Derivation

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

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

We will calculate the gradient and Hessian of this $f(\beta)$.

Gradient

$\frac{\lambda}{2}\cdot 2\beta_k$

We will take derivative with respect to β_k (∇_{β_k}): $\downarrow$

$$\nabla_{\beta_k} f(\beta) = -\sum_{i=1}^{n}\left\{\sum_{t=0}^{K-1} \mathbb{1}(y_i=t)\cdot\nabla_{\beta_k}\log p_t(x_i;\beta)\right\} + \lambda\cdot\beta_k$$

and for $$\nabla_{\beta_k}\log p_t(x_i;\beta) = \nabla_{\beta_k}\log\left(\frac{e^{x_i^T\beta_k}}{\sum_{l=0}^{K-1} e^{x_i^T\beta_l}}\right) =$$

If $t=k$:

Derivative of numerator of $p_k;\beta$ wrt β_k: $e^{x_i^T\beta_k}\cdot x_i$

Derivative of denominator of $p_k:\beta$ wrt β_k ($k=l$): $e^{x_i^T\beta_k}\cdot x_i$

By Quotient rule: $$\nabla_{\beta_k} p_k(x_i;\beta) = \frac{\left(\sum_{l=0}^{K-1} e^{x_i^T\beta_l}\right)\left(e^{x_i^T\beta_k}\cdot x_i\right) - \left(e^{x_i^T\beta_k}\cdot e^{x_i^T\beta_k}\cdot x_i\right)}{\left(\sum_{l=0}^{K-1} e^{x_i^T\beta_l}\right)^2}$$

$$= \frac{\left(e^{x_i^T\beta_k}\, x_i\right)}{\underbrace{\left(\sum_{l=0}^{K-1} e^{x_i^T\beta_l}\right)}_{p_k(x_i;\beta)}} \cdot \underbrace{\boxed{\frac{\left(\sum_{l=0}^{K-1} e^{x_i^T\beta_l} - e^{x_i^T\beta_k}\right)}{\left(\sum_{l=0}^{K-1} e^{x_i^T\beta_l}\right)}}}_{1-p_k(x_i;\beta)}$$

$$\nabla_{\beta_k} p_k(x_i;\beta) = p_k(x_i;\beta)\cdot x_i\cdot(1-p_k(x_i;\beta))$$

If $t \neq k$:

Derivative of numerator of p_k wrt β_k: $\frac{d}{d\beta_k} e^{x_i^T \beta_t} = 0$ (no dependence on k)

Derivative of denominator of p_k wrt β_k $(k=l)$: $e^{x_i^T \beta_k} \cdot x_i$

By Quotient rule: $\nabla_{\beta_k} p_k(x_i;\beta) = \dfrac{\left(\sum_{l=0}^{K-1} e^{x_i^T \beta_l}\right) \cdot 0 - e^{x_i^T \beta_t} \cdot e^{x_i^T \beta_k} \cdot x_i}{\left(\sum_{l=0}^{K-1} e^{x_i^T \beta_l}\right)^2}$

$$\nabla_{\beta_k} p_k(x_i;\beta) = -p_t(x_i;\beta) \cdot p_k(x_i;\beta) \cdot x_i$$

we will substitute back into $\nabla_{\beta_k} f(\beta)$:

$$\nabla_{\beta_k} f(\beta) = -\sum_{i=1}^{n} \left\{ \sum_{t=0}^{K-1} \mathbb{1}_{(y_i=t)} \cdot \nabla_{\beta_k} \log p_t(x_i;\beta) \right\} + \lambda \cdot \beta_k$$

$$= -\sum_{i=1}^{n} \left\{ \sum_{t=0}^{K-1} \mathbb{1}_{(y_i=t)} \cdot \frac{\nabla_{\beta_k} p_t(x_i;\beta)}{p_t(x_i;\beta)} \right\} + \lambda \cdot \beta_k$$

$$= -\sum_{i=1}^{n} \left\{ \sum_{t=0}^{K-1} \mathbb{1}_{(y_i=t)} \left[\underbrace{\frac{\cancel{p_k(x_i;\beta)} \cdot x_i \cdot (1-p_k(x_i;\beta))}{\cancel{p_k(x_i;\beta)}}}_{t=k} + \underbrace{\frac{-x_i \cdot \cancel{p_t(x_i;\beta)} \cdot p_k(x_i;\beta)}{\cancel{p_t(x_i;\beta)}}}_{t \neq k} \right] \right\}$$

$$+ \lambda \cdot \beta_k$$

$$= -\sum_{i=1}^{n} \left\{ \underbrace{\mathbb{1}_{(y_i=k)} x_i (1-p_k(x_i;\beta))}_{t=k} - \underbrace{\sum_{t \neq k}^{K-1} \mathbb{1}_{(y_i=t)} x_i \cdot p_k(x_i;\beta)}_{t \neq k} \right\} + \lambda \cdot \beta_k$$

$$= -\sum_{i=1}^{n} \left\{ \mathbb{1}_{(y_i=k)} x_i - \left(\mathbb{1}_{(y_i=k)} p_k(x_i;\beta) \cdot x_i + \sum_{t \neq k}^{K-1} \mathbb{1}_{(y_i=t)} \cdot x_i \cdot p_k(x_i;\beta) \right) \right\} + \lambda \cdot \beta_k$$

$$= -\sum_{i=1}^{n} \left\{ \mathbb{1}_{(y_i=k)} x_i - \left(\boxed{p_k(x_i;\beta) \cdot x_i} \underbrace{\left(\mathbb{1}_{(y_i=k)} + \sum_{t \neq k}^{K-1} \mathbb{1}_{(y_i=t)} \right)}_{1} \right) \right\} + \lambda \cdot \beta_k$$

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

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

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

$\mathbb{1}(Y=k)$ is 1 if $Y=k$ (class), 0 if $Y \neq k$ (class), P_k is vector for each $p_k(x_i;\beta)$, X is design matrix, $\lambda \beta_k$ is regularization term

Hessian

To derive the Hessian we need to differentiate the gradient $\nabla_{\beta_k} f(\beta)$

$$\nabla_{\beta_k} f(\beta) = -\sum_{i=1}^{n} \left\{ x_i \left(\mathbb{1}_{(y_i=k)} - P_k(x_i;\beta) \right) \right\} + \lambda \beta_k$$

Since $\mathbb{1}_{(y_i=k)}$ does not depend on β_k, derivative $\to 0$.
So we focus on $x_i \cdot P_k(x_i;\beta)$ for the two cases when $\underset{(\text{same class})}{t=k}$ and $\underset{(\text{different class})}{t \neq k}$.

$t=k$ case:

$$\frac{d^2 P_k(x_i;\beta)}{d\beta_k^2} = \frac{d}{d\beta_k} \nabla_{\beta_k} f(\beta) = x_i \cdot P_k(x_i;\beta)(1-P_k(x_i;\beta))x_i' + P_k(x_i;\beta)\cdot 0 + \lambda$$

$$= P_k(x_i;\beta)(1-P_k(x_i;\beta)) \cdot x_i \cdot x_i' + \lambda$$

$t \neq k$ case:

$$\frac{d^2 P_k(x_i;\beta)}{d\beta_k^2} = \frac{d}{d\beta_k} \nabla_{\beta_k} f(\beta) = x_i \cdot -P_t(x_i;\beta) \cdot P_k(x_i;\beta) \cdot x_i' + P_k(x_i;\beta) \cdot 0$$

$$= -P_t(x_i;\beta) \cdot P_k(x_i;\beta)\, x_i \cdot x_i'$$

Now summing over all samples:

$$\text{Hessian}(\beta) = \begin{cases} \sum_{i=1}^{n} \underbrace{P_k(x_i;\beta) \cdot (1-P_k(x_i;\beta))}_{W_{kk}} \cdot x_i x_i' + \lambda I, & \text{if } t=k \leftarrow \text{(diagonal entries of Hessian matrix)} \\ \sum_{i=1}^{n} -P_t(x_i;\beta) \cdot P_k(x_i;\beta)\, x_i x_i', & \text{if } t \neq k \leftarrow \text{(off-diagonal entries of Hessian matrix)} \end{cases}$$

So for diagonal entries ($t=k$) we can write Hessian(β) as

$$H_{kk} = X^T W_k X + \lambda I, \quad W_{kii} = P_k(x_i;\beta)(1-P_k(x_i;\beta)) \text{ as given}$$

Now let us plug in Gradient and Hessian into Damped Newton's update.

Damped Newton's Update

$$\beta_k^{(t+1)} = \beta_k^{(t)} - \eta \left(H_{kk} \right)^{-1} \left[\nabla_{\beta_k} f(\beta) \right], \quad k = 0, \ldots, K-1$$

Substituting in:

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

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

Final Result