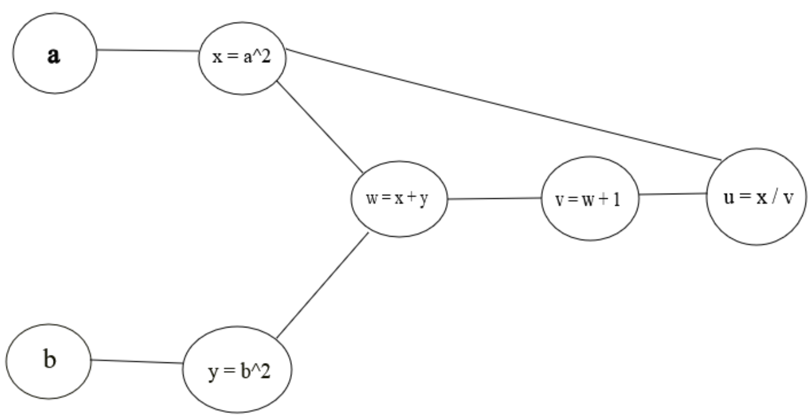
NNDL Exercise set 2 – Mathematical exercises

1.

a)



b)

$$\begin{aligned} z(a,b) &= \frac{a^2}{a^2+b^2+1} \\ &= \frac{x}{x+b^2+1} \\ &= \frac{x}{x+y+1} \\ &= \frac{x}{w+1} \\ &= \frac{x}{v} \\ &= u \end{aligned}$$

$$\begin{aligned} & x = a^2, \quad \frac{\partial x}{\partial a} = 2a \\ & y = b^2, \quad \frac{\partial y}{\partial b} = 2b \\ & w = x+y, \quad \frac{\partial w}{\partial x} = 1, \quad \frac{\partial w}{\partial y} = 1 \\ & v = w+1, \quad \frac{\partial v}{\partial w} = 1 \\ & u = \frac{x}{v}, \quad \frac{\partial u}{\partial x} = \frac{1}{v}, \quad \frac{\partial u}{\partial v} = -\frac{x}{v^2} \end{aligned}$$

c)

For $a = 2$, $b = 1$ we have

$$x = 4 \qquad y = 1 \qquad w = 5 \qquad v = 6 \qquad u = \frac{4}{6} = \frac{2}{3}.$$

Starting from the node u in the computational graph and going backwards using the chain rule, we obtain

$$\begin{aligned} \frac{\partial z}{\partial u} &= 1 \\ \frac{\partial z}{\partial v} &= \frac{\partial z}{\partial u}\frac{\partial u}{\partial v} = 1 \cdot \left(-\frac{x}{v^2}\right) = -\frac{1}{9} \\ \frac{\partial z}{\partial w} &= \frac{\partial z}{\partial v}\frac{\partial v}{\partial w} = -\frac{x}{v^2} \cdot 1 = -\frac{1}{9} \\ \frac{\partial z}{\partial y} &= \frac{\partial z}{\partial w}\frac{\partial w}{\partial y} = -\frac{x}{v^2} \cdot 1 = -\frac{1}{9} \\ \frac{\partial z}{\partial b} &= \frac{\partial z}{\partial y}\frac{\partial y}{\partial b} = -\frac{x}{v^2} \cdot 2b = -\frac{2a^2 b}{(a^2+b^2+1)^2} = -\frac{2}{9} \qquad \text{(analytical result)} \\ \frac{\partial z}{\partial x} &= \frac{\partial z}{\partial w}\frac{\partial w}{\partial x} + \frac{\partial z}{\partial u}\frac{\partial u}{\partial x} = -\frac{x}{v^2} \cdot 1 + 1 \cdot \frac{1}{v} = -\frac{1}{9} + \frac{1}{6} = \frac{1}{18} \\ \frac{\partial z}{\partial a} &= \frac{\partial z}{\partial x}\frac{\partial x}{\partial a} = \left(-\frac{x}{v^2} + \frac{1}{v}\right) \cdot 2a = \frac{2a(b^2+1)}{(a^2+b^2+1)^2} = \frac{2}{9} \qquad \text{(analytical result)} \end{aligned}$$

2.

We have a neuron $y(\mathbf{x})$ with M inputs, with weights $w_m \sim \mathcal{N}(0, \sigma^2)$ and independently distributed inputs $x_m \sim \mathcal{N}(0, 1)$.

a)

With no bias, before the activation function the neuron is

$$y = \sum_{m=1}^{M} w_m x_m. \tag{1}$$

The expectation is then

$$\mathrm{E}\left[\sum_{m=1}^{M} w_m x_m\right] = \sum_{m=1}^{M} \mathrm{E}[w_m x_m] = \sum_{i=1}^{M} \underbrace{\mathrm{E}[w_m]}_{=0} \underbrace{\mathrm{E}[x_m]}_{=0} = 0. \tag{2}$$

For two independent variables X and Y we have the result

$$\mathrm{Var}(XY) = \mathrm{E}[X]^2 \mathrm{Var}(Y) + \mathrm{E}[Y]^2 \mathrm{Var}(X) + \mathrm{Var}(X)\mathrm{Var}(Y), \tag{3}$$

with which we obtain

$$\begin{aligned}
\mathrm{Var}\left(\sum_{m=1}^{M} w_m x_m\right) &= \sum_{m=1}^{M} \mathrm{Var}(w_m x_m) \\
&= \sum_{m=1}^{M} \big(\mathrm{E}[w_m]^2 \mathrm{Var}(x_m) + \mathrm{E}[x_m]^2 \mathrm{Var}(w_m) + \mathrm{Var}(w_m)\mathrm{Var}(x_m)\big) \\
&= \sum_{m=1}^{M} \underbrace{\mathrm{Var}(w_m)}_{=\sigma^2} \underbrace{\mathrm{Var}(x_m)}_{=1} \\
&= \sum_{m=1}^{M} \sigma^2 \\
&= M\sigma^2.
\end{aligned} \tag{4}$$

b)

Now we have M neurons $y_1 = y$ in the first layer with $\mathrm{E}[y_1] = 0$ and $\mathrm{Var}(y_1) = M\sigma^2$. Assuming identity activation, the mean of a neuron in the second layer is

$$\mathrm{E}[y_2] = \mathrm{E}\left[\sum_{i=1}^{M} w_i y_{1,i}\right] = \sum_{i=1}^{M} \mathrm{E}[w_i]\mathrm{E}[y_{1,i}] = 0, \tag{5}$$

and the variance is

$$
\begin{aligned}
\mathrm{Var}(y_2) &= \mathrm{Var}\left(\sum_{i=1}^{M} w_i y_{1,i}\right) \\
&= \sum_{i=1}^{M} \mathrm{Var}(w_i y_{1,i}) \\
&= \sum_{i=1}^{M} \underbrace{\mathrm{Var}(w_i)}_{=\sigma^2} \mathrm{Var}(y_{1,i}) \\
&= \sum_{i=1}^{M} M\sigma^4 \\
&= M^2\sigma^4.
\end{aligned} \tag{6}
$$

c)

Let us denote the output of layer k by y_k. To retain the variance across layers, we must have

$$
\begin{aligned}
\mathrm{Var}(y_k) &= \mathrm{Var}(y_{k+1}) \\
&= \mathrm{Var}\left(\sum_{i=1}^{M} w_i y_k\right) \\
&= \sum_{i=1}^{M} \mathrm{Var}(w_i y_k) \\
&= \sum_{i=1}^{M} \mathrm{Var}(w_i)\mathrm{Var}(y_k) \\
&= \sum_{i=1}^{M} \sigma^2 \mathrm{Var}(y_k) \\
&= M\sigma^2 \mathrm{Var}(y_k).
\end{aligned} \tag{7}
$$

From the above we can deduce that

$$
M\sigma^2 = 1 \implies \sigma^2 = \frac{1}{M}. \tag{8}
$$

Having a value of σ^2 significantly larger/smaller than the above threshold leads to the value of the activation to increase/decrease with the number of the layers, leading to exploding/vanishing gradients. This leads to the network converging slowly or not converging at all.

d)

Using Eqs. (3) and (7) we find that the variance before the activation is

$$\begin{aligned}\mathrm{Var}(y_k) &= \sum_{m=1}^{M}\left(\underbrace{\mathrm{E}[w_m]^2}_{=0}\mathrm{Var}(x_m) + \mathrm{E}[x_m]^2\mathrm{Var}(w_m) + \mathrm{Var}(w_m)\mathrm{Var}(x_m)\right) \\ &= M\sigma^2(\mathrm{Var}(x) + \mathrm{E}[x]^2)\end{aligned} \tag{9}$$

The variance is defined as

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2. \tag{10}$$

And since $\mathrm{E}[x] = 0 \implies \mathrm{Var}(x) = \mathrm{E}[x^2]$, we find

$$\mathrm{Var}(y_k) = M\sigma^2\mathrm{E}[x^2]. \tag{11}$$

Now we have

$$\mathrm{E}[x^2] = \int_{-\infty}^{\infty} x^2 p(x)dx. \tag{12}$$

The input is ReLU applied to the output of the previous layer, e.g. $x = \max(0, y_{k-1})$, so we get

$$\begin{aligned}\mathrm{E}[x^2] &= \int_{0}^{\infty} y_{k-1}^2 p(y_{k-1})dy_{k-1} \\ &= \frac{1}{2}\int_{-\infty}^{\infty} y_{k-1}^2 p(y_{k-1})dy_{k-1} \\ &= \frac{1}{2}\mathrm{Var}(y_{k-1}).\end{aligned} \tag{13}$$

Therefore

$$\mathrm{Var}(y_k) = \frac{1}{2}M\sigma^2\mathrm{Var}(y_{k-1}), \tag{14}$$

from which we obtain

$$\frac{1}{2}M\sigma^2 = 1 \implies \sigma^2 = \frac{2}{M}. \tag{15}$$

Didn't have time to compute the expectation after the activation, but what can be said about it is that it is non-zero.