TASK 1 :

· Gradient : firstly rewrite $f(\beta) = \left[-\sum_{i=1}^{n} \sum_{k=0}^{K-1} \mathbb{1}\{y_i = k\} \log p(\vec{x}_i ; \vec{\beta}_k) + \frac{\lambda}{2} \sum_{k=0}^{K-1} \|\vec{\beta}_k\|_2^2 \right]$

$$\nabla_{\vec{\beta}_k} f(\beta) = - \sum_{i=1}^{n} \sum_{j=0}^{K-1} \mathbb{1}\{y_i = j\} \nabla_{\vec{\beta}_k}^{\log} p(\vec{x}_i ; \vec{\beta}_j) + \frac{\lambda}{2} \cdot 2 \vec{\beta}_k$$

Since $\nabla_{\vec{\beta}_k}^{\log} p(\vec{x}_i ; \vec{\beta}_k) \overset{\text{chain}}{\underset{\text{Rule}}{=\!=}} \nabla_{\vec{\beta}_k} (\vec{x}_i^T \vec{\beta}_k) \cdot \frac{\partial \log P}{\partial \vec{x}_i^T \vec{\beta}_k} \overset{C-R}{=} \nabla_{\vec{\beta}_k} (\vec{x}_i^T \vec{\beta}_k) \frac{\partial \exp(\vec{x}_i^T \vec{\beta}_k)}{\partial \vec{x}_i^T \vec{\beta}_k} \cdot \frac{\partial p}{\partial e^{\vec{x}_i^T \vec{\beta}_k}} \cdot \frac{\partial \log P}{\partial p}$

$$= \frac{\vec{x}_i \cdot \exp(\vec{x}_i^T \vec{\beta}_k)}{p(\vec{x}_i ; \vec{\beta}_k)} \frac{\sum_{\ell \neq k} e^{\vec{x}_i^T \vec{\beta}_\ell}}{(\sum_0^{K-1} e^{\vec{x}_i^T \vec{\beta}_\ell})^2} = \frac{\vec{x}_i \cdot p(\vec{x}_i ; \vec{\beta}_k) [1 - p(\vec{x}_i ; \vec{\beta}_k)]}{p(\vec{x}_i ; \vec{\beta}_k)}$$

when $j \neq k$ $\quad = \vec{x}_i \cdot [1 - p(\vec{x}_i ; \vec{\beta}_k)]$

Similarly $\nabla_{\vec{\beta}_k} \log p(\vec{x}_j ; \vec{\beta}_j) \overset{=\!=}{C-R} \vec{x}_i \exp(\vec{x}_i^T \vec{\beta}_k) \cdot \frac{1}{p} \left(- \frac{\exp(\vec{x}_i^T \vec{\beta}_j)}{(\sum_{\ell=0}^{K-1} e^{\vec{x}_i^T \vec{\beta}_\ell})^2} \right] = - \frac{\vec{x}_i e^{\vec{x}_i^T \vec{\beta}_k}}{\sum_{\ell=0}^{K-1} e^{\vec{x}_i^T \vec{\beta}_\ell}}$

Thus, $\sum_{j=0}^{K-1} \mathbb{1}\{y_i = j\} \nabla_{\vec{\beta}_k} \log p(\vec{x}_i ; \vec{\beta}_j) = \vec{x}_i \mathbb{1}\{y_i = k\} - \vec{x}_i \sum_j \mathbb{1}_{\dots} \cdot p \quad = - \vec{x}_i \cdot p(\vec{x}_i ; \vec{\beta}_k)$

$$= \vec{x}_i (\mathbb{1}\{y_i = k\} - p(\vec{x}_i ; \vec{\beta}_k))$$

$\therefore \nabla_{\vec{\beta}_k} f(\beta) = - \sum_i \vec{x}_i (\mathbb{1}(y_i = k) - p(\vec{x}_i ; \vec{\beta}_k)) + \lambda \vec{\beta}_k = X^T (P_k(\vec{x}_i) - \mathbb{1}\{\vec{Y} = k\}) + \lambda \vec{\beta}_k$

Hessian : With the results of $\nabla_{\vec{\beta}_k} f(\beta)$, we can directly derive $H_{\vec{\beta}_k} f(\beta) = \frac{\partial}{\partial \vec{\beta}_k^T} X^T ($ $P_k(\vec{x}_i) - \mathbb{1}\{\vec{Y} = k\}) + \lambda I$, with the former $= X^T W_k X$, since for the i^{th} row, $\frac{\partial}{\partial \vec{\beta}_k^T} p(\vec{x}_i ; \vec{\beta}_k) = \vec{x}_i^T \cdot p(\vec{x}_i ; \vec{\beta}_k) [1 - p(\vec{x}_i ; \vec{\beta}_k)]$. (Collect to get $W_k X$)

Therefore, $-\underset{\substack{\text{damped} \\ \text{rate}}}{\eta} [\lambda I + X^T W_k X]^{-1} [X^T (P_k(\vec{x}_i) - \mathbb{1}\{\vec{Y} = k\} + \lambda \vec{\beta}_k^{(t)}]$ is a damped Newton step.