Jacobian

$$\frac{\partial f}{\partial \beta_{kj}}\left(-\sum_{i=1}^{n}\left[\sum_{m=0}^{K-1} \mathbb{1}(y_i=m)\log p_m(x_i,\beta)\right]+\frac{\lambda}{2}\sum\sum \beta_{kj}^2\right)$$

$$=\frac{\partial f}{\partial \beta_{kj}}\left(-\sum_{i=1}^{n}\left[\left(\sum_{m\neq k}^{K-1} \mathbb{1}(y_i=m)\log p_m(x_i,\beta)\right)+\mathbb{1}(y_i=k)\log p_k(x_i,\beta)\right)\right)+\lambda\beta_{kj}$$

$$=-\sum_{i=1}^{n}\left[\frac{\partial}{\partial \beta_{kj}}\left(\mathbb{1}_{(y_i\neq k)}\log p_{y_i}(x_i;\beta)\right)+\frac{\partial}{\partial \beta_{kj}}\left(\mathbb{1}_{(y_i=k)}\log p_k(x_i;\beta)\right)\right]+\lambda\beta_{kj}$$

(1) (2)

$$(1) := \mathbb{1}_{(y_i\neq k)}\cdot\frac{1}{p_{y_i}(x_i,\beta)}\cdot\frac{\partial}{\partial\beta_{kj}}\left(\frac{e^{x_i^T\beta_{y_i}}}{\sum_l e^{x_i^T\beta_l}}\right)$$

Note: $=0$ if $y_i=k$, $\therefore$ assume $y_i\neq k$

$$=\mathbb{1}_{(y_i\neq k)}\underbrace{\frac{e^{x_i^T\beta_{y_i}}}{p_{y_i}(x_i,\beta)}\cdot(-1)\left(\sum_l e^{x_i^T\beta_l}\right)^{-2}}_{=(-1)\left(\sum_l e^{x_i^T\beta_l}\right)^{-1}}\underbrace{\frac{\partial}{\partial\beta_{kj}}\left(\sum_l e^{x_i^T\beta_l}\right)}_{=x_{ij}e^{x_i^T\beta_k}}$$

$$=\mathbb{1}_{(y_i\neq k)}\,x_{ij}\,p_k(x_i;\beta)\cdot(-1)$$

$$② \quad \frac{\partial}{\partial B_{kj}} \left(\mathbb{1}_{(y_i=k)} \log p_k(x_i ; B) \right)$$

$$= \mathbb{1}_{(y_i=k)} \left(p_k(x_i ; B) \right)^{-1} \underbrace{\frac{\partial}{\partial B_{kj}} \left(\frac{e^{x_i^T B_k}}{\sum_l e^{x_i^T B_l}} \right)}_{③}$$

$$③ = \frac{1}{\sum_l e^{x_i^T B_l}} \frac{\partial}{\partial B_{kj}} \left(e^{x_i^T B_k} \right) + e^{x_i^T B_k} (-1) \left(\sum_l e^{x_i^T B_l} \right)^{-2} \frac{\partial}{\partial B_{kj}} \left(\sum_l e^{x_i^T B_l} \right)$$

$$= x_{ij} \frac{e^{x_i^T B_k}}{\sum_l e^{x_i^T B_l}} + (-1)\, p_k(x_i , B) \cdot x_{ij} \cdot \left(\sum_l e^{x_i^T B_l} \right)^{-1} e^{x_i^T B_k}$$

$$= x_{ij} \left(p_k(x_i ; B) - p_k(x_i ; B)^2 \right)$$

$$\therefore ② = \mathbb{1}_{(y_i=k)} \left(p_k(x_i ; B) \right)^{-1} x_{ij} \; p_k(x_i ; B) \cdot \left(1 - p_k(x_i ; B) \right)$$

$$= x_{ij} \mathbb{1}_{(y_i=k)} \left(1 - p_k(x_i ; B) \right)$$

Continued

$$① + ② = (-1)\mathbb{1}_{(y_i \neq k)} x_{ij} p_k(x_i;B) + \mathbb{1}_{(y_i=k)} x_{ij}(1 - p_k(x_i;B))$$

$$= x_{ij}\left[(-1)\mathbb{1}_{(y_i \neq k)} p_k(x_i;B) + \mathbb{1}_{(y_i=k)} + (-1)\mathbb{1}_{(y_i=k)} p_k(x_i;B)\right]$$

$$= x_{ij}\left((-1)p_k(x_i;B) + \mathbb{1}_{(y_i=k)}\right)$$

Finally, we have:

$$\frac{\partial f}{\partial B_{kj}} = -\sum_{i=1}^{n}\left[x_{ij}(-1 \cdot p_k(x_i;B) + \mathbb{1}_{(y_i=k)}\right] + \lambda B_{kj}$$

$$= \sum_{i=1}^{n}\left[x_{ij} p_k(x_i;B) - \mathbb{1}_{(y_i=k)}\right] + \lambda B_{kj}$$

$$\therefore \nabla B_k = X^T\left\{P_k - \mathbb{1}_{(Y=k)}\right\} + \lambda B_k$$

Given: $\nabla_{\beta_k} = X^T\{P_k - \mathbb{1}_{(Y=k)}\} + \lambda \beta_k$

Column j of Hessian^T $= \frac{\partial}{\partial \beta_{kj}}\left(\nabla_{\beta_k}\right)$

$$= \frac{\partial}{\partial \beta_{kj}}\left[X^T P_k - X^T \mathbb{1}_{(Y=k)} + \lambda \beta_k\right]$$

$$= \underbrace{\frac{\partial}{\partial \beta_{kj}}\left(X^T P_k\right)}_{} - 0 + \lambda e_j$$

$$= X^T \frac{\partial}{\partial \beta_{kj}}\left(P_k\right) = X^T \frac{\partial}{\partial \beta_{kj}} \begin{bmatrix} \frac{e^{x_1^T \beta_k}}{\sum_l e^{x_1^T \beta_l}} \\ \vdots \\ \frac{e^{x_n^T \beta_k}}{\sum_l e^{x_n^T \beta_l}} \end{bmatrix}$$

Continued on
next page

$$\frac{\partial}{\partial \beta_{kj}}\left(\frac{e^{x_i^T\beta_k}}{\sum_l e^{x_i^T\beta_l}}\right)$$

$$= \underbrace{\frac{\partial}{\partial \beta_{kj}}\left(e^{x_i^T\beta_k}\right)\cdot\frac{1}{\sum_l e^{x_i^T\beta_l}}}_{①} + \underbrace{\frac{\partial}{\partial \beta_{kj}}\left(\sum_l e^{x_i^T\beta_l}\right)^{-1}\cdot e^{x_i^T\beta_k}}_{②}$$

$$① = x_{ij}e^{x_i^T\beta_k}\cdot\frac{1}{\sum_l e^{x_i^T\beta_l}}$$

$$= x_{ij}\, p_k(x_i;\beta)$$

$$② = (-1)\left(\sum_l e^{x_i^T\beta_l}\right)^{-2}\frac{\partial}{\partial \beta_{kj}}\left(\sum_l e^{x_i^T\beta_l}\right)\cdot e^{x_i^T\beta_k}$$

$$= (-1)\left(\sum_l e^{x_i^T\beta_l}\right)^{-2} x_{ij}\left(e^{x_i^T\beta_k}\right)^2$$

$$= (-1)\, x_{ij}\, p_k(x_i;\beta)^2$$

$$①+② = x_{ij}\, p_k(x_i;\beta) - x_{ij}\, p_k(x_i;\beta)^2$$

$$= x_{ij}\, p_k(x_i;\beta)\left(1-p_k(x_i;\beta)\right)^2$$

Finally, we have column j of Hessian^T is equal to:

$$X^T\left\{\operatorname{diag}\left(p_k(x_i;B)\left(1-p_k(x_i;B)\right)\right)X\right\} + \lambda e_j$$

And therefore Hessian^T is equal to

$$X^T \operatorname{diag}\left(p_k(x_i;B)\left(1-p_k(x_i;B)\right)\right)X + \lambda I_p$$

Note that Hessian^T is the sum of symmetric matrices, and is therefore symmetric.

And if we let $W = \operatorname{diag}\left(p_k(x_i;B)\left(1-p_k(x_i;B)\right)\right)$

Then Hessian $= X^T W X + \lambda I_p$