Multi-Modal Latent Dirichlet Allocation

Nanbo Sun

May 25, 2018

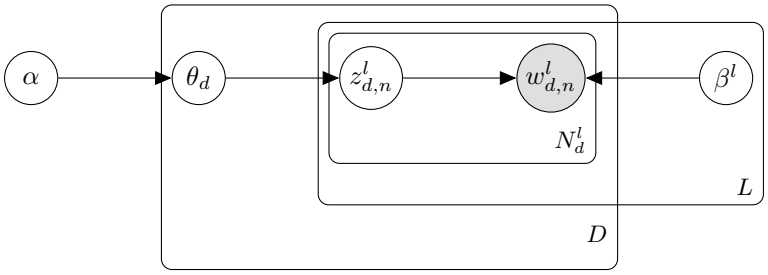

Figure 1: Multi-Modal LDA model.

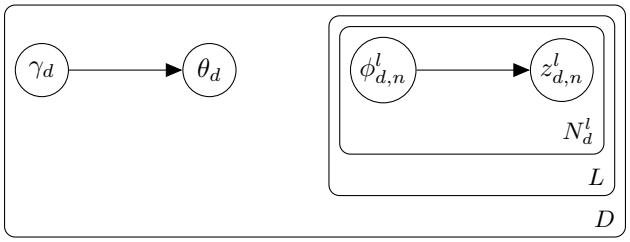

Figure 2: Variational distribution.

1 Generative Process

For patient $d = 1...D$ with K factors and V voxels,

1. draw a factor mixture $\theta_d \sim \text{Dir}(\alpha)$, where θ_d is a K-vector, and $\sum_k \theta_{dk} = 1$;
2. for each of the modality $l = 1...L$
3. for each of the voxel count $n = 1...N^l_d$ independently
   (a) draw a factor $z^l_{dn} \sim \text{Mult}(\theta_d)$;
   (b) draw a voxel $w^l_{dn} \sim p\left(w^l_{dn} \mid z^l_{dn}, \beta^l\right)$, where β^l is a $K \times V$ matrix, each row of which defines a multinomial distribution over all the voxels. β^l_{ij} is the probability that voxel j appears given factor i.

2 Constructing the Lower Bound

From Figure 2, the variational distribution used to approximate the true posterior is factorizable as

$$q(\theta, \mathbf{z} \mid \gamma, \phi) = q(\theta \mid \gamma) \prod_{l=1}^{L} \prod_{n=1}^{N^l} q\left(z_n^l \mid \phi_n^l\right).$$

The lower bound $\mathcal{L}(\gamma, \phi \mid \alpha, \beta)$ of the single-document[1] log likelihood $\log p(\mathbf{w} \mid \alpha, \beta)$ is computed using Jensen's inequality as follows

$$\begin{aligned}
\log p(\mathbf{w} \mid \alpha, \beta) &= \log \int \sum_{\mathbf{z}} p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) \, \mathrm{d}\theta \\
&= \log \int \sum_{\mathbf{z}} \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) \, q(\theta, \mathbf{z} \mid \gamma, \phi)}{q(\theta, \mathbf{z} \mid \gamma, \phi)} \mathrm{d}\theta \\
&= \log \int \sum_{\mathbf{z}} q(\theta, \mathbf{z} \mid \gamma, \phi) \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{q(\theta, \mathbf{z} \mid \gamma, \phi)} \mathrm{d}\theta \\
&= \log E_q \left\{ \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{q(\theta, \mathbf{z} \mid \gamma, \phi)} \right\} \\
&\geq E_q \{\log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)\} - E_q \{\log q(\theta, \mathbf{z} \mid \gamma, \phi)\} \\
&\triangleq \mathcal{L}(\gamma, \phi \mid \alpha, \beta).
\end{aligned} \tag{1}$$

The difference between the log likelihood and its lower bound can be proven to be in fact the KL divergence between the variational distribution and the true posterior.

$$\begin{aligned}
&\log p(\mathbf{w} \mid \alpha, \beta) - \mathcal{L}(\gamma, \phi \mid \alpha, \beta) \\
&= E_q \{\log p(\mathbf{w} \mid \alpha, \beta)\} - E_q \{\log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)\} + E_q \{\log q(\theta, \mathbf{z} \mid \gamma, \phi)\} \\
&= E_q \left\{ \log \frac{p(\mathbf{w} \mid \alpha, \beta) \, q(\theta, \mathbf{z} \mid \gamma, \phi)}{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)} \right\} \\
&= E_q \left\{ \log \frac{q(\theta, \mathbf{z} \mid \gamma, \phi)}{p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)} \right\} \\
&= \mathrm{D}_{\mathrm{KL}}(q(\theta, \mathbf{z} \mid \gamma, \phi) \parallel p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)).
\end{aligned}$$

Therefore, maximizing the lower bound is equivalent to minimizing the KL divergence $\mathrm{D}_{\mathrm{KL}}(q(\theta, \mathbf{z} \mid \gamma, \phi) \parallel p(\theta, \mathbf{z} \mid \mathbf{w}$ That is, the variational distribution automatically approaches to the real posterior as we maximize the lower bound.

[1] This also explains why the document subscript is dropped for simplicity hereafter.

3 Expanding the Lower Bound

To maximize the lower bound, we first need to spell out the lower bound $\mathcal{L}(\gamma, \phi \mid \alpha, \beta)$ in terms of the model parameters (α, β) and the variational parameters (γ, ϕ). Continuing from (1), we have

$$\begin{aligned}
\mathcal{L}(\gamma, \phi \mid \alpha, \beta) &= E_q\{\log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)\} - E_q\{\log q(\theta, \mathbf{z} \mid \gamma, \phi)\} \\
&= E_q\left\{\log \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{q(\theta, \mathbf{z} \mid \gamma, \phi)}\right\} \\
&= E_q\left\{\log \frac{p(\theta \mid \alpha)\, p(\mathbf{z} \mid \theta)\, p(\mathbf{w} \mid \mathbf{z}, \beta)}{q(\theta \mid \gamma)\, q(\mathbf{z} \mid \phi)}\right\} \qquad (2) \\
&= E_q\{\log p(\theta \mid \alpha)\} + E_q\{\log p(\mathbf{z} \mid \theta)\} + E_q\{\log p(\mathbf{w} \mid \mathbf{z}, \beta)\} \\
&\quad - E_q\{\log q(\theta \mid \gamma)\} - E_q\{\log q(\mathbf{z} \mid \phi)\}.
\end{aligned}$$

We now further expand each of the five terms in (2).

The first term is

$$\begin{aligned}
E_q\{\log p(\theta \mid \alpha)\} &= E_q\left\{\log \frac{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_k^{\alpha_k - 1}\right\} \\
&= E_q\left\{\log \Gamma\left(\sum_{i=1}^K \alpha_i\right) + \sum_{k=1}^K (\alpha_k - 1)\log \theta_k - \sum_{k=1}^K \log \Gamma(\alpha_k)\right\} \\
&= \sum_{k=1}^K (\alpha_k - 1) E_q\{\log \theta_k\} + \log \Gamma\left(\sum_{i=1}^K \alpha_i\right) - \sum_{k=1}^K \log \Gamma(\alpha_k) \\
&= \sum_{k=1}^K (\alpha_k - 1)\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^K \gamma_i\right)\right) + \log \Gamma\left(\sum_{i=1}^K \alpha_i\right) - \sum_{k=1}^K \log \Gamma(\alpha_k),
\end{aligned}$$

where $\Psi(\cdot)$ is the digamma function, the first derivative of the log Gamma function. The final line is due to the following property of the Dirichlet distribution as a member of the exponential family. If $\theta \sim Dir(\alpha)$, then $E_{p(\theta|\alpha)}\{\log \theta_i\} = \Psi(\alpha_i) - \Psi(\sum_{i=1}^K \alpha_i)$.

The second term is

$$\begin{aligned}
E_q\{\log p(\mathbf{z} \mid \theta)\} &= E_q\left\{\log \prod_{l=1}^L \prod_{n=1}^N p\left(z_n^l \mid \theta\right)\right\} \\
&= E_q\left\{\log \prod_{l=1}^L \prod_{n=1}^N \prod_{k=1}^K \theta_k^{\mathbb{1}_z(n,k)}\right\} \\
&= \sum_{l=1}^L \sum_{n=1}^N \sum_{k=1}^K E_q\{\mathbb{1}_z(n,k) \log \theta_k\} \\
&= \sum_{l=1}^L \sum_{n=1}^N \sum_{k=1}^K E_q\{\mathbb{1}_z(n,k)\}\, E_q\{\log \theta_k\} \\
&= \sum_{l=1}^L \sum_{n=1}^N \sum_{k=1}^K \phi_{n,k}^l \left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^K \gamma_i\right)\right),
\end{aligned}$$

where $\phi_{n,k}$ is the probability of the nth word being produced by topic k, and $\mathbb{1}(\cdot)$ is the indicator function.

We expand **the third term** as

$$\begin{aligned}
E_q\left\{\log p\left(\mathbf{w} \mid \mathbf{z}, \beta\right)\right\} &= E_q\left\{\log \prod_{l=1}^{L} \prod_{n=1}^{N} p\left(w_n^l \mid z_n^l, \beta^l\right)\right\} \\
&= E_q\left\{\log \prod_{l=1}^{L} \prod_{n=1}^{N} \prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{k,v}^{\mathbb{1}_z(n,k)\mathbb{1}_w(n,v)}\right\} \\
&= E_q\left\{\sum_{l=1}^{L} \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{v=1}^{V} \mathbb{1}_z(n,k)\mathbb{1}_w(n,v) \log \beta_{k,v}\right\} \\
&= \sum_{l=1}^{L} \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{v=1}^{V} E_q\left\{\mathbb{1}_z(n,k)\right\} \mathbb{1}_w(n,v) \log \beta_{k,v} \\
&= \sum_{l=1}^{L} \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{v=1}^{V} \phi_{n,k}^l \mathbb{1}_{w^l}(n,v) \log \beta_{k,v}^l.
\end{aligned}$$

Very similar to the first term, **the fourth term** is expanded as

$$E_q\left\{\log q\left(\theta \mid \gamma\right)\right\} = \sum_{k=1}^{K} (\gamma_k - 1)\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K} \gamma_i\right)\right) + \log \Gamma\left(\sum_{k=1}^{K} \gamma_k\right) - \sum_{k=1}^{K} \log \Gamma\left(\gamma_k\right).$$

Finally, **the fifth term** is expanded as

$$\begin{aligned}
E_q\left\{\log q\left(\mathbf{z} \mid \phi\right)\right\} &= E_q\left\{\log \prod_{l=1}^{L} \prod_{n=1}^{N} q\left(z_n^l \mid \phi_n^l\right)\right\} \\
&= E_q\left\{\log \prod_{l=1}^{L} \prod_{n=1}^{N} \prod_{k=1}^{K} \phi_{n,k}^{\mathbb{1}_z(n,k)}\right\} \\
&= \sum_{l=1}^{L} \sum_{n=1}^{N} \sum_{k=1}^{K} E_q\left\{\mathbb{1}_z(n,k)\right\} \log \phi_{n,k} \\
&= \sum_{l=1}^{L} \sum_{n=1}^{N} \sum_{k=1}^{K} \phi_{n,k}^l \log \phi_{n,k}^l.
\end{aligned}$$

Therefore, the fully expanded lower bound is

$$\begin{aligned}
\mathcal{L}\left(\gamma, \phi \mid \alpha, \beta\right) = &\sum_{k=1}^{K} (\alpha_k - 1)\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K} \gamma_i\right)\right) + \log \Gamma\left(\sum_{i=1}^{K} \alpha_i\right) - \sum_{k=1}^{K} \log \Gamma\left(\alpha_k\right) \\
&+ \sum_{l=1}^{L} \sum_{n=1}^{N} \sum_{k=1}^{K} \phi_{n,k}^l \left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K} \gamma_i\right)\right) \\
&+ \sum_{l=1}^{L} \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{v=1}^{V} \phi_{n,k}^l \mathbb{1}_w(n,v) \log \beta_{k,v}^l
\end{aligned} \tag{3}$$

$$
\begin{aligned}
&- \sum_{k=1}^{K} (\gamma_k - 1) \left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K} \gamma_i \right) \right) - \log \Gamma \left(\sum_{k=1}^{K} \gamma_k \right) + \sum_{k=1}^{K} \log \Gamma \left(\gamma_k \right) \\
&- \sum_{l=1}^{L} \sum_{n=1}^{N} \sum_{k=1}^{K} \phi_{n,k}^{l} \log \phi_{n,k}^{l}.
\end{aligned}
$$

4 Maximizing the Lower Bound

In this section, we maximize the lower bound w.r.t. the variational parameters ϕ and γ. Recall that as the maximization runs, the KL divergence between the variational distribution and the true posterior drops (E-step of the variational EM algorithm).

4.1 Variational Multinomial

We first maximize Equation (3) w.r.t. $\phi_{n,k}$. Since $\sum_{k=1}^{K} \phi_{n,k} = 1$, this is a constrained optimization problem that can be solved by the Lagrange multiplier method. The Lagrangian w.r.t. $\phi_{n,k}$ is

$$
\mathcal{L}_{[\phi_{n,k}]} = \phi_{n,k} \left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K} \gamma_i \right) \right) + \phi_{n,k} \log \beta_{k,v} - \phi_{n,k} \log \phi_{n,k} + \lambda_n \left(\sum_{i=1}^{K} \phi_{n,i} - 1 \right),
$$

where λ_n is the Lagrange multiplier. Taking the derivative, we get

$$
\frac{\partial}{\partial \phi_{n,k}} \mathcal{L}_{[\phi_{n,k}]} = \Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K} \gamma_i \right) + \log \beta_{k,v} - \log \phi_{n,k} - 1 + \lambda_n.
$$

Setting this derivative to zero yields

$$
\begin{aligned}
\phi_{n,k}^{l} &= \beta_{k,v}^{l} \exp \left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K} \gamma_i \right) + \lambda_n - 1 \right) \\
&\propto \beta_{k,v}^{l} \exp \left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K} \gamma_i \right) \right).
\end{aligned}
$$

4.2 Variational Dirichlet

Now we maximize Equation (3) w.r.t. γ_k, the kth component of the Dirichlet parameter. Only the terms containing γ_k are retained.

$$
\begin{aligned}
\mathcal{L}_{[\gamma]} = &\sum_{k=1}^{K} (\alpha_k - 1) \left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K} \gamma_i \right) \right) \\
&+ \sum_{n=1}^{N} \phi_{n,k} \left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K} \gamma_i \right) \right) \\
&- \sum_{k=1}^{K} (\gamma_k - 1) \left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K} \gamma_i \right) \right) - \log \Gamma \left(\sum_{i=1}^{K} \gamma_i \right) + \sum_{k=1}^{K} \log \Gamma \left(\gamma_k \right)
\end{aligned}
$$

Taking the derivative w.r.t. γ_k, we have

$$\begin{aligned}
\frac{\partial}{\partial \gamma_k}\mathcal{L}_{[\gamma]} &= \left(\Psi'(\gamma_k) - \Psi'\left(\sum_{i=1}^{K}\gamma_i\right)\right)(\alpha_k - 1) \\
&+ \left(\Psi'(\gamma_k) - \Psi'\left(\sum_{i=1}^{K}\gamma_i\right)\right)\sum_{n=1}^{N}\phi_{n,k} \\
&- \left(\Psi'(\gamma_k) - \Psi'\left(\sum_{i=1}^{K}\gamma_i\right)\right)(\gamma_k - 1) - \left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right) \\
&- \frac{\Psi\left(\sum_{i=1}^{K}\gamma_i\right)}{\Gamma\left(\sum_{i=1}^{K}\gamma_i\right)} + \frac{\Psi(\gamma_k)}{\Gamma(\gamma_k)} \\
&= \left(\Psi'(\gamma_k) - \Psi'\left(\sum_{i=1}^{K}\gamma_i\right)\right)\left(\alpha_k + \sum_{n=1}^{N}\phi_{n,k} - \gamma_k\right) - \Psi(\gamma_k) + \Psi\left(\sum_{i=1}^{K}\gamma_i\right) \\
&- \Psi\left(\sum_{i=1}^{K}\gamma_i\right) + \Psi(\gamma_k) \\
&= \left(\Psi'(\gamma_k) - \Psi'\left(\sum_{i=1}^{K}\gamma_i\right)\right)\left(\alpha_k + \sum_{n=1}^{N}\phi_{n,k} - \gamma_k\right).
\end{aligned}$$

Setting it to zero, we have

$$\gamma_k = \alpha_k + \sum_{l=1}^{L}\sum_{n=1}^{N}\phi_{n,k}^{l}.$$

5 Estimating Model Parameters

The previous section is the E-step of the variational EM algorithm, where we tighten the lower bound w.r.t. the variational parameters; this section is the M-step, where we maximize the lower bound w.r.t. the model parameters α and β. Now we add back the document subscript to consider the whole corpus.

By the assumed exchangeability of the documents, the overall log likelihood of the corpus is just the sum of all the documents' log likelihoods, and the overall lower bound is just the sum of the individual lower bounds. Again, only the terms involving β are left in the overall lower bound. Adding the Lagrange multipliers, we obtain

$$\mathcal{L}_{[\beta]} = \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{k=1}^{K}\sum_{v=1}^{V}\phi_{d,n,k}\mathbb{1}_w(d,n,v)\log\beta_{k,v} + \sum_{k=1}^{K}\lambda_k\left(\sum_{v=1}^{V}\beta_{k,v} - 1\right).$$

Taking the derivative w.r.t. $\beta_{k,v}$ and setting it to zero, we have

$$\frac{\partial}{\partial \beta_{k,v}}\mathcal{L}_{[\beta]} = \sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{d,n,k}\mathbb{1}_w(d,n,v)\frac{1}{\beta_{k,v}} + \lambda_k = 0$$

$$\Rightarrow \beta_{k,v} = -\frac{1}{\lambda_k} \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{d,n,k} \mathbb{1}_w(d,n,v)$$

$$\Rightarrow \beta_{k,v}^{l} \propto \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{d,n,k}^{l} \mathbb{1}_{w^l}(d,n,v).$$

Similarly, for α, we have

$$\mathcal{L}_{[\alpha]} = \sum_{d=1}^{D} \left(\sum_{k=1}^{K} (\alpha_k - 1) \left(\Psi(\gamma_{d,k}) - \Psi\left(\sum_{i=1}^{K} \gamma_{d,i} \right) \right) + \log \Gamma \left(\sum_{i=1}^{K} \alpha_i \right) - \sum_{k=1}^{K} \log \Gamma(\alpha_k) \right)$$

$$\begin{aligned} \frac{\partial}{\partial \alpha_k} \mathcal{L}_{[\alpha]} &= \sum_{d=1}^{D} \left(\Psi(\gamma_{d,k}) - \Psi\left(\sum_{i=1}^{K} \gamma_{d,i} \right) + \Psi\left(\sum_{i=1}^{K} \alpha_i \right) - \Psi(\alpha_k) \right) \\ &= \sum_{d=1}^{D} \left(\Psi(\gamma_{d,k}) - \Psi\left(\sum_{i=1}^{K} \gamma_{d,i} \right) \right) + D \left(\Psi\left(\sum_{i=1}^{K} \alpha_i \right) - \Psi(\alpha_k) \right). \end{aligned}$$

Since the derivative also depends on other $\alpha_{k' \neq k}$, we compute the Hessian

$$\frac{\partial^2}{\partial \alpha_k \partial \alpha_{k'}} \mathcal{L}_{[\alpha]} = D\Psi'\left(\sum_{i=1}^{K} \alpha_i \right) - D\delta(k - k')\Psi(\alpha_k),$$

and notice that its form allows for the linear-time Newton-Raphson algorithm.