$n_1 = 7 \rightarrow$ Hidden layer

$n_2 = 10 \rightarrow$ Output Layer

Logistic sigmoid for hidden layers : $\sigma(z) = \dfrac{1}{1+e^{-z}}$

$$\sigma'(z) = \sigma(z)(1-\sigma(z))$$

Softmax for output layer : $S(z_i^{(2)}) = \dfrac{e^{z_i^{(2)}}}{\sum_{k=1}^{n_2} e^{z(2)}}$

Cross-entropy loss : $E = -\dfrac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{n_2} t_{ij} \log(y_{ij})$

$\dfrac{\partial E}{\partial z^{(2)}} = y^{(2)} - t$, where $y^{(2)}$ is the softmax output

Using chain rule : $\dfrac{\partial E}{\partial W^{(2)}} = \dfrac{\partial E}{\partial z^{(2)}} \cdot (y^{(1)})^T$

, where $y^{(1)}$ is the output of the hidden layer after applying the sigmoid function.

$\Rightarrow \dfrac{\partial E}{\partial W^{(2)}} = (y^{(2)} - t) \cdot (y^{(1)})^T$: output layer weights

$z^{(1)}$ is : $\dfrac{\partial E}{\partial z^{(1)}} = (W^{(2)})^T \cdot \dfrac{\partial E}{\partial z^{(2)}} \circ \sigma'(z^{(1)})$

$\Rightarrow \dfrac{\partial E}{\partial z^{(1)}} = (W^{(2)})^T \cdot (y^{(2)} - t) \circ y^{(1)} \cdot (1 - y^{(1)})$

Input Vector

$W^{(1)}$ is : $\dfrac{\partial E}{\partial W^{(1)}} = \dfrac{\partial E}{\partial z^{(1)}} \cdot x^T$

$\Rightarrow \dfrac{\partial E}{\partial W^{(1)}} = \left[(W^{(2)})^T (y^{(2)} - t) \cdot y^{(1)} \cdot (1 - y^{(1)}) \right] \cdot x^T$

: Hidden layer weights

When weight decay is included, the loss function :

$E = \dfrac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{n_2} t_{ij} \log(y_{ij}) + \dfrac{\lambda}{2} \|W\|^2$ then, $\dfrac{\partial E}{\partial W^{(l)}} \rightarrow \dfrac{\partial E}{\partial W^{(l)}} + \lambda W^{(l)}$