1. (a) $$I(X,Y) = \sum_y \sum_x p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$
$$= \sum_y \sum_x p(x,y) [\log \frac{p(x,y)}{p(x)} - \log p(y)]$$
$$= \sum_y \sum_x p(x,y) [\log p(y|x) - \log p(y)]$$
$$= \sum_y \sum_x p(x,y) (-\log p(y)) + \sum_y \sum_x p(x,y) \log p(y|x)$$
$$= - \sum_y \log p(y) (\sum_x p(x,y)) - [-(\sum_y \sum_x p(y|x) p(x) \log p(y|x))]$$
$$= - \sum_y \log p(y) \cdot p(y) - [- \sum_x p(x) \sum_y p(y|x) \log p(y|x)]$$
$$= H(Y) - \sum_x p(x) H(Y|X=x) = H(Y) - H(Y|X)$$

Since $I(X,Y)$ is symmetric for X, Y, Similarly we can get $I(X,Y) = H(X) - H(X|Y)$

(b) From (a) we know, $I(X,Y) = H(X) - H(X|Y)$, $\Rightarrow$

$I(X,Y) = H(X) + \sum_x p(x) \sum_y p(y|x) \log p(y|x)$. Since $X = f(Y)$, then $p(y|x) = 0$ or 1, for each case, $\sum_x p(x) \sum_y p(y|x) \log p(y|x) = 0$, then $I(X,Y) = H(X) + 0 = H(X)$. Similarly, $I(X,Y) = H(Y)$

(c) $$\min_\theta D_{KL}(\hat{p} \| q) = H(\hat{p}, q) - H(\hat{p}) \propto H(\hat{p}, q)$$

(Noticing only q contains θ here)

$$H(\hat{p}, q) = -\sum_x \hat{p}(x) \log q(x|\theta)$$
$$= -\sum_x \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{\{x=x_i\}} \log q(x|\theta)$$
$$= -\frac{1}{N} \sum_{i=1}^N \log(q(x_i|\theta))$$

Thus $$\min_\theta D_{KL}(\hat{p} \| q) \Leftrightarrow \min_\theta -\frac{1}{N} \sum_{i=1}^N \log q(x_i|\theta)$$
$$\Leftrightarrow \max_\theta \sum_{i=1}^N \log q(x_i|\theta)$$
$$\Leftrightarrow \max_\theta \prod_{i=1}^N q(x_i|\theta) \longrightarrow \text{likelyhood}$$

Thus the minimum Kullback-Leibler divergence is obtained by the maximum likelihood.

(d) For continuous variable,

we have $H(x) = -\int p(x) \ln p(x) dx$,

From the question, it has three constraint:

$$\text{s.t.} \begin{cases} \int p(x) dx = 1 \\ \int x p(x) dx = \mu \\ \int (x-\mu)^2 p(x) dx = 6^2 \end{cases}$$

Use Lagrange multipliers to max $H(x)$:

$$L = -\int p(x) \ln p(x) dx + \lambda_1 \left(\int p(x) dx - 1\right) + \lambda_2 \left(\int x p(x) dx - \mu\right) + \lambda_3 \left(\int (x-\mu)^2 p(x) dx - 6^2\right)$$

$$\frac{df}{dp(x)} = -\ln p(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 = 0$$

$$\Rightarrow p(x) = e^{-1+\lambda_1+\lambda_2 x+\lambda_3 (x-\mu)^2}$$

with
$$\begin{cases} \int e^{-1+\lambda_1+\lambda_2 x+\lambda_3 (x-\mu)^2} dx = 1 \\ \int x e^{-1+\lambda_1+\lambda_2 x+\lambda_3 (x-\mu)^2} dx = \mu \\ \int (x-\mu)^2 e^{-1+\lambda_1+\lambda_2 x+\lambda_3 (x-\mu)^2} dx = 6^2 \end{cases} \Rightarrow p(x) = \frac{1}{(2\pi 6^2)^{\frac{1}{2}}} e^{-\frac{(x-\mu)^2}{2 6^2}}$$

$\therefore$ $p = N(\mu, 6^2)$ is the solution of max $H(x)$

$\Leftrightarrow$ $H(q) \leq H(p)$ for q be any probability density with mean μ and variance 6^2.

2. (a) $\text{Dirichlet}(p|\alpha) = \frac{\Gamma(\sum_1^m \alpha_k)}{\prod_{k=1}^m \Gamma(\alpha_k)} \prod_{k=1}^m p_k^{\alpha_k - 1}$

$= \exp\left(\log \frac{\Gamma(\sum_1^m \alpha_k)}{\prod_{k=1}^m \Gamma(\alpha_k)} \prod_{k=1}^m p_k^{\alpha_k - 1} \right)$

$= \exp[\log\Gamma(\sum_1^m \alpha_k) + \sum_1^m (\alpha_k - 1)\log p_k - \sum_1^m \log\Gamma(\alpha_k)]$

$= \exp[\sum_1^m (\alpha_k - 1)\log p_k - (\sum_1^m \log\Gamma(\alpha_k) - \log\Gamma(\sum_1^m \alpha_k))]$

Thus, set $\eta(\alpha) = \begin{pmatrix} \alpha_1 - 1 \\ \alpha_2 - 1 \\ \vdots \\ \alpha_m - 1 \end{pmatrix}$, $T(p) = \begin{pmatrix} \log p_1 \\ \log p_2 \\ \vdots \\ \log p_m \end{pmatrix}$,

$A(\alpha) = \sum_1^m \log\Gamma(\alpha_k) - \log\Gamma(\sum_1^m \alpha_k)$. Then $D(p|\alpha) = \exp[\eta(\alpha)^T T(p) - A(\alpha)]$

(b) The Dirichlet log-likehood function $F(\alpha) = \log P(D|\alpha) = \sum_{j=1}^N \log D(p^j|\alpha)$

$\therefore F(\alpha) = \sum_{j=1}^N [\sum_{k=1}^m (\alpha_k - 1)\log p_k^j + A(\alpha)]$

$= \sum_{k=1}^m (\alpha_k - 1)\sum_{j=1}^N \log p_k^j + N A(\alpha)$

$= N\sum_{k=1}^m (\alpha_k - 1)\frac{1}{N}\sum_{j=1}^N \log p_k^j + N[\log\Gamma(\sum_{k=1}^m \alpha_k) - \sum_{k=1}^m \log\Gamma(\alpha_k)]$

$= N[\sum_{k=1}^m (\alpha_k - 1)\hat{t}_k + \log\Gamma(\sum_{k=1}^m \alpha_k) - \sum_{k=1}^m \log\Gamma(\alpha_k)]$, $\hat{t}_k = \frac{1}{N}\sum_{j=1}^N \log p_k^j$

(c) $\frac{\partial F}{\partial \alpha_k} = N\left(\hat{t}_k + \frac{\partial \log\Gamma(\sum_{k=1}^m \alpha_k)}{\partial \sum_{k=1}^m \alpha_k} \cdot 1 - \frac{\partial \log\Gamma(\alpha_k)}{\partial \alpha_k}\right)$

$= N(\hat{t}_k + \Psi(\sum_{k=1}^m \alpha_k) - \Psi(\alpha_k))$, $k = 1, \cdots m$

$\therefore \frac{\partial F}{\partial \alpha} = \left(\frac{\partial F}{\partial \alpha_1}, \frac{\partial F}{\partial \alpha_2}, \cdots \frac{\partial F}{\partial \alpha_m}\right)$

(d) $\frac{\partial^2 F}{\partial \alpha_k^2} = N(\Psi'(\sum_{k=1}^m \alpha_k) - \Psi'(\alpha_k))$, $\frac{\partial^2 F}{\partial \alpha_k \alpha_j} = N\Psi'(\sum_{k=1}^m \alpha_k)$

Thus set $q_{kk} = -N\Psi'(\alpha_k)$, $c = N\Psi'(\sum_{k=1}^m \alpha_k)$.

then $\nabla_\alpha^2 F(\alpha) = \left(\frac{\partial F}{\partial \alpha_k \alpha_j}\right)_{k,j=1,\cdots m} = Q + c\mathbb{1}\mathbb{1}^T$

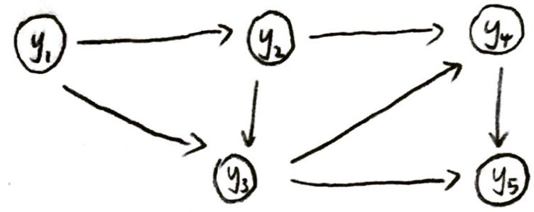
(e) $\alpha^{new} = \alpha^{old} - [H_F(\alpha^{old})]^{-1} \nabla F(\alpha^{old})$

$= \alpha^{old} - [Q + c\mathbb{1}\mathbb{1}^T]^{-1} \cdot \nabla F(\alpha^{old})$

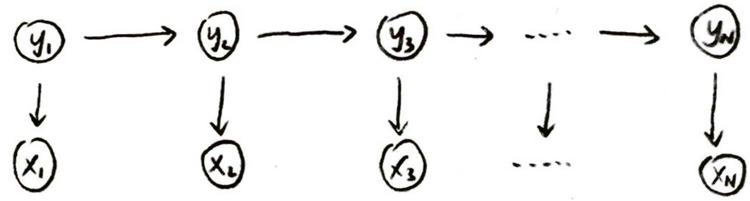
$= \alpha^{old} - \left(Q^{-1} - \frac{Q^{-1}c\mathbb{1}\mathbb{1}^T Q^{-1}}{1 + \mathbb{1}^T Q^{-1} c \mathbb{1}}\right)\nabla F(\alpha^{old})$

3. (a)

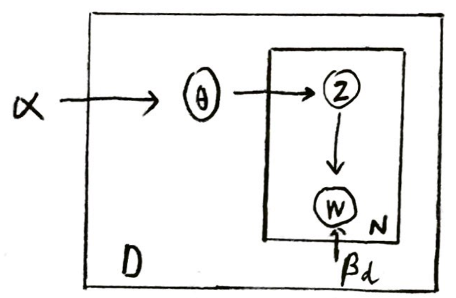
(i) $p(y_1, y_2, y_3, y_4, y_4, y_5) = p(y_1)p(y_2)\prod_{k=3}^{5} p(y_k \mid y_{k-1}, y_{k-2})$



(ii) $P(x_1, \cdots x_N, y_1, \cdots y_N) = P(y_1)\prod_{k=2}^{N} P(y_k \mid y_{k-1}) \prod_{k=1}^{N} P(x_k \mid y_k)$



(b)



(c) ①: $p(r, \theta, \phi, z) = \prod_{i=1}^{M} \left[p(r_i)\, p(\theta_i \mid r_i) \prod_{j=1}^{N} p(\phi_{ij})\, p(z_{ij} \mid \phi_{ij}) \right]$

②: $p(z, W) = \prod_{i=1}^{M} \left[p(z_i) \prod_{j=1}^{N} p(W_{ij} \mid z_i) \right]$

5. (a) $f(y|x;\theta) = \sum_{k=1}^{K} \pi_k \phi(y; w_k^T x + b_k, \sigma_k^2)$

$\Rightarrow f(\underline{y}|\underline{x};\theta) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \phi(y_i; w_k^T x_i + b_k, \sigma_k^2)$

$\therefore \log f(\underline{y}|\underline{x};\theta) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} \pi_k \phi(y_i; w_k^T x_i + b_k, \sigma_k^2)$

(b) $f(y|x, z=k;\theta) = \phi(y; w_k^T x + b_k, \sigma_k^2)$,

$f(y, z|x;\theta) = \pi_z \phi(y; w_z^T x + b_z, \sigma_z^2)$, Let $\delta_{ik} = 1_{\{z_i=k\}}$

Then $\log f(y, z|x;\theta) = \sum_{k=1}^{K} \sum_{i=1}^{N} 1_{\{z_i=k\}} \log \pi_{z_i} \phi(y_i; w_{z_i}^T x_i + b_{z_i}, \sigma_{z_i}^2)$

$:= \sum_{k=1}^{K} \sum_{i=1}^{N} \delta_{ik} \log \pi_k \phi(y_i; w_k^T x_i + b_k, \sigma_k^2)$

(c) $Q(\theta, \theta^{old}) = E_z[\log f(y, z|x;\theta) | y, x; \theta^{old}]$

Let $\gamma_{ik} = p(z_i=k | y_i, x_i; \theta^{old})$

$\gamma_{ik} = \frac{f(z_i, y_i | x_i; \theta^{old})}{f(y_i | x_i; \theta^{old})} = \frac{\pi_k^{old} \phi(y_i, w_k^{T\,old} x_i + b_k^{old}, \sigma_k^{old})}{\sum_{i=1}^{K} \pi_k^{old} \phi(y_i, w_k^{T\,old} x_i + b_k^{old}, \sigma_k^{old})}$

$Q(\theta, \theta^{old}) = \sum_{i=1}^{N} \sum_{k=1}^{K} E_q[1_{\{z_i=k\}} \log \pi_k \phi(y, w_k^T x + b_k, \sigma_k^2]$

$= \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik} \log \pi_k \phi(y, w_k^T x + b_k, \sigma_k^2)$

$= \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik} \log \pi_k + \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik} \log \phi(y, w_k^T x + b_k, \sigma_k^2)$

(d) $\theta^{new} = \arg\max_\theta Q(\theta, \theta^{old})$, s.t. $\sum_{k=1}^{K} \pi_k = 1$, $\forall \pi_k > 0$ for $k = 1, \cdots K$

let $f = Q(\theta, \theta^{old}) + \lambda(\sum_{k=1}^{K} \pi_k - 1) - \sum_{k=1}^{K} \alpha_k \pi_k$, $\alpha_k = 0$, $\forall k = 1, \cdots K$

$\frac{\partial f}{\partial \pi_k} = \sum_{i=1}^{N} \frac{\gamma_{ik}}{\pi_k} + \lambda = 0 \Rightarrow \pi_k = \frac{\sum_{i=1}^{N} \gamma_{ik}}{-\lambda}$

$\frac{\partial f}{\partial \lambda} = \sum_{k=1}^{K} \pi_k - 1 = 0 \Rightarrow \sum_{k=1}^{K} \sum_{i=1}^{N} \gamma_{ik} / -\lambda - 1 = 0 \Rightarrow \lambda = -N \Rightarrow \pi_k = \frac{\sum^{N} \gamma_{ik}}{K}$

$\frac{\partial f}{\partial w_k} = \sum_{i=1}^{N} \gamma_{ik}^2 \left(\frac{\log \phi(y, w_k^T x + b_k, \sigma_k^2)}{\partial w_k} \right) = \sum_{i=1}^{N} \gamma_{ik} \frac{\partial (\log \frac{1}{\sqrt{2\pi}\sigma_k} + \frac{y_i - (w_k^T x_i + b_k))^2}{2\sigma_k})}{\partial w_k}$

$= \sum_{i=1}^{N} \gamma_{ik} \frac{2(y_i - w_k^T x_i + b_k)}{2\sigma_k} (-x_i) = 0$

$\frac{\partial f}{\partial b_k} = \sum_{i=1}^{N} \gamma_{ik} \frac{\partial (\log \frac{1}{\sqrt{2\pi}\sigma_k} + \frac{(y_i - (w_k^T x_i + b_k))^2}{2\sigma_k})}{\partial b_k} = \sum_{i=1}^{N} \gamma_{ik} \frac{y_i - (w_k^T x_i + b_k)}{\sigma_k} \cdot (-1) = 0$

Set $\tilde{w}_k = \begin{bmatrix} w_k \\ b_k \end{bmatrix}$, $\tilde{x} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{bmatrix} = [x, 1]$, $y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$

then we can combine $\frac{\partial f}{\partial w_k}$ & $\frac{\partial f}{\partial b_k}$ to get $\sum_{i=1}^{N} \gamma_{ik} (y_i - \tilde{w}_k^T \tilde{x}_i) \tilde{x}_i = 0$

Use closed form to represent the whole question, we can set the objective function as

$\sum_{i=1}^{N} r_{ik}(y_i - \tilde{w}_k^T \tilde{x}_i)^2$ can be written as $(\tilde{X}\tilde{w}_k - y)^T R(\tilde{X}\tilde{w}_k - y)$

where $R = \text{diag}(r_{1k}, r_{2k}, \cdots r_{Nk})$

$$\therefore \sum_{i=1}^{N} r_{ik}(y_i - \tilde{w}_k^T \tilde{x}_i)\tilde{x}_i = 2X^T R X w - 2X^T R y = 0$$

$$\Rightarrow \tilde{w}_k = (X^T R X)^{-1} X^T R y$$

$$\frac{\partial f}{\partial \sigma_k^2} = \frac{\partial \sum_{i=1}^{N} r_{ik}\left(\log \frac{1}{\sqrt{2\pi\sigma_k^2}} + \left(-\frac{1}{2\sigma_k^2}(y_i - \tilde{w}_k^T \tilde{x}_i)^2\right)\right)}{\partial \sigma_k^2}$$

$$= \sum_{i=1}^{N} r_{ik} + \frac{1}{\sigma_k^2}\sum_{i=1}^{N}(y_i - \tilde{w}_k^T \tilde{x}_i)^2 r_{ik} = 0$$

$$\Rightarrow \sigma_k^2 = \frac{1}{\sum_{i=1}^{N} r_{ik}} \sum_{i=1}^{N} (y_i - \tilde{w}_k^T \tilde{x}_i)^2 r_{ik}$$

$$= \frac{1}{\sum_{i=1}^{N} r_{ik}} (\tilde{X}\tilde{w}_k - y)^T R(\tilde{X}\tilde{w}_k - y)$$