e^{\theta^T T(x)}dx} \\
&= E_{p_\theta}[T(x)]
\end{aligned} \tag{19}$$

6.2 Gamma Function

In mathematics, the gamma function (represented by the capital Greek letter Γ) is an extension of the factorial function, with its argument shifted down by 1, to real and complex numbers. That is, if n is a positive integer: $\Gamma(n) = (n-1)!$ The gamma function is defined for all complex numbers except the negative integers and zero. For complex numbers with a positive real part, it is defined via a convergent improper integral:

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,\mathrm{d}t$$

The digamma function is defined as the logarithmic derivative of the gamma function:

$$\Psi(x) = \frac{d}{dx}log\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)} \tag{20}$$

6.3 Variational Inference

So let's back to the Eq.(17),and Eq(1). The Eq.(1) can be rewritten in the form of exponential family of distribution:

$$p(\theta|\alpha) = e^{(\sum_{i=1}^k(\alpha_i-1)\log\theta_i)+\log\Gamma(\sum_{j=1}^k\alpha_j)-\sum_{i=1}^k\log\Gamma(\alpha_i)}$$

To break it to k components,with the form of Eq.(18), we can obtain $T(x_i) = \log\theta_i$, $\theta_i = \alpha_i - 1$, $A(\theta_i) = \log\Gamma(\alpha_i) - \log\Gamma(\sum_{j=1}^k\alpha_j)$

With the Eq.(20), If we put the $T(x_i), \theta_i$ and $A(\theta_i)$ into the Eq.(19), we

can easily obtain the Eq.(17).

6.3.1 $E_q[\log p(\theta|\alpha)]$

Let's first compute the $E_q[\log p(\theta|\alpha)]$ in which the q is the probability of θ under the condition of γ_i:

$$
\begin{aligned}
&E_q[\log p(\theta|\alpha)] \\
&= E_q[\log \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}...\theta_k^{\alpha_k-1}] \\
&= E_q[\sum_{i=1}^{k}(\alpha_i-1)\log\theta_i + \log\Gamma(\sum_{i=1}^{k}\alpha_i) - \sum_{i=1}^{k}\log\Gamma(\alpha_i)] \\
&= \sum_{i=1}^{k}E_q[\log\theta_i] \\
&+ \log\Gamma(\sum_{i=1}^{k}\alpha_i) - \sum_{i=1}^{k}\log\Gamma(\alpha_i) \\
&= \sum_{i=1}^{k}(\alpha_i-1)\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)\right) \\
&+ \log\Gamma(\sum_{i=j}^{k}\alpha_j) - \sum_{i=1}^{k}\log\Gamma(\alpha_i)
\end{aligned}
$$

6.3.2 $E_q[\log p(\mathbf{z}|\theta)]$

θ is a k-dimensional vector,z_n is a topic generated by the $Multinomial(\theta)$. Let's image there is a k-face die. Every component in the θ vector is the probability of the face with the same index. So $p(\boldsymbol{z}|\theta)$ is simply θ_i for the unique i such that $z_n^i = 1$. The q here is probability of z under the condition

of ϕ, so we can obtain:

$$
\begin{aligned}
&E_q[\log p(\mathbf{z}|\theta)] \\
&= E_q[\sum_{n=1}^{N}\sum_{i=1}^{k} z_n^i \log\theta_i] \\
&= \sum_{n=1}^{N}\sum_{i=1}^{k} E_q[z_n^i]E_q[\log\theta_i] \\
&= \sum_{n=1}^{N}\sum_{i=1}^{k} \phi_{ni}\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)\right)
\end{aligned}
$$

6.3.3 $E_q[\log p(\mathbf{w}|\mathbf{z},\beta)]$

With the z_n, and the vocabulary of V words, we say β_{ij} means we have the ith topic($z_n^i = 1$) and we jth word($w_n^j = 1$) in the vocabulary.But how can we get the w_n^j?,It is the same approach that we obtain the ith topic. we sample from a Dirichlet with V dimensions. And with the multinomial distribution we get the jth word in the vocabulary. The process above just generate only one word,To generate a document with N word,we have to do this N times.The $E_q[\log p(\mathbf{w}|\mathbf{z},\beta)]$ shows the process:

$$
\begin{aligned}
&E_q[\log p(\boldsymbol{w}|\boldsymbol{z},\beta)] \\
&= E_q[\sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{V} z_n^i w_n^j \log\beta_{ij}] \\
&= \sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{V} E_q[z_n^i]E_q[w_n^j]E_q[\log\beta_{ij}] \\
&= \sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{V} \phi_{ni} w_n^j \log\beta_{ij}
\end{aligned}
$$

6.3.4 $E_q[\log q(\theta|\gamma)]$

This is about how θ generated by Dirichlet distribution with the γ parameter mentioned in Eq.(11). so we can expand it like we do with the $E_q[\log p(\theta|\alpha)]$

before:

$$
\begin{aligned}
&E_q[\log q(\theta|\gamma)] \\
&= \sum_{i=1}^{k}(\gamma_i - 1)E_q[\log\log\theta_i] + \log\Gamma(\sum_{i=1}^{k}\gamma_i) - \sum_{i=1}^{k}\log\Gamma(\gamma_i) \\
&= \sum_{i=1}^{k}(\gamma_i - 1)[\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)] \\
&+ \log\Gamma(\sum_{j=1}^{k}\gamma_j) - \sum_{i=1}^{k}\log\Gamma(\gamma_i)
\end{aligned}
$$

6.3.5 $E_q[\log q(\mathbf{z}|\phi)]$

In Eq.(11),we have said that the $(phi_1, ...\phi_N)$ are parameters of multinomial distribution that generate the $\boldsymbol{z}$. That is alike the generating process of $E_q[\log p(\boldsymbol{z}|\theta)]$,so it also can be expanded in the same form:

$$
\begin{aligned}
&E_q[\log q(\mathbf{z}|\phi)] \\
&= E_q[\sum_{n=1}^{N}\sum_{i=1}^{k} z_n^i \log\phi_{ni}] \\
&= \sum_{n=1}^{N}\sum_{i=1}^{k} E_q[z_n^i]E_q[\log\phi_{ni}] \\
&= \sum_{n=1}^{N}\sum_{i=1}^{k} \phi_{ni}\log\phi_{ni}
\end{aligned}
$$

6.4 Computing $L(\gamma, \phi, \alpha, \beta)$

$L(\gamma, \phi, \alpha, \beta)$ in Eq.(16) has been broken in five terms,and each term has been expanded.Finally we can compute the $L(\gamma, \phi, \alpha, \beta)$ by composing these terms

together:

$$
\begin{aligned}
& L(\gamma, \phi, \alpha, \beta) \\
& = \sum_{i=1}^{k}(\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)) + \log\Gamma(\sum_{i=1}^{k}\alpha_i) - \sum_{i=1}^{k}\log\Gamma(\alpha_i) \\
& + \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)) \\
& + \sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{V}\phi_{ni}w_n^j\log\beta_{ij} \\
& - \sum_{i=1}^{k}(\gamma_i - 1)[\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)] - \log\Gamma(\sum_{j=1}^{k}\gamma_j) + \sum_{i=1}^{k}\log\Gamma(\gamma_i) \\
& - \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}\log\phi_{ni}
\end{aligned}
\tag{21}
$$

Where each of the five lines expands one of the five terms in the 6.3 section.The next step we will use the Lagrangian with respect to the variational parameters ϕ and γ.

6.4.1 Variational Multinomial

We first maximize Eq.(21) with respect to ϕ_{ni} , the probability that the nth word is generated by latent topic i. Observe that this is a constrained maximization since $\sum_{i=1}^{k}\phi_{ni} = 1$.

We form the Lagrangian by isolating the terms which contain ϕ_{ni} and adding the appropriate Lagrange multipliers. Let β_{iv} be $p(w_n^v = 1|z_i = 1)$ for the appropriate v. (Recall that each w_n is a vector of size V with exactly one component equal to one; we can select the unique v such that $w_n^v = 1$, so we can get rid of the $\sum_{j=1}^{V}$)):

$$
L_{[\phi_{ni}]} = \phi_{ni}(\Psi(\gamma_i)) - \Psi(\sum_{j=1}^{k}\gamma_j) + \phi_{ni}\log\beta_{iv} - \phi_{ni}\log\phi_{ni} + \lambda_n(\sum_{i=1}^{k}\phi_{ni} - 1)
$$

Where the $L_{\phi_{ni}}$ denotes that we only care the terms in $L(\gamma, \phi, \alpha, \beta)$ that are

a function of ϕ_{ni}.Taking derivatives with respect to ϕ_{ni},we obtain:

$$\frac{\partial L_{\phi_{ni}}}{\partial \phi_{ni}} = \Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) + \log \beta_{iv} - \log \phi_{ni} - 1 - \lambda_n$$

Setting the derivative to zero yield the maximizing value of the variational parameter ϕ_{ni}:

$$\phi_{ni} \propto \beta_{iv} exp(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)) \tag{22}$$

6.4.2 Variational Dirichlet

Next,we maximize Eq.(21) with respect to γ_i, the ith component of the posterior Dirichlet parameter. The terms containing γ_i are:

$$\begin{aligned}
&L_{[\gamma_i]} \\
&= (\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)) \\
&+ \sum_{n=1}^{N} \phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)) \\
&- (\gamma_i - 1)[\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)] - \log \Gamma(\sum_{j=1}^{k} \gamma_j) + \sum_{i=1}^{k} \log \Gamma(\gamma_i) \\
&= (\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j))(\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i) - \log \Gamma(\sum_{j=1}^{k} \gamma_j) + \log \Gamma(\gamma_i)
\end{aligned}$$

We take the derivative with respect to γ_i:

$$\begin{aligned}
&\frac{\partial L_{\gamma_i}}{\partial \gamma_i} \\
&= \Psi'(\gamma_i)(\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i) - \Psi'(\sum_{j=1}^{k} \gamma_j)(\alpha_j + \sum_{n=1}^{N} \phi_{nj} - \gamma_j)
\end{aligned}$$

Setting the equation to zero yields a maximum at:

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni} \tag{23}$$

Since Eq.(22) depends on the variational multinomial ϕ, full variational inference requires alternating between Eqs.(22) and (23) until the bound converges.

7 Parameter Estimation

We have thus far considered the log likelihood for a single document. Given our assumption of exchangeability for the documents,the overall log likelihood of a corpus $D = \{\boldsymbol{w}_1, \boldsymbol{w}_2, ...\boldsymbol{w}_M\}$ is the sum of the individual documents;moreover the variational lower bound is the sum of the individual variational bounds. To maximize with respect to β, we isolate terms and add Lagrange multipliers:

$$L_\beta = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \sum_{j=1}^{V} \phi_{dni} w_{dn}^j \log \beta_{ij} + \sum_{i=1}^{k} \lambda_i (\sum_{j=1}^{V} \beta_{ij} - 1)$$

Let's rewrite the equation with respect to β_{ij}:

$$L_{\beta_{ij}} = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^j \log \beta_{ij} + \lambda_i \beta_{ij} - \sum_{i=1}^{k} \lambda_i$$

Taking the derivative with β_{ij}, we obtain:

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^j$$