Northeastern University
2016 EECE PhD Qualifying Exam

By : Chieh Wu
ID : 000 176 171
Cell : 978 944 7388
Email : wu.chie@husky.neu.edu

March / 7 / 2016

Oral Presentation

I have a class on Monday and Thursday at 11:30 to 1:30.
Although I am allowed to miss a class, I would prefer not
to miss any classes. Besides the class, the only date
that I cannot attend is March/24/2016.

Thoughts on Latent Dirichlet Allocation

BY CHIEH WU

March / 7 / 2016

Abstract

The purpose of this paper is to provide an introduction into the theory and mathematical understanding of LDA. This paper will first present LDA through the lens of its ancestral algorithms; unigram, mixture of unigram and pLSI. The paper will also present a variational inference derivation that is different from the common approach. Lastly, the paper will discuss limitations to the algorithm as well as researches that has been accomplish to overcome these limitation.

1 Introduction

A useful interpretation of Latent Dirichlet Allocation (LDA) is as a dimensionality reduction technique. Given data x with d features, it is not guaranteed that all features will be helpful. In those cases, the extra features provide no extra information at best, and, at worst, confounds and buries the information in noise. To combat the curse of dimensionality, dimensionality reduction is an approach to concentrate the useful portion of the features into to a lower dimensional representation.

The dimensionality reduction characteristic of LDA has wide implications in various fields. A common examples of LDA implementation is document labeling. Given the explosion of text data in the world wide web, the amount of text content on the web is simply astronomical. To organize this vast amount of information, it would be extremely helpful if the data could be collected into specific topics and subtopics for fast search purpose. Unfortunately, the aggregation of similar topics can be done only if the topic of a document is known a priori. Yet, since the content for the vast majority of the websites is unknown, trying to cluster them into similar sites without knowing their content is impossible.

One obvious way to resolve this problem, and simultaneously end unemployment, is to hire billions of people to read the documents online and label each document with its content. Unfortunately for unemployment, the cost of this endeavor would be prohibitively expensive even for Google.

This is where LDA come into the picture to provide an automatic labeling algorithm. As mentioned previously, LDA is a dimensionality reduction technique. From the LDA perspective, each document is analogous to a single sample x with d words. For each document, LDA is capable of finding the hidden list of g topics. For example, article one of 1000 words might be about money and sports and article two of 1000 words might be about money and politics. Each document is initially represented as a 1000 dimensional dataset of word. The effect of LDA significantly reduces the documents into 2 dimension dataset of topics.

By understanding the dimensionality perspective of LDA, the technique could easily be applied to other fields such as genetic biology. Similar to text documents, the human genome are also made up of alphabets and strings. Unlike the 26 letter English alphabet, the human genome only consists of 4 letters, ATGC. By employing dynamic programming techniques such as the Longest Common Subsequence [10], words could be discovered within the DNA. The DNA of a person is similar to a document and certain key genetic sequences are the words. Applying these analogies to LDA, the entire genetic makeup of a person could be reduced down to a handful of dominant genetic variations.

Similar ideas has also been proposed even in the computer vision community [11]. In image analysis, parts of the image could be broken down into visual signals analogous to words in a document. With each image as a single document, LDA could be applied to discover the hidden themes within the image.

Although LDA can be in the context of genetics data, image data, or text data, the rest of the paper will refer to the text data as the base example. Since the original paper on LDA used text data as motivation, to provide a consistent language, this paper will follow the historical convention and notations.

Although, viewing LDA as a dimensionality reduction technique is a convenient interpretation, a deeper understanding points towards LDA's ability to infer the hidden themes from observed data through Bayesian statistics. Section 2 of this paper will begin the process of unfolding LDA's probabilistic insight by presenting historical techniques leading up to LDA. Previous text modeling techniques such as the unigram, mixture of unigrams, and pLSI will be presented here. Section 3 of the paper will provide the essential mathematical background required for LDA. Namely, section 3 will concentrate on a technique called Variational Inference, which provides the key insight to infer the posterior distribution of the hidden variables. Section 4 will combine the mathematical insight from section 3 to demonstrate how LDA uses variational inference to discover its latent variables. The paper will discuss in section 5 some of the limitations of LDA, as well as potential solutions that could resolve them.

2 Techniques Leading up to LDA

To understand the various probabilistic techniques leading up to LDA, it is essential to first step back and look at the generative model at large. The key idea in a generative model is that the data we observe came from a probabilistic model we wish to discover. For example, imagine observing multiple lists of random words.

$$\boldsymbol{w}_1 = [\text{ cat, car, sock, hat, shirt }]$$
$$\boldsymbol{w_2} = [\text{ bird, boat, goat, pants, shirt }]$$
$$\vdots$$

The generative model assumes that these list of words are generated from some internal mechanism. By observing many examples of these lists, the goal is to reverse engineer the underlying mechanism that generated the list. If the proposed mechanism is accurately depicting the real system, then our system should be able to regenerate very similar lists compared to the ones we have observed.

To this end, a simple technique called the unigram model was proposed. Imagine going through thousands of documents and count the word frequency of every word in the dictionary. The result of this endeavor is a histogram of words. By normalizing the histogram, it yields a probability mass function of each word within the dictionary. In other words, the more often a word appears, the higher the probability.

word	cat	car	hat	...	shirt
probability	0.01	0.04	0.2	...	0.01

The table above shows an example of calculating the probability of each word.

Table 2.1.

The unigram model proposes that the list of words was generated from the table above. If we have a bag of words, each with its own corresponding probability, a list of words could be generated if we reach into the bag and randomly grab 5 words. If we let w_n be the nth word drawn, and $\boldsymbol{w}$ as the list of words. The unigram model assumes that N words within the list are independently drawn. Therefore, the probability of $\boldsymbol{w}$ would be modeled as :

$$\boldsymbol{w} = (w_1, w_2, w_3, w_4, w_5) \qquad p(\boldsymbol{w}) = \prod_{n=1}^{N} p(w_n) \tag{2.1}$$

It is easy to see the limitation of the unigram model because it naively assumes that all documents came from the same distribution of words. Yet, everyday experience suggests that the likelihood of a word heavily depends on the topic. To take this feature into consideration, a mixture of unigram was proposed. Under this model, instead of having only a single distribution of words, there are multiple distributions depending on the topic. Mixture of unigrams allows for each document to come from a different topic. The list of words is drawn by first initializing on a topic z, and base on the chosen topic, the words are drawn from the corresponding distribution. This model is formulated with the following equation.

$$p(\boldsymbol{w}) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n|z) \tag{2.2}$$

To transition our thinking, terms such as **topic** and **documents** can take on a more mathematical interpretation. Since a document is literally a list of words, this paper will denote $\boldsymbol{w}$ as a **document**. The term **topic** is simply a distribution of words as demonstrated in table 2.1. Depending on the topic, a different distribution of the same words will arise. It is with this perspective that the shortcoming the mixture of unigram becomes obvious. While allowing different documents to exhibit different topics, each document is still restricted to a single topic. The improved flexibility of mixture of unigram neglected the fact that a document can actually touch upon multiple topics. For example, articles that talk about overpaid athletes is not an article only about sport, but also about salary and even social justice. Therefore, while the mixture of unigram is an improvement, it still lack the ability to represent a single document as containing combination of issues. It is with this limitation that motivated techniques such as the probabilistic latent semantic indexing (pLSI).

With pLSI, a set of M training documents is previously chosen, analyzed and topic labeled. This initial dataset will be referred to as the training set. From this model, a document of N words is generated by picking N documents from the training set. From the N documents chosen, N words are picked from each document. From this process, notice that a document of 10 words may have mixtures of topics, e.g. 3 words from money topic, 5 words from sports topic, and 2 words from salary topic. Given d as the original training documents, and z as the the chosen topic, this generative process is described with the following equation.

$$p(d, w_n) = p(d) \sum_z p(w_n|z)p(z|d) \tag{2.3}$$

At this point of the evolution, the advantage of one generation of algorithm improves upon the previous by including neglected assumption into the new model. LDA provides the next generation of improvements in that it addresses major flaws in the pLSI model. First, since the pLSI model requires an initial training set, it is not clear how the model can handle documents outside of the training set. In this sense, the pLSI models neglected to model how the documents themselves are generated. Without a document level model, it forces the model to include a large training set to cover more topic possibilities. Yet, as the size of the training set grows, the necessary parameters to estimate also grows linearly with the number of training documents. This expansion of the arguments quickly and easily lead into the problem of over fitting.

Upgrading to LDA, the fixed training documents from the pLSI can now be explained by a Dirichlet distribution. Where pLSI only has one training set, LDA extend this idea to infinite number of potential training sets. To generate a document, LDA first draw a single training set, it is from this randomly drawn training set, that it then perform pLSI to generate the document. To model this process, the Dirichlet distribution has the following form.

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\sum_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \ldots \theta_k^{\alpha_k - 1} \qquad \begin{array}{l} \alpha : \text{hyper} - \text{parameter} \\ k : \text{number of topics} \end{array} \tag{2.4}$$

Once a training set θ is chosen, N topics are then chosen from it. This process can be represented in a categorical distribution z.

$$p(z|\theta) = \prod_{i=1}^{k} \prod_{j=1}^{N} \theta_{i,j}^{z_{i,j}} \qquad k : \text{number of topics} \tag{2.5}$$

Lastly, for each topic drawn from the training set, a word is drawn to place into the document. Given V as the number of words in the dictionary and β be a $k \times V$ probability matrix. The probability of each word given the topic has the following form.

$$p(\boldsymbol{w}|\boldsymbol{z}, \beta) = \prod_{n=1}^{N} \beta_{z_n, w_n} \tag{2.6}$$

Given these formulations, the total joint distribution of the LDA modeled as :

$$p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) = p(\boldsymbol{w}|\boldsymbol{z}, \beta)p(\boldsymbol{z}|\theta)p(\theta|\alpha) \tag{2.7}$$

Writing the entire formula out would yield :

$$p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) = \left(\frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\sum_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} ... \theta_k^{\alpha_k - 1} \right) \prod_{n=1}^{N} \beta_{z_n, w_n} \theta_{z_n} \tag{2.8}$$

Similar to previous models, it is desirable to find $p(\boldsymbol{w}|\alpha, \beta)$. To accomplish this, it requires the marginalization of the variables θ and $\boldsymbol{z}$.

$$p(\boldsymbol{w}|\alpha, \beta) = \int \left(\frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\sum_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} ... \theta_k^{\alpha_k - 1} \right) \prod_{n=1}^{N} \sum_{z} \prod_{j=1}^{V} (\theta_z \beta_{i,j})^{w_n^j} \, d\theta \tag{2.9}$$

It is easy to notice the intractability of calculating the marginal distribution. Even in other cases where the discovery of $p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta)$ may be more applicable, it is not possible to escape the need to calculate $p(\boldsymbol{w}|\alpha, \beta)$. This reason can be easily seen from the Bayes theorem shown below.

$$p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta) = \frac{p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)}{p(\boldsymbol{w}|\alpha, \beta)} \tag{2.10}$$

Regardless whether the desirable target is $p(\boldsymbol{w}|\alpha, \beta)$ or $p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta)$ it is necessary to calculate and solve an extremely difficult marginal distribution. Due to the difficulty of this problem, various techniques have been invented over time to estimate and approximate the posterior distribution using only the joint distribution. Techniques such as MCMC and variational inference are among the top choices that has yield great sucess. Since variational inference was the approach suggested by the original paper, this paper will also concentrate on using variational inference to solve for the posterior distribution. The basic motivation and concept of variational inference will be introduced in the next section.

3 Variational Inference

A key feature of probabilistic modeling is to infer some unknow mechanism of a system z through various observations of x. Since the probability of x depends on z, it is normally a simple task to produce the joint pdf, i.e, $p(x, z) = p(x|z)p(z)$. However the posterior distribution of z given x requires an integral of the joint distribution in the denominator.

$$p(z|x) = \frac{p(x, z)}{\int_z p(x, z) \, dz} \tag{3.1}$$

If the integral can be solved, the posterior can be easily found. However, the complexity of the system might yield an intractable integral. This identical situation is seen in the LDA case where the denominator cannot be solved. The purpose of variational inference is to find an estimation function $q(z)$ that approximate $p(z|x)$ as closely as possible. To accomplish this task, it is useful to defines the KL divergence between $q(z)$ and $p(z|x)$ shown in the following equation.

$$\mathrm{KL}(q(z)||p(z|x)) = -\sum q(z) \ln \frac{p(z|x)}{q(z)} \qquad \text{Given:} \quad p(z|x) = \frac{p(z, x)}{p(x)} \tag{3.2}$$

Using the KL divergence above along with the given information, it can be manipulated into a more useful relationship.

$$\ln p(x) = \mathrm{KL}(q(z)||p(z|x)) + \sum q(z) \ln \frac{p(x, z)}{q(z)} \tag{3.3}$$

To simplify this equation, the second portion is normally denoted with $\mathcal{L}$ symbol. This symbol stands for the lowerbound, and the simplified relation is often written as :

$$\ln p(x) = \mathrm{KL}(q||p) + \mathcal{L} \tag{3.4}$$

It is worth emphasizing that the goal is to approximate $p(z|x)$ using $q(z)$. This could be done through the minimization of the KL divergence. However, the equation above demonstrates that given a constant $\ln p(x)$, maximizing the lowerbound $\mathcal{L}$ is equivalent to minimizing the KL divergence. Therefore, solving for the optimize approximation $q^*(z)$ is reduced into the following problem.

$$q^*(z) = \underset{q(z)}{\operatorname{argmax}} \sum q(z) \ln \frac{p(x,z)}{q(z)} \tag{3.5}$$

Notice that since the joint pdf of $p(x, z)$ is assumed to be know, the optimization in (3.5) is significantly easier than attempting to minimize the KL divergence itself. Although not always necessary, Bishop suggests that equation (3.5) can be further simplified using mean field theory to factorize $q(z)$ into independent distributions [5].

$$q(z) = \prod_{i=1}^{M} q(z_i) \tag{3.6}$$

Using this simplification, the optimal solution for each $q(z_i)$ can be expressed in the following equation.

$$\ln q^*(z_j) = E_{i\neq j}[\ln p(x,z)] + \text{const} \tag{3.7}$$

Finally, the optimal solution can be solved with the following equation.

$$q^*(z_j) = \frac{\exp\left(E_{i\neq j}[\ln p(x,z)]\right)}{\int \exp\left(E_{i\neq j}[\ln p(x,z)]\right) dz_j} \tag{3.8}$$

Given this understanding of the variation inference, we are now ready to apply these concepts in the next section to LDA.

4 Applying variational inference to LDA

As stated in the previous sections, a key objective of LDA is to estimate the posterior distribution.

$$p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta) = \frac{p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)}{p(\boldsymbol{w}|\alpha, \beta)} \tag{4.1}$$

Due to the difficulty of calculating the denominator term, variational inference uses an estimator, $q(z, \theta)$, to approximate the posterior. Furthermore, applying equation (3.5) to the LDA formulation, the optimal estimator could be achieved from maximizing the lowerbound, $\mathcal{L}$. This representation is shown in the equation below.

$$q^*(z,\theta) = \underset{q(z,\theta)}{\operatorname{argmax}} \sum q(z) \ln \frac{p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)}{q(z,\theta)} \tag{4.2}$$

At this point, two strategies emerge to maximize the lowerbound. The first approach follows the original paper by formulating the entire lower bound and maximizing it as an optimization problem. Using the first approach, the parameters of $q(z, \theta)$ can be found by taking the derivative of the lowerbound and setting it to zero. After reading much of the literature, this approach appears to be the standard approach taught among tutorials and presentations. Although this approach has historical precedence, the proof itself is long and difficult to understand. The proof also requires a reasonable guess of the posterior distribution.

For these reasons, this paper will propose a different perspective using Bishop's formulation of variational inference. By viewing the minimization of the lowerbound as a form of KL divergence, the proof could leverage equation (3.7) to directly solve for the estimator distributions without making any assumptions on its form and arguments. After having solved the problem in both ways, it is this paper's opinion that Bishop's formulation is a superior approach to present the derivation. It is therefore surprising that this formulation is not seen anywhere among the literatures. Since the original approach is already covered by various papers and tutorials, the contribution of this paper will present an alternative perspective from the traditional method. This paper will also use the same notations of LDA presented in section two of the paper.

Let k be the number of topics. The first step is to use the mean field theory to separate out dependencies within the estimator.

$$q(z|\theta) = q(\theta)\prod_{n=1}^{N} q(z_n) \tag{4.3}$$

Next, we use equation (3.7) to calculate each independent component of the estimator.

$$\ln q^*(\theta) = E_{q(z)}[\ln p(\boldsymbol{z}, \theta, \boldsymbol{w}|\alpha, \beta)] + \text{const} \tag{4.4}$$

For now, let's ignore the constant term and concentrate on the first portion of the equation (4.4).

$$E_{q(z)}[\ln p(\boldsymbol{z}, \theta, \boldsymbol{w}|\alpha, \beta)] = E_{q(z)}[\ln [p(\theta|\alpha) \ln p(\boldsymbol{z}|\theta)\, p(\boldsymbol{w}|\boldsymbol{z}, \beta)]] \tag{4.5}$$

To simplify some notations, let : $\bar{\Gamma}(\alpha) = \frac{\Gamma(\sum \alpha_i)}{\prod \Gamma(\alpha_i)}$, and N be the number of words in a document.

$$E_{q(z)}[\ln p(\boldsymbol{z}, \theta, \boldsymbol{w}|\alpha, \beta)] = E_{q(z)}\left[\ln \left[\left[\overline{\Gamma(\alpha)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right] \left[\prod_{i=1}^{k} \prod_{n=1}^{N} \theta_i^{z_{i,n}} \right] \left[\prod_{i=1}^{k} \prod_{n=1}^{N} \beta^{z_{i,n}} \right] \right] \right] \tag{4.6}$$

$$E_{q(z)}[\ln p(\boldsymbol{z}, \theta, \boldsymbol{w}|\alpha, \beta)] = E_{q(z)}\left[\ln \left[\overline{\Gamma(\alpha)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \prod_{n=1}^{N} \theta_i^{z_{i,n}} \beta^{z_{i,n}} \right] \right]$$

$$E_{q(z)}[\ln p(\boldsymbol{z}, \theta, \boldsymbol{w}|\alpha, \beta)] = E_{q(z)}\left[\ln \bar{\Gamma} + \sum_{i=1}^{k} \left[(\alpha_i - 1) \ln \theta_i + \sum_{n=1}^{N} z_{i,n} \ln (\theta_i \beta) \right] \right]$$

$$E_{q(z)}[\ln p(\boldsymbol{z}, \theta, \boldsymbol{w}|\alpha, \beta)] = \ln \overline{\Gamma(\alpha)} + \sum_{i=1}^{k} \left[(\alpha_i - 1) \ln \theta_i + \sum_{n=1}^{N} E_{q(z)}[z_{i,n}] \ln (\theta_i \beta) \right]$$

Since $q(z)$ is unknown, we let $\phi_{i,n}$ for now denote $E_{q(z)}[z_{i,n}]$, and $\mathcal{C}$ to denote the constant.

$$\ln q^*(\theta) = \ln \overline{\Gamma(\alpha)} + \sum_{i=1}^{k} \left[(\alpha_i - 1) \ln \theta_i + \sum_{n=1}^{N} \phi_{i,n} \ln (\theta_i \beta) \right] + \mathcal{C} \tag{4.7}$$

$$q^*(\theta) = (\overline{\Gamma(\alpha)}\, e^{\mathcal{C}})\, e^{\left(\sum_{i=1}^{k} [(\alpha_i - 1) \ln \theta_i + \sum_{n=1}^{N} \phi_{i,n} \ln (\theta_i \beta)]\right)} \tag{4.8}$$

$$q^*(\theta) = (\overline{\Gamma(\alpha)}\, e^{\mathcal{C}}) \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \prod_{n=1}^{N} \theta_i^{\phi_{i,n}} \beta^{\phi_{i,n}} \tag{4.9}$$

$$q^*(\theta) = \left(\overline{\Gamma(\alpha)}\, e^{\mathcal{C}} \prod_{n=1}^{N} \beta^{\phi_{i,n}} \right) \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \prod_{n=1}^{N} \theta_i^{\phi_{i,n}} \tag{4.10}$$

Since this is a function of θ, we can treat the entire first term as a constant $\mathcal{K}$.

$$q^*(\theta) = \mathcal{K} \prod_{i=1}^{k} \theta_i^{\left(\alpha_i + \sum_{n=1}^{N} \phi_{i,n}\right) - 1} \tag{4.11}$$

From equation (4.11), it is clear that the posterior distribution is a Dirichlet distribution. If we denote the parameters of this distribution with γ, we would get the following distribution.

$$q^*(\theta) = \overline{\Gamma(\gamma)} \prod_{i=1}^{k} \theta_i^{\gamma_i - 1} \qquad \text{where:} \quad \gamma = \alpha_i + \sum_{n=1}^{N} \phi_{i,n} \tag{4.12}$$

A similar approach could be taken to find $q(z_i)$.

$$\ln q^*(z_i) = E_{q \neq q(z_i)}[\ln p(\boldsymbol{z}, \theta, \boldsymbol{w} | \alpha, \beta)] + \text{const} \tag{4.13}$$

Similar to equation (4.6), except now we find the expectation of $q(\theta)$.

$$E_{q(\theta)}[\ln p(\boldsymbol{z}, \theta, \boldsymbol{w} | \alpha, \beta)] = E_{q(\theta)}\left[\ln \left[\left[\overline{\Gamma(\alpha)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right] \left[\prod_{i=1}^{k} \prod_{n=1}^{N} \theta_i^{z_{i,n}} \right] \left[\prod_{i=1}^{k} \prod_{n=1}^{N} \beta^{z_{i,n}} \right] \right] \right] \tag{4.14}$$

$$E_{q(\theta)}[\ln p(\boldsymbol{z}, \theta, \boldsymbol{w} | \alpha, \beta)] = E_{q(\theta)}\left[\ln \bar{\Gamma} + \sum_{i=1}^{k} \left[(\alpha_i - 1) \ln \theta_i + \sum_{n=1}^{N} z_{i,n} \ln (\theta_i \beta) \right] \right]$$

$$E_{q(\theta)}[\ln p(\boldsymbol{z}, \theta, \boldsymbol{w} | \alpha, \beta)] = \ln \bar{\Gamma} + \sum_{i=1}^{k} \left[(\alpha_i - 1)\, E_{q(\theta)}[\ln \theta_i] + \sum_{n=1}^{N} z_{i,n}\, E_{q(\theta)}[\ln \theta_i] + z_{i,n} \ln \beta \right]$$

$$\ln \prod_{n=1}^{N} q^*(z_n) = \ln \bar{\Gamma} + \sum_{i=1}^{k} \left[(\alpha_i - 1)\, E_{q(\theta)}[\ln \theta_i] + \sum_{n=1}^{N} z_{i,n}\, E_{q(\theta)}[\ln \theta_i] + z_{i,n} \ln \beta \right] + \mathcal{C} \tag{4.15}$$

$$\prod_{n=1}^{N} q^*(z_n) = (\bar{\Gamma}\, e^{\mathcal{C}})\, e^{\sum_{i=1}^{k} [(\alpha_i - 1)\, E_{q(\theta)}[\ln \theta_i] + \sum_{n=1}^{N} z_{i,n}\, E_{q(\theta)}[\ln \theta_i] + z_{i,n} \ln \beta]} \tag{4.16}$$

Since this function is in terms of z, we can move everything into the constant term at the front as $\mathcal{K}$, and the form of a categorical distribution emerges.

$$\prod_{n=1}^{N} q^*(z_n) = \mathcal{K} \prod_{n=1}^{N} \left(\beta_{i,n}\, e^{E_{q(\theta)}[\ln \theta_i]} \right)^{z_{i,n}} \tag{4.17}$$

If we let $\phi_{n,i}$ be the parameters of the parameters of this distribution, it is clear from equation above that $\phi_{n,i}$ is proportional to the following.

$$\phi_{n,i} \propto \beta_{i,n}\, e^{E_{q(\theta)}[\ln \theta_i]} \tag{4.18}$$

To find $E_{q(\theta)}[\ln \theta_i]$, we note that $p(\theta)$ was previously discovered as a Dirichlet distribution. It can also be re-written into the form of an exponential distribution.

$$p(\theta|\alpha) = \exp\left\{ \sum_{i=1}^{k} (\alpha_i - 1) \ln \theta_i - \left(-\ln \Gamma\left(\sum_{i=1}^{k} \alpha_i \right) + \sum_{i=1}^{k} \ln \Gamma(\alpha_i) \right) \right\}$$

In this form, the various components can be separated out.

$$\begin{aligned} &\text{Natural parameter}: && (\alpha_i - 1) \\ &\text{Sufficient statistic}: && \ln \theta_i \\ &\text{Log normalizer}: && -\ln \Gamma\left(\sum_{i=1}^{k} \alpha_i \right) + \sum_{i=1}^{k} \ln \Gamma(\alpha_i) \end{aligned}$$

From statistics, we know that the derivative of the log normalizer with respected to the natural parameter is equal to the expectation of the sufficient statistics.

$$\frac{d}{dx}\left\{ -\ln \Gamma\left(\sum_{i=1}^{k} \alpha_i \right) + \sum_{i=1}^{k} \ln \Gamma(\alpha_i) \right\} = E[\ln \theta_i]$$

The first derivative of a gamma function is called the digamma function ψ. Therefore, the expectation of $\ln \theta_i$ is the difference between two digamma functions. It should be noted that the digamma function can be solved using the Taylor approximation.

$$\psi(\alpha_i) - \psi\left(\sum_{i=1}^{k} \alpha_i \right) = E[\ln \theta_i]$$

From the results of the variational inference, the form of the two estimators emerges.

$$q^*(\theta) = \overline{\Gamma(\gamma)} \prod_{i=1}^{k} \theta_i^{\gamma_i - 1} \qquad \text{where}: \quad \gamma = \alpha_i + \sum_{n=1}^{N} \phi_{i,n}$$

$$\prod_{n=1}^{N} q^*(z_n) = \mathcal{K} \prod_{n=1}^{N} (\phi_{n,i})^{z_{i,n}} \qquad \text{where}: \quad \phi_{n,i} \propto \beta_{i,n}\, e^{\psi(\alpha_i) - \psi(\sum_{i=1}^{k} \alpha_i)}$$

To solve for both γ and $\phi_{n,i}$, we can initialize both values and iteratively calculate one variable while using it to calculate for the other variable. For the purpose of completeness, the following algorithm below shows the standard LDA algorithm.

Algorithm 4.1

(1) initialize $\phi^0_{n,i} := 1/k$ for all i and n
(2) initialize $\gamma_i := \alpha_i + N/k$ for all i
(3) repeat
(4) for $n = 1$ to N
(5) for $i = 1$ to k
(6) $\phi^{t+1}_{n,i} = \beta_{i,n}\, e^{\psi(\alpha_i) - \psi(\sum_{i=1}^{k} \alpha_i)}$

(7) normalize $\phi_{n,1}^{t+1}$ to sum to 1
(8) $\gamma^{t+1} = \alpha + \sum_{n=1}^{N} \phi_n^{t+1}$
(9) until convergence

5 A Discussion on the Limitations of LDA

LDA has seen a great deal of success since its inception, and various improvement has been proposed over the years. The limitations of LDA present future possibilities for improvement and ideas for the next generation of algorithms.

One potential area of improvement is to automatically discover the number of topics along with LDA. This is normally denoted as the k value of the Dirichlet topic prior. The approach commonly employed is to simply try a number of values and check for the best results. However, as suggested by Cao in her paper [13], this k value could actually be learned from the data base on the cosine distance between the topics, and the density of each topic cluster. By studying the relationship between these two values against the optimal clustering, her approach iteratively converges towards an optimal k value using an improving density profile.

Another minor, yet reasonable shortcoming for LDA comes from the inaccurate nature of variational inference. Since this approach to calculate the posterior is an approximation based on the assumption of mean field theory, The accuracy of the result heavily depends on the dependency of the approximation distributions. A strongly dependent distribution would render the mean field theory completely inappropriate. To improve upon this inadequacy, copula variational inference have been suggested by Dustin Tran, an David Blei [14]. Since copulas are capable of maintaining the separable product formulation from while capturing the interdependency among the different variables. Incorporating copulas into variational inference is a logical approach to improve accuracy while maintaining the computation feasibility. Other possibilities includes using MCMC methods such as Gibbs sampling to achieve a more accurate result. The limitation with Gibbs sampling has historically been the convergence rate. Since Gibbs sampling uses sampling to estimate the posterior, a large set of samples are required. Fortunately, the paper on Fast Collapsed Gibbs sampling for LDA solved this problem. By introducing a new algorithmic change, the new algorithm is able to significantly reduce the inner loop sampling time. Without compromising the accuracy of the final result, the Fast Collapsed Gibbs sampling is capable of improving the speed by 8 fold.

Besides the limitations previously mentioned, perhaps the most questionable assumption for LDA is its usage of the bag-of-words model. For the purpose of topic modeling, the assumption to neglect the ordering of the words within the model might be arguably sufficient. For computational tractability, the compromise is not egregious despite its obviously inaccuracy. The family of ancestral models prior to LDA such as the unigram model, mixture of unigram, and the pLSI model all made the same assumption. There's been several suggestions on combining concepts of n-gram with LDA. A paper written by Hanna Wallach made this exact suggestion in her paper, "Beyond bag of words [16]."

The purpose of the bag-of-words assumption is to maintain the independence between the words. According to De Finetti's theory, the assumption simplifies the joint pdf as an iid distribution. However, as it was mentioned previously, this assumption could be eased by the utilization of the copula formulation. Since the copula formulation is capable of capturing the dependency between variable while maintaining the simpler multiplication of iid formulation.

[1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *the Journal of machine Learning research* 3 (2003): 993-1022.‘

[2] Hofmann, Thomas. "Probabilistic latent semantic indexing."*Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999.‘

[3] Darling, William M. "A theoretical and practical implementation tutorial on topic modeling and gibbs sampling." *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 2011.‘

[4] Colorado, Reed. *Latent Dirichlet Allocation : Towards a Deeper Understanding.*. Jan. 2012. Web. 5 March 2016.

[5] Bishop, Christopher M. "Pattern Recognition." *Machine Learning* (2006).‘

[6] Wikipedia contributors. "Language model." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 21 Feb. 2016. Web. 5 Mar. 2016.‘

[7] Wikipedia contributors. "Latent Dirichlet allocation." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 17 Feb. 2016. Web. 5 Mar. 2016.‘

[8] Burton, Matt. *The Joy of Topic Modeling*. May. 2013. Web. 5 March 2016.

[9] Oneata, Dan. *A Tutorial on Probabilistic Latent Semantic Analysis.* arXiv:1212.3900 [stat.ML]

[10] Hirschberg, Daniel S. "Algorithms for the longest common subsequence problem." *Journal of the ACM (JACM)* 24.4 (1977): 664-675.‘

[11] Blei, David M. "Introduction to Probabilistic Topic Models."‘

[12] Dehua, Model Selection for Topic Models via Spectral Decomposition. arXiv:1410.6466 [stat.ML]

[13] caoJuan, XiaTian. "Adensity? based method for adaptive LDA model selection."*neurocomputing* 72.7 (2009): r9.‘

[14] Tran12, Dustin, David M. Blei, and Edoardo M. Airoldi. "Copula variational inference."‘

[15] Porteous, Ian, et al. "Fast collapsed gibbs sampling for latent dirichlet allocation." *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008.‘

[16] Wallach, Hanna M. "Topic modeling: beyond bag-of-words." *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006.‘