Task 1: Prove that the formula above for β_k update indeed corresponds to damped Newton's method for minimizing $f(\beta)$ (e.g. derive gradient and Hessian, and plug in). Upload your derivations to this repository as a pdf file called **Derivations.pdf**

Proof:

Let

$$f(\beta) = \left[-\sum_{i=1}^{n} \left(\sum_{k=0}^{K-1} 1(y_i = k) \log p_k(x_i; \beta) \right) + \frac{\lambda}{2} \sum_{k=0}^{K-1} \sum_{j=1}^{p} \beta_{kj}^2 \right],$$

where $p_k(x_i; \beta) = \dfrac{e^{x_i^T \beta_k}}{\sum_{l=0}^{K-1} e^{x_i^T \beta_l}}$

be the objective function for β.

We want to show that

$$\beta_k^{(t+1)} = \beta_k^{(t)} - \eta (X^T W_k X + \lambda I)^{-1} \left[X^T (P_k - 1(Y = k)) + \lambda \beta_k^{(t)} \right],$$

$k = 0, \ldots, K-1$

(or β_k update)
corresponds to damped Newton's method for minimizing $f(\beta)$.

We first derive the gradient which we will denote as $\nabla_{\beta_k} f(\beta)$, or the derivative with respect to β_k.

We find that

$$\frac{\delta f(\beta)}{\delta \beta_k} = \frac{\delta}{\delta \beta_k}\left[-\sum_{i=1}^{n}\left(\sum_{m=0}^{K-1} 1(y_i=m)\log p_m(x_i;\beta)\right)+\frac{\lambda}{2}\beta_k^2\right]$$

$$= -\sum_{i=1}^{n}\left[\sum_{m=0}^{K-1}\left(1(y_i=m)\frac{\delta}{\delta\beta_k}\log p_m(x_i;\beta)\right)\right]+\lambda\beta_k$$

For $\frac{\delta}{\delta\beta_k}\log p_m(x_i;\beta)$, we work by cases.

Case 1: $m \neq k$

We find that

$$\frac{\delta}{\delta\beta_k}\log p_m(x_i;\beta)$$

$$= \frac{\delta}{\delta\beta_k}\log\left(e^{x_i^T\beta_m}\right)-\frac{\delta}{\delta\beta_k}\log\left(\sum_{l=0}^{K-1}e^{x_i^T\beta_l}\right) \qquad \left(\text{By log quotient rule}\right)$$

$$= -\frac{1}{\sum_{l=0}^{K-1}e^{x_i^T\beta_l}}\cdot x_i\cdot e^{x_i^T\beta_k} = -x_i\cdot p_k(x_i;\beta) \qquad \left(\text{By vector gradient rule}\right)$$

Case 2: $m = k$

We find that

$$\frac{\delta}{\delta\beta_k}\log p_{m=k}(x_i;\beta)$$

$$= \frac{\delta}{\delta\beta_k}\log\left(e^{x_i^T\beta}\right)-\frac{\delta}{\delta\beta_k}\log\left(\sum_{l=0}^{K-1}e^{x_i^T\beta_l}\right) \qquad \left(\text{By log quotient rule}\right)$$

$$= \frac{1}{e^{x_i^T \beta_k}} x_i \cdot e^{x_i^T \beta_k} - \frac{1}{\sum_{\ell=0}^{K-1} e^{x_i^T \beta_\ell}} x_i \cdot e^{x_i^T \beta_k} \quad \begin{pmatrix} \text{By vector} \\ \text{gradient} \\ \text{rule} \end{pmatrix}$$

$$= x_i \cdot (1 - p_k(x_i ; \beta))$$

Thus,

$\frac{\delta f(\beta)}{\delta \beta_k}$ can be written as

$$-\sum_{i=1}^{n} \left[x_i (1 - p_k(x_i ; \beta)) \cdot 1(y_i = k) + \sum_{m \neq k} x_i (-p_k(x_i ; \beta)) 1(y_i = m) \right] + \lambda \beta_k$$

$$= -\sum_{i=1}^{n} \left[x_i 1(y_i = k) - x_i 1(y_i = k) p_k(x_i ; \beta) - \sum_{m \neq k} x_i p_k(x_i ; \beta) 1(y_i = m) \right] + \lambda \beta_k$$

$$= -\sum_{i=1}^{n} \left[x_i 1(y_i = k) - \left(x_i 1(y_i = k) p_k(x_i ; \beta) + \sum_{m \neq k} x_i p_k(x_i ; \beta) 1(y_i = m) \right) \right] + \lambda \beta_k$$

$$= -\sum_{i=1}^{n} \left[x_i 1(y_i = k) - x_i p_k(x_i ; \beta) \left(1(y_i = k) + \sum_{m \neq k} 1(y_i = m) \right) \right] + \lambda \beta_k$$

$$= -\sum_{i=1}^{n} \left[x_i 1(y_i = k) - x_i p_k(x_i ; \beta) \left(\sum_{m=0}^{K-1} 1(y_i = m) \right) \right] + \lambda \beta_k$$

$$= -\sum_{i=1}^{n} \left[x_i 1(y_i = k) - x_i p_k(x_i ; \beta) \cdot 1 \right] + \lambda \beta_k$$

$$= -\sum_{i=1}^{n} \left[x_i \left(1(y_i = k) - p_k(x_i ; \beta) \right) \right] + \lambda \beta_k$$

We can then define the gradient as the compact form of $\frac{\delta f(\beta)}{\delta \beta_k}$:

$$\nabla f(\beta) = -X^T\left(1(Y=k) - P_k\right) + \lambda \beta_k$$
$$= X^T\left(P_k - 1(Y=k)\right) + \lambda \beta_k$$

We now derive the Hessian:

Recall

$$\nabla f(\beta) = X^T\left(P_k - 1(Y=k)\right) + \lambda \beta_k$$

We will then define the Hessian to be the derivative of $\nabla f(\beta)$ with respect to β_k.

i.e. $\frac{\delta^2 f(\beta)}{\delta \beta_k \delta \beta_k}$

So,

$$\frac{\delta^2 f(\beta)}{\delta \beta_k \delta \beta_k} = \frac{\delta}{\delta \beta_k} X^T\left(P_k - 1(Y=k)\right) + \lambda \beta_k$$

This can be re-written as the following:

$$\frac{\delta^2 f(\beta)}{\delta\beta_k\,\delta\beta_k} = \frac{\delta}{\delta\beta_k} - \sum_{i=1}^{n}\left[x_i\left(1(y_i=k)-p_k(x_i\,;\beta)\right)\right] + \lambda\beta_k$$

$$= -\sum_{i=1}^{n}\left[x_i\left(-\frac{\delta}{\delta\beta_k}p_k(x_i\,;\beta)\right)\right] + \lambda\frac{\delta}{\delta\beta_k}\beta_k$$

(since $1(y_i=k)$ does not depend on β_k)

We find that

$$\frac{\delta}{\delta\beta_k}p_k(x_i\,;\beta) = \frac{\delta}{\delta\beta_k}\frac{e^{x_i^T\beta_k}}{\sum_{l=0}^{k-1}e^{x_i^T\beta_l}}$$

$$= \left(\frac{x_i e^{x_i^T\beta_k}\sum_{l=0}^{k-1}e^{x_i^T\beta_l} - e^{x_i^T\beta_k}\cdot e^{x_i^T\beta_k}\cdot x_i}{\left(\sum_{l=0}^{k-1}e^{x_i^T\beta_l}\right)^2}\right)$$

(By quotient rule and by rules of vector gradients)

$$= \left(\frac{x_i e^{x_i^T\beta_k}\left(\sum_{l=0}^{k-1}e^{x_i^T\beta_l} - e^{x_i^T\beta_k}\right)}{\left(\sum_{l=0}^{k-1}e^{x_i^T\beta_l}\right)\left(\sum_{l=0}^{k-1}e^{x_i^T\beta_l}\right)}\right)$$

$$= \left(\frac{x_i\,e^{x_i^T\beta_k}}{\left(\sum_{l=0}^{k-1}e^{x_i^T\beta_l}\right)}\cdot\left(\frac{\sum_{l=0}^{k-1}e^{x_i^T\beta_l}}{\left(\sum_{l=0}^{k-1}e^{x_i^T\beta_l}\right)} - \frac{e^{x_i^T\beta_k}}{\left(\sum_{l=0}^{k-1}e^{x_i^T\beta_l}\right)}\right)\right)$$

$$= x_i\cdot p_k(x_i\,;\beta)\left(1-p_k(x_i\,;\beta)\right)$$

Thus, the Hessian can be written as

$$\frac{\delta^2 f(\beta)}{\delta\beta_k \, \delta\beta_k}$$

$$= -\sum_{i=1}^{n}\left[x_i \left(-x_i \cdot p_k(x_i;\beta)\right)\left(1 - p_k(x_i;\beta)\right)\right] + \lambda I_k$$

$$= \sum_{i=1}^{n}\left[x_i \left(p_k(x_i;\beta)\right)\left(1 - p_k(x_i;\beta)\right) x_i \right] + \lambda I_k$$

$$= \sum_{i=1}^{n}\left[x_i \, w_i \, x_i \right] + \lambda I_k \qquad \left(\begin{array}{c}\text{By definition of} \\ W_k \text{ in slides}\end{array}\right)$$

We can then define the Hessian as the compact form of $\frac{\delta^2 f(\beta)}{\delta\beta_k \, \delta\beta_k}$:

$$\nabla^2 f(\beta) = X^T W X + \lambda I$$

By the definition of Newton's method for minimizing the objective function (over β), we find that

$$\beta^{(t+1)} = \beta^{(t)} - \eta \left\{ \nabla^2 f(\beta^{(t)}) \right\}^{-1} \nabla f(\beta^{(t)})$$

$$= \beta^{(t)} - \eta\left(X^T W_k X + \lambda I\right)^{-1}\left[X^T\left(P_k - 1(Y=k)\right) + \lambda \beta_k^{(t)}\right],$$

$k = 0, \ldots, K-1$

and where η is a parameter denoting the learning rate or "step size" in optimization determining behavior of gradient descent.

Therefore, we find that the formula for β_k update corresponds to damped Newton's method for minimizing $f(\beta)$. $\square$