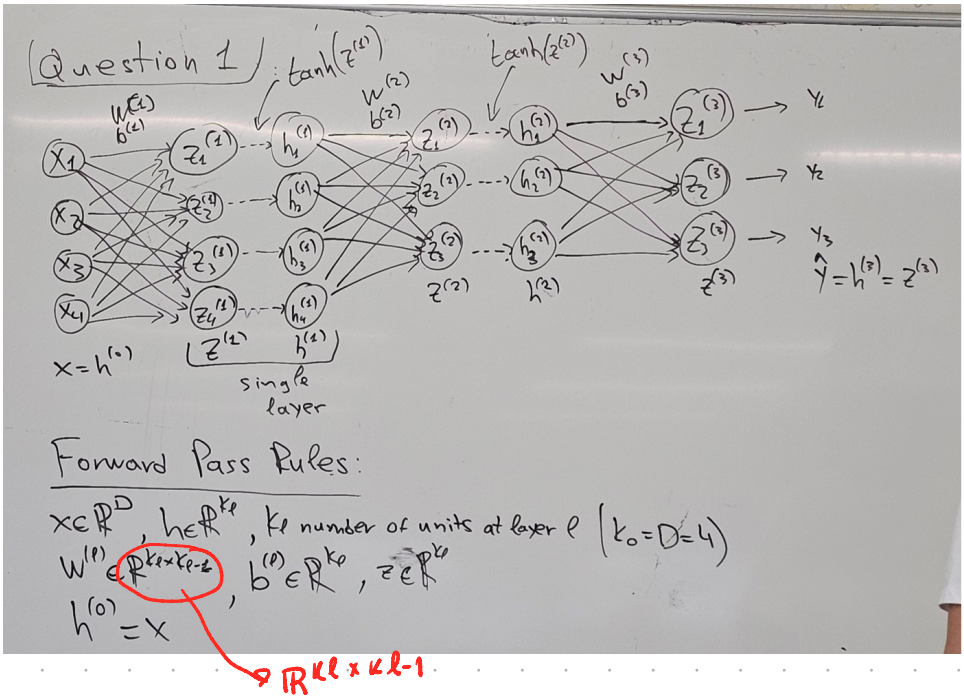
QUESTION 1

4433 (including input & output)

All except output include hyperbolic tangent activation functions

- Initialize all weights & bias to 0.1
- Do SGD using squared error loss
- learning rate $\eta = 0.1$



z preactivation: $z^{(l)} = W^{(l)} h^{(l-1)}(x) + b^{(l)}$

h representation: $h^{(l)} = g(z^{(l)})$, g: activation function

output: $\hat{y} = h^{(l=L)}$, L is the last layer

loss: $L(\hat{y}, y)$

weight init:

$$W^{(1)}_{4\times4} = \begin{bmatrix} w_{11} & \cdots & w_{14} \\ \vdots & \ddots & \vdots \\ w_{41} & \cdots & w_{44} \end{bmatrix} = \begin{bmatrix} 0.1 & \cdots & 0.1 \\ \vdots & \ddots & \vdots \\ 0.1 & \cdots & 0.1 \end{bmatrix}$$

$$b^{(1)}_{4\times1} = \begin{bmatrix} 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \end{bmatrix}$$

And also $g = \tanh$

$$W^{(2)}_{3\times4} = \begin{bmatrix} 0.1 & \cdots \\ \vdots & \ddots \\ \vdots & \ddots \end{bmatrix}, \quad b^{(2)}_{3\times1} = \begin{bmatrix} 0.1 \\ 0.1 \\ 0.1 \end{bmatrix}$$

$$W^{(3)}_{3\times3} = \begin{bmatrix} \cdots \\ \cdots \end{bmatrix}, \quad b^{(3)}_{3\times1} = \begin{bmatrix} \cdot \\ \vdots \end{bmatrix}$$

$x = [1\ 0\ 1\ 0]^T$, $y = [0\ 1\ 0]^T$

Run the forward pass:

$h^{(0)} = x = [1\ 0\ 1\ 0]^T$

$z^{(1)} = W^{(1)} h^{(0)}(x) + b^{(1)} = [0.3\ 0.3\ 0.3\ 0.3]^T$

$h^{(1)} = \tanh(z^{(1)}) = [0.2913 \dots 0.2913]^T$

$z^{(2)} = W^{(2)} h^{(1)} + b^{(2)} = [0.2165 \dots 0.2165]^T$

$h^{(2)} = \tanh(z^{(2)}) = [\dots]^T$

$z^{(3)} = W^{(3)} h^{(2)} + b^{(3)} = [0.16936 \dots 0.16936]^T$

$h^{(3)} = z^{(3)} = \hat{y}$

Loss: $L(\hat{y}, y) = \frac{1}{2} \sum_{i=1}^{3} (\hat{y}_i - y_i)^2 = \dots = 0.376$

(sum of squares)

(in mean squared error we would normalize w/ $\frac{1}{N}$ but here no!)

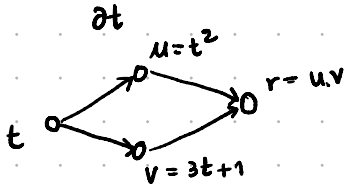
Backpropagation (Recap): $\frac{\partial L}{\partial W^{(l)}}$, $\frac{\partial L}{\partial b^{(l)}}$, $\frac{\partial L}{\partial h^{(l)}}$, $\frac{\partial L}{\partial z^{(l)}}$

Recap of the chain Rule: (slide 53)

Example: If a function $r(t)$ can be written in terms of intermediate results $q_i(t)$ then we have:

$$\frac{\partial r(t)}{\partial t} = \sum_i \frac{\partial r}{\partial q_i} \frac{\partial q_i(t)}{\partial t}$$

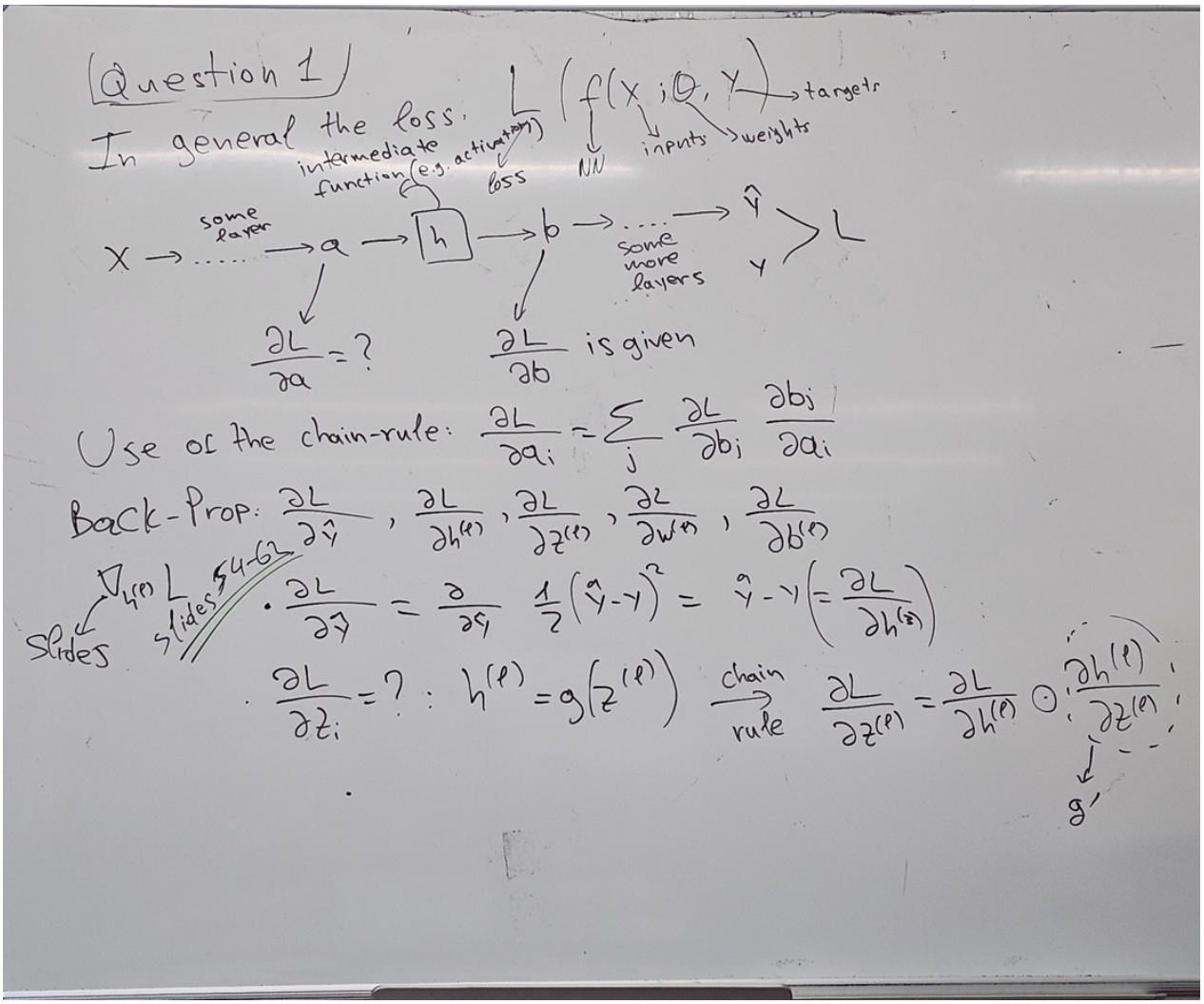
so for example:



$$\frac{\partial r(t)}{\partial t} = \frac{\partial r(u)}{\partial u} \frac{\partial u(t)}{\partial t} + \frac{\partial r(v)}{\partial v} \frac{\partial v(t)}{\partial t} = \dots = 9t^2 + 2t$$

In general the loss: $L(f(x; \theta, y))$

loss, NN, inputs, weights, targets



For $l=3$: g^l is the identity

so $\frac{\partial L}{\partial z^{(3)}} = \frac{\partial L}{\partial h^{(3)}}$

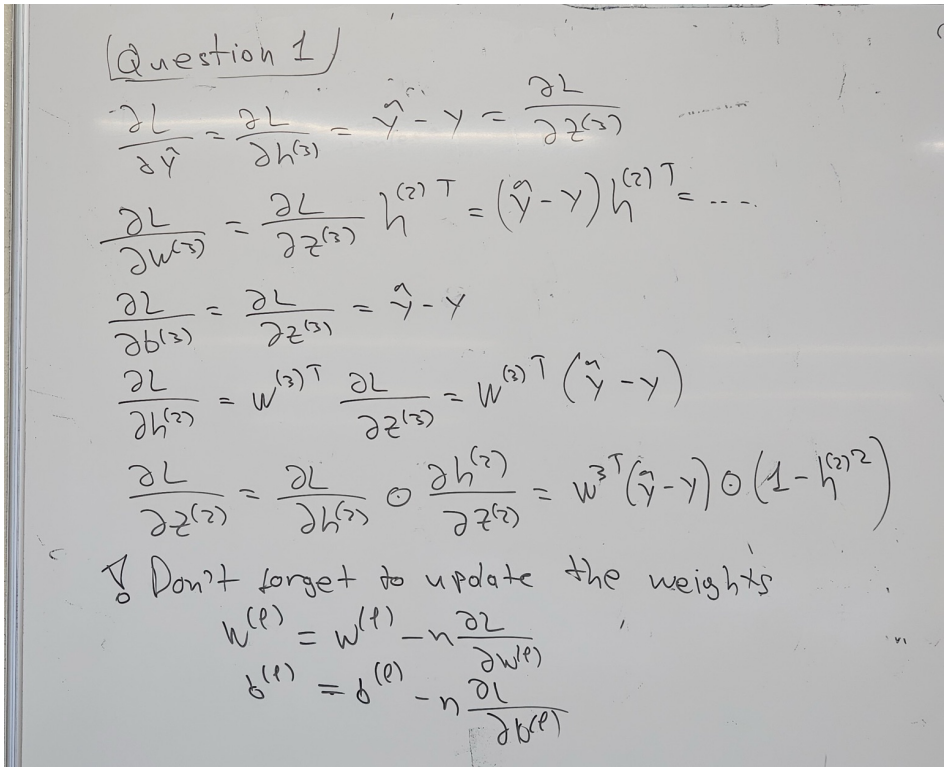
For $l \neq 3$: $g = \tanh$, so $g' = 1-\tanh^2$

$h^{(l)} = \tanh(z^{(l)})$

so $\frac{\partial L}{\partial z^{(l)}} = \frac{\partial L}{\partial h^{(l)}} \odot (1-\tanh^2(z^{(l)}))$

$= \frac{\partial L}{\partial h^{(l)}} \odot (1-h^{(l)2})$

- $\frac{\partial L}{\partial b^{(l)}} = \frac{\partial L}{\partial z^{(l)}}$ (slide 61)
- $\frac{\partial L}{\partial W^{(l)}} = \frac{\partial L}{\partial z^{(l)}} h^{(l-1)T}$ (slide 61)
- $\frac{\partial L}{\partial h^{(l-1)}} = W^{(l)T} \frac{\partial L}{\partial z^{(l)}}$



2) we know how to compute the grad for 1 example
what if we are given more?

$$L(\theta) = \sum_i L(f(x_i; \theta), y_i)$$

$$\rightarrow \nabla_\theta L(\theta) = \nabla_\theta \sum_i L(f(x_i;\theta), y_i) =$$

$$= \sum_i \nabla_\theta L(f(x_i;\theta), y_i)$$

$$L(\theta) = \frac{1}{N} \sum_i L_i$$

$$\nabla_\theta L(\theta) = \frac{1}{N} \sum_i \nabla_\theta L_i$$

the grad of the sum is the sum of the grads ...

$\nabla L = \nabla(L_1 + L_2)$
$\nabla L = \nabla L_1 + \nabla L_2$

QUESTION 2

- softmax activation
- error f. is cross-entropy

Same as question 1 but:

$$\hat{y} = h^{(3)} = \text{softmax}(z^{(3)})$$

cross-entropy: $L(\hat{y}, y) = -\sum_k y_k \log \hat{y}_k$

Tip: you can skip the computation of $\frac{\partial L}{\partial h^{(3)}}$
its easier to compute directly $\frac{\partial L}{\partial z^{(3)}}$

$$L(z, y) = -\sum_k y_k \log \hat{y}_k =$$

$$= -\sum_k y_k \underbrace{\log}_{\text{log}} \text{softmax}(z_k) =$$

$$= -\sum_k y_k \log \left[\frac{\exp(z_k)}{\sum_j \exp(z_j)} \right]$$

$$= -\sum y_k \left[z_k - \log \sum_j \exp(z_j) \right] =$$

$$= \sum y_k \left[\log \left(\sum_j \exp(z_j) \right) - z_k \right]$$

because derivative w.r. respect to z_i will give zero in all the other records

$$\frac{\partial L}{\partial z_i} = \frac{\partial}{\partial z_i} \left[\sum_k y_k \log \sum_j \exp(z_i) - \sum_k y_k z_k \right]$$

$$= \frac{\partial}{\partial z_i} \sum_k y_k \log \left(\sum_i \exp(z_j) \right) - y_i$$

$\frac{\partial}{\partial z_i}$ ($y_1 z_1 + y_2 z_2 + \dots$ $\dots y_i z_i + \dots + y_k z_k$)

$\frac{\partial}{\partial z_i}(y_i z_i) = y_i$

$1 =$ $\sum_k y_k$ $\frac{\partial}{\partial z_i} \log \sum_i \exp(z_j) - y_i$

$$= \frac{\partial}{\partial z_i} \log \left(\sum_i \exp(z_j) \right) - y_i$$

$$= \frac{1}{\sum_j \exp(z_j)} \cdot \frac{\partial}{\partial z_i} \sum_j \exp(z_j) - y_i =$$

$$= \frac{\exp(z_i)}{\sum_j \exp(z_j)} - y_i =$$

$$= \text{softmax}(z_i) - y_i \quad \text{(slide 48)}$$

So $\frac{\partial L}{\partial z^{(3)}} = \text{softmax}(z^{(3)}) - y = \hat{y} - y$

$$\nabla_{z^{(3)}} L = \text{softmax}(z^{(3)}) - y$$

And we continue by applying exactly the same rules as in question 1

2) As in Q1 the derivative of the sum is the sum of the derivatives ...

$\frac{\partial L}{\partial w}$ matrix $\mathbb{R}^{k_e \times k_{e-1}}$

$w^{(e)} = w^{(e)} - \eta \frac{\partial L}{\partial w}$