Rules of Differentiation

1) Addition Rule

$$\frac{\partial (f(x) + g(x))}{\partial x} = \frac{\partial f(x)}{\partial x} + \frac{\partial g(x)}{\partial x}$$

$$\Rightarrow \frac{\partial}{\partial x} \sum_{i}^{n} f_i(x) = \sum_{i}^{n} \frac{\partial f_i(x)}{\partial x}$$

2) Product Rule

$$\frac{\partial}{\partial x}\left[f(x) \cdot g(x)\right] = f(x)\frac{\partial g(x)}{\partial x} + g(x)\frac{\partial f(x)}{\partial x}$$

3) Chain Rule

$$\frac{\partial}{\partial x}\left[f(g(x))\right] = \frac{\partial f(g(x))}{\partial g(x)} \cdot \frac{\partial g(x)}{\partial x}$$

$$\underbrace{\mathcal{L}(\omega) = NLL(\omega)}_{\text{this is just notation}} = \sum_{i=1}^{n} \log\left(1+\exp(\omega^T z_i)\right) - y_i \omega^T z_i, \text{ where}$$

$z_i \in \mathbb{R}^{(p+1)\times 1}$ $\quad i \in \{1, 2, \dots, n\}$

$y_i \in \{0, 1\}$

Use $\underline{\omega}$ and ω interchangeably to mean the weight vector. ω_j and ω_k are scalars.

$\rightarrow \underline{\omega} = \omega \in \mathbb{R}^{(p+1)\times 1}$

1) derive the gradient vector $\nabla_{\underline{\omega}} L$. Note that $\nabla_{\omega} L$ is another way of writing $\frac{\partial L}{\partial \underline{\omega}}$.

→ what we want is a vector that looks like:

$$\frac{\partial L}{\partial \underline{\omega}} = \begin{bmatrix} \frac{\partial L}{\partial \omega_1} \\ \vdots \\ \frac{\partial L}{\partial \omega_j} \\ \vdots \\ \frac{\partial L}{\partial \omega_{p+1}} \end{bmatrix} \in \mathbb{R}^{(p+1)\times 1}. \quad \text{Here, } j \in \{1, 2, \dots, p+1\}$$

→ That is, we want a column vector that has the same dimensions as ω_i, where each element is the partial derivative of our loss function $L(\omega)$, with respect to each component of $\underline{\omega}$.

→ A start to do this is to look at what the partial derivative w.r.t each component looks like.

So we want to evaluate $\frac{\partial L}{\partial \omega_j}$

* Note that we are using a different index 'j' here, as opposed to 'i'. This is because 'j' iterates over each component of $\underline{w}$, while 'i' iterates over each data point.

$$\rightarrow \frac{\partial L}{\partial w_j} = \frac{\partial}{\partial w_j}\left[\sum_{i=1}^{n} \log(1+\exp(w^T z_i)) - y_i w^T z_i\right]$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial w_j}\left[\log(1+\exp(w^T z_i)) - y_i w^T z_i\right] \quad \text{Addition Rule}$$

$$① = \sum_{i=1}^{n} \frac{\partial}{\partial w_j}\left[\log(1+\exp(w^T z_i))\right] - \frac{\partial}{\partial w_j}\left[y_i w^T z_i\right] \quad \text{Addition Rule}$$

$\rightarrow$ let's look at each derivative individually.

First suppose $f(w) = w^T z_i$,

$$g(w) = \exp(w^T z_i) = \exp(f(w)) = g(f)$$

$$h(w) = 1 + \exp(w^T z_i) = 1 + g(w) = h(g) = h(g(f))$$

$$m(w) = \log(1+\exp(w^T z_i) = \log(h(w)) = \log(h(g(f)))$$

These mean that:

$$\frac{\partial}{\partial w_j}\left[\log(1+\exp(w^T z_i))\right] = \frac{\partial}{\partial w_j} m(w) = \frac{\partial}{\partial w_j} m(h(g(f)))$$

$$\frac{\partial}{\partial w_j} m(h(g(f))) = \frac{\partial m}{\partial h}\frac{\partial h}{\partial g}\frac{\partial g}{\partial f}\frac{\partial f}{\partial w_j}$$ Chain Rule

$$\frac{\partial m}{\partial h} = \frac{\partial \log(h(w))}{h(w)} = \frac{1}{h(w)} = \frac{1}{1+\exp(w^T z_i)}$$

$$\frac{\partial h}{\partial g} = \frac{\partial\, 1+g(w)}{\partial g(w)} = 1$$

$$\frac{\partial g}{\partial f} = \frac{\partial \exp(f)}{\partial f(w)} = \exp(f) = \exp(w^T z_i)$$

$$\frac{\partial f}{\partial w_j} = \frac{\partial w^T z_i}{\partial w_j} = z_i^j$$

* Note that z_i^j is the jth component of the ith data point. We get this because :-

1

$$\mathbb{R} \ni w^T z_i = \begin{bmatrix} w_1 & w_2 & \cdots & w_j & \cdots & w_{p+1} \end{bmatrix} \begin{bmatrix} z_i^1 \\ z_i^2 \\ \vdots \\ z_i^j \\ \vdots \\ z_i^{p+1} \end{bmatrix} = w_1 z_i^1 + w_2 z_i^2 + \cdots + w_j z_i^j + \cdots + w_{p+1} z_i^{p+1}$$

$$\Rightarrow \frac{\partial w^T z_i}{\partial w_j} = \frac{\partial (w_1 z + w_2 z_i^2 \cdots w_j z_i^j + \cdots + w_{p+1} z_i^{p+1})}{\partial w_j}$$

$$= \frac{\partial w_1 z_i^1}{\partial w_j} + \frac{\partial w_2 z_i^2}{\partial w_j} + \cdots + \frac{\partial w_j z_i^j}{\partial w_j} + \cdots + \frac{\partial w_{p+1} z_i^{p+1}}{\partial w_j}$$ Addition Rule

$$\frac{\partial w_{k\neq j} z_i^k}{\partial w_j} = 0$$, since $w_{k\neq j} z_i$ is not a function of w_j

$$\frac{\partial w_j z_i^j}{\partial w_j} = z_i^j \Rightarrow \frac{\partial w^T z_i}{\partial w_j} = z_i^j$$

$$\Rightarrow \frac{\partial}{\partial w_j}\left[\log\left(1+\exp(w^T z_i)\right)\right] = \frac{\exp(w^T z_i)}{1+\exp(w^T z_i)} z_i^j$$

similarly, we can differentiate the last term of ①,

$$\frac{\partial}{\partial w_j}\left(y_i w^T z_i\right) = y_i \frac{\partial}{\partial w_j}\left(w^T z_i\right),$$ since y_i isn't a function of w_j

$$= y_i z_i^j \quad \because ①$$

$\Rightarrow$ Back to ①

$$\Rightarrow \frac{\partial}{\partial w_j} L(w) = \sum_{i=1}^{n}\left[\frac{\partial}{\partial w_j}\left(\log\left(1+\exp(w^T z_i)\right)\right) - \frac{\partial}{\partial w_j}\left(y_i w^T z_i\right)\right]$$

$$= \sum_{i=1}^{n}\left[\frac{\exp(w^T z_i)}{1+\exp(w^T z_i)} z_i^j - y_i z_i^j\right]$$

$$= \sum_{i=1}^{n}\left[\frac{\exp w^T z_i}{1+\exp(w^T z_i)} - y_i\right] z_i^j$$

since from the problem sheet, $\mu_i = \frac{\exp(w^T z_i)}{1+\exp(w^T z_i)}$

$$\Rightarrow \frac{\partial L(w)}{\partial w_j} = \sum_{i=1}^{n}(\mu_i - y_i) z_i^j \longrightarrow ②$$

This gives us the partial derivative of $L(w)$ w.r.t each component of w_j. Thus, we can fill in our gradient vector $\nabla_{\underline{w}} L(w)$ as:

$$\frac{\partial L(w)}{\partial \underline{w}} = \begin{bmatrix} \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_j} \\ \vdots \\ \frac{\partial L}{\partial w_{p+1}} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} (\mu_i - y_i) z_i^1 \\ \vdots \\ \sum_{i=1}^{n} (\mu_i - y_i) z_i^j \\ \vdots \\ \sum_{i=1}^{n} (\mu_i - y_i) z_i^{p+1} \end{bmatrix}$$

$$= \begin{bmatrix} (\mu_1 - y_1) z_1^1 + \dots + (\mu_n - y_n) z_n^1 \\ \vdots \\ (\mu_1 - y_1) z_1^j + \dots + (\mu_n - y_n) z_n^j \\ \vdots \\ (\mu_1 - y_1) z_1^{p+1} + \dots + (\mu_n - y_n) z_n^{p+1} \end{bmatrix} = \begin{bmatrix} (\mu_1 - y_1) z_1^1 \\ \vdots \\ (\mu_1 - y_1) z_1^j \\ \vdots \\ (\mu_1 - y_1) z_1^{p+1} \end{bmatrix} + \dots + \begin{bmatrix} (\mu_n - y_n) z_n^1 \\ \vdots \\ (\mu_n - y_n) z_n^j \\ \vdots \\ (\mu_n - y_n) z_n^{p+1} \end{bmatrix}$$

$$= \sum_i^n (\mu_i - y_i) \begin{bmatrix} z_i^1 \\ \vdots \\ z_i^j \\ \vdots \\ z_i^{p+1} \end{bmatrix} = \sum_i^n (\mu_i - y_i) z_i,$$

remember, this is a vector, $z_i \in \mathbb{R}^{(p+1)\times 1}$

$$\therefore \boxed{\frac{\partial L(w)}{\partial \underline{w}} = \sum_{i=1}^{n} (\mu_i - y_i) z_i},$$

as in the notes.

Notice that the solution comes to this form because we wanted:

a) a column vector for the output. i.e. $\frac{\partial L}{\partial \underline{w}} \in \mathbb{R}^{(p+1)\times 1}$

b) we wrote z_i as a column vector. i.e $z_i \in \mathbb{R}^{(p+1)\times 1}$

This is different from the way that we write Z, the data matrix, since each data point z_i is a <u>row</u> vector. So the matrix derivative that is invoked in the answer sheet,

$$\frac{\partial(\exp(w^T z_i))}{\partial \underline{w}} = \exp(w^T z_i)\, z_i \text{ , is a}$$

result of both a) and b).

2) Find the Hessian $H(w)$ of $L(w)$.

→ As before, lets look at what the Hessian is supposed to look like.

$H(w)$ is supposed to contain all possible combinations of the second partial derivatives of a function w.r.t the parameters. Since is a $\mathbb{R}$-valued function (that is, $L(w) \in \mathbb{R}$), there are $(p+1)^2$ different derivatives.

This can be nicely summarised as a matrix.

In particular, we choose to write this as:-

$$H(w) = \begin{bmatrix} \frac{\partial L}{\partial w_1 \partial w_1} & \cdots & \frac{\partial L}{\partial w_1 \partial w_j} & \cdots & \frac{\partial L}{\partial w_1 \partial w_{p+1}} \\ \vdots & \ddots & & & \vdots \\ \frac{\partial L}{\partial w_k \partial w_1} & & \frac{\partial L}{\partial w_k \partial w_j} & & \vdots \\ \vdots & & & \ddots & \vdots \\ \frac{\partial L}{\partial w_{p+1} \partial w_1} & & & & \frac{\partial L}{\partial w_{p+1} \partial w_{p+1}} \end{bmatrix}.$$

* Note that, if we say:-

$\underline{f}(x) \in \mathbb{R}^{p+1}$, i.e is a vector valued function, and

$x \in \mathbb{R}$, i.e a scalar, and

$$\frac{\partial \underline{f}(x)}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x} \\ \vdots \\ \frac{\partial f_j}{\partial x} \\ \vdots \\ \frac{\partial f_{n+1}}{\partial x} \end{bmatrix},$$ for each component of $\underline{f}(x)$

Then we can see that if we substitute $\underline{f}(x) = \nabla_w L(w)$, and $x = w_k$,

$$\frac{\partial (\nabla_w L(w))}{\partial w_k} = \begin{bmatrix} \frac{\partial \nabla_{w_1} L(w)}{\partial w_k} \\ \vdots \\ \frac{\partial \nabla_{w_j} L(w)}{\partial w_k} \\ \vdots \\ \frac{\partial \nabla_{w_{p+1}} L(w)}{\partial w_k} \end{bmatrix} = \begin{bmatrix} \frac{\partial L}{\partial w_k \partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_k \partial w_j} \\ \vdots \\ \frac{\partial L}{\partial w_k \partial w_{p+1}} \end{bmatrix}$$

From the matrix of $H(w)$, we can see that this column vector looks like the kth row of $H(w)$.

So we can say that the kth row is:

$$H_k(w) = \left(\frac{\partial \nabla_w L(w)}{\partial k}\right)^T, = \frac{\partial [\nabla_w L(w)]^T}{\partial k}.$$

As such, this shows why the definition of the $H(w)$ is

$$H(\underline{w}) = \frac{\partial (\nabla_w L(w))^T}{\partial \underline{w}}.$$

→ As before, we can look at an individual component of $H(w)$.

$$\frac{\partial L(w)}{\partial w_k \partial w_j} = \frac{\partial}{\partial w_k}\left(\frac{\partial L(w)}{\partial w_j}\right) = \frac{\partial}{\partial w_k}\left(\sum_i^n (\mu_i - y_i) z_i^j\right)$$

from (2)

$$\frac{\partial}{\partial w_k}\left(\sum_{i=1}^n (\mu_i - y_i) z_i^j\right) = \sum_{i=1}^n \frac{\partial}{\partial w_k}(\mu_i z_i^j) - \frac{\partial}{\partial w_k}(y_i z_i^j) \rightarrow (3)$$

Addition Rule

→ looking at each partial derivative individually:

$$\frac{\partial}{\partial w_k}(y_i z_i^j) = 0$$ since neither y_i nor w_i^j depend on w_k.

$$\frac{\partial}{\partial w_k}(\mu_i z_i^j) = z_i^j \frac{\partial}{\partial w_k}(\mu_i)$$ since $\frac{\partial\, a f(x)}{\partial x} = a \frac{\partial f(x)}{\partial x}$

$$\frac{\partial}{\partial w_k}(\mu_i) = \frac{\partial}{\partial w_k}\left[\frac{\exp(w^T z_i)}{1 + \exp(w^T z_i)}\right]$$

$$= \frac{\partial}{\partial w_k}\left[\exp(w^T z_i)(1 + \exp(w^T z_i))^{-1}\right]$$

$$= \exp(w^T z_i) \frac{\partial}{\partial w_k}\left[\left(1+\exp(w^T z_i)\right)^{-1}\right] + \left(\exp(w^T z_i)+1\right)^{-1} \frac{\partial \exp w^T z}{\partial w_k}$$

Product Rule.

$$= \underbrace{\frac{\exp(w^T z_i) \times (-1) \exp(w^T z_i)}{\left(1+\exp(w^T z_i)\right)^2}}_{\text{use chain rule and Addition rule.}} z_i^k$$

this last term is the same as $\frac{\partial}{\partial w_j}\left(1+\exp(w^T z_i)\right)$ from before, but with w_k instead of w_j.

$$+ \left(\exp(w^T z_i)+1\right)^{-1} \underbrace{\exp(w^T z_i)}_{\text{used chain rule}} z_i^k$$

again, it's the same as $\frac{\partial \exp(w^T z)}{\partial w_j}$ from before, but with w_k, not w_j

$$= -\left[\frac{\exp(w^T z_i)}{1+\exp(w^T z_i)}\right]^2 z_i^k + \frac{\exp(w^T z_i)}{1+\exp(w^T z_i)} z_i^k$$

$$= -N_i^2 z_i^k + N_i z_i^k$$

$\rightarrow$ substituting the definition of $N_i = \frac{\exp(w^T z_i)}{1+\exp(w^T z_i)}$

$$= N_i(1-N_i) z_i^k$$

$$\Rightarrow z_i^j \frac{\partial N_i}{\partial w_k} = z_i^j N_i (1-N_i) z_i^k$$

$$= z_i^k N_i (1-N_i) z_i^j$$

because all values are scalar.

The reason to do this will become clear shortly.

Plugging back into 3,

$$\frac{\partial}{\partial w_k}\left(\sum_{i=1}^{n}(\mu_i - y_i) z_i^j\right) = \sum_{i=n}^{n} z_i^k \mu_i (1-\mu_i) z_i^j$$

$$\Rightarrow \frac{\partial L(w)}{\partial w_k \partial w_j} = \sum_{i=1}^{n} z_i^k \mu_i (1-\mu_i) z_i^j$$

→ As before, we can start to fill in the full matrix. Lets look at what a row looks like.

$$\frac{\partial L}{\partial w_k \partial \underline{w}} = \left[\frac{\partial L}{\partial w_k \partial w_1} \cdots \frac{\partial L}{\partial w_k \partial w_j} \cdots \frac{\partial L}{\partial w_k \partial w_{p+1}}\right]$$

Similarly to when we were finding the gradient, we can see that

$$\frac{\partial L}{\partial w_k \partial \underline{w}} = \left[\sum_{i=1}^{n} z_i^k \mu_i (1-\mu_i) z_i^1 \cdots \sum_{i=1}^{n} z_i^k \mu_i (1-\mu_i) z_i^j \cdots \sum_{i=1}^{n} z_i^k \mu_i (1-\mu_i) z_i^{p+1}\right]$$

$$= \sum_{i=1}^{n} z_i^k \mu_i (1-\mu_i) \left[z_i^1 \cdots z_i^j \cdots z_i^{p+1}\right]$$

$$= \sum_{i=1}^{n} z_i^k \mu_i (1-\mu_i) z_i^T, \text{ since } z_i \in \mathbb{R}^{(p+1)\times 1}$$

→ then, the full matrix is

$$\frac{\partial L}{\partial \underline{w}\,\partial \underline{w}} = \begin{bmatrix} \frac{\partial L}{\partial w_1 \partial \underline{w}} \\ \vdots \\ \frac{\partial L}{\partial w_k \partial \underline{w}} \\ \vdots \\ \frac{\partial L}{\partial w_{p+1} \partial \underline{w}} \end{bmatrix}$$

→ Follow a similar argument, to say that:

$$\frac{\partial L}{\partial \underline{w}\,\partial \underline{w}} = H(w) = \sum_{i=1}^{n} z_i \, N_i (1 - N_i) \, z_i^T$$

not transposed because we stacked $\frac{\partial L}{\partial w_k \partial \underline{w}}$ column-wise.

$$= \sum_{i=1}^{n} N_i (1 - N_i) \, z_i z_i^T$$

since $N_i(1-N_i) \in \mathbb{R}$.

→

$$z_i z_i^T = \begin{bmatrix} z_i^1 \\ \vdots \\ z_i^{p+1} \end{bmatrix} \begin{bmatrix} z_i^1 & \cdots & z_i^{p+1} \end{bmatrix}$$

since $z_i \in \mathbb{R}^{(p+1)\times 1}$,

$$= \begin{bmatrix} z_i^1 z_i^1 & \cdots & z_i^1 z_i^{p+1} \\ \vdots & \ddots & \vdots \\ z_i^{p+1} z_i^1 & & z_i^{p+1} z_i^{p+1} \end{bmatrix}$$

$$\Rightarrow \sum_{i=1}^{n} \mu_i (1-\mu_i) z_i z_i^T = \sum_{i=1}^{n} \mu_i (1-\mu_i) \begin{bmatrix} z_i^1 z_i^1 & \cdots & z_i^1 z_i^{p+1} \\ \vdots & \ddots & \vdots \\ z_i^{p+1} z_i^1 & & z_i^{p+1} z_i^{p+1} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{n} \mu_i (1-\mu_i) z_i^1 z_i^1 & \cdots & \sum_{i=1}^{n} \mu_i (1-\mu_i) z_i^1 z_i^{p+1} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} \mu_i (1-\mu_i) z_i^{p+1} z_i^1 & \cdots & \sum_{i=1}^{n} \mu_i (1-\mu_i) z_i^{p+1} z_i^{p+1} \end{bmatrix}$$

$$= \begin{bmatrix} z_1^1 & \cdots & z_n^1 \\ \vdots & \ddots & \\ z_1^{p+1} & \cdots & z_n^{p+1} \end{bmatrix} \begin{bmatrix} \mu_1(1-\mu_1) & & 0 \\ & \ddots & \\ 0 & & \mu_n(1-\mu_n) \end{bmatrix} \begin{bmatrix} z_1^1 & \cdots & z_1^{p+1} \\ \vdots & \ddots & \\ z_n^1 & \cdots & z_n^{p+1} \end{bmatrix}$$

convince yourself of this by multiplying them out.

$$\therefore \; H(w) = \sum_{i=1}^{n} z_i \, \mu_i (1-\mu_i) \, z_i^T$$

$$\boxed{H(w) = Z^T S Z}, \quad \text{where,} \quad H(w) \in \mathbb{R}^{(p+1)\times(p+1)}$$

$$Z \in \mathbb{R}^{n \times (p+1)}$$

$$S = \operatorname{diag}\left(\left[\mu_1(1-\mu_1), \ldots, \mu_n(1-\mu_n)\right]\right)$$

→ Note that $H(w)$ is symmetric, as $(Z^T S Z)^T = (SZ)^T (Z^T)^T = Z^T S Z$. This also makes sense because the second partial derivative does not care about the order of differentiation.

→ Also, note that if we stacked the data column-wise, i.e $Z^* \in \mathbb{R}^{(p+1)\times n}$, instead of $Z \in \mathbb{R}^{n \times (p+1)}$, then the above equation would be

$$H(w) = Z S Z^T.$$