Technical Note Series **Spring 2013**

Note 1: Varitional Methods for Latent Dirichlet Allocation

Version 1.0 *Wayne Xin Zhao (batmanfly@gmail.com)*

Disclaimer: *The focus of this note was to reorganize the content in the original Blei's paper and add more detailed derivations. For convenience, in some part, I fully copied Blei's content. I hope it can help the beginners to vEM of LDA.*

First of all, let us make some claims about the parameters and variables in the model.

Let K be the number of topics, D be the number of documents and V be the number of terms in the vocabulary. We use i to index a topic [1], d to index a document [2], n index a word [3] and w (or v) to denote a word. In LDA, $\boldsymbol{\alpha}_{K\times 1}$ and $\boldsymbol{\beta}_{K\times V}$ are *model parameters*, while $\boldsymbol{\theta}_{D\times K}$ and $\boldsymbol{z}$ [4] are *hidden variables*.

As a variational distribution $q(\cdot)$, we use a fully factorized model, where all the variables are independently governed by a different distribution,

$$q(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{\phi}) = q(\boldsymbol{\theta}|\boldsymbol{\gamma})q(\boldsymbol{z}|\boldsymbol{\phi}), \tag{1.1}$$

where Dirichlet parameter $\boldsymbol{\gamma}_{D\times K}$ and the multinomial parameters $\boldsymbol{\phi}$ [5] are *variational parameters*. The topic assignments of words and the documents are exchangeable, i.e., conditionally independent on the parameters (either model parameters or variational parameters). Note that all the variational distributions $q(\cdot)$ here are conditional distributions and should be written as $q(\cdot|\boldsymbol{w})$, for simplicity we write it as $q(\cdot)$.

The main idea is that we use variational expectation-maximization (EM): In the E-step variational EM, we use the variational approximation to the posterior described in the previous section and find the optimal values of variational parameters. In the M-step, we maximize the bound with respect to the model parameters. In a more condense way, we perform variational inference for learning variational parameters in E-step while perform parameter estimation in M-step. These two steps alternate in a iteration. We will optimize the lower bound w.r.t variational parameters and model parameters one by one, and this is to perform optimization using a coordinate ascent algorithm.

1.1 Variational objective function

1.1.1 Finding a lower bound for $\log p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})$

Jensens inequality. Let X be a random variable, and f is a convex function. Then we have $f(E(X)) \leq E(f(x))$. If f is a concave function we have $f(E(X)) \geq E(f(x))$.

[1] $\sum_i$ means $\sum_{i=1}^{K}$

[2] $\sum_d$ means $\sum_{d=1}^{D}$

[3] $\sum_n$ means $\sum_{n=1}^{N_d}$, where N_d is the length of current document

[4] It is represented as a three dimension matrix, each entry is indexed by a triplet $< d, n, i >$, indicating whether the topic assignment of the nth word in the dth document is the ith topic.

[5] Corresponding to $\boldsymbol{z}$, it is represented as a three dimension matrix, each entry is indexed by a triplet $< d, n, i >$, indicating the probability of the nth word in the dth document in the ith topic.

We use Jensens inequality to bound the log probability of ***a document***[6],

$$\begin{aligned}
& \log p(\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta}) \\
= & \log \int \sum_{\boldsymbol{z}} p(\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta}) d\boldsymbol{\theta}, \\
= & \log \int \sum_{\boldsymbol{z}} \frac{p(\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta}) q(\boldsymbol{\theta},\boldsymbol{z})}{q(\boldsymbol{\theta},\boldsymbol{z})} d\boldsymbol{\theta}, \\
\geq & \int \sum_{\boldsymbol{z}} q(\boldsymbol{\theta},\boldsymbol{z}) \log p(\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta}) d\boldsymbol{\theta} - \int \sum_{\boldsymbol{z}} q(\boldsymbol{\theta},\boldsymbol{z}) \log q(\boldsymbol{\theta},\boldsymbol{z}) d\boldsymbol{\theta}, \\
= & \int \sum_{\boldsymbol{z}} q(\boldsymbol{\theta},\boldsymbol{z}) \log p(\boldsymbol{\theta}|\boldsymbol{\alpha}) d\boldsymbol{\theta} + \int \sum_{\boldsymbol{z}} q(\boldsymbol{\theta},\boldsymbol{z}) \log p(\boldsymbol{z}|\boldsymbol{\theta}) d\boldsymbol{\theta} + \\
& \int \sum_{\boldsymbol{z}} q(\boldsymbol{\theta},\boldsymbol{z}) \log p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta}) d\boldsymbol{\theta} - \int \sum_{\boldsymbol{z}} q(\boldsymbol{\theta},\boldsymbol{z}) \log q(\boldsymbol{\theta},\boldsymbol{z}) d\boldsymbol{\theta}, \\
= & E_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + E_q[\log p(\boldsymbol{z}|\boldsymbol{\theta})] + E_q[\log p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta})] + H(q),
\end{aligned}$$

We introduce a function to denote the right side of the last line

$$L(\boldsymbol{\gamma},\boldsymbol{\phi}|\boldsymbol{\alpha},\boldsymbol{\beta}) = E_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + E_q[\log p(\boldsymbol{z}|\boldsymbol{\theta})] + E_q[\log p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta})] + H(q). \tag{1.2}$$

We can easily verify that

$$\log p(\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta}) = L(\boldsymbol{\gamma},\boldsymbol{\phi}|\boldsymbol{\alpha},\boldsymbol{\beta}) + D(q(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})||p(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{w})). \tag{1.3}$$

We indeed find a lower bound for $\log p(\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta})$, i.e., $L(\boldsymbol{\gamma},\boldsymbol{\phi}|\boldsymbol{\alpha},\boldsymbol{\beta})$. This shows that maximizing the lower bound $L(\boldsymbol{\gamma},\boldsymbol{\phi}|\boldsymbol{\alpha},\boldsymbol{\beta})$ with respect to $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ is equivalent to minimizing the KL divergence between the variational posterior probability and the true posterior probability, the optimization problem presented earlier in Eq. 1.5. [7]

1.1.2 Expanding the lower bound

In the lower bound defined above, we have four items to specify. We will present a detailed derivations for them next.

For the first item, we have

$$\begin{aligned}
& E_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] \\
= & E_q[\log \left(\exp\left\{ \left(\sum_{i=1}^{K} (\alpha_i - 1) \log \theta_i \right) + \log \Gamma(\sum_{i=1}^{K} \alpha_i) - \sum_{i=1}^{K} \log \Gamma(\alpha_i) \right\} \right)], \\
= & \left(\sum_{i=1}^{K} (\alpha_i - 1) E_q[\log \theta_i] \right) + \log \Gamma(\sum_{i=1}^{K} \alpha_i) - \sum_{i=1}^{K} \log \Gamma(\alpha_i), \\
= & \left(\sum_{i=1}^{K} (\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j} \gamma_j)) \right) + \log \Gamma(\sum_{i=1}^{K} \alpha_i) - \sum_{i=1}^{K} \log \Gamma(\alpha_i).
\end{aligned}$$

[6] Currently, we ignore the document index d, since all the hidden variables and variational parameters are document-specific. But we will use the document index d explicitly in the part of parameter estimation since these model parameters are related to all the documents.

[7] When we learn the variational parameters, we fix all the model parameters. So that $\log p(\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta})$ can be considered as a fixed value and it is the sum of the lower bound and the KL divergence.

Let $\boldsymbol{z}_n$ denote the topic assignment of the nth word in current document ,and it is a vector. $z_{n,i} = 1$ when the topic assignment is the ith topic, otherwise, $z_{n,i} = 0$.

$$\begin{aligned}
E_q[\log p(\boldsymbol{z}|\boldsymbol{\theta})] &= \sum_n E_q[\log p(\boldsymbol{z}_n|\boldsymbol{\theta})], \\
&= \sum_{n,i} E_q[\log p(z_{n,i}|\theta_i)], \\
&= \sum_{n,i} E_q[\log \theta_i^{z_{n,i}}], \\
&= \sum_{n,i} E_q[z_{n,i} \log \theta_i], \\
&= \sum_{n,i} E_q[z_{n,i}] E_q[\log \theta_i], \\
&= \sum_{n,i} \phi_{n,i} \Big(\Psi(\gamma_i) - \Psi(\sum_{i=1}^{K} \gamma_i) \Big).
\end{aligned}$$

For the third item, we have

$$\begin{aligned}
E_q[\log p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\beta})] &= \sum_n E_q[\log p(w_n|z_n, \boldsymbol{\beta})] \\
&= \sum_{n,i} E_q[\log \beta_{i,w_n}^{z_{n,i}}] \\
&= \sum_{n,i} \phi_{n,i} \log \beta_{i,w_n}.
\end{aligned}$$

$$\begin{aligned}
H(q) &= -\int \sum_{\boldsymbol{z}} q(\boldsymbol{\theta}, \boldsymbol{z}) \log q(\boldsymbol{\theta}, \boldsymbol{z}) d\boldsymbol{\eta}, \\
&= -\int q(\boldsymbol{\theta}) \log q(\boldsymbol{\theta}) d\boldsymbol{\theta} - \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log q(\boldsymbol{z}), \\
&= -\Big(\sum_{i=1}^{K} (\gamma_i - 1) E_q[\log \theta_i] \Big) - \log \Gamma(\sum_{i=1}^{K} \gamma_i) + \sum_{i=1}^{K} \log \Gamma(\gamma_i) - \sum_{n,i} \phi_{n,i} \log \phi_{n,i}, \\
&= -\Big(\sum_{i=1}^{K} (\gamma_i - 1) \Big(\Psi(\gamma_i) - \Psi(\sum_{i=1}^{K} \gamma_i) \Big) \Big) - \log \Gamma(\sum_{i=1}^{K} \gamma_i) + \sum_{i=1}^{K} \log \Gamma(\gamma_i) - \sum_{n,i} \phi_{n,i} \log \phi_{n,i},
\end{aligned}$$

Having these detailed derivations of these four items, we can have an expanded formulation for the lower bound

$$
\begin{aligned}
& L(\boldsymbol{\gamma}, \boldsymbol{\phi} | \boldsymbol{\alpha}, \boldsymbol{\beta}) \qquad (1.4)\\
= & \left(\sum_{i=1}^{K} (\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_j \gamma_j)) \right) + \log \Gamma(\sum_{i=1}^{K} \alpha_i) - \sum_{i=1}^{K} \log \Gamma(\alpha_i) \\
& + \sum_{n,i} \phi_{n,i} \left(\Psi(\gamma_i) - \Psi(\sum_{i=1}^{K} \gamma_i) \right) \\
& + \sum_{n,i} \phi_{n,i} \log \beta_{i,w_n} \\
& - \left(\sum_{i=1}^{K} (\gamma_i - 1) \left(\Psi(\gamma_i) - \Psi(\sum_{i=1}^{K} \gamma_i) \right) \right) - \log \Gamma(\sum_{i=1}^{K} \gamma_i) + \sum_{i=1}^{K} \log \Gamma(\gamma_i) - \sum_{n,i} \phi_{n,i} \log \phi_{n,i}.
\end{aligned}
$$

1.2 Variational inference

The aim of variational inference is to learn the values of variational parameters $\boldsymbol{\gamma}, \boldsymbol{\phi}$. With the learnt variational parameters, we can evaluate the posterior probabilities of hidden variables. Having specified a simplified family of probability distributions, the next step is to set up an optimization problem that determines the values of the variational parameters: $\boldsymbol{\gamma}, \boldsymbol{\phi}$. We can obtain a solution for these variational variables by solve the following optimization problem:

$$
(\boldsymbol{\gamma}^*, \boldsymbol{\phi}^*) = \arg\min_{\boldsymbol{\gamma}, \boldsymbol{\phi}} D(q(\boldsymbol{\theta}, \boldsymbol{z} | \boldsymbol{\gamma}, \boldsymbol{\phi}) || p(\boldsymbol{\theta}, \boldsymbol{z} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w})). \qquad (1.5)
$$

With Eq. 1.3, we can achieve minimize $D(q(\boldsymbol{\theta}, \boldsymbol{z} | \boldsymbol{\gamma}, \boldsymbol{\phi}) || p(\boldsymbol{\theta}, \boldsymbol{z} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}))$ by maximizing the lower bound $L(\boldsymbol{\gamma}, \boldsymbol{\phi} | \boldsymbol{\alpha}, \boldsymbol{\beta})$. **Note that in this section, variational parameters are learnt in terms of a document, so we ignore the document index d here.**

1.2.1 Learning the variational parameters

We have expanded each item of the lower bound $L(\boldsymbol{\gamma}, \boldsymbol{\phi} | \boldsymbol{\alpha}, \boldsymbol{\beta})$ in Eq. 1.2. Then we maximize the bound with respect to the variational parameters: $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$.

We first maximize Eq. 1.4 with respect to $\phi_{n,i}$, the probability that the nth word is generated by latent topic i. Observe that this is a constrained maximization since $\sum_i \phi_{n,i} = 1$, so we incorporate Lagrange Multipliers λ_ns for that

$$
\begin{aligned}
& L_{\phi_{n,i}} \qquad (1.6)\\
= & \sum_{n,i} \phi_{n,i} \Big(\Psi(\gamma_i) - \Psi(\sum_{i=1}^{K} \gamma_i) \Big) \\
& + \sum_{n,i} \phi_{n,i} \log \beta_{i,w_n} \\
& - \sum_{n,i} \phi_{n,i} \log \phi_{n,i} \\
& + \lambda_n (\sum_i \phi_{n,i} - 1).
\end{aligned}
$$

Taking the derivative of the $L_{\phi_{n,i}}$

$$
\begin{aligned}
& dL_{\phi_{n,i}}/d\phi_{n,i} \qquad (1.7)\\
= & \Big(\Psi(\gamma_i) - \Psi(\sum_{i=1}^{K} \gamma_i) \Big) \\
& + \log \beta_{i,w_n} \\
& - \log \phi_{n,i} - 1 \\
& + \lambda_n .
\end{aligned}
$$

By setting $dL_{[\gamma]}/d\gamma_i$ to zero, we can get

$$\phi_{n,i} = \beta_{i,w_n} \exp(\Psi(\gamma_i) - \Psi(\sum_{i=1}^{K} \gamma_i) + \lambda_n),$$

where we do not need to compute λ_n for $\phi_{n,i}$ since λ_n is the same for all is. Thus we have

$$\phi_{n,i} \propto \beta_{i,w_n} \exp(\Psi(\gamma_i)).$$

Next, we maximize Eq. 1.4 with respect to γ_i, the ith component of the posterior Dirichlet parameter. The terms containing γ_is are:

$$
\begin{aligned}
& L_{[\gamma]} \\
= & \Big(\sum_{i=1}^{K} (\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_j \gamma_j)) \Big) \\
& + \sum_{n,i} \phi_{n,i} \Big(\Psi(\gamma_i) - \Psi(\sum_{i=1}^{K} \gamma_i) \Big) \\
& - \Big(\sum_{i=1}^{K} (\gamma_i - 1) \Big(\Psi(\gamma_i) - \Psi(\sum_{i=1}^{K} \gamma_i) \Big) \Big) - \log \Gamma(\sum_{i=1}^{K} \gamma_i) + \sum_{i=1}^{K} \log \Gamma(\gamma_i).
\end{aligned}
$$

Taking the derivative of the $L_{[\gamma]}$

$$
\begin{aligned}
& dL_{[\gamma]}/d\gamma_i \\
= & (\alpha_i - 1)(\Psi'(\gamma_i) - \Psi'(\sum_j \gamma_j)) \\
& + \sum_n \phi_{n,i} \Big(\Psi'(\gamma_i) - \Psi'(\sum_{i=1}^{K} \gamma_i) \Big) \\
& - (\Psi(\gamma_i) - \Psi(\sum_j \gamma_j)) - (\gamma_i - 1)(\Psi'(\gamma_i) - \Psi'(\sum_j \gamma_j)) - \Psi(\sum_j \gamma_j) + \Psi(\gamma_i) \\
& = (\Psi'(\gamma_i) - \Psi'(\sum_j \gamma_j))(\alpha_i - 1 + \sum_n \phi_{n,i} - (\gamma_i - 1)).
\end{aligned}
$$

By setting $dL_{[\gamma]}/d\gamma_i$ to zero, we can get

$$\gamma_i = \alpha_i + \sum_n \phi_{n,i}. \tag{1.8}$$

1.3 Parameter estimation

In the previous section, we have discussed how to estimate the variational parameters. Next, we continue to estimate our model parameters, i.e., $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. We solve this problem by using the variational lower bound as a surrogate for the (intractable) marginal log likelihood, with the variational parameters. Note that we need to first aggregate document-specific lower bounds defined in Eq.1.2. And in this part, we will use the document index d.

We first rewrite the lower bound by only keeping the items which contain $\boldsymbol{\beta}$ with the lagrange multipliers ρ_is

$$L_{[\boldsymbol{\beta}]} = \sum_{d,n,i} \phi_{d,n,i} \log \beta_{i,w_n} + \sum_{i=1}^{K} \rho_i (\sum_{v=1}^{V} \beta_{i,v} - 1). \tag{1.9}$$

By taking the derivative $L_{[\boldsymbol{\beta}]}$, we can have

$$dL_{[\boldsymbol{\beta}]}/d\beta_{i,v} = \sum_{d,n} \frac{\phi_{d,n,i}\mathbf{1}(v = w_n)}{\beta_{i,v}} + \rho_i, \tag{1.10}$$

where $\mathbf{1}(v = w_n)$ is an indicator function which returns 1 when the condition is true otherwise returns 0. We can set $\sum_{d,n} \frac{\phi_{d,n,i}\mathbf{1}(v=w_n)}{\beta_{i,v}} + \rho_i$ to zero, and solve ρ_i: $\rho_i = -\sum_{d,n,v} \phi_{d,n,i}\mathbf{1}(v = w_n)$. Since we have $\sum_v \beta_{i,v} = 1$, we can ignore ρ_i to estimate an un-normalized value of $\beta_{i,v}$

$$\hat{\beta_{i,v}} \propto \sum_{d,n} \phi_{d,n,i}\mathbf{1}(v = w_n). \tag{1.11}$$

Similarly, we can rewrite the lower bound by only keeping the items which contain $\boldsymbol{\alpha}$

$$L_{\boldsymbol{\alpha}} = \sum_{d=1}^{D} \left(\sum_{i=1}^{K} (\alpha_i - 1)(\Psi(\gamma_{d,i}) - \Psi(\sum_j \gamma_{d,j})) + \log\Gamma(\sum_{i=1}^{K} \alpha_i) - \sum_{i=1}^{K} \log\Gamma(\alpha_i) \right)$$

.

By taking the derivative $L_{[\boldsymbol{\alpha}]}$, we can have

$$dL_{[\boldsymbol{\alpha}]}/d\alpha_i = \sum_d \left(\Psi(\gamma_{d,i}) - \Psi(\sum_j \gamma_{d,j}) \right) + D\left(\Psi(\sum_{i=1}^{K} \alpha_i) - \Psi(\alpha_i) \right),$$

This derivative of α_i depends on α_js, where $j \neq i$, and we therefore must use an iterative method to find the maximal α. In particular, the Hessian is in the form:

$$dL_{[\boldsymbol{\alpha}]}/d\alpha_i\alpha_j = D\Psi'(\sum_{i=1}^{K} \alpha_i) - D\Psi'(\alpha_i)\mathbf{1}[i = j],$$

having these derivatives, we can use the optimization algorithm in Blei's paper, and I will not discuss it in current version.

1.4 Summary of Variational EM algorithm for LDA

Having the variational inference part and parameter estimation, now we present a summary of the variational EM algorithm for LDA. The derivation yields the following iterative algorithm:

- **E-step**: For each document, running the following iterative algorithm to find the optimizing values of the variational parameters

initialize $\phi_{ni}^{(0)} \leftarrow \frac{1}{K}$ for all i and n;
initialize $\gamma_i^{(0)} \leftarrow \alpha_i + \frac{N_d}{K}$ for all i and n;
while *not converge* **do**
  for $n = 1$ *to* N_d **do**
    for $i = 1$ *to* K **do**
      $\phi_{n,i}^{(t+1)} \leftarrow \beta_{i,w_n} \exp(\Psi(\gamma_i))$;
    end
    normalize $\phi_{n,i}^{(t+1)}$ sum to 1;
  end
  $\gamma_i^{(t+1)} \leftarrow \alpha_i + \sum_n \phi_{n,i}^{(t+1)}$;
end

Algorithm 1: Iterative variational inference algorithm for a document.

- (M-step) Maximize the resulting lower bound on the log likelihood with respect to the model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. This corresponds to finding maximum likelihood estimates with expected sufficient statistics for each document under the approximate posterior which is computed in the E-step.

Appendix

In this part, we will go over some preliminary points for previous derivations. (This part is copied from Blei' paper)

The need to compute the expected value of the log of a single probability component under the Dirichlet arises repeatedly in deriving the inference and parameter estimation procedures for LDA. This value can be easily computed from the natural parameterization of the exponential family representation of the Dirichlet distribution. Recall that a distribution is in the exponential family if it can be written in the form:

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

where η is the natural parameter, $T(x)$ is the sufficient statistic, and $A(\eta)$ is the log of the normalization factor. We can write the Dirichlet in this form by exponentiating the log:

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

From this form, we immediately see that the natural parameter of the Dirichlet is $\eta_i = \alpha_i - 1$ and the sufficient statistic is $T(\theta_i) = \log \theta_i$. Furthermore, using the general fact that the derivative of the log normalization factor with respect to the natural parameter is equal to the expectation of the sufficient statistic, we obtain:

$$E[\log \theta_i | \alpha_i] = \Psi(\alpha_i) - \Psi(\sum_j \alpha_j),$$

where $\Psi(\cdot)$ is the digamma function, the first derivative of the log Gamma function. This is a very important point for the derivations in this note.