$\nabla_{W_1}(J)$

$\nabla_{b_1}(J)$

$\nabla_{W_3} J = \nabla_{W_3}(L(\hat{y}, y) + \lambda \Omega(\theta))$

$= \nabla_{a_3}(L(\hat{y}, y)) \nabla_{W_3}(a_3) + \lambda \nabla_{W_3}(\Omega(\theta))$

$= \delta^{(3)} f'(a_3) h_2^T + \lambda \nabla_{W_3}(\Omega(\theta))$

$\nabla_{b_3} J =$

$\delta^{(3)} f'(a_3) 1$

$+ \nabla_{b_3}(\Omega(\theta))$

Dont worry

I moved the text to the

next page

$$\nabla_{h_2}(L(\hat{y},y)) = \nabla_{h_2} a_3 \nabla_{a_3} L(\hat{y},y)$$

$$= \nabla_{h_2}(w_3^T h_2 + b_3)(g \odot f'(a_3))$$

$$= w_3^T (g \odot f'(a_3))$$

$$g \leftarrow \nabla_{h_2}(L(\hat{y},y)) = w_3^T (g \odot f'(a_3))$$

$$\nabla_{a_2}(L(\hat{y},y)) = \nabla_{h_2}(L(\hat{y},y))\ \nabla_{a_2}(h)$$

$$= g\, \nabla_{a_2}(f(a_2)) = g \cdot f'(a_2)$$

$$\nabla_{w_2} J = \nabla_{w_2}(L(\hat{y},y)) + \lambda w_2 (\Omega(\theta))\ (\text{2 layer})$$

$$= \nabla_{a_2}(L(\hat{y},y))\, \nabla_{w_2}(a_2) + \lambda \nabla_{w_2}(\Omega(\theta))$$

$$= (g \cdot f'(a_2))(\nabla_{w_2}(w_2^T h_1 + b_2)) + \lambda \nabla_{w_2}(\Omega(\theta))$$

$$\nabla_{b_2}(J) = \nabla_{b_2}(L(\hat{y},y) + \lambda \Omega(\theta))$$

$$= \nabla_{a_2}(L(\hat{y},y))\, \nabla_{b_2}(a_2) + \nabla_{b_2}(\lambda \Omega(\theta))$$

$$= (g \cdot f'(a_2)) \nabla_{b_2}(w_2^T h_1 + b_2) + \nabla_{b_2}(\lambda \Omega(\theta))$$

$$= (g \cdot f'(a_2)) + \lambda \nabla_{b_2}(\Omega(\theta))$$

$$\nabla_{h_1}(L(\hat{y},y)) = \nabla_{a_2}(L(\hat{y},y))\, \nabla_{h_1}(a_2)$$

$$= (g \cdot f'(a_2))\, \nabla_{h_1}(w_2^T h_1 + b_1)$$

$$= (g \cdot f'(a_2))\, w_2^T$$

$$g \leftarrow \nabla_h(L(\hat{y},y))$$

III $\nabla_{w_1}(J)$

$\nabla_{b_1}(J)$ can be found thru chain rule

in textbook he takes gradient of loss function with regularization

$g \leftarrow \nabla_{\hat{y}}(J)$ this is incorrect!

$$\nabla_{\hat{y}}(J) = \nabla_{\hat{y}}\left(L(y,\hat{y}) + \lambda\,\Omega(w,b)\right)$$

$$= \nabla_{\hat{y}}\left(L(y,\hat{y})\right) + \lambda\,\nabla_{\hat{y}}\left(\Omega(w,b)\right)$$

both $\hat{y}$ ←

& Ω are functions of w_3!!

This gradient is not zero!

HOWEVER

You may say that TB considers regularization as part of the loss function i.e. $L(\hat{y}, y)$ has the regularized term inside

eg: $L(\hat{y}, y) = MSE + \lambda\,\Omega(\theta)$