
Derivation of Gibbs Sampling Equation for Hierarchical Latent Dirichlet Allocation

Freddy Chong Tat Chua
School of Information Systems
Singapore Management University
80 Stamford Road Singapore 178902
freddy.chua.2009@smu.edu.sg

1 Introduction

We let L denote the maximum depth of the tree. Each document takes a path in the tree where each path has a length of L . Suppose the tree has T number of nodes, there are potentially T possible paths for each document to choose from. There are two kinds of path, paths from root node to a leaf node and paths from root node to an internal node. When the internal node is chosen, we spawn new leaf nodes below.

1.1 Sampling the Path using Nested Chinese Restaurant Process

The nice property about tree is that there is only one path from root node to any other node in the tree. That means, there are T paths in the tree and we pick one out of T paths for the documents. The following shows the equation for performing the sampling,

$$P(c_d = t | w, c_{-d}, z, \eta, \gamma) \propto P(c_d | c_{-d}, \gamma) P(w_d | c, w_{-d}, z, \eta) \quad (1)$$

Two factors influence the probability that a document belongs to a path. The first factor is the number of documents allocated to a path, a document is more likely to belong to popular paths. The second factor is due to the likelihood of seeing the words in the document generated from a path. Equation 1 highlights these two factors.

1.1.1 Nested Chinese Restaurant Process

The $P(c_d | c_{-d}, \gamma)$ is a Nested Chinese Restaurant Process (NCRP). The NCRP is a special case of Dirichlet Process and its validity can be proven by the Kolmogorov Consistency Theorem. The Kolmogorov Consistency Theorem simply proves that clusters can be divided into sub-clusters. Let h_t denote the number of documents that had selected the path t . The NCRP can be described using the following,

$$P(c_d = t | c_{-d}, \gamma) = \frac{h_t}{N - 1 + \gamma} \quad (2)$$

$$P(c_d = new | c_{-d}, \gamma) = \frac{\gamma}{N - 1 + \gamma} \quad (3)$$

where N is the total number of documents in the corpus. Equation 2 shows the likelihood of choosing this node. Equation 3 is the likelihood of creating a new cluster at this level.

Suppose the sampled node is an internal node instead of a leaf node, then it means we spawn new leaf nodes until we reach the maximum depth as defined. Suppose for a given path c_d , the path has topic levels $1, \dots, K$ and there are T number of word topic distributions. Let's use the following,

$$P(w_{d,n} = v | c_d, z_{d,n}, B) = b_{t,k,v} \quad (4)$$

It is important to note here that t denote the path and k selects the level in the path, a tuple (t, k) selects a topic.

$$B_{t,k} = (b_{t,k,1}, \dots, b_{t,k,V}) \quad (5)$$

$$P(B_{t,k}|\eta) = \frac{\Gamma(\sum_{v=1}^V \eta)}{\prod_{v=1}^V \Gamma(\eta)} \prod_{v=1}^V b_{t,k,v}^{\eta-1} \quad (6)$$

Now that the basic distributions and definitions are there, we shall proceed to do some tough stuff.

$$P(w_d|c_d, z_d, \eta) = \int P(w_d, B|c_d, z_d, \eta) dB \quad (7)$$

$$= \int P(w_d|c_d, z_d, B) P(B|\eta) dB \quad (8)$$

$$= \int \left[\prod_{n=1}^N P(w_{d,n}|c_d, z_{d,n}, B) \right] P(B|\eta) dB \quad (9)$$

$$= \int \left(\prod_{t=1}^T \prod_{v=1}^V b_{t,k,v}^{f_{t,k,v}} \right) P(B|\eta) dB \quad (10)$$

$$= \prod_{t=1}^T \left[\frac{\Gamma(V\eta)}{\Gamma(\eta)^V} \frac{\prod_{v=1}^V \Gamma(f_{t,k,v} + \eta)}{\Gamma(V\eta + \sum_{v=1}^V f_{t,k,v})} \right] \quad (11)$$

$$\propto \prod_{t=1}^T \frac{\prod_{v=1}^V \Gamma(f_{t,k,v} + \eta)}{\Gamma(V\eta + \sum_{v=1}^V f_{t,k,v})} \quad (12)$$

Now for even tougher stuff,

$$P(w_d|c_d, w_{-d}, z_d, \eta) = \frac{P(w|c_d, z_d, \eta)}{P(w_{-d}|c_d, z_d, \eta)} \quad (13)$$

$$= \prod_{t=1}^T \left[\frac{\Gamma(V\eta + \sum_{v=1}^V g_{t,k,v})}{\prod_{v=1}^V \Gamma(\eta + g_{t,k,v})} \frac{\prod_{v=1}^V \Gamma(g_{t,k,v} + f_{t,k,v} + \eta)}{\Gamma(V\eta + \sum_{v=1}^V (g_{t,k,v} + f_{t,k,v}))} \right] \quad (14)$$

And expressing in Logarithm form,

$$\begin{aligned} \log P(w_d|c_d, w_{-d}, z_d, \eta) &= \log \left[\prod_{t=1}^T \left[\frac{\Gamma(V\eta + \sum_{v=1}^V g_{t,k,v})}{\prod_{v=1}^V \Gamma(\eta + g_{t,k,v})} \frac{\prod_{v=1}^V \Gamma(g_{t,k,v} + f_{t,k,v} + \eta)}{\Gamma(V\eta + \sum_{v=1}^V (g_{t,k,v} + f_{t,k,v}))} \right] \right] \\ &= \sum_{t=1}^T \left[\log \Gamma(V\eta + \sum_{v=1}^V g_{t,k,v}) - \sum_{v=1}^V \log \Gamma(\eta + g_{t,k,v}) \right. \\ &\quad \left. + \sum_{v=1}^V \log \Gamma(g_{t,k,v} + f_{t,k,v} + \eta) - \log \Gamma(V\eta + \sum_{v=1}^V (g_{t,k,v} + f_{t,k,v})) \right] \end{aligned} \quad (15)$$

When sampling whether to branch off, the equations look like the following, it is pretty similar to the one above except that $g_{t,k,v}$ is always zero.

$$\log P(w_d|c_d, w_{-d}, z_d, \eta) = \log \left[\prod_{t=1}^T \left[\frac{\Gamma(V\eta + \sum_{v=1}^V \eta)}{\prod_{v=1}^V \Gamma(\eta)} \frac{\prod_{v=1}^V \Gamma(f_{t,k,v} + \eta)}{\Gamma(V\eta + \sum_{v=1}^V f_{t,k,v})} \right] \right] \quad (17)$$

$$\begin{aligned} &= \sum_{t=1}^T \left[\log \Gamma(V\eta) - \sum_{v=1}^V \log \Gamma(\eta) \right. \\ &\quad \left. + \sum_{v=1}^V \log \Gamma(f_{t,k,v} + \eta) - \log \Gamma(V\eta + \sum_{v=1}^V f_{t,k,v}) \right] \end{aligned} \quad (18)$$

1.2 Sampling the Topics in the Path using Stick Breaking Construction

Suppose we have D documents and (d, N) words in document d . For each word n in document d , we choose the topic it belongs to as follows,

$$V_i \sim \text{Beta}(m\pi, (1-m)\pi) \quad (19)$$

$$P(z_{d,n} = k | V_1, \dots, V_k) = V_k \prod_{i=1}^{k-1} (1 - V_i) \quad (20)$$

Suppose we let $e_{d,k}$ denote the counts of occurrence for each k in document d . Then suppose we want to derive the posterior distribution of V_1, \dots, V_k

$$V_1 | e_{d,1}, \dots, e_{d,K} \sim \text{Beta} \left(m\pi + e_{d,1}, (1-m)\pi + \sum_{i=2}^K e_{d,i} \right) \quad (21)$$

$$V_2 | e_{d,1}, \dots, e_{d,K} \sim \text{Beta} \left(m\pi + e_{d,2}, (1-m)\pi + \sum_{i=3}^K e_{d,i} \right) \quad (22)$$

$$V_k | e_{d,1}, \dots, e_{d,K} \sim \text{Beta} \left(m\pi + e_{d,k}, (1-m)\pi + \sum_{i=k+1}^K e_{d,i} \right) \quad (23)$$

Hence,

$$P(z_{d,n} = k | z_{d,-n}, m, \pi) = E \left[V_k \prod_{i=1}^{k-1} (1 - V_i) \right] \quad (24)$$

$$= \frac{m\pi + e_{d,k}}{\pi + \sum_{i=k}^K e_{d,i}} \prod_{i=1}^{k-1} \frac{(1-m)\pi + \sum_{j=i+1}^K e_{d,j}}{\pi + \sum_{j=i}^K e_{d,j}} \quad (25)$$

As for the word, it goes as follows, suppose we have V number of words in the vocabulary, let $d_{k,v}$ be the number of times word v is allocated to topic k .

$$P(w_{d,n} = v | z, c, w_{d,-n}) = \frac{d_{k,v} + \eta}{\sum_{v=1}^V d_{k,v} + V\eta} \quad (26)$$

So to sample a topic, the expression is as follows

$$P(z_{d,n} = k | z_{d,-n}, c, w, m, \pi, \eta) = \left[\frac{m\pi + e_{d,k}}{\pi + \sum_{i=k}^K e_{d,i}} \prod_{i=1}^{k-1} \frac{(1-m)\pi + \sum_{j=i+1}^K e_{d,j}}{\pi + \sum_{j=i}^K e_{d,j}} \right] \frac{d_{k,v} + \eta}{\sum_{v=1}^V d_{k,v} + V\eta} \quad (27)$$