2. **Softmax regression gradient calculation**

Consider a simple Softmax regression model,

$$\hat{\mathbf{y}} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b})$$

→ Multiclass classification problem

$$\mathbf{x} \in \mathbb{R}^d, \mathbf{W} \in \mathbb{R}^{k \times d}, \mathbf{b} \in \mathbb{R}^k$$

where d is the input dimension, k is the number of classes, σ is the softmax function:

$$\sigma(\mathbf{a})_i = \frac{\exp(a_i)}{\sum_j \exp(a_j)}$$

(a) Given the cross-entropy loss → Between y true label and $\hat{y}$ softmax

$$l(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_i y_i \log \hat{y}_i$$

$\mathbf{y}$ is the one-hot vector representing true labels ($[0, 0, \cdots, 1, 0, 0, \cdots]^T$ with the 1 corresponding to the true label), derive $\frac{\partial l}{\partial W_{i,j}}$. (You can use your results from Assignment 1)

(b) What happens to the loss function and the gradients when $y_{c_1} = 1, \hat{y}_{c_2} = 1, c_1 \neq c_2$? Why there is no need to worry about this situation?

Remember Sigmoid is used for binary classification methods and softmax is used on multiclass problems.

Softmax is an extension of the Sigmoid function.

Sigmoid function: $\frac{1}{1 + e^{-x}}$ Softmax function: $\frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}}$

2(a) Find $\frac{\partial l}{\partial W_{i,j}}$, derivative of scalar by matrix

$$Wx + b = \sum_{j=1}^{d} (w_{i,j} x_j) + b$$

2 (b) When y_{c_1} and $\hat{y}_{c_2} = 1$ and $c_1 \neq c_2$ it means that Softmax model is classifying c_1 true class like c_2 class

$$l(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_i y_i \log \hat{y}_i$$

There is no need to worry because $\log(\hat{y}_{c_2}) = 0$ $y_i \cdot 0 = 0$ and for that

Class