Gibbs Sampling for LDA with Asymmetric Dirichlet Priors

Andrey Kuzmenko
for.akuz@gmail.com
http://akuz.me

July 1, 2011

Abstract

This note presents a detailed derivation of the collapsed Gibbs sampling formulae for the Latent Dirichlet Allocation (LDA) model. The additional feature of the results presented here is that we allow topics to have asymmetric Dirichet priors over words (referred as "beta" in LDA literature), and documents to have assymetric Dirichlet priors over the topics (referred as "alpha" in the LDA literature). The resulting Gibbs sampling formulae are essentially equivalent to the symmetric case, with the addition of relevant indices for alpha and beta.

1 Introduction

We borrowed the idea of presenting the detailed derivation for a known scientific result from many self-published papers available on the Internet, and in particular from [3] for this note. However, we also incorporate new details such as allowing different topic concentrations and making sure Gibbs sampling can still be used in this case. We believe the availability of the derivation can stimulate the creation of new models for text analysis, and of their solutions.

We do not describe LDA in detail; please refer to paper [1] for an introduction. Gibbs sampling was first suggested for finding the parameters of the LDA model in [2]. In this paper we only provide a detailed derivation of the formulae that are needed for sampling from the marginal conditional distribution $p(Z|W, \alpha, \beta\}$, while allowing different for topics to have different concentrations (instead of treating it as a single common parameter for all topics).

We refer to a vector parameter of a Dirichlet distribution as *concentration* when all of its values are equal to the same value. However, in the derivation below we *still keep the parameter as vector*, even if all of its values are the same. This will be useful for declaring auxiliary functions in a systematic way. The notaton used in this note may slightly differ from the cited works, so please review it carefully below.

2 Notation Setup

The fixed parameters of the LDA model described here are:

- M - the number of documents
- Q - the total number of word occurences in all documents
- T - the number of terms in the dictionary
- K - the number of topics
- $\beta = \{\beta_k : 1 \leq k \leq K\}$ - topic concentrations over terms; each $\beta_k$is a T-dimensional vector (individual $\beta_k$can have *different values*; however the values *within each* of these vectors should be specified to the same value due to the principle of indifference, unless we have some prior information about specific term frequecies in certain topics)
- $\alpha = \{\alpha_m : 1 \leq m \leq M\}$ - document concentrations over topics; each $\alpha_m$is a K-dimensional vector (in general we would preset *all the values within all* of these vectors to the same global value, unless we have some prior information that a particular document should express more topics than any other one, or information that some of the documents should express cirtain topics more than other topics)

The random variables defined within the LDA model are:

- $\Phi = \{\Phi_k : 1 \leq k \leq K\}$- a set of independent T-dimentional vector-valued random variables, each of which is distributed according to the Dirichlet distribution with a corresponding concentration parameter β_k
- $\Theta = \{\Theta_m : 1 \leq m \leq M\}$- a set of independent K-dimentional vector-valued random variables, each of which is distributed according to the Dirichlet distribution with a corresponding concentration parameter α_m
- $Z = \{z_i : 1 \leq i \leq Q\}$ - a set of discrete random variables with a range of $\{1...K\}$ representing the index of the topic, which generates the word in the global position i
- $W = \{w_i : 1 \leq i \leq Q\}$ - a set of discrete random variables with a range of $\{1...T\}$ representing the index of the term generated by the topic in global position i

3 Joint Probability

We will first obtain an explicit formula for the joint probability $p(W, Z|\alpha, \beta)$, which we will later use to derive the formula needed for Gibbs sampling. According to the assumptions of the LDA model, the joint posterior factorizes as follows:

$$p(W, Z|\alpha, \beta) = p(W|Z, \beta)\, p(Z|\alpha) \tag{1}$$

Let's first consider $p(W|Z, \beta)$ from (1) separately. The expression for this factor can be obtained by integrating out Φ from the factor of complete probability function that depends on it:

$$p(W|Z, \beta) = \int p(W|Z, \Phi)\, p(\Phi, \beta)\, d\Phi \tag{2}$$

The first factor under the integral in (2) can be expressed as a product of probabilities of all words occuring given the topics in the positions of these words:

$$p(W|Z, \Phi) = \prod_{i=1}^{Q} p(w_i|z_i) = \prod_{i=1}^{Q} \Phi_{(z_i, w_i)} = \prod_{k=1}^{K} \prod_{t=1}^{T} \Phi_{k,t}^{n_{k,t}} \tag{3}$$

We used the assumption of the LDA model that all words are generated independently. We then reorganized the product over all word occurances into a double product over the topics and terms, with a vector n_k such that its elements, $n_{k,t}$, contain the counts of the number of times term t has occured together with topic k.

The second factor under the integral in (2) can be expressed as a product of probabilities under the assumed independent Dirichlet distributions over the topics-term distributions:

$$p(\Phi|\beta) = \prod_{k=1}^{K} \frac{1}{\Delta(\beta_k)} \prod_{t=1}^{T} \Phi_{k,t}^{\beta_{k,t}-1} \tag{4}$$

where, inspired by [3], we denoted $\Delta(x)$ as the following function on an arbitrary vector x:

$$\Delta(x) = \frac{\prod_{j=1}^{dim(x)} \Gamma(x_j)}{\Gamma(\sum_{j=1}^{dim(x)} x_j)}$$

By substituting (3) and (4) into (2) we get:

$$\begin{aligned}
p(W|Z, \beta) &= \int \prod_{k=1}^{K} \frac{1}{\Delta(\beta_k)} \prod_{t=1}^{T} \Phi_{k,t}^{n_{k,t}+\beta_{k,t}-1} d\Phi \\
&= \int_{dom(\Phi_K)} \dots \int_{dom(\Phi_1)} \prod_{k=1}^{K} \frac{1}{\Delta(\beta_k)} \prod_{t=1}^{T} \Phi_{k,t}^{n_{k,t}+\beta_{k,t}-1} d\Phi_1 \dots d\Phi_K \\
&= \prod_{k=1}^{K} \frac{1}{\Delta(\beta_k)} \int_{dom(\Phi_k)} \prod_{t=1}^{T} \Phi_{k,t}^{n_{k,t}+\beta_{k,t}-1} d\Phi_k
\end{aligned}$$

We now observe that the expression under the inner integral exactly corresponds to the formula of the probability under the Dirichlet distribution with a T-dimensional parameter vector Ψ_k with elements equal to $\Psi_{k,t} = n_{k,t} + \beta_{k,t}$. The domain of integration $dom(\Phi_k)$ exactly corresponds to the range of the random variable distributed under the Dirichlet. Therefore, the integral should evaluate to the normalization constant of the corresponding Dirichlet distribution. Using the notation of $\Delta(x)$ introduced above, this results in:

$$p(W|Z,\beta) \quad = \quad \prod_{k=1}^{K} \frac{\Delta(n_k + \beta_k)}{\Delta(\beta_k)}$$

Using exactly the same approach, the second factor in (1) can be evaluated to:

$$p(Z|\alpha) = \prod_{m=1}^{M} \frac{\Delta(n_m + \alpha_m)}{\Delta(\alpha_m)}$$

where the vector n_m is such that its elements, $n_{m,k}$, contain the counts of the number of times a word has been assigned to topic k within document m.

Substituting the last two results into (1), we obtain:

$$p(W,Z|\alpha,\beta) = \prod_{k=1}^{K} \frac{\Delta(n_k + \beta_k)}{\Delta(\beta_k)} \prod_{m=1}^{M} \frac{\Delta(n_m + \alpha_m)}{\Delta(\alpha_m)} \tag{5}$$

4 Marginal Probability

We can now derive the formula we need for sampling from the marginal conditional distribution $p(Z|W,\alpha,\beta)$. The key idea behind Gibbs sampling is that to get a sample from the joint distribution of variables Z, we sample from each of the separate random variables z_i, conditional on the state of all other variables. The resulting collection of sampled values of z_i constitute a sample that approximates the sample from the joint distribution of all variables in Z. Please refer to other publications about Gibbs sampling, in particular for LDA [2], for more details.

So, to execute the Gibbs sampling procedure described above, we need to derive the formula for $p(z_i|Z_{-i},W,\alpha,\beta)$, where Z_{-i} represents a subset of random variables Z with the variable at the global word position i excluded. A reminder: index i varies over all word occurances withing all the texts: $1 \leq i \leq Q$.

Using the probability chain rule several times, and omiting the parameters α and β, we obtain:

$$
\begin{aligned}
p(z_i|Z_{-i},W) &= \frac{p(W,Z)}{p(W,Z_{-i})} \\
&= \frac{p(W,Z)}{p(W|Z_{-i})\,p(Z_{-i})} \\
&= \frac{p(W,Z)}{p(W_{-i}|Z_{-i})\,p(w_i|Z_{-i})\,p(Z_{-i})} \\
&\propto \frac{p(W,Z)}{p(W_{-i}|Z_{-i})\,p(Z_{-i})} \quad \text{because } p(w_i|Z_{-i}) = p(w_i) = const \\
&\propto \frac{p(W,Z)}{p(W_{-i},Z_{-i})}
\end{aligned}
$$

We will now substitute the formula (5) into the last equation to obtain the expression for $p(z_i|Z_{-i},W)$. Below, the vectors n_k^{-i} and n_m^{-i} denote the same vectors of counts as n_k and n_m, but *exclude* the count for the occurence of term *and* topic as the global position i; these vectors are used in the expression for $p(W_{-i},Z_{-i})$. So, by substituting (5) we obtain:

$$
p(z_i|Z_{-i},W) \quad \propto \quad \prod_{k=1}^{K} \frac{\Delta(n_k+\beta_k)}{\Delta(n_k^{-i}+\beta_k)} \prod_{m=1}^{M} \frac{\Delta(n_m+\alpha_m)}{\Delta(n_m^{-1}+\alpha_m)}
$$

We now observe that, in each of the two product expressions above, only one of the factors will be different from unity. The non-unity factor within the first product expression above corresponds to the index of the topic j, for which we are calculating the probability $p(z_i = j|Z_{-i},W)$ for. The non-unity factor within the *second* product expression above corresponds to the index of the document d, within which the word at the global position i occurs. Therefore, we obtain:

$$
p(z_i = j|Z_{-i},W) \quad \propto \quad \frac{\Delta(n_j+\beta_j)}{\Delta(n_j^{-i}+\beta_j)} \frac{\Delta(n_d+\alpha_d)}{\Delta(n_d^{-1}+\alpha_d)}
$$

We can now substitute the definition of Δ to get:

$$
\begin{aligned}
p(z_i = j|Z_{-i},W) \quad \propto \quad & \frac{\prod_{t=1}^{T}\Gamma(n_{j,t}+\beta_{j,t})}{\prod_{t=1}^{T}\Gamma(n_{j,t}^{-i}+\beta_{j,t})} \frac{\Gamma(\sum_{t=1}^{T}(n_{j,t}^{-i}+\beta_{j,t}))}{\Gamma(\sum_{t=1}^{T}(n_{j,t}+\beta_{j,t}))} \\
& \frac{\prod_{k=1}^{K}\Gamma(n_{d,k}+\alpha_{d,k})}{\prod_{k=1}^{K}\Gamma(n_{d,k}^{-i}+\alpha_{d,k})} \frac{\Gamma(\sum_{k=1}^{K}(n_{d,k}^{-i}+\beta_{d,k}))}{\Gamma(\sum_{k=1}^{K}(n_{d,k}+\beta_{d,k}))}
\end{aligned}
$$

Observe that in the fractions of product expressions, only one of the factors will be different from unity, corresponding to the index w of the term that

actually occurs in global position i in the first case, and corresponding to the index j of the topic we are calculating the probability for in the second case. To obtain the formula below we also express the counts that account for the cooccurance of topic and term at global position i via the counts that do not:

$$p(z_i = j|Z_{-i}, W) \quad \propto \quad \frac{\Gamma(n_{j,w}^{-i} + \beta_{j,w} + 1)}{\Gamma(n_{j,w}^{-i} + \beta_{j,w})} \frac{\Gamma(\sum_{t=1}^{T}(n_{j,t}^{-i} + \beta_{j,t}))}{\Gamma(\sum_{t=1}^{T}(n_{j,t}^{-i} + \beta_{j,t}) + 1)} \\ \frac{\Gamma(n_{d,j}^{-i} + \alpha_{d,j} + 1)}{\Gamma(n_{d,j}^{-i} + \alpha_{d,j})} \frac{\Gamma(\sum_{k=1}^{K}(n_{d,j}^{-i} + \beta_{d,j}))}{\Gamma(\sum_{k=1}^{K}(n_{d,j}^{-i} + \beta_{d,j}) + 1)}$$

We now use the property of the Gamma function $\Gamma(x+1) = x\Gamma(x)$ to obtain:

$$p(z_i = j|Z_{-i}, W) \quad \propto \quad \frac{n_{j,w}^{-i} + \beta_{j,w}}{\sum_{t=1}^{T}(n_{j,t}^{-i} + \beta_{j,t})} \frac{n_{d,j}^{-i} + \alpha_{d,j}}{\sum_{k=1}^{K}(n_{d,j}^{-i} + \beta_{d,j})}$$

And finally note that $\sum_{k=1}^{K}(n_{d,j}^{-i} + \beta_{d,j})$ is a constant, and therefore can be excluded from calculating the (unnormalized) probability:

$$p(z_i = j|Z_{-i}, W) \propto \frac{(n_{j,w}^{-i} + \beta_{j,w})(n_{d,j}^{-i} + \alpha_{d,j})}{\sum_{t=1}^{T}(n_{j,t}^{-i} + \beta_{j,t})} \tag{6}$$

According to the Gibbs sampling approach, formula (6) can now be used for sampling from $p(z_i|Z_{-i}, W, \alpha, \beta)$ in order to approximate the samples from $p(Z|W, \alpha, \beta)$.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proc Natl Acad Sci U S A*, 101 Suppl 1:5228–5235, April 2004.

[3] Gregor Heinrich. Parameter estimation for text analysis. *Technical Note, vsonix GmbH + University of Leipzig*, Germany, version 2.4: August 2008.