变分推断
Variational Inference

西安交通大学管理学院
信息管理与电子商务系
智能决策与机器学习研究中心
刘佳鹏

EM算法回顾

- EM 算法是一种迭代算法，用于含有隐变量的概率模型参数的极大似然估计，或极大后验概率估计

- 问题描述：
- 观测变量 x，隐变量 z，参数 θ
- 观测数据 $X = (x_1, x_2, \ldots, x_n)$
- 未观测数据 $Z = (z_1, z_2, \ldots, z_n)$
- 注：x_i，z_i 和 θ 可能包含多个分量

$$
\begin{aligned}
\theta_{MLE} &= \arg\max_{\theta} \log p(X \mid \theta) \\
&= \arg\max_{\theta} \sum_{i=1}^{n} \log p(x_i \mid \theta)
\end{aligned}
$$

EM算法回顾

- EM算法：
- 初始化参数值 $\theta^{(1)}$，然后交替迭代以下两步骤直至收敛
- for $t = 1, 2, \ldots$
- E步：计算当前参数 $\theta^{(t)}$ 下隐变量 z_i 的后验分布

$$p\left(z_i \mid x_i, \theta^{(t)}\right), \qquad i = 1, \cdots, n$$

并计算完全数据的对数似然函数关于隐变量的后验分布的期望

$$\sum_{i=1}^{n} \sum_{z_i} p\left(z_i \mid x_i, \theta^{(t)}\right) \log p\left(x_i, z_i | \theta\right)$$

- M步：上述期望关于参数 θ 求极大，更新参数 θ

$$\theta^{(t+1)} = \arg\max_{\theta} \sum_{i=1}^{n} \sum_{z_i} p\left(z_i \mid x_i, \theta^{(t)}\right) \log p\left(x_i, z_i | \theta\right)$$

EM算法回顾

- EM算法可以归纳为

$$\begin{aligned}\theta^{(t+1)} &= \arg\max_{\theta} E_{Z|X,\theta^{(t)}}[\log p(X, Z \mid \theta)] \\ &= \arg\max_{\theta} \int_Z \log p(X, Z \mid \theta) \cdot p(Z \mid X, \theta^{(t)}) \,\mathrm{d}\, Z\end{aligned}$$

- E步：计算当前参数 $\theta^{(t)}$ 下隐变量的后验分布

$$p(Z \mid X, \theta^{(t)})$$

  并计算完全数据的对数似然函数关于隐变量的后验分布的期望

$$E_{Z|X,\theta^{(t)}}[\log p(X, Z \mid \theta)] = \int_Z \log p(X, Z \mid \theta) \cdot p(Z \mid X, \theta^{(t)})$$

- M步：上述期望关于参数 θ 求极大，更新参数 θ

$$\theta^{(t+1)} = \arg\max_{\theta} \int_Z \log p(X, Z \mid \theta) \cdot p(Z \mid X, \theta^{(t)}) \,\mathrm{d}\, Z$$

EM算法回顾

- **为什么能保证** $\log p\left(X \mid \theta^{(t+1)}\right) \geqslant \log p\left(X \mid \theta^{(t)}\right)$ **?**

- 第一种证明方式：
- 对等式两边 $\log p(X \mid \theta) = \log p(X, Z \mid \theta) - \log p(Z \mid X, \theta)$ 分别关于隐变量的后验分布求期望
- 左边得到

$$
\begin{aligned}
Left &= \int_Z p(Z \mid X, \theta^{(t)}) \cdot \log p(X \mid \theta) \, \mathrm{d}\, Z \\
&= \log p(X \mid \theta) \int_Z p(Z \mid X, \theta^{(t)}) \, \mathrm{d}\, Z \\
&= \log p(X \mid \theta) \cdot 1 \\
&= \log p(X \mid \theta)
\end{aligned}
$$

EM算法回顾

- 右边得到

$$Right = \underbrace{\int_Z p(Z \mid X, \theta^{(t)}) \cdot \log p(X, Z \mid \theta) \,\mathrm{d}\, Z}_{Q(\theta,\theta^{(t)})} - \underbrace{\int_Z p(Z \mid X, \theta^{(t)}) \cdot \log p(Z \mid X, \theta) \,\mathrm{d}\, Z}_{H(\theta,\theta^{(t)})}$$

- 这里 $Q(\theta, \theta^{(t)}) = \int_Z p(Z \mid X, \theta^{(t)}) \cdot \log p(X, Z \mid \theta) \,\mathrm{d}\, Z$ 即为EM算法中M步的优化目标，因此有

$$Q(\theta^{(t+1)}, \theta^{(t)}) \geqslant Q(\theta^{(t)}, \theta^{(t)})$$

- 而对于 $H(\theta, \theta^{(t)}) = \int_Z p(Z \mid X, \theta^{(t)}) \cdot \log p(Z \mid X, \theta) \,\mathrm{d}\, Z$，可以证明

$$\begin{aligned}
& H(\theta^{(t+1)}, \theta^{(t)}) - H(\theta^{(t)}, \theta^{(t)}) \\
= & \int_Z p(Z \mid X, \theta^{(t)}) \cdot \log p(Z \mid X, \theta^{(t+1)}) \,\mathrm{d}\, Z - \int_Z p(Z \mid X, \theta^{(t)}) \cdot \log p(Z \mid X, \theta^{(t)}) \,\mathrm{d}\, Z \\
= & \int_Z p(Z \mid X, \theta^{(t)}) \cdot \log \frac{p(Z \mid X, \theta^{(t+1)})}{p(Z \mid X, \theta^{(t)})} \,\mathrm{d}\, Z \\
\leqslant & \log \int_Z p(Z \mid X, \theta^{(t)}) \cdot \frac{p(Z \mid X, \theta^{(t+1)})}{p(Z \mid X, \theta^{(t)})} \,\mathrm{d}\, Z \qquad \text{(Jensen不等式)} \\
= & \log \int_Z p(Z \mid X, \theta^{(t+1)}) \,\mathrm{d}\, Z \\
= & \log 1 = 0
\end{aligned}$$

EM算法回顾

- 从而得到

$$
\begin{aligned}
&\log p(X \mid \theta^{(t+1)}) - \log p(X \mid \theta^{(t)}) \\
=&\left[Q(\theta^{(t+1)}, \theta^{(t)}) - H(\theta^{(t+1)}, \theta^{(t)})\right] - \left[Q(\theta^{(t)}, \theta^{(t)}) - H(\theta^{(t)}, \theta^{(t)})\right] \\
=&\left[Q(\theta^{(t+1)}, \theta^{(t)}) - Q(\theta^{(t)}, \theta^{(t)})\right] - \left[H(\theta^{(t+1)}, \theta^{(t)}) - H(\theta^{(t)}, \theta^{(t)})\right] \\
\geqslant&0
\end{aligned}
$$

- 命题得证

EM算法回顾

- 第二种证明方式：
- 引入隐变量 Z 的某种分布 $q(Z)$

$$
\begin{aligned}
\log p(X \mid \theta) &= \log p(X, Z \mid \theta) - \log p(Z \mid X, \theta) \\
&= \log \frac{p(X, Z \mid \theta)}{q(Z)} - \log \frac{p(Z \mid X, \theta)}{q(Z)} \qquad q(Z) \neq 0
\end{aligned}
$$

- 对上式两边分别关于分布 $q(Z)$ 求期望
- 左边得到

$$
\begin{aligned}
Left &= \int_Z q(Z) \cdot \log p(X \mid \theta) \, \mathrm{d}\, Z \\
&= \log p(X \mid \theta) \int_Z q(Z) \, \mathrm{d}\, Z \\
&= \log p(X \mid \theta) \cdot 1 \\
&= \log p(X \mid \theta)
\end{aligned}
$$

- 右边得到

$$
Right = \int_Z q(Z) \log \frac{p(X, Z \mid \theta)}{q(Z)} \, \mathrm{d}\, Z - \int_Z q(Z) \log \frac{p(Z \mid X, \theta)}{q(Z)} \, \mathrm{d}\, Z
$$

EM算法回顾

- 联立得到

$$\begin{aligned}\underbrace{\log p(X \mid \theta)}_{evidence} &= \int_Z q(Z) \log \frac{p(X, Z \mid \theta)}{q(Z)} \,\mathrm{d}\, Z - \int_Z q(Z) \log \frac{p(Z \mid X, \theta)}{q(Z)} \,\mathrm{d}\, Z \\ &= \underbrace{\int_Z q(Z) \log \frac{p(X, Z \mid \theta)}{q(Z)} \,\mathrm{d}\, Z}_{ELBO} + \underbrace{\int_Z q(Z) \log \frac{q(Z)}{p(Z \mid X, \theta)} \,\mathrm{d}\, Z}_{KL(q(Z)||p(Z|X,\theta))}\end{aligned}$$

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

EM算法回顾

- KL散度（Kullback-Leibler divergence, KL divergence）：描述两个概率分布 $q(x)$ 和 $p(x)$ 相似度的一种方式，记为 $KL(q||p)$
- 对于离散随机变量 x

$$KL(q||p) = \sum_i q(i) \log \frac{q(i)}{p(i)}$$

- 对于连续随机变量 x

$$KL(q||p) = \int q(x) \log \frac{q(x)}{p(x)} \,\mathrm{d}\, x$$

- 容易证明KL散度具有性质：$KL(q||p) \geqslant 0$，当且仅当 $q = p$ 时 $KL(q||p) = 0$

$$\begin{aligned} KL(q||p) &= \int q(x) \log \frac{q(x)}{p(x)} \,\mathrm{d}\, x \\ &= -\int q(x) \log \frac{p(x)}{q(x)} \,\mathrm{d}\, x \\ &\geqslant -\log \int q(x) \frac{p(x)}{q(x)} \,\mathrm{d}\, x \\ &= -\log \int p(x) \,\mathrm{d}\, x = 0 \end{aligned}$$

- KL散度是非对称的，也不满足三角不等式，不是严格意义上的距离度量

EM算法回顾

- 回到刚才的推导

$$\underbrace{\log p(X \mid \theta)}_{evidence} = \int_Z q(Z) \log \frac{p(X, Z \mid \theta)}{q(Z)} \,\mathrm{d}\, Z - \int_Z q(Z) \log \frac{p(Z \mid X, \theta)}{q(Z)} \,\mathrm{d}\, Z$$

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

- 从而得到

$$\underbrace{\log p(X \mid \theta)}_{evidence} \geqslant \underbrace{\int_Z q(Z) \log \frac{p(X, Z \mid \theta)}{q(Z)} \,\mathrm{d}\, Z}_{ELBO}$$

上式取等号当且仅当 $q(Z) = p(Z \mid X, \theta)$

EM算法回顾

$$\underbrace{\log p(X \mid \theta)}_{evidence} \geqslant \underbrace{\int_Z q(Z) \log \frac{p(X, Z \mid \theta)}{q(Z)} \, \mathrm{d}\, Z}_{ELBO}$$

$$(\text{等号当且仅当} q(Z) = p(Z \mid X, \theta))$$

- 由此引出EM算法
- E步：固定参数 $\theta^{(t)}$，取 $q(z) = p(Z \mid X, \theta^{(t)})$，此时有

$$\underbrace{\log p(X \mid \theta)}_{evidence} = \underbrace{\int_Z p(Z \mid X, \theta^{(t)}) \log \frac{p(X, Z \mid \theta)}{p(Z \mid X, \theta^{(t)})} \, \mathrm{d}\, Z}_{ELBO}$$

- M步：ELBO关于参数 θ 求最大，更新参数

$$\begin{aligned}
\theta^{(t+1)} &= \arg\max_{\theta} \int_Z p(Z \mid X, \theta^{(t)}) \log \frac{p(X, Z \mid \theta)}{p(Z \mid X, \theta^{(t)})} \, \mathrm{d}\, Z \\
&= \arg\max_{\theta} \int_Z p(Z \mid X, \theta^{(t)}) \log p(X, Z \mid \theta) \, \mathrm{d}\, Z \\
&= \arg\max_{\theta} E_{Z \mid X, \theta^{(t)}} [\log p(X, Z \mid \theta)]
\end{aligned}$$

从狭义EM算法到广义EM算法

- EM算法的目标是通过极大似然估计找到 θ 的最优值，使得 $p(X \mid \theta)$ 达到最大

$$
\begin{aligned}
\theta_{MLE} &= \arg\max_{\theta} \log p(X \mid \theta) \\
&= \arg\max_{\theta} \sum_{i=1}^{n} \log p(x_i \mid \theta)
\end{aligned}
$$

- 证据（evidence）可以分解为

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

- 证据下界（ELBO）可以看成是分布 $q(Z)$ 和参数 θ 的函数

$$
ELBO = \mathcal{L}(q(Z), \theta)
$$

从狭义EM算法到广义EM算法

- 在狭义的EM算法的E步中，将 $q(Z)$ 取为当前参数值 $\theta^{(t)}$ 下隐变量 Z 的后验分布

$$q(Z) = p(Z \mid X, \theta^{(t)})$$

- 这里要求后验分布 $p(Z \mid X, \theta^{(t)})$ 必须有解析解，但这种理想情况只有对于简单模型才成立（比如GMM模型、概率潜在语义分析），而在复杂模型中后验分布 $p(Z \mid X, \theta^{(t)})$ 往往没有解析解（intractable），由此引出下列广义EM算法
- E步：固定参数 θ，证据 $\log p(X \mid \theta)$ 为固定值，此时寻找分布 $q(Z)$ 使得 $KL(q(Z)||p(Z \mid X, \theta))$ 最小，相当于寻找分布 $q(Z)$ 使得ELBO最大

$$\begin{aligned} q(Z)^* &= \arg\min_{q(Z)} KL(q(Z)||p(Z \mid X, \theta)) \\ &= \arg\max_{q(Z)} ELBO \\ &= \arg\max_{q(Z)} \mathcal{L}(q(Z), \theta) \end{aligned}$$

- M步：固定分布 $q(Z)$，ELBO关于参数 θ 求最大

$$\begin{aligned} \theta^* &= \arg\max_{\theta} ELBO \\ &= \arg\max_{\theta} \mathcal{L}(q(Z), \theta) \end{aligned}$$

从狭义EM算法到广义EM算法

- 广义EM算法（Generalized Expectation-Maximization Algorithm, GEM Algorithm）：
- E步：

$$q(Z)^{(t+1)} = \arg\max_{q(Z)} \mathcal{L}\left(q(Z), \theta^{(t)}\right)$$

- M步：

$$\theta^{(t+1)} = \arg\max_{\theta} \mathcal{L}\left(q(Z)^{(t+1)}, \theta\right)$$

- 广义EM算法亦被称为极大-极大算法（Maximization-Maximization Algorithm, MM Algorithm）

从广义EM算法到狭义EM算法

- 狭义EM算法是广义EM算法的特殊情况
- 当隐变量 Z 的后验分布 $p(Z \mid X, \theta^{(t)})$ 有解析解时
- E步：

$$
\begin{aligned}
q(Z)^{(t+1)} &= \arg\max_{q(Z)} \mathcal{L}\left(q(Z), \theta^{(t)}\right) \\
&= p(Z \mid X, \theta^{(t)})
\end{aligned}
$$

- M步：

$$
\begin{aligned}
\theta^{(t+1)} &= \arg\max_{\theta} \mathcal{L}\left(p(Z \mid X, \theta^{(t)}), \theta\right) \\
&= \arg\max_{\theta} \int_Z p(Z \mid X, \theta^{(t)}) \log \frac{p(X, Z \mid \theta)}{p(Z \mid X, \theta^{(t)})} \, \mathrm{d}\, Z \\
&= \arg\max_{\theta} \int_Z p(Z \mid X, \theta^{(t)}) \log p(X, Z \mid \theta) \, \mathrm{d}\, Z - \int_Z p(Z \mid X, \theta^{(t)}) \log p(Z \mid X, \theta^{(t)}) \, \mathrm{d}\, Z \\
&= \arg\max_{\theta} \int_Z p(Z \mid X, \theta^{(t)}) \log p(X, Z \mid \theta) \, \mathrm{d}\, Z \\
&= \arg\max_{\theta} E_{Z \mid X, \theta^{(t)}} [\log p(X, Z \mid \theta)]
\end{aligned}
$$

变分推断

- 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
- 变分推断和马尔可夫链蒙特卡罗法（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）
    - 变分推断则通过解析的方法计算模型的后验概率的近似值（优化）
    - 变分推断更适合解决数据规模很大的学习和推断问题

变分推断

- **为什么关心后验概率 $p(\theta \mid X)$？**
- （1）**推断**（Bayesian 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 X 中推断用户的偏好模型 θ
- （2）**决策**（Bayesian decision theory）：对于新样本 $\tilde{x}$，求 $p(\tilde{x} \mid X)$

$$
\begin{aligned}
p(\tilde{x} \mid X) &= \int_{\theta} p(\tilde{x}, \theta \mid X) \,\mathrm{d}\,\theta \\
&= \int_{\theta} p(\tilde{x} \mid \theta) p(\theta \mid X) \,\mathrm{d}\,\theta \\
&= E_{\theta \mid X}\left[p(\tilde{x} \mid \theta)\right]
\end{aligned}
$$

被称为后验预测分布（Posterior predictive distribution）

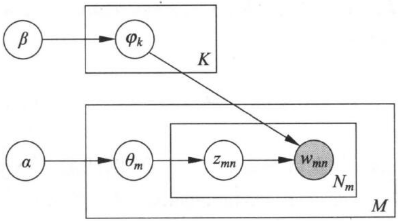
  - 例如：根据用户的历史评分信息 X 中预测用户对于新电影 $\tilde{x}$ 的评分

变分推断

- **变分推断**

- 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：这里和EM算法中的表述略有区别
- (X, Z) 完全数据
- 目标：学习后验分布

$$p(Z \mid X)$$



- 在LDA模型中，观测数据是 $X = \{w_{mn} \mid m = 1, \cdots, M,\ n = 1, \cdots, N_m\}$，隐变量+参数包括
  $Z = \{z_{mn}, \varphi_k, \theta_m \mid k = 1, \cdots, K,\ m = 1, \cdots, M,\ n = 1, \cdots, N_m\}$，超参数为 α 和 β

变分推断

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

$$\Downarrow$$

$$\underbrace{\int_Z q(Z) \log \frac{q(Z)}{p(Z \mid X, \theta)} \, \mathrm{d}\, Z}_{KL(q(Z)||p(Z|X,\theta))} = \underbrace{\log p(X \mid \theta)}_{evidence} - \underbrace{\int_Z q(Z) \log \frac{p(X, Z \mid \theta)}{q(Z)} \, \mathrm{d}\, Z}_{ELBO}$$

- 上式蕴含了变分推断的思想：通过最小化 $KL(q(Z)||p(Z \mid X, \theta))$ 寻找与后验分布 $p(Z \mid X, \theta)$ 最相似的变分分布 $q(Z)$ [1]

$$q(Z)^* = \arg\min_{q(Z)} KL(q(Z)||p(Z \mid X, \theta))$$

  - 后验分布 $p(Z \mid X, \theta)$ 太复杂，直接估计其参数很困难

[1] 数学上把函数的函数称为泛函，求泛函的极值问题称为变分问题

变分推断

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

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

变分推断

- 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
- 一种简单常用的方法是假设 $q(Z)$ 对 $Z = (Z_1, Z_2, \cdots, Z_d)$ 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

$$q(Z) = q(Z_1)q(Z_2)\cdots q(Z_d)$$

  这时的变分分布被称为**平均场**（mean field）[2]
- 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

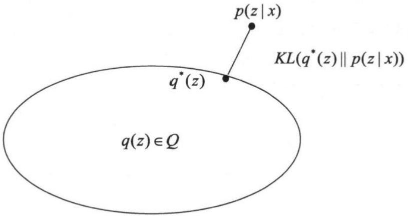
$$q(Z)^* = \arg\max_{q(Z)\in Q} E_{q(Z)}[\log p(X, Z \mid \theta)] - E_{q(Z)}[\log q(Z)]$$

[2] 平均场的概念最初来自于物理学

变分推断

- 变分推断的原理

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



变分推断

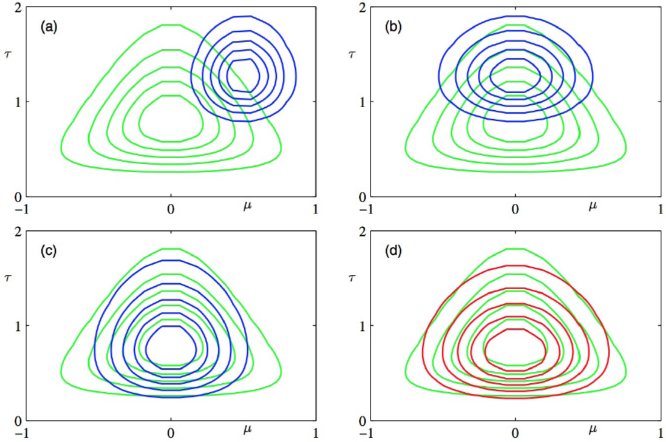
- 假设模型是联合概率分布 $p(X,Z\mid\theta)$，其中 X 是观测变量，Z 是隐变量和参数，θ 是超参数
- 目标是通过观测数据的概率（证据） $\log p(X\mid\theta)$ 的最大化，估计模型的超参数 θ 和变分分布 $q(Z)$
- 应用广义EM算法得到**变分EM算法**（Variational Expectation-Maximization Algorithm）
- 引入平均场 $q(Z)=\prod_{j=1}^{d} q(Z_j)$，定义证据下界

$$\mathcal{L}\left(q(Z),\theta\right)=E_{q(Z)}\left[\log p(X,Z\mid\theta)\right]-E_{q(Z)}\left[\log q(Z)\right]$$

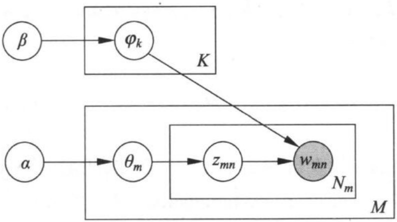
- **变分EM算法：**
- 循环执行以下E步和M步，直至收敛
- E步：固定 θ，求 $\mathcal{L}\left(q(Z),\theta\right)$ 对 $q(Z)$ 的最大化
- M步：固定 $q(Z)$，求 $\mathcal{L}\left(q(Z),\theta\right)$ 对 θ 的最大化

变分推断

- 变分EM示意图



变分推断在LDA模型中的应用



- 在LDA模型中，观测数据是 $X = \{w_{mn} \mid m = 1, \cdots, M,\ n = 1, \cdots, N_m\}$，隐变量+参数包括 $Z = \{z_{mn}, \varphi_k, \theta_m \mid k = 1, \cdots, K,\ m = 1, \cdots, M,\ n = 1, \cdots, N_m\}$，超参数为 α 和 β
- 完全数据 (X, Z) 的对数似然函数

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

变分推断在LDA模型中的应用

- 定义基于平均场的变分分布

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

  其中 $\mu_k = (\mu_{k1}, \mu_{k2}, \cdots, \mu_{kV})$ 和 $\gamma_m = (\gamma_{m1}, \gamma_{m2}, \cdots, \gamma_{mK})$ 是狄利克雷分布的参数，$\eta_{mn} = (\eta_{mn1}, \eta_{mn2}, \cdots, \eta_{mnK})$ 是多项分布的参数
- 在变分分布 $q(\boldsymbol{z}, \varphi_{1:K}, \theta_{1:M})$ 中，变量 $\boldsymbol{z}, \varphi_{1:K}, \theta_{1:M}$ 的各个分量之间都是条件独立的
- 目标：求KL散度意义下与LDA模型的后验分布 $p(\boldsymbol{z}, \varphi_{1:K}, \theta_{1:M} \mid \boldsymbol{w}, \alpha, \beta)$ 最相似的变分分布 $q(\boldsymbol{z}, \varphi_{1:K}, \theta_{1:M} \mid \mu_{1:K}, \gamma_{1:M}, \{\eta_{mn}\})$

变分推断在LDA模型中的应用

- 定义证据下界

$$
\begin{aligned}
ELBO =& E_{q(\boldsymbol{z},\varphi_{1:K},\theta_{1:M}\mid\mu_{1:K},\gamma_{1:M},\{\eta_{mn}\})}\left[\log p\left(\boldsymbol{w},\boldsymbol{z},\varphi_{1:K},\theta_{1:M}\mid\alpha,\beta\right)\right] \\
&- E_{q(\boldsymbol{z},\varphi_{1:K},\theta_{1:M}\mid\mu_{1:K},\gamma_{1:M},\{\eta_{mn}\})}\left[\log q(\boldsymbol{z},\varphi_{1:K},\theta_{1:M}\mid\mu_{1:K},\gamma_{1:M},\{\eta_{mn}\})\right] \\
=& \sum_{m=1}^{M} E_{q(\theta_m\mid\gamma_m)}\left[\log p(\theta_m\mid\alpha)\right] \qquad (1) \\
&+ \sum_{k=1}^{K} E_{q(\varphi_k\mid\mu_k)}\left[\log p(\varphi_k\mid\beta)\right] \qquad (2) \\
&+ \sum_{m=1}^{M}\sum_{n=1}^{N_m} E_{q(z_{mn},\theta_m\mid\eta_{mn},\gamma_m)}\left[\log p(z_{mn}\mid\theta_m)\right] \qquad (3) \\
&+ \sum_{m=1}^{M}\sum_{n=1}^{N_m} E_{q(\varphi_{1:K},z_{mn}\mid\mu_{1:K},\eta_{mn})}\left[\log p(w_{mn}\mid\varphi_{1:K},z_{mn})\right] \qquad (4) \\
&- \sum_{k=1}^{K} E_{q(\varphi_k\mid\mu_k)}\left[\log q(\varphi_k\mid\mu_k)\right] \qquad (5) \\
&- \sum_{m=1}^{M} E_{q(\theta_m\mid\gamma_m)}\left[\log q(\theta_m\mid\gamma_m)\right] \qquad (6) \\
&- \sum_{m=1}^{M}\sum_{n=1}^{N_m} E_{q(z_{mn}\mid\eta_{mn})}\left[\log q(z_{mn}\mid\eta_{mn})\right] \qquad (7)
\end{aligned}
$$

变分推断在LDA模型中的应用

- 第（1）项

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

此处用到狄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[3]，ψ 是digamma函数，即对数伽马函数的一阶导数

[3] 详见李航《统计学习方法》第2版P455

变分推断在LDA模型中的应用

- 第（2）项

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

变分推断在LDA模型中的应用

- 第（3）项

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

变分推断在LDA模型中的应用

- 第（4）项

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

其中 $i(w_{mn}) \in \{1,\cdots,V\}$ 表示单词 w_{mn} 的索引

变分推断在LDA模型中的应用

- 第（5）项

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

变分推断在LDA模型中的应用

- 第（6）项

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

变分推断在LDA模型中的应用

- 第（7）项

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

变分推断在LDA模型中的应用

- 上述七部分合并得到

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

变分推断在LDA模型中的应用

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

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

分别关于 μ_{kv}, $v = 1, \cdots, V$ 求偏导，得到

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

令偏导数为零，得到 μ_{kv} 的更新公式

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

变分推断在LDA模型中的应用

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

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

分别关于 γ_{mk}, $k = 1, \cdots, K$ 求偏导，得到

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

令偏导数为零，得到 γ_{mk} 的更新公式

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

变分推断在LDA模型中的应用

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

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

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}, $k = 1, \cdots, K$ 求偏导，得到

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

令偏导数为零，得到 η_{mnk} 的更新公式

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

变分推断在LDA模型中的应用

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

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

  分别关于 α_k, $k = 1, \cdots, K$ 求一阶和二阶偏导，得到

$$\frac{\partial\mathcal{L}}{\partial\alpha_k} = M\left[\psi\left(\sum_{l=1}^{K}\alpha_l\right) - \psi\left(\alpha_k\right)\right] + \sum_{m=1}^{M}\left[\psi(\gamma_{mk}) - \psi(\sum_{l=1}^{K}\gamma_{ml})\right]$$

$$\frac{\partial^2\mathcal{L}}{\partial\alpha_k\alpha_t} = M\left[\psi'\left(\sum_{l=1}^{K}\alpha_l\right) - \mathbb{I}\left(k = t\right)\psi'\left(\alpha_k\right)\right]$$

  由此得到目标函数 $\mathcal{L}(\mu_{1:K}, \gamma_{1:M}, \{\eta_{mn}\}, \alpha, \beta)$ 关于 α 的梯度 $g(\alpha)$ 和Hessian矩阵 $H(\alpha)$
- 应用牛顿法求目标函数 $\mathcal{L}(\mu_{1:K}, \gamma_{1:M}, \{\eta_{mn}\}, \alpha, \beta)$ 关于 α 的最大化，根据以下公式迭代

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

变分推断在LDA模型中的应用

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

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

分别关于 β_v, $v = 1, \cdots, V$ 求一阶和二阶偏导，得到

$$\frac{\partial\mathcal{L}}{\partial\beta_v} = K\left[\psi\left(\sum_{s=1}^{V}\beta_s\right) - \psi(\beta_v)\right] + \sum_{k=1}^{K}\left[\psi(\mu_{kv}) - \psi(\sum_{s=1}^{V}\mu_{ks})\right]$$

$$\frac{\partial^2\mathcal{L}}{\partial\beta_v\beta_l} = K\left[\psi'\left(\sum_{s=1}^{V}\beta_s\right) - \mathbb{I}(v = l)\,\psi'(\beta_v)\right]$$

由此得到目标函数 $\mathcal{L}(\mu_{1:K}, \gamma_{1:M}, \{\eta_{mn}\}, \alpha, \beta)$ 关于 β 的梯度 $g(\beta)$ 和Hessian矩阵 $H(\beta)$

- 应用牛顿法求目标函数 $\mathcal{L}(\mu_{1:K}, \gamma_{1:M}, \{\eta_{mn}\}, \alpha, \beta)$ 关于 β 的最大化，根据以下公式迭代

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

变分推断在LDA模型中的应用

- **LDA模型的变分EM算法**
- 输入：文本的单词序列 $\boldsymbol{w} = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_m, \cdots, \boldsymbol{w}_M\}$, $\boldsymbol{w}_m = (w_{m1}, \cdots, w_{mn}, \cdots, w_{mN_m})$
- 输出：变分分布的参数 $\mu_{1:K}, \gamma_{1:M}, \{\eta_{mn}\}$，模型超参数 α, β
- 交替迭代E步和M步，直至收敛
- E步：固定模型超参数 α, β，按下式更新变分分布的参数 $\mu_{1:K}, \gamma_{1:M}, \{\eta_{mn}\}$

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

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

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

$$m = 1, \cdots, M, \quad n = 1, \cdots, N_m, \quad k = 1, \cdots, K$$

- M步：固定变分分布的参数 $\mu_{1:K}, \gamma_{1:M}, \{\eta_{mn}\}$，使用牛顿法更新模型超参数 α, β

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

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

MCMC和变分推断的比较

- MCMC和变分推断（VI）都是贝叶斯模型中用来近似参数和隐变量的后验分布的近似推断方法
- MCMC属于随机近似方法，通过采样一组满足后验分布的样本来近似后验分布（本质上是采样问题）
  - 理论上保障只要马尔可夫链到达平稳分布，采样得到的样本一定是符合后验分布的
  - 当模型复杂时（先验和似然不共轭），采样接受率比较低（MH），马尔可夫链难以到达平稳分布（MH & Gibbs）
  - MH和Gibbs的改进：哈密顿蒙特卡罗（Hamiltonian Monte Carlo）
    - Pyro [4] – Pyro is a universal probabilistic programming language (PPL) written in Python and supported by PyTorch
- VI属于确定近似方法，通过变分分布来近似后验分布（本质上是优化问题）
  - 得到的结果可能是有偏的（设定了与真实后验分布差异很大的变分分布，平均场变分分布往往会低估了方差）
  - 可以采用优化技巧（并行化、SGD、Adam等非凸优化算法），适合解决大数据场景下的问题
- 模型精度（Accuracy）和求解效率（Efficiency）之间的权衡（Trade-off）

[4] http://pyro.ai/