Note: $\frac{\partial h_i}{\partial z_j} = \frac{\partial(\mathrm{Relu}(z_i))}{\partial z_j} = \delta_{ij} \underbrace{\mathbb{1}(z_i > 0)}_{\substack{1 \text{ if } z_i > 0 \\ 0 \text{ otherwise}}}$

d) $\frac{\partial L}{\partial z_2} = \frac{\partial L}{\partial h_2} \frac{\partial h_2}{\partial z_2} = (-0.01799)\cdot(1) = -.0180$

$\frac{\partial L}{\partial z_4} = \frac{\partial L}{\partial h_4} \frac{\partial h_4}{\partial z_4} = .0360$

$z_i = a_{ij} x_j \qquad \frac{\partial z_k}{\partial a_{ij}} = \delta_{ik} x_j$

e) $\frac{\partial L}{\partial a_{21}} = \frac{\partial L}{\partial z_2} \frac{\partial z_2}{\partial a_{21}} = -0.0180 * x_1 = -0.0180 \cdot -1 = .0180$

$\frac{\partial L}{\partial a_{22}} = \frac{\partial L}{\partial z_2} \frac{\partial z_2}{\partial a_{22}} = -0.0180 \cdot x_2 = -0.0180 \cdot 3 = -0.0540$

6)

a) # Weights $\sim O(H^2 L)$

b) # computation for $\frac{\partial L}{\partial w_{ij}^{(0)}} \sim O(H^2 L)$

c) Following the product of (a) and (b) ~ our number of computations should simply approximately $O((H^2 L)^2)$ thus our number of computations is quadratic with respect to weights

7)

a)

$$h_i^{(l+1)} = \sum_j W_{ij}^l h_j^l$$

$$\frac{\partial h_i^{(l+1)}}{\partial W_{jk}^{(l)}} = \frac{\partial}{\partial W_{jk}^l} \sum_j W_{im}^l h_m = \delta_{ij} \delta_{mk} h_m^l$$

$$= \delta_{ij} h_k^l$$

b)

$$\frac{\partial L}{\partial W_{jk}^{(l)}} = \frac{\partial L}{\partial h_i^{(l+1)}} \frac{\partial h_i^{(l+1)}}{\partial W_{jk}^{(l)}} = \frac{\partial L}{\partial h_i^{(l+1)}} h_k \delta_{ij}$$

$$\frac{\partial L}{\partial W_{jk}^{(l)}} = \frac{\partial L}{\partial h_j^{(l+1)}} h_k^l \qquad \frac{\partial L}{\partial h^{(l+1)}} \text{ is given}$$

H^2 for weights between L and $L+1$

∴ We only need to compute 1 derivative

c)

$$\frac{\partial L}{\partial W^{(l)}} = \begin{bmatrix} \frac{\partial L}{\partial h_1^{(l+1)}} \\ \vdots \\ \frac{\partial L}{\partial h_H^{(l+1)}} \end{bmatrix} \begin{bmatrix} h_1^l & h_2^l & \cdots & h_H^l \end{bmatrix}$$

$$= \frac{\partial L}{\partial \vec{h}^{(l+1)}} (\vec{h}^l)^T = \frac{\partial L}{\partial h^{l+1}} \otimes \vec{h}^l$$

outer product notation

total computations = $O(H^2)$

d) $h_i^{(l+1)} = W_{ij}^{(l)} h_j^{(l)}$

$$\frac{\partial h_i^{(l+1)}}{\partial h_j^{(l)}} = W_{ij}^{(l)}$$

e) $\frac{\partial L}{\partial h_j^{(l)}} = \sum_i \frac{\partial L}{\partial h_i^{(l+1)}} \frac{\partial h_i^{(l+1)}}{\partial h_j^{(l)}} = \sum_i \frac{\partial L}{\partial h_i^{(l+1)}} W_{ij}^{(l)}$

i ranges from 1 to H

H computations additional

f) $\frac{\partial L}{\partial \bar{h}^{(l)}} = \left(W^{(l)}\right)^T \frac{\partial L}{\partial \bar{h}^{(l+1)}}$

layer of weight

total number of computations is H^2

Big O doesn't care about coefficients

g) $O(H^2L + H^2L) \sim O(H^2L)$

need to make $O(H^2L)$ computations for one iteration of backpropagation. This number is proportional to the total number of weights in the network

h) Foward pass requires $\sim H^2L$ computations computations.

Back pass requires $\sim 2H^2L$ computations

Ratio of forward pass to backward pass computation : $\sim \frac{1}{2}$