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

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

$$-\sum_{i=1}^{n} \sum_{k=\emptyset}^{K-1} 1(y_i = k) \log p_k(x_i; \beta)$$

Gradient

$$-\sum_{i=1}^{n} 1(y_i = k) \log p_k(x_i; \beta)$$

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

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

$$\frac{\partial}{\partial \beta_k} (x_i^T \beta_k) = x_i$$

$$\frac{\partial}{\partial \beta_k} \log \left(\sum_{l=\emptyset}^{K-1} e^{x_i^T \beta_l} \right) = \frac{1}{\sum_{l=\emptyset}^{K-1} e^{x_i^T \beta_l}} \cdot \frac{\partial}{\partial \beta_k} \sum_{l=\emptyset}^{K-1} e^{x_i^T \beta_l}$$

$$\frac{\partial}{\partial \beta_k} \sum_{l=\emptyset}^{K-1} e^{x_i^T \beta_l} = e^{x_i^T \beta_k} \cdot x_i$$

So:

$$\frac{\partial}{\partial \beta_k} \log \left(\sum_{l=\emptyset}^{K-1} e^{x_i^T \beta_l} \right) = \frac{e^{x_i^T \beta_k}}{\sum_{l=\emptyset}^{K-1} e^{x_i^T \beta_l}} \cdot x_i$$

$$\frac{\partial}{\partial \beta_{lk}} \log\left(\sum_{l=0}^{K-1} e^{x_i^T \beta_l}\right) = \frac{e^{x_i^T \beta_k}}{\sum_{l=0}^{K-1} e^{x_i^T \beta_l}} x_i = p_k(x_i;\beta) \cdot x_i$$

$$\frac{\partial}{\partial \beta_{lk}} \left(-1(y_i = k) \log p_k(x_i;\beta)\right) = -1(y_i = k) x_i + p_{lk}(x_i;\beta) x_i$$

Sum over all i:

$$\nabla_{\beta_k} L(\beta) = -\sum_{i=1}^{n} (1(y_i = k) - p_{lk}(x_i;\beta)) x_i$$

Regularization term

$$\frac{\lambda}{2} \sum_{k=0}^{K-1} \sum_{j=1}^{p} \beta_{kj}^2$$

$$\frac{\partial}{\partial \beta_k} \left(\frac{\lambda}{2} \sum_{k=0}^{K-1} \sum_{j=1}^{p} \beta_{kj}^2\right) = \frac{\lambda}{2} \cdot 2 \sum_{j=1}^{p} \beta_{kj} = \lambda \sum_{j=1}^{p} \beta_{kj}$$

$$= \lambda \beta_k$$

Total Gradient:

$$\nabla_{\beta_k} L(\beta) = -\sum_{i=1}^{n} (1(y_i = k) - p_k(x_i;\beta)) x_i$$

$$1(y = k) = \begin{bmatrix} 1(y_1 = k) \\ \vdots \\ 1(y_n = k) \end{bmatrix} \quad p_k = \begin{bmatrix} p_k(x_1;\beta) \\ \vdots \\ p_k(x_n;\beta) \end{bmatrix}$$

$$\nabla_{\beta_K} L(\beta) = \sum_{i=1}^{n} x_i \left(p_{iK}(x_i;\beta) - 1(y_i = K) \right)$$

$$\nabla_{\beta_K} L(\beta) = X^T \left(p_K - 1(Y = K) \right)$$

Add regularization term:

$$\nabla_{\beta_K} f(\beta) = X^T \left(p_K - 1(Y = K) \right) + \lambda \beta_K$$

Hessian Calculation

$$\nabla_{\beta_K} L(\beta) = -\sum_{i=1}^{n} \left(1(y_i = K) - p_K(x_i;\beta) \right) x_i$$

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

Apply quotient rule:

$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{g(x) f'(x) - f(x) g'(x)}{g(x)^2}$$

$f(x) = e^{x_i^T \beta_K}$ $\quad$ $g(x) = \sum_{l=0}^{K-1} e^{x_i^T \beta_l}$

$f'(x) = e^{x_i^T \beta_K} x_i$ $\quad$ $g'(x) = e^{x_i^T \beta_K} x_i$

$$\frac{\partial}{\partial \beta_K} p_K(x_i;\beta) = \frac{\left(\sum_{l=0}^{K-1} e^{x_i^T \beta_l} \cdot e^{x_i^T \beta_K} x_i \right) - e^{x_i^T \beta_K} \cdot e^{x_i^T \beta_K} x_i}{\left(\sum_{l=0}^{K-1} e^{x_i^T \beta_l} \right)^2}$$

$$\frac{\partial}{\partial \beta_K} p_K(x_i;\beta) = \frac{e^{x_i^T \beta_K} x_i \left(\sum_{l=\emptyset}^{K-1} e^{x_i^T \beta_l} - e^{x_i^T \beta_K} \right)}{\left(\sum_{l=\emptyset}^{K-1} e^{x_i^T \beta_l} \right)^2}$$

$$\frac{\partial}{\partial \beta_K} p_K(x_i;\beta) = p_K(x_i;\beta)\left(1 - p_K(x_i;\beta)\right) x_i$$

Sub into simplified product rule expression:

$$\frac{\partial}{\partial \beta_K} \left(p_K(x_i;\beta) x_i \right) = \left(p_K(x_i;\beta)\left(1 - p_K(x_i;\beta)\right) x_i \right) x_i$$

$$\frac{\partial}{\partial \beta_K} \left(p_K(x_i;\beta) x_i \right) = p_K(x_i;\beta)\left(1 - p_K(x_i;\beta)\right) x_i x_i^T$$

Hessian: $\sum_{i=1}^{n} p_K(x_i;\beta)\left(1 - p_K(x_i;\beta)\right) x_i x_i^T$

$x_i \in \mathbb{R}^p$ is a feature vector

$X \in \mathbb{R}^{n \times p}$ is the matrix of input features where each row corresponds to feature vector x_i^T

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}$$

Weights $p_K(x_i;\beta)\left(1 - p_K(x_i;\beta)\right)$ apply to each outer product individually. Use diagonal matrix $W_K \in \mathbb{R}^{n \times n}$ so:

$$W_K = \begin{bmatrix} p_K(x_1;\beta)(1-p_K(x_1;\beta)) & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & p_K(x_n;\beta)(1-p_K(x_n;\beta)) \end{bmatrix}$$

Hessian Regularization Term

$$\frac{\partial}{\partial \beta_K}\left(\frac{\lambda}{2}\sum_{j=1}^{p}\beta_{Kj}^2\right) = \lambda\beta_K \Rightarrow \frac{d}{d\beta_K}\lambda\beta_K = \lambda I$$

Thus Hessian matrix:

$$H_{\beta_K} = X^T W_K X + \lambda I$$

Newton's Update Formula

$$\beta_K^{(t+1)} = \beta_K^{t} - \eta H_{\beta_K}^{-1} \nabla_{\beta_K} f(\beta)$$

$$\beta_K^{(t+1)} = \beta_K^{t} - \eta\left(X^T W_K X + \lambda I\right)^{-1}\left(X^T(p_K - 1(Y=K)) + \lambda\beta_K^{(t)}\right)$$

Case 1 : $t = K$ (when $y_i = K$)

For $t = K$, the indicator function $1(y_i = K) = 1$:

$$\nabla_{\beta_K} f(\beta) = X^T(p_K - 1) + \lambda\beta_K$$

Apply Newton's Update:

$$\beta_K^{(t+1)} = \beta_K^{(t)} - \eta\underbrace{\left(X^T W_K X + \lambda I\right)^{-1}}_{\text{Hessian, capturing the}}\underbrace{\left(X^T(p_K - 1) + \lambda\beta_K^{(t)}\right)}$$

curvature ... function ... gradient ...

Update accounts for both the predicted probabilities & the true labels

$$\nabla_{\beta_K} f(\beta) = X^T (P_K - 1) + \lambda \beta_K$$

Case 2: $t \neq K$ (when $y_i \neq K$)

For $t \neq K$, indicator function $\mathbb{1}(y_i = K) = 0$

The gradient for this case simplifies to

$$\nabla_{\beta_K} f(\beta) = X^T P_K + \lambda \beta_K$$

Newtons update:

$$\beta_K^{(t+1)} = \beta_K^{(t)} - \eta (X^T W_K X + \lambda I)^{-1} (X^T P_K + \lambda \beta_K^{(t)})$$

Update is simpler involving only the predicted probabalities as

$$\nabla_{\beta_K} f(\beta) = X^T P_K + \lambda \beta_K$$