Theoretical part: Question 1

First denote $f(x,w)_i = z_i$ From the chain rule we get

$$\frac{\partial L(\hat{y},y)}{\partial w} = \sum_i \frac{\partial L(\hat{y},y)}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial w}$$

$$\frac{\partial \hat{y}_i}{\partial w} = \sum_k \frac{\partial \hat{y}_i}{\partial z_k} \frac{\partial z_k}{\partial w}$$

and since

$$\hat{y}_i = softmax(x)_i = \frac{e^{f(x,w)_i}}{\sum_j e^{f(x,w)_j}} = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

we get

$$\frac{\partial \hat{y}_i}{\partial z_k} = \begin{cases} \frac{e^{z_i}}{\sum_j e^{z_j}} \left(1 - \frac{e^{z_i}}{\sum_j e^{z_j}}\right) = \hat{y}_i (1 - \hat{y}_i) & \text{, i = k} \\ -\frac{e^{z_i}}{\sum_j e^{z_j}} \frac{e^{z_k}}{\sum_j e^{z_j}} = -\hat{y}_i \hat{y}_k & \text{, i} \neq \text{k} \end{cases}$$

$$\frac{\partial \hat{y}_i}{\partial w} = \sum_{k \neq i} -\hat{y}_i \hat{y}_k \frac{\partial z_k}{\partial w} + \hat{y}_i (1 - \hat{y}_i) \frac{\partial z_i}{\partial w} = \sum_k -\hat{y}_i \hat{y}_k \frac{\partial z_k}{\partial w} + \hat{y}_i \frac{\partial z_i}{\partial w}$$

$$\frac{\partial z_i}{\partial w} = \frac{\partial f(x,w)_i}{\partial w}$$

and so we get

$$\frac{\partial L(\hat{y},y)}{\partial w} = \sum_i \left(\sum_k -\hat{y}_i \hat{y}_k \frac{\partial z_k}{\partial w} + \hat{y}_i \frac{\partial z_i}{\partial w} \right) \frac{\partial L(\hat{y},y)}{\partial \hat{y}_i}$$