Study Notes on the Latent Dirichlet Allocation

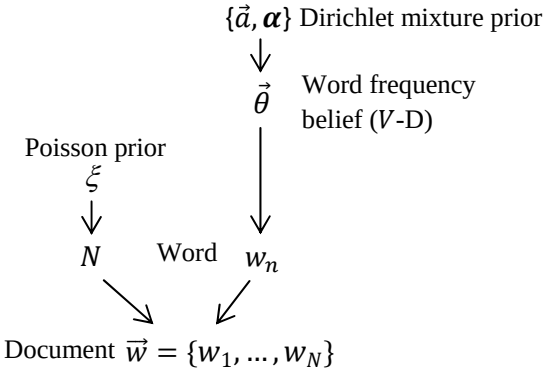
Xugang Ye

1. Model Framework

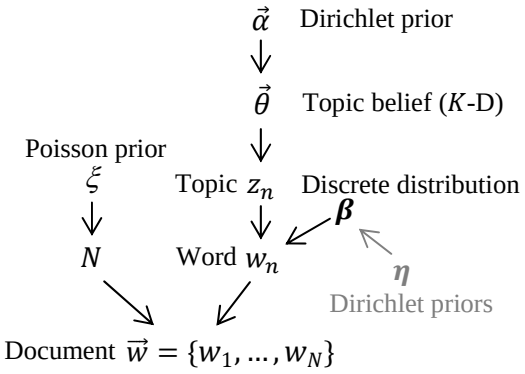
A word w is an element of dictionary $\{1, \dots, V\}$.

A document $\vec{w}$ is represented by a sequence of N words: $\vec{w} = (w_1, \dots, w_N)$, $w_n \in \{1, \dots, V\}$.

A corpus $\boldsymbol{D}$ is a collection of M documents: $\boldsymbol{D} = \{\vec{w}^{(1)}, \dots, \vec{w}^{(M)}\}$.



Graphical representation of two-level model



Graphical representation of LDA

In two-level generative model, the probability of seeing an N-words document $\vec{w}$ is parameterized by hyper parameters $\{\vec{a}, \boldsymbol{\alpha}\}$ of the Dirichlet mixture prior, where $\vec{a}$ contains m mixture parameters and $\boldsymbol{\alpha}$ contains m V-dimensional pseudo-count vectors. The probability can be evaluated via conditioning on the word frequency belief $\vec{\theta}$ as follows

$$
\begin{aligned}
P(\vec{w}|\vec{a}, \boldsymbol{\alpha}) &= \int P(\vec{w}|\vec{\theta}, \vec{a}, \boldsymbol{\alpha}) P(\vec{\theta}|\vec{a}, \boldsymbol{\alpha}) d\vec{\theta} \\
&= \int P(\vec{w}|\vec{\theta}) P(\vec{\theta}|\vec{a}, \boldsymbol{\alpha}) d\vec{\theta} \\
&= \int \left(\prod_{n=1}^{N} P(w_n|\vec{\theta})\right) P(\vec{\theta}|\vec{a}, \boldsymbol{\alpha}) d\vec{\theta} \\
&= \sum_{i=1}^{m} a_i \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i + N)} \prod_{j=1}^{V} \frac{\Gamma(\alpha_{ij} + c_j(\vec{w}))}{\Gamma(\alpha_{ij})}, \qquad (1)
\end{aligned}
$$

which is an explicit formula and where a_i is the i-th mixture parameter, α_{ij} is the pseudo-count of the j-th keyword of the i-th component, $\alpha_i = \sum_{j=1}^{V} \alpha_{ij}$, $c_j(\vec{w})$ is the count of the j-th keyword, and Γ stands for gamma function. The two-level model does not explicitly consider the topic issue. Each Dirichlet component may represent a mix of topics.

Latent Dirichlet allocation (LDA) aggressively considers the topic issue by incorporating the distribution of the word over topic, that is $P(w|z, \boldsymbol{\beta})$, where the matrix $\boldsymbol{\beta} = [\beta_{ij}]$ parameterizes this distribution and β_{ij} stands for the probability of seeing the j-th keyword under the i-th topic. LDA is a three-level generative model in which there is a topic level between the word level and the belief level and in LDA, $\vec{\theta}$ becomes topic belief. LDA assumes that whenever a key word w is observed, there is an associated topic z, hence an N-words document $\vec{w}$ is associated with a topic sequence $\vec{z}$ of length N. By conditioning on $\vec{z}$ and conditional independence, the parameterized probability $P(\vec{w}|\vec{\alpha}, \boldsymbol{\beta})$ of seeing $\vec{w}$ can be written as

$$\begin{aligned} P(\vec{w}|\vec{\alpha}, \boldsymbol{\beta}) &= \sum_{\vec{z}} P(\vec{w}|\vec{z}, \vec{\alpha}, \boldsymbol{\beta}) P(\vec{z}|\vec{\alpha}, \boldsymbol{\beta}) \\ &= \sum_{\vec{z}} P(\vec{w}|\vec{z}, \boldsymbol{\beta}) P(\vec{z}|\vec{\alpha}). \end{aligned} \tag{2}$$

By chain rule and conditional independence, we have

$$\begin{aligned} P(\vec{w}|\vec{z}, \boldsymbol{\beta}) &= P(w_1, \dots, w_N|z_1, \dots, z_N, \boldsymbol{\beta}) \\ &= P(w_1|w_2, \dots, w_N, z_1, \dots, z_N, \boldsymbol{\beta}) P(w_2, \dots, w_N|z_1, \dots, z_N, \boldsymbol{\beta}) \\ &= P(w_1|z_1, \boldsymbol{\beta}) P(w_2, \dots, w_N|z_2, \dots, z_N, \boldsymbol{\beta}) \\ &= \cdots \\ &= P(w_1|z_1, \boldsymbol{\beta}) \cdots P(w_N|z_N, \boldsymbol{\beta}). \end{aligned} \tag{3}$$

Plugging (3) into (2) yields

$$P(\vec{w}|\vec{\alpha}, \boldsymbol{\beta}) = \sum_{\vec{z}} \prod_{n=1}^{N} P(w_n|z_n, \boldsymbol{\beta})\, P(\vec{z}|\vec{\alpha}). \tag{4}$$

By conditioning on $\vec{\theta}$ and conditional independence, we have

$$\begin{aligned} P(\vec{z}|\vec{\alpha}) &= \int P\left(\vec{z}\middle|\vec{\theta}, \vec{\alpha}\right) P(\vec{\theta}|\vec{\alpha}) d\vec{\theta} \\ &= \int P\left(\vec{z}\middle|\vec{\theta}\right) P(\vec{\theta}|\vec{\alpha}) d\vec{\theta} \\ &= \int \prod_{n=1}^{N} P\left(z_n\middle|\vec{\theta}\right) P(\vec{\theta}|\vec{\alpha}) d\vec{\theta}. \end{aligned} \tag{5}$$

Plugging (5) into (4) and exchanging the order of integration and summation yields

$$P(\vec{w}|\vec{\alpha},\boldsymbol{\beta}) = \sum_{\vec{z}} \prod_{n=1}^{N} P(w_n|z_n,\boldsymbol{\beta}) \int \prod_{n=1}^{N} P\left(z_n\middle|\vec{\theta}\right) P(\vec{\theta}|\vec{\alpha}) d\vec{\theta}$$

$$= \sum_{\vec{z}} \int \prod_{n=1}^{N} P(w_n|z_n,\boldsymbol{\beta}) \prod_{n=1}^{N} P\left(z_n\middle|\vec{\theta}\right) P(\vec{\theta}|\vec{\alpha}) d\vec{\theta}$$

$$= \sum_{\vec{z}} \int \prod_{n=1}^{N} P(w_n|z_n,\boldsymbol{\beta}) P\left(z_n\middle|\vec{\theta}\right) P(\vec{\theta}|\vec{\alpha}) d\vec{\theta}$$

$$= \int \left(\sum_{\vec{z}} \prod_{n=1}^{N} P(w_n|z_n,\boldsymbol{\beta}) P\left(z_n\middle|\vec{\theta}\right)\right) P(\vec{\theta}|\vec{\alpha}) d\vec{\theta}. \quad (6)$$

Note that

$$\sum_{\vec{z}} \prod_{n=1}^{N} P(w_n|z_n,\boldsymbol{\beta}) P\left(z_n\middle|\vec{\theta}\right) = \sum_{z_1} P(w_1|z_1,\boldsymbol{\beta}) P\left(z_1\middle|\vec{\theta}\right) \sum_{z_2,\dots,z_n} \prod_{n=2}^{N} P(w_n|z_n,\boldsymbol{\beta}) P\left(z_n\middle|\vec{\theta}\right)$$

$$= \cdots$$

$$= \prod_{n=1}^{N} \sum_{z_n} P(w_n|z_n,\boldsymbol{\beta}) P\left(z_n\middle|\vec{\theta}\right). \quad (7)$$

Plugging (7) into (6) yields the final simplified form

$$P(\vec{w}|\vec{\alpha},\boldsymbol{\beta}) = \int \left(\prod_{n=1}^{N} \sum_{z_n} P(w_n|z_n,\boldsymbol{\beta}) P\left(z_n\middle|\vec{\theta}\right)\right) P(\vec{\theta}|\vec{\alpha}) d\vec{\theta}$$

$$= \int \left(\prod_{n=1}^{N} \sum_{z} P(w_n|z,\boldsymbol{\beta}) P\left(z\middle|\vec{\theta}\right)\right) P(\vec{\theta}|\vec{\alpha}) d\vec{\theta}, \quad (8)$$

which unfortunately does not have explicit formula (the integral cannot be removed because z is latent).

Note that $P(w|\vec{\theta},\boldsymbol{\beta}) = \sum_z P(w|z,\boldsymbol{\beta}) P\left(z\middle|\vec{\theta}\right)$, we have

$$P(\vec{w}|\vec{\alpha},\boldsymbol{\beta}) = \int \left(\prod_{n=1}^{N} P(w_n|\vec{\theta},\boldsymbol{\beta})\right) P(\vec{\theta}|\vec{\alpha}) d\vec{\theta}\ , \quad (9)$$

which can be understood as a form of two level model. But in LDA, there is no direct arc from $\vec{\theta}$ to w_n, it has to go through the intermediate latent topic level and apply the distribution of word over topic. The derivation of the probability (8) shows that each document can exhibit multiple topics and each word can also associate with multiple topics. This flexibility gives LDA strong power to model a large collection of documents, as compared with some old models like unigram and mixture of unigram. Compared with the pLSI, LDA has natural way of assigning probability to a previously unseen document.

To estimate $\vec{\alpha},\boldsymbol{\beta}$ from a corpus $\boldsymbol{D} = \{\vec{w}^{(1)}, \dots, \vec{w}^{(M)}\}$, one needs to maximize the log likelihood of data:

$$\ln P(\boldsymbol{D}|\vec{\alpha},\boldsymbol{\beta}) = \sum_{d=1}^{M} \ln P\left(\vec{w}^{(d)}|\vec{\alpha},\boldsymbol{\beta}\right), \quad (10)$$

which is computationally intractable. However, the variational inference method can provide a computationally tractable lower bound.

Given a set of estimated parameters $\{\vec{\alpha}, \boldsymbol{\beta}\}$, one way to test it is to calculate the perplexity of a test set of T documents $\boldsymbol{D'} = \{\vec{w}^{(1)}, \dots, \vec{w}^{(T)}\}$. The perplexity has the form

$$\text{perplexity } (\boldsymbol{D'}) = \exp\left\{-\frac{\sum_{d=1}^{T} \ln P(\vec{w}^{(d)}|\vec{\alpha},\boldsymbol{\beta})}{\sum_{d=1}^{T} N_d}\right\}, \tag{11}$$

where N_d is the total number of keywords in d-th document. In order to evaluate (11), one has to compute $\ln P\left(\vec{w}^{(d)}|\vec{\alpha},\boldsymbol{\beta}\right)$ for each d. Let $g(\vec{\theta}) = \prod_{n=1}^{N} \sum_z P(w_n|z,\boldsymbol{\beta}) P\left(z|\vec{\theta}\right)$. By (8), we have

$\ln P(\vec{w}|\vec{\alpha},\boldsymbol{\beta}) = \ln \int g(\vec{\theta}) P(\vec{\theta}|\vec{\alpha}) d\vec{\theta} = \ln E(g(\vec{\theta}))$. An estimator is $\ln \frac{1}{L} \prod_{l=1}^{L} g(\vec{\theta}^{(l)})$, where all $\vec{\theta}^{(l)}$ are drawn from the Dirichlet distribution Dir($\vec{\alpha}$).

2. Variational Inference

The key of the variational inference method is to apply the Jensen's inequality to obtain a lower bound of $\ln P(\vec{w}|\vec{\alpha},\boldsymbol{\beta})$. Let $q(\vec{z},\vec{\theta})$ be a joint probability density function of $\vec{z},\vec{\theta}$, applying the idea of importance sampling yields

$$\begin{aligned}
\ln P(\vec{w}|\vec{\alpha},\boldsymbol{\beta}) &= \ln \int \sum_{\vec{z}} P\left(\vec{w},\vec{z},\vec{\theta}|\vec{\alpha},\boldsymbol{\beta}\right) d\vec{\theta} \\
&= \ln \int \sum_{\vec{z}} \frac{P(\vec{w},\vec{z},\vec{\theta}|\vec{\alpha},\boldsymbol{\beta})}{q(\vec{z},\vec{\theta})} q(\vec{z},\vec{\theta})\, d\vec{\theta} \\
&\geq \int \sum_{\vec{z}} q(\vec{z},\vec{\theta}) \ln \frac{P(\vec{w},\vec{z},\vec{\theta}|\vec{\alpha},\boldsymbol{\beta})}{q(\vec{z},\vec{\theta})} d\vec{\theta} = L(\vec{\alpha},\boldsymbol{\beta}),
\end{aligned} \tag{12}$$

where the last step is by Jensen's inequality. Note that

$$\begin{aligned}
L(\vec{\alpha},\boldsymbol{\beta}) &= \int \sum_{\vec{z}} q(\vec{z},\vec{\theta}) \ln \frac{P(\vec{w},\vec{z},\vec{\theta}|\vec{\alpha},\boldsymbol{\beta})}{q(\vec{z},\vec{\theta})} d\vec{\theta} \\
&= \int \sum_{\vec{z}} q(\vec{z},\vec{\theta}) \ln \frac{P(\vec{z},\vec{\theta}|\vec{w},\vec{\alpha},\boldsymbol{\beta}) P(\vec{w}|\vec{\alpha},\boldsymbol{\beta})}{q(\vec{z},\vec{\theta})} d\vec{\theta} \\
&= \int \sum_{\vec{z}} q(\vec{z},\vec{\theta}) \left(\ln \frac{P(\vec{z},\vec{\theta}|\vec{w},\vec{\alpha},\boldsymbol{\beta})}{q(\vec{z},\vec{\theta})} + \ln P(\vec{w}|\vec{\alpha},\boldsymbol{\beta})\right) d\vec{\theta} \\
&= -\int \sum_{\vec{z}} q(\vec{z},\vec{\theta}) \ln \frac{q(\vec{z},\vec{\theta})}{P(\vec{z},\vec{\theta}|\vec{w},\vec{\alpha},\boldsymbol{\beta})} d\vec{\theta} + \int \sum_{\vec{z}} q(\vec{z},\vec{\theta}) \ln P(\vec{w}|\vec{\alpha},\boldsymbol{\beta})\, d\vec{\theta} \\
&= -KL\left(q(\vec{z},\vec{\theta}) || P\left(\vec{z},\vec{\theta}|\vec{w},\vec{\alpha},\boldsymbol{\beta}\right)\right) + \ln P(\vec{w}|\vec{\alpha},\boldsymbol{\beta}).
\end{aligned}$$

Hence,

$$\ln P(\vec{w}|\vec{\alpha},\boldsymbol{\beta}) = L(\vec{\alpha},\boldsymbol{\beta}) + KL\left(q(\vec{z},\vec{\theta})||P(\vec{z},\vec{\theta}|\vec{w},\vec{\alpha},\boldsymbol{\beta})\right), \tag{13}$$

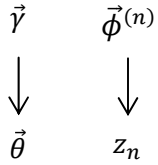
which says $\ln P(\vec{w}|\vec{\alpha},\boldsymbol{\beta})$ is the sum of the variational lower bound and the KL-distance between the variational posterior and the true posterior. In variational method, ones wants to find a $L(\vec{\alpha},\boldsymbol{\beta})$ that is as much close to $\ln P(\vec{w}|\vec{\alpha},\boldsymbol{\beta})$ as possible. And this is equivalent to find a (parameterized) $q(\vec{z},\vec{\theta})$ that has as small KL-distance to $P(\vec{z},\vec{\theta}|\vec{w},\vec{\alpha},\boldsymbol{\beta})$ as possible. Using the fact (factorization by conditional independence) that

$$\begin{aligned} P(\vec{w},\vec{z},\vec{\theta}|\vec{\alpha},\boldsymbol{\beta}) &= P(\vec{w},\vec{z}|\vec{\theta},\vec{\alpha},\boldsymbol{\beta})P(\vec{\theta}|\vec{\alpha},\boldsymbol{\beta}) \\ &= P(\vec{w},\vec{z}|\vec{\theta},\vec{\alpha},\boldsymbol{\beta})P(\vec{\theta}|\vec{\alpha}) \\ &= P(\vec{w}|\vec{z},\vec{\theta},\vec{\alpha},\boldsymbol{\beta})P(\vec{z}|\vec{\theta},\vec{\alpha},\boldsymbol{\beta})P(\vec{\theta}|\vec{\alpha}) \\ &= P(\vec{w}|\vec{z},\boldsymbol{\beta})P(\vec{z}|\vec{\theta})P(\vec{\theta}|\vec{\alpha}), \end{aligned} \tag{14}$$

we have

$$\begin{aligned} L(\vec{\alpha},\boldsymbol{\beta}) &= \int \sum_{\vec{z}} q(\vec{z},\vec{\theta}) \ln \frac{P(\vec{w}|\vec{z},\boldsymbol{\beta})P(\vec{z}|\vec{\theta})P(\vec{\theta}|\vec{\alpha})}{q(\vec{z},\vec{\theta})} d\vec{\theta} \\ &= E_q(\ln P(\vec{w}|\vec{z},\boldsymbol{\beta})) + E_q\left(\ln P(\vec{z}|\vec{\theta})\right) + E_q\left(\ln P(\vec{\theta}|\vec{\alpha})\right) - E_q\left(\ln q(\vec{z},\vec{\theta})\right). \end{aligned} \tag{15}$$

Now, let $q(\vec{z},\vec{\theta})$ be parameterized:



Graphical representation of variational distribution $q(\vec{z},\vec{\theta})$

$$\begin{aligned} q(\vec{z},\vec{\theta}) &= q(\vec{z},\vec{\theta}|\vec{\phi}^{(1)},\dots,\vec{\phi}^{(N)},\vec{\gamma}) \\ &= q(\vec{z}|\vec{\theta},\vec{\phi}^{(1)},\dots,\vec{\phi}^{(N)},\vec{\gamma})q(\vec{\theta}|\vec{\phi}^{(1)},\dots,\vec{\phi}^{(N)},\vec{\gamma}) \\ &= q(\vec{z}|\vec{\phi}^{(1)},\dots,\vec{\phi}^{(N)})q(\vec{\theta}|\vec{\gamma}) \\ &= q(z_1,\dots,z_N|\vec{\phi}^{(1)},\dots,\vec{\phi}^{(N)})q(\vec{\theta}|\vec{\gamma}) \\ &= q(z_1|z_2,\dots,z_N,\vec{\phi}^{(1)},\dots,\vec{\phi}^{(N)})q(z_2,\dots,z_N|\vec{\phi}^{(1)},\dots,\vec{\phi}^{(N)})q(\vec{\theta}|\vec{\gamma}) \end{aligned}$$

$$= q\left(z_1|\vec{\phi}^{(1)}\right)q\left(z_2,\dots,z_N|\vec{\phi}^{(2)},\dots,\vec{\phi}^{(N)}\right)q\left(\vec{\theta}|\vec{\gamma}\right)$$

$$= \cdots$$

$$= q\left(\vec{\theta}|\vec{\gamma}\right)\prod_{n=1}^{N} q\left(z_n|\vec{\phi}^{(n)}\right), \tag{16}$$

where $\vec{\gamma}$ is Dirichlet parameter and $\vec{\phi}^{(n)}$ is multinomial parameter. Applying this variational distribution to (15) yields the expansion of each term as

$$\begin{aligned}
E_q(\ln P(\vec{w}|\vec{z},\boldsymbol{\beta})) &= \int \sum_{\vec{z}} q(\vec{z},\vec{\theta}) \ln P(\vec{w}|\vec{z},\boldsymbol{\beta})\, d\vec{\theta} \\
&= \int \sum_{\vec{z}} \left(q\left(\vec{\theta}|\vec{\gamma}\right)\prod_{n=1}^{N} q\left(z_n|\vec{\phi}^{(n)}\right)\right) \ln P(\vec{w}|\vec{z},\boldsymbol{\beta})\, d\vec{\theta} \\
&= \sum_{\vec{z}} \left(\prod_{n=1}^{N} q\left(z_n|\vec{\phi}^{(n)}\right)\right) \ln P(\vec{w}|\vec{z},\boldsymbol{\beta}) \int q\left(\vec{\theta}|\vec{\gamma}\right) d\vec{\theta} \\
&= \sum_{\vec{z}} \left(\prod_{n=1}^{N} q\left(z_n|\vec{\phi}^{(n)}\right)\right) \ln P(\vec{w}|\vec{z},\boldsymbol{\beta}) \\
&= \sum_{\vec{z}} \left(\prod_{n=1}^{N} q\left(z_n|\vec{\phi}^{(n)}\right)\right) (\ln \prod_{n=1}^{N} P(w_n|z_n,\boldsymbol{\beta})) \\
&= \sum_{\vec{z}} \left(\prod_{n=1}^{N} q\left(z_n|\vec{\phi}^{(n)}\right)\right) (\sum_{n=1}^{N} \ln P(w_n|z_n,\boldsymbol{\beta})) \\
&= \sum_{\vec{z}} \left(\prod_{n=1}^{N} q\left(z_n|\vec{\phi}^{(n)}\right)\right) \left(\sum_{n=1}^{N} \ln \beta_{z_n,w_n}\right) \\
&= \sum_{\vec{z}} \left(\prod_{n=1}^{N} q\left(z_n|\vec{\phi}^{(n)}\right)\right) \left(\ln \beta_{z_1,w_1} + \cdots + \ln \beta_{z_N,w_N}\right) \\
&= \sum_{\vec{z}} \left(\ln \beta_{z_1,w_1}\right) \prod_{n=1}^{N} q\left(z_n|\vec{\phi}^{(n)}\right) + \cdots + \sum_{\vec{z}} \left(\ln \beta_{z_N,w_N}\right) \prod_{n=1}^{N} q\left(z_n|\vec{\phi}^{(n)}\right) \\
&= \sum_{n=1}^{N} \left(\sum_{z_n} q\left(z_n|\vec{\phi}^{(n)}\right) \ln \beta_{z_n,w_n}\right) \sum_{\vec{z}\setminus z_n} \prod_{\substack{t=1 \\ t\neq n}}^{N} q\left(z_t|\vec{\phi}^{(t)}\right) \\
&= \sum_{n=1}^{N} \sum_{z_n} q\left(z_n|\vec{\phi}^{(n)}\right) \ln \beta_{z_n,w_n} \\
&= \sum_{n=1}^{N} \sum_{z_n} \phi_{z_n}^{(n)} \ln \beta_{z_n,w_n} \\
&= \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_i^{(n)} \ln \beta_{i,w_n}, \qquad (17)
\end{aligned}$$

$$\begin{aligned}
E_q\left(\ln P\left(\vec{z}|\vec{\theta}\right)\right) &= \int \sum_{\vec{z}} q(\vec{z},\vec{\theta}) \ln P\left(\vec{z}|\vec{\theta}\right) d\vec{\theta} \\
&= \int \sum_{\vec{z}} \left(q\left(\vec{\theta}|\vec{\gamma}\right)\prod_{n=1}^{N} q\left(z_n|\vec{\phi}^{(n)}\right)\right) \ln P\left(\vec{z}|\vec{\theta}\right) d\vec{\theta} \\
&= \int q\left(\vec{\theta}|\vec{\gamma}\right) d\vec{\theta} \sum_{\vec{z}} \left(\prod_{n=1}^{N} q\left(z_n|\vec{\phi}^{(n)}\right)\right) \ln P\left(\vec{z}|\vec{\theta}\right) \\
&= \int q\left(\vec{\theta}|\vec{\gamma}\right) d\vec{\theta} \sum_{\vec{z}} \left(\prod_{n=1}^{N} q\left(z_n|\vec{\phi}^{(n)}\right)\right) \left(\ln \prod_{n=1}^{N} P\left(z_n|\vec{\theta}\right)\right) \\
&= \int q\left(\vec{\theta}|\vec{\gamma}\right) d\vec{\theta} \sum_{\vec{z}} \left(\prod_{n=1}^{N} q\left(z_n|\vec{\phi}^{(n)}\right)\right) \left(\sum_{n=1}^{N} \ln P\left(z_n|\vec{\theta}\right)\right)
\end{aligned}$$

$$= \int q(\vec{\theta}|\vec{\gamma}) d\vec{\theta} \sum_{\vec{z}} \left(\prod_{n=1}^{N} q(z_n|\vec{\phi}^{(n)})\right) \left(\ln P(z_1|\vec{\theta}) + \cdots + \ln P(z_N|\vec{\theta})\right)$$

$$= \int q(\vec{\theta}|\vec{\gamma}) d\vec{\theta} \sum_{n=1}^{N} \sum_{\vec{z}} \left(\ln P(z_n|\vec{\theta})\right) \prod_{t=1}^{N} q(z_t|\vec{\phi}^{(t)})$$

$$= \int q(\vec{\theta}|\vec{\gamma}) d\vec{\theta} \sum_{n=1}^{N} \sum_{z_n} \left(\ln P(z_n|\vec{\theta})\right) q(z_n|\vec{\phi}^{(n)}) \sum_{\vec{z}\setminus z_n} \prod_{\substack{t=1 \\ t\neq n}}^{N} q(z_t|\vec{\phi}^{(t)})$$

$$= \int q(\vec{\theta}|\vec{\gamma}) d\vec{\theta} \sum_{n=1}^{N} \sum_{z_n} \left(\ln P(z_n|\vec{\theta})\right) q(z_n|\vec{\phi}^{(n)})$$

$$= \int q(\vec{\theta}|\vec{\gamma}) d\vec{\theta} \sum_{n=1}^{N} \sum_{z_n} \left(\ln \theta_{z_n}\right) \phi_{z_n}^{(n)}$$

$$= \int q(\vec{\theta}|\vec{\gamma}) d\vec{\theta} \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_i^{(n)} \ln \theta_i$$

$$= \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_i^{(n)} \int q(\vec{\theta}|\vec{\gamma}) d\vec{\theta} \ln \theta_i \,.$$

$$= \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_i^{(n)} \left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{K} \gamma_j\right)\right). \tag{18}$$

$$E_q\left(\ln P(\vec{\theta}|\vec{\alpha})\right) = \int \sum_{\vec{z}} q(\vec{z}, \vec{\theta}) \ln P(\vec{\theta}|\vec{\alpha}) \, d\vec{\theta}$$

$$= \int \sum_{\vec{z}} \left(q(\vec{\theta}|\vec{\gamma}) \prod_{n=1}^{N} q(z_n|\vec{\phi}^{(n)})\right) \ln P(\vec{\theta}|\vec{\alpha}) \, d\vec{\theta}$$

$$= \int q(\vec{\theta}|\vec{\gamma}) \ln P(\vec{\theta}|\vec{\alpha}) \, d\vec{\theta} \sum_{\vec{z}} \left(\prod_{n=1}^{N} q(z_n|\vec{\phi}^{(n)})\right)$$

$$= \int q(\vec{\theta}|\vec{\gamma}) \ln P(\vec{\theta}|\vec{\alpha}) \, d\vec{\theta}$$

$$= \int q(\vec{\theta}|\vec{\gamma}) (\ln \Gamma(\sum_{i=1}^{K} \alpha_i) - \sum_{i=1}^{K} \ln \Gamma(\alpha_i) + \sum_{i=1}^{K} (\alpha_i - 1) \ln \theta_i) d\vec{\theta}$$

$$= \ln \Gamma(\sum_{i=1}^{K} \alpha_i) - \sum_{i=1}^{K} \ln \Gamma(\alpha_i) + \sum_{i=1}^{K} (\alpha_i - 1) \int q(\vec{\theta}|\vec{\gamma}) \ln \theta_i \, d\vec{\theta}$$

$$= \ln \Gamma(\sum_{i=1}^{K} \alpha_i) - \sum_{i=1}^{K} \ln \Gamma(\alpha_i) + \sum_{i=1}^{K} (\alpha_i - 1) \left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{K} \gamma_j\right)\right), \tag{19}$$

$$E_q\left(\ln q(\vec{z}, \vec{\theta})\right) = E_q\left(\ln\left(q(\vec{\theta}|\vec{\gamma}) \prod_{n=1}^{N} q(z_n|\vec{\phi}^{(n)})\right)\right)$$

$$= E_q\left(\ln q(\vec{\theta}|\vec{\gamma})\right) + \sum_{n=1}^{N} E_q\left(\ln q(z_n|\vec{\phi}^{(n)})\right),$$

$$E_q\left(\ln q(\vec{\theta}|\vec{\gamma})\right) = \ln \Gamma(\sum_{i=1}^{K} \gamma_i) - \sum_{i=1}^{K} \ln \Gamma(\gamma_i) + \sum_{i=1}^{K} (\gamma_i - 1) \left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{K} \gamma_j\right)\right), \tag{20}$$

$$\sum_{n=1}^{N} E_q\left(\ln q(z_n|\vec{\phi}^{(n)})\right) = \sum_{n=1}^{N} \int \sum_{\vec{z}} q(\vec{z}, \vec{\theta}) \ln q(z_n|\vec{\phi}^{(n)}) \, d\vec{\theta}$$

$$= \sum_{n=1}^{N} \int \sum_{\vec{z}} \left(q(\vec{\theta}|\vec{\gamma}) \prod_{t=1}^{N} q(z_t|\vec{\phi}^{(t)})\right) \ln q(z_n|\vec{\phi}^{(n)}) \, d\vec{\theta}$$

$$= \sum_{n=1}^{N} \sum_{\vec{z}} \left(\prod_{t=1}^{N} q(z_t|\vec{\phi}^{(t)})\right) \ln q(z_n|\vec{\phi}^{(n)}) \int q(\vec{\theta}|\vec{\gamma}) d\vec{\theta}$$

$$= \sum_{n=1}^{N} \sum_{\vec{z}} \left(\prod_{t=1}^{N} q(z_t|\vec{\phi}^{(t)})\right) \ln q(z_n|\vec{\phi}^{(n)})$$

$$= \sum_{n=1}^{N} \sum_{z_n} q\left(z_n|\vec{\phi}^{(n)}\right) \ln q\left(z_n|\vec{\phi}^{(n)}\right) \sum_{\vec{z}\setminus z_n} \left(\prod_{\substack{t=1 \\ t\neq n}}^{N} q\left(z_t|\vec{\phi}^{(t)}\right)\right)$$

$$= \sum_{n=1}^{N} \sum_{z_n} q\left(z_n|\vec{\phi}^{(n)}\right) \ln q\left(z_n|\vec{\phi}^{(n)}\right)$$

$$= \sum_{n=1}^{N} \sum_{z_n} \phi_{z_n}^{(n)} \ln \phi_{z_n}^{(n)}$$

$$= \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_i^{(n)} \ln \phi_i^{(n)}, \tag{21}$$

where Ψ is digamma function. Therefore,

$$L(\vec{\alpha}, \boldsymbol{\beta}) = L\left(\vec{w}; \vec{\phi}^{(1)}, \dots, \vec{\phi}^{(N)}, \vec{\gamma}; \vec{\alpha}, \boldsymbol{\beta}\right)$$

$$\begin{aligned}
&= \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_i^{(n)} \ln \beta_{i,w_n} \\
&\quad + \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_i^{(n)} \left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{K} \gamma_j\right)\right) \\
&\quad + \ln \Gamma\left(\sum_{i=1}^{K} \alpha_i\right) - \sum_{i=1}^{K} \ln \Gamma(\alpha_i) + \sum_{i=1}^{K} (\alpha_i - 1) \left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{K} \gamma_j\right)\right) \\
&\quad - \ln \Gamma\left(\sum_{i=1}^{K} \gamma_i\right) + \sum_{i=1}^{K} \ln \Gamma(\gamma_i) - \sum_{i=1}^{K} (\gamma_i - 1) \left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{K} \gamma_j\right)\right) \\
&\quad - \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_i^{(n)} \ln \phi_i^{(n)}.
\end{aligned} \tag{22}$$

3 EM Algorithm

The purpose of the EM algorithm is to maximize (10) by maximizing its variational lower bound. Since $\ln P\left(\vec{w}^{(d)}|\vec{\alpha}, \boldsymbol{\beta}\right) \geq L\left(\vec{w}^{(d)}; \vec{\phi}^{(1,d)}, \dots, \vec{\phi}^{(N_d,d)}, \vec{\gamma}^{(d)}; \vec{\alpha}, \boldsymbol{\beta}\right)$, we have

$$\begin{aligned}
\ln P(\boldsymbol{D}|\vec{\alpha}, \boldsymbol{\beta}) &= \sum_{d=1}^{M} \ln P\left(\vec{w}^{(d)}|\vec{\alpha}, \boldsymbol{\beta}\right) \\
&\geq \sum_{d=1}^{M} L\left(\vec{w}^{(d)}; \vec{\phi}^{(1,d)}, \dots, \vec{\phi}^{(N_d,d)}, \vec{\gamma}^{(d)}; \vec{\alpha}, \boldsymbol{\beta}\right) \\
&= \boldsymbol{L}(\boldsymbol{\phi}, \boldsymbol{\gamma}; \vec{\alpha}, \boldsymbol{\beta}),
\end{aligned} \tag{23}$$

which is overall variational lower bound and where $\boldsymbol{\phi} = \left[\vec{\phi}^{(n,d)}\right]$ is a tensor and $\boldsymbol{\gamma} = [\vec{\gamma}^{(d)}]$ is a matrix.

The idea of the EM algorithm is to start from an initial $(\vec{\alpha}, \boldsymbol{\beta})$, and iteratively improve the estimate via the following alternating E-step and M-step:

E-step: Given $(\vec{\alpha}, \boldsymbol{\beta})$, find $(\boldsymbol{\phi}, \boldsymbol{\gamma})$ to maximize $\boldsymbol{L}(\boldsymbol{\phi}, \boldsymbol{\gamma}; \vec{\alpha}, \boldsymbol{\beta})$,

M-step: Given $(\boldsymbol{\phi}, \boldsymbol{\gamma})$ found from the E-step, find $(\vec{\alpha}, \boldsymbol{\beta})$ to maximize $\boldsymbol{L}(\boldsymbol{\phi}, \boldsymbol{\gamma}; \vec{\alpha}, \boldsymbol{\beta})$.

The two steps are repeated until the value of $\boldsymbol{L}(\boldsymbol{\phi}, \boldsymbol{\gamma}; \vec{\alpha}, \boldsymbol{\beta})$ converges. The update equations can be derived (using Lagrange multipliers) from the overall variational lower bound

$$\begin{aligned}
&\boldsymbol{L}(\boldsymbol{\phi}, \boldsymbol{\gamma}; \vec{\alpha}, \boldsymbol{\beta}) \\
&= \sum_{d=1}^{M} L\left(\vec{w}^{(d)}; \vec{\phi}^{(1,d)}, \dots, \vec{\phi}^{(N_d,d)}, \vec{\gamma}^{(d)}; \vec{\alpha}, \boldsymbol{\beta}\right) \\
&= \sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{i=1}^{K} \phi_i^{(n,d)} \ln \beta_{i,w_n^{(d)}} \\
&\quad + \sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{i=1}^{K} \phi_i^{(n,d)} \left(\Psi\left(\gamma_i{}^{(d)}\right) - \Psi\left(\sum_{j=1}^{K} \gamma_j{}^{(d)}\right)\right) \\
&\quad + \sum_{d=1}^{M} \left(\ln \Gamma\left(\sum_{i=1}^{K} \alpha_i\right) - \sum_{i=1}^{K} \ln \Gamma(\alpha_i) + \sum_{i=1}^{K} (\alpha_i - 1) \left(\Psi\left(\gamma_i{}^{(d)}\right) - \Psi\left(\sum_{j=1}^{K} \gamma_j{}^{(d)}\right)\right)\right) \\
&\quad - \sum_{d=1}^{M} \left(\ln \Gamma\left(\sum_{i=1}^{K} \gamma_i{}^{(d)}\right) - \sum_{i=1}^{K} \ln \Gamma\left(\gamma_i{}^{(d)}\right) + \sum_{i=1}^{K} \left(\gamma_i{}^{(d)} - 1\right) \left(\Psi\left(\gamma_i{}^{(d)}\right) - \Psi\left(\sum_{j=1}^{K} \gamma_j{}^{(d)}\right)\right)\right) \\
&\quad - \sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{i=1}^{K} \phi_i^{(n,d)} \ln \phi_i^{(n,d)} \qquad (24)
\end{aligned}$$

and the constraints $\sum_{j=1}^{V} \beta_{ij} = 1, i = 1, \dots, K; \sum_{i=1}^{K} \phi_i^{(n,d)} = 1, n = 1, \dots, N_d, d = 1, .., M.$

In E-step, the update equations for $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$ are

$$\phi_i^{(n,d)} \propto \beta_{i,w_n^{(d)}} \exp\left(\Psi\left(\gamma_i{}^{(d)}\right) - \Psi\left(\sum_{j=1}^{K} \gamma_j{}^{(d)}\right)\right), \qquad (25)$$

$$\gamma_i{}^{(d)} = \alpha_i + \sum_{n=1}^{N_d} \phi_i^{(n,d)} . \qquad (26)$$

And in order for the variational bound converges, it requires alternating between these two equations.

Once $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$ are obtained in E-step, the update equation for $\boldsymbol{\beta}$ can be determined by using Lagrange multipliers, it's

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_i^{(n,d)} I(w_n^{(d)} = j), \qquad (27)$$

where I is indicator function.

Finally, there is no analytical form of the update equation for $\vec{\alpha}$ in the M-step, the Newton's method can be employed for obtaining updated $\vec{\alpha}$ by maximizing

$$\boldsymbol{L}_{\vec{\alpha}} = \sum_{d=1}^{M} \left(\ln \Gamma\left(\sum_{i=1}^{K} \alpha_i\right) - \sum_{i=1}^{K} \ln \Gamma(\alpha_i) + \sum_{i=1}^{K} (\alpha_i - 1) \left(\Psi\left(\gamma_i{}^{(d)}\right) - \Psi\left(\sum_{j=1}^{K} \gamma_j{}^{(d)}\right)\right)\right). \qquad (28)$$

Once the EM algorithm converges, it can return two target distributions: $\boldsymbol{\beta}$, which is word distribution over topic and $\vec{\alpha}/(\sum_{i=1}^{K} \alpha_i)$, which is topic distribution.

4 Gibbs Sampling

Given the data $\boldsymbol{D} = \{\vec{w}^{(1)}, \dots, \vec{w}^{(M)}\}$, let $\boldsymbol{Z} = \{\vec{z}^{(1)}, \dots, \vec{z}^{(M)}\}$ be the corresponding latent variables such that $z_n^{(d)}$ is associated with $w_n^{(d)}$. The central quantity of the Gibbs sampling methods for learning a LDA model is the full conditional posterior distribution $P(z_n^{(d)}|\boldsymbol{Z}\backslash z_n^{(d)}, \boldsymbol{D})$. Note that

$$P\left(z_n^{(d)}\middle|\boldsymbol{Z}\backslash z_n^{(d)}, \boldsymbol{D}\right) = \frac{P(\boldsymbol{Z},\boldsymbol{D})}{P(\boldsymbol{Z}\backslash z_n^{(d)},\boldsymbol{D})}. \tag{29}$$

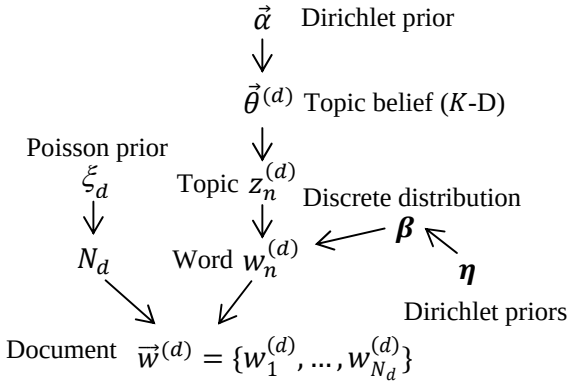
And most Gibbs sampling methods put a prior $\boldsymbol{\eta}$ on $\boldsymbol{\beta}$ so that (29) is parameterized as

$$P\left(z_n^{(d)}\middle|\boldsymbol{Z}\backslash z_n^{(d)}, \boldsymbol{D}, \vec{\alpha}, \boldsymbol{\eta}\right) = \frac{P(\boldsymbol{Z},\boldsymbol{D}|\vec{\alpha},\boldsymbol{\eta})}{P(\boldsymbol{Z}\backslash z_n^{(d)},\boldsymbol{D}|\vec{\alpha},\boldsymbol{\eta})}. \tag{30}$$

The nominator of right hand side of (30) is

$$P(\boldsymbol{Z}, \boldsymbol{D}|\vec{\alpha}, \boldsymbol{\eta}) = P(\boldsymbol{D}|\boldsymbol{Z}, \vec{\alpha}, \boldsymbol{\eta})P(\boldsymbol{Z}|\vec{\alpha}, \boldsymbol{\eta}) = P(\boldsymbol{D}|\boldsymbol{Z}, \boldsymbol{\eta})P(\boldsymbol{Z}|\vec{\alpha}). \tag{31}$$

To evaluate (31), one may refer to the conditional independence represented by the following graph:



Graphical representation of LDA for corpus

By the conditional independence for $\boldsymbol{D}$ and $\boldsymbol{Z}$, we have

$$\begin{aligned} P(\boldsymbol{Z}|\vec{\alpha}) &= \int P(\boldsymbol{Z}|\boldsymbol{\theta}, \vec{\alpha})P(\boldsymbol{\theta}|\vec{\alpha})d\boldsymbol{\theta} \quad (\boldsymbol{\theta} = \{\vec{\theta}^{(1)}, \dots, \vec{\theta}^{(M)}\}) \\ &= \int P(\boldsymbol{Z}|\boldsymbol{\theta})P(\boldsymbol{\theta}|\vec{\alpha})d\boldsymbol{\theta} \\ &= \int \left(\prod_{d=1}^{M}\prod_{n=1}^{N_d} P(z_n^{(d)}|\vec{\theta}^{(d)})\right)\left(\prod_{d=1}^{M} P(\vec{\theta}^{(d)}|\vec{\alpha})\right)d\vec{\theta}^{(1)}\cdots d\vec{\theta}^{(M)} \end{aligned}$$

$$= \prod_{d=1}^{M} \int \left(\prod_{n=1}^{N_d} P(z_n^{(d)} | \vec{\theta}^{(d)}) \right) P(\vec{\theta}^{(d)} | \vec{\alpha}) d\vec{\theta}^{(d)}$$

$$= \prod_{d=1}^{M} \int \left(\prod_{n=1}^{N_d} \theta_{z_n^{(d)}}^{(d)} \right) P(\vec{\theta}^{(d)} | \vec{\alpha}) d\vec{\theta}^{(d)}$$

$$= \prod_{d=1}^{M} \int \left(\prod_{i=1}^{K} (\theta_i^{(d)})^{c_i^{(d)}} \right) P(\vec{\theta}^{(d)} | \vec{\alpha}) d\vec{\theta}^{(d)}$$

$$= \prod_{d=1}^{M} \left(\frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i + \sum_{i=1}^{K} c_i^{(d)}\right)} \prod_{i=1}^{K} \frac{\Gamma\left(\alpha_i + c_i^{(d)}\right)}{\Gamma(\alpha_i)} \right)$$

$$= \prod_{d=1}^{M} \frac{\mathrm{B}(\vec{\alpha} + \vec{c}^{(d)})}{\mathrm{B}(\vec{\alpha})}, \tag{32}$$

where $c_i^{(d)}$ is count of topic i in in document d, and B is Beta function,

$$P(\boldsymbol{D} | \boldsymbol{Z}, \boldsymbol{\eta})$$

$$= \int P(\boldsymbol{D} | \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\eta}) P(\boldsymbol{\beta} | \boldsymbol{Z}, \boldsymbol{\eta}) d\boldsymbol{\beta}$$

$$= \int P(\boldsymbol{D} | \boldsymbol{Z}, \boldsymbol{\beta}) P(\boldsymbol{\beta} | \boldsymbol{\eta}) d\boldsymbol{\beta}$$

$$= \int \left(\prod_{d=1}^{M} \prod_{n=1}^{N_d} P(w_n^{(d)} | z_n^{(d)}, \boldsymbol{\beta}) \right) P(\boldsymbol{\beta} | \boldsymbol{\eta}) d\boldsymbol{\beta}$$

$$= \int \left(\prod_{d=1}^{M} \prod_{n=1}^{N_d} \beta_{z_n^{(d)}, w_n^{(d)}} \right) P(\boldsymbol{\beta} | \boldsymbol{\eta}) d\boldsymbol{\beta}$$

$$= \int \left(\prod_{i=1}^{K} \prod_{j=1}^{V} \beta_{ij}^{c_{ij}} \right) P(\boldsymbol{\beta} | \boldsymbol{\eta}) d\boldsymbol{\beta}$$

$$= \int \left(\prod_{i=1}^{K} \prod_{j=1}^{V} \beta_{ij}^{c_{ij}} \right) \left(\prod_{i=1}^{K} P(\vec{\beta}^{(i)} | \vec{\eta}^{(i)}) \right) d\vec{\beta}^{(1)} \cdots d\vec{\beta}^{(K)}$$

$$= \prod_{i=1}^{K} \int \left(\prod_{j=1}^{V} \beta_{ij}^{c_{ij}} \right) P(\vec{\beta}^{(i)} | \vec{\eta}^{(i)}) d\vec{\beta}^{(i)}$$

$$= \prod_{i=1}^{K} \left(\frac{\Gamma\left(\sum_{j=1}^{V} \eta_j^{(i)}\right)}{\Gamma\left(\sum_{j=1}^{V} \eta_j^{(i)} + \sum_{j=1}^{V} c_{ij}\right)} \prod_{j=1}^{V} \frac{\Gamma\left(\eta_j^{(i)} + c_{ij}\right)}{\Gamma\left(\eta_j^{(i)}\right)} \right)$$

$$= \prod_{i=1}^{K} \frac{\mathrm{B}(\vec{\eta}^{(i)} + \vec{c}_i)}{\mathrm{B}(\vec{\eta}^{(i)})}, \tag{33}$$

where c_{ij} is count of topic i to word j in all documents.

Plugging (32) and (33) into (31) yields

$$P(\boldsymbol{Z}, \boldsymbol{D} | \vec{\alpha}, \boldsymbol{\eta}) = \left(\prod_{i=1}^{K} \frac{\mathrm{B}(\vec{\eta}^{(i)} + \vec{c}_i)}{\mathrm{B}(\vec{\eta}^{(i)})} \right) \cdot \left(\prod_{d=1}^{M} \frac{\mathrm{B}(\vec{\alpha} + \vec{c}^{(d)})}{\mathrm{B}(\vec{\alpha})} \right). \tag{34}$$

For the denominator of the right hand side of (30), note that

$$
\begin{aligned}
P\left(\boldsymbol{Z}\backslash z_n^{(d)}, \boldsymbol{D}\middle|\vec{\alpha},\boldsymbol{\eta}\right) &= P\left(w_n^{(d)}, \boldsymbol{Z}\backslash z_n^{(d)}, \boldsymbol{D}\backslash w_n^{(d)}\middle|\vec{\alpha},\boldsymbol{\eta}\right) \\
&= P\left(w_n^{(d)}\middle|\boldsymbol{Z}\backslash z_n^{(d)}, \boldsymbol{D}\backslash w_n^{(d)}, \vec{\alpha},\boldsymbol{\eta}\right) P\left(\boldsymbol{Z}\backslash z_n^{(d)}, \boldsymbol{D}\backslash w_n^{(d)}\middle|\vec{\alpha},\boldsymbol{\eta}\right) \\
&= \left(\sum_{z_n^{(d)}=1}^{K} P\left(w_n^{(d)}\middle|z_n^{(d)},\boldsymbol{\eta}\right) P\left(z_n^{(d)}\middle|\vec{\alpha}\right)\right) P\left(\boldsymbol{Z}\backslash z_n^{(d)}, \boldsymbol{D}\backslash w_n^{(d)}\middle|\vec{\alpha},\boldsymbol{\eta}\right) \\
&\propto P\left(\boldsymbol{Z}\backslash z_n^{(d)}, \boldsymbol{D}\backslash w_n^{(d)}\middle|\vec{\alpha},\boldsymbol{\eta}\right) \\
&= \left(\prod_{i=1}^{K} \frac{\mathrm{B}\left(\vec{\eta}^{(i)}+\vec{c}_i'\right)}{\mathrm{B}\left(\vec{\eta}^{(i)}\right)}\right) \cdot \left(\prod_{m=1}^{M} \frac{\mathrm{B}\left(\vec{\alpha}+\vec{c}^{(m)\prime}\right)}{\mathrm{B}\left(\vec{\alpha}\right)}\right), \qquad (35)
\end{aligned}
$$

where $\vec{c}_i'$ and $\vec{c}^{(m)\prime}$ respectively correspond to $\vec{c}_i$ and $\vec{c}^{(m)}$,with count associated with $z_n^{(d)}$ excluded. Let $k = z_n^{(d)}$and $v = w_n^{(d)}$, then (30) is reduced to

$$
\begin{aligned}
&P\left(z_n^{(d)} = k\middle|\boldsymbol{Z}\backslash z_n^{(d)}, w_n^{(d)} = v, \boldsymbol{D}\backslash w_n^{(d)}, \vec{\alpha},\boldsymbol{\eta}\right) \\
&\propto \frac{\mathrm{B}\left(\vec{\eta}^{(k)}+\vec{c}_k\right)}{\mathrm{B}\left(\vec{\eta}^{(k)}+\vec{c}_k'\right)} \cdot \frac{\mathrm{B}\left(\vec{\alpha}+\vec{c}^{(d)}\right)}{\mathrm{B}\left(\vec{\alpha}+\vec{c}^{(d)\prime}\right)} \\
&= \frac{\frac{\prod_{j=1}^{V}\Gamma\left(\eta_j^{(k)}+c_{kj}\right)}{\Gamma\left(\sum_{j=1}^{V}\eta_j^{(k)}+\sum_{j=1}^{V}c_{kj}\right)}}{\frac{\prod_{j=1}^{V}\Gamma\left(\eta_j^{(k)}+c_{kj}'\right)}{\Gamma\left(\sum_{j=1}^{V}\eta_j^{(k)}+\sum_{j=1}^{V}c_{kj}'\right)}} \cdot \frac{\frac{\prod_{i=1}^{K}\Gamma\left(\alpha_i+c_i^{(d)}\right)}{\Gamma\left(\sum_{i=1}^{K}\alpha_i+\sum_{i=1}^{K}c_i^{(d)}\right)}}{\frac{\prod_{i=1}^{K}\Gamma\left(\alpha_i+c_i^{(d)\prime}\right)}{\Gamma\left(\sum_{i=1}^{K}\alpha_i+\sum_{i=1}^{K}c_i^{(d)\prime}\right)}} \\
&= \frac{\Gamma\left(\eta_v^{(k)}+c_{kv}\right)}{\Gamma\left(\eta_v^{(k)}+c_{kv}-1\right)} \cdot \frac{\Gamma\left(\sum_{j=1}^{V}\eta_j^{(k)}+\sum_{j=1}^{V}c_{kj}-1\right)}{\Gamma\left(\sum_{j=1}^{V}\eta_j^{(k)}+\sum_{j=1}^{V}c_{kj}\right)} \cdot \frac{\Gamma\left(\alpha_k+c_k^{(d)}\right)}{\Gamma\left(\alpha_k+c_k^{(d)}-1\right)} \cdot \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i+\sum_{i=1}^{K}c_i^{(d)}-1\right)}{\Gamma\left(\sum_{i=1}^{K}\alpha_i+\sum_{i=1}^{K}c_i^{(d)}\right)} \\
&= \frac{\eta_v^{(k)}+c_{kv}-1}{\sum_{j=1}^{V}\eta_j^{(k)}+\sum_{j=1}^{V}c_{kj}-1} \cdot \frac{\alpha_k+c_k^{(d)}-1}{\sum_{i=1}^{K}\alpha_i+\sum_{i=1}^{K}c_i^{(d)}-1} \\
&\propto \frac{\left(\eta_v^{(k)}+c_{kv}-1\right)\cdot\left(\alpha_k+c_k^{(d)}-1\right)}{\sum_{j=1}^{V}\eta_j^{(k)}+\sum_{j=1}^{V}c_{kj}-1}. \qquad (36)
\end{aligned}
$$

Given this full conditional posterior, the Gibbs sampling procedure is as straightforward as follows:

Step 0: Initialize $\vec{\alpha},\boldsymbol{\eta}$, $\boldsymbol{Z}$

Step 1: Iteratively update $\boldsymbol{Z}$ by drawing $z_n^{(d)}$ according to (36)

Step 2: Update $\vec{\alpha},\boldsymbol{\eta}$ by maximizing the joint likelihood function (34). Go to step (1)

The procedure above is generic. The step 1 usually requires huge number of iterations. The objective function in step 2 is tractable compared with (10) and there is no need for a variational lower bound.

One can use Newton's method to maximize the log of this likelihood function. Once the burn-out phase is reached, the two target distributions can be estimated as

$\hat{E}(\vec{\beta}^{(i)}) = (\vec{\eta}^{(i)} + \vec{c}_i)/(\sum_{j=1}^{V} \eta_j^{(i)} + \sum_{j=1}^{V} c_{ij})$, which represents the word distribution over topic i and $\hat{E}(\vec{\theta}^{(d)}) = (\vec{\alpha} + \vec{c}^{(d)})/(\sum_{i=1}^{K} \alpha_i + \sum_{i=1}^{K} c_i^{(d)})$, which represents the topic distribution in document d.

As a special case, $\boldsymbol{\beta}$ is fixed, then the full conditional posterior (36) can be reduced to

$$P(z_n^{(d)} = k | \boldsymbol{Z} \backslash z_n^{(d)}, w_n^{(d)} = v, \boldsymbol{D} \backslash w_n^{(d)}, \vec{\alpha}, \boldsymbol{\beta}) \propto \beta_{kv} \cdot (\alpha_k + c_k^{(d)} - 1). \quad (37)$$

In fact, by reducing (34) and (35), we have

$$\begin{aligned} P(\boldsymbol{Z}, \boldsymbol{D} | \vec{\alpha}, \boldsymbol{\beta}) &= P(\boldsymbol{D} | \boldsymbol{Z}, \vec{\alpha}, \boldsymbol{\beta}) P(\boldsymbol{Z} | \vec{\alpha}, \boldsymbol{\beta}) \\ &= P(\boldsymbol{D} | \boldsymbol{Z}, \boldsymbol{\beta}) P(\boldsymbol{Z} | \vec{\alpha}) \\ &= \left(\prod_{i=1}^{K} \prod_{j=1}^{V} \beta_{ij}^{c_{ij}} \right) \cdot \prod_{d=1}^{M} \frac{\mathrm{B}(\vec{\alpha} + \vec{c}^{(d)})}{\mathrm{B}(\vec{\alpha})}, \end{aligned} \quad (38)$$

$$\begin{aligned} P(\boldsymbol{Z} \backslash z_n^{(d)}, \boldsymbol{D} | \vec{\alpha}, \boldsymbol{\beta}) &= P(w_n^{(d)}, \boldsymbol{Z} \backslash z_n^{(d)}, \boldsymbol{D} \backslash w_n^{(d)} | \vec{\alpha}, \boldsymbol{\beta}) \\ &= P(w_n^{(d)} | \boldsymbol{Z} \backslash z_n^{(d)}, \boldsymbol{D} \backslash w_n^{(d)}, \vec{\alpha}, \boldsymbol{\beta}) P(\boldsymbol{Z} \backslash z_n^{(d)}, \boldsymbol{D} \backslash w_n^{(d)} | \vec{\alpha}, \boldsymbol{\beta}) \\ &= \left(\sum_{z_n^{(d)}=1}^{K} P(w_n^{(d)} | z_n^{(d)}, \boldsymbol{\beta}) P(z_n^{(d)} | \vec{\alpha}) \right) P(\boldsymbol{Z} \backslash z_n^{(d)}, \boldsymbol{D} \backslash w_n^{(d)} | \vec{\alpha}, \boldsymbol{\beta}) \\ &\propto P(\boldsymbol{Z} \backslash z_n^{(d)}, \boldsymbol{D} \backslash w_n^{(d)} | \vec{\alpha}, \boldsymbol{\beta}) \\ &= \left(\frac{\prod_{i=1}^{K} \prod_{j=1}^{V} \beta_{ij}^{c_{ij}}}{\beta_{kv}} \right) \cdot \left(\prod_{m=1}^{M} \frac{\mathrm{B}(\vec{\alpha} + \vec{c}^{(m)\prime})}{\mathrm{B}(\vec{\alpha})} \right), \end{aligned} \quad (39)$$

and dividing (38) by (39) yields (37).

Compared with the variational inference method, which is deterministic, the Gibbs sampling is stochastic (in step 1), hence, it's easier for Gibbs sampling's iterations to jump out of local optima. Moreover, it's easier for the Gibbs sampling algorithm to incorporate the uncertainty on $\boldsymbol{\beta}$. The variational inference method (together with its EM algorithm) introduced in sections 2, 3 assumes that $\boldsymbol{\beta}$ is fixed, whereas it's easy for the Gibbs sampling algorithm to relax this assumption (as the full conditional posterior formula (36) has shown). And as $\boldsymbol{\beta}$ is fixed, the full conditional posterior formula has simpler form. Another advantage is that the Gibbs sampling algorithm relaxes the assumption that documents are independent. In variational inference method, the likelihood of seeing the corpus is the product of the likelihoods of seeing individual documents, however no such assumption is made in the inference based on the Gibbs sampling method. Given the hyperparameters $\vec{\alpha}, \boldsymbol{\eta}$, the step 1 of the Gibbs sampling procedure can be applied to a corpus $\boldsymbol{D} = \{\vec{w}^{(1)}, \ldots, \vec{w}^{(M)}\}$ to calculate the perplexity. The burn-out phases of multiple runs can be averaged to

obtain the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\theta} = \{\vec{\theta}^{(1)}, \dots, \vec{\theta}^{(M)}\}$. Given $\boldsymbol{\beta}, \boldsymbol{\theta}$, the joint likelihood of seeing $\boldsymbol{D}$ is simply expressed as

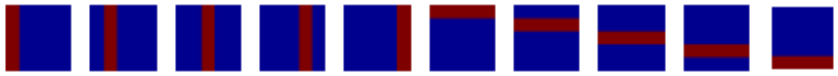
$$
\begin{aligned}
P(\boldsymbol{D}|\boldsymbol{\beta},\boldsymbol{\theta}) &= \textstyle\sum_{\boldsymbol{Z}} P(\boldsymbol{D}|\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\theta}) P(\boldsymbol{Z}|\boldsymbol{\beta},\boldsymbol{\theta}) \\
&= \textstyle\sum_{\boldsymbol{Z}} P(\boldsymbol{D}|\boldsymbol{Z},\boldsymbol{\beta}) P(\boldsymbol{Z}|\boldsymbol{\theta}) \\
&= \textstyle\sum_{\boldsymbol{Z}} \left(\prod_{d=1}^{M}\prod_{n=1}^{N_d} \beta_{z_n^{(d)},w_n^{(d)}}\right)\left(\prod_{d=1}^{M}\prod_{n=1}^{N_d} \theta_{z_n^{(d)}}^{(d)}\right) \\
&= \textstyle\sum_{\boldsymbol{Z}} \prod_{d=1}^{M}\prod_{n=1}^{N_d} \beta_{z_n^{(d)},w_n^{(d)}} \theta_{z_n^{(d)}}^{(d)} \\
&= \textstyle\sum_{z_1^{(1)},\boldsymbol{Z}\setminus z_1^{(1)}} \beta_{z_1^{(1)},w_1^{(1)}} \theta_{z_1^{(1)}}^{(1)} \left(\prod_{d=1}^{M}\prod_{n=1}^{N_d} \beta_{z_n^{(d)},w_n^{(d)}} \theta_{z_n^{(d)}}^{(d)}\right) \Big/ \left(\beta_{z_1^{(1)},w_1^{(1)}} \theta_{z_1^{(1)}}^{(1)}\right) \\
&= \left(\textstyle\sum_{z_1^{(1)}} \beta_{z_1^{(1)},w_1^{(1)}} \theta_{z_1^{(1)}}^{(1)}\right) \cdot \textstyle\sum_{\boldsymbol{Z}\setminus z_1^{(1)}} \left(\prod_{d=1}^{M}\prod_{n=1}^{N_d} \beta_{z_n^{(d)},w_n^{(d)}} \theta_{z_n^{(d)}}^{(d)}\right) \Big/ \left(\beta_{z_1^{(1)},w_1^{(1)}} \theta_{z_1^{(1)}}^{(1)}\right) \\
&= \left(\textstyle\sum_{i=1}^{K} \beta_{i,w_1^{(1)}} \theta_i^{(1)}\right) \cdot \textstyle\sum_{\boldsymbol{Z}\setminus z_1^{(1)}} \left(\prod_{d=1}^{M}\prod_{n=1}^{N_d} \beta_{z_n^{(d)},w_n^{(d)}} \theta_{z_n^{(d)}}^{(d)}\right) \Big/ \left(\beta_{z_1^{(1)},w_1^{(1)}} \theta_{z_1^{(1)}}^{(1)}\right) \\
&= \cdots \\
&= \textstyle\prod_{d=1}^{M}\prod_{n=1}^{N_d} \sum_{i=1}^{K} \beta_{i,w_n^{(d)}} \theta_i^{(d)} \\
&= \textstyle\prod_{d=1}^{M}\prod_{j=1}^{V} \left(\sum_{i=1}^{K} \beta_{ij} \theta_i^{(d)}\right)^{n_j^{(d)}}, \qquad (40)
\end{aligned}
$$

where $n_j^{(d)}$ is number of j-th keyword in document d. This joint likelihood of corpus leads to a new perplexity formula as

$$
\text{perplexity } (\boldsymbol{D}) = \exp\left\{-\frac{\sum_{d=1}^{M}\sum_{j=1}^{V} n_j^{(d)} \ln\left(\sum_{i=1}^{K} \beta_{ij}\theta_i^{(d)}\right)}{\sum_{d=1}^{M} N_d}\right\}. \qquad (41)
$$

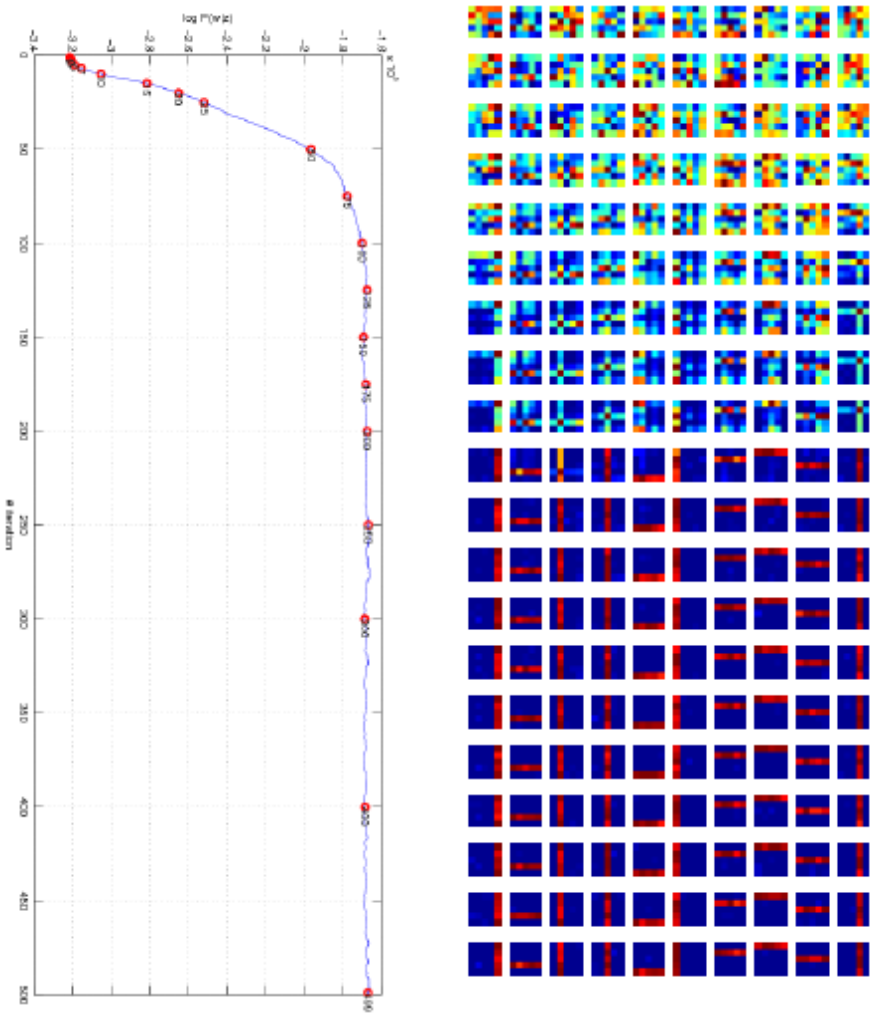
Particularly, $\prod_{j=1}^{V}\left(\sum_{i=1}^{K} \beta_{ij}\theta_i^{(d)}\right)^{n_j^{(d)}} = P(\vec{w}^{(d)}|\boldsymbol{\beta},\vec{\theta}^{(d)})$. Although this perplexity formula assumes the independence between documents, it does not deal with the integral over $\boldsymbol{\theta}$. However, the estimation method is still based on sampling. To obtain the "expected" perplexity, the multiple runs of drawing $z_n^{(d)}$ are required for obtaining the expected $\boldsymbol{\beta}, \boldsymbol{\theta}$ corresponding to the target corpus.

There is a very elegant toy example for testing any implementation of an LDA learning algorithm. Suppose there are 10 topics and 25 words, and the ground truth topic-to-word distribution $\boldsymbol{\beta}$ (10×25 matrix) is represented by the following figure



The ground truth of $\boldsymbol{\beta}$ (10×25 matrix)

Also suppose $\vec{\alpha} = \vec{1}$, which means each topic is expected to have equal chance to appear. After generating 2000 documents with each has length 100, it is expected to recover the ground truth of $\boldsymbol{\beta}$ from the artificially generated data. The following figure shows that it's well recovered by the Gibbs sampling procedure.



Gibbs sampling results for learning $\boldsymbol{\beta}$

5 Application

5.1 Classification

For the classification purpose, suppose an LDA model $\{\vec{\alpha}^{(i)}, \boldsymbol{\beta}^{(i)}\}$ is trained from a corpus that is labeled i, then for an unlabeled document $\vec{w}$, the posterior probability that it belongs to class i is

$$P(i|\vec{w}) = \frac{P(\vec{w}|i)P(i)}{\sum_j P(\vec{w}|j)P(j)} = \frac{P(\vec{w}|\vec{\alpha}^{(i)},\boldsymbol{\beta}^{(i)})P(i)}{\sum_j P(\vec{w}|\vec{\alpha}^{(j)},\boldsymbol{\beta}^{(j)})P(j)}, \tag{42}$$

where $P(\vec{w}|\vec{\alpha}^{(i)}, \boldsymbol{\beta}^{(i)})$ is the i-th conditional likelihood that can be evaluated according to (8), and $P(i)$ is the portion of documents that is labeled i. Practically, due to the small scale of the joint probability of the keywords sequence of a document, only $\ln P(\vec{w}|\vec{\alpha}^{(j)}, \boldsymbol{\beta}^{(j)})$ is available for each j. Let $\tau_j = \ln P(\vec{w}|\vec{\alpha}^{(j)}, \boldsymbol{\beta}^{(j)}) + \ln P(j)$ and $\tau_M = \max_j \tau_j$, then

$$P(i|\vec{w}) = \frac{\exp(\tau_i - \tau_M)}{\sum_j \exp(\tau_j - \tau_M)}, \tag{43}$$

which is an usual formula for avoiding numerical overflow.

Since the size of the dictionary could be very large, the problem of feature selection arises. An ingenious idea that David Blei proposed is to take all the labeled and unlabelled documents as a single corpus and apply the variational inference method for the purpose of feature selection. Note that the objective function is $\boldsymbol{L}(\boldsymbol{\phi}, \boldsymbol{\gamma}; \vec{\alpha}, \boldsymbol{\beta}) = \sum_{d=1}^{M} L(\vec{w}^{(d)}; \vec{\phi}^{(1,d)}, \dots, \vec{\phi}^{(N_d,d)}, \vec{\gamma}^{(d)}; \vec{\alpha}, \boldsymbol{\beta})$. Upon the termination of the algorithm, we will have $\vec{\gamma}^{*(d)}$ for each d. And $\vec{\gamma}^{*(d)}$ is a document-specific parameter vector of the Dirichlet posterior of the topic belief . Since $\vec{\gamma}^{*(d)}$ has much lower dimension than $\vec{w}^{(d)}$, we can use $\vec{\gamma}^{*(d)}$ as the new representation of the document d. Compared with the usual feature selection methods like mutual information and χ^2 statistics, which are basically greedy selection methods that represent each feature with a single word, each feature generated by LDA's variational inference process is represented by a topic that is related to many words. This implies that the LDA's feature vector of K dimensions carry more information than selecting K words. Without the worry about which key-word is informative for whether assigning a document to a class, the LDA features are theoretically superior.

Moreover, this feature selection idea can also be implemented when the Gibbs sampling algorithm is used. That is, the Gibbs sampling algorithm can be applied on all the labeled and unlabelled documents. consequently, for each document d, we can estimate the expected topic composition $E(\vec{\theta}^{(d)})$ as $\hat{E}(\vec{\theta}^{(d)}) = (\vec{\alpha} + \vec{c}^{(d)})/(\sum_{i=1}^{K} \alpha_i + \sum_{i=1}^{K} c_i^{(d)})$, where $\vec{c}^{(d)}$ is counted from $\vec{z}^{(d)}$ that is generated by the Gibbs sampling. We can use $\hat{E}(\vec{\theta}^{(d)})$ as the new dimension-reduced representation of the document d for the purpose of document classification.

Sometimes it's interesting to estimate $P(i|z)$ where z is a topic hidden in unlabelled documents. By conditioning on unlabelled document, we have

$$P(i|z) = \sum_{\vec{w}} P(i|\vec{w}, z)P(\vec{w}|z)$$

$$= \sum_{\vec{w}} P(i|\vec{w})P(\vec{w}|z)$$

$$= \sum_{\vec{w}} \frac{P(i|\vec{w})P(z|\vec{w})}{P(z)} P(\vec{w})$$

$$= E\left(\frac{P(i|\vec{w})P(z|\vec{w})}{P(z)}\right). \quad (44)$$

Practically there are only limited number of unlabelled documents, say $\boldsymbol{D} = \{\vec{w}^{(1)}, \dots, \vec{w}^{(M)}\}$. But one estimator of (44) can be built as

$$\hat{P}(i|z) = \frac{1}{M}\sum_{d=1}^{M} \frac{P\left(i|\vec{w}^{(d)}\right)P\left(z|\vec{w}^{(d)}\right)}{P(z)}$$

$$= \frac{1}{M}\sum_{d=1}^{M} \frac{P\left(i|\vec{w}^{(d)}\right)P(\vec{w}^{(d)}|z)}{P(\vec{w}^{(d)})}$$

$$= \frac{1}{M}\sum_{d=1}^{M} \frac{P\left(i|\vec{w}^{(d)}\right)\prod_{n=1}^{N_d} P(w_n^{(d)}|z)}{P(\vec{w}^{(d)})}$$

$$= \frac{1}{M}\sum_{d=1}^{M} \frac{P\left(i|\vec{w}^{(d)}\right)\prod_{n=1}^{N_d} \beta_{z,w_n^{(d)}}}{P(\vec{w}^{(d)})}, \quad (45)$$

where $P\left(i\middle|\vec{w}^{(d)}\right)$ can be computed using (43) and $P(\vec{w}^{(d)})$ can be computed as $P(\vec{w}^{(d)}) = \sum_j P\left(\vec{w}^{(d)}\middle|\vec{\alpha}^{(j)}, \boldsymbol{\beta}^{(j)}\right)P(j)$. One may need special care for computing $\rho_d = \frac{\prod_{n=1}^{N_d} \beta_{z,w_n^{(d)}}}{P(\vec{w}^{(d)})}$ in order to avoid numerical overflow. Let $\tau'_d = \ln \prod_{n=1}^{N_d} \beta_{z,w_n^{(d)}} = \sum_{n=1}^{N_d} \ln \beta_{z,w_n^{(d)}}$ and $\tau_d = \ln P(\vec{w}^{(d)})$, then $\rho_d = \frac{\exp(\tau'_d)}{\exp(\tau_d)} = \exp(\tau'_d - \tau_d)$. Another method for computing (45) is to apply the Gibbs sampling to infer the latent topics sequence $\vec{z}^{(d)}$ for $\vec{w}^{(d)}$. Given $\vec{z}^{(d)}$, $P\left(z\middle|\vec{w}^{(d)}\right)$ can be estimated as the portion of the topic z in $\vec{z}^{(d)}$. And finally $P(z)$ can be estimated as the portion of the topic z in $\vec{z}^{(1)}, \dots, \vec{z}^{(M)}$. Since $P(z) = \sum_{\vec{w}} P(z|\vec{w})P(\vec{w}) = E\left(P(z|\vec{w})\right)$, one estimator of $P(z)$ can be $\frac{1}{M}\sum_{d=1}^{M} P\left(z\middle|\vec{w}^{(d)}\right)$, which yields

$$\hat{P}(i|z) = \frac{1}{M}\sum_{d=1}^{M} \frac{P\left(i|\vec{w}^{(d)}\right)P\left(z|\vec{w}^{(d)}\right)}{\frac{1}{M}\sum_{d'=1}^{M} P\left(z|\vec{w}^{(d')}\right)} = \sum_{d=1}^{M} \frac{P\left(z|\vec{w}^{(d)}\right)}{\sum_{d'=1}^{M} P\left(z|\vec{w}^{(d')}\right)} P\left(i\middle|\vec{w}^{(d)}\right), \quad (46)$$

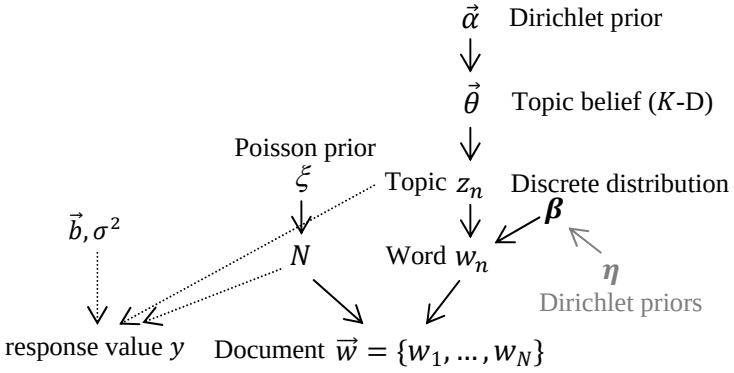
which is the weighted average of the $P(i|\vec{w})$ over $\vec{w}^{(1)}, \dots, \vec{w}^{(M)}$. It may need the model averaging in burn-out phase of the Gibbs sampling procedure to obtain averaged estimate of $P\left(z\middle|\vec{w}^{(d)}\right)$.

5.2 Similarity

The similarity between two documents $\vec{w}^{(d)}$ and $\vec{w}^{(f)}$ can be measured by the "distance" between $E(\vec{\theta}^{(d)})$ and $E(\vec{\theta}^{(f)})$, each of which is the expected topic composition. Given the LDA hyper-

parameters $\{\vec{\alpha}, \boldsymbol{\eta}\}$, they are evaluated as $\hat{E}\left(\vec{\theta}^{(d)}\right) = \left(\vec{\alpha} + \vec{c}^{(d)}\right)/\left(\sum_{i=1}^{K} \alpha_i + \sum_{i=1}^{K} c_i^{(d)}\right)$ and $\hat{E}\left(\vec{\theta}^{(f)}\right) = \left(\vec{\alpha} + \vec{c}^{(f)}\right)/\left(\sum_{i=1}^{K} \alpha_i + \sum_{i=1}^{K} c_i^{(f)}\right)$, where $\vec{c}^{(d)}$ and $\vec{c}^{(f)}$ are counted from $\vec{z}^{(d)}$ and $\vec{z}^{(f)}$ that are generated from Gibbs sampling. One can calculate the similarity measure of $\hat{E}\left(\vec{\theta}^{(d)}\right)$ and $\hat{E}\left(\vec{\theta}^{(f)}\right)$. If the variational inference method is used, then one can calculate the similarity measure of $\vec{\gamma}^{*(d)}/\sum_i \gamma_i^{*(d)}$ and $\vec{\gamma}^{*(f)}/\sum_i \gamma_i^{*(f)}$.

5.3 Supervised LDA



Graphical representation of sLDA

$$P\left(\vec{w}, y|\vec{\alpha}, \boldsymbol{\beta}, \vec{b}, \sigma^2\right)$$

$$= \sum_{\vec{z}} P\left(\vec{w}, y|\vec{z}, \vec{\alpha}, \boldsymbol{\beta}, \vec{b}, \sigma^2\right) P\left(\vec{z}|\vec{\alpha}, \boldsymbol{\beta}, \vec{b}, \sigma^2\right)$$

$$= \sum_{\vec{z}} P\left(y|\vec{z}, \vec{b}, \sigma^2\right) P(\vec{w}|\vec{z}, \boldsymbol{\beta}) P(\vec{z}|\vec{\alpha})$$

$$= \sum_{\vec{z}} P\left(y|\vec{z}, \vec{b}, \sigma^2\right) P(\vec{w}|\vec{z}, \boldsymbol{\beta}) \int P\left(\vec{z}|\vec{\theta}\right) P\left(\vec{\theta}|\vec{\alpha}\right) d\vec{\theta}$$

$$= \int \left(\sum_{\vec{z}} P\left(y|\vec{z}, \vec{b}, \sigma^2\right) P(\vec{w}|\vec{z}, \boldsymbol{\beta}) P\left(\vec{z}|\vec{\theta}\right)\right) P\left(\vec{\theta}|\vec{\alpha}\right) d\vec{\theta}$$

$$= \int P\left(\vec{\theta}|\vec{\alpha}\right) d\vec{\theta} \sum_{\vec{z}} P\left(y|\vec{z}, \vec{b}, \sigma^2\right) \prod_{n=1}^{N} \left(P(w_n|z_n, \boldsymbol{\beta}) P\left(z_n|\vec{\theta}\right)\right). \quad (47)$$

$$y|\vec{z}, \vec{b}, \sigma^2 \sim N\left(\frac{1}{N} \vec{b}^T \vec{c}, \sigma^2\right), \quad (48)$$

where $\vec{c}$ is the counts vector of $\vec{z}$.

$$P\left(y|\vec{w}, \vec{\alpha}, \boldsymbol{\beta}, \vec{b}, \sigma^2\right) = \frac{P\left(\vec{w}, y|\vec{\alpha}, \boldsymbol{\beta}, \vec{b}, \sigma^2\right)}{P(\vec{w}|\vec{\alpha}, \boldsymbol{\beta})}. \quad (49)$$

It's not quite clear how this method (by David Blei and Jon McAliffe, 2007) is superior to the method of first finding the LDA feature representations and then do regression. Also, compared with (8), the summation $\Sigma_{\vec{z}}$ in (47) cannot be simplified. However, the experimental results reported in the 2007 paper shows better results than the regression using LDA features and Lasso.

(to be continued …)