Derivation of Inference and Parameter Estimation Algorithm for Latent Dirichlet Allocation (LDA)

Kittipat "Bot" Kampa

June 16, 2010

Abstract

In this paper, I just want to show the details of how to derive the inference and parameter estimation algorithm for Latent Dirichlet Allocation (LDA) in [1]. So the best way to use this note is to read the original paper, and use this note when you get stuck in the technical details. Eventhough I carefully write the derivations, but this is the first version, so there might be some typos and mistakes. Please feel free to give any comments or suggestions via my e-mail address given below[1].

1 Joint distribution

The joint given the parameters are given by the term

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta) \tag{1}$$

and the word distribution given the topic is given by

$$\begin{aligned} p(w_n|z_n, \beta) &= \prod_{i=1}^{k}\prod_{j=1}^{V} p(w_n^j = 1|z_n^i = 1)^{w_n^j z_n^i} \\ &= \prod_{i=1}^{k}\prod_{j=1}^{V} \beta_{ij}^{w_n^j z_n^i} \end{aligned} \tag{2}$$

where $p(w_n^j = 1|z_n^i = 1) = \beta_{ij}$. For the topic distribution we can write it as $p(z_n^i = 1|\theta) = \theta_i$. Therefore, we can write

$$\begin{aligned} p(z_n|\theta) &= \prod_{i=1}^{k} p(z_n^i = 1|\theta)^{z_n^i} \\ &= \prod_{i=1}^{k} \theta_i^{z_n^i} \end{aligned} \tag{3}$$

[1] kittipat@gmail.com, Electrical and Computer Engineering Department, University of Florida, Gainesville, FL, USA.

2 Posterior

What we want is the posterior

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \tag{4}$$

The denominator is given by

$$\begin{aligned}
p(\mathbf{w}|\alpha, \beta) &= \int \sum_{\mathbf{z}} p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) d\theta \\
&= \int \sum_{\mathbf{z}} p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) d\theta \\
&= \int p(\theta|\alpha) \left(\sum_{\mathbf{z}} \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta \\
&= \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta \\
&= \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(w_n, z_n|\beta, \theta) \right) d\theta \\
&= \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} \prod_{i=1}^{k} \prod_{j=1}^{V} p(w_n^j = 1, z_n^i = 1|\beta, \theta)^{w_n^j z_n^i} \right) d\theta \\
&= \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n=1}^{k} \prod_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j z_n^i} \right) d\theta
\end{aligned} \tag{5}$$

and we can simplify the term by proving (in 7.2) that $\sum_{z_n=1}^{k} \prod_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j z_n^i} = \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j}$, so finally we will get that

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta \tag{6}$$

and in the corpus that there are D documents, which all is assumed independent. We will have that

$$\begin{aligned}
p(D|\alpha, \beta) &= \prod_{d=1}^{D} p(\mathbf{w}_d|\alpha, \beta) \\
&= \prod_{d=1}^{D} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_{dn}^j} \right) d\theta_d
\end{aligned} \tag{7}$$

3 Variational Inference

The coupling between θ and β is a big problem, and that is caused by the edges between z,θ and $\mathbf{w}$. So we will drop the edges and $\mathbf{w}$ to make the structure easier. Note that in the true posterior we have $p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$, so we will use the variational distribution of similar form that is $q(\theta, \mathbf{z}|\gamma, \phi)$ and neglect $\mathbf{w}$. For simplicity, we assume the variational distribution to be completely factorized which is given by

$$\begin{aligned}
q(\theta, \mathbf{z}|\gamma, \phi) &= q(\theta|\gamma) q(\mathbf{z}|\phi) \\
&= q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n),
\end{aligned}$$

where γ denotes the Dirichlet parameter and $\phi = (\phi_1, \ldots, \phi_N)$ denote multinomial parameters. We start with log likelihood of the document (the denominator in posterior term):

$$\begin{aligned}
\log p(\mathbf{w}|\alpha,\beta) &= \log \int \sum_{\mathbf{z}} p(\theta, \mathbf{z}, \mathbf{w}|\alpha,\beta) d\theta \\
&= \log \int \sum_{\mathbf{z}} \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha,\beta)}{q(\theta, \mathbf{z}|\gamma,\phi)} q(\theta, \mathbf{z}|\gamma,\phi) d\theta \\
&= \log E_{q(\theta,\mathbf{z}|\gamma,\phi)} \left[\frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha,\beta)}{q(\theta, \mathbf{z}|\gamma,\phi)} \right] \\
&\geq E_{q(\theta,\mathbf{z}|\gamma,\phi)} \left[\log \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha,\beta)}{q(\theta, \mathbf{z}|\gamma,\phi)} \right] \quad (8) \\
&= E_{q(\theta,\mathbf{z}|\gamma,\phi)} \left[\log p(\theta, \mathbf{z}, \mathbf{w}|\alpha,\beta) \right] \quad (9) \\
&\quad - E_{q(\theta,\mathbf{z}|\gamma,\phi)} \left[\log q(\theta, \mathbf{z}|\gamma,\phi) \right] \quad (10) \\
&= -KL(q(\theta, \mathbf{z}|\gamma,\phi)||p(\theta, \mathbf{z}, \mathbf{w}|\alpha,\beta)) \quad (11) \\
&= L(\gamma,\phi;\alpha,\beta) \quad (12)
\end{aligned}$$

That is $L(\gamma,\phi;\alpha,\beta)$ is a lower bound of the true log likelihood $\log p(\mathbf{w}|\alpha,\beta)$, and the difference (the "bound" gap) is given by

$$\begin{aligned}
\Delta(\gamma,\phi) &= \log p(\mathbf{w}|\alpha,\beta) - L(\gamma,\phi;\alpha,\beta) \quad (13) \\
&= \log p(\mathbf{w}|\alpha,\beta) - E_{q(\theta,\mathbf{z}|\gamma,\phi)} \left[\log \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha,\beta)}{q(\theta, \mathbf{z}|\gamma,\phi)} \right] \quad (14) \\
&= \log p(\mathbf{w}|\alpha,\beta) + E_{q(\theta,\mathbf{z}|\gamma,\phi)} \left[\log \frac{q(\theta, \mathbf{z}|\gamma,\phi)}{p(\theta, \mathbf{z}, \mathbf{w}|\alpha,\beta)} \right] \quad (15) \\
&= \log p(\mathbf{w}|\alpha,\beta) + E_{q(\theta,\mathbf{z}|\gamma,\phi)} \left[\log \frac{q(\theta, \mathbf{z}|\gamma,\phi)}{p(\theta, \mathbf{z}|\mathbf{w},\alpha,\beta) p(\mathbf{w}|\alpha,\beta))} \right] \quad (16) \\
&= E_{q(\theta,\mathbf{z}|\gamma,\phi)} \left[\log p(\mathbf{w}|\alpha,\beta) \right] + E_{q(\theta,\mathbf{z}|\gamma,\phi)} \left[\log \frac{q(\theta, \mathbf{z}|\gamma,\phi)}{p(\theta, \mathbf{z}|\mathbf{w},\alpha,\beta) p(\mathbf{w}|\alpha,\beta))} \right] \quad (17) \\
&= E_{q(\theta,\mathbf{z}|\gamma,\phi)} \left[\log \frac{q(\theta, \mathbf{z}|\gamma,\phi)}{p(\theta, \mathbf{z}|\mathbf{w},\alpha,\beta)} \right] \quad (18) \\
&= KL(q(\theta, \mathbf{z}|\gamma,\phi)||p(\theta, \mathbf{z}|\mathbf{w},\alpha,\beta)) \quad (19)
\end{aligned}$$

It's interesting but quite intuitive to know that the gap, $\Delta(\gamma,\phi)$, is actually KL divergence:

$$\Delta(\gamma,\phi) = \log p(\mathbf{w}|\alpha,\beta) - L(\gamma,\phi;\alpha,\beta) = KL(q(\theta, \mathbf{z}|\gamma,\phi)||p(\theta, \mathbf{z}|\mathbf{w},\alpha,\beta)) \quad (20)$$

, therefore, when we maximize the lower bound $L(\gamma,\phi;\alpha,\beta)$ w.r.t γ and ϕ, we automatically minimize the KL w.r.t. γ and ϕ as well. Another form that can be useful is

$$\log p(\mathbf{w}|\alpha,\beta) = KL(q(\theta, \mathbf{z}|\gamma,\phi)||p(\theta, \mathbf{z}|\mathbf{w},\alpha,\beta)) - KL(q(\theta, \mathbf{z}|\gamma,\phi)||p(\theta, \mathbf{z}, \mathbf{w}|\alpha,\beta)) \quad (21)$$

Now we will expand the lower bound L by using the factorization od p and q:

$$\begin{aligned}
L(\gamma,\phi;\alpha,\beta) &= E_q \left[\log p(\theta, \mathbf{z}, \mathbf{w}|\alpha,\beta) \right] - E_q \left[\log q(\theta, \mathbf{z}|\gamma,\phi) \right] \quad (22) \\
&= E_q \left[\log p(\theta|\alpha) \right] + E_q \left[\log p(\mathbf{z}|\theta) \right] + E_q \left[\log p(\mathbf{w}|\mathbf{z},\beta) \right] \quad (23) \\
&\quad - E_q \left[\log q(\theta|\gamma) \right] - E_q \left[\log q(\mathbf{z}|\phi) \right] \quad (24)
\end{aligned}$$

and with some details in 8, we obtain the lower bound as

$$
\begin{aligned}
L(\gamma,\phi;\alpha,\beta) &= \log\Gamma(\sum_{i=1}^{k}\alpha_i) - \sum_{i=1}^{k}\log\Gamma(\alpha_i) + \sum_{i=1}^{k}(\alpha_i-1)\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)\right) && (25)\\
&+ \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_n^i\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)\right) && (26)\\
&+ \sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{V} w_n^j\phi_n^i\log\beta_{ij} && (27)\\
&- \log\Gamma(\sum_{i=1}^{k}\gamma_i) + \sum_{i=1}^{k}\log\Gamma(\gamma_i) - \sum_{i=1}^{k}(\gamma_i-1)\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)\right) && (28)\\
&- \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_n^i\log\phi_n^i && (29)
\end{aligned}
$$

The goal here is to make an inference by maximizing the lower bound $L(\gamma,\phi;\alpha,\beta)$ w.r.t. all the variational parameters (γ,ϕ). The big picture of inference is that the true parameters (α,β) are assumed known, we are trying to adjust the variational parameters (γ,ϕ) so that the gap between the true posterior and the variational posterior is minimized. That is, we are trying to maximize the lower bound $L(\gamma,\phi;\alpha,\beta)$. A good way to do maximization is to take derivative w.r.t. each of the variational parameters.

Once we get the optimal variational parameters, we will plug the optimal parameters (γ^*,ϕ^*) back into the variational distribution $q(\theta,\mathbf{z}|\gamma,\phi)$, then we will get the distribution at each hidden variable. That is why this is an inference algorithm.

3.1 Inference for variational multinomial parameter ϕ

We will take derivative of $L(\gamma,\phi;\alpha,\beta)$ w.r.t. each of ϕ_n^i, so the term that does not contains ϕ_n^i will be eventually eliminated. So we will just pick the term containing ϕ_n^i to take derivative of:

$$L_{[\phi_n^i]} = \phi_n^i\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)\right) + \sum_{j=1}^{V} w_n^j\phi_n^i\log\beta_{ij} - \phi_n^i\log\phi_n^i \tag{30}$$

, but don't forget that w_n is observed, so we will know which index j, make $w_n^j = 1$. Let's say $w_n^v = 1$, and $w_n^j = 0$ for all $j \neq v$. We will have that

$$L_{[\phi_n^i]} = \phi_n^i\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)\right) + \phi_n^i\log\beta_{iv} - \phi_n^i\log\phi_n^i \tag{31}$$

, and remember that we have for each n a constrain on ϕ_n^i such that

$$\sum_{i=1}^{k}\phi_n^i = 1 \tag{32}$$

Consequently, again we will have the Lagrange multipliers $\lambda_n\left(\sum_{j=1}^{k}\phi_n^j - 1\right)$ for each n. When incorporate all the constrains, the objective function (Lagrangian) becomes

$$\mathcal{L}(\gamma,\phi,\lambda) = L(\gamma,\phi;\alpha,\beta) + \sum_{n'=1}^{k}\lambda_{n'}\left(\sum_{j=1}^{k}\phi_{n'}^j - 1\right) \tag{33}$$

we incorporate the constrain into the objective function using Lagrange multiplier:

$$\mathcal{L}_{[\phi_n^i]} = \phi_n^i \left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) \right) + \phi_n^i \log \beta_{iv} - \phi_n^i \log \phi_n^i + \lambda_n \left(\sum_{j=1}^{k} \phi_n^j - 1 \right) \tag{34}$$

, hence we will have

$$\frac{\partial \mathcal{L}}{\partial \phi_n^i} = \left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) \right) + \log \beta_{iv} - \log \phi_n^i - 1 + \lambda_n. \tag{35}$$

Then setting the derivative to zero yields the maximizing value of the variational paramater ϕ_n^i :

$$\log \phi_n^i = \left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) \right) + \log \beta_{iv} - 1 + \lambda_n \tag{36}$$

$$\phi_n^i = \exp(-1 + \lambda_n) \beta_{iv} \exp \left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) \right) \tag{37}$$

where, you will realize that, $\exp(-1 + \lambda_n)$ is just a constant, so it is more convenient to write the update equation using "$\propto$":[2]

$$\phi_n^i \propto \beta_{iv} \exp \left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) \right). \tag{38}$$

3.2 Inference for variational Dirichlet parameter γ

We follow the same process as in the previous section[3], except that we don't have any constrain on γ_i as it is just a parameter, not a probability distribution like ϕ_n^i. Therefore, we will have that

$$\begin{aligned}
L_{[\gamma]} &= \sum_{i=1}^{k} (\alpha_i - 1) \left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) \right) && (39)\\
&\quad + \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_n^i \left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) \right) && (40)\\
&\quad - \log \Gamma(\sum_{j=1}^{k} \gamma_j) + \log \Gamma(\gamma_i) - \sum_{i=1}^{k} (\gamma_i - 1) \left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) \right) && (41)\\
&= \sum_{i=1}^{k} \left(\alpha_i + \sum_{n=1}^{N} \phi_n^i - \gamma_i \right) \left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) \right) - \log \Gamma(\sum_{j=1}^{k} \gamma_j) + \log \Gamma(\gamma_i) \\
&= \sum_{i=1}^{k} \left(\alpha_i + \sum_{n=1}^{N} \phi_n^i - \gamma_i \right) \Psi(\gamma_i) && (42)\\
&\quad - \Psi(\sum_{j=1}^{k} \gamma_j) \sum_{i=1}^{k} \left(\alpha_i + \sum_{n=1}^{N} \phi_n^i - \gamma_i \right) && (43)
\end{aligned}$$

[2]Note that when using the proportional symbol, we don't have to pay attention to the normalization constant. Please see details in 7.4

[3]One good exercise here is to ask: In equ (39)-(41), as we want the term containing only γ_i, then why can't we just pick the term $(\alpha_i - 1) \left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) \right)$ instead of $\sum_{i=1}^{k} (\alpha_i - 1) \left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) \right)$? See details in 7.3

$$- \log \Gamma(\sum_{j=1}^{k} \gamma_j) + \log \Gamma(\gamma_i) \tag{44}$$

We take the derivative w.r.t. γ_i :

$$\begin{aligned}
\frac{\partial L}{\partial \gamma_i} &= \left(\alpha_i + \sum_{n=1}^{N} \phi_n^i - \gamma_i\right) \Psi'(\gamma_i) - \Psi(\gamma_i) \\
&\quad \Psi(\sum_{j=1}^{k} \gamma_j) - \Psi'(\sum_{j=1}^{k} \gamma_j) \sum_{i=1}^{k} \left(\alpha_i + \sum_{n=1}^{N} \phi_n^i - \gamma_i\right) \\
&\quad -\Psi(\sum_{j=1}^{k} \gamma_j) + \Psi(\gamma_i) \\
&= \left(\alpha_i + \sum_{n=1}^{N} \phi_n^i - \gamma_i\right) \Psi'(\gamma_i) \qquad (45) \\
&\quad -\Psi'(\sum_{j=1}^{k} \gamma_j) \sum_{j=1}^{k} \left(\alpha_j + \sum_{n=1}^{N} \phi_n^j - \gamma_j\right) \qquad (46)
\end{aligned}$$

Setting the derivative to zero, and solve for γ_i. It's pretty obvious to see that one good solution is to pick for each i,

$$\alpha_i + \sum_{n=1}^{N} \phi_n^i - \gamma_i = 0 \tag{47}$$

, but there are several choices of solution since γ_i's are coupled and cannot be separated. In order to get a more general solution, we may need some analysis:

- From [2]we know that $\Psi'(x) \geq 0$ for all x
- Normally the Dirichlet parameters $\alpha_1, ..., \alpha_k > 0$ [4]

but, you will not get a closed form solution from this approach either. So let's stick with the simple appraoch in equ (47) and we will have that

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_n^i \tag{48}$$

4 Parameter estimation

In this section, we will learn the model parameters α and β from a corpus $D = \{\mathbf{w}_1, \ldots, \mathbf{w}_M\}$ containing M documents. In the previous section we infer ϕ and γ as if α and β are known (fixed) by maximizing the lower bound $L(\gamma, \phi; \alpha, \beta)$ w.r.t. ϕ and γ. In this section, we will maximizing the lower bound w.r.t. the model parameters α and β as if ϕ and γ are known (fixed). Therefore, the big picture of the inference in section and parameter estimation in this section is like EM algorithm, so this type of algorithm is so called variational EM algorithm where

- E-step: maximizing the lower bound w.r.t. the variational parameters (ϕ and γ in this case) by fixing the model parameters (α and β in this case)
- M-step: maximizing the lower bound w.r.t. the model parameters (α and β in this case) by fixing the variational parameters (ϕ and γ in this case)

Each step will give a set of iterative equations. By running the iterative equations until converges, finally we can learn the parameters and obtain the distribution of each hidden variable at the same stroke. The overall procedure can be viewed as coordinate ascent in L.

Now we will proceed with parameter estimation. For simplicity, we assume the variational distribution for the corpus to be completely factorized as product of distribution of each document d which is given by

$$q(\theta, \mathbf{z}|\gamma, \phi) = \prod_{d=1}^{M} q(\theta_d, \mathbf{z}_d|\gamma_d, \phi_d) \tag{49}$$

[4]where

$$\begin{aligned} q(\theta_d, \mathbf{z}_d|\gamma_d, \phi_d) &= q(\theta_d|\gamma_d)q(\mathbf{z}_d|\phi_d) \\ &= q(\theta_d|\gamma_d) \prod_{n=1}^{N} q(z_{dn}|\phi_{dn}), \end{aligned}$$

We will do the similar process with the previous section by starting with the log likelihood of the corpus and assuming that each document d in the corpus D is independent, so we will get

$$\log p(\mathbf{w}|\alpha, \beta) = \sum_{d=1}^{D} \log p(\mathbf{w}_d|\alpha, \beta) \tag{50}$$

Note that the model parameters α and β does not depend on the document, but depends on the whole corpus, so there is no index d at the parameters. Remember that we abuse the notation a little bit: $\theta = \{\theta_1, \ldots, \theta_M\}$ and $\mathbf{z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_M\}$ and we will have

$$\begin{aligned} \log p(\mathbf{w}|\alpha, \beta) &\geq E_{q(\theta,\mathbf{z}|\gamma,\phi)}\left[\log \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{q(\theta, \mathbf{z}|\gamma, \phi)}\right] \\ &= E_{q(\theta,\mathbf{z}|\gamma,\phi)}\left[\log p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)\right] \\ &\quad -E_{q(\theta,\mathbf{z}|\gamma,\phi)}\left[\log q(\theta, \mathbf{z}|\gamma, \phi)\right] \\ &= E_{q(\theta,\mathbf{z}|\gamma,\phi)}\left[\sum_d \log p(\theta_d, \mathbf{z}_d, \mathbf{w}_d|\alpha, \beta)\right] \\ &\quad -E_{q(\theta,\mathbf{z}|\gamma,\phi)}\left[\sum_d \log q(\theta_d, \mathbf{z}_d|\gamma_d, \phi_d)\right] \\ &= \sum_d \Big(E_{q(\theta,\mathbf{z}|\gamma,\phi)}\left[\log p(\theta_d, \mathbf{z}_d, \mathbf{w}_d|\alpha, \beta)\right] - E_{q(\theta,\mathbf{z}|\gamma,\phi)}\left[\log q(\theta_d, \mathbf{z}_d|\gamma_d, \phi_d)\right]\Big) \\ &= \sum_d \Big(E_{q(\theta_d,\mathbf{z}_d|\gamma_d,\phi_d)}\left[\log p(\theta_d, \mathbf{z}_d, \mathbf{w}_d|\alpha, \beta)\right] - E_{q(\theta_d,\mathbf{z}_d|\gamma_d,\phi_d)}\left[\log q(\theta_d, \mathbf{z}_d|\gamma_d, \phi_d)\right]\Big) \\ &= -\sum_d KL(q(\theta_d, \mathbf{z}_d|\gamma_d, \phi_d)||p(\theta_d, \mathbf{z}_d, \mathbf{w}_d|\alpha, \beta)) \\ &= \sum_d L(\gamma_d, \phi_d; \alpha, \beta) \end{aligned} \tag{51}$$

What we learn from here is that the lower bound of the corpus D is the summation of the lower bound of each document d. Note that in the previous section L represents the lower bound for each document d, but here we abuse the notation L to represent the overall lower bound. That is

$$L(\gamma, \phi; \alpha, \beta) = \sum_d L(\gamma_d, \phi_d; \alpha, \beta) \tag{52}$$

[4]I don't understand why don't we say γ does not depend on the document d? γ is analogous to α, so it should be the parameter of the whole corpus, not of each document?

Hence, we will have the overall lower bound:

$$
\begin{aligned}
L(\gamma,\phi;\alpha,\beta) &= \sum_{d=1}^{M}\left[\log\Gamma(\sum_{i=1}^{k}\alpha_i)-\sum_{i=1}^{k}\log\Gamma(\alpha_i)+\sum_{i=1}^{k}(\alpha_i-1)\left(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj})\right)\right] && (53)\\
&+\sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\phi_{dn}^{i}\left(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj})\right) && (54)\\
&+\sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\sum_{j=1}^{V}w_{dn}^{j}\phi_{dn}^{i}\log\beta_{ij} && (55)\\
&\sum_{d=1}^{M}\left[-\log\Gamma(\sum_{i=1}^{k}\gamma_{di})+\sum_{i=1}^{k}\log\Gamma(\gamma_{di})-\sum_{i=1}^{k}(\gamma_{di}-1)\left(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj})\right)\right] && (56)\\
&-\sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\phi_{dn}^{i}\log\phi_{dn}^{i} && (57)
\end{aligned}
$$

In the next step, we will take derivative w.r.t. model parameters α and β.

4.1 Conditional Multinomials parameter β

Recall that $\beta_{ij}=p(w_n^j=1|z_n^i=1)$, therefore there is a constrain $\sum_{j=1}^{V}\beta_{ij}=1$ for each i. Consequently, again we will have the Lagrange multipliers $\lambda_i\left(\sum_{j=1}^{V}\beta_{ij}-1\right)$ for each i. When incorporate all the constrains, the objective function (Lagrangian) becomes

$$\mathcal{L}(\alpha,\beta,\lambda)=L(\gamma,\phi;\alpha,\beta)+\sum_{i'=1}^{k}\lambda_{i'}\left(\sum_{j=1}^{V}\beta_{i'j}-1\right) \tag{58}$$

$$\mathcal{L}_{[\beta_{ij}]}=\sum_{d=1}^{M}\sum_{n=1}^{N_d}w_{dn}^{j}\phi_{dn}^{i}\log\beta_{ij}+\lambda_i\left(\sum_{j=1}^{V}\beta_{ij}-1\right) \tag{59}$$

and we will take derivative of $\mathcal{L}$ w.r.t. each β_{ij}:

$$\frac{\partial\mathcal{L}}{\partial\beta_{ij}}=\left(\frac{1}{\beta_{ij}}\right)\sum_{d=1}^{M}\sum_{n=1}^{N_d}w_{dn}^{j}\phi_{dn}^{i}+\lambda_i \tag{60}$$

Setting the derivative to zero:

$$\beta_{ij}=\left(\frac{1}{-\lambda_i}\right)\sum_{d=1}^{M}\sum_{n=1}^{N_d}w_{dn}^{j}\phi_{dn}^{i} \tag{61}$$

and, again, we prefer to write

$$\beta_{ij}\propto\sum_{d=1}^{M}\sum_{n=1}^{N_d}w_{dn}^{j}\phi_{dn}^{i} \tag{62}$$

4.2 Dirichlet parameter α

$$L_{[\alpha_i]}=\sum_{d=1}^{M}\left[\log\Gamma(\sum_{j=1}^{k}\alpha_j)-\log\Gamma(\alpha_i)+(\alpha_i-1)\left(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj})\right)\right] \tag{63}$$

We will take derivative of $L_{[\alpha_i]}$ w.r.t. α_i:

$$\frac{\partial L}{\partial \alpha_i} = \sum_{d=1}^{M}\left[\Psi(\sum_{j=1}^{k}\alpha_j) - \Psi(\alpha_i) + \Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k}\gamma_{dj})\right] \tag{64}$$

$$= M\left(\Psi(\sum_{j=1}^{k}\alpha_j) - \Psi(\alpha_i)\right) + \sum_{d=1}^{M}\left[\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k}\gamma_{dj})\right] \tag{65}$$

By setting the derivative to zero, we obtain the update equation. However, in this case, α_i coupled with other α_j's : $j \neq i$, that is, α_i depends on other α_j's. So we won't get a closed form solution from this derivative. Instead, we must use an iterative method to find a stationary point of α_i. Newton-Raphson algorithm is a good candidate, but it requires us to calculate the Hessian of L:

$$\frac{\partial L}{\partial \alpha_i \alpha_j} = M\left(\Psi'(\sum_{j=1}^{k}\alpha_j) - \delta(i,j)\Psi'(\alpha_i)\right) \tag{66}$$

[5]There is a trick to make linear-time Newton-Raphson algorithm described in the original paper [1].

5 Summary

The variational EM algorithm is described as follows:

- E-step: (inference on variational parameters)

$$\phi_n^i \propto \beta_{iv}\exp\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)\right) \tag{67}$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N}\phi_n^i \tag{68}$$

- M-step: (estimating model parameters)

$$\beta_{ij} \propto \sum_{d=1}^{M}\sum_{n=1}^{N_d} w_{dn}^j \phi_{dn}^i \tag{69}$$

$$\alpha_{new} = \alpha_{old} - H(\alpha_{old})^{-1} g(\alpha_{old}) \tag{70}$$

We will start with E-step, please read Figure 6. in [1]. We will iterate in E-step until convergence, then go to M-step. Then iterate in M-step until convergence, then go to E-step again. All the parameters except α can be calculated in one stroke as they have closed form solution. In contrast, α don't have a closed from solution, instead its value relies on iterative process like Newton-Raphson.

[5]I did not get the same form as in the original paper [1] though

6 Discussion and Questions

1. Why do the authors approximate the distribution q like that? I have an intuitive idea that the approximate distribution should be setup such that all the hidden variables that we want to know in posterior (θ and $\mathbf{z}$) are factorized (decoupling). That's because when they are decoupled we can deal with each expectation term separately which make the computation tractable. A good example would be in section 8.2, you will see that choosing q completely factorized can make things much easier.

2. Is there any better way to come up with such an approximate distribution q?

3. Why don't we minimize $KL(q(\theta, \mathbf{z}|\gamma, \phi)||p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta))$ instead of maximize L? Is there is anything to do with generative or discriminative model?

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[2] Eric W. Weisstein. Trigamma function. from mathworld–a wolfram web resource., May 2010. `http://mathworld.wolfram.com/TrigammaFunction.html`.

[3] Wikipedia.org. Digamma function, June 2010. `http://en.wikipedia.org/wiki/Digamma_function`.

[4] Wikipedia.org. Dirichlet distribution, May 2010. `http://en.wikipedia.org/wiki/Dirichlet_distributic`

[5] Wikipedia.org. Exponential family, May 2010. `http://en.wikipedia.org/wiki/Exponential_family`.

7 Appendix A: Interesting Techniques

There are some interesting techniques in this paper that can be useful for other applications.

7.1 Interchangeable summation and product form

One trick we learn from here is that if we make a binary random variable as an indicator, then we can interchange between product and sum. This can be very useful when taking log to the term. Suppose that it's originally a summation which, of course, log cannot go through, but if we change the summation to product then log can be distributed inside the summation

$$\prod_{i=1}^{k} \theta_i^{z_n^i} = \sum_{i=1}^{k} z_n^i \theta_i$$

For example, from equ (2)

$$p(w_n|z_n, \beta) = \prod_{i=1}^{k} \prod_{j=1}^{V} \beta_{ij}^{w_n^j z_n^i}$$

Actually another way to write this is

$$p(w_n|z_n,\beta) = \sum_{i=1}^{k}\sum_{j=1}^{V} w_n^j z_n^i \beta_{ij}$$

since w_n^j and z_n^i are binary random variables, and, at the end, there will be only one β_{ij} survives.

7.2 Sum of Product in binary random variable

Let $\prod_{j=1}^{V}(\theta_i\beta_{ij})^{w_n^j} = Y_i$ and we will have that

$$\begin{aligned}
\sum_{z_n=1}^{k}\prod_{i=1}^{k}\prod_{j=1}^{V}(\theta_i\beta_{ij})^{w_n^j z_n^i} &= \sum_{z_n=1}^{k}\prod_{i=1}^{k} Y_i^{z_n^i} && (71)\\
&= \sum_{z_n=1}^{k} Y_1^{z_n^1} Y_2^{z_n^2}\ldots Y_k^{z_n^k} && (72)\\
&= Y_1^{z_n^1=1} Y_2^{z_n^2=0}\ldots Y_k^{z_n^k=0} && (73)\\
&\quad + Y_1^{z_n^1=0} Y_2^{z_n^2=1}\ldots Y_k^{z_n^k=0} && (74)\\
&\quad + \ldots && (75)\\
&\quad + Y_1^{z_n^1=0} Y_2^{z_n^2=0}\ldots Y_k^{z_n^k=1} && (76)\\
&= Y_1^{z_n^1=1} + Y_2^{z_n^2=1} + \ldots + Y_k^{z_n^k=1} && (77)\\
&= Y_1 + Y_2 + \ldots + Y_k && (78)\\
&= \sum_{i=1}^{k} Y_i && (79)\\
&= \sum_{i=1}^{k}\prod_{j=1}^{V}(\theta_i\beta_{ij})^{w_n^j} && (80)
\end{aligned}$$

7.3 Taking derivative of a function of summation

One good exercise here is to ask: In equ (39)-(41), as we want the term containing only γ_i, then why can't we just pick the term $(\alpha_i-1)\left(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{k}\gamma_j)\right)$ instead of $\sum_{i=1}^{k}(\alpha_i-1)\left(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{k}\gamma_j)\right)$?

The answer is that we cannot neglect the summation because we have the term $\Psi(\sum_{j=1}^{k}\gamma_j)$ which includes all $i\in\{1,\ldots,k\}$. Therefore, we will have that

$$\frac{\partial}{\partial\gamma_i}(\alpha_i-1)\left(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{k}\gamma_j)\right) \neq \frac{\partial}{\partial\gamma_i}\sum_{i=1}^{k}(\alpha_i-1)\left(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{k}\gamma_j)\right) \quad (81)$$

, and that's why we cannot neglect the summation. So it's good to remember that we have to be more careful than usual when taking derivative w.r.t. γ_i on the term like $\mathcal{F}(\sum_{j=1}^{k}\gamma_j)$.

7.4 Using "$\propto$" instead of "=" on update equations derived from constraints

Notice that the constrain occurs when dealing with parameters representing probability like ϕ and β. In that case, the objective function involves Lagrange multipliers, and we usually use the proportional

symbol "$\propto$" in the update equations due to the normalization factor which is usually the function of the Lagrange multiplier. However, in the the normal parameters not representing probability like α and γ, there is no need for the constrain...so it's easier. For example, in 3.1 we have

$$\phi_n^i = \exp(-1+\lambda_n)\,\beta_{iv}\exp\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)\right)$$

where, you will realize that, $\exp(-1+\lambda_n)$ is just a constant, and to be more precise, it is the normalization constant:

$$\exp(-1+\lambda_n) = \frac{1}{\sum_{i'=1}^{k}\beta_{i'v}\exp\left(\Psi(\gamma_{i'}) - \Psi(\sum_{j=1}^{k}\gamma_j)\right)}$$

Therefore it's more convenient to write

$$\phi_n^i \propto \beta_{iv}\exp\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)\right).$$

Note that when using the proportional symbol, we don't have to pay attention to the normalization constant. Instead, we will just have to evaluate, for each $i \in \{1,\ldots,k\}$, unnormalized ϕ_n^i denoted by $\hat{\phi}_n^i$:

$$\hat{\phi}_n^i = \beta_{iv}\exp\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)\right)$$

and after that we can simply evaluate the normalized, ϕ_n^i by

$$\phi_n^i = \frac{\hat{\phi}_n^i}{\sum_{i'=1}^{k}\hat{\phi}_n^{i'}}$$

8 Appendix B: Expectations

8.1 Computing $E_q[\log p(\theta|\alpha)]$

Start with

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1} \tag{82}$$

and taking log, we will get

$$\log p(\theta|\alpha) = \sum_{i=1}^{k}(\alpha_i-1)\log\theta_i + \log\Gamma(\sum_{i=1}^{k}\alpha_i) - \sum_{i=1}^{k}\log\Gamma(\alpha_i) \tag{83}$$

Take the expectation w.r.t. $q(\theta,\mathbf{z}|\gamma,\phi)$ (note that q is a function of only θ and $\mathbf{z}$) and we will get

$$\langle\log p(\theta|\alpha)\rangle_q = \sum_{i=1}^{k}(\alpha_i-1)\langle\log\theta_i\rangle_q + \log\Gamma(\sum_{i=1}^{k}\alpha_i) - \sum_{i=1}^{k}\log\Gamma(\alpha_i). \tag{84}$$

In order to calculate the expectation $\langle\log\theta_i\rangle_q$, we will exploit a general fact in exponential family[6].

[6] A good reference for this section is [5, 3]

Recall that a distribution in exponential family can be written in the form:

$$p(x|\eta) = h(x) \exp\left\{\eta^T T(x) - A(\eta)\right\},$$

and Dirichlet is in exponential family as it has the form

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

where the exponential family parameters are given by:

- The sufficient statistic: $T(\theta_i) = \log \theta_i$
- $h(\theta) = 1$
- The natural parameter: $\eta_i = \alpha_i - 1$
- The normalization factor $A(\eta) = -\log \Gamma(\sum_{i=1}^{k} \alpha_i) + \sum_{i=1}^{k} \log \Gamma(\alpha_i)$

In the section "Differential identities for cumulants" in [5], we will find that

$$E_p[T(\theta_i)] = \frac{\partial A(\eta)}{\partial \eta_i}, \tag{85}$$

and so, we will have that

$$\begin{aligned}
E_{p(\theta|\alpha)}[\log \theta_i] &= \frac{\partial}{\partial(\alpha_i - 1)} \left[-\log \Gamma(\sum_{j=1}^{k} \alpha_j) + \sum_{j=1}^{k} \log \Gamma(\alpha_j)\right] && (86)\\
&= \frac{\partial}{\partial \alpha_i} \left[-\log \Gamma(\sum_{j=1}^{k} \alpha_j) + \sum_{j=1}^{k} \log \Gamma(\alpha_j)\right] && (87)\\
&= -\frac{1}{\Gamma(\sum_{j=1}^{k} \alpha_j)} \frac{\partial \Gamma(\sum_{j=1}^{k} \alpha_j)}{\partial \left(\sum_{j=1}^{k} \alpha_j\right)} + \frac{1}{\Gamma(\alpha_i)} \frac{\partial \Gamma(\alpha_i)}{\partial \alpha_i} && (88)\\
&= -\Psi(\sum_{j=1}^{k} \alpha_j) + \Psi(\alpha_i) && (89)
\end{aligned}$$

where Ψ is the digamma function. In order to take logarithmic derivative of the the gamma function, please refer to [3].

Now we will use equ. (89) to calculate the expectation $\langle \log \theta_i \rangle_q$:

$$\begin{aligned}
\langle \log \theta_i \rangle_{q(\theta, \mathbf{z}|\gamma, \phi)} &= \langle \log \theta_i \rangle_{q(\theta|\gamma)} && (90)\\
&= -\Psi(\sum_{j=1}^{k} \gamma_j) + \Psi(\gamma_i) && (91)
\end{aligned}$$

Therefore, the final expectation will be

$$E_q\left[\log p(\theta|\alpha)\right] = \sum_{i=1}^{k} (\alpha_i - 1) \left(-\Psi(\sum_{j=1}^{k} \gamma_j) + \Psi(\gamma_i)\right) + \log \Gamma(\sum_{i=1}^{k} \alpha_i) - \sum_{i=1}^{k} \log \Gamma(\alpha_i) \tag{92}$$

8.2 Computing $E_q[\log p(\mathbf{z}|\theta)]$

$$p(\mathbf{z}|\theta) = \prod_{n=1}^{N} p(z_n|\theta) \tag{93}$$

$$= \prod_{n=1}^{N}\prod_{i=1}^{k} \theta_i^{z_n^i} \tag{94}$$

$$\log p(\mathbf{z}|\theta) = \sum_{n=1}^{N}\sum_{i=1}^{k} z_n^i \log\theta_i \tag{95}$$

$$E_q[\log p(\mathbf{z}|\theta)] = \sum_{n=1}^{N}\sum_{i=1}^{k} \left\langle z_n^i \log\theta_i \right\rangle_q \tag{96}$$

$$= \sum_{n=1}^{N}\sum_{i=1}^{k} \left\langle z_n^i \right\rangle_{q(\mathbf{z}|\phi)} \left\langle \log\theta_i \right\rangle_{q(\theta|\gamma)} \tag{97}$$

In order to calculate the expectation terms on RHS, we have to know the followings:

$$q(z_n|\phi) = \prod_{i=1}^{k} \left(\phi_n^i\right)^{z_n^i} \tag{98}$$

where $q(z_n^i = 1|\phi) = \phi_n^i$. That leads to

$$\left\langle z_n^i \right\rangle_{q(\mathbf{z}|\phi)} = \left\langle z_n^i \right\rangle_{q(z_n|\phi)} \tag{99}$$

$$= \sum_{z_n} z_n^i q(z_n|\phi) \tag{100}$$

$$= \sum_{z_n} z_n^i \prod_{j=1}^{k} \left(\phi_n^j\right)^{z_n^j} \tag{101}$$

$$= \phi_n^i \tag{102}$$

Finally, we will get that

$$E_q[\log p(\mathbf{z}|\theta)] = \sum_{n=1}^{N}\sum_{i=1}^{k} \phi_n^i \left(-\Psi(\sum_{j=1}^{k}\gamma_j) + \Psi(\gamma_i)\right) \tag{103}$$

8.3 Computing $E_q[\log p(\mathbf{w}|\mathbf{z},\beta)]$

From that $p(w_n|z_n,\beta) = \prod_{i=1}^{k}\prod_{j=1}^{V} \beta_{ij}^{w_n^j z_n^i}$, and we will get that

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

$$= \prod_{n=1}^{N}\prod_{i=1}^{k}\prod_{j=1}^{V} \beta_{ij}^{w_n^j z_n^i}$$

Hence we will have

$$E_q\left[\log p(\mathbf{w}|\mathbf{z},\beta)\right] = \sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{V} w_n^j \left\langle z_n^i \right\rangle_{q(\mathbf{z}|\phi)} \log \beta_{ij} \tag{104}$$

$$= \sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{V} w_n^j \phi_n^i \log \beta_{ij} \tag{105}$$

8.4 Computing $E_q\left[\log q(\theta|\gamma)\right]$

From form of $q(\theta|\gamma)$ is determined by Dirichlet distribution, so it will be

$$q(\theta|\gamma) = \frac{\Gamma(\sum_{i=1}^{k}\gamma_i)}{\prod_{i=1}^{k}\Gamma(\gamma_i)}\theta_1^{\gamma_1-1}\cdots\theta_k^{\gamma_k-1} \tag{106}$$

and taking log, we will get

$$\log q(\theta|\gamma) = \sum_{i=1}^{k}(\gamma_i-1)\log\theta_i + \log\Gamma(\sum_{i=1}^{k}\gamma_i) - \sum_{i=1}^{k}\log\Gamma(\gamma_i) \tag{107}$$

when taking expectation we will get

$$E_q\left[\log q(\theta|\gamma)\right] = \sum_{i=1}^{k}(\gamma_i-1)\left\langle\log\theta_i\right\rangle_{q(\theta|\gamma)} + \log\Gamma(\sum_{i=1}^{k}\gamma_i) - \sum_{i=1}^{k}\log\Gamma(\gamma_i) \tag{108}$$

$$= \sum_{i=1}^{k}(\gamma_i-1)\left(-\Psi(\sum_{j=1}^{k}\gamma_j) + \Psi(\gamma_i)\right) + \log\Gamma(\sum_{i=1}^{k}\gamma_i) - \sum_{i=1}^{k}\log\Gamma(\gamma_i) \tag{109}$$

8.5 Computing $E_q\left[\log q(\mathbf{z}|\phi)\right]$

From form of $q(\mathbf{z}|\phi)$ is determined by multinomial distribution, so it will be

$$q(z_n^i = 1|\phi) = \phi_n^i \tag{110}$$

$$q(z_n|\phi) = \prod_{i=1}^{k}\left[\phi_n^i\right]^{z_n^i} \tag{111}$$

$$q(\mathbf{z}|\phi) = \prod_{n=1}^{N} q(z_n|\phi) \tag{112}$$

$$= \prod_{n=1}^{N}\prod_{i=1}^{k}\left[\phi_n^i\right]^{z_n^i} \tag{113}$$

Hence, we will have the expectation:

$$E_q\left[\log q(\mathbf{z}|\phi)\right] = \sum_{n=1}^{N}\sum_{i=1}^{k}\left\langle z_n^i\right\rangle_{q(\mathbf{z}|\phi)}\log\phi_n^i \tag{114}$$

$$= \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_n^i\log\phi_n^i \tag{115}$$