第一章　话题模型之LDA(Latent Dirichlet Allocation)介绍

最近我阅读了文献[1]的一部分，在这里说说自己对LDA的理解，希望有助于读者对LDA的学习。有不正确的地方，请留言赐教。

LDA是一种话题建模的方法,他认为每个文档是按算法1产生的：

Algorithm 1 LDA的文档生成过程

生成本文档的先验话题分布，$\boldsymbol{\theta} \sim Dirichlet(\boldsymbol{\alpha})$。
for $n = 1, 2, \ldots, N$ **do**
  从先验话题分布中选择一个话题，$z_n \sim multinomial(\boldsymbol{\theta})$。
  由话题z_n产生文档中的第n个单词w_n，$w_n \sim multinomial(\boldsymbol{\beta}_{z_n})$。
end for

其中$\boldsymbol{\alpha}, \boldsymbol{\theta}$都是$k$维向量，$k$是预设的话题个数，$\theta_i$表示本文档属于话题$i$的概率。$N$是当前文档的单词总数。$z_n$ 表示第n个单词来自哪个话题。$\boldsymbol{\beta}_i$是V维向量，V表示由全体单词组成的词库中的单词总数。接下来令$\boldsymbol{\beta} = [\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \ldots, \boldsymbol{\beta}_k^T]$，$w_n$表示文档中华第$n$ 个单词是词库中的第几个词。词库中第j个单词出现在话题i中的概率为β_{ij}。从而有如下联合概率：

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

其中，$\boldsymbol{z}$是N长的向量，z_n表示第n个单词来自哪个话题。$\boldsymbol{w}$是N长的向量，表示当前文档的所有单词，w_n表示文档的第n个单词。并且，

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \sim Dirichlet(\boldsymbol{\alpha}) \quad (2)$$

$$p(z_n|\boldsymbol{\theta}) \sim multinomial(\boldsymbol{\theta}) \Rightarrow p(z_n = i|\boldsymbol{\theta}) = \theta_i \quad (3)$$

$$p(w_n|\boldsymbol{\beta}_{z_n}) \sim multinomial(\boldsymbol{\beta}_{z_n}) \Rightarrow p(w_n = j|\boldsymbol{\beta}_i) = \beta_{ij} \quad (4)$$

$Dirichlet$表示Dirichlet分布，$multinomial$表示多项分布(multinomial distribution)，Dirichlet分布的具体形式如下：

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \sim Dirichlet(\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1} \quad (5)$$

我们的目的是给定文档$\boldsymbol{w}$时，求出该文档每种话题分布的概率以及文档中每个单词来自于每个话题的概率，也就是要求解概率分布$p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$。

我自己理解(也就是这段话我在原文中没有找到根据)，一旦得到这样的概率分布，那么若，

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_{\boldsymbol{z}} p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

$\boldsymbol{\theta}^*$便可以表示本文档属于每个话题的概率。同样的，

$$\int_{\boldsymbol{z}_{-n}} \int p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{\theta} d\boldsymbol{z}_{-n}$$

就表示该文档中单词w_n属于话题z_n的概率。$\boldsymbol{z}_{-n}$表示去掉$\boldsymbol{z}$中第n个分量后剩下分量构成的向量。

另外，我们还想将得到的话题表示出来，LDA用单词的分布来表示每个话题，也就是我们希望求得参数$\boldsymbol{\beta}$，他表示每个话题中每个单词的出现概率。

我们现在来考察如何求得$p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$。因为

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

其中，

$$\begin{aligned} p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \int \sum_{\boldsymbol{z}} p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{\theta} \\ &= \int \sum_{\boldsymbol{z}} p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{n=1}^{N} \left(p(z_n|\boldsymbol{\theta}) p(w_n|\boldsymbol{\beta}_{z_n}) \right) d\boldsymbol{\theta} \end{aligned}$$

其中涉及到很大的求和($\boldsymbol{z}$的维数可能很高)和复杂的积分，精确计算比较困难的。所以我们采用变分法(variational methods)近似计算。这里所谓的变分法的主要是思想是用带参数(变分参数)的简单分布去逼近一个复杂分布，我们可以不断调整变分参数使逼近得更好。

$$
\begin{aligned}
\log p(\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta}) &= \log \int \sum_{\boldsymbol{z}} p(\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta}) d\boldsymbol{\theta} \\
&= \log \int \sum_{\boldsymbol{z}} \frac{p(\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta}) q(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})}{q(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})} d\boldsymbol{\theta} \\
&\geq \int \sum_{\boldsymbol{z}} q(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi}) \log \frac{p(\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta})}{q(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})} d\boldsymbol{\theta} \\
&= \int \sum_{\boldsymbol{z}} q(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi}) \log p(\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta}) d\boldsymbol{\theta} \\
&\quad - \int \sum_{\boldsymbol{z}} q(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi}) \log q(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi}) d\boldsymbol{\theta} \\
&\triangleq L(\boldsymbol{\gamma},\boldsymbol{\phi};\boldsymbol{\alpha},\boldsymbol{\beta}) \quad (6)
\end{aligned}
$$

Jesen不等式使得其中的$\geq$成立，$\triangleq$表示定义为。$q(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})$是一个带参数的任意分布。可以很容易地获得如下结论:

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

其中，$KL(q\|p)$表示两个分布之间的KL散度(Kullback - Leibler divergence)。将KL散度的定义带入便可以验证这个结论。

我们可以看出$L(\boldsymbol{\gamma},\boldsymbol{\phi};\boldsymbol{\alpha},\boldsymbol{\beta})$越大，$KL(q(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi}) \| p(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta}))$就越小，那么$q(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})$就越接近$p(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta})$。至此，我们得到了如下结论：

因为$p(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta})$难以计算，故我们用一个简单分布$q(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})$逼近他。逼近的效果，我们用这两个分布的KL距离$KL(q\|p)$表示，我们希望最小化这个距离，而最小化这个距离等价于最大化$L(\boldsymbol{\gamma},\boldsymbol{\phi};\boldsymbol{\alpha},\boldsymbol{\beta})$。于是我们需要调整变分参数$\boldsymbol{\gamma},\boldsymbol{\phi}$使得$L(\boldsymbol{\gamma},\boldsymbol{\phi};\boldsymbol{\alpha},\boldsymbol{\beta})$最大。

那么我们究竟采用什么样的$q(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})$呢？我们采用最简单的形式：

$$q(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi}) = q(\boldsymbol{\theta}|\boldsymbol{\gamma}) q(\boldsymbol{z}|\boldsymbol{\phi}) = q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \prod_{i=1}^{N} q(z_n|\phi_n) \quad (7)$$

$\boldsymbol{\gamma}$是k维向量。$\boldsymbol{\phi}$是$k \times N$的矩阵，ϕ_n是其第n列。并且，

$$q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \sim Dirichlet(\boldsymbol{\gamma}) \quad (8)$$

$$q(z_n|\phi_n) \sim multinomial(\phi_n) \Rightarrow q(z_n = i|\phi_n) = \phi_{ni} \quad (9)$$

接下来，我们展开$L(\boldsymbol{\gamma},\boldsymbol{\phi};\boldsymbol{\alpha},\boldsymbol{\beta})$的第一项，因为

$$\begin{aligned}
&\int\sum_{\boldsymbol{z}} q(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})\log p(\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta})d\boldsymbol{\theta}\\
&=\int\sum_{\boldsymbol{z}} q(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta})d\boldsymbol{\theta}\\
&=\int\sum_{\boldsymbol{z}} q(\boldsymbol{\theta}|\boldsymbol{\gamma})q(\boldsymbol{z}|\boldsymbol{\phi})\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta})d\boldsymbol{\theta}\\
&=\int\sum_{\boldsymbol{z}} q(\boldsymbol{\theta}|\boldsymbol{\gamma})q(\boldsymbol{z}|\boldsymbol{\phi})\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})d\boldsymbol{\theta}\\
&+\int\sum_{\boldsymbol{z}} q(\boldsymbol{\theta}|\boldsymbol{\gamma})q(\boldsymbol{z}|\boldsymbol{\phi})\log p(\boldsymbol{z}|\boldsymbol{\theta})d\boldsymbol{\theta}\\
&+\int\sum_{\boldsymbol{z}} q(\boldsymbol{\theta}|\boldsymbol{\gamma})q(\boldsymbol{z}|\boldsymbol{\phi})\log p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta})d\boldsymbol{\theta} \quad (10)
\end{aligned}$$

第一个等号是因为带入式(1)而得到的。第二个等号是因为带入式(7)而得到的。第三个等号则是由对数函数的和差公式而得到的。类似地，我们展开$L(\boldsymbol{\gamma},\boldsymbol{\phi};\boldsymbol{\alpha},\boldsymbol{\beta})$的第二项,

$$\begin{aligned}
&\int\sum_{\boldsymbol{z}} q(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})\log q(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})d\boldsymbol{\theta}\\
&=\int\sum_{\boldsymbol{z}} q(\boldsymbol{\theta}|\boldsymbol{\gamma})q(\boldsymbol{z}|\boldsymbol{\phi})\log q(\boldsymbol{\theta}|\boldsymbol{\gamma})q(\boldsymbol{z}|\boldsymbol{\phi})d\boldsymbol{\theta}\\
&=\int\sum_{\boldsymbol{z}} q(\boldsymbol{\theta}|\boldsymbol{\gamma})q(\boldsymbol{z}|\boldsymbol{\phi})\log q(\boldsymbol{\theta}|\boldsymbol{\gamma})d\boldsymbol{\theta}\\
&+\int\sum_{\boldsymbol{z}} q(\boldsymbol{\theta}|\boldsymbol{\gamma})q(\boldsymbol{z}|\boldsymbol{\phi})\log q(\boldsymbol{z}|\boldsymbol{\phi})d\boldsymbol{\theta} \quad (11)
\end{aligned}$$

现在我们把$L(\boldsymbol{\gamma},\boldsymbol{\phi};\boldsymbol{\alpha},\boldsymbol{\beta})$展开成五项，接下来我们继续化简这五项。继续化简前，我们先介绍两个定理。

定理1.可以表示成$p(\boldsymbol{x})=h(\boldsymbol{x})exp\left(\sum_{j=1}^{k}\eta_jT_j(\boldsymbol{x})-A(\boldsymbol{\eta})\right)$的概率分布称为指数族分布，对于指数族分布有，

$$E(T_j(\boldsymbol{x}))=\int p(\boldsymbol{x})T_j(\boldsymbol{x})d\boldsymbol{x}=\frac{\partial A(\boldsymbol{\eta})}{\partial\eta_j}$$

定理1说的是指数族分布的一个重要性质，他的证明比较复杂。因为,

$$
\begin{aligned}
q(\boldsymbol{\theta}|\boldsymbol{\gamma}) &= \exp\left(\log q(\boldsymbol{\theta}|\boldsymbol{\gamma})\right) \\
&= \exp\left(\log \frac{\Gamma(\sum_{i=1}^{k}\gamma_i)}{\prod_{i=1}^{k}\Gamma(\gamma_i)}\theta_1^{\gamma_1-1}\cdots\theta_k^{\gamma_k-1}\right) \\
&= \exp\left(\sum_{i=1}^{k}(\gamma_i-1)\log\theta_i + \log\frac{\Gamma(\sum_{i=1}^{k}(\gamma_i-1+1))}{\prod_{i=1}^{k}\Gamma(\gamma_i-1+1)}\right)
\end{aligned}
$$

如果，我们令$\eta_i = \gamma_i - 1, \log\theta_i = T_i(\boldsymbol{\theta})$，那么Dirichlet分布也是指数族分布，且由定理1还有，

$$
\begin{aligned}
\int q(\boldsymbol{\theta}|\boldsymbol{\gamma})\log\theta_i d\boldsymbol{\theta} &= \frac{\partial}{\partial\gamma_i}\left(-\log\frac{\Gamma(\sum_{i=1}^{k}\gamma_i)}{\prod_{i=1}^{k}\Gamma(\gamma_i)}\right) \\
&= -\frac{\partial}{\partial\gamma_i}\log\Gamma(\sum_{i=1}^{k}\gamma_i) + \frac{\partial}{\partial\gamma_i}\sum_{i=1}^{k}\log\Gamma(\gamma_i) \\
&= \Psi(\gamma_i) - \Psi(\sum_{i=1}^{k}\gamma_i) \quad (12)
\end{aligned}
$$

其中，$\Psi(\cdot)$是Digamma函数，是$\log\Gamma(\cdot)$的一阶导数。

接着，我们介绍另外一个定理，

定理2.设$q(\boldsymbol{z}), p(\boldsymbol{w})$是两个$N$维的概率密度函数，且满足

$$
q(\boldsymbol{z}) = \prod_{n=1}^{N} q(z_n), p(\boldsymbol{w}) = \prod_{n=1}^{N} p(w_n)
$$

那么，

$$
\sum_{\boldsymbol{z}} q(\boldsymbol{z})\log p(\boldsymbol{w}) = \sum_{n=1}^{N}\sum_{z_n} q(z_n)\log p(w_n)
$$

proof.

定理2的证明比较容易，

$$
\begin{aligned}
&\sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log p(\boldsymbol{z}) = \sum_{\boldsymbol{z}} \prod_{n=1}^{N} q(z_n) \log \prod_{n=1}^{N} p(z_n) \\
&= \sum_{z_1} \cdots \sum_{z_N} \prod_{n=1}^{N} q(z_n) \log \prod_{n=1}^{N} p(z_n) \\
&= \sum_{z_1} \cdots \sum_{z_{N-1}} \left(\sum_{z_N} \prod_{n=1}^{N} q(z_n) \log \prod_{n=1}^{N} p(z_n) \right) \\
&= \sum_{z_1} \cdots \sum_{z_{N-1}} \prod_{n=1}^{N-1} q(z_n) \left(\sum_{z_N} q(z_N) \log \prod_{n=1}^{N} p(z_n) \right) \\
&= \sum_{z_1} \cdots \sum_{z_{N-1}} \prod_{n=1}^{N-1} q(z_n) \left(\sum_{z_N} q(z_N) \log \prod_{n=1}^{N-1} p(z_n) + \sum_{z_N} q(z_N) \log p(z_N) \right) \\
&= \sum_{z_1} \cdots \sum_{z_{N-1}} \prod_{n=1}^{N-1} q(z_n) \left(\log \prod_{n=1}^{N-1} p(z_n) \sum_{z_N} q(z_N) + \sum_{z_N} q(z_N) \log p(z_N) \right) \\
&= \sum_{z_1} \cdots \sum_{z_{N-1}} \prod_{n=1}^{N-1} q(z_n) \left(\log \prod_{n=1}^{N-1} p(z_n) + \sum_{z_N} q(z_N) \log p(z_N) \right) \\
&= \sum_{z_1} \cdots \sum_{z_{N-1}} \prod_{n=1}^{N-1} q(z_n) \log \prod_{n=1}^{N-1} p(z_n) + \sum_{z_1} \cdots \sum_{z_{N-1}} \prod_{n=1}^{N-1} q(z_n) \left(\sum_{z_N} q(z_N) \log p(z_N) \right) \\
&= \sum_{z_1} \cdots \sum_{z_{N-1}} \prod_{n=1}^{N-1} q(z_n) \log \prod_{n=1}^{N-1} p(z_n) + \sum_{z_N} q(z_N) \log p(z_N)
\end{aligned}
$$

第一个等号成立是因为定理2中的前提条件。第二个等号成立是因为$pmbz$是N维向量，对他进行累加就等于依次对其各个分量进行累加。第三个等号成立是因为加法结合律，也就是先对z_N进行累加。第四个等号成立是因为我们提取了所有加法项的公共因子，且这个公共因子相对于z_N是常数。第五个等号成立时因为对数函数的和差公式。第六个等号同样是提取所有加法项的公共常数因子。第七个等号成立是因为概率归一化条件，

$$\sum_{z_N} q(z_N) = 1$$

第八个等号成立的还是是因为，

$$\sum_i (a_i + b_i) = \sum_i a_i + \sum_i b_i$$

第九个等式成立，也是因为概率归一化条件，

$$\sum_{z_1}\cdots\sum_{z_{N-1}}\prod_{n=1}^{N-1}q(z_n)=1$$

至此，我们得到了递推公式：

$$\sum_{z_1}\cdots\sum_{z_N}\prod_{n=1}^{N}q(z_n)\log\prod_{n=1}^{N}p(z_n)=\sum_{z_1}\cdots\sum_{z_{N-1}}\prod_{n=1}^{N-1}q(z_n)\log\prod_{n=1}^{N-1}p(z_n)+\sum_{z_N}q(z_N)\log p(z_N)$$

反复使用该递推公式N次，便可的到定理2.■

我们再给出一个有用的结论，他的证明很容易，这里就不给出细节了，

$$\sum_{z_1}\cdots\sum_{z_N}\left(\prod_{n=1}^{N}p(z_n)\right)=\prod_{n=1}^{N}\left(\sum_{z_n}p(z_n)\right) \tag{13}$$

现在，我们继续化简$L(\boldsymbol{\gamma},\boldsymbol{\phi};\boldsymbol{\alpha},\boldsymbol{\beta})$展开成的五项。我们逐项地化简。

首先化简第一项，

$$\begin{aligned}
&\int\sum_{\boldsymbol{z}}q(\boldsymbol{\theta}|\boldsymbol{\gamma})q(\boldsymbol{z}|\boldsymbol{\phi})\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})d\boldsymbol{\theta}\\
&=\int q(\boldsymbol{\theta}|\boldsymbol{\gamma})\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})\sum_{\boldsymbol{z}}q(\boldsymbol{z}|\boldsymbol{\phi})d\boldsymbol{\theta}\\
&=\int q(\boldsymbol{\theta}|\boldsymbol{\gamma})\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})d\boldsymbol{\theta}\\
&=\int q(\boldsymbol{\theta}|\boldsymbol{\gamma})\log\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1}d\boldsymbol{\theta}\\
&=\int q(\boldsymbol{\theta}|\boldsymbol{\gamma})\log\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}d\boldsymbol{\theta}+\sum_{i=1}^{k}(\alpha_i-1)\int q(\boldsymbol{\theta}|\boldsymbol{\gamma})\log\theta_i d\boldsymbol{\theta}\\
&=\log\Gamma(\sum_{i=1}^{k}\alpha_i)-\sum_{i=1}^{k}\log\Gamma(\alpha_i)+\sum_{i=1}^{k}(\alpha_i-1)(\Psi(\gamma_i)-\Psi(\sum_{i=1}^{k}\gamma_i))
\end{aligned}$$

第三个等号成立是因为式(5)。最后一个等号成立是因为式(9)。其他等号成立则是因为概率归一化条件或者对数和差公式。

再来化简第二项，首先

$$
\begin{aligned}
&\sum_{\boldsymbol{z}} q(\boldsymbol{z}|\boldsymbol{\phi}) \log p(\boldsymbol{z}|\boldsymbol{\theta}) \\
&= \sum_{z_1} \cdots \sum_{z_N} \prod_{n=1}^{N} q(z_n|\phi_n) \log \prod_{n=1}^{N} p(z_n|\boldsymbol{\theta}) \\
&= \sum_{n=1}^{N} \sum_{z_n} q(z_n|\phi_n) \log p(z_n|\boldsymbol{\theta}) \\
&= \sum_{n=1}^{N} \sum_{i=1}^{k} q(z_n = i|\phi_n) \log p(z_n = i|\boldsymbol{\theta}) \\
&= \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \log \theta_i
\end{aligned}
$$

第一个等号成立是因为式(1)和式(7)。第二个等号成立是因为定理2。第四个等号成立是因为式(3)和式(8)。那么第二项有，

$$
\begin{aligned}
&\int \sum_{\boldsymbol{z}} q(\boldsymbol{\theta}|\boldsymbol{\gamma}) q(\boldsymbol{z}|\boldsymbol{\phi}) \log p(\boldsymbol{z}|\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \sum_{\boldsymbol{z}} q(\boldsymbol{z}|\boldsymbol{\phi}) \log p(\boldsymbol{z}|\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \log \theta_i d\boldsymbol{\theta} \\
&= \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \int q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \log \theta_i d\boldsymbol{\theta} \\
&= \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} (\Psi(\gamma_i) - \Psi(\sum_{i=1}^{k} \gamma_i))
\end{aligned}
$$

最后一个等号成立是因为式(9)。

我们现在来化简第三项，

$$
\begin{aligned}
&\int \sum_{\boldsymbol{z}} q(\boldsymbol{\theta}|\boldsymbol{\gamma}) q(\boldsymbol{z}|\boldsymbol{\phi}) \log p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta}) d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \sum_{\boldsymbol{z}} q(\boldsymbol{z}|\boldsymbol{\phi}) \log p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta}) d\boldsymbol{\theta} \\
&= \sum_{\boldsymbol{z}} q(\boldsymbol{z}|\boldsymbol{\phi}) \log p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta}) \\
&= \sum_{\boldsymbol{z}} \prod_{n=1}^{N} q(z_n|\phi_n) \log \prod_{n=1}^{N} p(w_n|\beta_{z_n}) \\
&= \sum_{n=1}^{N} \sum_{z_n} q(z_n|\phi_n) \log p(w_n|\beta_{z_n}) \\
&= \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \sum_{j=1}^{V} w_n^j \log \beta_{ij}
\end{aligned}
$$

第四个等号成立是因为定理2。第五个等号成立是因为式(4)和式(8)。w_n^j当本文档中第n个单词是词库中第j个单词时为1，其他时候为0。

我们接着来化简第四项。

$$
\begin{aligned}
&\int \sum_{\boldsymbol{z}} q(\boldsymbol{\theta}|\boldsymbol{\gamma}) q(\boldsymbol{z}|\boldsymbol{\phi}) \log q(\boldsymbol{\theta}|\boldsymbol{\gamma}) d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \log q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \sum_{\boldsymbol{z}} q(\boldsymbol{z}|\boldsymbol{\phi}) d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \log q(\boldsymbol{\theta}|\boldsymbol{\gamma}) d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \log \frac{\Gamma(\sum_{i=1}^{k} \gamma_i)}{\prod_{i=1}^{k} \Gamma(\gamma_i)} \theta_1^{\gamma_1 - 1} \cdots \theta_k^{\gamma_k - 1} d\boldsymbol{\theta} \\
&= \log \frac{\Gamma(\sum_{i=1}^{k} \gamma_i)}{\prod_{i=1}^{k} \Gamma(\gamma_i)} + \sum_{i=1}^{k} (\gamma_i - 1) \int q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \log \theta_i d\boldsymbol{\theta} \\
&= \log \Gamma(\sum_{i=1}^{k} \gamma_i) - \sum_{i=1}^{k} \log \Gamma(\gamma_i) + \sum_{i=1}^{k} (\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{i=1}^{k} \gamma_i))
\end{aligned}
$$

第三个等号成立是因为式(8)。最后一个等号成立是因为式(12)。

最后，我们化简第五项，

$$
\begin{aligned}
&\int \sum_{\boldsymbol{z}} q(\boldsymbol{\theta}|\boldsymbol{\gamma}) q(\boldsymbol{z}|\boldsymbol{\phi}) \log q(\boldsymbol{z}|\boldsymbol{\phi}) d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \sum_{\boldsymbol{z}} q(\boldsymbol{z}|\boldsymbol{\phi}) \log q(\boldsymbol{z}|\boldsymbol{\phi}) d\boldsymbol{\theta} \\
&= \sum_{\boldsymbol{z}} q(\boldsymbol{z}|\boldsymbol{\phi}) \log q(\boldsymbol{z}|\boldsymbol{\phi}) \\
&= \sum_{n=1}^{N} \sum_{z_n} q(z_n|\phi_n) \log p(z_n|\phi_n) \\
&= \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \log \phi_{ni}
\end{aligned}
$$

第三个等号成立是因为定理2。

现在，我们终于可以写出$L(\boldsymbol{\gamma},\boldsymbol{\phi};\boldsymbol{\alpha},\boldsymbol{\beta})$的具体表达式了，

$$
\begin{aligned}
L(\boldsymbol{\gamma},\boldsymbol{\phi};\boldsymbol{\alpha},\boldsymbol{\beta}) = {} & \log\Gamma(\sum_{i=1}^{k}\alpha_i) - \sum_{i=1}^{k}\log\Gamma(\alpha_i) + \sum_{i=1}^{k}(\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{i=1}^{k}\gamma_i)) \\
& + \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{i=1}^{k}\gamma_i)) \\
& + \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}\sum_{j=1}^{V} w_n^j \log\beta_{ij} \\
& - \log\Gamma(\sum_{i=1}^{k}\gamma_i) + \sum_{i=1}^{k}\log\Gamma(\gamma_i) - \sum_{i=1}^{k}(\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{i=1}^{k}\gamma_i)) \\
& - \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}\log\phi_{ni}
\end{aligned}
$$

现在，我们调整变分参数$\boldsymbol{\gamma},\boldsymbol{\phi}$使$L(\boldsymbol{\gamma},\boldsymbol{\phi};\boldsymbol{\alpha},\boldsymbol{\beta})$达到最大。我们先固定参数$\boldsymbol{\gamma}$使，优化参数$\boldsymbol{\phi}$，则得到如下问题

$$
\begin{cases}
\max\limits_{\boldsymbol{\phi}} \quad L(\boldsymbol{\phi}) \\
s.t \quad \sum\limits_{i=1}^{k}\phi_{ni} = 1, n = 1, 2, \ldots, N
\end{cases}
$$

这个问题的Lagrange函数为：

$$\mathcal{L} = L(\boldsymbol{\phi}) + \sum_{n=1}^{N} \lambda_n (\sum_{i=1}^{k} \phi_{ni} - 1)$$

则，容易得到，

$$\begin{aligned}
&\frac{\partial \mathcal{L}}{\partial \phi_{ni}} = \Psi(\gamma_i) - \Psi(\sum_{i=1}^{k} \gamma_i) + \sum_{j=1}^{V} w_n^j \log \beta_{ij} - \log \phi_{ni} - 1 + \lambda_n = 0 \\
&\Rightarrow \log \phi_{ni} = \Psi(\gamma_i) - \Psi(\sum_{i=1}^{k} \gamma_i) + \sum_{j=1}^{V} w_n^j \log \beta_{ij} - 1 + \lambda_n \\
&\Rightarrow \log \phi_{ni} = \Psi(\gamma_i) - \Psi(\sum_{i=1}^{k} \gamma_i) + \log \beta_{iv} - 1 + \lambda_n \\
&\Rightarrow \phi_{ni} = \exp(\Psi(\gamma_i) - \Psi(\sum_{i=1}^{k} \gamma_i) + \log \beta_{iv} - 1 + \lambda_n) \\
&= \exp(\Psi(\gamma_i) - \Psi(\sum_{i=1}^{k} \gamma_i)) \exp(\log \beta_{iv}) \exp(-1 + \lambda_n)) \\
&\Rightarrow \phi_{ni} = \exp(\Psi(\gamma_i) - \Psi(\sum_{i=1}^{k} \gamma_i)) \beta_{iv} \exp(-1 + \lambda_n) \\
&\Rightarrow \phi_{ni} \propto \exp(\Psi(\gamma_i) - \Psi(\sum_{i=1}^{k} \gamma_i)) \beta_{iv}
\end{aligned}$$

第三步我们令$v = \arg_j(w_n^j = 1)$。而最后一步则是将第六步得到的ϕ_{ni}带入下述约束条件得到的，

$$\sum_{i=1}^{k} \phi_{ni} = 1$$

接着，我们先固定参数使$\boldsymbol{\phi}$，优化参数$\boldsymbol{\gamma}$。则得到如下问题

$$\left\{ \max_{\boldsymbol{\gamma}} \quad L(\boldsymbol{\gamma}) \right.$$

注意尽管有

$$\sum_{i=1}^{k} \gamma_i = 1$$

但是这个约束是Dirichlet分布自动满足的，不需要额外关注。

$$
\begin{aligned}
&\frac{\partial L(\boldsymbol{\gamma})}{\partial \gamma_i} = (\alpha_i - 1)\Psi'(\gamma_i) - \sum_{i=1}^{k}(\alpha_i - 1)\Psi'(\sum_{i=1}^{k}\gamma_i) \\
&+ \sum_{n=1}^{N}\phi_{ni}\Psi'(\gamma_i) - \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}\Psi'(\sum_{i=1}^{k}\gamma_i) \\
&- \Psi(\sum_{i=1}^{k}\gamma_i) + \Psi(\gamma_i) - \Psi(\gamma_i) - (\gamma_i - 1)\Psi'(\gamma_i) + \sum_{i=1}^{k}(\gamma_i - 1)\Psi'(\sum_{i=1}^{k}\gamma_i) + \Psi(\sum_{i=1}^{k}\gamma_i) \\
&= \Psi'(\gamma_i)(\alpha_i + \sum_{n=1}^{N}\phi_{ni} - \gamma_i) - \Psi'(\sum_{i=1}^{k}\gamma_i)\sum_{i=1}^{k}(\alpha_i + \sum_{n=1}^{N}\phi_{ni} - \gamma_i))
\end{aligned}
$$

从而，观察上述最后一步，我们可以发现一个零点，即，

$$\frac{\partial L(\boldsymbol{\gamma})}{\partial \gamma_i} = 0 \Rightarrow \gamma_i = \alpha_i + \sum_{n=1}^{N}\phi_{ni}$$

至此，我们完成了变分近似的算法推导。然而参数$\boldsymbol{\alpha}, \boldsymbol{\beta}$目前未知，所以接下来，我们需要继续推导求解$\boldsymbol{\alpha}, \boldsymbol{\beta}$的算法。我们采用MLE来估计这两个参数。因为直接最大化似然函数比较困难，不过我们有，

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \sum_{d=1}^{M}\log p(\boldsymbol{w}_d|\boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&\geq \sum_{d=1}^{M} L(\boldsymbol{\gamma}_d, \boldsymbol{\phi}_d; \boldsymbol{\alpha}, \boldsymbol{\beta}) \triangleq L(D)
\end{aligned}
$$

其中$\boldsymbol{w}_d$是文档集D中第d个文档的单词向量，$\boldsymbol{\gamma}_d, \boldsymbol{\phi}_d$则是第$d$个文档对应的变分参数。一共$M$个文档。所以我们可以最大化似然函数的下界进而最大化似然函数，不过问题是上述下界中还有变分参数。所以，我们交替执行下面两个步骤，

E步 固定参数$\boldsymbol{\alpha}, \boldsymbol{\beta}$，调整变分参数使得$L(D)$最大。

M步 固定变分参数，调整$\boldsymbol{\alpha}, \boldsymbol{\beta}$使得$L(D)$最大。

其中，E步就是计算每个文档的最优变分参数，我们前面已经推导完成了。现在，我们来看M步。先固定参数$\boldsymbol{\alpha}$，那么再加上β_{ij}上的概率归一化条件，

$$\sum_{j=1}^{V}\beta_{ij} = 1$$

我们最大化Lagrange函数，

$$\mathcal{L}(D)=\sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\phi_{dni}\sum_{j=1}^{V}w_{dn}^{j}\log\beta_{ij}+\sum_{i=1}^{k}\lambda_i(\sum_{j=1}^{V}\beta_{ij}-1)+c$$

c表示常数项，N_d表示第d个文档中的单词个数，则，

$$\begin{aligned}\frac{\partial\mathcal{L}(D)}{\partial\beta_{ij}}&=\frac{1}{\beta_{ij}}\sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^{j}+\lambda_i=0\\ \Rightarrow\beta_{ij}&=\frac{\sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^{j}}{-\lambda_i}\end{aligned}$$

带入β_{ij}上的概率归一化条件，有

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

接下里，我们固定参数$\boldsymbol{\beta}$，优化参数$\boldsymbol{\alpha}$，则

$$\mathcal{L}(D)=\sum_{d=1}^{M}\left(\log\Gamma(\sum_{i=1}^{k}\alpha_i)-\sum_{i=1}^{k}\log\Gamma(\alpha_i)+\sum_{i=1}^{k}(\alpha_i-1)(\Psi(\gamma_{di})-\Psi(\sum_{i=1}^{k}\gamma_{di}))\right)+c$$

最大化上述函数时，没有解析的表示式，所以采用Newton-Raphson方法求解上述函数的最大点。Newton-Raphson求解上述问题的迭代公式为：

$$\boldsymbol{\alpha}_{new}=\boldsymbol{\alpha}_{old}-H(\boldsymbol{\alpha}_{old})^{-1}g(\boldsymbol{\alpha}_{old})$$

$H(\cdot)$为$\mathcal{L}(D)$的Hessian矩阵，$g(\cdot)$为$\mathcal{L}(D)$的梯度。注意，尽管这里需要求逆矩阵，不过$\mathcal{L}(D)$的Hessian矩阵的逆矩阵可以线性地求得。

最后，我们做一个总结。LDA的参数训练算法为算法2。 当完成上述参数训练算法后，我们可以得到文档d的话题分布

$$Dirichlet(\boldsymbol{\theta}|\boldsymbol{\gamma}_d)$$

文档d的第n个单词的话题分布

$$multinomial(z_{dn}|\phi_{dn})$$

话题i的单词分布

$$multinomial(w|\boldsymbol{\beta}_i)$$

Algorithm 2 LDA的参数训练算法

初始化：

$\boldsymbol{\alpha} = \boldsymbol{\alpha}^0, \boldsymbol{\beta} = \boldsymbol{\beta}^0$

$\phi_{dni} = \frac{1}{k}, d = 1, 2, \ldots, M, n = 1, 2, \ldots, N_d, i = 1, 2, \ldots, k,$

$\gamma_{di} = \alpha_i + \frac{N}{k}, d = 1, 2, \ldots, M, i = 1, 2, \ldots, k$

while 收敛 **do**

E步：

while 收敛 **do**

for $d = 1, 2, \ldots, M$ **do**

for $n = 1, 2, \ldots, N_d$ **do**

for $i = 1, 2, \ldots, k$ **do**

$\phi_{dni} = \beta_{if(w_{dn})} \exp(\Psi(\gamma_{di}) - \Psi(\sum_{i=1}^{k} \gamma_{di}))$

end for

归一化$\phi_{dni}, = 1, 2, \ldots, k$

for $i = 1, 2, \ldots, k$ **do**

$\gamma_{di} = \alpha_i + \sum_{n=1}^{N} \phi_{dni}$

end for

end for

end for

end while

M步：

while 收敛 **do**

for $i = 1, 2, \ldots, k$ **do**

for $j = 1, 2, \ldots, V$ **do**

$\beta_{ij} = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j}$

end for

归一化$\beta_{ij}, j = 1, 2, \ldots, V$

end for

调用Newton-Raphson法，求解$\boldsymbol{\alpha}$

end while

end while

参考文献

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