1) Single Output, single sample

$$\hat{y} = w^T x + b \quad , L = (\hat{y} - y)^2$$

$$\frac{dL}{dw} = \frac{dL}{d\hat{y}} \cdot \frac{d\hat{y}}{dw} = 2(\hat{y} - y)x$$

$$\frac{dL}{db} = \frac{dL}{d\hat{y}} \cdot \frac{d\hat{y}}{db} = 2(\hat{y} - y)$$

2) Single Output, Many Samples, N samples

$$MSE: L = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

$$\frac{dL}{dw} = \frac{d}{dw}\left(\frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2\right) = \frac{1}{N} \sum_{i=1}^{N} \frac{d}{dw} (\hat{y}_i - y_i)^2 = \frac{1}{N} \sum_{i=1}^{N} 2(\hat{y}_i - y_i)x$$

$$= \frac{2}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i) x_i$$

$$\frac{dL}{db} = \frac{2}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)$$

3) Many Outputs

N: # of samples

M: # of input features

D: # of outputs

$X_{N,M}$ $W_{M,D}$ $\hat{y}_{N,D}$

For a sample i and output j, $\hat{y}_{i,j} = \sum_{k=1}^{M} X_{i,k} W_{k,j} + b_j$

$$\text{MSE}: L = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{D} (\hat{y}_{i,j} - y_{i,j})^2$$

$$\frac{dL}{W_{k,j}} = \frac{d}{dW_{k,j}} \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{D} (\hat{y}_{i,j} - y_{i,j})^2 \right) = \frac{1}{N} \sum_{i=1}^{N} \frac{d}{dW_{k,j}} (\hat{y}_{i,j} - y_{i,j})^2$$

$$\frac{d}{dW_{kj}} (\hat{y}_{ij} - y_{ij})^2 = 2(\hat{y}_{ij} - y_{ij}) \frac{d\hat{y}_{ij}}{dW_{kj}} = 2(\hat{y}_{ij} - y_{ij}) X_{ik}$$

$$\frac{dL}{W_{kj}} = \frac{1}{N} \sum_{i=1}^{N} 2(\hat{y}_{ij} - y_{ij}) X_{ik} = 2/N \sum_{i=1}^{N} (\hat{y}_{ij} - y_{ij}) X_{iK}$$

$$\underset{(M \times D)}{\frac{dL}{W}} = 2/N \underset{(M \times N)}{X^T} \cdot (\underset{(N \times D)}{\hat{y}} - \underset{(N \times D)}{y})$$

$$\frac{dL}{b_j} = \frac{1}{N} \sum_{i=1}^{N} 2(\hat{y}_{ij} - y_{ij}) \cdot \frac{dy_{i,j}}{db_j} = \frac{2}{N} \sum_{i=1}^{N} \underset{(N \times D)}{(\hat{y}_{ij} - y_{ij})}$$

$$\underset{(1 \times D)}{\frac{dL}{b}} = \frac{2}{N} \underset{(1 \times N)}{\text{ALL-ONES}} * \underset{(N \times D)}{(\hat{y} - y)}$$

$1_{1 \times N}$