MULTI-LAYER NEURAL NETWORK WITH LINEAR ACTIVATION FUNCTION

$s(x) = x$

SUPPOSE WE HAVE A TWO LAYER N.N.:

$x \quad z_1 \quad y$

$W_1 \quad 2\times 4$
$b_1 \quad 2\times 1$

$W_2 \quad 2\times 2$
$b_2 \quad 2\times 1$

→ IT'S COMPLETELY GENERAL FOR ANY NUMBER OF LAYERS IN THE N.N.

$z_1 = s(W_1 x + b_1)$

$y = s(W_2 z_1 + b_2)$

BUT $s(x) = x \Rightarrow z_1 = s(W_1 x + b_1)$
$= W_1 x + b_1$

AND $y = s(W_2 z_1 + b_2) = W_2 z_1 + b_2$
$= W_2 W_1 x + W_2 b_1 + b_2$

IF I CALL $W = W_2 W_1 \quad [2\times 2 \cdot 2\times 4 = 2\times 4]$
$b = W_2 b_1 + b_2 \quad [2\times 2 \cdot 2\times 1 + 2\times 1 = 2\times 1]$

I OBTAIN $y = Wx + b$

WHICH CAN BE DESIGNED WITH A SINGLE LAYER N.N.

$s(x) = \lambda x$

$z_1 = \lambda W_1 x + \lambda b_1$

$y = \lambda^2 W_2 W_1 x + \lambda^2 W_2 b_1 + \lambda b_2$

$\Rightarrow W = \lambda^2 W_2 W_1$
$b = \lambda^2 W_2 b_1 + \lambda b_2$

$\Rightarrow$ SAME FOR ANY ACTIVATION FUNCTION

ASSUME $z_i \sim N(0, \sigma^2)$, WITH $\sigma << 1$

FOR WHICH OF THE FOLLOWING ACTIVATION FUNCTIONS IS A D.N.N. EQUIVALENT TO A LINEAR NETWORK FOR THE GIVEN DISTRIBUTION?

① $s(x) = \frac{1}{1+e^{-x}}$

② $s(x) = \tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^{\cancel{x}} - e^{-x}}{e^x + e^{-x}}$

③ $s(x) = \text{ReLU}(x) = \max(0, x)$

④ $s(x) = \text{SELU}(x) = \begin{cases} \lambda x & \text{IF } x > 0 \\ \alpha[e^x - 1] & \text{IF } x \leq 0 \end{cases}$

SINCE I HAVE INPUTS THAT ARE GAUSSIAN WITH ZERO MEAN AND VERY SMALL VARIABILITY, ALL THE FUNCTIONS WILL BE ACTIVATED FOR VALUES NEAR ZERO $\Rightarrow x \in [-\sigma, \sigma]$, $\sigma << 1$

SO I HAVE TO FIND A FUNCTION THAT ACTS AS A LINEAR FUNCTION AROUND ZERO

$\Rightarrow$ TAYLOR EXPANSION!

① $\frac{1}{1+e^{-x}} \sim \frac{1}{1+1-x} = \frac{1}{2-x}$

② $\tanh = \frac{\sinh}{\cosh} = \frac{x}{1} = x$

$= \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{\cancel{1} + x - \cancel{1} + x}{1 \cancel{+x} + 1 \cancel{-x}} = \frac{2x}{2} = x$

③ $\text{ReLU}(x) = \max(0, x)$ IT'S LINEAR ONLY FOR INPUTS THAT ARE IN THE POSITIVE INTERVAL $(0, \sigma]$

④ $\text{SELU}(x) = \begin{cases} \lambda x & x > 0 \\ \alpha[e^x - 1] \end{cases}$ $\rightarrow$ LINEAR $\forall \lambda \in \mathbb{R}$

$\rightarrow \alpha[\cancel{1} + x \cancel{-1}] = \alpha x$

$\Rightarrow$ LINEAR $\forall \alpha \in \mathbb{R}$

SO ACTIVATION FUNCTIONS

$$s(x) = \tanh(x)$$

$$s(x) = \mathrm{SELU}(x)$$

BOTH MAKES A DEEP NETWORK EQUIVALENT TO A LINEAR NETWORK

ALSO $s(x) = \mathrm{ReLU}(x)$ ACTS AS LINEAR BUT ONLY FOR POSITIVE VALUES OF x