- **Task 1 (25 p):** Show that

$$\delta^{[l]} = g^{[l]'}(Z^{[l]}) \odot \left(W^{[l+1]^T} * \delta^{[l+1]}\right), \quad l = 2, \ldots, L-1 \qquad \text{(i)}$$

and

$$\delta^{[L]} = J'(A^{[L]}) \odot g^{[L]'}(Z^{[L]}) \qquad \text{(ii)}$$

where $\odot$ denotes element-wise multiplication and $*$ denotes ordinary matrix multiplication. Here $g^{[L]'}(Z^{[L]})$ and $J'(A^{[L]})$ denote the derivatives of the univariate functions applied element-wise to the vectors. If any operations with tensors are used, these need to be clearly defined.

(i)

$$\frac{\partial J}{\partial z^{[l]}} = \underbrace{\frac{\partial J}{\partial z^{[L]}} \frac{\partial z^{[L]}}{\partial g^{[L-1]}} \frac{\partial g^{[L-1]}}{\partial z^{[L-1]}} \frac{\partial z^{[L-1]}}{\partial g^{[L-2]}} \cdots\cdots \frac{\partial z^{[l+1]}}{\partial g^{[l]}}}_{= \frac{\partial J}{\partial g^{[l]}} = \frac{\partial J}{\partial z^{[l+1]}} \frac{\partial z^{[l+1]}}{\partial g^{[l]}}} \underbrace{\frac{\partial g^{[l]}}{\partial z^{[l]}}}_{= g^{[l]'}(z^{[l]})}$$

$$= \delta^{[l+1]} \frac{\partial}{\partial g^{l}} \left[W^{[l+1]} g^{[l]}(z^{[l]}) + b \right]$$

$$= \delta^{[l+1]} \; W^{[l+1]}$$

$$\Rightarrow \quad \delta^{[l]} = \frac{\partial g^{[l]}}{\partial z^{[l]}} \odot \left(W^{[l+1]^T} * \delta^{[l+1]} \right)$$

$\blacksquare$

(ii) show $\delta^{[L]} = \dfrac{\partial J}{\partial z^{[L]}} = \dfrac{\partial J}{\partial g^{[L]}} \dfrac{\partial g^{[L]}}{\partial z^{[L]}}$ $\blacksquare$

simple application of chain rule.

- **Task 2 (10 p):** Derive the corresponding expression for the gradients under the following setting:
  a) – Cost function: Mean Squared Error (MSE)
  b) – Activation function for hidden layers: Rectified Linear Unit (ReLU)
  c) – Activation function for the output layer: Identity function

a) $$MSE(\theta) = \frac{1}{m}\sum_{i=1}^{m} \| Y - \hat{Y} \|_2^2 = \left\{ \begin{array}{l} \text{from task } (c) \text{ we know } \hat{Y} = Z^{[L]} \\ \Rightarrow \frac{1}{m}\sum_{i=1}^{m} \| Y - Z^{[L]} \|_2^2 \end{array} \right\}$$

$$\frac{\partial J}{\partial Z^{[L]}} = \frac{\partial J}{\partial Z^{[L]}} \left[\frac{1}{m} \sum_{i=1}^{m} \sum_{p=1}^{D} (y_i - z^{[L]})_i^2 \right] = -\frac{2}{m} \sum \sum_{p=1}^{D} (y_i - z_i^{[L]}) \quad (*)$$

$$\nabla_\theta J = \frac{\partial J}{\partial \theta} = \left(\frac{\partial J}{\partial W^{[1]}}, \ldots, \frac{\partial J}{\partial W^{[L]}}, \frac{\partial J}{b^{[1]}}, \ldots, \frac{\partial J}{\partial b^{[L]}} \right)$$

(i) $$\frac{\partial J}{\partial W^{[l]}} = \frac{\partial J}{\partial Z^{[l]}} \frac{\partial Z^{[l]}}{\partial W^{[l]}} = \left(\frac{\partial g^{[l]}}{\partial Z^{[l]}} \right)^T \left(W^{[l+1]^T} \delta^{[l+1]} \right) \cdot \frac{\partial Z^{[l]}}{\partial W^{[l]}}$$

$$= \delta^{[l]} \cdot g^{[l-1]}(Z^{[l-1]})$$

$$Z^{[l]} = W^{[l]} A^{[l-1]} + b^{[l]}$$

(ii) $$\frac{\partial J}{\partial b^{[l]}} = \frac{\partial J}{\partial Z^{[l]}} \frac{\partial Z^{[l]}}{\partial b^{[l]}} = \frac{\partial g^{[l]}}{\partial Z^{[l]}} \left(W^{[l+1]^T} \delta^{[l+1]} \right) = \delta^{[l]}$$

(iii) $$\frac{\partial J}{\partial W^{[L]}} = \frac{\partial J}{\partial Z^{[L]}} \frac{\partial Z^{[L]}}{\partial W^{[L]}} = \frac{\partial J}{\partial Z^{[L]}} \cdot g^{[L-1]}(Z^{[L-1]}) = \delta^{[L]} \cdot g^{[L-1]}(Z^{[L-1]})$$

(iv) $$\frac{\partial J}{\partial b^{[L]}} = \frac{\partial J}{\partial Z^{[L]}} \frac{\partial Z^{[L]}}{\partial b^{[L]}} = \ldots = \delta^{[L]}$$

b) $\frac{\partial g^{[L]}}{\partial z^{[L]}} = \frac{\partial}{\partial z^{[L]}}\left[\max(0, z^{[L]})\right] = \max(0,1) = 1$

c) $g(z^{[L]}) = z^{[L]}$ identity

plug into (i) - (iv) to get final gradient

- **Task 3 (5 p):** How would you avoid an exponential blowup of computation when computing the gradients?

Use back-propagation.

Refer to page 205, equations (6.49) - (6.52) in the course literature, and also what we did in the previous task. Several parts need only be computed once and then stored in memory.

$$\frac{\partial z}{\partial w} \tag{6.49}$$

$$= \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}\frac{\partial x}{\partial w} \tag{6.50}$$

$$= f'(y)f'(x)f'(w) \tag{6.51}$$

← we compute this

$$= f'(f(f(w)))f'(f(w))f'(w). \tag{6.52}$$

← not this.

exponential blowup.