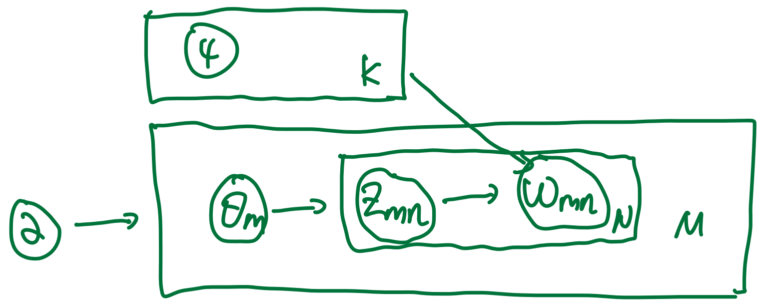
LDA



$W=\{w_1, w_2, \cdots w_N\}$

$Z=\{z_1, z_2, \cdots z_N\}$

· 先考虑一篇文本，即 $M=1$

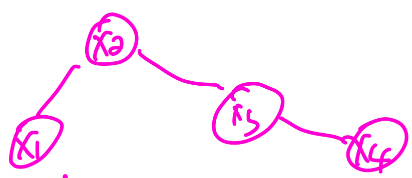
联合：

$$p(\theta, z, w \mid \alpha, \varphi) = p(\theta \mid \alpha) \cdot \prod_{n=1}^{N} p(z_n \mid \theta) \cdot p(w_n \mid z_n, \varphi)$$

其中 w 是 observed，其余是 latent

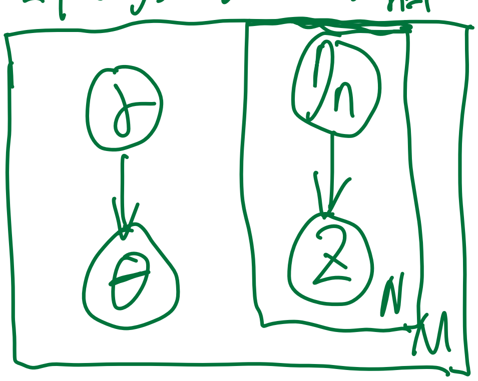
后验：$p(\theta, z \mid w, \alpha, \varphi)$

· 由 mean-field $q(z) = \prod_{i=1}^{M} q_i\, z_i$ （之前是用来表示无向图的）



$$P(X) = \frac{1}{Z} \varphi_{12}(x_1, x_2)\varphi_{23}(x_2, x_3)\varphi_{34}(x_3, x_4)$$

$$q(\theta, z \mid \gamma, \eta) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \eta_n)$$



γ: Dir 参数

η_n: $\{\eta_1, \eta_2, \cdots \eta_n\}$ 多项分布

此时 θ 和 z 独立，而原后验不独立

$\Downarrow$

这个就是构造的 q 函数，用来近似的 $P(\theta, z \mid w, \alpha, \varphi)$

$\log P(x) = \text{ELBO} + \text{KL divergence}$

$$= E_{q(x)} \log \frac{P(x,y)}{q(x)} + E_{q(x)} \log \frac{q(x) \leftarrow \text{构造的 } q(x)}{P(y \mid x) \leftarrow \text{后验}}$$

此时，ELBO：

$$L(\gamma, \eta, \theta, z) = E_{q(\theta, z \mid \gamma, \eta)} \log \frac{P(\theta, z, w \mid \alpha, \varphi)}{q(\theta, z \mid \gamma, \eta)}$$

$$= E_q \log P(\theta, z, w \mid \alpha, \varphi) - E_q \log q(\theta, z \mid \gamma, \eta)$$

展开：

$$L(\alpha, \eta, \theta, z) = E_q \log P(\theta \mid \alpha) + E_q \log P(\boldsymbol{z} \mid \theta)$$
$$+ E_q \log(\boldsymbol{W} \mid \boldsymbol{z}, \varphi) - E_q \log P(\theta \mid \gamma) - E_q \log P(\boldsymbol{z} \mid \eta)$$

$\boldsymbol{z}$ 和 $\boldsymbol{W}$ 都是 vector

Δ：这式融合了模型参数 α, β 和变分参数 φ 和 η

对 $L(\alpha, \eta, \theta, z)$ 一项一项研究：

第1项： $E_q \log P(\theta \mid \alpha)$

$P(\theta \mid \alpha)$ 服从 Dir 分布

$$P(\theta \mid \alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$$

$$\log p(\theta|\alpha) = \log\left[\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)\right] - \sum_{i=1}^{K}\log\Gamma(\alpha_k) + \sum_{i=1}^{K}(\alpha_k-1)\log\theta_i$$

$$E_q \log p(\theta|\alpha) = E_q \log\left[\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)\right] - E_q\left(\sum_{i=1}^{K}\log\Gamma(\alpha_k)\right]$$

$$+ E_q\left(\sum_{k=1}^{K}(\alpha_k-1)\log\theta_k\right)$$

$q \to q(\theta, z|\gamma, \varphi)$

只有该项有θ

$$= \log\left[\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)\right] - \sum_{k=1}^{K}\log\Gamma(\alpha_k)$$

$$+ \sum_{k=1}^{K}(\alpha_k-1)\, E_q \log\theta_k$$

公式：$E_{q(\theta|\gamma)} \ln(\theta_i) = \Psi(\gamma_i) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)$

where Ψ is digamma func.

易混：

$p(\theta|\alpha)$

$q(\theta|\gamma)$

由公式：

$$E_{q(\theta|\gamma)} \log\theta_k = \Psi(\gamma_k) - \Psi\left(\sum_{k=1}^{K}\gamma_k\right)$$

$$\therefore E_q \log p(\theta|\alpha) = \log\left[\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)\right] - \sum_{k=1}^{K}\log\Gamma(\alpha_k)$$

$$+ \sum_{k=1}^{K}(\alpha_k-1)\left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K}\gamma_l\right)\right]$$

其中，α_k，γ_k 分别是第 k 个文档的 dir 参数

第2项：$E_q \log p(z|\theta)$

$$E_q \log p(z|\theta) = \sum_{n=1}^{N} E_q \log p(z_n|\theta)$$

$q \to q(\theta, z|\gamma, \varphi)$

$$= \sum_{n=1}^{N} E_{q(\theta, z|\gamma, \varphi)} \log p(z_n|\theta)$$

$z_n \sim \mathrm{multi}(\theta)$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} q(z_n|\varphi) E_{q(\theta|\gamma)} \log \theta_k \qquad \theta_k^{z_{nk}} \quad \left(z_{nk}=1, =0\right)$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} q(z_{nk}|\varphi)\left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K} \gamma_l\right)\right] \qquad E(x) = \sum_{k=1}^{K} p_k x_k$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \eta_{nk} \left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K} \gamma_l\right)\right] \qquad z_n \sim \text{multi}(\theta)$$

$$\Rightarrow \theta^{z_i}$$

η_{nk}：文档第 n 个单词由第 k 个话题产生的概率

第3项： $E_q \log(W|Z, \varphi)$

$$= \sum_{n=1}^{N} E_q \log(W|z_n, \varphi) \qquad q \to q(\theta, z|\gamma, \varphi)$$

$$= \sum_{n=1}^{N} E_{q(z_n|\eta)} \log(w|z_n, \varphi)$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} q(z_{nk}|\eta) \log(w|z_{nk}, \varphi) \qquad E(x) = \sum_{k=1}^{K} x_k p_k$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{v=1}^{V} \eta_{nk} w_n^v \log \varphi_{kv}$$

式中 η_{nk} 表示文档第 n 个位置的单词由第 k 个话题产生的概率，w_n^v 在第 n 个位置的单词是单词集合的第 v 个单词时取值为 1，否则取值为 0，φ_{kv} 表示第 k 个话题生成单词集合中第 v 个单词的概率。

第4项：$E_q \log p(\theta|\gamma)$

$q \to q(\theta, z|\gamma, \varphi)$

$\theta \sim Dir(\gamma)$ 类似和项

$$E_q[\log q(\theta|\gamma)] = \log\Gamma\left(\sum_{l=1}^{K}\gamma_l\right) - \sum_{k=1}^{K}\log\Gamma(\gamma_k) + \sum_{k=1}^{K}(\gamma_k - 1)\left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K}\gamma_l\right)\right] \tag{20.54}$$

式中 γ_k 表示第 k 个话题的狄利克雷分布参数。

第5项：$E_q \log p(z|\eta)$

$$= \sum_{n=1}^{N} E_{q(z_n|\eta)} \log p(z_n|\eta)$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} q(z_{nk}|\eta) \log p(z_{nk}|\eta)$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} \eta_{nk} \log \eta_{nk}$$

∴ 联合

$$\begin{aligned}
L(\gamma,\eta,\alpha,\varphi) = {} & \log\Gamma\left(\sum_{l=1}^{K}\alpha_l\right) - \sum_{k=1}^{K}\log\Gamma(\alpha_k) + \sum_{k=1}^{K}(\alpha_k - 1)\left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K}\gamma_l\right)\right] + \\
& \sum_{n=1}^{N}\sum_{k=1}^{K}\eta_{nk}\left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K}\gamma_l\right)\right] + \\
& \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{v=1}^{V}\eta_{nk}w_n^v \log\varphi_{kv} - \\
& \log\Gamma\left(\sum_{l=1}^{K}\gamma_l\right) + \sum_{k=1}^{K}\log\Gamma(\gamma_k) - \sum_{k=1}^{K}(\gamma_k - 1)\left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K}\gamma_l\right)\right] - \\
& \sum_{n=1}^{N}\sum_{k=1}^{K}\eta_{nk}\log\eta_{nk}
\end{aligned} \tag{20.47}$$

式中 $\Psi(\alpha_k)$ 是对数伽马函数的导数，即

$$\Psi(\alpha_k) = \frac{\mathrm{d}}{\mathrm{d}\alpha_k}\log\Gamma(\alpha_k) \tag{20.48}$$

VI EM algorithm

在E步 骤，我们使用近似于原始 LDA 后验分布的变分分布去寻找变分参数（variational parameter）的最优解。在M步骤，使用E步骤中已获得的变分参数估计去估计（比如极大似然估计（parameter estimate））原模型参数（model parameter）， 从而进一步最大化模型的下界（bound）。

算法 overview:

算法 20.5（**LDA 的变分 EM 算法**）

输入：给定文本集合 $D = \{\mathbf{w}_1, \cdots, \mathbf{w}_m, \cdots, \mathbf{w}_M\}$；

输出：变分参数 γ，η，模型参数 α，φ。

交替迭代 E 步和 M 步，直到收敛。

（1）E 步

固定模型参数 α，φ，通过关于变分参数 γ，η 的证据下界的最大化，估计变分参数 γ，η。具体见算法 20.4。

（2）M 步

固定变分参数 γ，η，通过关于模型参数 α，φ 的证据下界的最大化，估计模型参数 α，φ。具体算法见式 (20.63) 和式 (20.67)。

根据变分参数 (γ, η) 可以估计模型参数 $\theta = (\theta_1, \cdots, \theta_m, \cdots, \theta_M)$，$\mathbf{z} = (\mathbf{z}_1, \cdots, \mathbf{z}_m, \cdots, \mathbf{z}_M)$。■

算法 9.1（**EM 算法**）

输入：观测变量数据 Y，隐变量数据 Z，联合分布 $P(Y, Z|\theta)$，条件分布 $P(Z|Y, \theta)$；

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 θ 的估计值，在第 $i+1$ 次迭代的 E 步，计算

$$\begin{aligned} Q(\theta, \theta^{(i)}) &= E_Z[\log P(Y, Z|\theta)|Y, \theta^{(i)}] \\ &= \sum_Z \log P(Y, Z|\theta) P(Z|Y, \theta^{(i)}) \end{aligned} \tag{9.9}$$

这里，$P(Z|Y, \theta^{(i)})$ 是在给定观测数据 Y 和当前的参数估计 $\theta^{(i)}$ 下隐变量数据 Z 的条件概率分布；

（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 $\theta^{(i+1)}$

$$\theta^{(i+1)} = \arg\max_{\theta} Q(\theta, \theta^{(i)}) \tag{9.10}$$

（4）重复第 (2) 步和第 (3) 步，直到收敛。■

式 (9.9) 的函数 $Q(\theta, \theta^{(i)})$ 是 EM 算法的核心，称为 Q 函数（Q function）。

对于GMM来说，后验$P(Z|Y,\theta^{(i)})$是可求的，但对于LDA来说，后验不可求，用构造的q逼近真实后验P。q由参数γ,η决定。E步估算出γ,η。具体方法：写出联合，固定模型参数，引进lagrangian乘子。

E步算法 overview:

(VI章的平均场中有叠代算法)

算法 20.4（**LDA 的变分参数估计算法**）

（1）初始化：对所有 k 和 n，$\eta_{nk}^{(0)} = 1/K$

（2）初始化：对所有 k，$\gamma_k = \alpha_k + N/K$

（3）重复

（4）　　对 $n = 1$ 到 N

（5）　　　　对 $k = 1$ 到 K

（6）　　　　　　$\eta_{nk}^{(t+1)} = \varphi_{kv} \exp\left[\Psi(\gamma_k^{(t)}) - \Psi\left(\sum_{l=}^{K} \gamma_l^{(t)}\right)\right]$

（7）　　　　规范化 $\eta_{nk}^{(t+1)}$ 使其和为 1

（8）　　$\gamma^{(t+1)} = \alpha + \sum_{n=1}^{N} \eta_n^{(t+1)}$

（9）直到收敛

推导:

$L(\gamma, \eta, \alpha, \varphi)$ 包含 η 的项：第2、3、6项

$$L_{(\eta)} = \sum_{n=1}^{N}\sum_{k=1}^{K} \eta_{nk}\left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K}\gamma_l\right)\right] + \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{v=1}^{V} \eta_{nk} w_n^v \log \varphi_{kv} + \sum_{n=1}^{N}\sum_{k=1}^{K} \eta_{nk} \log \eta_{nk}$$

η_{nk}：第n个位置的单词是由第k个话题生成的概率

Besides，有约束如下：$\sum_{l=1}^{K} \eta_{nl} = 1$

∴ Lagrangian：

$$L(\eta) = \underbrace{\eta_{nk}\left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K}\gamma_l\right)\right]}_{2} + \underbrace{\eta_{nk}\log\varphi_{kv}}_{3} - \underbrace{\eta_{nk}\log\eta_{nk}}_{5} + \lambda_n\left(\sum_{l=1}^{K}\eta_{nl} - 1\right)$$

φ_{kv}：在第n个位置由第k个话题生成第v个单词的概率