A supplment material to understand Latent Dirichlet Allocation

Saki Kuzushima

2019-05-29

The Latent Dirichlet Allocation model is very popular among the political scientists, and many people have summarized the model and the inference. However, understanding LDA requires some basis. This material aims to provide essential definitions and theorems to understand LDA and their variants. First section sketches the model and the infernece of LDA, and the second section describes some definitions and theorems essential to understand the model and the vaious inferences.

Part I

1. Latent Dirichlet Allocation model(s)

1.1 The original model

We firstly define the key terms used in the model.

- Word: an element of a vocabulary.
- Vocabulary: a set of V unique words appear in the entire corpus.
- Document: a sequence of N words, denoted by $d = (w_1, w_2, , ..w_N)$
- Corpus: a set of J documents, denoted by $D = (d_1, d_2...d_M)$
- Topic: a distribution of words. We assume that the number of topics is K.

LDA is the following generating probabilistic model.

- For each document, $d_j = (w_{j,1}, w_{j,2}, , ..w_{j,N}), j \in [1...J]$,
  - $\theta_j \sim Dirichlet(\alpha)$
  - For each word, $w_n, n \in [1...N]$
    - $z_{j,n} \sim Mutlnimoial(\theta_j)$
    - $w_{j,n} \sim Multinomial(\phi_{z_{j,n}})$

For instance, we have three topics on a news article: sports, politics, and technology ($K = 3$). θ_j, which is the distribution of these topics over the document d_j was drawn from the Dirichlet distribution. We assume that the sampled $\theta_j = [\frac{1}{3}, \frac{1}{4}, \frac{5}{12}]$, where the order of topics are [sports, politics, technology]. Substantively, this means that the document d_j is a mixture of 1/3 "sports", 1/4 "poliitcs", and 5/12 "technology".

Next, we draw the topic of each word $z_{j,n}$ from $Multinomial(\theta)$. We assume this happens to be "politics." Then, we sample each word $w_{j,n}$ from $Multinomial(\phi_{zj,n})$.

Assuming that the documents and the words are independent, $d_j \perp\!\!\!\perp d_{j'}$ and $w_{j,n} \perp\!\!\!\perp w_{j,n'}$, we can write the joint distribution of W, Z, Θ given α, Φ in the following way.

$$p(W, Z, \Theta|\alpha, \Phi) = \prod_{j=1}^{J} p(\theta_j|\alpha) \prod_{j=1}^{K} p(z_{j,n}|\theta_j)p(w_{j,n}|\phi_{z_{j,n}}) \tag{1}$$

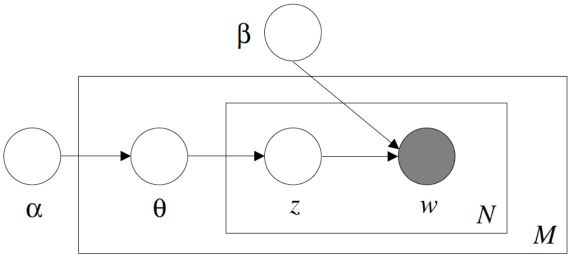

Figure 1: LDA (Source: Blei et al. 2003)

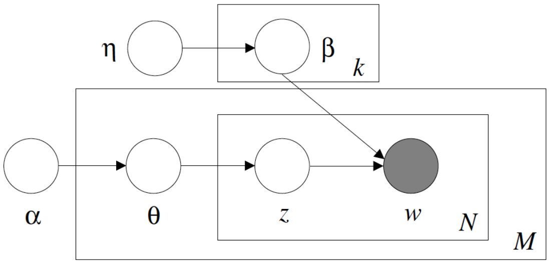

Figure 2: Smoothed LDA (Source: Blei et al. 2003)

1.2 Smoothed LDA

Alternatively, we can assign a prior to ϕ. This model is called smoothed LDA.

- For each topic, $\phi_i \sim Dirichlet(\beta), i \in [1, ..., K]$
- For each document, $d_j = (w_{j,1}, w_{j,2}, , ..w_{j,N}), j \in [1...J]$,
  - $\theta_j \sim Dirichlet(\alpha)$
  - For each word, $w_n, n \in [1...N]$
    - $z_{j,n} \sim Mutlnimoial(\theta_j)$
    - $w_{j,n} \sim Multinomial(\phi_{z_{j,n}})$

In this model, we can write the joint distribution of W, Z, Θ given α, β.

$$p(W, Z, \Theta|\alpha, \beta) = \prod_{i=1}^{K} p(\phi_j|\beta) \prod_{j=1}^{J} p(\theta_j|\alpha) \prod_{j=1}^{K} p(z_{j,n}|\theta_j) p(w_{j,n}|\phi_{z_{j,n}}) \tag{2}$$

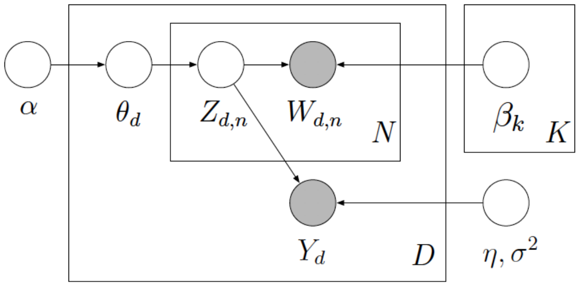

Figure 3: supervised LDA (Source: Blei and McAuliffe 2008)

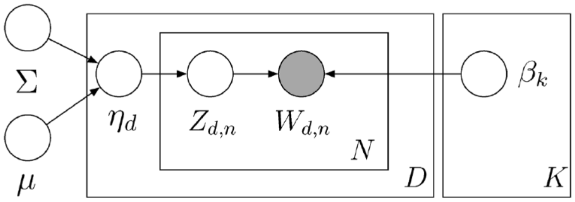

Figure 4: Correlated Topic Model (Source: Blei and Lafferty 2007)

1.3 Supervised LDA

LDA does not guarantee that it returns the topics we are substantively interested in. For instance, we might be interested in the sentiment of movie review, but the topics generated by the LDA might represents categories of movies (e.g. comedy, drama, romance etc. . .). Supervised LDA aims to ameliorate this problem, by incorporating humans' input. (Blei and McAuliffe 2008). The model is the following.

- For each document, $d_j = (w_{j,1}, w_{j,2}, , ..w_{j,N}), j \in [1...J]$,
  - $\theta_j \sim Dirichlet(\alpha)$
  - For each word, $w_n, n \in [1...N]$
    - $z_{j,n} \sim Mutlnimoial(\theta_j)$
    - $w_{j,n} \sim Multinomial(\phi_{z_{j,n}})$
  - $y|z_{1:N}, \eta, \sigma^2 \sim N(\eta^\intercal \bar{z}, \sigma^2)$

where $\bar{z} = \frac{1}{N}\sum_{n=1}^{N} z_n$. The last line is the change from the original LDA. It is a normal linear model, but we could change it to a generalizing linear model, particularly when y is a categorical variable. The covariate is the average number of each topic of a parcitular document.

1.4 Correlated Topipc Model (CTM)

One of the problems of LDA is that the topics are assumed to be nearly independent. However, topics are likely to be correlated. For instance, in the corpus of academic journal articles, a document with a topic about public opinion is more likely to contain a topic about election than a topic about Bayesian statistics. This problem arises from the Dirichlet prior to the topic-word parameter. CTM replaces Dirichlet to logit-normal distribution, and the covariance matrix of normal distribution can control the correlation across topics. The drawback of this model is the loss of conjugacy between Multinomial and Dirichlet. Blei and Lafferty (2007) uses varitaionl inference to approximate posterior distribution.

- $\eta_d|\mu, \Sigma \sim N(\mu, \Sigma)$
- For $n \in 1...N_d$,
  - $Z_{d,n}|\eta_d \sim Multinomial(f(\eta_d))$
  - $W_{d,n}|z_{d,n,\beta_{1:K}} \sim Multinomial(\beta_{z_{d,n}})$

where $f(\eta) = \frac{\exp(\eta)}{\sum_i \exp(\eta_i)}$.

2. Inference

This section discuss the inference methods of the original LDA. First it is useful to check the dimension of the parameters.

Assuming that

- J: the number of documents
- K: the number of topics
- V: the number of unique words in all documents
- N: the number of words in each document. [1]

Then, the dimension of the parameters are

- α: $K \times 1$: Dirichlet hyperparameter
- Θ: $J \times K$: document - topic matrix
- Φ: $K \times V$: topic - vocabulary matrix
- Z: $J \times N$: document - word matrix ($z_{j,i} = k \in \{1...K\}$)
- W: $J \times N$: document - word matrix ($w_{j,i} = v \in \{1...V\}$)

What we are interested in is the posterior distribution of Z (latent variable) and Θ (unknown parameter). Note that, although Φ is an unknown paraeter too, we cannot estimate this in the Bayesian framework because we do not have a prior to Φ in the original LDA. So, we estimate Φ by maximum likelihood estimation (we skip this for a moment). In constrast, in the smoothed LDA we have a prior β to Φ, so we can also estimate Φ in the same way as Θ. [2]

By the definition of conditional density,

$$p(Z, \Theta|W, \Phi, \alpha) = \frac{p(W, Z, \Theta|\Phi, \alpha)}{p(W|\Phi, \alpha)} \tag{3}$$

We want to obtain this density. First we start with writing out the joint posterior distribution.

$$p(W, Z, \Theta|\Phi, \alpha) = p(\Theta|\alpha)p(Z|\Theta)p(W|Z, \Phi) \tag{4}$$

Because $\theta_j \perp\!\!\!\perp \theta_{j'}, j \neq j'$ and $Z_{ij} \perp\!\!\!\perp Z_{i'j}, i \neq i'$, we can rewrite each of the three terms in the RHS as

$$p(\Theta|\alpha) = \prod_{j=1}^{J} p(\theta_j|\alpha)$$

$$p(Z|\Theta) = \prod_{j=1}^{J}\prod_{i=1}^{N} p(z_{ij}|\theta_j)$$

$$p(W|Z, \Phi) = \prod_{j=1}^{J}\prod_{i=1}^{N} p(w_{ij}|\phi_k, z_{ij} = k)$$

[1] N could vary across documents, but we assume the number of words is the same across documents for simplicity. Relaxing this assumption does not affect the following inference.

[2] Now we can see why Geigle used the notation Φ although Blei used β...

Therefore, the joint posterior distribution (4) becomes

$$p(W, Z, \Theta|\Phi, \alpha) = \prod_{j=1}^{J} p(\theta_j|\alpha) \prod_{i=1}^{N} \ p(z_{ij}|\theta_j) \ p(w_{ij}|\phi_k, z_{ij} = k) \tag{5}$$

We obtained the numerator of (3). The denominator is

$$\begin{aligned} p(W|\Phi, \alpha) &= \prod_{j=1}^{J} p(w_j|\Phi, \alpha) && (6)\\ &= \prod_{j=1}^{J} \int_{\theta_j} \sum_{z_j} p(w_j, \theta_j, z_j, |\Phi, \alpha) && (7)\\ &= \prod_{j=1}^{J} \int_{\theta_j} \sum_{z_j} p(\theta_j|\alpha) p(z_j|\theta_j) p(w_j|z_j, \Phi) \ d\theta_j && (8)\\ &= \prod_{j=1}^{J} \int_{\theta_j} p(\theta_j|\alpha) \sum_{z_j} p(z_j|\theta_j) p(w_j|z_j, \Phi) \ d\theta_j && (9)\\ & && (10) \end{aligned}$$

Focusing on the parts from the summation

$$\begin{aligned} &\sum_{z_j} \prod_{i=1}^{N} \ p(z_{ij}|\theta_j) \ p(w_{ij}|\phi_k, z_{ij}) && (11)\\ &= \sum_{z_{1j}} \cdots \sum_{z_{Nj}} \prod_{i=1}^{N} \ p(z_{ij}|\theta_j) \ p(w_{ij}|\phi_k, z_{ij}) && (12)\\ &= \sum_{z_{1j}} p(z_{1j}|\theta_j) \ p(w_{1j}|\phi_k, z_{1j}) \cdots \sum_{z_{Nj}} p(z_{Nj}|\theta_j) \ p(w_{Nj}|\phi_k, z_{Nj}) && (13)\\ &= \prod_{i=1}^{N} \sum_{z_{ij}} p(z_{ij}|\theta_j) \ p(w_{ij}|\phi_k, z_{ij}) && (14)\\ &= \prod_{i=1}^{N} \sum_{k=1}^{K} p(z_{ij} = k|\theta_j) \ p(w_{ij}|\phi_k, z_{ij} = k) && (15)\\ & && (16) \end{aligned}$$

Plugging this back,

$$
\begin{align}
p(W|\Phi,\alpha) &= \prod_{j=1}^{J}\int_{\theta_j} p(\theta_j|\alpha)\sum_{z_j} p(z_j|\theta_j)p(w_j|z_j,\phi_j)\ d\theta_j \tag{17}\\
&= \prod_{j=1}^{J}\int_{\theta_j} p(\theta_j|\alpha)\prod_{i=1}^{N}\sum_{k=1}^{K} p(z_{ij}=k|\theta_j)\ p(w_{ij}=v|\phi_k, z_{ij}=k)\ d\theta_j \tag{18}\\
&= \prod_{j=1}^{J}\int_{\theta_j} \frac{1}{B(\alpha)}\prod_{k'=1}^{K}\theta_{k'}^{\alpha_{k'}-1}\prod_{i=1}^{N}\sum_{k=1}^{K}(\theta_{j,k}\phi_{k,v})^{\sum_{v=1}^{V}1\{w_{ij}=v\}}\ d\theta_j \tag{19}\\
&= \prod_{j=1}^{J}\int_{\theta_j} \frac{1}{B(\alpha)}\prod_{k'=1}^{K}\theta_{k'}^{\alpha_{k'}-1}\prod_{i=1}^{N}\sum_{k=1}^{K}\prod_{v=1}^{V}(\theta_{j,k}\phi_{k,v})^{1\{w_{ij}=v\}}\ d\theta_j \tag{20}\\
& \tag{21}
\end{align}
$$

The last expression is "intractable". There are two main ways to get around this problem: Variational Inference and Gibbs sampler.

[3] [4]

2.1 Variational Inference (the original method by Blei et al.)

The original paper uses variational inference. Variational inference approximates the intractable posterior with another distribution that varies across some parameters, and tune the parameters to minimize the "distance" between the posterior and the approximate distribution. The "distance" of the posterior and the approximate distributions are measured by the Kullback-Leibler divergence. [5]

For the two pmf p and q, which are defined on the same probability space, the Kullback-Leibler divergence between q and p is defined as the following.

$$D_{KL}(q||p) = \mathbb{E}_q[\log q - \log p] = \sum_{i=1}^{N} q(x_i)(\log q(x_i) - \log p(x_i)) = \sum_{i=1}^{N} q(x_i)\ \log\frac{q(x_i)}{p(x_i)}$$

if x_i is discrete. Summation is replaced by integration if x_i is continuous.[6]

Let's focus on the document level distribution. We want a nice approximate distribution for $p(z_j,\theta_j|w_j,\Phi,\alpha)$. Define the approximate distribution q as

$$q \equiv q(z_j,\theta_j|\gamma_j,\pi_j) = q(\theta_j|\gamma_j)\prod_{i=1}^{N} q(z_{ij}|\pi_{ij})$$

where

$$\theta_j \sim Dirichlet(\gamma_j)$$
$$z_{ij} \sim Multinomial(\pi_j)$$

[3] Let's assume the computational time for $\prod_{v=1}^{V}(\theta_{j,k}\phi_{k,v})^{1\{w_{ij}=v\}}$ is t, then the computational time for $\prod_{i=1}^{N}\sum_{k=1}^{K}\prod_{v=1}^{V}(\theta_{j,k}\phi_{k,v})^{1\{w_{ij}=v\}}$ is tK^N. This is prohibitively large because N is the number of words in each document. Also, we have to integarate over θ_j, which is a $1\times K$ vector and it is a $K-1$ simplex (the summation of all elements must be 1 and each elemnt must be positive.). It is again prohibitively costly to compute many grid values over this simplex.

[4] According to wikipedia, "A problem that can be solved in theory (e.g. given large but finite resources, especially time), but for which in practice any solution takes too many resources to be useful, is known as an intractable problem." However, Blei et al (2003) wrote "is intractable due to the coupling between θ and β (ϕ in our case) in the summation over latent topics (Dickey, 1983)." This means that the summation inside the integral does not allow us to obtain the expression for the integral (if we do not have summation, the integrant is Dirichlet kernel.)

[5] KL divergence is not exactly a distance because $D_{KL}(q||p) \neq D_{KL}(p||q)$, so it is called divergence.

[6] For more information about KL divergence, this blog is helpful.

The goal is to obtain

$$(\pi_j^*, \gamma_j^*) = \text{argmin}_{\pi_j, \gamma_j}\ D_{KL}(q||p) \tag{22}$$

$$\begin{align}
D_{KL}(q||p) &= \mathbb{E}_{q(z_j, \theta_j|\gamma_j, \pi_j)}[\log \frac{q(z_j, \theta_j|\gamma_j, \pi_j)}{p(z_j, \theta_j|w_j, \Phi, \alpha)}] \tag{23}\\
&= \mathbb{E}_{q(z_j, \theta_j|\gamma_j, \pi_j)}[\log \frac{q(z_j, \theta_j|\gamma_j, \pi_j)p(w_j|\Phi, \alpha)}{p(z_j, \theta_j, w_j|, \Phi, \alpha)}] \tag{24}\\
&= \mathbb{E}_{q(z_j, \theta_j|\gamma_j, \pi_j)}[\log q(z_j, \theta_j|\gamma_j, \pi_j) + \log p(w_j|\Phi, \alpha) - \log p(z_j, \theta_j, w_j|, \Phi, \alpha)] \tag{25}\\
&= \mathbb{E}_{q(z_j, \theta_j|\gamma_j, \pi_j)}[\log q(z_j, \theta_j|\gamma_j, \pi_j) - \log p(z_j, \theta_j, w_j|, \Phi, \alpha)]\ \ + \log p(w_j|\Phi, \alpha) \tag{26}\\
& \tag{27}
\end{align}$$

The last term do not depend on γ_j, π_j, so we want to minimize the expectaiton. Define L

$$\begin{align}
L &= -\mathbb{E}_{q(z_j, \theta_j|\gamma_j, \pi_j)}[\log q(z_j, \theta_j|\gamma_j, \pi_j) - \log p(z_j, \theta_j, w_j|, \Phi, \alpha)] \tag{28}\\
&= \mathbb{E}_{q(z_j, \theta_j|\gamma_j, \pi_j)}[\log p(z_j, \theta_j, w_j|, \Phi, \alpha) - \log q(z_j, \theta_j|\gamma_j, \pi_j)] \tag{29}\\
&= \mathbb{E}_q[\log\{p(\theta_j|\alpha)p(z_j|\theta_j)p(w_j|z_j, \Phi)\} - \log\{q(z_j|\pi_j)q(w_j|z_j, \Phi)\}] \tag{30}\\
&= \mathbb{E}_q[\log p(\theta_j|\alpha)] + \mathbb{E}_q[\log p(z_j|\theta_j)] + \mathbb{E}_q[\log p(w_j|z_j, \Phi)] - \mathbb{E}_q[\log q(z_j|\pi_j)] - \mathbb{E}_q[\log q(\theta_j|\gamma_j)] \tag{31}
\end{align}$$

Then the problem (21) becomes

$$(\pi_j^*, \gamma_j^*) = \text{argmax}_{\pi_j, \gamma_j}\ L \tag{32}$$

We inspect this expression one term by one term. The first term can be simplified as the follwoing.

$$\begin{align}
\mathbb{E}_q[\log p(\theta_j|\alpha)] &\propto \mathbb{E}_q\Big[\log \prod_{k=1}^{K} \theta_{j,k}^{\alpha_k - 1}\Big] \tag{33}\\
&= \mathbb{E}_q[\sum_{k=1}^{K}(\alpha_k - 1)\log \theta_{j,k}] \tag{34}\\
&= \sum_{k=1}^{K}(\alpha_k - 1)\log \mathbb{E}_q[\log \theta_{j,k}] \tag{35}\\
& \tag{36}
\end{align}$$

We compute the expectation. Because in the expectation $\theta_j \sim \mathit{Dirichlet}(\gamma_j)$, and Dirichlet distribution belongs to the exponential family [7] we can use the theorem about the expectation of the sufficient statistic. [8]

The theorem states that

$$\mathbb{E}_\theta[t(\theta)] = \frac{\partial}{\partial \eta} A(\eta)$$

[7] Proof in Part II section 2

[8] Proof in Part II section 3

where $t(\theta)$ is the sufficient statistic, η is the natual parameter, $A(\eta)$ is the log partition (log of normalizing factor). In our case,

$$t(\theta) = \log \theta_j$$
$$\eta = \gamma_j$$
$$A(\eta) = \log(\frac{1}{B(\gamma_j)})$$

Simplyfuing $A(\eta)$,

$$\begin{aligned}
A(\eta) &= \log(\frac{1}{B(\gamma)}) \\
&= \log\Big(\frac{\prod_{k=1}^{K}\Gamma(\gamma_k)}{\Gamma(\sum_{k=1}^{K}\gamma_k)}\Big) \\
&= \log\prod_{k=1}^{K}\Gamma(\gamma_k) - \log\Gamma(\sum_{k=1}^{K}\gamma_k) \\
&= \sum_{k=1}^{K}\log\Gamma(\gamma_k) - \log\Gamma(\sum_{k=1}^{K}\gamma_k)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}_q[\log(\theta_{j,k})] &= \frac{\partial}{\partial\gamma_{j,k}}\sum_{k=1}^{K}\log\Gamma(\gamma_{j,k}) - \log\Gamma(\sum_{k=1}^{K}\gamma_{j,k}) \\
&= \sum_{k=1}^{K}\Psi(\gamma_{j,k}) - \Psi(\sum_{k=1}^{K}\gamma_{j,k})
\end{aligned}$$

where $\Psi(\gamma_{j,k}) \equiv \frac{\partial}{\partial\gamma_{j,k}}\log\Gamma(\gamma_{j,k})$.

2.2 Collapsed Gibbs Sampler

The posterior we want to obtain is

$$p(Z_{i^*j^*} = k^*|Z_{\neg i^*j^*}W, \alpha, \beta) \propto p(W, Z|\alpha, \beta)$$

So, we first derive the joint distribution of W, Z. By the definition of conditional probability,

$$p(W, Z|\alpha, \beta) = p(Z|\alpha)p(W|Z, \beta) \tag{37}$$

We derive the expression of the each term of the RHS.

First, we derive $p(Z|\alpha)$.

$$\begin{aligned}
p(Z|\alpha) &= \int_{\Theta} p(Z, \Theta|\alpha)\ d\Theta &(38)\\
&= \int_{\Theta} p(\Theta|\alpha)p(Z|\Theta)d\Theta &(39)
\end{aligned}$$

Recall that

$$\theta_j \overset{i.i.d}{\sim} Dirichlet(\alpha),\ j \in [1, J]$$
$$z_{ij} \overset{i.i.d}{\sim} Multinomial(\theta_j),\ i \in [1, N]$$

Therefore,

$$p(\Theta|\alpha) = \prod_{j=1}^{J} p(\theta_j|\alpha) \tag{40}$$

$$= \prod_{j=1}^{J} \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_{j,k}^{\alpha_k - 1} \tag{41}$$

$$p(Z|\Theta) = \prod_{j=1}^{J} \prod_{i=1}^{N} p(Z_{i,j} = k|\theta_j) \tag{42}$$

$$= \prod_{j=1}^{J} \prod_{i=1}^{N} \prod_{k=1}^{K} \theta_{j,k}^{1\{Z_{i,j}=k\}} \tag{43}$$

$$= \prod_{j=1}^{J} \prod_{k=1}^{K} \theta_{j,k}^{\sum_{i=1}^{N} 1\{Z_{i,j}=k\}} \tag{44}$$

$$= \prod_{j=1}^{J} \prod_{k=1}^{K} \theta_{j,k}^{\sigma_{j,k}} \tag{45}$$

(43) and (44) are doing the same thing, but they can be interpreted in different ways. In (43), we check if $Z_{ij} = k$ for each k, and repeat this for all words i in all the documents j. In (44), we counts the number of words with topic k for each k in each document j, and repeat this for all j. (45) follows if we define $\sigma_{j,k} \equiv \sum_{i=1}^{N} 1\{Z_{i,j} = k\}$. This is the number of words with topic k in a document j.

Therefore, (38) is expressed in the following way.

$$p(Z|\alpha) = \int_{\Theta} p(\Theta|\alpha) p(Z|\Theta) d\Theta \tag{46}$$

$$= \int_{\Theta} \Big(\prod_{j=1}^{J} \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_{j,k}^{\alpha_k - 1} \Big) \Big(\prod_{j=1}^{J} \prod_{k=1}^{K} \theta_{j,k}^{\sigma_{j,k}} \Big) d\Theta \tag{47}$$

$$= \prod_{j=1}^{J} \int_{\theta_j} \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_{j,k}^{\alpha_k + \sigma_{j,k} - 1} d\theta_j \tag{48}$$

$$= \prod_{j=1}^{J} \frac{B(\alpha + \sigma_j)}{B(\alpha)} \tag{49}$$

(49) follows because the integrant (except $\frac{1}{B(\alpha)}$) is the dirichlet kernel.

Next, we derive $p(W|Z, \beta)$. A similar result follows.

$$p(W|Z, \beta) = \int_{\Phi} p(W, \Phi|Z, \beta) \; d\Phi \tag{50}$$

$$= \int_{\Phi} p(\Phi|\beta) p(W|\Phi, Z) \; d\Phi \tag{51}$$

Recall that the model is

$$\phi_k \overset{i.i.d}{\sim} Dirichlet(\beta),\ k \in [1, K]$$
$$w_{ij} \overset{i.i.d}{\sim} Multinomial(\phi_k),\ i \in [1, N]$$

Therefore,

$$\begin{aligned} p(\Phi|\beta) &= \prod_{k=1}^{K} p(\phi_k|\beta) && (52) \\ &= \prod_{k=1}^{K} \frac{1}{B(\beta)} \prod_{v=1}^{V} \theta_{k,v}^{\beta_v - 1} && (53) \end{aligned}$$

$$\begin{aligned} p(W|\Phi, Z) &= \prod_{j=1}^{J} \prod_{i=1}^{N} p(W_{i,j} = v|\phi_k, Z_{i,j} = k) && (54) \\ &= \prod_{j=1}^{J} \prod_{i=1}^{N} \Big(\prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{k,v}^{1\{W_{ij}=v\}1\{Z_{ij}=k\}} \Big) && (55) \\ &= \prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{k,v}^{\sum_{j=1}^{J} \sum_{i=1}^{N} 1\{W_{ij}=v\}1\{Z_{ij}=k\}} && (56) \end{aligned}$$

Therefore, (50) is expressed in the following way.

$$\begin{aligned} p(W|Z, \beta) &= \int_{\Phi} p(\Phi|\beta) p(W|\Phi, Z)\ d\Phi \\ &= \int_{\Phi} \Big(\prod_{k=1}^{K} \frac{1}{B(\beta)} \prod_{v=1}^{V} \phi_{k,v}^{\beta_v - 1} \Big) \Big(\prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{k,v}^{\sum_{j=1}^{J} \sum_{i=1}^{N} 1\{W_{ij}=v\}1\{Z_{ij}=k\}} \Big) d\Phi \\ &= \prod_{k=1}^{K} \frac{B(\beta + \delta_k)}{B(\beta)} \end{aligned}$$

where $\delta_{v,k} = \sum_{j=1}^{J} \sum_{i=1}^{N} 1\{W_{ij} = v\}1\{Z_{ij} = k\}$, and δ_k is a vector of length V. $\delta_{v,k}$ is the number of unique word v with topic k in the entire corpus.

Therefore, the joint distribution of W and Z can be expressed as

$$p(W, Z|\alpha, \beta) = p(Z|\alpha, \beta) p(W|Z, \alpha, \beta) = \Big(\prod_{j=1}^{J} \frac{B(\alpha + \sigma_j)}{B(\alpha)} \Big) \Big(\prod_{k=1}^{K} \frac{B(\beta + \delta_k)}{B(\beta)} \Big) \qquad (57)$$

Now that we obtained the joint distribuion, let's omit the terms that do not change if we change $Z_{i^* j^*} = k^*$ because what we wanted to have is $p(Z_{i^* j^*} = k^* | Z_{\neg i^* j^*} W, \alpha, \beta)$. [9] Let us focus on the prior $p(Z|\alpha, \beta)$ first and then move on to the likelihood $p(W|Z, \alpha, \beta)$.

- Prior

[9] Be careful when we drop terms if the variable is discrete. Although the variable itself is Z, we want to focus on the case when $Z_{i^*, j^*} = k^*$. So, the terms we are supposed to omit may not contain Z_{i^*, j^*}, and instead, it may be written as a function of k^*. To see which variable is a function of $Z_{i^*, j^*} = k^*$ the diagram of data generating process (Figure 2) may be helpful.

First, we can omit $j' \neq j^*$ because we focus on the document j^*. We can also omit the denominator because it only depends on the hyperparameter.

$$\prod_{j=1}^{J} \frac{B(\alpha+\sigma_j)}{B(\alpha)} \propto \frac{B(\alpha+\sigma_j^*)}{B(\alpha)} \propto B(\alpha+\sigma_j^*)$$

Using the relaitonship between beta and gamma function,

$$B(\alpha+\sigma_j^*) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k+\sigma_{j^*,k})}{\Gamma(\sum_{k=1}^{K} \alpha_k+\sigma_{j^*,k})}$$

From here, we use ‘Gamma trick’. First, we isolate k^* from the iteration over k

$$\frac{\prod_{k=1}^{K} \Gamma(\alpha_k+\sigma_{j^*,k})}{\Gamma(\sum_{k=1}^{K} \alpha_k+\sigma_{j^*,k})} = \frac{\Gamma(\alpha_{k^*}+\sigma_{j^*,k^*}) \prod_{k\neq k^*}^{K} \Gamma(\alpha_k+\sigma_{j^*,k})}{\Gamma(\alpha_{k^*}+\sigma_{j^*,k^*}+\sum_{k\neq k^*}^{K}(\alpha_k+\sigma_{j^*,k}))}$$

Let’s define

$$\sigma_{j^*k}^{\neg} = \begin{cases} \sigma_{j^*,k}-1, & if\ k=k^* \\ \sigma_{j^*,k}, & else \end{cases}$$

Replacing σ_{j^*k} with $\sigma_{j^*k}^{\neg}$,

$$\frac{\Gamma(\alpha_{k^*}+\sigma_{j^*,k^*}) \prod_{k\neq k^*}^{K} \Gamma(\alpha_k+\sigma_{j^*,k})}{\Gamma(\alpha_{k^*}+\sigma_{j^*,k^*}+\sum_{k\neq k^*}^{K}(\alpha_k+\sigma_{j^*,k}))} = \frac{\Gamma(\alpha_{k^*}+\sigma_{j^*,k^*}^{\neg}+1) \prod_{k\neq k^*}^{K} \Gamma(\alpha_k+\sigma_{j^*,k}^{\neg})}{\Gamma(\alpha_{k^*}+\sigma_{j^*,k^*}^{\neg}+1+\sum_{k\neq k^*}^{K}(\alpha_k+\sigma_{j^*,k}^{\neg}))}$$

Using the fact that $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$,

$$\begin{aligned}\frac{\Gamma(\alpha_{k^*}+\sigma_{j^*,k^*}^{\neg}+1) \prod_{k\neq k^*}^{K} \Gamma(\alpha_k+\sigma_{j^*,k}^{\neg})}{\Gamma(\alpha_{k^*}+\sigma_{j^*,k^*}^{\neg}+1+\sum_{k\neq k^*}^{K}(\alpha_k+\sigma_{j^*,k}^{\neg}))} &= \frac{(\alpha_{k^*}+\sigma_{j^*,k^*}^{\neg})\Gamma(\alpha_{k^*}+\sigma_{j^*,k^*}^{\neg}) \prod_{k\neq k^*}^{K} \Gamma(\alpha_k+\sigma_{j^*,k}^{\neg})}{(\alpha_{k^*}+\sigma_{j^*,k^*}^{\neg}+\sum_{k\neq k^*}^{K}(\alpha_k+\sigma_{j^*,k}^{\neg}))\Gamma(\alpha_{k^*}+\sigma_{j^*,k^*}^{\neg}+\sum_{k\neq k^*}^{K}(\alpha_k+\sigma_{j^*,k}^{\neg}))} \\ &= \frac{(\alpha_{k^*}+\sigma_{j^*,k^*}^{\neg}) \prod_{k=1}^{K} \Gamma(\alpha_k+\sigma_{j^*,k}^{\neg})}{(\sum_{k=1}^{K}(\alpha_k+\sigma_{j^*,k}^{\neg}))\Gamma(\sum_{k=1}^{K}(\alpha_k+\sigma_{j^*,k}^{\neg}))}\end{aligned}$$

Notice that the terms except $\alpha_{k^*}+\sigma_{j^*,k^*}^{\neg}$ do not change if we chagne from $Z_{i^*j^*}=k^*$ to $Z_{i^*j^*}=k'$ because they iterate over all k anyways. But, let’s keep one term $\sum_{k=1}^{K}(\alpha_k+\sigma_{j^*,k}^{\neg})$ because it makes the interpretation clearer. As a result, we reached the following final expression for the prior

$$\frac{\alpha_{k^*}+\sigma_{j^*,k^*}^{\neg}}{\sum_{k=1}^{K}(\alpha_k+\sigma_{j^*,k}^{\neg})} = \frac{\alpha_{k^*}+\sigma_{j^*,k^*}-1}{\sum_{k=1}^{K}(\alpha_k+\sigma_{j^*,k})-1}$$

What does this expression means? Recall that σ_{j^*,k^*} is the number of words with the topic k^* in the document j^*. So, this expression is essentially the ratio of the words with topic k^* among all the topic assignment $k=1...K$ in the document j^*, adjusted by the hyperparameter α, and -1 represents the exclusion of the current word we are focusing on i^*. Having this expression in the posterior makes sense because if the documents have many words whose topic assignment is k^*, then it is more likely that a given word in such a document also have the topic k^*.

Part II

1. The Dirichlet is a conjugate prior of the multinomial

The Dirichelt distribution and the multinomial distribution are the key distributions for the latent Dirichlet allocation. First we review the key features of the two distributions, and confirm that the Dirichlet distribution is a conjugate prior to the multinomial distribution.

The multinomial distribution has the following pmf. If

$$X \sim Multinomial(\theta, n)$$

where $X = (X_1, ..., X_n)$, $\theta = (\theta_1, ...\theta_k)$. Then,

$$p(x|\theta, n) = \frac{n!}{\prod_{i=1}^{K} x_i!} \prod_{i=1}^{K} \theta_i^{x_i} \tag{58}$$

where $x = (x_1, ..., x_n)$, and $\sum_{i=1}^{K} x_i = n$.

The Dirichlet distribution has the following pdf. If

$$\theta \sim Dirichlet(\alpha)$$

Then,

$$p(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \tag{59}$$

where $\sum_{i=1}^{K} \theta_i = 1$ and $\theta_i \geq 0, \forall i$. (i.e. θ belongs to the $K - 1$ simplex.)

We can intuitively think how these distributions are related. We can think of the multinomial distribution as a result of n consecutive rolls of a die, where $X = (X_1, ..., X_6)$ is the number of each face showing up during the n rolls. Then, the Dirichlet distribution is the distribution over the probability of each face showing up in each roll. (If a die is fair, then $\theta = (1/6, ...1/6)$).

We can mathmatically show that the Dirichlet distribution is a conjugate prior of the multinomial distribution. The posterior of θ is

$$p(\theta|X, \alpha) \propto p(X|\theta, \alpha)p(\theta|\alpha)$$

Then, the RHS is

$$\begin{aligned}
&p(X|\theta, \alpha)p(\theta|\alpha) \\
&= \frac{n!}{\prod_{i=1}^{K} x_i!} \prod_{i=1}^{K} \theta_i^{x_i} \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \\
&\propto \prod_{i=1}^{K} \theta_i^{x_i + \alpha_i - 1}
\end{aligned}$$

This is the kernel of the Dirichlet pdf. So we find that the posterior is also the Dirichlet distribution. By adjusting the constant term, we find that

$$p(\theta|X) = \frac{1}{B(\alpha + x)} \prod_{i=1}^{K} \theta_i^{x_i+\alpha_i-1} \tag{60}$$

Therefore, the posterior distribution of θ is

$$\theta^{post} \sim Dirichlet(\alpha + x)$$

1.1 The relationship of the kernel of Dirichlet and Beta and Gamma functions

It is also helpful to see the integral of the kernel of Dirichlet is a Beta function, and therefore a ratio of gamma functions. This equation is useful in the derivation of the collapsed gibbs sampler.

$$\begin{aligned}
& p(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{\alpha_i-1} \\
\iff & 1 = \int_\theta p(\theta|\alpha) d\theta = \int_\theta \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{\alpha_i-1} d\theta \\
\iff & B(\alpha) = \int_\theta \prod_{i=1}^{K} \theta_i^{\alpha_i-1} d\theta \\
\iff & \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)} = \int_\theta \prod_{i=1}^{K} \theta_i^{\alpha_i-1} d\theta
\end{aligned}$$

2. The Dirichlet distribution belongs to the exponential family

The fact that the Dirichlet distribution belongs to the exponential family is useful. While the expression of the Dirichlet in terms of the exponential family is widely available, the derivation is not. The following sketches the conversion from the usual expression of the pdf of Dirichlet distribution to the form in terms of the exponential family.

The pdf of the Dirichlet distribution is

$$p(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{\alpha_i-1} \tag{61}$$

where $\theta = [\theta_1....\theta_K]^\intercal$, and $\sum_{i=1}^{K} \theta_i = 1$, $\theta_i \geq 0, \forall i \in \{1, 2, ..., K\}$.

The exponential family has the following general pdf

$$p(\theta|\eta) = h(\theta) \exp(\eta^\intercal t(\theta) - A(\eta)) \tag{62}$$

where $t(\theta)$ is the sufficient statistic, η is called the natural parameter, $A(\eta)$ is the log normalization factor, and $h(x)$ is the base measure.

It is known that the Dirichlet distribution belongs to the exponential family, and the parameters are

- The natural parameter: $\eta = \alpha = [\alpha_1 ... \alpha_K]^\intercal$
- The sufficient statistic: $t(\theta) = \log\theta = \log[\theta_1, ...\theta_K]^\intercal$.
- The log normalization factor: $A(\eta) = \sum_{i=1}^K \log\Gamma(\eta_i) - \log\Gamma(\sum_{i=1}^K \eta_i)$
- The base measure: $h(x) = \frac{1}{\prod_{i=1}^K}\theta_i$

The following sketches why the parameters are expressed in these ways.

First, recall that $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)})$. Then the pdf of the Dirichlet distribution is

$$\begin{aligned} p(\theta|\alpha) &= \frac{1}{B(\alpha)}\prod_{i=1}^K \theta_i^{\alpha_i - 1} \\ &= \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \cdot \frac{\prod_{i=1}^K \theta_i^{\alpha_i}}{\prod_{i=1}^K \theta_i} \end{aligned}$$

Taking the exponential and log,

$$\begin{aligned} p(\theta|\alpha) &= \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \cdot \frac{\prod_{i=1}^K \theta_i^{\alpha_i}}{\prod_{i=1}^K \theta_i} \\ &= \frac{1}{\prod_{i=1}^K \theta_i} \exp\Big[\log\frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)}\prod_{i=1}^K \theta_i^{\alpha_i}\Big] \\ &= \frac{1}{\prod_{i=1}^K \theta_i} \exp\Big[\log\Gamma(\sum_{i=1}^K \alpha_i) - \log\prod_{i=1}^K \Gamma(\alpha_i) + \log\prod_{i=1}^K \theta_i^{\alpha_i}\Big] \\ &= \frac{1}{\prod_{i=1}^K \theta_i} \exp\Big[\log\Gamma(\sum_{i=1}^K \alpha_i) - \sum_{i=1}^K \log\Gamma(\alpha_i) + \sum_{i=1}^K \alpha_i \log\theta_i\Big] \\ &= \frac{1}{\prod_{i=1}^K \theta_i} \exp\Big[\alpha^\intercal \log\theta - \big\{\sum_{i=1}^K \log\Gamma(\alpha_i) - \log\Gamma(\sum_{i=1}^K \alpha_i)\big\}\Big] \end{aligned}$$

Comparison of this expression with the general pdf of the exponential family yields the parameters listed above.

3. The expectation of the sufficient statistic of the exponential family is the derivative of the log normalizing factor

Note that this is useful in the derivation of the original variational inference for LDA (in Blei et al 2003).

Theorem:

$$\mathbb{E}_\theta[t(\theta)] = \frac{\partial}{\partial\eta}A(\eta) \tag{63}$$

Proof:

Using the fact that the pdf must integrate to one, we first express $a(\eta)$ by the other parameters.

$$
\begin{aligned}
1 = \int_\theta p(\theta|\eta) d\theta &= \int_\theta h(\theta)\exp(\eta\ t(\theta) - A(\eta))d\theta \\
\iff 1 &= \frac{1}{\exp A(\eta)} \int_\theta h(\theta)\exp(\eta\ t(\theta))d\theta \\
\iff \exp A(\eta) &= \int_\theta h(\theta)\exp(\eta\ t(\theta))d\theta \\
\iff A(\eta) &= \log \int_\theta h(\theta)\exp(\eta\ t(\theta))d\theta
\end{aligned}
$$

Let $g(\eta) = \frac{1}{\exp A(\eta)}$. Then this is also equivalent to

$$
1 = g(\eta) \int_\theta h(\theta)\exp(\eta\ t(\theta))d\theta \tag{64}
$$

Taking the derivative of Eq (3) with respect to η,

$$
\begin{aligned}
0 &= \frac{\partial}{\partial\eta}\Big[g(\eta)\int_\theta h(\theta)\exp(\eta\ t(\theta))d\theta\Big] \\
&= g'(\eta)\int_\theta h(\theta)\exp(\eta\ t(\theta))d\theta + g(\eta)\frac{\partial}{\partial\eta}\int_\theta h(\theta)\exp(\eta\ t(\theta))d\theta \\
&= \frac{g'(\eta)}{g(\eta)} + g(\eta)\int_\theta h(\theta)\exp(\eta\ t(\theta))t(\theta)d\theta \\
&= \frac{g'(\eta)}{g(\eta)} + \int_\theta g(\eta)h(\theta)\exp(\eta\ t(\theta))t(\theta)d\theta \\
&= \frac{g'(\eta)}{g(\eta)} + \mathbb{E}_\theta[t(\theta)]
\end{aligned}
$$

i.e.

$$
\mathbb{E}_\theta[t(\theta)] = -\frac{g'(\eta)}{g(\eta)} = -\frac{\partial}{\partial\eta}log(g(\eta)) \tag{65}
$$

because $\frac{\partial}{\partial\eta}log(g(\eta)) = \frac{1}{g(\eta)} \cdot \frac{\partial}{\partial\eta}g(\eta)$.

Because $g(\eta) = \frac{1}{\exp(A\eta))} = -\exp(A(\eta))$, Eq (4) is equivalent to

$$
\mathbb{E}[t(\theta)] = -\frac{\partial}{\partial\eta}log(g(\eta)) = \frac{\partial}{\partial\eta}A(\eta) \tag{66}
$$

4. Derivation steps for the collapsed Gibbs Sampler for the smoothed LDA

This section describes the steps for the derivation of the collapsed Gibbs Sampler for the smoothed LDA. It aims to fill the gaps of the lecture slides by Shiraito for POLSCI798. The model is specified in Part I, smoothed LDA. (Notation has changed. Need to fix.) Because the following equations involve many products, let us assume that the product term,$\prod$, ends at the ×.

Joint posterior distribution

$$
\begin{aligned}
& p(Z, \theta, \eta | W, \alpha, \beta) && (67)\\
& \propto \underbrace{p(\eta|\beta) p(\theta|\alpha)}_{\text{prior}} \times \underbrace{p(Z|\theta) p(W|Z, \eta)}_{\text{likelihood}} && (68)\\
& = \prod_{k=1}^{K} p(\eta_k|\beta) \times \prod_{j=1}^{J} p(\theta_j|\alpha) \times \prod_{j=1}^{J}\prod_{i=1}^{N_j} p(Z_{ij}|\theta_j) \times \prod_{j=1}^{J}\prod_{i=1}^{N_j} p(W_{ij}|Z_{ij}, \eta_j) && (69)\\
& = \prod_{k=1}^{K} p(\eta_k|\beta) \times \prod_{j=1}^{J} p(\theta_j|\alpha) \prod_{i=1}^{N_j} p(Z_{ij}|\theta_j) p(W_{ij}|Z_{ij}, \eta_j) && (70)\\
& && (71)
\end{aligned}
$$

Joint posterior distribution after collapsing θ and η

$$
\begin{aligned}
p(Z|W) & \propto \int_\theta \int_\eta p(Z, \theta, \eta | W) \; d\eta d\theta && (72)\\
& = \int_\theta \int_\eta \prod_{k=1}^{K} p(\eta_k|\beta) \times \prod_{j=1}^{J} p(\theta_j|\alpha) \prod_{i=1}^{N_j} p(Z_{ij}|\theta_j) \; d\eta d\theta && (73)\\
& = \int_\theta \prod_{j=1}^{J} p(\theta_j|\alpha) \prod_{i=1}^{N_j} p(Z_{ij}|\theta_j) \; d\theta \;\; \times \int_\eta \prod_{k=1}^{K} p(\eta_k|\beta) \times \prod_{j=1}^{J}\prod_{i=1}^{N_j} p(Z_j|\theta_j) p(W_{ij}|Z_j, \eta) \; d\eta && (74)
\end{aligned}
$$

Conditional posterior

$$
\begin{aligned}
& p(Z_{i^*j^*} = k | Z_{-i^*j^*}, W)\\
& \propto \int_{\theta_{j^*}} p(\theta_{j^*}|\alpha) \prod_{i=1}^{N_j} p(Z_{i^*j^*} = k, \theta_{j^*}) \prod_{i \neq i^*} p(Z_{ij}|\theta_{j^*}) \; d\theta_{j^*}\\
& \quad \times \int_{\eta_k} p(\eta_k|\beta) \times p(W_{i^*j^*}|Z_{i^*j^*} = k, \eta_k) \times \prod_{ij \neq i^*j^*, Z_{ij} = k} p(W_ij | Z_{ij} = k, \eta_k) \; d\eta_k
\end{aligned}
$$

NOTE: diagram of data generating process is helpful - insert here

Reference

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), pp.993-1022.

Lafferty, J.D. and Blei, D.M., 2006. Correlated topic models. In Advances in neural information processing systems (pp. 147-154).

Mcauliffe, J.D. and Blei, D.M., 2008. Supervised topic models. In Advances in neural information processing systems (pp. 121-128).

Geigle, C., 2016. Inference Methods for Latent Dirichlet Allocation.

Gelman, A., Stern, H.S., Carlin, J.B., Dunson, D.B., Vehtari, A. and Rubin, D.B., 2013. Bayesian data analysis. Chapman and Hall/CRC.

Will Kurt. A blog post about KL divergence

Lecture Slides by Yuki Shiraito, POLSCI798, 2019.