Trinity College Dublin

Directed Studies

Topic Models - Latent Dirichlet Allocation

This report, produced as a part of the requirement for the Directed studies, provides the theory on the properties of Dirichlet, Dirichlet priors and Latent Dirichlet Allocation for Topic modelling. LDA by Variational Inference and Gibbs sampling are covered and their update equations are completely derived.

Author:
Arun Jayapal

Supervisor:
Dr. Martin Emms

November 27, 2014

Contents

1 Introduction

Topic models are the probabilistic models to discover the topics from a collection of documents. These probabilistic models help in analyzing text from a large collection of documents and automatically identify the topics from this collection. Further these probabilistic models do not have any prior information on what topics are in the collection such as medical science, computer, statistics, government, tax, etc., There are different topic models in the state of the art such as probabilistic latent semantic indexing (PLSI), Non-negative matrix factorization (NNMF) and Latent Dirichlet Allocation (LDA) while LDA is the most commonly used topic model to discover topics from text. Therefore, LDA will be studied for this directed studies module. There are two variants of the LDA (1) by variational approximation and (2) by collapsed gibbs sampling, which are discussed in detail in this report. The LDA by variational approximation was originally put forward by Blei *et al.* [2003], while the variant of LDA by collapsed gibbs sampling was put forward by Griffiths and Steyvers [2004]. Later on there are many extensions to LDA based on application, introduced by different researchers such as hierarchical mixture model and continuous time dynamic topic model.

This report is organized as follows: As an introduction to the discussion on LDA, dirichlet distribution and its properties are discussed in section 2. In the next section 3 we will discuss about the LDA in general, and then provide the LDA inference in section 3.1 and further discuss LDA by variational approximation in section 3.2 and collapsed gibbs sampling in section 3.4 with their algorithms provided. The details of the derivations for the update equations used for the algorithm are provided in the appendix 5. Then we conclude the discussion on LDA by providing the differences between LDA by variational approximation and by collapsed gibbs sampling in section 4.

2 Dirichlet distribution

Dirichlet distribution can be considered as the multi-variate generalization of the 'Beta' distribution. The density function of 'Beta' distribution is defined by (1), where x is the random variable and α, β are the parameters.

$$f(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{1}$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is a Beta function.

We can see 'Dirichlet' distribution as a distribution over distributions. Let us consider a vector of random probability distributions $\boldsymbol{X} = [x_1, x_2, \dots, x_k]$ where $\sum_{i=1}^{k} x_k = 1$ and is parameterized by $\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_k]$. So, the density function of dirichlet is given by

$$f(x_1, x_2...x_k; \alpha_1, \alpha_2...\alpha_k) = \frac{1}{\beta(\boldsymbol{\alpha})} \prod_{i=1}^{k} x_i^{\alpha_i - 1} \tag{2}$$

where,

(1) $\beta(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)}$

(2) $\sum_{i=1}^{k} x_i = 1$ and $0 < x_i < 1$,

(3) $\Gamma(n)$ denotes the gamma function, which is a generalization of the factorial function ie., $\Gamma(n) = (n-1)!$ The beauty of this generalization is the gamma function works with n being a real number.

(4) $\boldsymbol{\alpha}$ is a vector of parameters and each value in the vector is greater than 0. Further in this section, we will now proceed to understand the properties of dirichlet distribution.

2.1 Properties of Dirichlet

The properties (mean and mode) of Dirichlet distribution are discussed in this section.

2.1.1 Mean

Before, we write down the expectation or mean of the Dirichlet distribution, we derive the expectation for Beta distribution as dirichlet distribution is considered the multi-variate generalization of beta distribution. This proof closely follows the proof from `http://www.statlect.com/beta_distribution.htm`

The expectation of a vector of random variables X is given by

$$\mathbb{E}[X] \quad = \quad \int_0^1 \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} x.dx$$

we define $B(\alpha, \beta)$

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$\mathbb{E}\left[X\right] = \frac{1}{B(\alpha,\beta)} \int_0^1 x^{\alpha}(1-x)^{\beta-1}dx$$

$$= \frac{1}{B(\alpha,\beta)} \int_0^1 x^{(\alpha-1)+1}(1-x)^{\beta-1}dx$$

$$= \frac{1}{B(\alpha,\beta)} B(\alpha+1,\beta) \quad \text{(by integral representation of beta function)}^1$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \quad \text{(from definition of Beta function)}$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+1)} \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \quad \text{(re-arranging from previous step)}$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta)(\alpha+\beta)} \frac{\Gamma(\alpha)(\alpha)}{\Gamma(\alpha)} \quad \text{(gamma's property } \Gamma(z) = \Gamma(z-1)(z-1)\text{)}$$

$$\mathbb{E}\left[X\right] = \frac{\alpha}{\alpha+\beta}$$

Expectation for Dirichlet: Consider X as a vector of random variables $x_1, x_2, \ldots, x_n$. Here, we show the expectation or mean of the distribution as

$$\mathop{\mathbb{E}}_{k}\left[x\right] = \frac{\boldsymbol{\alpha_k}}{\sum_k \alpha_k} \tag{3}$$

This derivation is similar to the one for estimating the Expectation of the beta distribution.

$$\mathbb{E}\left[x_1\right] = \int_x \frac{1}{\beta(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1} x_1 .dx$$

$$= \frac{1}{\beta(\boldsymbol{\alpha})} \int_x \prod_{k=2}^{K} x_k^{\alpha_k - 1} x_1^{(\alpha_1+1)-1} .dx$$

[1] The integral representation of Beta function is provided at `http://www.statlect.com/subon2/betfun1.htm`

consider $\boldsymbol{\alpha'_{k\neq1}} = \boldsymbol{\alpha_k}$ and $\boldsymbol{\alpha'_1} = \boldsymbol{\alpha} + 1$

$$\mathbb{E}\left[x_1\right] = \frac{1}{\beta(\boldsymbol{\alpha})} \int_x \frac{\beta(\boldsymbol{\alpha'})}{\beta(\boldsymbol{\alpha'})} \prod_{k=1}^{K} x_k^{\alpha'_k - 1}.dx$$

$$= \frac{\beta(\boldsymbol{\alpha'})}{\beta(\boldsymbol{\alpha})}(1)$$

$$= \frac{\prod_{k=2}^{K} \Gamma(\alpha_k)\Gamma(\alpha_1 + 1)}{\gamma(\sum_k \alpha_k + 1)} \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)}$$

using gamma's property $\Gamma(n+1) = n\Gamma(n)$, we get the following step

$$\mathbb{E}\left[x_1\right] = \frac{\alpha_1}{\sum_k \alpha_k}$$

Generalizing this, we get

$$\mathbb{E}\left[x_j\right] = \frac{\alpha_j}{\sum_k \alpha_k}$$

2.1.2 Mode

Mode is the value of the distribution where the density is the highest. In other words, mode is the Maximum A Posteriori (MAP) of the distribution and its formula for Dirichlet is:

$$x_k = \frac{\alpha_k - 1}{\sum_k (\alpha_k - 1)}, \text{ when } \ (\alpha > 1) \tag{4}$$

Here, we first derive the mode for beta distribution and extend it to the dirichlet distribution as beta distribution is the generalization of dirichlet distribution. To derive the mode of the beta distribution, we have get the partial derivative of the density function see equation 1 with respect to x, set it to zero and solve for x.

$$\frac{\partial}{\partial x} f(x; \alpha, \beta) = \frac{\partial}{\partial x} \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} = 0$$

Considering the beta function $\frac{1}{B(\alpha,\beta)}$ as a constant C,

$$\frac{\partial}{\partial x} f(x; \alpha, \beta) = C\left((\alpha - 1)x^{\alpha-2}(1-x)^{\beta-1} + x^{\alpha-1}(-1)(\beta-1)(1-x)^{\beta-2}\right) = 0$$

$$= (\alpha-1)x^{\alpha-1}x^{-1}(1-x)^{\beta-1} - x^{\alpha-1}(\beta-1)(1-x)^{\beta-1}(1-x)^{-1} = 0$$

$$= x^{\alpha-1}(1-x)^{\beta-1}\left[\frac{\alpha-1}{x} - \frac{\beta-1}{(1-x)}\right] = 0$$

$$= x^{\alpha-1}(1-x)^{\beta-1}\left[\frac{(1-x)(\alpha-1)-(\beta-1)x}{x(1-x)}\right] = 0$$

$$= (1-x)(\alpha-1)-(\beta-1)x = 0$$

Now, solving for x, we get

$$x = \frac{\alpha-1}{\alpha+\beta-2} \qquad \text{[Mode of Beta distribution]}$$

We know that α and β are the two parameters of the beta distribution, but dirichlet distribution has k number of α parameters. Therefore, we can extend the mode function obtained from the beta distribution as given below

$$\boldsymbol{x}_k = \frac{\boldsymbol{\alpha}_k - 1}{\sum_k(\alpha_k - 1)}$$

For dirichlet distribution with $(\alpha > 1)$, the mode will be the maximum value, but for the distribution with $(\alpha < 1)$, the mode will be the minimum value (as they would be the stationary points in the distribution).

2.2 Dirichlet as Conjugate prior

By now we know that Dirichlet distribution is a generalization of the beta distribution. As we use 'beta' distribution as prior for binomial distribution, it is a good idea to use 'Dirichlet' distribution as prior for multi-nomial distribution. Following is the proof to prove multi-nomial and dirichlet distributions form conjugate prior[2]. The multinomial distribution is defined by

$$\text{multi}[x;\theta] = \frac{(\sum_{k=1}^{K} x_k)!}{\prod_{k=1}^{k}(x_k!)} \prod_{k=1}^{K} \theta_k^{x_k} \tag{5}$$

where the parameters θ are the probability values to get into one of the K-categories.

[2] A conjugate prior of a likelihood function is the prior when both posterior and prior distributions are of the same distribution

The posterior probability function is defined by

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{6}$$

But, as $p(x)$ will act as a normmalizing constant, we can exclude this and rewrite the posterior probability distribution as

$$\begin{aligned}
p(\theta|x) &= p(x|\theta)p(\theta) \\
&= \text{multi}[x;\theta] \ \text{Dir}(\theta|x) \\
&\approx \prod_{k=1}^{K} \theta_k^{x_k} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \\
&\approx \prod_{k=1}^{K} \theta_k^{(x_k + \alpha_k - 1)} \\
&= \text{Dir}(x + \alpha)
\end{aligned}$$

This section followed the report from Huang [2005].

2.3 Dirichlet plots - a demonstration

A short demonstration of the dirichlet distribution with their plots are discussed here. It is difficult to visualize plots that are greater than 2 or 3 dimensional. Therefore, two dimensional and three dimensional plots are provided to depict the behaviour of dirichlet distribution.

Figure 1 provides 15 random draws at each of four different α settings and the α's are considered symmentric for these plots. The graph has the topics plotted on x-axis and the probability of the topics on the y-axis. Here, we can observe the following (1) as the alpha goes below 0, the graph gets uneven and sparse (2) for each draw, the same distribution is maintained but the proportion of topics are not always same for each draw. We will utilize this property while modelling LDA.

Figure 2 provides three-dimensional plots of dirichlet distribution with α's at different settings. These plots can be reproduced by changing the values of the variables $a1, a2$ and $a3$ (corresponding to alphas) provided in the code at Appendix 5.2. The plots provided in these figures will help in understanding LDA better, is discussed in section 3

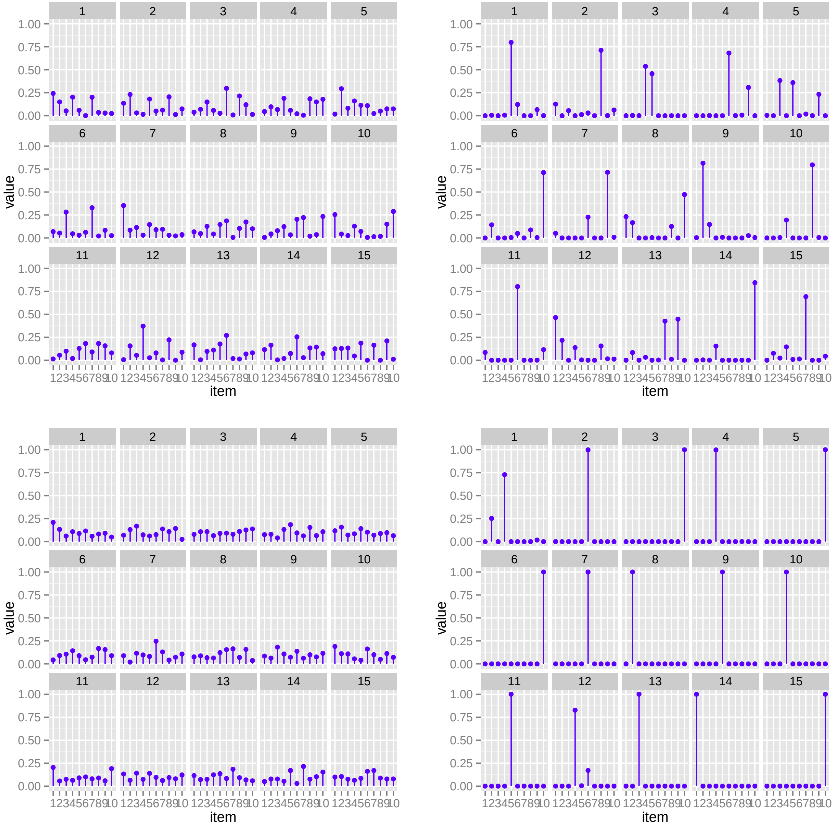
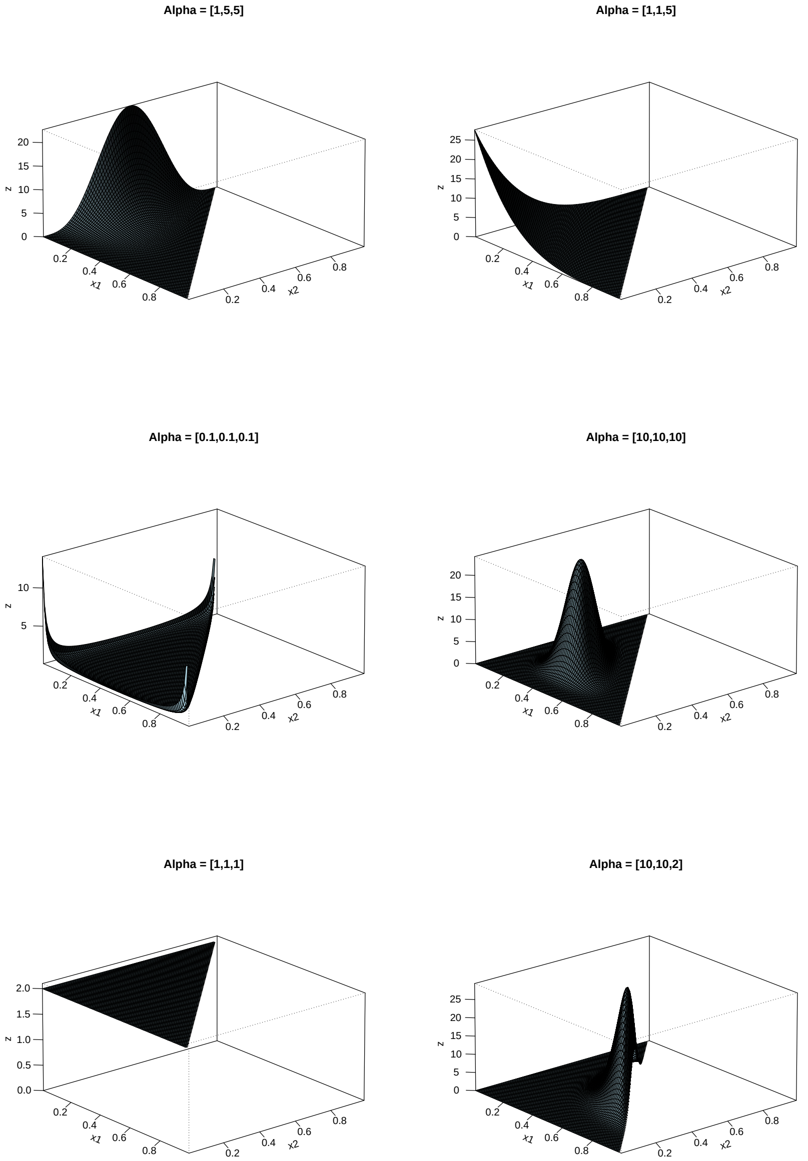


Figure 1: 2D plots of Dirichlet distribution, (from top left) in the clockwise direction, α at different settings: 1. $\alpha = 1$, 2. $\alpha = 0.1$, 3. $\alpha = 0.01$, 4. $\alpha = 5$. The figure provides 15 random draws at each of four different α settings. The graph has the number of topics plotted on x-axis and the probability of the topic on the y-axis. This graph can be reproduced by executing the R script provided in 5.1



Alpha = [1,5,5]

Alpha = [1,1,5]

Alpha = [0.1,0.1,0.1]

Alpha = [10,10,10]

Alpha = [1,1,1]

Alpha = [10,10,2]

Figure 2: 3D plots of Dirichlet distribution. Each of these plots are produced at dimension $k = 3$ and the respective alpha settings are titled for each plot. It can observed that when $\alpha = 1$ (symmentric), the distribution is uniform, while $\alpha = 10$ (symmentric), the distribution has a mode at the maximum and the hump is dense. But when we have α's set asymmentric, the distributions are skewed. It should also be noted that when $\alpha < 1$ (symmentric), the mode is at the minimum.

3 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation is a generative model to discover topics from the text. LDA considers each document from a document collection to be a mixture of multiple topics, where a topic can be defined to be a distribution over a fixed vocabulary terms. Each document from the document collection is considered to be a different proportion of the topics. Say for example we analyze a science document collection, which has documents from a variety of fields. One document might exhibit topics from medicine and technology, another document may have topics of medicine and bio science and a different document with topics from bio science and technology. The major challenge is that these topics are not known in advance, but should be learned automatically from the documents. The current section will follow Blei *et al.* [2003].

The LDA model is provided in the plate diagram in figure 3. The notations used in the model are introduced here:
$\boldsymbol{\alpha}$: a vector of symmentric dirichlet priors of size k, where k is the number of topics which is fixed
$\boldsymbol{\beta}$: the conditional probability table of the words to topics
$\boldsymbol{\theta^d}$: the multinomial variable (a vector of topic probabilities) is selected once for each document
$\boldsymbol{w^d}$: vector of words in the document, which is the only observed variable in the model
$\boldsymbol{z^d}$: the topic choice for each word (vector) in the document

The plate diagram can be interpreted in conjunction with the generative process assumed by LDA for each document $\boldsymbol{w}$ in document collection D.

(1) choose $\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha})$
(2) For each of the N words w_n:
(a) Choose a topic $z_n \sim \text{Multinomial}(\boldsymbol{\theta})$
(b) Choose a word w_n from $p(w_n|z_n, \boldsymbol{\beta})$, a multinomial probability conditioned on the topic z_n

Given the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the joint probability of $\boldsymbol{\theta^d}$, $\boldsymbol{z^d}$ and $\boldsymbol{w^d}$, is given by:

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

The probability of words in a document is obtained by marginalizing out z and θ, so

$$p(\boldsymbol{w^d}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int_{\boldsymbol{\theta^d}} p(\boldsymbol{\theta^d}; \boldsymbol{\alpha}) \prod_{n=1}^{N} \sum_{z_n^d} p(z_n^d; \boldsymbol{\theta^d}) p(w_n^d | z_n^d; \boldsymbol{\beta}) d\boldsymbol{\theta^d} \tag{8}$$

To get the probability of all the documents in the corpus, we take a product of the marginal distributions, is given by

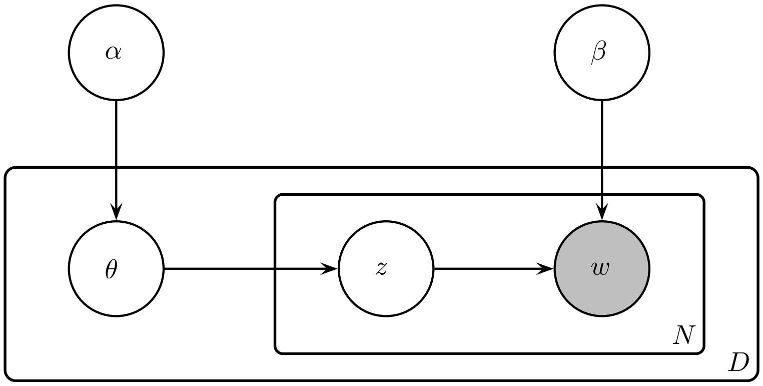


Figure 3: LDA - Graphical model

$$p(D; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{d=1}^{D} \int_{\boldsymbol{\theta^d}} p(\boldsymbol{\theta^d}; \boldsymbol{\alpha}) \prod_{n=1}^{N} \sum_{z_n^d} p(z_n^d; \boldsymbol{\theta^d}) p(w_n^d | z_n^d; \boldsymbol{\beta}) d\boldsymbol{\theta^d} \tag{9}$$

3.1 Inference

Now, for a particular document we want to compute the posterior of the hidden variables $\boldsymbol{\theta^d}$ and $\boldsymbol{z^d}$ given its words $\boldsymbol{w^d}$ and the parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$, which is given by

$$p(\boldsymbol{\theta^d}, \boldsymbol{z^d} | \boldsymbol{w^d}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{\theta^d}, \boldsymbol{z^d}, \boldsymbol{w^d}; \boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\boldsymbol{w^d}; \boldsymbol{\alpha}, \boldsymbol{\beta})} \tag{10}$$

See (7) for the values for the numerator and (8) for the values corresponding to the denominator of the posterior distribution (10). The computation of the denominator of the posterior is intractable: because, both the variables $\boldsymbol{\theta^d}$ and $\boldsymbol{z^d}$ coupled together are latent in nature. So it is not possible to compute the expectation of the posterior distribution, which leads to variational inference, where the edges between $\boldsymbol{\theta}$, $\boldsymbol{z}$ and $\boldsymbol{w}$ are removed and a simpler distribution without many dependancies, are shown in figure 4. The approximate posterior distribution based on the plate diagram provided in figure 4 is given by:

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

We are going to use the variational distribution: $\boldsymbol{q(.)}$ as an approximate posterior function for the real posterior see - (10). As the posterior distribution is intractable, we will use a variational inference algorithm (discussed in section 3.2).The EM algorithm using the variational inference is:

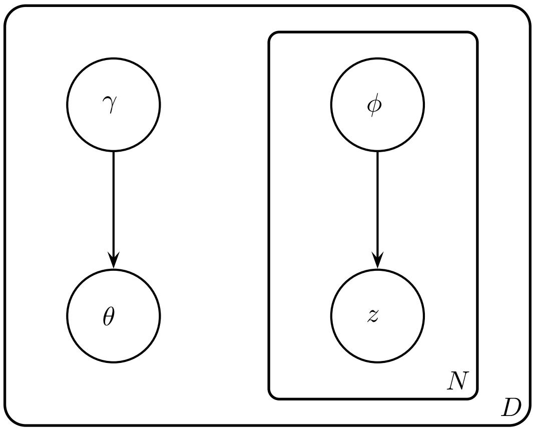


Figure 4: Variational Inference - Graphical model

E-step: For each document, get the best approximate posterior $q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma^d}, \boldsymbol{\phi^d})$ using the Variational inference algorithm.

M-step: We maximize the bounds of $\boldsymbol{q(.)}$ function with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

3.2 Variational inference

The variational inference algorithm uses the variational distribution, see (11) attempting to find the tighest lower bound on the data. Here, we will derive the lower bound for the variational EM procedure. The further part of this section will closely follow Kampa [2010] and Zhao [2013]. Here, we attempt to find parameters α and β that maximize the log likelihood of the data $\boldsymbol{w^d}$; we start by defining the probability of the data given its parameters:

$$p(\boldsymbol{w^d}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{\theta^d}, \boldsymbol{z^d}, \boldsymbol{w^d}; \boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\boldsymbol{\theta^d}, \boldsymbol{z^d} | \boldsymbol{w^d}; \boldsymbol{\alpha}, \boldsymbol{\beta})}$$

Ignoring the denominator as its just a normalizing constant, we get

$$p(\boldsymbol{w^d}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{\theta^d}, \boldsymbol{z^d}, \boldsymbol{w^d}; \boldsymbol{\alpha}, \boldsymbol{\beta})$$

To get the log likelihood of the document, we first get the marginal distribution and then take log

$$log\ p(\boldsymbol{w^d}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = log \int_{\boldsymbol{\theta^d}} \sum_{\boldsymbol{z^d}} p(\boldsymbol{\theta^d}, \boldsymbol{z^d}, \boldsymbol{w^d}; \boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{\theta^d}$$

We know that, $\mathbb{E}\left[X\right] = \int xf(x)dx$

$$= log \int_{\boldsymbol{\theta^d}} \sum_{\boldsymbol{z^d}} p(\boldsymbol{\theta^d}, \boldsymbol{z^d}, \boldsymbol{w^d}; \boldsymbol{\alpha}, \boldsymbol{\beta}) \frac{q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma^d}, \boldsymbol{\phi^d})}{q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma^d}, \boldsymbol{\phi^d})} d\boldsymbol{\theta^d}$$

$$= log \underset{q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \gamma^d, \boldsymbol{\phi^d})}{\mathbb{E}} \left[\frac{p(\boldsymbol{\theta^d}, \boldsymbol{z^d}, \boldsymbol{w^d}; \boldsymbol{\alpha}, \boldsymbol{\beta})}{q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma^d}, \boldsymbol{\phi^d})} \right]$$

applying Jenson's inequality where $f(\mathbb{E}\left[X\right]) \geq \mathbb{E}\left[f(X)\right]$ if $f(\mathbb{E}\left[X\right])$ is a convex function

$$\geq \underset{q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \gamma^d, \boldsymbol{\phi^d})}{\mathbb{E}} \left[log \left(\frac{p(\boldsymbol{\theta^d}, \boldsymbol{z^d}, \boldsymbol{w^d}; \boldsymbol{\alpha}, \boldsymbol{\beta})}{q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma^d}, \boldsymbol{\phi^d})} \right) \right]$$

$$\geq \int_{\boldsymbol{\theta^d}} \sum_{\boldsymbol{z^d}} q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma^d}, \boldsymbol{\phi^d}) \ log \ p(\boldsymbol{\theta^d}, \boldsymbol{z^d}, \boldsymbol{w^d}; \boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{\theta^d} - \int_{\boldsymbol{\theta^d}} \sum_{\boldsymbol{z^d}} q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma^d}, \boldsymbol{\phi^d}) \ log \ q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma^d}, \boldsymbol{\phi^d}) \ d\boldsymbol{\theta^d}$$

$$= \mathbb{E}\left[log \ p(\boldsymbol{\theta^d}, \boldsymbol{z^d}, \boldsymbol{w^d}; \boldsymbol{\alpha}, \boldsymbol{\beta}) \right] - \mathbb{E}\left[log \ q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma}, \boldsymbol{\phi}) \right]$$

We could factorize this further

$$= \int_{\boldsymbol{\theta^d}} \sum_{\boldsymbol{z^d}} q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \gamma, \boldsymbol{\phi}) \ log \ p(\boldsymbol{\theta^d}; \boldsymbol{\alpha}) \ d\boldsymbol{\theta^d} +$$
$$\int_{\boldsymbol{\theta^d}} \sum_{\boldsymbol{z^d}} q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \gamma, \boldsymbol{\phi}) \ log \ p(\boldsymbol{z^d} | \boldsymbol{\theta^d}) \ d\boldsymbol{\theta^d} +$$
$$\int_{\boldsymbol{\theta^d}} \sum_{\boldsymbol{z^d}} q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \gamma, \boldsymbol{\phi}) \ log \ p(\boldsymbol{w^d} | \boldsymbol{z^d}; \boldsymbol{\beta}) \ d\boldsymbol{\theta^d} +$$
$$\int_{\boldsymbol{\theta^d}} \sum_{\boldsymbol{z^d}} q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \gamma, \boldsymbol{\phi}) \ log \ q(\boldsymbol{\theta^d}; \gamma) \ d\boldsymbol{\theta^d} +$$
$$\int_{\boldsymbol{\theta^d}} \sum_{\boldsymbol{z^d}} q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \gamma, \boldsymbol{\phi}) \ log \ q(\boldsymbol{z^d}; \boldsymbol{\phi}) \ d\boldsymbol{\theta^d}$$

The lower bound is defined by

$$\begin{aligned} L(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = & \underset{q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma}, \boldsymbol{\phi})}{\mathbb{E}} \left[log\ p(\boldsymbol{\theta^d}; \boldsymbol{\alpha}) \right] + \underset{q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma}, \boldsymbol{\phi})}{\mathbb{E}} \left[log\ p(\boldsymbol{z^d} | \boldsymbol{\theta^d}) \right] \\ & + \underset{q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma}, \boldsymbol{\phi})}{\mathbb{E}} \left[log\ p(\boldsymbol{w^d} | \boldsymbol{z^d}; \boldsymbol{\beta}) \right] - \underset{q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma}, \boldsymbol{\phi})}{\mathbb{E}} \left[log\ q(\boldsymbol{\theta^d}; \boldsymbol{\gamma}) \right] \\ & - \underset{q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma}, \boldsymbol{\phi})}{\mathbb{E}} \left[log\ q(\boldsymbol{z^d}; \boldsymbol{\phi}) \right] \end{aligned} \tag{12}$$

The expectation for each component from the lower bounds equation (12) is derived below.

3.2.1 Expectations for lower bound

Solving for $\mathbb{E}_{q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma}, \boldsymbol{\phi})} \left[log\ p(\boldsymbol{\theta^d}; \boldsymbol{\alpha}) \right]$

$$p(\boldsymbol{\theta^d}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_i)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

Now, taking log over $p(\boldsymbol{\theta^d}; \boldsymbol{\alpha})$, we get

$$log\ p(\boldsymbol{\theta^d}; \boldsymbol{\alpha}) = \sum_k (\alpha_k - 1)\ log\ \theta_k + log\ \Gamma(\sum_k \alpha_k) - \sum_k^K\ log\ \Gamma(\alpha_k)$$

Now, we take expectation with respect to the q-function

$$\underset{q}{\mathbb{E}} \left[log\ p(\boldsymbol{\theta^d}; \boldsymbol{\alpha}) \right] = \sum_k (\alpha_k - 1) \underset{q}{\mathbb{E}} [log\ \theta_k] + log\ \Gamma(\sum_k \alpha_k) - \sum_k^K\ log\ \Gamma(\alpha_k) \tag{13}$$

$$\underset{q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma}, \boldsymbol{\phi})}{\mathbb{E}} [log\ \theta_k] = -\Psi \left(\sum_k \gamma_k \right) + \Psi(\gamma_k) \tag{14}$$

To know how we got the $\mathbb{E}_{q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma}, \boldsymbol{\phi})} [log\ \theta_k]$, please visit Appendix:5.4. Now, applying (14) in (13), we get,

$$\begin{aligned} \underset{q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma}, \boldsymbol{\phi})}{\mathbb{E}} \left[log\ p(\boldsymbol{\theta^d}; \boldsymbol{\alpha}) \right] = & \sum_k (\alpha_k - 1) \left[-\Psi \left(\sum_k \gamma_k \right) + \Psi(\gamma_k) \right] \\ & + log\ \Gamma \left(\sum_k \alpha_k \right) - \sum_k\ log\ \Gamma(\alpha_k) \end{aligned}$$

Now, solving for $\mathbb{E}_{q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma}, \boldsymbol{\phi})} \left[log\ p(\boldsymbol{w^d} | \boldsymbol{z^d}; \boldsymbol{\beta}) \right]$, we formulate

w^d_{nj}: equals 1 in doc d, if vocab item j is at position n
z^d_{nk}: equals 1 in doc d, if topic k is at position n

$$\text{so, } p(w^d_n|z^d_n, \boldsymbol{\beta}) = \prod_{k=1}^{K}\prod_{j=1}^{V} p(w^j_{n_d} = 1|z^k_{n_d} = 1)^{w^j_{n_d}, z^k_{n_d}}$$

For simplicity reasons, we represent

$$p(\boldsymbol{w^d}|\boldsymbol{z^d}, \boldsymbol{\beta}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\prod_{j=1}^{V} p(w^j_n = 1|z^j_n = 1)^{w^j_n, z^k_n}$$

$$= \prod_{n=1}^{N}\left\{\prod_{k=1}^{K}\prod_{j=1}^{V} \beta_{kj}^{w^j_n z^k_n}\right\}$$

Taking log and expectation, we get

$$\mathop{\mathbb{E}}_{q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma}, \boldsymbol{\phi})}\left[log\ p(\boldsymbol{w^d}|\boldsymbol{z^d}; \boldsymbol{\beta})\right] = \sum_{n=1}^{N}\left\{\sum_{k=1}^{K}\sum_{j=1}^{V}\mathop{\mathbb{E}}_{q}\left[log\ \beta_{kj}^{w^j_n z^k_n}\right]\right\}$$

$$= \sum_{n=1}^{N}\left\{\sum_{k=1}^{K}\sum_{j=1}^{V} w^j_n \mathop{\mathbb{E}}_{q}\left[z^k_n\right]\ log\ \beta_{kj}\right\}$$

Now, we will compute the expectation of z^k_n with respect to the q function

$$\mathop{\mathbb{E}}_{q(\boldsymbol{z}|\phi)}\left[z^k_n\right] = \sum_{z_n} z^k_n q(z_n|\phi)$$

$$= \sum_{z_n} z^k_n \prod_{j=1}^{K}(\phi^j_n)^{z^j_n} = \phi^k_n$$

$$\mathop{\mathbb{E}}_{q(\boldsymbol{\theta^d}, \boldsymbol{z^d}; \boldsymbol{\gamma}, \boldsymbol{\phi})}\left[log\ p(\boldsymbol{w^d}|\boldsymbol{z^d}; \boldsymbol{\beta})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{V} w^j_n \phi^k_n\ log\ \beta_{kj}$$

Now, solving for $\mathbb{E}_{q(\boldsymbol{\theta^d},\boldsymbol{z^d};\boldsymbol{\gamma},\boldsymbol{\phi})}\left[log\ p(\boldsymbol{z^d}|\boldsymbol{\theta})\right]$

$$p(\boldsymbol{z^d};\boldsymbol{\theta}) = \prod_{n=1}^{N} p(z_n^d;\boldsymbol{\theta})$$

$$= \prod_{n=1}^{N}\prod_{k=1}^{K} \theta_i^{z_{nd}^k}$$

Taking log, we get

$$log\ p(\boldsymbol{z^d};\boldsymbol{\theta}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nd}^k\ log\ \theta_k$$

Now, taking expectation over q-function

$$\mathop{\mathbb{E}}_{q(\boldsymbol{\theta^d},\boldsymbol{z^d};\boldsymbol{\gamma},\boldsymbol{\phi})}\left[log\ p(\boldsymbol{z^d};\boldsymbol{\theta})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} \mathop{\mathbb{E}}_{q}\left[z_{nd}^k\ log\ \theta_k\right]$$

We factorize this further to get,

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} \mathop{\mathbb{E}}_{q(\boldsymbol{z};\boldsymbol{\phi})}\left[z_n^{kd}\right] \mathop{\mathbb{E}}_{q(\boldsymbol{\theta};\boldsymbol{\gamma})}\left[log\ \theta_k\right]$$

$$\mathop{\mathbb{E}}_{q(\boldsymbol{z};\boldsymbol{\phi})}\left[z_n^k\right] = \sum_{z_n} z_n^k\ q(\boldsymbol{z_n};\boldsymbol{\phi})$$

$$= \sum_{z_n} z_n^k \prod_{j=1}^{K} (\phi_n^j)^{z_n^j} = \phi_n^k$$

$$\mathop{\mathbb{E}}_{q(\boldsymbol{\theta};\boldsymbol{\gamma})}\left[log\ \boldsymbol{\theta_k}\right] = \left[-\Psi\left(\sum_{j=1}^{K}\gamma_j\right) + \Psi(\gamma_k)\right]$$

Now, the factorized expectations are derived and substituted back to get

$$\mathop{\mathbb{E}}_{q(\boldsymbol{\theta^d},\boldsymbol{z^d};\boldsymbol{\gamma},\boldsymbol{\phi})}\left[log\ p(\boldsymbol{z^d}|\boldsymbol{\theta})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K}(\phi_n^k)\left[-\Psi\left(\sum_{j=1}^{K}\gamma_j\right) + \Psi(\gamma_k)\right]$$

Now, solving for $\mathbb{E}_{q(\boldsymbol{\theta^d},\boldsymbol{z^d};\boldsymbol{\gamma},\boldsymbol{\phi})}\left[log\ q(\theta;\gamma)\right]$

This is similar to the proof of $\mathbb{E}_{q(\boldsymbol{\theta^d},\boldsymbol{z^d};\boldsymbol{\gamma},\boldsymbol{\phi})}\left[log\ p(\boldsymbol{\theta^d};\boldsymbol{\alpha})\right]$

$$\mathop{\mathbb{E}}_{q(\boldsymbol{\theta^d},\boldsymbol{z^d};\boldsymbol{\gamma},\boldsymbol{\phi})}\left[log\ q(\theta;\gamma)\right] = \sum_{k=1}^{K}(\gamma_k - 1)\left[-\Psi\left(\sum_{k=1}^{K}\gamma_k\right) + \Psi(\gamma_k)\right] + log\ \Gamma\left(\sum_{k=1}^{K}\gamma_k\right) - \sum_{k=1}^{K} log\ \Gamma(\alpha_k)$$

Now, solving for $\mathbb{E}_{q(\boldsymbol{\theta^d},\boldsymbol{z^d};\boldsymbol{\gamma},\boldsymbol{\phi})}\left[log\ q(\boldsymbol{z^d};\boldsymbol{\phi})\right]$

$$q(\boldsymbol{z_n};\boldsymbol{\phi}) = \prod_{k=1}^{K}(\phi_n^k)^{z_n^k}$$

$$log\ q(\boldsymbol{z_n};\boldsymbol{\phi}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_n^k\ log(\phi_n^k)$$

$$\mathop{\mathbb{E}}_{q(\boldsymbol{\theta^d},\boldsymbol{z^d};\boldsymbol{\gamma},\boldsymbol{\phi})}\left[log\ q(\boldsymbol{z^d};\boldsymbol{\phi})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K}\mathop{\mathbb{E}}_{q(\boldsymbol{z};\boldsymbol{\phi})}\left[z_n^k\right]\ log(\phi_n^k)$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K}\phi_n^k\ log(\phi_n^k)$$

Applying all the derived expectations in (12), we get the overall lower bounds.

$$\begin{aligned} L(\gamma,\phi;\alpha,\beta) = {} & \sum_{k}(\alpha_k - 1)\left[-Psi\left(\sum_{k}\gamma_k\right) + \psi(\gamma_k)\right] \\ & + log\ \Gamma\left(\sum_{k}\alpha_k\right) - \sum_{k}\ log\ \Gamma(\alpha_k) \\ & + \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{V} w_n^j\phi_n^j\ log\ \beta_{kj} + \sum_{n=1}^{N}\sum_{k=1}^{K}(\phi_n^k)\left[-\Psi\left(\sum_{k=1}^{K}\gamma_k\right) + \Psi(\gamma_k)\right] \\ & - \sum_{k=1}^{K}(\gamma_k - 1)\left[-\Psi\left(\sum_{k=1}^{K}\gamma_k\right) + \Psi(\gamma_k)\right] + log\ \Gamma\left(\sum_{k=1}^{K}\gamma_k\right) \\ & - \sum_{k=1}^{K} log\ \Gamma(\alpha_k) - \sum_{n=1}^{N}\sum_{k=1}^{K}\phi_n^k\ log(\phi_n^k) \end{aligned}$$

Further, the derivations of the update equations for the variational Expectation Maximization are worked out at Appendix 5.3. Following is the algorithm used for the LDA variational expectation maximization.

Algorithm: Iterative Variational EM

E-step: For each document, the following iterative algorithm is used to identify the values for the variational parameters

initialize $\phi_n^{k(0)}$ with equal probability values for all k and n at 0^{th} iteration

initialize $\gamma_k^{(0)} \leftarrow \alpha_k + \frac{N_d}{K}$ for all k and n at 0^{th} iteration

```
do until convergence
    for $n = 1$ to $N_d$ do
        for $k = 1$ to $K$ do
            $\phi_n^{k,(t+1)} \leftarrow \beta_{kj} + exp\left\{\Psi(\gamma_k) - \Psi\left(\sum_{k=1}^{K} \gamma_k\right)\right\}$
        end
    Normalize $\phi_n^{k,(t+1)}$ sum to 1
    end
  $\gamma_k^{(t+1)} \leftarrow \alpha_k + \sum_{n=1}^{N} \phi_n^{k,(t+1)}$
end
```

M-step: Maximize the lower bounds with respect the model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

Until now we discussed LDA by variational approximation which is just an approximation of the posterior. Therefore we will discuss an alternative approach to variation EM, called as LDA by collapsed Gibbs sampling, but before we try to understand LDA by collapsed gibbs sampling, it is a good idea to understand about Gibbs sampling in general. In the next section 3.3 we will discuss about the Gibbs sampling technique in general and in section 3.4 we will discuss about LDA by collapsed gibbs sampling.

3.3 Gibbs sampling

Definition: Gibbs sampling is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution (i.e. from the joint probability distribution of two or more random variables), when direct sampling is difficult[3].

This section will closely follow Resnik and Hardisty [2010] and Bishop [2006]. The idea behind using this algorithm is to get the desired posterior distribution after iterating through a number of sampling steps from the conditional distribution.

Consider a probability distribution $p(\boldsymbol{Z}) = p(z_1, z_2, \ldots, z_n)$, from which we want to sample. Gibbs sampling is used to generate a sequence of samples from such a probability distribution. The gibbs sampling procedure can work with some initial state. So we initialize state values for the variables $z_1, z_2, \ldots, z_n$. Each step of the gibbs sampling would involve replacing the value of one of the variable z_i with a value sampled from a distribution of the variable conditioned on the remaining variables ie., $p(z_i|\boldsymbol{z}_{-i})$. We get one gibbs sample once we sample for all the variables in the distribution. This procedure is defined in the following pseudo code.

Gibbs sampling

```
1. Initialize $\{z_i : i = 1, \ldots, M\}$
2. for $\tau = 1, \ldots, T$:
   – Sample $z_1^{\tau+1} \sim p(z_1|z_2^{(\tau)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)}$
   – Sample $z_2^{\tau+1} \sim p(z_2|z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)}$
       ⋮
   – Sample $z_M^{\tau+1} \sim p(z_M|z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)}$
```

This procedure is repeated a number of times until the samples begin to converge to what would be sampled from the true distribution. Although the number of sampling steps required to get a desired (stationary) distribution is not known, but its theoretically proved that Gibbs sampling method will reach the desired distribution after many number of sampling steps.

The next section 3.4 will closely follow William [2011], Griffiths and Steyvers [2004] and Carpenter [2010].

3.4 Collapsed gibbs sampling for LDA

LDA by Collapsed gibbs sampling is applied to a slight extension of the LDA model given earlier, is represented as a plate diagram in figure 5. As described

[3] The definition taken from `http://en.wikipedia.org/wiki/Gibbs_sampling`

earlier in section 3, we view the documents as mixtures of topics, where each document has a different mixture from the document collection. From the plate diagram (figure 5), it can be observed that, the only change from the plate diagram presented in figure 3 is the addition of dirichlet prior η over the parameter β.

Lets start by understanding the notations being used here:

$\boldsymbol{w}$ - A vector of words per document
$\boldsymbol{z}$ - A vector of topics corresponding to each word in the document
$\boldsymbol{Z}$ - topics across all documents
$\boldsymbol{W}$ - words across all documents
$\boldsymbol{\theta}$ - A vector of topic proportions for the document
$\boldsymbol{\alpha}$ - A vector of (symmentric) dirichlet parameters over $\boldsymbol{\theta}$
$\boldsymbol{\beta}$ - the conditional probability table of the words to topics (A vector of vectors)
$\boldsymbol{\eta}$ - A vector of (symmentric) dirichlet parameters over $\boldsymbol{\beta}$

For LDA, we are interested only in the latent document-topic proportions θ_d, topic-word distributions $\beta^{(z)}$ and topic-index assignments for each word z_i. LDA gibbs sampling algorithm can be used to get each of the said latent variables and it has been noted in William [2011], θ_d and $\beta^{(z)}$ can be computed with just the topic-index assignments z_i. Therefore, a simpler algorithm called collapsed Gibbs sampler could be derived by integrating out the multi-nomial parameters and just sample z_i.

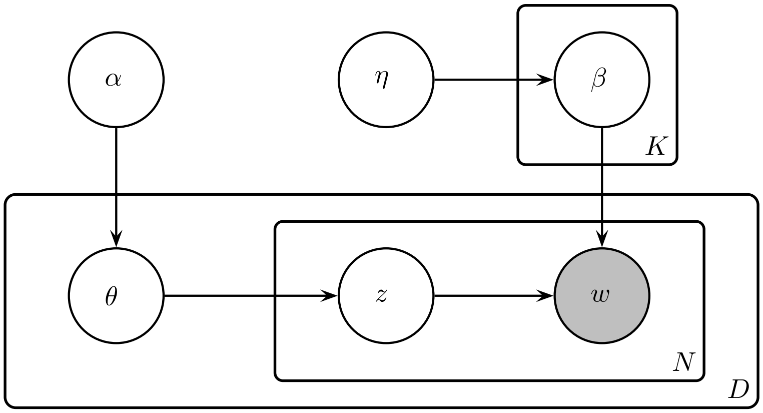


Figure 5: LDA - Gibbs sampling

Let i be some position in a document. Also, let Z_i stand for the topic at that position (which can also be written as z_n^d) and $\boldsymbol{Z}_{-i}$ stand for topic choices at all positions except Z_i

So, the idea of gibbs sampler is to compute the probability of topic z_i being assigned to w_i, given all other topic assignments $\boldsymbol{Z}_{-i}$, is given by $p(Z_i|\boldsymbol{Z}_{-i}, \boldsymbol{W}; \alpha, \eta)$.

$$p(Z_i|\boldsymbol{Z}_{-i}, \boldsymbol{W}; \alpha, \eta) = \frac{p(\boldsymbol{Z}, \boldsymbol{W}; \alpha, \eta)}{p(\boldsymbol{Z}_{-i}, \boldsymbol{W}; \alpha, \eta)} \tag{15}$$

Now, considering the denominator, from (15)

$$\begin{aligned} p(\boldsymbol{Z}_{-i}, \boldsymbol{W}; \alpha, \eta) = p(\boldsymbol{Z}_{-i}, \boldsymbol{W}_{-i}; \alpha, \eta) \times \\ \sum_k p(\boldsymbol{Z}_{-i} = k, \boldsymbol{W}_{-i}) \end{aligned} \tag{16}$$

Ignoring second multiplicant of (16), $p(Z_i|\boldsymbol{Z}_{-i}, \boldsymbol{W}; \alpha, \eta)$ is a ratio of two joint probabilities, which is

$$p(Z_i|\boldsymbol{Z}_{-i}, \boldsymbol{W}; \alpha, \eta) \propto \frac{p(\boldsymbol{Z}, \boldsymbol{W}; \alpha, \eta)}{p(\boldsymbol{Z}_{-i}, \boldsymbol{W}_{-i}; \alpha, \eta)} \tag{17}$$

Before proceeding to derive the update equations for gibbs sampler, lets introduce a few more notations for counts.

$\Omega_{d,k}$ - counts of topic k in document d
$\Psi_{k,v}$ - counts of word v in document d

Let Ω and Ψ be the counts assuming at i, topic t is chosen and $(\ )^{-i}$ are the counts of $[(\ ) - 1]$ (ie., the count of i^{th} position is excluded). If i is in document d and v is the word at position i, then

$$\Omega^{-i}_{d,t} = \Omega_{d,t} - 1 \text{ , else same}$$
$$\Psi^{-i}_{k,v} = \Psi_{k,v} - 1 \text{ , else same}$$

Given the model, the joint probability of all the parameters are provided by

$$\begin{aligned} p(\boldsymbol{W}^{1:D}, \boldsymbol{Z}^{1:D}, \boldsymbol{\theta}^{1:D}, \boldsymbol{\beta}_{1:K}; \boldsymbol{\alpha}, \boldsymbol{\eta}) &= p(\boldsymbol{\theta}; \boldsymbol{\alpha}) \times p(\boldsymbol{Z}|\boldsymbol{\theta}) \times p(\boldsymbol{\beta}; \boldsymbol{\eta}) \times p(\boldsymbol{W}|\boldsymbol{Z}, \boldsymbol{\beta}) \\ &= \prod_{d=1}^{D} p(\boldsymbol{\theta}^d; \boldsymbol{\alpha}) \prod_{k=1}^{K} p(\beta_k; \boldsymbol{\eta}) \prod_{d=1}^{D} \prod_{n=1}^{N} p(z^d_n|\boldsymbol{\theta}^d) \\ &\prod_{d=1}^{D} \prod_{n=1}^{N} p(w^d_n|z^d_n, \boldsymbol{\beta}) \end{aligned}$$

By definition, $p(\boldsymbol{W}, \boldsymbol{Z}; \boldsymbol{\alpha}, \boldsymbol{\eta}) = \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\beta}} p(\boldsymbol{W}^{1:D}, \boldsymbol{Z}^{1:D}, \boldsymbol{\theta}^{1:D}, \boldsymbol{\beta}_{1:K}; \boldsymbol{\alpha}, \boldsymbol{\eta}) d\boldsymbol{\theta} d\boldsymbol{\beta}$

$$\begin{aligned} &= \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\beta}} \prod_{d=1}^{D} Dir(\boldsymbol{\theta}^d; \boldsymbol{\alpha}) \prod_{k=1}^{K} Dir(\beta_k; \boldsymbol{\eta}) \prod_{d=1}^{D} \prod_{k=1}^{K} (\boldsymbol{\theta}^d_k)^{\Omega^d_k} \\ &\prod_{k=1}^{K} \prod_{n=1}^{N} (\beta_{k,n})^{\Psi_{k,n}}) d\boldsymbol{\theta} d\boldsymbol{\beta} \end{aligned}$$

After re-arrangement ...

$$p(\boldsymbol{W}, \boldsymbol{Z}; \boldsymbol{\alpha}, \boldsymbol{\eta}) = \int_{\boldsymbol{\beta}_{1:K}} \prod_k Dir(\boldsymbol{\beta}_k; \boldsymbol{\eta}) \prod_k (\beta_{k,n})^{\Psi_{k,n}} d\boldsymbol{\beta}$$
$$\int_{\boldsymbol{\theta}_{1:D}} \prod_d Dir(\boldsymbol{\theta}^d; \boldsymbol{\alpha}) \prod_d (\boldsymbol{\theta}^d_k)^{\Omega^d_k} d\boldsymbol{\theta}$$

$$= \int_{\beta_{1:K}} \prod_k \frac{1}{B(\boldsymbol{\eta})} \prod_k \beta_{k,n}^{\Psi_{k,n}+\eta_k-1} d\beta$$
$$\int_{\theta_{1:D}} \prod_d \frac{1}{B(\boldsymbol{\alpha})} \prod_d \theta_{k,d}^{\Omega_{k,d}+\alpha_k-1} d\theta$$

Multiply and divide by $B(\Psi_k + \eta)$ and $B(\Omega_d + \alpha)$

$$= \prod_k \frac{B(\Psi_k + \boldsymbol{\eta})}{B(\boldsymbol{\eta})} \int_{\boldsymbol{\beta}_{1:K}} \left[\frac{1}{B(\Psi_k + \boldsymbol{\eta})} \prod_k \beta_{k,n}^{\Psi_{k,n}+\boldsymbol{\eta}_k-1} \right] d\boldsymbol{\beta}$$
$$\prod_d \frac{B(\Omega_d + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})} \int_{\boldsymbol{\theta}_{1:D}} \left[\frac{1}{B(\Omega_d + \boldsymbol{\alpha})} \prod_d \theta_{d,k}^{\Omega_{k,d}+\boldsymbol{\alpha}_k-1} \right] d\boldsymbol{\theta}$$

The integrals from the above equation are the PDF's of dirichlet. Therefore they sum to 1.

$$= \prod_k \frac{B(\Psi_k + \eta)}{B(\eta)} \prod_d \frac{B(\Omega_d + \alpha)}{B(\alpha)}$$

The numerator of (17) is

$$p(\boldsymbol{Z}, \boldsymbol{W}; \boldsymbol{\alpha}, \boldsymbol{\eta}) = \prod_k \frac{B(\Psi_k + \boldsymbol{\eta})}{B(\boldsymbol{\eta})} \prod_d \frac{B(\Omega_d + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})}$$

By analogy, the denominator of (17) can be written as

$$p(\boldsymbol{Z}_{-i}, \boldsymbol{W}_{-i}; \boldsymbol{\alpha}, \boldsymbol{\eta}) = \prod_k \frac{B(\Psi_k^{-i} + \boldsymbol{\eta})}{B(\boldsymbol{\eta})} \prod_d \frac{B(\Omega_d^{-i} + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})}$$

$$\therefore \quad p(Z_i | \boldsymbol{Z}_{-i}, \boldsymbol{W}) = \prod_k \frac{B(\Psi_k + \boldsymbol{\eta})}{B(\Psi_k^{-i} + \boldsymbol{\eta})} \times \prod_d \frac{B(\Omega_d + \boldsymbol{\alpha})}{B(\Omega_d^{-i} + \boldsymbol{\alpha})} \tag{18}$$

Lets remember the Beta function and the gamma rule which we will be using for further deriving (simplying the above equation) the update equation.

$$B(\boldsymbol{\alpha}) = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)} \qquad \Bigg| \qquad \Gamma(\alpha_k + 1) = \alpha_k \Gamma(\alpha_k)$$

Now, lets work out the first and second multiplicants of (18) seperately

First multiplicant:

$$\prod_k \frac{B(\Psi_k + \boldsymbol{\eta})}{B(\Psi_k^{-i} + \boldsymbol{\eta})} \Rightarrow \frac{B(\boldsymbol{\eta} + \Psi_t)}{B(\boldsymbol{\eta} + \Psi_t^{-i})}$$

We got the above step because in LHS, the $\bar{t}$ terms have $\frac{B(\boldsymbol{\eta}+\Psi_{\bar{t}})}{B(\boldsymbol{\eta}+\Psi_{\bar{t}})}$. Now, recall v is a word at position i. Define 1_v as $(0, ..., 1, ..., 0)$, with 1 at position v, then $\Psi_t = \Psi_t^{-i} + 1_v$

$$\Rightarrow \frac{B(\boldsymbol{\eta} + \Psi_t^{-i} + 1_v)}{B(\boldsymbol{\eta} + \Psi_t^{-i})}$$

Now, consider $\boldsymbol{x} = \boldsymbol{\eta} + \Psi_t^{-i}$

$$\Rightarrow \frac{B(\boldsymbol{x} + 1_v)}{B(\boldsymbol{x})} = \frac{\prod_{\bar{v}} \Gamma(\boldsymbol{x}_{\bar{v}})\Gamma(x_v + 1)/\Gamma(\sum_v x_v + 1)}{\prod_{\bar{v}} \Gamma(\boldsymbol{x}_{\bar{v}})\Gamma(x_v)/\Gamma(\sum_v x_v)}$$

After cancellation and applying gamma rule, we get

$$\Rightarrow \frac{\Gamma(\sum_v x_v)(x_v)(\Gamma(x_v))}{\Gamma(x_v)(\sum_v x_v)(\Gamma(\sum_v x_v))} = \frac{x_v}{\sum_v x_v}$$

$$\Rightarrow \frac{\Psi_{tv}^{-i} + \eta_t}{\sum_{v'}(\eta_t + \Psi_{tv'}^{-i})}$$

$$\text{First Multiplicant } \Rightarrow \frac{\Psi_{tv}^{-i} + \eta_t}{(\sum_{v'} \Psi_{tv'}^{-i}) + V\eta} \quad \text{(when symmentric)}$$

Second multiplicant:

If d_i is doc at position i, for all $d' \neq d_i \quad \Omega_{d'}^{-i} = \Omega_{d'}$, so in $\prod_d \frac{B(\Omega_d+\boldsymbol{\alpha})}{B(\Omega_d^{-i}+\boldsymbol{\alpha})}$ the $\bar{d}_i$ terms has $\frac{B(\Omega_{\bar{d}_1}+\boldsymbol{\alpha})}{B(\Omega_{\bar{d}_i}+\boldsymbol{\alpha})}$.

$$\prod_d \frac{B(\Omega_d + \boldsymbol{\alpha})}{B(\Omega_d^{-i} + \boldsymbol{\alpha})} \Rightarrow \frac{B(\Omega_{d_i} + \boldsymbol{\alpha})}{B(\Omega_{d_i}^{-i} + \boldsymbol{\alpha})}$$

Lets define $1_t = (0, ..., 1, ..., 0)$ with 1 at position t, then $\Omega_{d_i} = \Omega_{d_i}^{-i} + 1_t$.

$$\Rightarrow \frac{B(\boldsymbol{\alpha} + \Omega_{d_i}^{-i} + 1_t)}{B(\boldsymbol{\alpha} + \Omega_{d_i}^{-i})}$$

Now, let $\boldsymbol{y} = \Omega_{d_i}^{-i} + \boldsymbol{\alpha}$.

$$\prod_d \frac{B(\Omega_d + \boldsymbol{\alpha})}{B(\Omega_d^{-i} + \boldsymbol{\alpha})} \Rightarrow \frac{B(\boldsymbol{y} + 1_t)}{B(\boldsymbol{y})}$$

By analogy from the first multiplicant,

$$\Rightarrow \frac{y_t}{\sum_{k'} y_{k'}} = \frac{\Omega_{d_it}^{-i} + \alpha_t}{\sum_{k'}(\alpha_d + \Omega_{d_ik'}^{-i})}$$

$$\text{Second multiplicant} \quad \Rightarrow \frac{\Omega_{d_it}^{-i} + \alpha}{(\sum_{k'} \Omega_{d_ik'}^{-i}) + (K\alpha)} \quad \text{(when symmentric)}$$

Substituting the first multiplicant and second multiplicant in (18), we get

$$p(Z_i|\boldsymbol{Z}_{-i}, \boldsymbol{W}) = \frac{\Psi_{tv}^{-i} + \eta_t}{(\sum_{v'} \Psi_{tv'}^{-i}) + V\eta} \times \frac{\Omega_{d_it}^{-i} + \alpha}{(\sum_{k'} \Omega_{d_ik'}^{-i}) + (K\alpha)} \tag{19}$$

The above equation will be the gibbs sampling update for the LDA. The topic probabilities for each iteration is stored and can be used for statistical analysis later. Following is the gibbs sampling algorithm for LDA.

```
Algorithm: LDA by Gibbs sampling

begin
    randomly initialize $\boldsymbol{Z}$ and increment counters (for each iteration)
    for $e$ach-iteration do
        for $e$ach-document, $n = 1 \dots N$ do
            word $\leftarrow W[n]$
            topic $\leftarrow Z[n]$
            $\Omega_{d,k} - = 1, \Psi_k, v - = 1$
            for $e$ach-topic do
                $p(Z_i|\boldsymbol{Z}_{-i}, \boldsymbol{W}) = \frac{\Psi_{tv}^{-i} + \eta_t}{(\sum_{v'} \Psi_{tv'}^{-i}) + V\eta} \times \frac{\Omega_{d_it}^{-i} + \alpha}{(\sum_{k'} \Omega_{d_ik'}^{-i}) + (K\alpha)}$
            end
            sample topic $Z[n]$ from $p(Z_i|\boldsymbol{Z}_{-i}, \boldsymbol{W})$
            $\Omega_{d,k} + = 1, \Psi_k, v + = 1$
        end
        store $Z_i$ value
    end
    return the stored $\boldsymbol{Z}$ values for analysis
end
```

This gets us to the end of the discussion on LDA by collapsed gibbs sampling. But this gibbs sampler ends up producing a number of samples (equal to the number of iterations), which can be used later for further statistical analysis. When the number of iterations is large, then the algorithm ends up producing label switching, which is a common problem that occurs with gibbs sampling. In the next section 3.4.1 we discuss the Label switching problem.

3.4.1 Label switching

Label switching is a problem identified in MCMC (Monte Carlo Markov Chain) sampling techniques on mixture models that does not always. Gibbs sampling, being a MCMC sampling technique, we will just try to understand the problem here.

To explain this problem, lets consider a dataset with word *play* having two different senses. *Sense 1* may represent a sentence or document with the word *play* being used in the context of a game, while *Sense 2* may represent a sentence or document with the word *play* being used in the context of music. The model may not distinguish all the game related documents as *Sense 1* and all the music related documents as *Sense 2*. For different samples, the model might label music related documents as *Sense 1* and game related documents as *Sense 2*. This is called the problem of label switching. Therefore, it would be dangerous to consider the mean or mode of the labels from multiple samples.

According to Stephens [2000], label switching problem arises when taking a Bayesian approach to parameter estimation and clustering using mixture models. In LDA, when we assign priors to the latent variables θ and β, and they are sampled from symmetric hyper-parameters, which will make the prior distributions of each sample to be symmetric. Therefore our posteriors will also be symmetric, which leads to label switching. Say for example, we have documents with topics related to *dogs* and *medicine*. In one sample, the *dogs* model will assign *Topic1* to words related to *dogs* while assigning *Topic2* to words related to *medicine* and in a different sample, the words related to *medicine* will take *Topic1* and *dogs* will take *Topic2*. A different illustrative example is provided in Stephens [2000] to explain the same problem.

Here, we will not discuss the various solutions that have been proposed.

4 Conclusion

LDA is a statistical model to discover topics from a huge collection of text. LDA by variational approximation and collapsed gibbs sampling were discussed in detail in this report. Although LDA by variational approximation was the very first proposed method to discover topics from text, the gibbs sampling approach is more promising as the former approach is just an approximation of the later one. The collapsed gibbs sampling approach will be very useful in getting the actual posterior rather than the approximate posterior. Also, in the Gibbs sampling approach, as we get a number of samples of the posterior, we have the freedom to pursue further statistical analysis with the samples. These LDA methods need not necessarily be used as a standalone application, bu the topic distributions inferred can be used for further tasks such as sense disambiguation, information retrieval and question answering.

The learning experience has been enriching and I have planned to use the knowledge earned by learning this technique in my PhD for *Diachronic analysis of word-sense*[4]. Additionally, I have performed experiments with different LDA toolkits used in the state-of-the-art and seperate report on the results obtained is submitted with this report.

[4] Diachronic analysis of word sense is to understand the semantic change of a word over time and space `http://en.wikipedia.org/wiki/Semantic_change`

5 Appendix

5.1 Generate 2D plots

```
require(MCMCpack)
library(ggplot2)
alpha <- 1
draws <- 15
dimen <- 10
x <- rdirichlet(draws, rep(alpha, dimen))
dat <- data.frame(item=factor(rep(1:10,15)),
       draw=factor(rep(1:15,each=10)),
       value=as.vector(t(x)))
library(ggplot2)
gplot <- ggplot(dat,aes(x=item,y=value,ymin=0,ymax=value)) +
     geom_point(colour=I(”blue”))       +
     geom_linerange(colour=I(”blue”))   +
     facet_wrap(~draw,ncol=5)           +
     scale_y_continuous(lim=c(0,1))
postscript(file = paste(‘test’,␣‘.eps’, sep=””), width = 5,
         height = 5)
print(gplot)
dev.off()
```

5.2 Generate 3D plots

```
a1 <- 0.1
a2 <- 0.1
a3 <- 0.1
x1 <- x2 <- seq(0.01, .99, by=.01)

f <- function(x1, x2){
        term1 <- gamma(a1+a2+a3)/(gamma(a1)*
                gamma(a2)*gamma(a3))
        term2 <- x1^(a1-1)*x2^(a2-1)*(1-x1-x2)^(a3-1)
        term3 <- (x1 + x2 < 1)
        term1*term2*term3
        }

z <- outer(x1, x2, f)
z[z<=0] <- NA
persp(x1, x2, z,
        main = "Alpha=[0.1,0.1,0.1]",
        col = "lightblue",
        theta = 50,
        phi = 20,
        r = 50,
        d = 0.1,
        expand = 0.5,
        ltheta = 90,
        lphi = 180,
        shade = 0.75,
        ticktype = "detailed",
        nticks = 5,
        zlim = if(length(na.omit
                (unique(as.vector(z))))<=1) {
                c(0,2.1)
        } else {
                range(z, na.rm = TRUE)
        })
```

5.3 Variational parameters for Update

We should now, learn the parameter update equations for the variation expectation. The variational parameters can be learnt by taking derivatives with respect to ϕ and γ.

$$L(\phi;\beta) = \phi_n^k \; log \; \beta_{kj} + \phi_n^k \left[-\Psi \left(\sum_{k=1}^{K} \gamma_k \right) + \Psi(\gamma_k) \right] - \phi_n^k \; log \; \phi_n^k$$

Applying Lagrangian, we get

$$= \phi_n^k \; log \; \beta_{kj} + \phi_n^k \left[-\Psi \left(\sum_{k=1}^{K} \gamma_k \right) + \Psi(\gamma_k) \right] - \phi_n^k \; log \; \phi_n^k + \lambda_n \left(\sum_k \phi_n^k - 1 \right)$$

Differenting with respect to ϕ_n^k

$$\frac{\partial L}{\partial \phi_n^k} = log \; \beta_{kj} - \Psi \left(\sum_{k=1}^{K} \gamma_k \right) + \Psi(\gamma_k) - log \; \phi_n^k + \lambda_n - 1 = 0$$

$$log \; \phi_n^k = log \; \beta_{kj} + \Psi(\gamma_k) - \Psi \left(\sum_{k=1}^{K} \gamma_k \right) + \lambda_n - 1$$

$$\phi_n^k = \beta_{kj} + exp \left\{ \Psi(\gamma_k) - \Psi \left(\sum_{k=1}^{K} \gamma_k \right) \right\} + exp \left\{ \lambda_n - 1 \right\}$$

As, $exp \left\{ \lambda_n - 1 \right\}$ is just a constant, we can ignore this

$$\phi_n^k \propto \beta_{kj} + exp \left\{ \Psi(\gamma_k) - \Psi \left(\sum_{k=1}^{K} \gamma_k \right) \right\}$$

Now, we derive the update equation for gamma.

$$
\begin{aligned}
L(\gamma;\alpha) = &\sum_{k=1}^{K}(\alpha_k - 1)\left[-\Psi\left(\sum_k \gamma_k\right) + \Psi(\gamma_k)\right] \\
&+ \sum_n \sum_k \Phi_n^k \left[-\Psi\left(\sum_{k=1}^{K} \gamma_k\right) + \Psi(\gamma_k)\right] \\
&- \sum_{k=1}^{K}(\gamma_k - 1)\left[-\Psi\left(\sum_k \gamma_k\right) + \Psi(\gamma_k)\right] + log\ \Gamma\left(\sum_k \gamma_k\right) - log\ \Gamma(\gamma_k)
\end{aligned}
$$

$$
\begin{aligned}
= &\sum_k \left[\Psi(\gamma_k) - \Psi\left(\sum_k \gamma_k\right)\right]\left[\alpha_k - 1 + \sum_{n=1}^{N} \phi_n^k - \gamma_k + 1\right] \\
&+ log\ \Gamma\left(\sum_k \gamma_k\right) - log\ \Gamma(\gamma_k) \quad (re - arrangement)
\end{aligned}
$$

$$
\begin{aligned}
= &\sum_{k=1}^{K}\left(\alpha_k + \sum_{n=1}^{N} \phi_n^k - \gamma_k\right)\Psi(\gamma_k) - \Psi\left(\sum_{k=1}^{K} \gamma_k\right)\sum_{k=1}^{K}\left(\alpha_k + \sum_{n=1}^{N} \phi_n^k - \gamma_k\right) \\
&+ log\ \Gamma\left(\sum_k \gamma_k\right) - log\ \Gamma(\gamma_k) \quad (re - arrangement)
\end{aligned}
$$

Differentiate with respect to γ_k

$$
\begin{aligned}
\frac{\partial L}{\partial \gamma_k} = &\sum_{k=1}^{K}\left(\alpha_k + \sum_{n=1}^{N} \phi_n^k - \gamma_k\right)\Psi(\gamma_k) - \Psi\left(\sum_{k=1}^{K} \gamma_k\right)\sum_{k=1}^{K}\left(\alpha_k + \sum_{n=1}^{N} \phi_n^k - \gamma_k\right) \\
&+ \Psi\left(\sum_k \gamma_k\right) - \Psi(\gamma_k) = 0
\end{aligned}
$$

With a further bit of re-arrangement and cancellation, we get

$$
\gamma_k = \alpha_k + \sum_{n=1}^{N} \phi_k
$$

For the maximization step, we get the parameter updates for α and β to maximize the lower bound with respect to the model parameters α and β.

For β

$$L(\gamma, \phi; \alpha, \beta) = L_{\beta_{kj}} = \sum_{d=1}^{D} \sum_{n=1}^{N_d} w_n^j \phi_n^k \; log \; \beta_{kj}$$

$$L(\alpha, \beta, \lambda) = L_{\beta_{kj}} + \lambda_k \left(\sum_{j=1}^{V} \beta_{kj} - 1 \right)$$

Taking derivative with respect to β_{kj}

$$\frac{\partial L}{\partial \beta_{kj}} = \frac{1}{\beta_{kj}} \sum_{d=1}^{D} \sum_{n=1}^{N_d} w_n^j \phi_n^k = 0$$

$$\beta_{kj} = \sum_{d=1}^{D} \sum_{n=1}^{N_d} w_n^j \phi_n^k$$

I have not provided the derivation for the update for α as I have not completely understood this yet.

5.4 Expectation of dirichlet in exponential family

We will now exploit a distribution in exponential family

$$p(x;\eta) = h(x)\ exp\{\eta^T T(x) - A(\eta)\}$$

Dirichlet, in its exponential family has the form

$$p(\boldsymbol{\theta^d};\boldsymbol{\alpha}) = exp\left\{\sum_k (\alpha_k - 1)\ log\ \theta_k + log\ \Gamma(\sum_k \alpha_k) - \sum_k^K\ log\ \Gamma(\alpha_k)\right\}$$

The exponential parameters are,

(1) $T(\theta_k) =\ log\ \theta_k$

(2) $h(\theta) = 1$

(3) $\eta_k = \alpha_k$

(4) $A(\eta) = -log\ \Gamma(\sum_k \alpha_k) + \sum_k\ log\ \gamma(\alpha_k)$

Further, $\underset{p(\boldsymbol{\theta};\boldsymbol{\alpha})}{\mathbb{E}}[T(\theta_k)] = \dfrac{\partial A(\eta)}{\partial \eta_k}$

$$\therefore \underset{p(\boldsymbol{\theta};\boldsymbol{\alpha})}{\mathbb{E}}[T(\theta_k)] = \frac{\partial}{\partial(\alpha_k - 1)}\left[-log\ \Gamma\left(\sum_k \alpha_k\right) + \sum_k\ log\ \Gamma(\alpha_k)\right]$$

$$= -\frac{1}{\Gamma(\sum_k \alpha_k)}\frac{\partial \Gamma(\sum_k \alpha_k)}{\partial(\sum_k \alpha_k)} + \frac{1}{\Gamma(\sum_k \alpha_k)}\frac{\partial \gamma(\alpha_k)}{\partial \alpha_k}$$

$$= -\Psi\left(\sum_k \alpha_k\right) + \Psi(\alpha_k)$$

References

Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, February 2006.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. In *Journal of Machine Learning Research*, volume 3, pages 993–1022,, March 2003.

Bob Carpenter. Integrating out multinomial parameters in latent dirichlet allocation and naive bayes for collapsed gibbs sampling, September 2010.

Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

Jonathan Huang. Maximum likelihood estimation of dirichlet distribution parameters. Technical report, Stanford University, 2005.

Kittipat “Bot” Kampa. Derivation of inference and parameter estimation algorithm for latent dirichlet allocation (lda), June 2010. A self-published tutorial.

Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

Philip Resnik and Eric Hardisty. Gibbs sampling for the uninitiated. Lamp-tr-153, University of Maryland, June 2010.

Matthew Stephens. Dealing with label switching in mixture models. In *Journal of the Royal Statistical Society*, volume 62 of *B (Statistical Methodology)*, pages 795 – 809. University of Oxford, UK, Blackwell Publishing for the Royal Statistical Society, 2000.

Darling M. William. A theoretical and practical implementation tutorial on topic modeling and gibbs sampling, December 2011.

Wayne Xin Zhao. Varitional methods for latent dirichlet allocation, 2013. Technical Note.

Wanchuang Zhu and Yanan Fan. Relabelling algorithms for large dataset mixture models. March 2014.