CS280 HW1

1. First we have three formula

(1)
$$\begin{cases} l(\theta) = \sum_{n=1}^{N} \log P(X_n|\theta) \\ P(X_n|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(X_n|\mu_k, \Sigma_k) \\ \mathcal{N}(X_n|\mu_k,\Sigma_k) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} \cdot \exp\{-\frac{1}{2}(X_n-\mu_k)^T \Sigma_k^{-1} (X_n-\mu_k)\} \end{cases}$$

and then we can derive the log-likelihood wrt μ_k.

$$\frac{\partial l(\theta)}{\partial \mu_k} = \sum_{n=1}^{N} \frac{1}{P(X_n|\theta)} \cdot \frac{\partial P(X_n|\theta)}{\partial \mu_k}$$ according to chain rule

$$= \sum_{n=1}^{N} \underbrace{\boxed{\frac{1}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(X_n|\mu_{k'},\Sigma_{k'})} \cdot \pi_k \cdot \mathcal{N}(X_n|\mu_k,\Sigma_k)}}_{= r_{nk}} \cdot (X_n-\mu_k) \cdot \Sigma_k^{-1}$$

$$= \sum_n r_{nk} \cdot \Sigma_k^{-1} \cdot (X_n-\mu_k)$$

(2) the process is same as (1) to derive π_k

① without constraints

$$\frac{\partial l(\theta)}{\partial \pi_k} = \sum_{n=1}^{N} \frac{1}{P(X_n|\theta)} \cdot \frac{\partial P(X_n|\theta)}{\partial \pi_k}$$ according to chain rule

$$= \sum_{n=1}^{N} \frac{1}{P(X_n|\theta)} \cdot \mathcal{N}(X_n|\mu_k,\Sigma_k)$$

$$= \sum_{n=1}^{N} r_{nk}$$

∴ finally, we get it

② with constraints

according to Lagrangian, we can have

$$L(\pi_k, \lambda) = l(\theta) + \lambda(1 - \sum_{k=1}^{K} \pi_k)$$

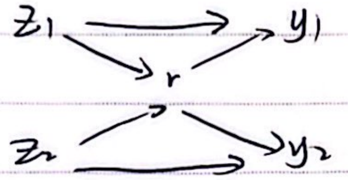
$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N} r_{nk} - \lambda = 0$$

$$\therefore \lambda = \frac{1}{K} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} = \frac{N}{K}$$, let $N = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk}$

$$\therefore \frac{\partial l(\theta)}{\partial \pi_k} = \sum_{n=1}^{N} r_{nk} - \frac{N}{K}$$

∴ we get it.

2. (a)



and assume we already have $\frac{\partial E}{\partial y_1}$ and $\frac{\partial E}{\partial y_2}$, depending on your definition

(b) when given $\overline{y_i}$, we can compute error easily, denoting it as $E(y_1, y_2)$

$$\frac{\partial E}{\partial r} = \frac{\partial E}{\partial y_1}\frac{\partial y_1}{\partial r} + \frac{\partial E}{\partial y_2}\frac{\partial y_2}{\partial r}, \quad \text{for } \frac{\partial y_i}{\partial r} = \frac{-e^{z_i}}{r^2}$$

$$= \frac{\partial E}{\partial y_1}\cdot\frac{-e^{z_1}}{r^2} + \frac{\partial E}{\partial y_2}\cdot\frac{-e^{z_2}}{r^2}$$

$$\frac{\partial E}{\partial z_1} = \frac{\partial E}{\partial y_1}\frac{\partial y_1}{\partial z_1} + \frac{\partial E}{\partial r}\frac{\partial r}{\partial z_1} \qquad \frac{\partial E}{\partial z_2} = \frac{\partial E}{\partial y_2}\frac{\partial y_2}{\partial z_2} + \frac{\partial E}{\partial r}\frac{\partial r}{\partial z_2}$$

$$= \frac{\partial E}{\partial y_1}\frac{e^{z_1}}{r} + \frac{\partial E}{\partial r}\cdot e^{z_1} \qquad = \frac{\partial E}{\partial y_2}\frac{e^{z_2}}{r} + \frac{\partial E}{\partial r}\cdot e^{z_2}$$

and then we can use $\frac{\partial E}{\partial z_1}, \frac{\partial E}{\partial z_2}$ to update $\overline{z_j}$, if $E(y_1, y_2)$ is determined.

Also, for more general case, this formula is also useful if $\frac{\partial E}{\partial y_j}$ is given

$$\begin{cases} \frac{\partial E}{\partial z_j} = \frac{\partial E}{\partial y_j}\frac{e^{z_j}}{r} + \frac{\partial E}{\partial r}\cdot e^{z_j} \\ \frac{\partial E}{\partial r} = \sum_i \frac{\partial E}{\partial y_i}\frac{-e^{z_i}}{r^2} \end{cases}$$

when given

$$\Longrightarrow \begin{cases} \frac{\partial E}{\partial z_i} = \overline{z_j} = \overline{y_j}\frac{e^{\overline{z_j}}}{r} + \overline{r}e^{\overline{z_j}} \\ \frac{\partial E}{\partial r} = \overline{r} = -\sum_j \overline{y_i}\frac{e^{\overline{z_i}}}{r^2} \end{cases}$$

(c)

```
def softmax_VJP(Z, Y_bar):
    R = np.sum(np.exp(Z), axis=1, keepdims=True)
    R_bar = -np.sum(Y_bar * np.exp(Z), axis=1, keepdims=True)
            /R**2
    Z_bar = Y_bar * (np.exp(Z)/R) + R_bar * np.exp(Z)
    return Z_bar
```