1. $x \in R^d$ $\quad f(x) = \|x\|_2^2 = \left(\left(\sum_{i=1}^{d} x_i^2\right)^{\frac{1}{2}}\right)^2 = \sum_{i=1}^{d} x_i^2$

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{pmatrix} = \begin{pmatrix} 2x_1 \\ \vdots \\ 2x_d \end{pmatrix} = 2\vec{x}$$

2. $f(x) = A^T x \in R^n$, $A \in R^{d \times n}$ $\quad x \in R^d$

$$\underset{n \times d \times d \times 1}{\frac{\partial f}{\partial x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_1} \\ \vdots & & & \vdots \\ \frac{\partial f_1}{\partial x_d} & & & \frac{\partial f_n}{\partial x_d} \end{bmatrix} = A$$

$f(x) = [f_1(x), \ldots, f_n(x)]^T$

3. $g(x) = A^T x \in R^n$ $\quad f(y) = \|y\|_2^2 = \sum_{i=1}^{n} y_i^2$

$$\nabla_x f(g(x)) = \underbrace{\frac{\partial g(x)}{\partial x}}_{\frac{\partial g(x)}{\partial (x)} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_1} \\ \vdots & & & \vdots \\ \frac{\partial f_1}{\partial x_d} & & & \frac{\partial f_n}{\partial x_d} \end{bmatrix}} \frac{\partial f}{\partial g(x)} = 2 \cdot \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_1} \\ \vdots & & & \vdots \\ \frac{\partial f_1}{\partial x_d} & & & \frac{\partial f_n}{\partial x_d} \end{bmatrix} \cdot A^T x = 2AA^T x$$

$$\frac{\partial f}{\partial g(x)} = 2g(x)$$

4. $g(A) = A^T x \in \mathbb{R}^n$ ($n \times d \cdot d \times 1$) $f(y) = \|y\|_2^2$

$\frac{\partial f}{\partial \vec{y}} = 2\vec{y} \Rightarrow \boxed{\frac{\partial f}{\partial y} = 2A^T x}$

$\nabla_A f(g(A)) = \frac{\partial g(A)}{\partial A} \cdot \frac{\partial f}{\partial g} =$

$f(g(A)) = g(A)^T g(A) =$

$= (A^T x)^T A^T x = x^T A A^T x = y$ Scalar

vector by matrix derivative

derivative of scalar by matrix!

$\nabla_A g(A) = x e_i^T$

combining the results

we get in matrix form!

$\nabla_A x^T A A^T x = \begin{bmatrix} \frac{\partial y}{\partial A_{11}} & \cdots & \frac{\partial y}{\partial A_{1n}} \\ & & \\ \frac{\partial y}{\partial A_{d1}} & \cdots & \frac{\partial y}{\partial A_{dn}} \end{bmatrix} =$

$\boxed{2xx^T A}$

more explicitly I use the trace Ar

the scalar: Since $f(g(A))$ is a scalar only y then:

$y = x^T A A^T x$

$y = \text{trace}(y) = \text{trace}(x^T A A^T x) = \text{trace}(A^T x x^T A)$ $X = xx^T$ - Symmetric

$d \times d \quad d \times d \quad d \times n$

$n \times n$

$\frac{\partial}{\partial A} tr(A^T x x^T A) = \frac{\partial}{\partial A} tr(A^T X A) =$

$= 2XA = \boxed{2xx^T A}$

B. $J = -y \log \hat{y} - (1-y) \log (1-\hat{y})$

1 $\hat{y} = G(x + h + b_2)$

$h = w_2 \cdot [\max(0, d)]$ (relu)

$d = w_1 \cdot x + b_1$

$= G\left(x + w_2 \cdot \max(0, [w_1 \cdot x + b_1]) + b_2\right) = G\left(x + w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2\right)$

$\hat{y} = \frac{1}{1 + e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}}$

2.

$$\frac{\partial J}{\partial w_2} = \underbrace{\frac{\partial J}{\partial \hat{y}}}_{(1)} \cdot \underbrace{\frac{\partial \hat{y}}{\partial w_2}}_{(2)} = \left[\frac{-y}{\hat{y}} + \frac{(1-y)}{1-\hat{y}}\right] \cdot \frac{e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}}{\left(1 + e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}\right)^2} \cdot \text{Relu}(w_1 x + b_1)$$

(1): $\frac{-y}{\hat{y}} + \frac{(1-y)}{1-\hat{y}}$

(2): $\frac{\partial}{w_2} \frac{1}{1 + e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}} = G'\left(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x\right) \cdot \text{Relu}(w_1 x + b_1) =$

$G'(x) = G(x) \cdot (1 - G(x))$

$= \frac{1}{1 + e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}} \cdot \frac{e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}}{1 + e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}} \cdot \text{Relu}(w_1 x + b_1) =$

$= \frac{e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}}{\left(1 + e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}\right)^2} \cdot \text{Relu}(w_1 x + b_1)$

$$\frac{\partial J}{\partial b_2} = \underbrace{\frac{\partial J}{\partial \hat{y}}}_{①} \cdot \underbrace{\frac{\partial \hat{y}}{\partial b_2}}_{②} = \left[\frac{-y}{\hat{y}} + \frac{(1-y)}{1-\hat{y}}\right] \cdot \frac{e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}}{\left(1 + e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}\right)^2}$$

① $\frac{-y}{\hat{y}} + \frac{(1-y)}{1-\hat{y}}$

② $\sigma'\left(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x\right) = \frac{e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}}{\left(1 + e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}\right)^2}$

3. $$\frac{\partial J}{\partial w_1} = \underbrace{\frac{\partial J}{\partial \hat{y}}}_{①} \cdot \underbrace{\frac{\partial \hat{y}}{\partial w_1}}_{②} = \left[\frac{-y}{\hat{y}} + \frac{(1-y)}{1-\hat{y}}\right] \cdot \frac{e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}}{\left(1 + e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}\right)^2} \cdot \begin{cases} w_2 \cdot x, & w_1 x + b_1 > 0 \\ 0, & \text{otherwise} \end{cases}$$

① $\frac{-y}{\hat{y}} + \frac{(1-y)}{1-\hat{y}}$

② $\frac{\partial}{w_1} \frac{1}{1 + e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}} = \frac{e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}}{\left(1 + e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}\right)^2} \cdot \frac{\partial}{w_1}\left[w_2 \cdot \text{Relu}(w_1 x + b_1)\right] =$

$\frac{\partial}{w_1}\left[w_2 \cdot \text{Relu}(w_1 x + b_1)\right] = w_2 \cdot x \cdot \underbrace{\text{Relu}'\left[w_1 x + b_1\right]}$

$$\begin{cases} 0, & w_1 x + b_1 \le 0 \\ 1, & w_1 x + b_1 > 0 \end{cases}$$

② $\frac{e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}}{\left(1 + e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}\right)^2} \cdot w_2 \cdot x$, $w_1 x + b_1 > 0$

0, otherwise

3b. $\frac{\partial J}{\partial b_1} = \underbrace{\frac{\partial J}{\partial \hat{y}}}_{(1)} \cdot \underbrace{\frac{\partial \hat{y}}{\partial b_1}}_{(2)}$

$$= \left[\frac{-y}{\hat{y}} + \frac{(1-y)}{1-\hat{y}}\right] \cdot \frac{e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}}{\left(1 + e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}\right)^2} \cdot w_2, \quad w_1 x + b_1 > 0$$

$$0, \quad \text{otherwise}$$

(1) $\frac{-y}{\hat{y}} + \frac{(1-y)}{1-\hat{y}}$

(2) $\frac{\partial \hat{y}}{\partial b_1} = \frac{\partial}{b_1} \frac{1}{1 + e^{-(w_2 \text{Relu}(w_1 x + b_1) + b_2 + x)}} = \frac{e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}}{\left(1 + e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}\right)^2} \cdot \frac{\partial}{b_1} \left[w_2 \cdot \text{Relu}(w_1 x + b_1)\right] =$

$$\frac{\partial}{b_1} \left[w_2 \cdot \text{Relu}(w_1 x + b_1)\right] = w_2 \cdot 1 \cdot \text{Relu}'\left[w_1 x + b_1\right]$$

$$\begin{cases} 0, & w_1 x + b_1 \le 0 \\ 1, & w_1 x + b_1 > 0 \end{cases}$$

3c. $\frac{\partial J}{\partial x}$ = going to repeat same steps as before with the difference of the derivative of relu

$$\frac{\partial}{x} \left[w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x\right] = w_2 \cdot w_1 \cdot \text{Relu}'\left[w_1 x + b_1\right] + 1$$

$$\begin{cases} 0, & w_1 x + b_1 \le 0 \\ 1, & w_1 x + b_1 > 0 \end{cases}$$

$$\frac{\partial J}{\partial x} = \left[\frac{-y}{\hat{y}} + \frac{(1-y)}{1-\hat{y}}\right] \cdot \frac{e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}}{\left(1 + e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}\right)^2} \cdot (w_2 w_1 + 1), \quad w_1 x + b_1 > 0$$

$$\left[\frac{-y}{\hat{y}} + \frac{(1-y)}{1-\hat{y}}\right] \cdot \frac{e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}}{\left(1 + e^{-(w_2 \cdot \text{Relu}(w_1 x + b_1) + b_2 + x)}\right)^2}, \quad \text{otherwise}$$

4. We need to cache : X, $Z_1 = W_1 X + b_1$, $Z_2 = W_2 \cdot \text{Relu}(Z_1) + b_2 + X$,

$G(x)$ and $G'(x) = G(x) \cdot (1 - G(x))$, $\frac{\partial J}{\partial y} = \frac{-y}{G(x)} + \frac{1-y}{1-y}$,

$G'(W_2 \text{Relu}(W_1 X + b_1) + b_2 + X) = \frac{e^{(W_2 \text{Relu}(W_1 X + b_1) + b_2 + X)}}{(1 + e^{(W_2 \text{Relu}(W_1 X + b_1) + b_2 + X)})^2}$

$$\text{Relu}'(x) = \begin{cases} 1, & W_1 X + b_1 > 0 \\ 0, & \text{otherwise} \end{cases} \quad (1) \quad \cdot \quad (2) \ \frac{\partial y}{\partial k}$$

(2): where y is a function of k

variable by which we want to differentiate