从狭义EM到变分自编码器

1. 狭义EM

Reference：[统计学习方法 第二版 第九章]

EM算法是一种迭代算法，1977 年由 Dempster 等人总结提出，用于含有隐变量（hidden variable）的概率模型参数的极大似然估计，或极大后验概率估计。EM算法的每次迭代由两步组成：E 步，求期望（expectation）；M 步，求极大（maximization）。所以这一算法称为期望极大算法（expectation maximization algorithm），简称 EM 算法。

1.1 EM 算法的引入

概率模型有时既含有观测变量（observable variable），又含有隐变量或潜在变量（latent variable）。如果概率模型的变量都是观测变量，那么给定数据，可以直接用极大似然估计法，或贝叶斯估计法估计模型参数（**此时只有固定的参数未知**）。EM 算法就是含有隐变量的概率模型参数的极大似然估计法，或**极大后验概率估计法**。

例1（三硬币模型）：假设有 3 枚硬币，分别记作 A，B，C。这些硬币正面出现的概率分别是 π，p，q。进行如下掷硬币试验：先掷硬币 A，根据其结果选出硬币 B 或硬币 C，正面选硬币 B，反面选硬币 C；然后掷选出的硬币，掷硬币的结果，出现正面记作 1，出现反面记作 0；独立地重复 n 次试验（这里，$n=10$），观测结果如下，如何估计三硬币模型的参数：

$$1,1,0,1,0,0,1,0,1,1$$

三硬币模型可以写作

$$P(y \mid \theta)=\sum_z P(y,z \mid \theta)=\sum_z P(z \mid \theta)P(y \mid z,\theta)=\pi p^y(1-p)^{1-y}+(1-\pi)q^y(1-q)^{1-y} \tag{1}$$

这里随机变量 y 是观测变量，表示一次试验观测到的结果是 1 或 0；随机变量 z 是隐变量，表示未观测到的掷硬币 A 的结果；$\theta=(\pi,p,q)$ 是模型参数。则所有观测数据的似然函数为

$$P(Y \mid \theta)=\sum_Z P(Z \mid \theta)P(Y \mid Z,\theta)=\prod_{j=1}^n[\pi p^{y_j}(1-p)^{1-y_j}+(1-\pi)q^{y_j}(1-q)^{1-y_j}] \tag{2}$$

考虑求模型参数 $\theta=(\pi,p,q)$ 的极大似然估计，即

$$\hat{\theta}=\arg\max_\theta \log P(Y \mid \theta) \tag{3}$$

这个问题没有解析解，因为 log 中有相加的两项，只有通过迭代的方法求解。

E 步：计算在模型参数 $\pi^{(i)},p^{(i)},q^{(i)}$ 下观测数据 y_j 来自掷硬币 B 的概率（将上一步的参数作为已知值，在算法一中应该为M步）

$$\mu_j^{(i+1)}=\frac{\pi^{(i)}(p^{(i)})^{y_j}(1-p^{(i)})^{1-y_j}}{\pi^{(i)}(p^{(i)})^{y_j}(1-p^{(i)})^{1-y_j}+(1-\pi^{(i)})(q^{(i)})^{y_j}(1-q^{(i)})^{1-y_j}} \tag{4}$$

M 步：计算模型参数的新估计值

$$\begin{aligned}\pi^{(i+1)}&=\frac{1}{n}\sum_{j=1}^n\mu_j^{(i+1)}\\ p^{(i+1)}&=\frac{\sum_{j=1}^n\mu_j^{(i+1)}y_j}{\sum_{j=1}^n\mu_j^{(i+1)}}\\ q^{(i+1)}&=\frac{\sum_{j=1}^n(1-\mu_j^{(i+1)})y_j}{\sum_{j=1}^n(1-\mu_j^{(i+1)})}\end{aligned} \tag{5}$$

按照上述迭代步骤直至收敛，若假设模型参数初值为 $\pi^{(0)}=0.5,p^{(0)}=0.5,q^{(0)}=0.5$，则模型参数的极大似然估计为 $\hat{\pi}=0.5,\hat{p}=0.6,\hat{q}=0.6$。若假设模型参数初值为 $\pi^{(0)}=0.4,p^{(0)}=0.6,q^{(0)}=0.6$，则模型参数的极大似然估计为 $\hat{\pi}=0.4064,\hat{p}=0.5368,\hat{q}=0.6432$。

一般地，用 Y 表示观测随机变量的数据，Z 表示隐随机变量的数据。Y 和 Z 连在一起称为完全数据（complete-data），观测数据 Y 又称为不完全数据（incomplete-data）。

首先写出所有观测的似然函数

$$P(Y \mid \theta)=\prod_{j=1}^n P(y_i \mid \theta)=\prod_{j=1}^n[\pi p+(1-\pi)q]^{y_j}[\pi(1-p)+(1-\pi)(1-q)]^{(1-y_i)} \tag{6}$$

计算 $P(z_j=1 \mid y_j,\theta^{(i)})$：

$$\mu_j^{(i+1)} = P(z_j = 1 \mid y_j, \theta^{(i)}) = \frac{P(y_j \mid z_j = 1, \theta^{(i)})P(z_j = 1 \mid \theta^{(i)})}{P(y_j \mid \theta^{(i)})} = \begin{cases} \frac{\pi^{(i)}p^{(i)}}{\pi^{(i)}p^{(i)}+(1-\pi^{(i)})q^{(i)}} & \text{if } y_j = 1 \\ \frac{\pi^{(i)}(1-p^{(i)})}{\pi^{(i)}(1-p^{(i)})+(1-\pi^{(i)})(1-q^{(i)})} & \text{if } y_j = 0 \end{cases} \tag{7}$$

计算完全数据的对数似然函数的期望

$$\begin{aligned} Q(\theta \mid \theta^{(i)}) =& \mathbb{E}_{P(Z|Y,\theta^{(i)})}[\log P(Y, Z \mid \theta)] \\ =& \mathbb{E}_{P(Z|Y,\theta^{(i)})}[\sum_{j=1}^{n} \log P(y_j, z_j \mid \theta)] \\ =& \sum_{j=1}^{n} \mathbb{E}_{P(z_j|y_j,\theta^{(i)})}[\log P(y_j, z_j \mid \theta)] \\ =& \sum_{j=1}^{n} \sum_{z_j} P(z_j \mid y_j, \theta^{(i)}) \log P(y_j, z_j \mid \theta) \\ =& \sum_{j=1}^{n} \left[\mu_j^{(i+1)} \log \left(\pi p^{y_j}(1-p)^{(1-y_j)}\right) + (1-\mu_j^{(i+1)}) \log \left((1-\pi) q^{y_j}(1-q)^{(1-y_j)}\right)\right] \end{aligned} \tag{8}$$

对各参数求导，并令其满足一阶条件可得公式（5）（可拆解验证其凹性）。

算法1（EM 算法）：
输入：观测变量数据 Y，隐变量数据 Z，联合分布 $P(Y, Z \mid \theta)$，条件分布 $P(Z \mid Y, \theta)$；
输出：模型参数 θ。
（1）选择参数的初值 $\theta^{(0)}$，开始迭代；
（2）E 步：记 $\theta^{(i)}$ 为第 i 次迭代参数的估计值，在第 $i+1$ 次迭代的 E 步，计算

$$Q(\theta, \theta^{(i)}) = E_Z[\log P(Y, Z \mid \theta) \mid Y, \theta^{(i)}] = \sum_Z P(Z \mid Y, \theta^{(i)}) \log P(Y, Z \mid \theta) \tag{9}$$

（3）M 步：求使 $Q(\theta, \theta^{(i)})$ 极大化的 θ，确定第 $i+1$ 次迭代的参数的估计值 $\theta^{(i+1)}$

$$\theta^{(i+1)} = \arg\max_\theta Q(\theta, \theta^{(i)}) \tag{10}$$

（4）重复第（2）步和第（3）步，直到收敛。

注意：参数的初值可以任意选择，但 EM 算法对初值是敏感的。

定义1：Q 函数（Q function）
完全数据的对数似然函数 $\log P(Y, Z \mid \theta)$ 关于在给定观测数据 Y 和当前参数 $\theta^{(i)}$ 下对未观测数据 Z 的条件概率分布 $P(Z \mid Y, \theta^{(i)})$ 的期望称为 Q 函数

$$Q(\theta, \theta^{(i)}) = E_Z[\log P(Y, Z \mid \theta) \mid Y, \theta^{(i)}] \tag{11}$$

1.2 EM 算法的导出（其本质都是对似然 $\log(Y \mid \theta)$ 的最大化）

1.2.1 方法一（正向推导，只需要有进步即可）

Reference：[统计学习方法 第二版 179页]
面对一个含有隐变量的概率模型，目标是极大化观测数据（不完全数据） Y 关于参数 θ 的对数似然函数，即极大化

$$L(\theta) = \log P(Y \mid \theta) = \log \sum_Z P(Y, Z \mid \theta) = \log \left(\sum_Z P(Y \mid Z, \theta) P(Z \mid \theta)\right) \tag{12}$$

上式极大化的主要困难在于未观测数据以及对数里的和（或者积分）。

EM 算法是通过迭代逐步近似极大化 $L(\theta)$ 的，假设在第 i 次迭代后 θ 的估计值是 $\theta^{(i)}$ 。我们希望新估计值 θ 能使 $L(\theta)$ 增加，即 $L(\theta) > L(\theta^{(i)})$，并逐步达到极大值。

$$L(\theta) - L(\theta^{(i)}) = \log \left(\sum_Z P(Y \mid Z, \theta) P(Z \mid \theta)\right) - \log P(Y \mid \theta^{(i)}) \tag{13}$$

利用 Jensen 不等式得到其下界：

$$
\begin{aligned}
L(\theta)-L(\theta^{(i)}) =&\log\left(\sum_Z P(Z\mid Y,\theta^{(i)})\frac{P(Y\mid Z,\theta)P(Z\mid\theta)}{P(Z\mid Y,\theta^{(i)})}\right)-\log P(Y\mid\theta^{(i)})\\
\geq&\sum_Z P(Z\mid Y,\theta^{(i)})\log\frac{P(Y\mid Z,\theta)P(Z\mid\theta)}{P(Z\mid Y,\theta^{(i)})}-\log P(Y\mid\theta^{(i)})\\
=&\sum_Z P(Z\mid Y,\theta^{(i)})\log\frac{P(Y\mid Z,\theta)P(Z\mid\theta)}{P(Z\mid Y,\theta^{(i)})}-\sum_Z P(Z\mid Y,\theta^{(i)})\log P(Y\mid\theta^{(i)})\\
=&\sum_Z P(Z\mid Y,\theta^{(i)})\log\frac{P(Y\mid Z,\theta)P(Z\mid\theta)}{P(Z\mid Y,\theta^{(i)})P(Y\mid\theta^{(i)})}
\end{aligned}\tag{14}
$$

令

$$
B(\theta,\theta^{(i)}) =L(\theta^{(i)})+\sum_Z P(Z\mid Y,\theta^{(i)})\log\frac{P(Y\mid Z,\theta)P(Z\mid\theta)}{P(Z\mid Y,\theta^{(i)})P(Y\mid\theta^{(i)})}\tag{15}
$$

则 $L(\theta)\geq B(\theta,\theta^{(i)})$，即函数 $B(\theta,\theta^{(i)})$ 是 $L(\theta)$ 的一个下界，而且 $B(\theta^{(i)},\theta^{(i)})=L(\theta^{(i)})$ 。因此，任何使 $B(\theta,\theta^{(i)})$ 相较于在 $\theta^{(i)}$ 处增大的 θ 也可以使相应的 $L(\theta)$ 增大，即 $L(\theta^{(i)})=B(\theta^{(i)},\theta^{(i)})\leq B(\theta^{(i+1)},\theta^{(i)})\leq L(\theta^{(i+1)})$ 。

$$
\begin{aligned}
\theta^{(i+1)} =&\arg\max_\theta B(\theta,\theta^{(i)})\\
=&\arg\max_\theta\left(L(\theta^{(i)})+\sum_Z P(Z\mid Y,\theta^{(i)})\log\frac{P(Y\mid Z,\theta)P(Z\mid\theta)}{P(Z\mid Y,\theta^{(i)})P(Y\mid\theta^{(i)})}\right)\\
=&\arg\max_\theta\left(\sum_Z P(Z\mid Y,\theta^{(i)})\log P(Y\mid Z,\theta)P(Z\mid\theta)\right)\\
=&\arg\max_\theta Q(\theta,\theta^{(i)})
\end{aligned}\tag{16}
$$

下图给出 EM 算法的直观解释：

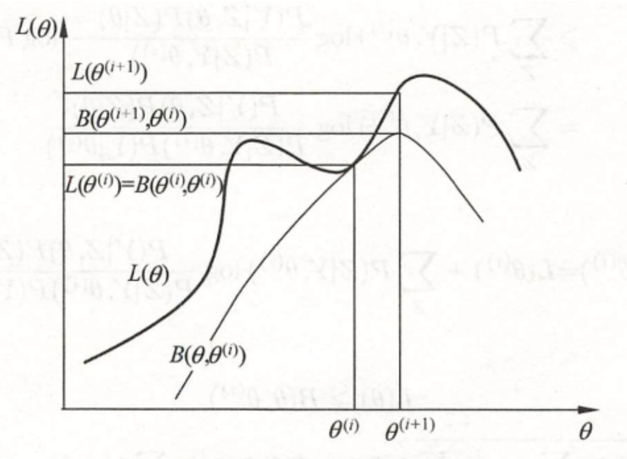


图 9.1　EM 算法的解释

1.2.2 方法二（通过条件概率公式引入隐变量）

Reference：[变分推断PPT]
对等式两边 $\log P(Y\mid\theta)=\log P(Y,Z\mid\theta)-\log P(Z\mid Y,\theta)$ 分别关于隐变量的后验分布求期望

左边得到

$$
\begin{aligned}
\text{Left} =&\sum_Z P(Z\mid Y,\theta^{(i)})\log P(Y\mid\theta)\\
=&\log P(Y\mid\theta)\sum_Z P(Z\mid Y,\theta^{(i)})\\
=&\log P(Y\mid\theta)
\end{aligned}\tag{17}
$$

右边得到

$$
\begin{aligned}
\text{Right} =&\sum_Z P(Z\mid Y,\theta^{(i)})\log P(Y,Z\mid\theta)-\sum_Z P(Z\mid Y,\theta^{(i)})\log P(Z\mid Y,\theta)\\
=&Q(\theta,\theta^{(i)})-H(\theta,\theta^{(i)})
\end{aligned}\tag{18}
$$

此处 $Q(\theta,\theta^{(i)})$ 即为 EM 算法中 M 步的优化目标，因此有 $Q(\theta^{(i+1)},\theta^{(i)})\geq Q(\theta^{(i)},\theta^{(i)})$ 。

而对于 $H(\theta,\theta^{(i)})$ ，可以证明

$$
\begin{aligned}
&H(\theta^{(i+1)},\theta^{(i)})-H(\theta^{(i)},\theta^{(i)})\\
=&\sum_Z P(Z\mid Y,\theta^{(i)})\log P(Z\mid Y,\theta^{(i+1)})-\sum_Z P(Z\mid Y,\theta^{(i)})\log P(Z\mid Y,\theta^{(i)})\\
=&\sum_Z P(Z\mid Y,\theta^{(i)})\log\frac{P(Z\mid Y,\theta^{(i+1)})}{P(Z\mid Y,\theta^{(i)})}\\
\le&\log\sum_Z P(Z\mid Y,\theta^{(i)})\cdot\frac{P(Z\mid Y,\theta^{(i+1)})}{P(Z\mid Y,\theta^{(i)})}\\
=&0
\end{aligned}
\tag{19}
$$

从而得到

$$
\begin{aligned}
&\log P(Y\mid\theta^{(i+1)})-\log P(Y\mid\theta^{(i)})\\
=&[Q(\theta^{(i+1)},\theta^{(i)})-H(\theta^{(i+1)},\theta^{(i)})]-[Q(\theta^{(i)},\theta^{(i)})-H(\theta^{(i)},\theta^{(i)})]\\
=&[Q(\theta^{(i+1)},\theta^{(i)})-Q(\theta^{(i)},\theta^{(i)})]-[H(\theta^{(i+1)},\theta^{(i)})-H(\theta^{(i)},\theta^{(i)})]\\
\ge&0
\end{aligned}
\tag{20}
$$

1.2.3 方法三（引入隐变量的近似分布，承接变分推断内容）

Reference：[变分推断PPT]
引入隐变量 Z 的某种分布 $q_\phi(Z)$

$$
\begin{aligned}
\log P(Y\mid\theta)=&\log P(Y,Z\mid\theta)-\log P(Z\mid Y,\theta)\\
=&\log\frac{P(Y,Z\mid\theta)}{q(Z)}-\log\frac{P(Z\mid Y,\theta)}{q(Z)}
\end{aligned}
\tag{21}
$$

对上式两边分别关于分布 $q(Z)$ 求期望，左边得到

$$
\begin{aligned}
\text{Left}=&\sum_Z q(Z)\log P(Y\mid\theta)\\
=&\log P(Y\mid\theta)
\end{aligned}
\tag{22}
$$

右边得到

$$
\text{Right}=\sum_Z q(Z)\log\frac{P(Y,Z\mid\theta)}{q(Z)}-\sum_Z q(Z)\log\frac{P(Z\mid Y,\theta)}{q(Z)}
\tag{23}
$$

联立得到

$$
\begin{aligned}
\underbrace{\log P(Y\mid\theta)}_{\text{evidence}}=&\sum_Z q(Z)\log\frac{P(Y,Z\mid\theta)}{q(Z)}-\sum_Z q(Z)\log\frac{P(Z\mid Y,\theta)}{q(Z)}\\
=&\underbrace{\sum_Z q(Z)\log\frac{P(Y,Z\mid\theta)}{q(Z)}}_{\text{ELBO}}+\underbrace{\sum_Z q(Z)\log\frac{q(Z)}{P(Z\mid Y,\theta)}}_{\text{KL}(q(Z))\|P(Z\mid Y,\theta)}
\end{aligned}
\tag{24}
$$

- $\log P(Y\mid\theta)$ 被称为证据（evidence）
- $\sum_Z q(Z)\log\frac{P(Y,Z\mid\theta)}{q(Z)}$ 被称为证据下界（evidence lower bound，ELBO）
- $\sum_Z q(Z)\log\frac{q(Z)}{P(Z\mid Y,\theta)}=KL(q(Z)\mid\mid P(Z\mid Y,\theta))$ 是分布 $q(Z)$ 相对于分布 $P(Z\mid Y,\theta)$ 的 KL散度（Kullback-Leibler divergence）

因为 KL 散度非负，从而得到下式，当且仅当 $q(Z)=P(Z\mid Y,\theta)$ 时取等号

$$
\underbrace{\log P(Y\mid\theta)}_{\text{evidence}}\ge\underbrace{\sum_Z q(Z)\log\frac{P(Y,Z\mid\theta)}{q(Z)}}_{\text{ELBO}}
\tag{25}
$$

E 步：固定参数 $\theta^{(i)}$， 取 $q(Z)=P(Z\mid Y,\theta^{(i)})$ ，此时有（**不严谨，为何此时取等号，疑为ppt错误**）

$$
\underbrace{\log P(Y\mid\theta)}_{\text{evidence}}=\underbrace{\sum_Z P(Z\mid Y,\theta^{(i)})\log\frac{P(Y,Z\mid\theta)}{P(Z\mid Y,\theta^{(i)})}}_{\text{ELBO}}
\tag{26}
$$

M 步：ELBO 关于参数 θ 求最大，更新参数

$$
\begin{aligned}
\theta^{(i+1)}=&\arg\max_\theta\sum_Z P(Z\mid Y,\theta^{(i)})\log\frac{P(Y,Z\mid\theta)}{P(Z\mid Y,\theta^{(i)})}\\
=&\arg\max_\theta\underbrace{\sum_Z P(Z\mid Y,\theta^{(i)})\log P(Y,Z\mid\theta)}_{Q(\theta,\theta^{(i)})}
\end{aligned}
\tag{27}
$$

笔者更正：固定参数 $\theta^{(i)}$，取 $q(Z)=P(Z\mid Y,\theta^{(i)})$，此时有（根据式24）

$$\log P(Y\mid\theta)=\underbrace{\sum_Z P(Z\mid Y,\theta^{(i)})\log\frac{P(Y,Z\mid\theta)}{P(Z\mid Y,\theta^{(i)})}}_{A}+\underbrace{\sum_Z P(Z\mid Y,\theta^{(i)})\log\frac{P(Z\mid Y,\theta^{(i)})}{P(Z\mid Y,\theta)}}_{B} \tag{28}$$

而当$\theta=\theta^{(i)}$ 时有

$$\log P(Y\mid\theta^{(i)})=\underbrace{\sum_Z P(Z\mid Y,\theta^{(i)})\log\frac{P(Y,Z\mid\theta^{(i)})}{P(Z\mid Y,\theta^{(i)})}}_{C}+\underbrace{\sum_Z P(Z\mid Y,\theta^{(i)})\log\frac{P(Z\mid Y,\theta^{(i)})}{P(Z\mid Y,\theta^{(i)})}}_{D} \tag{29}$$

由KL散度性质可知 $B\geq D=0$：

$$\begin{aligned}
&\theta^{(i+1)}=\arg\max_\theta\underbrace{\sum_Z P(Z\mid Y,\theta^{(i)})\log P(Y,Z\mid\theta)}_{Q(\theta,\theta^{(i)})}\\
&\qquad\quad=\arg\max_\theta A\\
\Rightarrow\quad &A(\theta^{(i+1)})\geq C\\
\Rightarrow\quad &\log P(Y\mid\theta^{(i+1)})\geq\log P(Y\mid\theta^{(i)})
\end{aligned} \tag{30}$$

1.3 EM 算法的收敛性

定理1：设 $L(\theta)=\log P(Y\mid\theta)$ 为观测数据的对数似然函数，$\theta^{(i)}(i=1,2,\cdots)$为 EM 算法得到的参数估计序列，$L\left(\theta^{(i)}\right)(i=1,2,\cdots)$ 为对应的对数似然函数序列。
(1) 如果 $P(Y\mid\theta)$ 有上界, 则 $L\left(\theta^{(i)}\right)=\log P\left(Y\mid\theta^{(i)}\right)$ 收敛到某一值 L^*；
(2) **在函数 $Q\left(\theta,\theta'\right)$ 与 $L(\theta)$ 满足一定条件下**，由 EM 算法得到的参数估计序列 $\theta^{(i)}$ 的收敛值 θ^* 是 $L(\theta)$ 的稳定点。

证明：
(1) 由 $L(\theta^{(i)})=\log P\left(Y\mid\theta^{(i)}\right)$ 的单调性及 $P(Y\mid\theta)$ 的有界性得到。
(2) 证明从略，参阅文献 [1983 On the convergence properties of the EM algorithm]。

1.4 EM 算法在高斯混合模型学习中的应用

定义2：高斯混合模型
高斯混合模型是指具有如下形式的概率分布模型：

$$P(y\mid\theta)=\sum_{k=1}^K\alpha_k\cdot\phi(y\mid\theta_k) \tag{31}$$

其中，α_k 是系数，$\alpha_k\geq 0$，$\sum_{k=1}^K\alpha_k=1$；$\phi(y\mid\theta_k)$ 是高斯分布密度，$\theta_k=(\mu_k,\sigma_k^2)$，

$$\phi(y\mid\theta_k)=\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right) \tag{32}$$

称为第 k 个分模型。一般混合模型可以由任意概率分布密度代替式（29）中的高斯分布密度，此处只介绍最常用的高斯混合模型。

1.4.1 高斯混合模型参数估计的 EM 算法

假设观测数据 $y_1,y_2,\ldots,y_N$ 由高斯混合模型生成，

$$P(y\mid\theta)=\sum_{k=1}^K\alpha_k\cdot\phi(y\mid\theta_k) \tag{33}$$

其中，$\theta=(\alpha_1,\alpha_2,\ldots,\alpha_K;\theta_1,\theta_2,\ldots,\theta_K)$。

观测数据的产生过程：首先依概率 $(\alpha_1,\ldots,\alpha_K)$ 选择第 k 个高斯分布模型，然后依第 k 个分模型的概率分布 $\phi(y\mid\theta_k)$ 生成观测数据 y_j。这时观测数据 y_j，$j=1,2,\ldots,N$ 是已知的；反映观测数据 y_j 来自第 k 个分模型的数据是未知的，$k=1,2,\ldots,K$，以隐变量 γ_{jk} 表示，其定义如下：

$$\gamma_{jk}=\begin{cases}1, & \text{第 } j \text{ 个观测来自第 } k \text{ 个分模型}\\ 0, & \text{否则}\end{cases} \tag{34}$$

有了观测数据 y_j 及未观测数据 γ_{jk}，那么完全数据是

$$(y_j,\gamma_{j1},\gamma_{j2},\ldots,\gamma_{jK}),\quad j=1,2,\ldots,N \tag{35}$$

于是可以写出完全数据的似然函数

$$
\begin{aligned}
P(y,\gamma \mid \theta) &= \prod_{j=1}^{N} P(y_j,\gamma_{j1},\gamma_{j2},\ldots,\gamma_{jK} \mid \theta) \\
&= \prod_{k=1}^{K}\prod_{j=1}^{N}[\alpha_k \cdot \phi(y_j \mid \theta_k)]^{\gamma_{jk}} \\
&= \prod_{k=1}^{K}\alpha_k^{n_k}\prod_{j=1}^{N}[\phi(y_j \mid \theta_k)]^{\gamma_{jk}} \\
&= \prod_{k=1}^{K}\alpha_k^{n_k}\prod_{j=1}^{N}\left[\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y_j-\mu_k)^2}{2\sigma_k^2}\right)\right]^{\gamma_{jk}}
\end{aligned}
\tag{36}
$$

式中，$n_k=\sum_{j=1}^{N}\gamma_{jk}$，$\sum_{k=1}^{K}n_k=N$。

那么，完全数据的对数似然函数为

$$
\log P(y,\gamma \mid \theta)=\sum_{k=1}^{K}\left\{n_k\log\alpha_k+\sum_{j=1}^{N}\gamma_{jk}\left[\log\left(\frac{1}{\sqrt{2\pi}}\right)-\log\sigma_k-\frac{1}{2\sigma_k^2}(y_j-\mu_k)^2\right]\right\}
\tag{37}
$$

进一步计算 Q 函数

$$
\begin{aligned}
Q(\theta,\theta^{(i)}) &= E_\gamma\left[\log P(y,\gamma \mid \theta) \mid y,\theta^{(i)}\right] \\
&= E_\gamma\left\{\sum_{k=1}^{K}\left\{n_k\log\alpha_k+\sum_{j=1}^{N}\gamma_{jk}\left[\log\left(\frac{1}{\sqrt{2\pi}}\right)-\log\sigma_k-\frac{1}{2\sigma_k^2}(y_j-\mu_k)^2\right]\right\}\right\} \\
&= \sum_{k=1}^{K}\left\{\sum_{j=1}^{N}(E[\gamma_{jk}])\log\alpha_k+\sum_{j=1}^{N}(E[\gamma_{jk}])\left[\log\left(\frac{1}{\sqrt{2\pi}}\right)-\log\sigma_k-\frac{1}{2\sigma_k^2}(y_j-\mu_k)^2\right]\right\}
\end{aligned}
\tag{38}
$$

这里需要计算 $E(\gamma_{jk} \mid y,\theta)$，记为 $\hat{\gamma}_{jk}$ 。

$$
\begin{aligned}
\hat{\gamma}_{jk} &= E\left(\gamma_{jk} \mid y,\theta\right)=P\left(\gamma_{jk}=1 \mid y,\theta\right) \\
&= \frac{P\left(\gamma_{jk}=1,y_j \mid \theta\right)}{\sum_{k=1}^{K}P\left(\gamma_{jk}=1,y_j \mid \theta\right)} \\
&= \frac{P\left(y_j \mid \gamma_{jk}=1,\theta\right)P\left(\gamma_{jk}=1 \mid \theta\right)}{\sum_{k=1}^{K}P\left(y_j \mid \gamma_{jk}=1,\theta\right)P\left(\gamma_{jk}=1 \mid \theta\right)} \\
&= \frac{\alpha_k\phi\left(y_j \mid \theta_k\right)}{\sum_{k=1}^{K}\alpha_k\phi\left(y_j \mid \theta_k\right)}, \quad j=1,2,\cdots,N; \quad k=1,2,\cdots,K
\end{aligned}
\tag{39}
$$

$\hat{\gamma}_{jk}$ 是在当前模型参数下第 j 个观测数据来自第 k 个分模型的概率，称为分模型 k 对观测数据 y_j 的响应度。将$\hat{\gamma}_{jk}=E[\gamma_{jk}]$ 及 $\hat{n}_k=\sum_{j=1}^{N}E[\gamma_{jk}]$ 代入式（38），即得

$$
Q\left(\theta,\theta^{(i)}\right)=\sum_{k=1}^{K}\left\{\hat{n}_k\log\alpha_k+\sum_{j=1}^{N}\hat{\gamma}_{jk}\left[\log\left(\frac{1}{\sqrt{2\pi}}\right)-\log\sigma_k-\frac{1}{2\sigma_k^2}(y_j-\mu_k)^2\right]\right\}
\tag{40}
$$

迭代的 M 步是求函数 $Q(\theta,\theta^{(i)})$ 对 θ 的极大值，即求新一轮迭代的模型参数

$$
\theta^{(i+1)}=\arg\max_{\theta}Q(\theta,\theta^{(i)})
\tag{41}
$$

用 $\hat{\mu}_k$，$\hat{\sigma}_k^2$ 及 $\hat{\alpha}_k$ ，$k=1,2,\ldots,K$，表示 $\theta^{(i+1)}$ 的各参数。求 $\hat{\mu}_k$，$\hat{\sigma}_k^2$ 只需将式（40）分别对 $\hat{\mu}_k$，$\hat{\sigma}_k^2$ 求偏导数并令其为 0，即可得到；求 $\hat{\alpha}_k$ 是在 $\sum_{k=1}^{K}\alpha_k=1$ 条件下求偏导数并令其为 0 得到的（可拆解验证其凹性）。结果如下

$$
\hat{\mu}_k=\frac{\sum_{j=1}^{N}\hat{\gamma}_{jk}\cdot y_j}{\sum_{j=1}^{N}\hat{\gamma}_{jk}}, \quad k=1,2,\ldots,K
\tag{42}
$$

$$
\hat{\sigma}_k^2=\frac{\sum_{j=1}^{N}\hat{\gamma}_{jk}(y_j-\mu_k)^2}{\sum_{j=1}^{N}\hat{\gamma}_{jk}}, \quad k=1,2,\ldots,K
\tag{43}
$$

$$
\hat{\alpha}_k=\frac{\hat{n}_k}{N}=\frac{\sum_{j=1}^{N}\hat{\gamma}_{jk}}{N}, \quad k=1,2,\ldots,K
\tag{44}
$$

重复以上计算，直到对数似然函数值不再有明显的变化为止。

1.5 EM 算法的推广（可参考变分推断PPT更简单易理解）

1.5.1 F 函数的极大-极大算法

定义3：F 函数

假设隐变量数据 Z 的概率分布为 $\tilde{P}(Z)$，定义分布 $\tilde{P}$ 与参数 θ 的函数 $F(\tilde{P},\theta)$ 如下

$$F(\tilde{P},\theta)=E_{\tilde{P}}[\log P(Y,Z\mid\theta)]+H(\tilde{P}) \tag{45}$$

称为 F 函数，式中 $H(\tilde{P})=-E_{\tilde{P}}\log\tilde{P}(Z)$ 是分布 $\tilde{P}(Z)$ 的熵。

在定义3中，通常假设 $P(Y,Z\mid\theta)$ 是 θ 的连续函数，因而 $F(\tilde{P},\theta)$ 是 $\tilde{P}$ 和 θ 的连续函数。函数 $F(\tilde{P},\theta)$ 还有以下重要性质。
引理1：
对于固定的 θ，存在唯一的分布 $\tilde{P}_\theta$ 极大化 $F(\tilde{P},\theta)$，这时 $\tilde{P}_\theta$ 由下式给出

$$\tilde{P}_\theta(Z)=P(Z\mid Y,\theta) \tag{46}$$

并且 $\tilde{P}_\theta$ 随 θ 连续变化。

证明：
对于固定的 θ，可以求得使 $F(\tilde{P},\theta)$ 达到极大的分布 $\tilde{P}_\theta(Z)$。为此，引进拉格朗日乘子 λ，拉格朗日函数为

$$L=E_{\tilde{P}}\log P(Y,Z\mid\theta)-E_{\tilde{P}}\log\tilde{P}(Z)+\lambda\left(1-\sum_Z\tilde{P}(Z)\right) \tag{47}$$

将其对 $\tilde{P}$ 求偏导数（针对特定 Z）

$$\frac{\partial L}{\partial\tilde{P}(Z)}=\log P(Y,Z\mid\theta)-\log\tilde{P}(Z)-1-\lambda \tag{48}$$

令偏导数等于 0，得出

$$\lambda=\log P(Y,Z\mid\theta)-\log\tilde{P}_\theta(Z)-1 \tag{49}$$

由此推出 $\tilde{P}_\theta(Z)$ 与 $P(Y,Z\mid\theta)$ 成比例

$$\frac{P(Y,Z\mid\theta)}{\tilde{P}_\theta(Z)}=\exp(1+\lambda) \tag{50}$$

再从约束条件 $\sum_Z\tilde{P}_\theta(Z)=1$ 得到式（46）。
由假设 $P(Y,Z\mid\theta)$ 是 θ 的连续函数，得到 $\tilde{P}_\theta$ 是 θ 的连续函数。

引理2：
若 $\tilde{P}_\theta(Z)=P(Z\mid Y,\theta)$，则

$$F(\tilde{P},\theta)=\log P(Y\mid\theta) \tag{51}$$

证明：

$$\begin{aligned}F(\tilde{P},\theta)=&E_{\tilde{P}}[\log P(Y,Z\mid\theta)]+H(\tilde{P})\\=&E_{\tilde{P}}[\log\frac{P(Y,Z\mid\theta)}{P(Z\mid Y,\theta)}]\\=&\sum_Z P(Z\mid Y,\theta)\cdot\log\frac{P(Y,Z\mid\theta)}{P(Z\mid Y,\theta)}\\=&\sum_Z P(Z\mid Y,\theta)\cdot\log P(Y\mid\theta)\\=&\log P(Y\mid\theta)\end{aligned} \tag{52}$$

由以上引理，可以得到关于 EM 算法用 F 函数的极大-极大算法的解释。
定理2：
设 $L(\theta)=\log P(Y\mid\theta)$ 为观测数据的对数似然函数，$\theta^{(i)},i=1,2,\ldots$，为 EM 算法得到的参数估计序列，函数 $F(\tilde{P},\theta)$ 由式（45）定义。如果 $F(\tilde{P},\theta)$ 在 $\tilde{P}^*$ 和 θ^* 有局部极大值，那么 $L(\theta)$ 也在 θ^* 有局部极大值。类似地，如果 $F(\tilde{P},\theta)$ 在 $\tilde{P}^*$ 和 θ^* 达到全局最大值，那么 $L(\theta)$ 也在 θ^* 达到全局最大值。
证明：
由引理1和引理2可知，$L(\theta)=\log P(Y\mid\theta)=F(\tilde{P}_\theta,\theta)$ **对任意 θ 成立**。特别地，对于使 $F(\tilde{P},\theta)$ 达到极大的参数 θ^*，有

$$L(\theta^*)=F(\tilde{P}_{\theta^*},\theta^*)=F(\tilde{P}^*,\theta^*) \tag{53}$$

为了证明 θ^* 是 $L(\theta)$ 的极大点，需要证明不存在接近 θ^* 的点 θ^{**}，使 $L(\theta^{**})>L(\theta^*)$。假如存在这样的点 θ^{**}，那么应有 $F(\tilde{P}^{**},\theta^{**})>F(\tilde{P}^*,\theta^*)$，这里 $\tilde{P}^{**}=\tilde{P}_{\theta^{**}}$。但因 $\tilde{P}_\theta$ 是随 θ 连续变化的，$\tilde{P}^{**}$ 应接近 $\tilde{P}^*$，这与 $\tilde{P}^*$ 和 θ^* 是 $F(\tilde{P},\theta)$ 的局部极大点的假设矛盾。类似可以证明关于全局最大值的讨论。

定理3：
EM 算法的一次迭代可由 F 函数的极大-极大算法实现。
设 $\theta^{(i)}$ 为第 i 次迭代参数 θ 的估计, $\tilde{P}^{(i)}$ 为第 i 次迭代函数 $\tilde{P}$ 的估计。在第 $i+1$次迭代的两步为：
(1) 对固定的 $\theta^{(i)}$, 求 $\tilde{P}^{(i+1)}$ 使 $F\left(\tilde{P},\theta^{(i)}\right)$ 极大化;

(2) 对固定的 $\tilde{P}^{(i+1)}$, 求 $\theta^{(i+1)}$ 使 $F\left(\tilde{P}^{(i+1)}, \theta\right)$ 极大化。
证明:
(1) 由引理 1, 对于固定的 $\theta^{(i)}$,

$$\tilde{P}^{(i+1)}(Z)=\tilde{P}_{\theta^{(i)}}(Z)=P(Z \mid Y, \theta^{(i)}) \tag{54}$$

使 $F(\tilde{P}, \theta^{(i)})$ 极大化。此时，

$$\begin{aligned} F(\tilde{P}^{(i+1)}, \theta) &= E_{\tilde{P}^{(i+1)}}[\log P(Y, Z \mid \theta)]+H(\tilde{P}^{(i+1)}) \\ &= \sum_{Z} \log P(Y, Z \mid \theta) P(Z \mid Y, \theta^{(i)})+H(\tilde{P}^{(i+1)}) \end{aligned} \tag{55}$$

由 $Q\left(\theta, \theta^{(i)}\right)$ 的定义式 11 有

$$F\left(\tilde{P}^{(i+1)}, \theta\right)=Q\left(\theta, \theta^{(i)}\right)+H\left(\tilde{P}^{(i+1)}\right) \tag{56}$$

(2) 固定 $\tilde{P}^{(i+1)}$，求 $\theta^{(i+1)}$ 使 $F\left(\tilde{P}^{(i+1)}, \theta\right)$ 极大化。得到

$$\theta^{(i+1)}=\arg \max_{\theta} F\left(\tilde{P}^{(i+1)}, \theta\right)=\arg \max_{\theta} Q\left(\theta, \theta^{(i)}\right) \tag{57}$$

通过以上两步完成了 EM 算法的一次迭代。由此可知，由 EM算法与 F 函数的极大-极大算法得到的参数估计序列 $\theta^{(i)}, i=1,2, \cdots$，是一致的。

1.5.2 GEM 算法

算法2：
输入：观测数据， F 函数;
输出：模型参数。
(1) 初始化参数 $\theta^{(0)}$，开始迭代;
(2) 第 $i+1$ 次迭代，第 1 步：记 $\theta^{(i)}$ 为参数 θ 的估计值，$\tilde{P}^{(i)}$ 为函数 $\tilde{P}$ 的估计，求 $\tilde{P}^{(i+1)}$ 使 $\tilde{P}$ 极大化 $F\left(\tilde{P}, \theta^{(i)}\right)$;
(3) 第 2 步：求 $\theta^{(i+1)}$ 使 $F\left(\tilde{P}^{(i+1)}, \theta\right)$ 极大化;
(4) 重复 (2) 和 (3)，直到收敛。

在 GEM 算法 1 中，有时求 $Q\left(\theta, \theta^{(i)}\right)$ 的极大化是很困难的。下面介绍的 GEM 算法 2 和 GEM 算法 3 并不是直接求 $\theta^{(i+1)}$ 使 $Q\left(\theta, \theta^{(i)}\right)$ 达到极大的 θ，而是找一个 $\theta^{(i+1)}$ 使得 $Q\left(\theta^{(i+1)}, \theta^{(i)}\right)>Q\left(\theta^{(i)}, \theta^{(i)}\right)$ 。

算法3：
输入：观测数据, Q 函数;
输出：模型参数。
(1) 初始化参数 $\theta^{(0)}$，开始迭代;
(2) 第 $i+1$ 次迭代，第 1 步：记 $\theta^{(i)}$ 为参数 θ 的估计值，计算

$$\begin{aligned} Q\left(\theta, \theta^{(i)}\right) &= E_{Z}\left[\log P(Y, Z \mid \theta) \mid Y, \theta^{(i)}\right] \\ &= \sum_{Z} P\left(Z \mid Y, \theta^{(i)}\right) \log P(Y, Z \mid \theta) \end{aligned} \tag{58}$$

(3) 第 2 步：求 $\theta^{(i+1)}$ 使

$$Q\left(\theta^{(i+1)}, \theta^{(i)}\right)>Q\left(\theta^{(i)}, \theta^{(i)}\right) \tag{59}$$

(4) 重复 (2) 和 (3)，直到收敛。

当参数 θ 的维数为 $d(d \geqslant 2)$ 时，可采用一种特殊的 GEM 算法，它将 EM 算法的 M 步分解为 d 次条件极大化，每次只改变参数向量的一个分量，其余分量不改变。

算法4：
输入：观测数据，Q 函数;
输出：模型参数。
(1) 初始化参数 $\theta^{(0)}=\left(\theta_{1}^{(0)}, \theta_{2}^{(0)}, \cdots, \theta_{d}^{(0)}\right)$，开始迭代;
(2) 第 $i+1$ 次迭代，第 1 步：记 $\theta^{(i)}=\left(\theta_{1}^{(i)}, \theta_{2}^{(i)}, \cdots, \theta_{d}^{(i)}\right)$ 为参数 $\theta=(\theta_{1}, \theta_{2}, \cdots, \theta_{d})$的估计值，计算

$$\begin{aligned} Q\left(\theta, \theta^{(i)}\right) &= E_{Z}\left[\log P(Y, Z \mid \theta) \mid Y, \theta^{(i)}\right] \\ &= \sum_{Z} P\left(Z \mid y, \theta^{(i)}\right) \log P(Y, Z \mid \theta) \end{aligned} \tag{60}$$

(3) 第 2 步：进行 d 次条件极大化：
首先，在 $\theta_2^{(i)},\cdots,\theta_d^{(i)}$ 保持不变的条件下求使 $Q\left(\theta,\theta^{(i)}\right)$ 达到极大的 $\theta_1^{(i+1)}$；然后，在 $\theta_1=\theta_1^{(i+1)},\theta_j=\theta_j^{(i)},j=3,4,\cdots,d$ 的条件下求使 $Q\left(\theta,\theta^{(i)}\right)$ 达到极大的 $\theta_2^{(i+1)}$；
如此继续，经过 d 次条件极大化，得到 $\theta^{(i+1)}=\left(\theta_1^{(i+1)},\theta_2^{(i+1)},\cdots,\theta_d^{(i+1)}\right)$ 使得

$$Q\left(\theta^{(i+1)},\theta^{(i)}\right)>Q\left(\theta^{(i)},\theta^{(i)}\right) \tag{61}$$

(4) 重复 (2) 和 (3)，直到收敛。

2. 变分推断

Reference：[变分推断PPT]

2.1 变分推断介绍

变分推断（Variational Inference，VI）是贝叶斯学习中常用的、含有隐变量模型的学习和推断方法。变分推断和马尔科夫链蒙特卡洛法（MCMC）属于不同的技巧：

- MCMC通过随机抽样的方法近似地计算模型的后验概率（采样），适合小数据集以及精确度更重要的场景
- 变分推断通过解析的方法计算模型的后验概率的近似值（优化），适合大数据集以及想快速测试多种模型的场景

为什么关心后验概率 $P(\theta\mid X)$？

1. 推断（Beyesian Inference）：后验分布 $P(\theta\mid X)$ 包含了模型的重要信息，描述了数据样本产生的过程，例如从用户的观影历史评分信息 Y 中推断用户的偏好模型 θ
2. 决策（Beyesian Dicision Theory）：对于新样本 $\tilde{x}$，求 $P(\tilde{x}\mid X)$

$$\begin{aligned}P(\tilde{x}\mid X)&=\int_\theta P(\tilde{x},\theta\mid X)\,\mathrm{d}\theta\\&=\int_\theta P(\tilde{x}\mid\theta)P(\theta\mid X)\,\mathrm{d}\theta\\&=E_{\theta\mid X}[P(\tilde{x}\mid\theta)]\end{aligned} \tag{62}$$

   被称为后验预测分布（Posterior predictive distribution），例如根据用户的历史评分信息 X 预测用户对于新电影 $\tilde{x}$ 的评分

2.2 变分推断推导

贝叶斯参数学习问题的描述：

- X 观测数据
- Z 隐变量+参数
- θ 超参数
  注意，这里的符号表示和 EM 算法中的表述有区别，贝叶斯参数学习需要推断的是 Z 中的参数，及学习后验分布 $P(Z\mid\theta)$

首先是 evidence 的分解

$$\underbrace{\log P(X\mid\theta)}_{\text{evidence}}=\underbrace{\int_Z q(Z)\log\frac{P(X,Z\mid\theta)}{q(Z)}\,\mathrm{d}Z}_{\text{ELBO}}+\underbrace{\int_Z q(Z)\log\frac{q(Z)}{P(Z\mid X,\theta)}\,\mathrm{d}Z}_{\mathrm{KL}(q(Z)\|P(Z\mid X,\theta))} \tag{63}$$

当我们知道超参数 θ 时，上式中 evidence 应是固定的，因为 $\log P(X\mid\theta)=\log\sum_Z P(X,Z\mid\theta)$，虽然这个值通常求不出来。

变分推断的目标是通过最小化 $\mathrm{KL}(q(Z)\,||\,P(Z\mid X,\theta))$ 来寻找与后验分布 $P(Z\mid X,\theta)$ 最相似的变分分布 $q(Z)$。

$$q(Z)^*=\arg\min_{q(Z)}\mathrm{KL}(q(Z)\,||\,P(Z\mid X,\theta)) \tag{64}$$

后验分布 $P(Z\mid X,\theta)$ 太复杂，直接估计其密度很困难，但利用 KL 散度和 ELBO 的和为常数，可以转而求

$$\begin{aligned}q(Z)^*&=\arg\min_{q(Z)}\mathrm{KL}(q(Z)\,||\,P(Z\mid X,\theta))\\&=\arg\max_{q(Z)}\mathrm{ELBO}\\&=\arg\max_{q(Z)}\int_Z q(Z)\log\frac{P(X,Z\mid\theta)}{q(Z)}\,\mathrm{d}Z\\&=\arg\max_{q(Z)}\int_Z q(Z)\log P(X,Z\mid\theta)\,\mathrm{d}Z-\int_Z q(Z)\log q(Z)\,\mathrm{d}Z\\&=\arg\max_{q(Z)}E_{q(Z)}[\log P(X,Z\mid\theta)]-E_{q(Z)}[\log q(Z)]\end{aligned} \tag{65}$$

变分分布 $q(Z)$ 有多种参数化方法，要求参数化后的 $q(Z)$ 使得上述优化问题容易求解，一种常用的方法是假设 $q(Z)$ 对 $Z=(Z_1,Z_2,\ldots,Z_d)$ 的所有分量 Z_j 都是相互独立的（实际是条件独立于参数），即满足

$$q(Z)=q(Z_1)\cdot q(Z_2)\cdots q(Z_d) \tag{66}$$

这时的变分分布被称为满足平均场（mean field）假设。

KL 散度的最小化或 ELBO 的最大化实际是在平均场的集合，即满足独立假设的分布集合 $Q=\{q(Z)\mid q(Z)=\prod_{j=1}^{d}q(Z_j)\}$ 之中进行的

$$q(Z)^*=\arg\max_{q(Z)\in Q}E_{q(Z)}[\log P(X,Z\mid\theta)]-E_{q(Z)}[\log q(Z)] \tag{67}$$

Reference：[intermediate_vb, PRML chapter 10]

现在我们将目标函数重新写为

$$\begin{aligned}\text{ELBO}&=\int q_\phi(\mathbf{z})\log\left(\frac{p(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z})}\right)\mathrm{d}\mathbf{z}\\&=\int q_\phi(\mathbf{z})\log(p(\mathbf{x},\mathbf{z}))\mathrm{d}\mathbf{z}-\int q_\phi(\mathbf{z})\log(q_\phi(\mathbf{z}))\mathrm{d}\mathbf{z}\\&=\underbrace{\int\prod_{i=1}^{M}q_i(z_i)\log(p(\mathbf{x},\mathbf{z}))\mathrm{d}\mathbf{z}}_{-H(q,p)}+\underbrace{\left(-\int\prod_{i=1}^{M}q_i(z_i)\sum_{i=1}^{M}\log(q_i(z_i))\mathrm{d}\mathbf{z}\right)}_{H(q)}\end{aligned} \tag{68}$$

首先考虑第一部分 $-H(q,p)$

$$\begin{aligned}-H(q,p)&=\int\prod_{i=1}^{M}q_i(z_i)\log(p(\mathbf{x},\mathbf{z}))\mathrm{d}\mathbf{z}\\&=\int_{Z_1}\int_{Z_2}\cdots\int_{Z_M}\prod_{i=1}^{M}q_i(z_i)\log(p(\mathbf{x},\mathbf{z}))\mathrm{d}\mathbf{z}_1\mathrm{d}\mathbf{z}_2\cdots\mathrm{d}\mathbf{z}_M\end{aligned} \tag{69}$$

只考虑其中一项 $q_j(z_j)$

$$\begin{aligned}-H(q,p)_j&=\int_{Z_j}q_j(z_j)\left(\int\cdots\int_{Z_{i\neq j}}\prod_{i\neq j}^{M}q_i(z_i)\log(p(\mathbf{x},\mathbf{z}))\prod_{i\neq j}^{M}\mathrm{d}z_i\right)\mathrm{d}z_j\\&=\int_{Z_j}q_j(z_j)\mathbb{E}_{\mathbf{z}\setminus z_j}[\log(p(\mathbf{x},\mathbf{z}))]\mathrm{d}z_j\end{aligned} \tag{70}$$

再考虑第二部分 $H(q)$

$$\begin{aligned}H(q)&=-\int\left(\prod_{i=1}^{M}q_i(z_i)\right)\sum_{i=1}^{M}\log(q_i(z_i))\mathrm{d}\mathbf{z}\\&=\sum_{i=1}^{M}\left(-\int_{Z_i}q_i(z_i)\log(q_i(z_i))\mathrm{d}z_i\right)\\&=\sum_{i=1}^{M}H(q_i(z_i))\end{aligned} \tag{71}$$

仅考虑其中一项 $q_j(z_j)$

$$\begin{aligned}H(q)_j&=-\int_{Z_j}q_j(z_j)\log(q_j(z_j))\mathrm{d}z_j+\text{Const.}\\&=H(q_j(z_j))+\text{Const.}\end{aligned} \tag{72}$$

针对 ELBO 只考虑优化 q_j

$$\begin{aligned}\text{ELBO}(q_j)&=-H(q,p)_j+H(q)_j\\&=\int_{Z_j}q_j(z_j)\mathbb{E}_{\mathbf{z}\setminus z_j}[\log(p(\mathbf{x},\mathbf{z}))]\mathrm{d}z_j-\int_{Z_j}q_j(z_j)\log(q_j(z_j))\mathrm{d}z_j+\text{Const.}1\\&=\int_{Z_j}q_j(z_j)\log\tilde{p}(\mathbf{x},z_j)\mathrm{d}z_j-\int_{Z_j}q_j(z_j)\log(q_j(z_j))\mathrm{d}z_j+\text{Const.}2\\&=\int_{Z_j}q_j(z_j)\log\left[\frac{\tilde{p}(\mathbf{x},z_j)}{q_j(z_j)}\right]\mathrm{d}z_j+\text{Const.}2\\&=-\mathbb{KL}\left(q_j(z_j)\,||\,\tilde{p}(\mathbf{x},z_j)\right)+\text{Const.}2\end{aligned} \tag{73}$$

这里我们定义了一个新分布 $\tilde{p}(\mathbf{x}, z_j)$

$$\log \tilde{p}(\mathbf{x}, z_j) = \mathbb{E}_{\mathbf{z}\setminus z_j}[\log(p(\mathbf{x}, \mathbf{z}))] + \text{Const.} \tag{74}$$

因此我们可以通过最小化下述 KL 散度来最大化 ELBO，而 KL 散度的性质可知其值为零时最小，即 $q_j^* = \tilde{p}(\mathbf{x}, z_j)$

$$\log q_j^*(z_j) = \mathbb{E}_{\mathbf{z}\setminus z_j}[\log(p(\mathbf{x}, \mathbf{z}))] + \text{Const.} \tag{75}$$

注意此处的 $\exp(\mathbb{E}_{\mathbf{z}\setminus z_j}[\log(p(\mathbf{x}, \mathbf{z}))])$ 是伪概率分布（pseudo distribution），只能满足概率分布的非负性而不能保证具有归一性，常数项为归一化常数 $\int \exp(\mathbb{E}_{\mathbf{z}\setminus z_j}[\log(p(\mathbf{x}, \mathbf{z}))]) \mathrm{d}z_j$，保证 $\tilde{p}$ 的归一性和非负性，因此有

$$q_j^*(z_j) = \frac{\exp(\mathbb{E}_{\mathbf{z}\setminus z_j}[\log(p(\mathbf{x}, \mathbf{z}))])}{\int \exp(\mathbb{E}_{\mathbf{z}\setminus z_j}[\log(p(\mathbf{x}, \mathbf{z}))]) \mathrm{d}z_j} \tag{76}$$

2.3 Gaussian-Gamma

可观测变量为 $\mathcal{D} = \{x_1, \ldots, x_n\}$，似然为

$$\begin{aligned} p(\mathcal{D} \mid \mu, \tau) &= \prod_{i=1}^{n} \left(\frac{\tau}{2\pi}\right)^{\frac{1}{2}} \exp\left(\frac{-\tau}{2}(x_i - \mu)^2\right) \\ &= \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left(\frac{-\tau}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \end{aligned} \tag{77}$$

假设先验为

$$\begin{aligned} p(\mu \mid \tau) &= \mathcal{N}(\mu_0, (\lambda_0\tau)^{-1}) \propto \exp\left(\frac{-\lambda_0\tau}{2}(\mu - \mu_0)^2\right) \\ p(\tau) &= \text{Gamma}(\tau \mid a_0, b_0) \propto \tau^{a_0 - 1}\exp(-b_0\tau) \end{aligned} \tag{78}$$

利用共轭性质可以计算解析后验（在2.4节介绍）

$$\begin{aligned} p(\mu, \tau \mid \mathcal{D}) \propto& p(\mathcal{D} \mid \mu, \tau) p(\mu \mid \tau) p(\tau) \\ =& \mathcal{N}(\mu_n, (\lambda_n\tau)^{-1}) \text{Gamma}(\tau \mid a_n, b_n) \end{aligned} \tag{79}$$

此处

$$\begin{aligned} \mu_n &= \frac{\lambda_0\mu_0 + n\bar{x}}{\lambda_0 + n} \\ \lambda_n &= \lambda_0 + n \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\lambda_0 n(\bar{x} - \mu_0)^2}{2(\lambda_0 + n)} \end{aligned} \tag{80}$$

但是如果我们不能计算其解析后验，可用变分推断来近似其后验。假设变分分布 $q(\mathbf{z})$ 为

$$q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau) \tag{81}$$

利用式（75）得出的结论

$$\begin{aligned} \log q_\mu^*(\mu) &= \mathbb{E}_{q_\tau(\tau)}\left[\log p(\mu, \tau, \mathcal{D})\right] \\ &= \mathbb{E}_{q_\tau(\tau)}\left[\log p(\mathcal{D} \mid \mu, \tau) + \log p(\mu \mid \tau)\right] + \text{Const.} \\ &= \mathbb{E}_{q_\tau(\tau)}\left[\frac{n}{2}\log\tau - \frac{\tau}{2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{\lambda_0\tau}{2}(\mu - \mu_0)^2\right] + \text{Const.} \\ &= -\frac{1}{2}\mathbb{E}_{q_\tau}[\tau] \underbrace{\left[\sum_{i=1}^{n}(x_i - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right]}_{\text{terms taking out of } \tau} + \text{Const.} \end{aligned} \tag{82}$$

将中括号内式子展开，形成高斯分布 $\mathcal{N}(\mu; \mu^*, \tau^*)$ 的形式（如果一个连续随机变量的对数概率密度函数是关于该变量的严格负二次函数，并且该函数可以被正规化（即对应的密度可积且积分有限），那么该随机变量必定服从高斯分布）

$$\begin{aligned} \sum_{i=1}^{n}(x_i - \mu)^2 + \lambda_0(\mu - \mu_0)^2 &= n\mu^2 - 2n\mu\bar{\mathbf{x}} + \lambda_0\mu^2 - 2\lambda_0\mu_0\mu + \text{Const.} \\ &= (n + \lambda_0)\mu^2 - 2\mu\left(n\bar{\mathbf{x}} + \lambda_0\mu_0\right) + \text{Const.} \\ &= (n + \lambda_0)\left(\mu^2 - \frac{2\mu\left(n\bar{\mathbf{x}} + \lambda_0\mu_0\right)}{n + \lambda_0}\right) + \text{Const.} \\ &= (n + \lambda_0)\left(\mu - \frac{n\bar{\mathbf{x}} + \lambda_0\mu_0}{n + \lambda_0}\right)^2 + \text{Const.} \end{aligned} \tag{83}$$

因此我们有

$$
\begin{aligned}
\log q_\mu^*(\mu) &= -\frac{\mathbb{E}_{q_\tau}[\tau]}{2}\left[\sum_{i=1}^{n}(x_i-\mu)^2+\lambda_0(\mu-\mu_0)^2\right]+\text{Const.}\\
&= -\frac{\mathbb{E}_{q_\tau}[\tau]\,(n+\lambda_0)}{2}\left(\mu-\frac{n\overline{\mathbf{x}}+\lambda_0\mu_0}{n+\lambda_0}\right)^2+\text{Const.}\\
&= -\frac{1}{2}\underbrace{\mathbb{E}_{q_\tau}[\tau]\,(n+\lambda_0)}_{\tau^*}\left(\mu-\underbrace{\frac{n\overline{\mathbf{x}}+\lambda_0\mu_0}{n+\lambda_0}}_{\mu^*}\right)^2+\text{Const.}\\
\Longrightarrow q_\mu^*(\mu) &= \mathcal{N}\left(\frac{n\overline{\mathbf{x}}+\lambda_0\mu_0}{n+\lambda_0},\mathbb{E}_{q_\tau}[\tau]\,(n+\lambda_0)\right)\quad \because -\frac{\tau}{2}(x-\mu)^2
\end{aligned}
\tag{84}
$$

利用式（82），去掉期望符号 $\mathbb{E}_{q_\tau}[\cdot]$，我们还可以得到 $p(\mu\mid\mathcal{D},\tau)$（注意，删掉期望值就是原分布的后验，因为 $p(\mathcal{D},\tau)$ 在常数项里）

$$
\begin{aligned}
\log p(\mathcal{D}\mid\mu,\tau)+\log p(\mu\mid\gamma) &= \underbrace{-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2}_{\log(p(\mathcal{D}\mid\mu,\tau))}\underbrace{-\frac{\lambda_0\tau}{2}(\mu-\mu_0)^2}_{\log p(\mu\mid\gamma)}+\text{Const.}\\
&= -\frac{\tau}{2}\left[\sum_{i=1}^{n}(x_i-\mu)^2+\lambda_0(\mu-\mu_0)^2\right]+\text{Const.}\\
&= -\frac{\tau\,(n+\lambda_0)}{2}\left(\mu-\frac{n\overline{\mathbf{x}}+\lambda_0\mu_0}{n+\lambda_0}\right)^2+\text{Const.}\\
\Longrightarrow p(\mu\mid\mathcal{D},\tau) &= \mathcal{N}\left(\frac{n\overline{\mathbf{x}}+\lambda_0\mu_0}{n+\lambda_0},\tau\,(n+\lambda_0)\right)
\end{aligned}
\tag{85}
$$

同理我们可以计算 $\log q_\tau^*(\tau)$

$$
\begin{aligned}
\log q_\tau^*(\tau) &= \mathbb{E}_{q_\mu}[\log p(\mu,\tau,\mathcal{D})]\\
&= \mathbb{E}_{q_\mu}[\log p(\mathcal{D}\mid\mu,\tau)+\log p(\mu\mid\tau)+\log p(\tau)]+\text{Const.}\\
&= \mathbb{E}_{q_\mu}[\underbrace{\frac{n}{2}\log(\tau)-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2}_{\log p(\mathcal{D}\mid\mu,\tau)}\underbrace{-\frac{\lambda_0\tau}{2}(\mu-\mu_0)^2}_{\log p(\mu\mid\gamma)}\underbrace{+(a_0-1)\log(\tau)-b_0\tau}_{\log p(\tau)}]+\text{Const.}\\
&= \frac{n}{2}\log\tau+(a_0-1)\log\tau-b_0\tau-\frac{\tau}{2}\mathbb{E}_{q_\mu(\mu)}\left[\sum_{i=1}^{n}(x_i-\mu)^2+\lambda_0(\mu-\mu_0)^2\right]+\text{Const.}\\
&= \left(\underbrace{\frac{n}{2}+a_0}_{a_n}-1\right)\log\tau-\tau\left(\underbrace{b_0+\frac{1}{2}\mathbb{E}_{q_\mu(\mu)}\left[\sum_{i=1}^{n}(x_i-\mu)^2+\lambda_0(\mu-\mu_0)^2\right]}_{b_n}\right)+\text{Const.}\\
&\Longrightarrow q_\tau^*(\tau)=\text{Gamma}(a_n,b_n)
\end{aligned}
\tag{86}
$$

可以将 b_n 展开写为

$$
\begin{aligned}
b_n &= b_0+\frac{1}{2}\mathbb{E}_{q_\mu}\left[\sum_{i=1}^{n}(x_i-\mu)^2+\lambda_0(\mu-\mu_0)^2\right]\\
&= b_0+\frac{1}{2}\mathbb{E}_{q_\mu}\left[-2\mu n\bar{x}+n\mu^2+\lambda_0\mu^2-2\lambda_0\mu_0\mu\right]+\sum_{i=1}^{n}x_i^2+\lambda_0\mu_0^2\\
&= b_0+\frac{1}{2}\left[(n+\lambda_0)\mathbb{E}_{q_\mu}\left[\mu^2\right]-2\,(n\bar{x}+\lambda_0\mu_0)\mathbb{E}_{q_\mu}[\mu]+\sum_{i=1}^{n}x_i^2+\lambda_0\mu_0^2\right]
\end{aligned}
\tag{87}
$$

因为前面已经知道 $q_\mu(\mu)$，可以计算这里的 $\mathbb{E}_{q_\mu}[\mu^2]$ 和 $\mathbb{E}_{q_\mu}[\mu]$。

同样地，也可以轻易地获得原分布的后验 $p(\tau\mid\mathcal{D},\mu)$

$$
\begin{aligned}
&\log p(\tau\mid\mathcal{D},\mu)=\log(p(\mathcal{D}\mid\mu,\tau))+\log p(\mu\mid\tau)+\log p(\tau)+\ \text{Const.}\\
&= \underbrace{\frac{n}{2}\log(\tau)-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2}_{\log(p(\mathcal{D}\mid\mu,\tau))}\underbrace{-\frac{\lambda_0\tau}{2}(\mu-\mu_0)^2}_{\log p(\mu\mid\gamma)}\underbrace{+(a_0-1)\log(\tau)-b_0\tau}_{\log p(\tau)}+\ \text{Const.}\\
&= (\underbrace{\frac{n}{2}+a_0}_{a_n}-1)\log(\tau)-\tau\underbrace{(b_0+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2+\lambda_0(\mu-\mu_0)^2)}_{b_n}+\ \text{Const.}\\
&\Longrightarrow p(\tau\mid\mathcal{D},\mu)=\text{Gamma}\,(a_n,b_n)\\
&\qquad\qquad\begin{cases} a_n &= \frac{n}{2}+a_0\\ b_n &= b_0+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2+\lambda_0(\mu-\mu_0)^2\end{cases}
\end{aligned}
\tag{88}
$$

2.4 指数族分布

2.4.1 概览

给定先验和似然都是指数族分布，则他们形成一个共轭对，则变分推断（平均场近似）有下列更新公式

$$\eta_j = \mathbb{E}_{q(\mathbf{z}\backslash z_j|\cdot)}[\eta_{\text{post}}(\mathbf{z}\backslash z_j)] \tag{89}$$

这里的 $\eta_{\text{post}}(\mathbf{z}\backslash z_j)$ 是和后验分布 $p(z_j \mid \cdot)$ 相关的自然参数。

和通用的更新公式相比

$$\log q_i^*(z_i) = \mathbb{E}_{i\neq j}[\log p(\mathbf{x}, \mathbf{z})] \tag{90}$$

使用指数族更新公式更加直接方便。

2.4.2 指数族

指数族分布通常用自然参数 η 表示为下列形式

$$\begin{aligned}
&h(x)\exp(T(x)^\top\eta - A(\eta)) \\
&= \underbrace{\exp(-A(\eta))}_{\text{normalization}} h(x)\exp(T(x)^\top\eta) \\
&\Longrightarrow \exp(-A(\eta))\int h(x)\exp(T(x)^\top\eta)\mathrm{d}x = 1 \\
&\Longrightarrow \int h(x)\exp(T(x)^\top\eta)\mathrm{d}x = \exp(A(\eta)) \\
&\Longrightarrow A(\eta) = \log\int h(x)\exp T(x)^\top\eta\mathrm{d}x
\end{aligned} \tag{91}$$

- η-自然参数（natural parameter）
- $T(x)$-充分统计量（sufficient statistic）
- $A(\eta)$-对数配分函数（log-partition function）：主要作用是归一化，为 η 的凸函数
- $h(x)$-基测度（base measure）：不依赖于参数 η 的部分，用于调整使得积分有意义

指数族分布具有易求最大似然估计的性质（可用下列高斯分布验证，其求解易于原参数形式）

$$\begin{aligned}
&\arg\max_\eta[\log p(X \mid \eta)] \\
&= \arg\max_\eta\left[\log\prod_{i=1}^n p(x_i \mid \eta)\right] \\
&= \arg\max_\eta\left[\log\left\{\prod_{i=1}^n h(x_i)\exp\left(\sum_{i=1}^n T(x_i)^\top\eta - nA(\eta)\right)\right\}\right] \\
&= \arg\max_\eta \underbrace{\left[\sum_{i=1}^n T(x_i)^\top\eta - nA(\eta)\right]}_{\mathcal{L}(\eta)} \\
&\Longrightarrow \frac{\partial\mathcal{L}(\eta)}{\partial\eta} = \sum_{i=1}^n T(x_i) - nA'(\eta) = 0 \\
&\Longrightarrow A'(\eta) = \sum_{i=1}^n \frac{T(x_i)}{n}
\end{aligned} \tag{92}$$

从另一个角度来看，指数分布族具有性质：对数规范化因子 $A(\eta)$ 对自然参数 η 的导数等于充分统计量 $T(x)$ 的数学期望，这是任何情况都成立的

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\eta}A(\eta) &= \frac{\mathrm{d}}{\mathrm{d}\eta}\log\int h(x)\exp\left\{\eta^{\mathrm{T}}T(x)\right\}\mathrm{d}x \\
&= \frac{\int T(x)\exp\left\{\eta^{\mathrm{T}}T(x)\right\}h(x)\mathrm{d}x}{\underbrace{\int h(x)\exp\left\{\eta^{\mathrm{T}}T(x)\right\}\mathrm{d}x}_{A(\eta)\text{ 见式(91)}}}(\text{交换积分与微分的顺序}) \\
&= \int T(x)\exp\left\{\eta^{\mathrm{T}}T(x) - A(\eta)\right\}h(x)\mathrm{d}x \\
&= \int T(x)p(x \mid \eta)\mathrm{d}x \\
&= \mathbb{E}[T(x)]
\end{aligned} \tag{93}$$

再对一阶导数求导，得到Hessian矩阵

$$
\begin{aligned}
\frac{\mathrm{d}A^2(\eta)}{\mathrm{d}^2\eta} &= \frac{\mathrm{d}\int T(x)p(x\mid\eta)\mathrm{d}x}{\mathrm{d}\eta} \\
&= \int T(x)\nabla_\eta p(x\mid\eta)\mathrm{d}x
\end{aligned} \tag{94}
$$

计算

$$
\begin{aligned}
\nabla_\eta p(x\mid\eta) &= \nabla_\eta\left(h(x)\exp\left(\eta^\top T(x)-A(\eta)\right)\right) \\
&= p(x\mid\eta)(T(x)-\nabla_\eta A(\eta)) \\
&= p(x\mid\eta)(T(x)-\mathbb{E}[T(x)]) \text{ 由式 } (93)
\end{aligned} \tag{95}
$$

代入式 (94)可得

$$
\begin{aligned}
\frac{\mathrm{d}A^2(\eta)}{\mathrm{d}^2\eta} &= \int T(x)p(x\mid\eta)(T(x)-\mathbb{E}[T(x)])^\top\mathrm{d}x \\
&= \int (T(x)T(x)^\top - T(x)\mathbb{E}[T(x)]^\top)p(x\mid\eta)\mathrm{d}x \\
&= \mathbb{E}[T(x)T(x)^\top]-\mathbb{E}[T(x)]\mathbb{E}[T(x)]^\top
\end{aligned} \tag{96}
$$

这就是协方差矩阵的定义公式，即 $A(\eta)$ 的二阶导数是充分统计量的协方差矩阵。进而可知，对任意非零向量 v，协方差矩阵 $\Sigma=\mathrm{Cov}(T(x))$ 满足

$$
\begin{aligned}
v^\top\Sigma v &= v^\top\mathbb{E}[(T(x)-\mathbb{E}[T(x)])(T(x)-\mathbb{E}[T(x)])^\top]v \\
&= \mathbb{E}[v^\top(T(x)-\mathbb{E}[T(x)])(T(x)-\mathbb{E}[T(x)])^\top v] \\
&= \mathbb{E}[(v^\top(T(x)-\mathbb{E}[T(x)]))^2]\geq 0
\end{aligned}
$$

由此可知 $\frac{\mathrm{d}A^2(\eta)}{\mathrm{d}^2\eta}$ 为半正定矩阵，则 $A(\eta)$ 为凸函数。当使用最大似然估计来求解指数族分布的参数时，需要最小化负对数似然函数（NLL）：

$$
\mathrm{NLL}(\eta) = -\eta^\top T(x) + A(\eta) - \log h(x) \tag{97}
$$

由于 $A(\eta)$ 是凸函数，且 $-\eta^T T(x)$ 是线性函数（也是凸函数），所以 NLL 也是凸函数。这意味着：指数族分布的最大似然估计问题，保证存在全局最优解，没有局部最优陷阱（如果优化的是常规参数 ordinary parameter 不一定能保证）。 这就是为什么逻辑回归、线性回归等模型训练如此稳定且高效的根本原因。

例如一维高斯分布，可以将其写为指数族分布的形式

$$
\begin{aligned}
\mathcal{N}\left(x;\mu,\sigma^2\right) &= \left(2\pi\sigma^2\right)^{-1/2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\
&= \exp\left(-\frac{x^2-2x\mu+\mu^2}{2\sigma^2}-\frac{1}{2}\log\left(2\pi\sigma^2\right)\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}x^2+\frac{\mu}{\sigma^2}x-\frac{\mu^2}{2\sigma^2}-\frac{1}{2}\log\left(2\pi\sigma^2\right)\right) \\
&= \exp\left(\begin{bmatrix}x & x^2\end{bmatrix}\begin{bmatrix}\frac{\mu}{\sigma^2} & -\frac{1}{2\sigma^2}\end{bmatrix}^\top-\frac{\mu^2}{2\sigma^2}-\frac{1}{2}\log\left(2\pi\sigma^2\right)\right)
\end{aligned} \tag{98}
$$

其中

$$
\begin{aligned}
T(x) &= \begin{bmatrix}x & x^2\end{bmatrix} \\
\boldsymbol{\eta} &= \begin{bmatrix}\eta_1 & \eta_2\end{bmatrix} \\
&= \begin{bmatrix}\frac{\mu}{\sigma^2} & -\frac{1}{2\sigma^2}\end{bmatrix} \\
\theta &= \begin{bmatrix}\mu \\ \sigma^2\end{bmatrix} = \begin{bmatrix}\frac{-\eta_1}{2\eta_2} \\ \frac{-1}{2\eta_2}\end{bmatrix}
\end{aligned} \tag{99}
$$

现在我们可以移除 μ 和 σ 得到一维高斯分布的指数族分布形式

$$
\begin{aligned}
\mathcal{N}_{\mathrm{nat}}(x,\boldsymbol{\eta}) &= \exp\left(\begin{bmatrix}x & x^2\end{bmatrix}\begin{bmatrix}\eta_1 & \eta_2\end{bmatrix}^\top-\frac{\mu^2}{2\sigma^2}-\frac{1}{2}\log\left(2\pi\sigma^2\right)\right) \\
&= \exp\left(\begin{bmatrix}x & x^2\end{bmatrix}\begin{bmatrix}\eta_1 & \eta_2\end{bmatrix}^\top-\frac{\left(\frac{-\eta_1}{2\eta_2}\right)^2}{2\left(\frac{-1}{2\eta_2}\right)}-\frac{1}{2}\log\left(2\pi\left(\frac{-1}{2\eta_2}\right)\right)\right) \\
&= \exp\left(T(x)^\top\boldsymbol{\eta}+\frac{\eta_1^2}{4\eta_2}-\frac{1}{2}\log\left(\frac{2\pi}{-2\eta_2}\right)\right) \\
&= \exp\left(T(x)^\top\boldsymbol{\eta}+\underbrace{\frac{\eta_1^2}{4\eta_2}+\frac{1}{2}\log\left(-2\eta_2\right)-\frac{1}{2}\log(2\pi)}_{A(\boldsymbol{\eta})}\right)
\end{aligned} \tag{100}
$$

2.4.3 共轭概率

共轭表示先验和后验是同种形式的概率分布，例如

$$\underbrace{p_{\eta_{\text{post}}}(\theta \mid \mathbf{x})}_{\text{same type}} \propto p(\mathbf{x} \mid \theta)\underbrace{p_{\eta_{\text{prior}}}(\theta)}_{\text{same type}} \tag{101}$$

使用指数族分布表示时，共轭的含义是先验和后验有相同的充分统计量 $T(\theta)$ 和 $h(\theta)$ （注意这里的 θ 是变量），不同的自然参数，即 $\eta_{\text{post}}, \eta_{\text{prior}}$ 以及不同的对数归一化因子。

证明：一个指数族分布的先验 $p(\theta|\alpha,\nu) = \frac{1}{Z(\alpha,\nu)}\exp\left(\alpha^T\theta - \nu A(\theta)\right)$ 必定有对应的似然使其拥有一个共轭的后验

$$\begin{aligned}
p(\theta \mid x) &\propto p(x \mid \theta)p(\theta \mid \alpha,\nu) \\
&= \prod_{i=1}^{N}\left[h(x_i)\exp\left(\theta^\top T(x_i) - A(\theta)\right)\right] \cdot \frac{1}{Z(\alpha,\nu)}\exp\left(\alpha^T\theta - \nu A(\theta)\right) \\
&\propto \exp\left(\theta^T\sum_{i=1}^{N}T(x_i) - N \cdot A(\theta)\right) \cdot \exp\left(\alpha^T\theta - \nu A(\theta)\right) \\
&= \exp\left(\underbrace{(\alpha + \sum_{i=1}^{N}T(x_i))}_{\alpha_{\text{new}}}{}^T\theta - \underbrace{(\nu + N)}_{\nu_{\text{new}}}A(\theta)\right)
\end{aligned} \tag{102}$$

观察最后的结果，会发现后验分布的形式与先验分布完全一致，只是参数发生了变化，$\alpha' = \alpha + \sum_{i=1}^{N}T(x_i)$，$\nu' = \nu + N$。

2.4.4 变分推断

隐变量 β 的后验分布，注意下式中的 $h(\beta)$ 和 $T(\beta)$ 是相同的，因为设置变分分布是和真实条件后验分布同一种分布，这是同分布的结果

$$\begin{aligned}
&p(\beta \mid z, x) \\
&= h(\beta)\exp\left(T(\beta)^\top\eta(z,x) - A_g(\eta(z,x))\right) \\
&\approx q(\beta \mid \lambda) = h(\beta)\exp\left(T(\beta)^\top\lambda - A_g(\lambda)\right)
\end{aligned} \tag{103}$$

隐变量 z 的后验分布

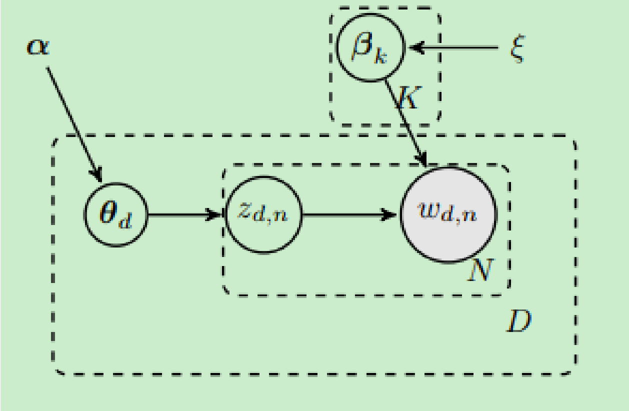
$$\begin{aligned}
&p(z \mid \beta, x) \\
&= h(z)\exp\left(T(z)^\top\eta(\beta,x) - A_l(\eta(\beta,x))\right) \\
&\approx q(z \mid \phi) = h(z)\exp\left(T(z)^\top\phi - A_l(\phi)\right)
\end{aligned} \tag{104}$$

固定 ϕ ，优化 λ ，每一步去除无关项（损失函数见式（65））

$$\begin{aligned}
&\mathcal{L}(\lambda,\phi) \\
&= E_{q(z,\beta)}[\log p(x,z,\beta)] - E_{q(z,\beta)}[\log q(z,\beta)] \\
&= E_{q(z,\beta)}[\log p(\beta \mid x,z) + \log p(z,x)] - E_{q(z,\beta)}[\log q(\beta)] - E_{q(z,\beta)}[\log q(z)] \\
&= E_{q(z,\beta)}[\log p(\beta \mid x,z)] - E_{q(z,\beta)}[\log q(\beta)] \\
&= E_{q(z,\beta)}[\log h(\beta)] + E_{q(z,\beta)}[T(\beta)^\top\eta(z,x)] - E_{q(z,\beta)}[A_g(\eta(x,z))] \\
&\quad - E_{q(z,\beta)}[\log h(\beta)] - E_{q(z,\beta)}[T(\beta)^\top\lambda] + E_{q(z,\beta)}[A_g(\lambda)] \\
&= E_{q(\beta)}[T(\beta)^\top]E_{q(z)}[\eta(z,x)] - E_{q(z)}[A_g(\eta(x,z))] - E_{q(\beta)}[T(\beta)^\top\lambda] + A_g(\lambda) \quad \text{using } \frac{\partial A(\eta)}{\partial\eta} = E_{p(x|\eta)}[T(x)] \\
&= A_g'(\lambda)^\top E_{q(z)}[\eta(z,x)] - \lambda A_g'(\lambda)^\top + A_g(\lambda) \quad \text{taking partial derivative with respect to } \lambda \\
&\Rightarrow A_g''(\lambda)^\top E_{q(z)}[\eta(z,x)] - A_g'(\lambda)^\top - \lambda A_g''(\lambda)^\top + A_g'(\lambda) = 0 \\
&\Rightarrow A_g''(\lambda)^\top(E_{q(z)}[\eta(x,z)] - \lambda) = 0 \\
&\Longrightarrow \lambda = E_{q(z)}[\eta(x,z)] \qquad \text{where } q(z) = q(z \mid \phi)
\end{aligned} \tag{105}$$

同理，固定 λ ，优化 ϕ，可得到更新公式 $\phi = E_{q(\beta|\lambda)}[\eta(\beta,x)]$

2.5 基于变分推断的LDA参数学习



For each topic k :

$$\boldsymbol{\beta}_k \sim \operatorname{Dir}(\xi, \ldots, \xi) \quad \text{for } k \in \{1, \ldots, K\} \tag{106}$$

For each document d :

$$\boldsymbol{\theta}_{\boldsymbol{d}} \sim \operatorname{Dir}(\alpha, \ldots, \alpha) \tag{107}$$

For each word $w \in \{1, \ldots, N\}$:

$$\begin{aligned} & z_{d,n} \sim \operatorname{Mult}\left(\boldsymbol{\theta}_d\right) \\ & w_{d,n} \sim \operatorname{Mult}\left(\beta_{z_{d,n}}\right) \end{aligned} \tag{108}$$

因为先验和似然是共轭的，而选取变分分布时应该选取和条件后验相同的分布有利于计算，所以此时选取和先验同样的分布（可以选择不同分布，但是会带来计算复杂性）

$$q(\boldsymbol{\beta}_k) = \operatorname{Dir}(\boldsymbol{\lambda}_k), \quad q(\boldsymbol{\theta}_d) = \operatorname{Dir}(\boldsymbol{\gamma}_d), \quad q(z_{d,n}) = \operatorname{Mult}(\boldsymbol{\phi}_{d,n}) \tag{109}$$

$z_{d,n}$ 的变分分布选取基于两个理由，一是离散的、有限类别的变量，Multinomial（或 Categorical）分布是唯一合理的描述方式；二是由式（75）可知

$$\begin{aligned} & \ln q^*(z_{d,n}) = \mathbb{E}_{q(\theta,\beta)}[\ln p(z_{d,n}|\theta_d) + \ln p(w_{d,n}|z_{d,n}, \beta)] + C \\ & \Rightarrow \ln q^*(z_{d,n} = k) = \mathbb{E}[\ln \theta_{d,k}] + \mathbb{E}[\ln \beta_{k,w_{d,n}}] + C \\ & \Rightarrow q^*(z_{d,n} = k) \propto \exp\left(\mathbb{E}[\ln \theta_{d,k}] + \mathbb{E}[\ln \beta_{k,w_{d,n}}]\right) \end{aligned} \tag{110}$$

这就完全符合 Multinomial 分布 的定义。

2.5.1 基于指数族分布的变分推断

2.5.1.1 更新 $\Phi_{D,N}$

首先找到后验 $p(z_{d,n} = k \mid \boldsymbol{\theta}_d, \boldsymbol{\varphi}_k, w_{d,n})$ 的自然参数

$$\begin{aligned} p(z_{d,n} = k \mid \boldsymbol{\theta}_d, \boldsymbol{\beta}_{1:K}, w_{d,n}) & \propto p(z_{d,n} = k \mid \boldsymbol{\theta}_d) \cdot p(w_{d,n} \mid z_{d,n} = k, \boldsymbol{\beta}_{1:K}) \\ & = \theta_{d,k} \cdot \beta_{k,w_{d,n}} \\ & = \exp\left(\underbrace{(\log \theta_{d,k} + \log \beta_{k,w_{d,n}})}_{\eta_l(\boldsymbol{\theta}_{d,k}, \boldsymbol{\beta}_{1:K}, w_{d,n})} \cdot \underbrace{1}_{T(z_{d,n})} \right) \end{aligned} \tag{111}$$

使用正常的多项式分布可以表达为

$$p(z_{d,n} \mid \boldsymbol{\theta}_d, \boldsymbol{\beta}_{1:K}, w_{d,n}) = \operatorname{Mult}(\theta_{d,1} \cdot \beta_{1,w_{d,n}}, \ldots, \theta_{d,k} \cdot \beta_{k,w_{d,n}}) \tag{112}$$

利用更新公式可知

$$\begin{aligned} \eta(\phi_{d,n}^k) & = \log(\phi_{d,n}^k) \quad \text{多项式分布的自然参数形式} \\ & \propto E_{q(\boldsymbol{\theta}_d, \boldsymbol{\beta}_k)}[\eta_l(\boldsymbol{\theta}_d, \boldsymbol{\beta}_{1:K}, w_{d,n})] \quad \text{由式（75）的梯度上升最优更新公式} \\ & = E_{q(\boldsymbol{\theta}_d)}[\log(\theta_{d,k})] + E_{q(\boldsymbol{\beta}_k)}[\log(\beta_{k,w_{d,n}})] \\ & = \Psi(\gamma_{d,k}) - \Psi\left(\sum_{k=1}^{K} \gamma_{d,k}\right) + \Psi(\lambda_{k,w_{d,n}}) - \Psi\left(\sum_{v=1}^{V} \lambda_{k,v}\right) \\ & \Longrightarrow \phi_{d,n}^k \propto \exp\left[\Psi(\gamma_{d,k}) - \Psi\left(\sum_{k=1}^{K} \gamma_{d,k}\right) + \Psi(\lambda_{k,w_{d,n}}) - \Psi\left(\sum_{v=1}^{V} \lambda_{k,v}\right)\right] \\ & \qquad \propto \exp\left[\Psi(\gamma_{d,k}) + \Psi(\lambda_{k,w_{d,n}}) - \Psi\left(\sum_{v=1}^{V} \lambda_{k,v}\right)\right] \end{aligned} \tag{113}$$

2.5.1.2 更新 Γ_D

同样先推导后验 $p(\boldsymbol{\theta}_d \mid \mathbf{z}_d)$ 的表达式

$$
\begin{aligned}
p(\boldsymbol{\theta}_d \mid \mathbf{z}_d) &= p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}) \prod_{n=1}^{N} p(z_{d,n} \mid \boldsymbol{\theta}_d) \\
&= \prod_{k=1}^{K} \left(\theta_{d,k}^{\alpha_k - 1} \prod_{n=1}^{N} \theta_{d,k}^{\mathbb{1}(z_{d,n}=k)} \right) \\
&= \exp \left[\log \left(\prod_{k=1}^{K} \left(\theta_{d,k}^{\alpha_k - 1} \prod_{n=1}^{N} \theta_{d,k}^{\mathbb{1}(z_{d,n}=k)} \right) \right) \right] \\
&= \exp \left[\sum_{k=1}^{K} \log \left(\theta_{d,k}^{\alpha_k - 1} \prod_{n=1}^{N} \theta_{d,k}^{\mathbb{1}(z_{d,n}=k)} \right) \right] \\
&= \exp \left[\sum_{k=1}^{K} \left(\log \theta_{d,k}^{\alpha_k - 1} + \sum_{n=1}^{N} \log \left(\theta_{d,k}^{\mathbb{1}(z_{d,n}=k)} \right) \right) \right] \\
&= \exp \left[\sum_{k=1}^{K} \left((\alpha_k - 1) \log \theta_{d,k} + \sum_{n=1}^{N} \mathbb{1}(z_{d,n}=k) \log \theta_{d,k} \right) \right] \\
&= \exp \left[\sum_{k=1}^{K} \left(\alpha_k - 1 + \sum_{n=1}^{N} \mathbb{1}(z_{d,n}=k) \right) \log \theta_{d,k} \right] \\
&= \exp \left(\underbrace{\begin{bmatrix} \alpha_1 - 1 + n_1 \\ \cdots \\ \alpha_K - 1 + n_K \end{bmatrix}^{\top}}_{\eta_l(\boldsymbol{\alpha}, \mathbf{z}_d)} \underbrace{\begin{bmatrix} \log \theta_{d,1} \\ \cdots \\ \log \theta_{d,K} \end{bmatrix}}_{T(\boldsymbol{\theta}_d)} \right) \quad \text{by letting } n_k = \sum_{n=1}^{N} \mathbb{1}(z_{d,n}=k) \\
&= \mathrm{Dir}(\alpha_1 + n_1, \ldots, \alpha_K + n_K)
\end{aligned} \tag{114}
$$

接下来用变分分布 $q(\eta(\boldsymbol{\gamma}_d)) = \mathrm{Dir}(\eta(\boldsymbol{\gamma}_d))$ 来近似 $p(\boldsymbol{\theta}_d \mid \mathbf{z}_d)$ ，利用更新公式

$$
\begin{aligned}
\eta(\boldsymbol{\gamma}_d) &= E_{q(\mathbf{z}_d \mid \boldsymbol{\phi}_d)}[\eta_l(\boldsymbol{\alpha}, \mathbf{z}_d)] \quad \text{狄利克雷分布的自然参数形式} \\
&= E_{q(\mathbf{z}_d \mid \boldsymbol{\phi}_d)}[(\alpha_1 - 1 + n_1) \ldots (\alpha_K - 1 + n_K)]
\end{aligned} \tag{115}
$$

计算这个期望

$$
\begin{aligned}
E_{q(\mathbf{z}_d \mid \boldsymbol{\phi}_d)} \left[\sum_{n=1}^{N} \mathbb{1}(z_{d,n}=k) \right] &= \sum_{n=1}^{N} E_{q(\mathbf{z}_d \mid \boldsymbol{\phi}_d)}[\mathbb{1}(z_{d,n}=k)] \\
&= \sum_{n=1}^{N} q(z_{d,n}=k) \\
&= \sum_{n=1}^{N} \phi_{d,n}^{k}
\end{aligned} \tag{116}
$$

因此有

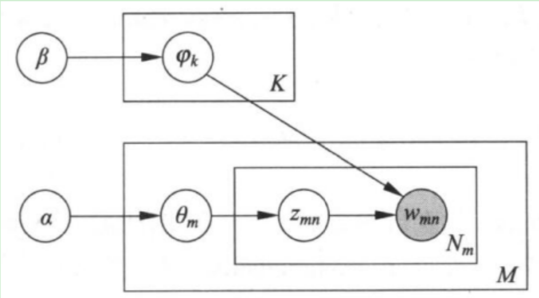
$$
\begin{aligned}
\eta(\boldsymbol{\gamma}_d) &= \left[\left(\alpha_1 - 1 + \sum_{n=1}^{N} \phi_{d,n}^{1} \right) \ldots \left(\alpha_K - 1 + \sum_{n=1}^{N} \phi_{d,n}^{K} \right) \right] \\
\Longrightarrow \boldsymbol{\eta} &= \left[\left(\alpha_1 + \sum_{n=1}^{N} \phi_{d,n}^{1} \right) \ldots \left(\alpha_K + \sum_{n=1}^{N} \phi_{d,n}^{K} \right) \right] \quad \text{迪利克雷分布的自然参数} \eta_i = \alpha_i - 1 \\
&= \boldsymbol{\alpha} + \sum_{n=1}^{N} \boldsymbol{\phi}_{d,n}
\end{aligned} \tag{117}
$$

2.5.1.3 更新 $\boldsymbol{\lambda}_K$

与 $\boldsymbol{\gamma}_d$ 更新公式类似，有

$$
\boldsymbol{\lambda}_k = \xi + \sum_{d=1}^{D} \sum_{n=1}^{N} w_{d,n} \cdot \phi_{d,n}^{k} \tag{118}
$$

2.5.2 基于展开式的推断



当超参数给定时，$\log p(X \mid \theta)$ 是常数，因此

$$\begin{aligned}
q(Z)^* &= \arg\min_{q(Z)} \mathrm{KL}(q(Z) \| p(Z \mid X, \theta)) \\
&= \arg\max_{q(Z)} \mathrm{ELBO} \\
&= \arg\max_{q(Z)} \int_Z q(Z) \log \frac{p(X, Z \mid \theta)}{q(Z)} \, \mathrm{d}Z \\
&= \arg\max_{q(Z)} \int_Z q(Z) \log p(X, Z \mid \theta) \, \mathrm{d}Z - \int_Z q(Z) \log q(Z) \, \mathrm{d}Z \\
&= \arg\max_{q(Z)} E_{q(Z)}[\log p(X, Z \mid \theta)] - E_{q(Z)}[\log q(Z)]
\end{aligned} \tag{119}$$

KL 散度的最小化或证据下界的最大化实际是在平均场的集合，即满足独立假设的分布集合 $Q = \{q(Z) \mid q(Z) = \prod_{j=1}^d q(Z_j)\}$ 之中进行的

$$\begin{aligned}
&\log p\left(\mathbf{w}, \mathbf{z}, \varphi_{1:K}, \theta_{1:M} \mid \alpha, \beta\right) \\
&= \log \left\{ \left[\prod_{m=1}^M p(\theta_m \mid \alpha) \right] \left[\prod_{k=1}^K p(\varphi_k \mid \beta) \right] \left[\prod_{m=1}^M \prod_{n=1}^{N_m} p(z_{mn} \mid \theta_m) \right] \left[\prod_{m=1}^M \prod_{n=1}^{N_m} p(w_{mn} \mid \varphi_{1:K}, z_{mn}) \right] \right\} \\
&= \sum_{m=1}^M \log p(\theta_m \mid \alpha) + \sum_{k=1}^K \log p(\varphi_k \mid \beta) + \sum_{m=1}^M \sum_{n=1}^{N_m} \log p(z_{mn} \mid \theta_m) + \sum_{m=1}^M \sum_{n=1}^{N_m} \log p(w_{mn} \mid \varphi_{1:K}, z_{mn})
\end{aligned} \tag{120}$$

定义基于平均场的变分分布

$$\begin{aligned}
q(\mathbf{z}, \varphi_{1:K}, \theta_{1:M} \mid \mu_{1:K}, \gamma_{1:M}, \{\eta_{mn}\}) &= \prod_{k=1}^K q(\varphi_k \mid \mu_k) \prod_{m=1}^M q(\theta_m \mid \gamma_m) \prod_{m=1}^M \prod_{n=1}^{N_m} q(z_{mn} \mid \eta_{mn}) \\
&= \prod_{k=1}^K \mathrm{Dir}(\varphi_k \mid \mu_k) \prod_{m=1}^M \mathrm{Dir}(\theta_m \mid \gamma_m) \prod_{m=1}^M \prod_{n=1}^{N_m} \mathrm{Mult}(z_{mn} \mid \eta_{mn})
\end{aligned} \tag{121}$$

展开证据下界

$$\begin{aligned}
\mathrm{ELBO} = & E_{q(\mathbf{z}, \varphi_{1:K}, \theta_{1:M} \mid \mu_{1:K}, \gamma_{1:M}, \{\eta_{mn}\})} \left[\log p(\mathbf{w}, \mathbf{z}, \varphi_{1:K}, \theta_{1:M} \mid \alpha, \beta)\right] \\
& - E_{q(\mathbf{z}, \varphi_{1:K}, \theta_{1:M} \mid \mu_{1:K}, \gamma_{1:M}, \{\eta_{mn}\})} \left[\log q(\mathbf{z}, \varphi_{1:K}, \theta_{1:M} \mid \mu_{1:K}, \gamma_{1:M}, \{\eta_{mn}\})\right] \\
= & \sum_{m=1}^M E_{q(\theta_m \mid \gamma_m)} \left[\log p(\theta_m \mid \alpha)\right] + \sum_{k=1}^K E_{q(\varphi_k \mid \mu_k)} \left[\log p(\varphi_k \mid \beta)\right] + \sum_{m=1}^M \sum_{n=1}^{N_m} E_{q(z_{mn}, \theta_m \mid \eta_{mn}, \gamma_m)} \left[\log p(z_{mn} \mid \theta_m)\right] \\
& + \sum_{m=1}^M \sum_{n=1}^{N_m} E_{q(\varphi_{1:K}, z_{mn} \mid \mu_{1:K}, \eta_{mn})} \left[\log p(w_{mn} \mid \varphi_{1:K}, z_{mn})\right] - \sum_{k=1}^K E_{q(\varphi_k \mid \mu_k)} \left[\log q(\varphi_k \mid \mu_k)\right] \\
& - \sum_{m=1}^M E_{q(\theta_m \mid \gamma_m)} \left[\log q(\theta_m \mid \gamma_m)\right] - \sum_{m=1}^M \sum_{n=1}^{N_m} E_{q(z_{mn} \mid \eta_{mn})} \left[\log q(z_{mn} \mid \eta_{mn})\right]
\end{aligned} \tag{122}$$

第一项：

$$\begin{aligned}
& \sum_{m=1}^M E_{q(\theta_m \mid \gamma_m)} \left[\log p(\theta_m \mid \alpha)\right] \\
&= \sum_{m=1}^M E_{q(\theta_m \mid \gamma_m)} \left[\log \left(\frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_{mk}^{\alpha_k - 1} \right) \right] \\
&= \sum_{m=1}^M \mathbb{E}_{q(\theta_m \mid \gamma_m)} \left[\log \Gamma\left(\sum_{k=1}^K \alpha_k\right) - \sum_{k=1}^K \log \Gamma(\alpha_k) + \sum_{k=1}^K (\alpha_k - 1) \log \theta_{mk} \right] \\
&= \sum_{m=1}^M \log \Gamma\left(\sum_{k=1}^K \alpha_k\right) - \sum_{m=1}^M \sum_{k=1}^K \log \Gamma(\alpha_k) + \sum_{m=1}^M \sum_{k=1}^K (\alpha_k - 1) E_{q(\theta_m \mid \gamma_m)} \left[\log \theta_{mk}\right] \\
&= \sum_{m=1}^M \log \Gamma\left(\sum_{k=1}^K \alpha_k\right) - \sum_{m=1}^M \sum_{k=1}^K \log \Gamma(\alpha_k) + \sum_{m=1}^M \sum_{k=1}^K (\alpha_k - 1) \left[\psi(\gamma_{mk}) - \psi\left(\sum_{l=1}^K \gamma_{ml}\right) \right]
\end{aligned} \tag{123}$$

此处用到迪利克雷分布作为指数族分布的性质：对数规范化因子对自然参数的导数等于充分统计量的数学期望，ψ 是 digamma 函数，即对数伽马函数的一阶导数。

第二项：

$$
\begin{aligned}
&\sum_{k=1}^{K} E_{q(\varphi_k|\mu_k)}\left[\log p\left(\varphi_k \mid \boldsymbol{\beta}\right)\right] \\
&= \sum_{k=1}^{K} \log \Gamma\left(\sum_{v=1}^{V} \beta_v\right) - \sum_{k=1}^{K}\sum_{v=1}^{V} \log \Gamma\left(\beta_v\right) + \sum_{k=1}^{K}\sum_{v=1}^{V}\left(\beta_v - 1\right)\left[\psi\left(\mu_{kv}\right) - \psi\left(\sum_{s=1}^{V}\mu_{ks}\right)\right]
\end{aligned}
\tag{124}
$$

第三项:

$$
\begin{aligned}
&\sum_{m=1}^{M}\sum_{n=1}^{N_m} E_{q(z_{mn},\theta_m|\eta_{mn},\gamma_m)}[\log p(z_{mn} \mid \theta_m)] \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N_m} E_{q(z_{mn},\theta_m|\eta_{mn},\gamma_m)}\left[\log \prod_{k=1}^{K}(\theta_{mk})^{\mathbb{I}(z_{mn}=k)}\right] \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N_m} E_{q(z_{mn},\theta_m|\eta_{mn},\gamma_m)}\left[\sum_{k=1}^{K}\mathbb{I}(z_{mn}=k)\log\theta_{mk}\right] \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{K} E_{q(z_{mn},\theta_m|\eta_{mn},\gamma_m)}[\mathbb{I}(z_{mn}=k)\log\theta_{mk}] \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{K} E_{q(z_{mn}|\eta_{mn})}[\mathbb{I}(z_{mn}=k)]E_{q(\theta_m|\gamma_m)}[\log\theta_{mk}] \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{K} \eta_{mnk}\left[\psi(\gamma_{mk}) - \psi\left(\sum_{l=1}^{K}\gamma_{ml}\right)\right]
\end{aligned}
\tag{125}
$$

第四项:

$$
\begin{aligned}
&\sum_{m=1}^{M}\sum_{n=1}^{N_m} E_{q(\varphi_{1:K},z_{mn}|\mu_{1:K},\eta_{mn})}\left[\log p\left(w_{mn} \mid \varphi_{1:K}, z_{mn}\right)\right] \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N_m} E_{q(\varphi_{1:K},z_{mn}|\mu_{1:K},\eta_{mn})}\left[\log \prod_{k=1}^{K}\varphi_{k,i(w_{mn})}^{\mathbb{I}(z_{mn}=k)}\right] \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N_m} E_{q(\varphi_{1:K},z_{mn}|\mu_{1:K},\eta_{mn})}\left[\sum_{k=1}^{K}\mathbb{I}\left(z_{mn}=k\right)\log\varphi_{k,i(w_{mn})}\right] \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{K} E_{q(\varphi_k,z_{mn}|\mu_k,\eta_{mn})}\left[\mathbb{I}\left(z_{mn}=k\right)\log\varphi_{k,i(w_{mn})}\right] \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{K} E_{q(z_{mn}|\eta_{mn})}\left[\mathbb{I}\left(z_{mn}=k\right)\right]E_{q(\varphi_k|\mu_k)}\left[\log\varphi_{k,i(w_{mn})}\right] \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{K} \eta_{mnk}\left[\psi\left(\mu_{k,i(w_{mn})}\right) - \psi\left(\sum_{s=1}^{V}\mu_{ks}\right)\right]
\end{aligned}
\tag{126}
$$

式中 $i(w_{mn}) \in \{1,\ldots,V\}$ 表示单词 w_{mn} 的索引。
第五项:

$$
\begin{aligned}
&-\sum_{k=1}^{K} E_{q(\varphi_k|\mu_k)}\left[\log q\left(\varphi_k \mid \mu_k\right)\right] \\
&= -\sum_{k=1}^{K} E_{q(\varphi_k|\mu_k)}\left[\log\left(\frac{\Gamma\left(\sum_{v=1}^{V}\mu_{kv}\right)}{\prod_{v=1}^{V}\Gamma\left(\mu_{kv}\right)}\prod_{v=1}^{V}\varphi_{kv}^{\mu_{kv}-1}\right)\right] \\
&= -\sum_{k=1}^{K} E_{q(\varphi_k|\mu_k)}\left[\log\Gamma\left(\sum_{v=1}^{V}\mu_{kv}\right) - \sum_{v=1}^{V}\log\Gamma\left(\mu_{kv}\right) + \sum_{v=1}^{V}\left(\mu_{kv}-1\right)\log\varphi_{kv}\right] \\
&= -\sum_{k=1}^{K}\log\Gamma\left(\sum_{v=1}^{V}\mu_{kv}\right) + \sum_{k=1}^{K}\sum_{v=1}^{V}\log\Gamma\left(\mu_{kv}\right) - \sum_{k=1}^{K}\sum_{v=1}^{V}\left(\mu_{kv}-1\right)E_{q(\varphi_k|\mu_k)}\left[\log\varphi_{kv}\right] \\
&= -\sum_{k=1}^{K}\log\Gamma\left(\sum_{v=1}^{V}\mu_{kv}\right) + \sum_{k=1}^{K}\sum_{v=1}^{V}\log\Gamma\left(\mu_{kv}\right) - \sum_{k=1}^{K}\sum_{v=1}^{V}\left(\mu_{kv}-1\right)\left[\psi\left(\mu_{kv}\right) - \psi\left(\sum_{s=1}^{V}\mu_{ks}\right)\right]
\end{aligned}
\tag{127}
$$

第六项:

$$
\begin{aligned}
&-\sum_{m=1}^{M} E_{q(\theta_m|\gamma_m)}\left[\log q\left(\theta_m \mid \gamma_m\right)\right] \\
&= -\sum_{m=1}^{M}\log\Gamma\left(\sum_{k=1}^{K}\gamma_{mk}\right) + \sum_{m=1}^{M}\sum_{k=1}^{K}\log\Gamma\left(\gamma_{mk}\right) - \sum_{m=1}^{M}\sum_{k=1}^{K}\left(\gamma_{mk}-1\right)\left[\psi\left(\gamma_{mk}\right) - \psi\left(\sum_{l=1}^{K}\gamma_{ml}\right)\right]
\end{aligned}
\tag{128}
$$

第七项:

$$
\begin{aligned}
& -\sum_{m=1}^{M}\sum_{n=1}^{N_m} E_{q(z_{mn}|\eta_{mn})}[\log q(z_{mn}\mid \eta_{mn})] \\
&= -\sum_{m=1}^{M}\sum_{n=1}^{N_m} E_{q(z_{mn}|\eta_{mn})}\left[\log \prod_{k=1}^{K} \eta_{mnk}^{\mathbb{I}(z_{mn}=k)}\right] \\
&= -\sum_{m=1}^{M}\sum_{n=1}^{N_m} E_{q(z_{mn}|\eta_{mn})}\left[\sum_{k=1}^{K} \mathbb{I}(z_{mn}=k)\log \eta_{mnk}\right] \\
&= -\sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{K} E_{q(z_{mn}|\eta_{mn})}[\mathbb{I}(z_{mn}=k)]\cdot \log \eta_{mnk} \\
&= -\sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{K} \eta_{mnk}\log \eta_{mnk}
\end{aligned} \tag{129}
$$

上述七项合并得到

$$
\begin{aligned}
& \operatorname{ELBO}(\mu_{1:K},\gamma_{1:M},\{\eta_{mn}\},\alpha,\beta) \\
& = \mathcal{L}(\mu_{1:K},\gamma_{1:M},\{\eta_{mn}\},\alpha,\beta) \\
& = \sum_{m=1}^{M}\log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) - \sum_{m=1}^{M}\sum_{k=1}^{K}\log\Gamma\left(\alpha_k\right) + \sum_{m=1}^{M}\sum_{k=1}^{K}\left(\alpha_k-1\right)\left[\psi\left(\gamma_{mk}\right)-\psi\left(\sum_{l=1}^{K}\gamma_{ml}\right)\right] \\
& + \sum_{k=1}^{K}\log\Gamma\left(\sum_{v=1}^{V}\beta_v\right) - \sum_{k=1}^{K}\sum_{v=1}^{V}\log\Gamma\left(\beta_v\right) + \sum_{k=1}^{K}\sum_{v=1}^{V}\left(\beta_v-1\right)\left[\psi\left(\mu_{kv}\right)-\psi\left(\sum_{s=1}^{V}\mu_{ks}\right)\right] \\
& + \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{K}\eta_{mnk}\left[\psi\left(\gamma_{mk}\right)-\psi\left(\sum_{l=1}^{K}\gamma_{ml}\right)\right] \\
& + \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{K}\eta_{mnk}\left[\psi\left(\mu_{k,i(w_{mn})}\right)-\psi\left(\sum_{s=1}^{V}\mu_{ks}\right)\right] \\
& - \sum_{k=1}^{K}\log\Gamma\left(\sum_{v=1}^{V}\mu_{kv}\right) + \sum_{k=1}^{K}\sum_{v=1}^{V}\log\Gamma\left(\mu_{kv}\right) - \sum_{k=1}^{K}\sum_{v=1}^{V}\left(\mu_{kv}-1\right)\left[\psi\left(\mu_{kv}\right)-\psi\left(\sum_{s=1}^{V}\mu_{ks}\right)\right] \\
& - \sum_{m=1}^{M}\log\Gamma\left(\sum_{k=1}^{K}\gamma_{mk}\right) + \sum_{m=1}^{M}\sum_{k=1}^{K}\log\Gamma\left(\gamma_{mk}\right) - \sum_{m=1}^{M}\sum_{k=1}^{K}\left(\gamma_{mk}-1\right)\left[\psi\left(\gamma_{mk}\right)-\psi\left(\sum_{l=1}^{K}\gamma_{ml}\right)\right] \\
& - \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{K}\eta_{mnk}\log\eta_{mnk}
\end{aligned} \tag{130}
$$

目标函数 $\mathcal{L}(\mu_{1:K},\gamma_{1:M},\{\eta_{mn}\},\alpha,\beta)$ 中关于 μ_k 的部分：

$$
\begin{aligned}
\mathcal{L}_{[\mu_k]} &= \sum_{v=1}^{V}(\beta_v-1)\left[\psi(\mu_{kv})-\psi\left(\sum_{s=1}^{V}\mu_{ks}\right)\right] + \sum_{m=1}^{M}\sum_{n=1}^{N_m}\eta_{mnk}\left[\psi(\mu_{k,i(w_{mn})})-\psi\left(\sum_{s=1}^{V}\mu_{ks}\right)\right] \\
&\quad -\log\Gamma\left(\sum_{v=1}^{V}\mu_{kv}\right)+\sum_{v=1}^{V}\log\Gamma(\mu_{kv})-\sum_{v=1}^{V}(\mu_{kv}-1)\left[\psi(\mu_{kv})-\psi\left(\sum_{s=1}^{V}\mu_{ks}\right)\right] \\
&= \sum_{v=1}^{V}\left[\psi(\mu_{kv})-\psi\left(\sum_{s=1}^{V}\mu_{ks}\right)\right]\left(\beta_v+\sum_{m=1}^{M}\sum_{n=1}^{N_m}\eta_{mnk}\mathbb{I}(i(w_{mn})=v)-\mu_{kv}\right) \\
&\quad -\log\Gamma\left(\sum_{v=1}^{V}\mu_{kv}\right)+\sum_{v=1}^{V}\log\Gamma(\mu_{kv})
\end{aligned} \tag{131}
$$

分别关于 μ_{kv}，$v=1,\dots,V$ 求偏导，得到

$$
\left[\sum_{m=1}^{M}\sum_{n=1}^{N_m}\mathbb{I}(i(w_{mn}=v))\cdot\eta_{mnk}+\beta_v-\mu_{kv}\right]\cdot\psi'(\mu_{kv})+\left[\sum_{m=1}^{M}\sum_{n=1}^{N_m}\eta_{mnk}+\sum_{s=1}^{V}(\mu_{ks}-\beta_s)\right]\cdot\psi'(\sum_{s=1}^{V}\mu_{ks}) \tag{132}
$$

因为 KL 散度为 $q(z)$ 的凸函数，而对单个变分参数的优化等价于优化一个 KL 散度加一个常数（见式（73）），所以令偏导数为零，得到 μ_{kv} 的更新公式

$$
\mu_{kv}=\beta_v+\sum_{m=1}^{M}\sum_{n=1}^{N_m}\eta_{mnk}\mathbb{I}(i(w_{mn}=v)) \tag{133}
$$

目标函数 $\mathcal{L}(\mu_{1:K},\gamma_{1:M},\{\eta_{mn}\},\alpha,\beta)$ 中关于 γ_m 的部分：

$$
\begin{aligned}
\mathcal{L}_{[\gamma_m]} &= \sum_{k=1}^{K}(\alpha_k-1)\left[\psi(\gamma_{mk})-\psi\left(\sum_{l=1}^{K}\gamma_{ml}\right)\right]+\sum_{n=1}^{N_m}\sum_{k=1}^{K}\eta_{mnk}\left[\psi(\gamma_{mk})-\psi\left(\sum_{l=1}^{K}\gamma_{ml}\right)\right] \\
&\quad -\log\Gamma\left(\sum_{k=1}^{K}\gamma_{mk}\right)+\sum_{k=1}^{K}\log\Gamma(\gamma_{mk})-\sum_{k=1}^{K}(\gamma_{mk}-1)\left[\psi(\gamma_{mk})-\psi\left(\sum_{l=1}^{K}\gamma_{ml}\right)\right] \\
&= \sum_{k=1}^{K}\left[\psi(\gamma_{mk})-\psi\left(\sum_{l=1}^{K}\gamma_{ml}\right)\right]\left(\alpha_k+\sum_{n=1}^{N_m}\eta_{mnk}-\gamma_{mk}\right) \\
&\quad -\log\Gamma\left(\sum_{k=1}^{K}\gamma_{mk}\right)+\sum_{k=1}^{K}\log\Gamma(\gamma_{mk})
\end{aligned} \tag{134}
$$

分别关于 γ_{mk}，$k=1,\ldots,K$ 求偏导，得到

$$
\left[\sum_{n=1}^{N_m}\eta_{mnk}+\alpha_k-\gamma_{mk}\right]\cdot\psi'(\gamma_{mk})+\left[-\sum_{n=1}^{N_m}\sum_{l=1}^{K}\eta_{mnl}-\sum_{l=1}^{K}(\alpha_l-1)+\sum_{l=1}^{K}(\gamma_{ml}-1)\right]\cdot\psi'(\sum_{l=1}^{K}\gamma_{ml}) \tag{135}
$$

令偏导数为零，得到 γ_{mk} 的更新公式

$$
\gamma_{mk}=\alpha_k+\sum_{n=1}^{N_m}\eta_{mnk} \tag{136}
$$

目标函数中关于 η_{mn} 的部分：

$$
\begin{aligned}
\mathcal{L}_{\{\eta_{mn}\}} &= \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{K}\eta_{mnk}\left[\psi(\gamma_{mk})-\psi\left(\sum_{l=1}^{K}\gamma_{ml}\right)\right]+\sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{K}\eta_{mnk}\left[\psi(\mu_{k,i(w_{mn})})-\psi\left(\sum_{s=1}^{V}\mu_{ks}\right)\right] \\
&\quad -\sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{K}\eta_{mnk}\log\eta_{mnk}
\end{aligned} \tag{137}
$$

考虑约束 $\sum_{l=1}^{K}\eta_{mnl}=1$，构造约束优化问题的拉格朗日函数，并分别关于 η_{mnk}，$k=1,\ldots,K$ 求偏导，得到

$$
\psi(\gamma_{mk})-\psi(\sum_{l=1}^{K}\gamma_{ml})+\psi(\mu_{k,i(w_{mn})})-\psi(\sum_{s=1}^{V}\mu_{ks})-\log\eta_{mnk}-1+\lambda \tag{138}
$$

令偏导数为零，得到 η_{mnk} 的更新公式

$$
\eta_{mnk}=\frac{\exp\left\{\psi(\gamma_{mk}-\psi(\sum_{l=1}^{K}\gamma_{ml})+\psi(\mu_{k,i(w_{mn})})-\psi(\sum_{s=1}^{V}\mu_{ks})\right\}}{\sum_{t=1}^{K}\left(\exp\left\{\psi(\gamma_{mt}-\psi(\sum_{l=1}^{K}\gamma_{ml})+\psi(\mu_{t,i(w_{mn})})-\psi(\sum_{s=1}^{V}\mu_{ts})\right\}\right)} \tag{139}
$$

目标函数中关于 α 的部分：

$$
\mathcal{L}_{[\alpha]}=\sum_{m=1}^{M}\log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)-\sum_{m=1}^{M}\sum_{k=1}^{K}\log\Gamma\left(\alpha_k\right)+\sum_{m=1}^{M}\sum_{k=1}^{K}(\alpha_k-1)\left[\psi(\gamma_{mk})-\psi\left(\sum_{l=1}^{K}\gamma_{ml}\right)\right] \tag{140}
$$

分别关于 α_k，$k=1,\ldots,K$ 求一阶和二阶偏导，得到

$$
\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\alpha_k} &= M\left[\psi\left(\sum_{l=1}^{K}\alpha_l\right)-\psi(\alpha_k)\right]+\sum_{m=1}^{M}\left[\psi(\gamma_{mk})-\psi\left(\sum_{l=1}^{K}\gamma_{ml}\right)\right] \\
\frac{\partial^2\mathcal{L}}{\partial\alpha_k\partial\alpha_t} &= M\left[\psi'\left(\sum_{l=1}^{K}\alpha_l\right)-\mathbb{I}(k=t)\psi'(\alpha_k)\right]
\end{aligned} \tag{141}
$$

由此得到目标函数关于 α 的梯度 $g(\alpha)$ 和 Hessian 矩阵 $H(\alpha)$，应用牛顿法求目标函数关于 α 的最大化，根据以下公式迭代

$$
\alpha_{\text{new}}=\alpha_{\text{old}}-H(\alpha_{\text{old}})^{-1}g(\alpha_{\text{old}}) \tag{142}
$$

目标函数中关于 β 的部分：

$$
\mathcal{L}_{[\beta]}=\sum_{k=1}^{K}\log\Gamma\left(\sum_{v=1}^{V}\beta_v\right)-\sum_{k=1}^{K}\sum_{v=1}^{V}\log\Gamma(\beta_v)+\sum_{k=1}^{K}\sum_{v=1}^{V}(\beta_v-1)\left[\psi(\mu_{kv})-\psi\left(\sum_{s=1}^{V}\mu_{ks}\right)\right] \tag{143}
$$

分别关于 β_v，$v=1,\ldots,V$ 求一阶和二阶偏导，得到

$$
\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\beta_v} &= K\left[\psi\left(\sum_{s=1}^{V}\beta_s\right)-\psi(\beta_v)\right]+\sum_{k=1}^{K}\left[\psi(\mu_{kv})-\psi\left(\sum_{s=1}^{V}\mu_{ks}\right)\right] \\
\frac{\partial^2\mathcal{L}}{\partial\beta_v\partial\beta_l} &= K\left[\psi'\left(\sum_{s=1}^{V}\beta_s\right)-\mathbb{I}(v=l)\psi'(\beta_v)\right]
\end{aligned} \tag{144}
$$

由此得到目标函数关于 β 的梯度 $g(\beta)$ 和 Hessian 矩阵 $H(\beta)$，应用牛顿法求目标函数关于 β 的最大化，根据以下公式迭代

$$\beta_{\text{new}} = \beta_{\text{old}} - H(\beta_{\text{old}})^{-1} g(\beta_{\text{old}}) \tag{145}$$

注意：超参数可以不进行更新，以及论文中推荐先更新局部参数至收敛再更新全局参数[2003 Latent dirichlet allocation]。

2.6 随机变分推断

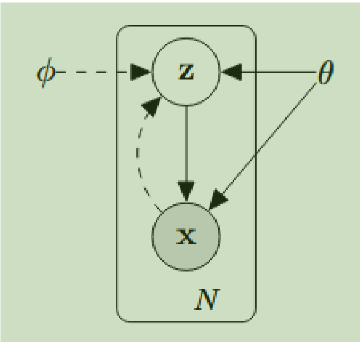
Reference: [2013 Stochastic Variational Inference]
利用指数族分布的良好性质，采用自然梯度代替一般梯度，解决参数向量之间的欧氏距离与它们所代表的概率分布之间的实际统计差异不成比例的问题，并使用批量更新提高效率。

2.7 结构化随机变分推断

Reference: [2015 Structured Stochastic Variational Inference]
放松平均场近似，以允许全局参数和局部隐藏变量之间的任意依赖关系，通过减少偏差、对局部最优解的敏感性和对超参数的敏感性来产生更好的参数估计。

3. 变分自编码器

Reference:[2022 Auto-Encoding Variational Bayes]
标准的变分自编码器（Variational autoencoder）主要用于建模连续的隐变量，对离散隐变量的建模因为采样过程无法传递梯度信息存在问题，但可利用Gumbel-Softmax、STE、梯度旋转（见RQ-VAE github）等技术扩展到离散隐变量，此外 VAE 支持对离散和连续的观测变量进行建模。



考虑 N 个独立同分布（i.i.d.）样本的数据集 $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$，样本 $\mathbf{x}^{(i)}$ 是连续或离散的变量，假设数据由一个随机过程生成，该过程涉及一个未被观测到的连续随机变量 $\mathbf{z}$，具体两步为：（1）从某个先验分布 $p_{\theta^*}(\mathbf{z})$ 中生成一个 $\mathbf{z}^{(i)}$；（2）在给定条件分布 $p_{\theta^*}(\mathbf{x} \mid \mathbf{z})$ 下生成观测样本 $\mathbf{x}^{(i)}$。假设先验分布 $p_{\theta^*}(\mathbf{z})$ 和条件分布 $p_{\theta^*}(\mathbf{x} \mid \mathbf{z})$ 来自参数化的分布族 $p_\theta(\mathbf{z})$ 和 $p_\theta(\mathbf{x} \mid \mathbf{z})$，并且他们的概率密度函数（PDFs）或概率质量函数（PMFs）几乎所有位置关于 θ 和 $\mathbf{z}$ 都是可微的。

VAE想解决两个挑战：
1、Intractability：当 $\mathbf{z}$ 是高维连续变量，且 $p_\theta(\mathbf{x} \mid \mathbf{z})$ 是复杂的函数时，边际似然 $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{z}) p_\theta(\mathbf{x} \mid \mathbf{z}) d\mathbf{z}$ 是难解的，导致无法直接使用以计算损失函数和更新 θ；推断隐变量的真实后验也存在困难，因为 $p_\theta(\mathbf{z} \mid \mathbf{x}) = p_\theta(\mathbf{x} \mid \mathbf{z}) p_\theta(\mathbf{z}) / p_\theta(\mathbf{x})$，导致 EM 算法无法使用，因为 E 步需要真实的后验分布（由式（9）可知）；在不具备共轭性质的复杂模型中，即使是使用平均场假设，变分推断过程中的积分式 $\mathbb{E}_q[\ln p(\mathbf{x}, \mathbf{z})]$（由式（65）可知）也无法得到解析解。
2、A large dataset：数据集过大，以至于批量（更准确地说是全量）优化成本过高；希望使用小批量甚至单个数据点来更新参数。蒙特卡洛 EM 等采样方法通常会很慢，因为它涉及为每个数据点执行昂贵的采样循环。

3.1 变分下界

由式（24）或式（63），可以得到每个数据点的边际似然：

$$\begin{aligned} \log p_\theta(\mathbf{x}^{(i)}) &= \int q_\phi(\mathbf{z} \mid \mathbf{x}^{(i)}) \log \frac{p_\theta(\mathbf{x}^{(i)}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x}^{(i)})} d\mathbf{z} - \int q_\phi(\mathbf{z} \mid \mathbf{x}^{(i)}) \log \frac{p_\theta(\mathbf{z} \mid \mathbf{x}^{(i)})}{q_\phi(\mathbf{z} \mid \mathbf{x}^{(i)})} d\mathbf{z} \\ &= \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}) + D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}^{(i)}) \mid\mid p_\theta(\mathbf{z} \mid \mathbf{x}^{(i)})) \end{aligned} \tag{146}$$

因为 KL 散度的非负性，所以

$$\log p_\theta(\mathbf{x}^{(i)}) \geq \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}[-\log q_\phi(\mathbf{z} \mid \mathbf{x}^{(i)}) + \log p_\theta(\mathbf{x}^{(i)}, \mathbf{z})] \tag{147}$$

变分下界还可以写为

$$\begin{aligned} \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}[-\log \frac{q_\phi(\mathbf{z} \mid \mathbf{x}^{(i)})}{p_\theta(\mathbf{z})} + \log \frac{p_\theta(\mathbf{x}^{(i)}, \mathbf{z})}{p_\theta(\mathbf{z})}] \\ &= -D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}^{(i)}) \mid\mid p_\theta(\mathbf{z})) + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}[\log p_\theta(\mathbf{x}^{(i)} \mid \mathbf{z})] \end{aligned} \tag{148}$$

欲针对变分参数 ϕ 和生成参数 θ 求导和优化变分下界 $\mathcal{L}(\theta, \phi; \mathbf{x}^{(i)})$，然而下界关于变分参数 ϕ 的梯度计算存在一些问题。通常使用一般的蒙特卡洛梯度估计（求导和积分的互换见[2020 Monte Carlo Gradient Estimation in Machine Learning]：

$$\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})}[f(\mathbf{z})] = \mathbb{E}_{q_\phi(\mathbf{z})}[f(\mathbf{z})\nabla_\phi \log q_\phi(\mathbf{z})] \simeq \frac{1}{L}\sum_{l=1}^{L} f(\mathbf{z})\nabla_\phi \log q_\phi(\mathbf{z}^{(l)}) \tag{149}$$

此处 $\mathbf{z}^{(l)} \sim q_\phi(\mathbf{z} \mid \mathbf{x}^{(i)})$，然而这样基于得分函数的蒙特卡洛估计方差非常大（见[2012 Variational Bayesian inference with Stochastic Search]），无法适用于此文研究（但在强化学习中常见，如REINFORCE）。

3.2 Stochastic Gradient Variational Bayes (SGVB) estimator

注意假设的近似后验具有 $q_\phi(\mathbf{z} \mid \mathbf{x})$ 的形式，但是同样可以假设 $q_\phi(\mathbf{z})$，主要用于全局模型参数。需注意式（65）中的 $q(Z)$ 如果是近似局部参数严格表述应为 $q(z_i; \lambda_i)$。

在相对温和的条件限制下，可以用一个可微分的转换 $g_\phi(\epsilon, \mathbf{z})$ 和一个辅助（噪声）变量 ϵ 为近似后验 $q_\phi(\mathbf{z} \mid \mathbf{x})$ 重参数化变量 $\tilde{\mathbf{z}} \sim q_\phi(\mathbf{z} \mid \mathbf{x})$：

$$\tilde{\mathbf{z}} = g_\phi(\epsilon, \mathbf{x}) \quad \text{with } \epsilon \sim p(\epsilon) \tag{150}$$

然后我们可以获得 $q_\phi(\mathbf{z} \mid \mathbf{x})$ 关于函数 $f(\mathbf{z})$ 的期望的蒙特卡洛估计：

$$\mathbb{E}_{q_\phi(\mathbf{z}\mid\mathbf{x}^{(i)})}[f(\mathbf{z})] = \mathbb{E}_{p(\epsilon)}\left[f(g_\phi(\epsilon, \mathbf{x}^{(i)}))\right] \simeq \frac{1}{L}\sum_{l=1}^{L} f(g_\phi(\epsilon^{(l)}, \mathbf{x}^{(i)})) \quad \text{where } \epsilon^{(l)} \sim p(\epsilon) \tag{151}$$

将该技术作用于式（147），得到第一种 SGVB 估计器 $\tilde{\mathcal{L}}^A(\theta, \phi; \mathbf{x}^{(i)}) \simeq \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)})$ （注意其为 ELBO 需最大化）：

$$\tilde{\mathcal{L}}^A(\theta, \phi; \mathbf{x}^{(i)}) = \frac{1}{L}\sum_{l=1}^{L}\left[\log p_\theta(\mathbf{x}^{(i)}, \mathbf{z}^{(i,l)}) - \log q_\phi(\mathbf{z}^{(i,l)} \mid \mathbf{x}^{(i)})\right] \quad \text{where } \mathbf{z}^{(i,l)} = g_\phi(\epsilon^{(i,l)}, \mathbf{x}^{(i)}), \epsilon^{(l)} \sim p(\epsilon) \tag{152}$$

通常式（148）中的 $D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}^{(i)}) \,||\, p_\theta(\mathbf{z}))$ 可以解析的计算，因此只有期望的重构误差 $\mathbb{E}_{q_\phi(\mathbf{z}\mid\mathbf{x}^{(i)})}[\log p_\theta(\mathbf{x}^{(i)} \mid \mathbf{z})]$ 需要使用采样进行估计。这个 KL 散度项可以被理解为规范变分参数 ϕ，促使估计后验尽量和先验 $p_\theta(\mathbf{z})$ 接近。因此针对式（148）可以得到第二种 SGVB 估计器 $\tilde{\mathcal{L}}^B(\theta, \phi; \mathbf{x}^{(i)}) \simeq \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)})$，相较于通用估计器具有更小的方差：

$$\begin{gathered}\tilde{\mathcal{L}}^B(\theta, \phi; \mathbf{x}^{(i)}) = -D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}^{(i)}) \,||\, p_\theta(\mathbf{z})) + \frac{1}{L}\sum_{l=1}^{L}\left[\log p_\theta(\mathbf{x}^{(i)} \mid \mathbf{z}^{(i,l)})\right] \\ \text{where } \mathbf{z}^{(i,l)} = g_\phi(\epsilon^{(i,l)}, \mathbf{x}^{(i)}), \epsilon^{(l)} \sim p(\epsilon)\end{gathered} \tag{153}$$

给定具有 N 个数据点的数据集 X 中的多个数据点，我们可以基于小批量构造全数据集的边际似然的下界（注意这是为了得到整个数据集的无偏估计，实践中使用 batch 内平均损失更稳健；使用平均损失时 Full VB 要注意 weight decay 的设置）：

$$\mathcal{L}(\theta, \phi; X) \simeq \tilde{\mathcal{L}}^M(\theta, \phi; X^M) = \frac{N}{M}\sum_{i=1}^{M}\tilde{\mathcal{L}}(\theta, \phi; \mathbf{x}^{(i)}) \tag{154}$$

作者在实验中发现只要小批量 M 足够大，例如 $M = 100$，每个数据点采样数 L 可以设置为1（因为批次内各个点的独立随机噪声会相互抵消）。

算法5：
初始化原始参数 θ 和变分参数 ϕ
重复：
从完整数据集 X 中选取小批量数据 X^M
从噪声分布 $p(\epsilon)$ 中抽取样本 ϵ
计算梯度 $g \leftarrow \nabla_{\theta,\phi}\tilde{\mathcal{L}}^M(\theta, \phi; X^M, \epsilon)$
利用梯度 g 更新 参数 (θ, ϕ)
直到参数 (θ, ϕ) 收敛

3.3 完全变分贝叶斯

目标是将 VAE 的重参数化技巧和随机梯度下降方法，从仅仅推断局部隐变量 z (latent variables) 扩展到同时推断模型的全局参数 θ (global parameters)。注意在前文中参数 θ 是被当做固定的未知量，通过最大似然（或 MAP）来寻找其点估计 。但在本节中将把 θ 也看作是一个随机变量，给它设定一个先验分布，并使用变分推断来求它的后验分布。

首先，假设参数 θ 服从一个超先验分布 $p_\alpha(\theta)$，其中 α 是超参数。

$$\log p_\alpha(X) = D_{KL}(q_\phi(\theta)||p_\alpha(\theta|X)) + \mathcal{L}(\phi; X) \tag{155}$$

要最大化整个数据集 X 的边际似然 $\log p_\alpha(X)$，只需要最大化后面的变分下界（由式（146）可知）：

$$\mathcal{L}(\phi; X) = \int q_\phi(\theta)\left(\log p_\theta(X) + \log p_\alpha(\theta) - \log q_\phi(\theta)\right)\mathrm{d}\theta \tag{156}$$

因为数据集 X 由 N 个独立同分布的样本组成，所以 $\log p_\theta(X) = \sum_{i=1}^{N} \log p_\theta(\mathbf{x}^{(i)})$，对于每一个单一样本再次使用变分推断来引入关于局部隐变量 $\mathbf{z}$ 的近似后验 $q_\phi(\mathbf{z} \mid \mathbf{x}^{(i)})$：

$$\log p_\theta(\mathbf{x}^{(i)}) = D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}^{(i)}) \mid\mid p_\theta(\mathbf{z}|\mathbf{x}^{(i)})) + \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}) \tag{157}$$

单一样本的变分下界为

$$\mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}) = \int q_\phi(\mathbf{z} \mid \mathbf{x}^{(i)}) \left(\log p_\theta(\mathbf{x}^{(i)} \mid \mathbf{z}) + \log p_\theta(\mathbf{z}) - \log q_\phi(\mathbf{z} \mid \mathbf{x}^{(i)})\right) \mathrm{d}\mathbf{z} \tag{158}$$

分别对隐变量 $\mathbf{z}$ 和全局参数 θ 进行重参数化，可得

$$\tilde{\mathbf{z}} = g_\phi(\boldsymbol{\epsilon}, \mathbf{x}^{(i)}) \quad 其中\ \boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}), \quad \tilde{\boldsymbol{\theta}} = h_\phi(\boldsymbol{\zeta}) \quad 其中\ \boldsymbol{\zeta} \sim p(\boldsymbol{\zeta}) \tag{159}$$

设简化记号（使用单样本估计，为了保证无偏性，必须乘回数据集大小 N）

$$f_\phi(\mathbf{x}, \mathbf{z}, \theta) = N \cdot (\log p_\theta(\mathbf{x} \mid \mathbf{z}) + \log p_\theta(\mathbf{z}) - \log q_\phi(\mathbf{z} \mid \mathbf{x})) + \log p_\alpha(\boldsymbol{\theta}) - \log q_\phi(\boldsymbol{\theta}) \tag{160}$$

最终得到蒙特卡洛估计器（注意到在最终的估计器中省略了数据集的似然中的KL散度）

$$\mathcal{L}(\phi; X) \simeq \frac{1}{L} \sum_{l=1}^{L} f_\phi\left(\mathbf{x}^{(l)}, g_\phi(\boldsymbol{\epsilon}^{(l)}, \mathbf{x}^{(l)}), h_\phi(\boldsymbol{\zeta}^{(l)})\right) \tag{161}$$

如果所有的先验和近似后验都是高斯分布，可以把式（160）中的一些项解析计算出来，从而进一步降低方差 。隐变量先验 $p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$，近似后验 $q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\mu_z, \sigma_z^2\mathbf{I})$ 。参数先验 $p_\alpha(\theta) = \mathcal{N}(\mathbf{0}, \mathbf{I})$，近似后验 $q_\phi(\theta) = \mathcal{N}(\mu_\theta, \sigma_\theta^2\mathbf{I})$ 。

此时，θ 和 $\mathbf{z}$ 都可以写成“均值 + 标准差 $\times$ 标准正态噪声”的形式 。最终的低方差估计器为

$$\begin{aligned}\mathcal{L}(\phi; \mathbf{X}) \simeq \frac{1}{L} \sum_{l=1}^{L} N \cdot &\left(\frac{1}{2} \sum_{j=1}^{J} \left(1 + \log\left((\sigma_{\mathbf{z},j}^{(l)})^2\right) - (\mu_{\mathbf{z},j}^{(l)})^2 - (\sigma_{\mathbf{z},j}^{(l)})^2\right) + \log p_\theta\left(\mathbf{x}^{(i)}\mathbf{z}^{(i)}\right)\right) \\ &+ \frac{1}{2} \sum_{j=1}^{J} \left(1 + \log\left((\sigma_{\theta,j}^{(l)})^2\right) - (\mu_{\theta,j}^{(l)})^2 - (\sigma_{\theta,j}^{(l)})^2\right)\end{aligned} \tag{162}$$

3.4 重参数化技巧

令 $\mathbf{z}$ 为一个连续的随机变量（注意重参数技巧无法直接应用于离散分布），并且 $\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x})$ 作为条件分布，通常可以将其表示为一个确定性变量 $\mathbf{z} = g_\phi(\epsilon, \mathbf{x})$，其中 ϵ 是一个辅助变量，具有边际分布 $p(\epsilon)$，$g_\phi(\cdot)$ 是以 ϕ 为参数的向量函数。重参数化技巧可以用来改写关于 $q_\phi(\mathbf{z} \mid \mathbf{x})$ 的期望，使得该期望的蒙特卡洛估计值关于 ϕ 可微。

首先根据变量代换定理，只要 $\mathbf{z}$ 和 ϵ 间为确定性映射，则

$$q_\phi(\mathbf{z}|\mathbf{x}) \prod_i dz_i = p(\boldsymbol{\epsilon}) \prod_i d\epsilon_i \tag{163}$$

改写期望

$$\int q_\phi(\mathbf{z}|\mathbf{x}) f(\mathbf{z}) d\mathbf{z} = \int p(\boldsymbol{\epsilon}) f(\mathbf{z}) d\boldsymbol{\epsilon} = \int p(\boldsymbol{\epsilon}) f(g_\phi(\boldsymbol{\epsilon}, \mathbf{x})) d\boldsymbol{\epsilon} \tag{164}$$

这样积分的变量里就不含参数 ϕ，采样不阻断梯度的传导

$$\int q_\phi(\mathbf{z}|\mathbf{x}) f(\mathbf{z}) d\mathbf{z} \simeq \frac{1}{L} \sum_{l=1}^{L} f(g_\phi(\boldsymbol{\epsilon}^{(l)}, \mathbf{x})) \quad \text{where } \boldsymbol{\epsilon}^{(l)} \sim p(\boldsymbol{\epsilon}) \tag{165}$$

以单变量高斯分布举例：使 $z \sim p(z \mid x) = \mathcal{N}(\mu, \sigma^2)$，一个有效的重参数化是 $z = \mu + \sigma\epsilon$，其中 ϵ 是辅助的噪声变量 $\epsilon \sim \mathcal{N}(0, 1)$，因此

$$\mathbb{E}_{\mathcal{N}(z;\mu,\sigma^2)}[f(z)] = \mathbb{E}_{\mathcal{N}(\epsilon;0,1)}[f(\mu + \sigma\epsilon)] \simeq \frac{1}{L} \sum_{l=1}^{L} f(\mu + \sigma\epsilon^{(l)}) \quad \text{where } \epsilon^{(l)} \sim \mathcal{N}(0, 1) \tag{166}$$

对于哪些 $q_\phi(\mathbf{z} \mid \mathbf{x})$ 可以选择这样的可微变换 $g_\phi(\cdot)$ 和辅助变量 $\epsilon \sim p(\epsilon)$？有三种基本方法：
1、易处理的逆 CDF。在这种情况下，令 $\epsilon \sim \mathcal{U}(0, 1)$，并令 $g_\phi(\epsilon, \mathbf{x})$ 是 $q_\phi(\mathbf{z} \mid \mathbf{x})$ 的逆 CDF。例如：指数分布、柯西分布、Logistic 分布、瑞利分布、帕累托分布、威布尔分布、倒数分布、Gompertz 分布、Gumbel 分布和 Erlang 分布。
2、类似于高斯分布的例子，对于任何“位置-尺度”（location-scale）分布族，可以选择标准分布（location=0, scale=1）作为辅助变量 ϵ，并令 $g(\cdot) = \text{location} + \text{scale} \cdot \epsilon$。例如拉普拉斯分布、椭圆分布、学生 t 分布、Logistic 分布、均匀分布、三角分布和高斯分布。
3、组合，通常可以将随机变量表达为辅助变量的不同变换。例子：对数正态分布（正态分布变量的指数化）、伽马分布（指数分布变量的总和）、狄利克雷分布（伽马变量的加权总和）、贝塔分布、卡方分布和 F 分布。

当所有三种方法都失败时，存在对逆累积分布函数的良好近似，其计算的时间复杂度与概率密度函数相当（见[1986 Sample-based non-uniform random variate generation]）。

3.5 示例：变分自编码器

假设隐变量的先验分布为 $p_\theta(\mathbf{z})=\mathcal{N}(\mathbf{z};\mathbf{0},\mathbf{I})$，这是简化的做法，通常出于
1、正态分布的良好计算性质，先验与近似后验的 KL 散度可以解析表示。
2、正则化，使其合理的聚集
3、协方差矩阵设为 $\mathbf{I}$ 保证各维度间相互独立，引导模型学习独立的特征

令 $p_\theta(\mathbf{x}\mid\mathbf{z})$ 为多元高斯，并且分布参数为 $\mathbf{z}$ 经过一个 MLP 得到的值，注意此时后验分布 $p_\theta(\mathbf{z}\mid\mathbf{x})$ 是没有解析解的。假设 $q_\phi(\mathbf{z}|\mathbf{x})$ 是一个具有对角协方差矩阵的多元高斯分布

$$\log q_\phi(\mathbf{z}\mid\mathbf{x}^{(i)})=\log\mathcal{N}(\mathbf{z};\boldsymbol{\mu}^{(i)},\boldsymbol{\sigma}^{2(i)}\mathbf{I})\tag{167}$$

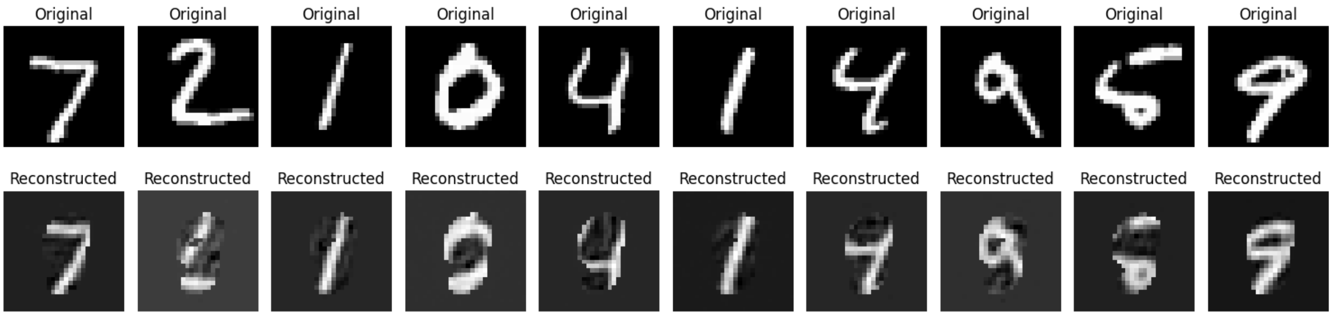
其中 $\boldsymbol{\mu}^{(i)},\boldsymbol{\sigma}^{2(i)}$ 是数据点 $\mathbf{x}^{(i)}$ 经过神经网络参数为变分参数 ϕ 的 MLP 得到的。因此可以采样得到
$\mathbf{z}^{(i,l)}=g_\phi(\mathbf{x}^{(i)},\boldsymbol{\epsilon}^{(i)})=\boldsymbol{\mu}^{(i)}+\boldsymbol{\sigma}^{(i)}\odot\boldsymbol{\epsilon}^{(l)}$ where $\boldsymbol{\epsilon}^{(l)}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$，并根据式（153）可以得到最终的损失函数

$$\mathcal{L}(\theta,\phi;\mathbf{x}^{(i)})\simeq\frac{1}{2}\sum_{j=1}^{J}\left(1+\log((\sigma_j^{(i)})^2)-(\mu_j^{(i)})^2-(\sigma_j^{(i)})^2\right)+\frac{1}{L}\sum_{l=1}^{L}\log p_\theta(\mathbf{x}^{(i)}|\mathbf{z}^{(i,l)})\tag{168}$$

前半部分为两个正态分布的 KL 散度解析表达式，后半部分为重构损失。

3.6 复现实验

Loss A（1000 epoch，z_dim=128）：



Loss B（1000 epoch，z_dim=128）：