Define $\delta_j^{(\ell)} = \dfrac{\partial c}{\partial z_j^{(\ell)}}$

We will later relate $\delta_j^{(\ell)}$ back to the desired gradients.

c is written as a function of $a^{(L)}$: $c = c(a^{(L)})$

last layer

Let's get $\delta_j^{(L)}$ first:

$$\delta_j^{(L)} = \frac{\partial c}{\partial z_j^{(L)}} = \frac{\partial c}{\partial a_j^{(L)}} \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}}$$

$$\boxed{\delta_j^{(L)} = \frac{\partial c}{\partial a_j^{(L)}} g_L'(z_j^{(L)})} \quad (1)$$

For layer $\ell < L$:

$$\delta_j^{(\ell)} = \frac{\partial c}{\partial z_j^{(\ell)}} = \sum_{k=1}^{n_{\ell+1}} \underbrace{\frac{\partial c}{\partial z_k^{(\ell+1)}}}_{=\delta_k^{(\ell+1)}} \frac{\partial z_k^{(\ell+1)}}{\partial z_j^{(\ell)}}$$

where
$$\frac{\partial z_k^{(\ell+1)}}{\partial z_j^{(\ell)}} = \frac{\partial}{\partial z_j^{(\ell)}} \left[\sum_{i=1}^{n_{\ell-1}} a_i^{(\ell)} W_{ik}^{(\ell+1)} + b_k^{(\ell+1)} \right]$$

$$= \sum_i g_\ell'(z_i^{(\ell)}) \frac{\partial z_i^{(\ell)}}{\partial z_j^{(\ell)}} W_{ik}^{(\ell+1)}$$

$$= g_\ell'(z_j^{(\ell)}) W_{jk}^{(\ell+1)}$$

$$\Rightarrow \boxed{\delta_j^{(\ell)} = \sum_k \delta_k^{(\ell+1)} W_{kj}^{(\ell+1)T} g_\ell'(z_j^{(\ell)})} \quad (2)$$

Now let's get the gradients:

$$\frac{\partial c}{\partial W_{ij}^{(\ell)}} = \sum_{k=1}^{n_{\ell}} \frac{\partial c}{\partial z_k^{(\ell)}} \frac{\partial z_k^{(\ell)}}{\partial W_{ij}^{(\ell)}}$$

$$= \sum_k \delta_k^{(\ell)} \frac{\partial}{\partial W_{ij}^{(\ell)}} \sum_m \left(a_m^{(\ell-1)} W_{mk}^{(\ell)} + b_k^{(\ell)} \right)$$

$$\boxed{\frac{\partial c}{\partial W_{ij}^{(\ell)}} = a_i^{(\ell-1)} \delta_j^{(\ell)}} \quad (3)$$

$$\frac{\partial c}{\partial b_j^{(\ell)}} = \sum_k \frac{\partial c}{\partial z_k^{(\ell)}} \frac{\partial z_k^{(\ell)}}{\partial b_j^{(\ell)}}$$

$$\boxed{\frac{\partial c}{\partial b_j^{(\ell)}} = \delta_j^{(\ell)}} \quad (4)$$

Summary: training algorithm (one "epoch")

- $\vec{a}^{(0)} = \vec{x}$
- \# forward pass:
  for $\ell = 1, \ldots, L$:
  $\quad \vec{a}^{(\ell)} = g_\ell(\vec{a}^{(\ell-1)} W^{(\ell)} + \vec{b}^{(\ell)})$
- \# backpropagation:
  get $\vec{\delta}^{(L)}$ from (1)
  for $\ell = L-1, \ldots, 1$:
  $\quad$ get $\vec{\delta}^{(\ell)}$ from (2)
  $\quad$ get $\partial c / \partial W_{ij}^{(\ell)}$ and $\partial c / \partial b_j^{(\ell)}$ from (3) & (4)
- update $W_{ij}^{(\ell)} = W_{ij}^{(\ell)} - \eta \dfrac{\partial c}{\partial W_{ij}^{(\ell)}}$
  $\qquad\quad b_j^{(\ell)} = b_j^{(\ell)} - \eta \dfrac{\partial c}{\partial b_j^{(\ell)}}$
  or use an alternative to GD