We wish to show:

$$\underbrace{\beta_K^{(t+1)}}_{p\times 1} = \underbrace{\beta_K^{(t)}}_{p\times 1} - \eta \underbrace{\left(\underbrace{X^T}_{p\times n}\underbrace{W_K}_{n\times n}\underbrace{X}_{n\times p} + \underbrace{\lambda I}_{p\times p}\right)^{-1}}_{p\times p}\underbrace{\left[\underbrace{X^T}_{p\times n}\left\{\underbrace{P_K}_{n\times 1} - \underbrace{\mathbb{1}_{\{y=K\}}}_{n\times 1}\right\} + \lambda \underbrace{\beta_K^t}_{p\times 1}\right]}_{p\times 1}, \quad K = 0, \dots, K-1$$

is given by

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

When $f(\beta) = \left[-\sum_{i=1}^{n}\sum_{K=0}^{K-1}\mathbb{1}_{\{y_i=K\}}\log(p_K(x_i;\beta)) + \frac{\lambda}{2}\sum_{K=0}^{K-1}\sum_{j=1}^{p}\beta_{K,j}^2\right]$, where $p_K(x_i;\beta) = \frac{\exp(x_i^T\beta_K)}{\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)}$

and W_K is a diagonal matrix with elements $(W_K)_{ii} = p_K(x_i;\beta)(1-p_K(x_i;\beta))$ and P_K is an $\mathbb{R}^n$ vector with $(P_K)_i = p_K(x_i;\beta)$

$$\log\left(\frac{\exp(x_i^T\beta_K)}{\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)}\right) = \left(\log(\exp(x_i^T\beta_K)) - \log\left(\textstyle\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)\right)\right) = x_i^T\beta_K - \log\left(\textstyle\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)\right)$$

So

$$\begin{aligned} f(\beta) &= \left[-\sum_{i=1}^{n}\sum_{K=0}^{K-1}\mathbb{1}_{\{y_i=K\}}\left(x_i^T\beta_K - \log\left(\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)\right)\right) + \lambda/2\sum_{K=0}^{K-1}\sum_{j=1}^{p}\beta_{K,j}^2\right] \\ &= \left[-\sum_{i=1}^{n}\sum_{K=0}^{K-1}\mathbb{1}_{\{y_i=K\}}(x_i^T\beta_K) + \sum_{i=1}^{n}\sum_{K=0}^{K-1}\mathbb{1}_{\{y_i=K\}}\log\left(\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)\right) + \lambda/2\sum_{K=0}^{K-1}\sum_{j=1}^{p}\beta_{K,j}^2\right] \\ &= \left[-\sum_{i=1}^{n}\sum_{K=0}^{K-1}\mathbb{1}_{\{y_i=K\}}(x_i^T\beta_K) + \sum_{i=1}^{n}\log\left(\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)\right) + \lambda/2\sum_{K=0}^{K-1}\sum_{j=1}^{p}\beta_{K,j}^2\right] \end{aligned}$$

Since only one term of "$\sum_{K=0}^{K-1}\mathbb{1}_{\{y_i=K\}}\log\left(\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)\right)$" will be non-zero.

I will need the following calculations to take the gradient of $f(\beta)$ above

(a) $$\nabla_{\beta_j}\left(\mathbb{1}_{\{y_i=K\}}(x_i^T\beta_K)\right) = \begin{cases} x_i & \text{if } j = K \\ 0 & \text{if } j \neq K \end{cases} = \mathbb{1}_{j=K}\cdot x_i$$

This holds since if $j=K$, $\nabla_{\beta_j}\left(\mathbb{1}_{\{y_i=K\}}x_i^T\beta_K\right) = \nabla_{\beta_K}(x_i^T\beta_K) = x_i$, and $\nabla_{\beta_j}\left(\mathbb{1}_{\{y_i=K\}}x_i^T\beta_K\right) = \nabla_{\beta_j}(0) = 0$ if $j\neq K$

Moreover

$$\nabla_{\beta_j}\log\left(\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)\right) = \left(\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)\right)^{-1}\nabla_{\beta_j}\left(\textstyle\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)\right)$$

Since for $f:\mathbb{R}\to\mathbb{R}$ and $g:\mathbb{R}^n\to\mathbb{R}$, $\nabla_t f(g(t)) = f'(g(t))\nabla_t g(t)$ (*)

However $\nabla_{\beta_j}\left(\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)\right) = \sum_{\ell=0}^{K-1}\nabla_{\beta_j}\exp(x_i^T\beta_\ell)$ where $\nabla_{\beta_j}\exp(x_i^T\beta_\ell) = \exp(x_i^T\beta_\ell)\nabla_{\beta_j}(x_i^T\beta_\ell)$ by applying (*) again

Hence $\nabla_{\beta_j}\exp(x_i^T\beta_\ell) = \begin{cases} x_i\exp(x_i^T\beta_j) & \text{if } j=\ell \\ 0 & \text{if } j\neq\ell \end{cases} = \mathbb{1}_{\{j=\ell\}}\cdot x_i\exp(x_i^T\beta_j)$

So

(b) $$\nabla_{\beta_j}\log\left(\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)\right) = \left(\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)\right)^{-1}\sum_{\ell=0}^{K-1}\mathbb{1}_{\{j=\ell\}}\cdot x_i\exp(x_i^T\beta_j) = \frac{\exp(x_i^T\beta_j)}{\left(\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)\right)}\cdot x_i = x_i\,p_j(x_i,\beta)$$

Finally $\nabla_{\beta_j}\left(\frac{\lambda}{2}\sum_{k=0}^{K-1}\sum_{i=1}^{p}\beta_{k,i}^2\right) = \nabla_{\beta_j}\left(\lambda/2\sum_{k=0}^{K-1}\|\beta_k\|^2\right) = \lambda/2\sum_{k=0}^{K-1}\nabla_{\beta_j}\|\beta_k\|^2$

However $\nabla_{\beta_j}\|\beta_k\|^2 = \begin{cases} 2\beta_j & \text{if } j=k \\ 0 & \text{if } j\neq k \end{cases}$

So

(c) $\nabla_{\beta_j}\left(\lambda/2\sum_{k=0}^{K-1}\sum_{i=1}^{p}\beta_{k,i}^2\right) = \lambda/2\sum_{k=0}^{K-1}\mathbb{1}_{\{j=k\}}(2\cdot\beta_k) = \lambda/2\cdot 2\beta_j = \lambda\beta_j$.

Using these results to calculate $\nabla_{\beta_j} f(\beta)$ for some fixed $j=0,\dots,K-1$ we have (a), (b) and (c) giving us

$$\nabla_{\beta_j}(f(\beta)) = \nabla_{\beta_j}\left[-\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{1}_{\{y_i=k\}}(x_i^T\beta_k) + \sum_{i=1}^{n}\log\left(\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)\right) + \lambda/2\sum_{k=0}^{K-1}\sum_{j=1}^{p}\beta_{k,j}^2\right]$$

$$= -\sum_{i=1}^{n}\sum_{k=0}^{K-1}\nabla_{\beta_j}\left(\mathbb{1}_{\{y_i=k\}}(x_i^T\beta_k)\right) + \sum_{i=1}^{n}\nabla_{\beta_j}\left[\log\left(\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)\right)\right] + \nabla_{\beta_j}\left[\sum_{k=0}^{K-1}\sum_{j=1}^{p}\beta_{k,j}^2\right]$$

$$= -\sum_{i=1}^{n}\sum_{k=0}^{K-1}x_i\cdot\mathbb{1}_{\{j=k\}} + \sum_{i=1}^{n}x_i\,p_j(x_i,\beta) + \lambda\beta_j$$

$$= -\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}x_i\,p_j(x_i,\beta) + \lambda\beta_j$$

$$= \sum_{i=1}^{n}x_i(p_j(x_i,\beta)-1) + \lambda\beta_j$$

$$= X^T\left(p_j - \mathbb{1}_{\{y=j\}}\right) + \lambda\beta_j \quad , \text{ for every } j=0,\dots,K-1$$

Now to calculate the Hessian of f denoted by $\nabla^2_{\beta_m} f(\beta)$ we observe that

$$\nabla^2_{\beta_m} f(\beta) = \frac{\partial^2}{\partial\beta_m\partial\beta_m} f(\beta) = \nabla_{\beta_m}\left[\nabla_{\beta_m} f(\beta)\right]$$

where we've shown $\nabla_{\beta_m} f(\beta) = \sum_{i=1}^{n}x_i(p_m(x_i,\beta)-1) + \lambda\beta_m$

It remains to calculate $\nabla_{\beta_m}\left[\sum_{i=1}^{n}x_i(p_m(x_i,\beta)-1) + \lambda\beta_m\right]$

I'll need the following calculations

$$\nabla_{\beta_j}\left[x_i(p_m(x_i,\beta)-1)\right] = \nabla_{\beta_j}\left[x_i\,p_m(x_i,\beta) - x_i\right] = x_i\,\nabla_{\beta_j}p_m(x_i,\beta)$$

However $\nabla_{\beta_j}p_m(x_i,\beta) = \nabla_{\beta_j}\left[\frac{\exp(x_i^T\beta_m)}{\sum_{k=0}^{K-1}\exp(x_i^T\beta_k)}\right] = \frac{\left(\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)\right)\nabla_{\beta_j}\left[\exp(x_i^T\beta_m)\right] - \exp(x_i^T\beta_m)\nabla_{\beta_j}\left[\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)\right]}{\left(\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)\right)^2}$

we've shown $\nabla_{\beta_j}\left[\exp(x_i^T\beta_m)\right] = \mathbb{1}_{\{j=k\}}\cdot x_i\exp(x_i^T\beta_j)$

and we've shown $\nabla_{\beta_j}\left[\sum_{\ell=0}^{K-1}\exp(x_i^T\beta_\ell)\right] = \sum_{\ell=0}^{K-1}\mathbb{1}_{\{\ell=j\}}\cdot x_i\exp(x_i^T\beta_j) = x_i\exp(x_i^T\beta_j)$

So $\nabla_{\beta_j} P_K(x_i,\beta) = \dfrac{\left(\sum_{c=0}^{K-1} \exp(x_i^T\beta_c)\right)\left(\mathbb{1}_{\{j=K\}} \cdot x_i \exp(x_i^T\beta_j)\right) - \exp(x_i^T\beta_K)\, x_i \exp(x_i^T\beta_j)}{\left(\sum_{c=0}^{K-1} \exp(x_i^T\beta_c)\right)^2}$

$$= \frac{\mathbb{1}_{\{j=K\}} x_i \exp(x_i^T\beta_j)}{\sum_{c=0}^{K-1} \exp(x_i^T\beta_c)} - \frac{\exp(x_i^T\beta_K)}{\left(\sum_{c=0}^{K-1} \exp(x_i^T\beta_c)\right)} \cdot \frac{x_i \exp(x_i^T\beta_j)}{\left(\sum_{c=0}^{K-1} \exp(x_i^T\beta_c)\right)}$$

$$= \mathbb{1}_{\{j=K\}} x_i P_j(x_i,\beta) - P_K(x_i,\beta)\left(x_i P_j(x_i,\beta)\right)$$

We consider only the case when $j=K$ since we're calculating $\nabla_{\beta_K}[\nabla_{\beta_K} f(\beta)]$

So we have $\nabla_{\beta_K}\left[\nabla_{\beta_m} P_K(x_i,\beta)\right] = x_i P_K(x_i,\beta) - x_i P_K^2(x_i,\beta) = x_i P_K(x_i,\beta)(1-P_K(x_i,\beta))$

So $\nabla_{\beta_K}\left[\nabla_{\beta_K} f(\beta)\right] = \sum_{i=1}^{n} \nabla_{\beta_K}\left[x_i(P_K(x_i,\beta) - 1)\right] + \nabla_{\beta_K}\left[\lambda\beta_K\right]$

$$= \sum_{i=1}^{n} x_i \left(P_K(x_i,\beta)\right)\left(1 - P_K(x_i,\beta)\right) x_i + \lambda$$

$$= X^T W_K X + \lambda I$$

Where $(W_K)_{ii} = P_K(x_i,\beta)(1-P_K(x_i,\beta))$ is a diagonal matrix

Hence by the damped Newton's Method:

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

$$= \beta_K^{(t)} - \eta\left[X^T W_K X + \lambda I\right]^{-1}\left(X^T\left(P_K - \mathbb{1}_{\{y=K\}}\right) + \lambda\beta_K\right)$$