Task 2

a) $y = x^3 + x - 5$

$\frac{dy}{dx} = 3x^{3-1} \cdot \frac{dx}{dx} + (1) - 0$

$= 3x^2 + (1) = 3x^2$

b) $y = (5x^3 - 2x)(2x)$

$\frac{dy}{dx} = f'(x) \cdot gx + f(x) g'(x)$

$f(x) = 5x^3 - 2x$

$g(x) = 2x$

$f'(x) = 5 \times 3x^2 - 2 = 15x^2 - 2$

$g'(x) = 2(1) = 2$

$\frac{dy}{dx} = (15x^2 - 2) \, 2x + (5x^3 - 2x) \cdot 2$

$= (30x^3 - 4x) + (10x^3 - 4x)$

$= 30x^3 - 4x + 10x^3 - 4x$

$\frac{dy}{dx} = 40x^3 - 8x$

c) $y = \frac{2x^2 + 3}{8x + 1}$

$y' = \frac{g'(x) \cdot f(x) - f(x) \cdot g'(x)}{(f(x))^2}$

$g(x) = 2x^2 + 3$

$f(x) = 8x + 1$

$g'(x) = 4x$

$f'(x) = 8$

$y' = \frac{(4x)(8x+1) - 8(2x^2+3)}{(8x+1)^2}$

$= \frac{32x^2 + 4x - 16x^2 - 24}{(8x+1)^2}$

$= \frac{16x^2 + 4x - 24}{(8x+1)^2}$

$=$

d) $y = (3x-2)^8$

$$\frac{dy}{dx} = 8(3x-2)^{8-1} \times 3(1)$$

$$= 8(3x-2)^7 * 3$$

$$= 24(3x-2)^7$$

e)

$$y = \log(x^2+x)$$

$$f(u) = \log(u) = \frac{1}{u}$$

$$g(x) = x^2+x$$

$$f'(u) = \frac{d}{du}(\log(u) = \frac{1}{u}$$

$$g'(x) = \frac{d}{dx}(x^2+x) = 2x+1$$

Now

$$\frac{dy}{dx} = f'(u) \cdot g'(x)$$

$$= \frac{1}{x^2+x} \cdot 2(x)+1$$

$$= \frac{2x+1}{x^2+x}$$

Task 1

$$X = \begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix}, \quad Y = \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \quad Z = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

a) $Y^T = \begin{bmatrix} 1 & 3 \end{bmatrix}, \quad Y^T Z = \underset{\substack{1\times 2 \\ m \; n}}{\begin{bmatrix} 1, 3 \end{bmatrix}} \underset{\substack{2\times 1 \\ n \; p}}{\begin{bmatrix} 2 \\ 3 \end{bmatrix}} = \begin{bmatrix} 2+9 \end{bmatrix} = \begin{bmatrix} 11 \end{bmatrix}$

$(1 \times p)$

b) $XY = \begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} = XY \ (2\times 1)$

$\quad\quad (2\times 2) \quad\quad (2\times 1)$

$= \begin{bmatrix} 2+12 \\ 1+9 \end{bmatrix} = \begin{bmatrix} 14 \\ 10 \end{bmatrix}_{(2\times 1)}$

c) $X^2 =$

$X \times X = \begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 4+4 & 8+12 \\ 2+3 & 4+9 \end{bmatrix}$

$= \begin{bmatrix} 8 & 20 \\ 5 & 13 \end{bmatrix}$

d) X is Invertible as its Determinant is non-zero

$X = \begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix} = det(X) = 6 - 4 = 2$

So now

$X^{-1} = \frac{1}{det(X)} = \begin{bmatrix} 3 & -4 \\ -1 & 2 \end{bmatrix} \div \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

$= \begin{bmatrix} \frac{3}{2} & -\frac{4}{2} \\ -\frac{1}{2} & \frac{2}{2} \end{bmatrix} = \begin{bmatrix} \frac{3}{2} & -2 \\ -\frac{1}{2} & 1 \end{bmatrix}$

P) Rank = ~~2~~

$$x = \begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix}$$

$$= \frac{1}{2}\begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix}$$ $\frac{1}{2} \times$ Row 1

$$= \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix}$$ Row 2 − Row 1

$$= \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$$ non zero Rows ⇒ 2

---

*Regulization

Solutions for Assignment 1, Task 3

Starting with, cross entropy,

$$\mathrm{CE}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_i y_i \log \hat{y}_i \tag{1}$$

We can derive the gradient of the cross-entropy function, using back propagation,

$$\frac{\partial(\mathrm{CE})}{\partial \hat{y}_i} = -\frac{y_j}{\hat{y}_i} \tag{2}$$

Thus,

$$\frac{\partial(\mathrm{CE})}{\partial \theta_k} = \frac{\partial(\mathrm{CE})}{\partial \hat{y}_i}\frac{\partial \hat{y}_i}{\partial \theta_k} = -\frac{y_j}{\hat{y}_i}\frac{\partial \hat{y}_i}{\partial \theta_k} \tag{3}$$

Calculating the partial derivative of $\hat{y}_i$ (where $i = k$)

$$\begin{aligned}\frac{\partial \hat{y}_i}{\partial \theta_i} =& \frac{\partial}{\partial \theta_i}\left(\frac{e^{\theta_i}}{\sum_{j=1} e^{\theta_j}}\right) & (4)\\ =& \frac{e^{\theta_i}}{\sum_{j=1} e^{\theta_j}} - \left(\frac{e^{\theta_i}}{\sum_{j=1} e^{\theta_j}}\right)^2 & (5)\\ =& \hat{y}_i \cdot (1 - \hat{y}_i) & (6)\end{aligned}$$

and (where $i \neq k$),

$$\begin{aligned}\frac{\partial \hat{y}_i}{\partial \theta_k} =& \frac{\partial}{\partial \theta_k}\left(\frac{e^{\theta_i}}{\sum_{j=1} e^{\theta_j}}\right) & (7)\\ =& -\left(\frac{e^{\theta_i} e^{\theta_k}}{\sum_{j=1} e^{\theta_j}}\right) & (8)\\ =& -\hat{y}_i \hat{y}_k & (9)\end{aligned}$$

Combining Equations 2, 6, 9, yields

$$\frac{\partial(\mathrm{CE})}{\partial \theta_k} = \begin{cases} -y_j(1 - \hat{y}_k) & \text{for } i = k \\ y_j \hat{y}_k & \text{for } i \neq k \end{cases} \tag{10}$$

Requiring y_j to be non-zero, imposes that the auxiliary condition, $k = j$ and $y_j = 1$, hence it follows immediately,

$$\frac{\partial(\mathrm{CE})}{\partial \theta_j} = \begin{cases} (\hat{y}_j - 1) & \text{for } i = j \\ \hat{y}_j & \text{for } i \neq j \end{cases} \tag{11}$$

Which is equivalent to (with the substitution y_j for 1 in the first case of Equation 11),

$$\frac{\partial(\mathrm{CE})}{\partial \boldsymbol{\theta}} = \hat{\mathbf{y}} - \mathbf{y} \tag{12}$$

Solutions for Assignment 1, Task 4a

In order to simplify the notation used to solve the problem, define the following terms:

$$\mathbf{x}^{(2)} \equiv \quad \mathbf{h} \tag{13}$$

$$\mathbf{z}_i \equiv \quad \mathbf{x}^{(\mathbf{i})}\mathrm{W}_{\mathbf{i}} + \mathbf{b}_{\mathbf{i}} \tag{14}$$

Now, to calculate $\partial J / \partial \mathbf{x}^{\mathbf{1}}$, one can use the back propagation algorithm.

$$\frac{\partial J}{\partial \mathbf{z}_2} = \hat{\mathbf{y}} - \mathbf{y} \tag{15}$$

and

$$\frac{\partial \mathbf{z}_i}{\partial \mathbf{x}^{(i)}} = \mathrm{W}_{\mathbf{i}}^{\top} \tag{16}$$

Sigmoid (σ) derivative can be found in Question 2(a), but we define:

$$\frac{\partial \mathbf{x}^{(2)}}{\partial \mathbf{z}_1} \equiv \sigma'(z_1) \tag{17}$$

Combining these, and using $\cdot$ to denote element-wise product:

$$\frac{\partial J}{\partial z_i} = (\hat{\mathbf{y}} - \mathbf{y})\mathrm{W}_{\mathbf{2}}^{\top} \cdot \sigma'(\mathbf{z_1}) \tag{18}$$

Finally, using the results from Equation 16 (but for the first layer):

$$\frac{\partial J}{\partial \mathbf{x}^{(1)}} = (\hat{\mathbf{y}} - \mathbf{y})\mathrm{W}_{\mathbf{2}}^{\top} \cdot \sigma'(\mathbf{z_1}) \cdot \mathrm{W}_{\mathbf{1}}^{\top} \tag{19}$$

Solutions for Assignment 1, Task 4b

$\mathrm{W}_{\mathbf{1}}$ must have dimensions: $\mathrm{D}_x \times \mathrm{H}$. The bias ($\mathbf{b_1}$) for the first layer must have dimensions H. Adding these two together, yields $(\mathrm{D}_x + 1) \times \mathrm{H}$. Proceeding to the second layer, there must be $\mathrm{H} \times \mathrm{D}_y$ parameters associated with the weight matrix W_2. The bias ($\mathbf{b_2}$) for the second layer must have dimensions D_y elements. This yields,

$$(\mathrm{D}_x + 1) \times \mathrm{H} + \mathrm{D}_y \times (\mathrm{H} + 1) \tag{20}$$

weights, for each input vector of dimensions D_x.