1.

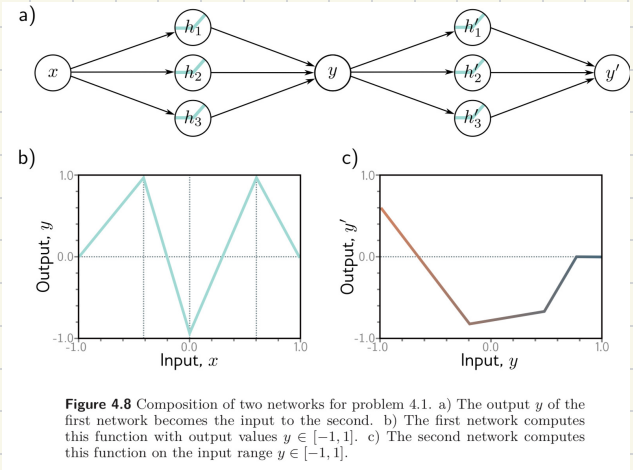
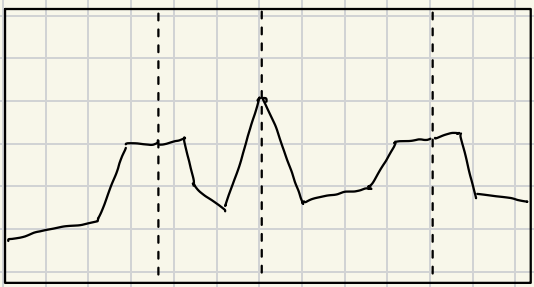


Figure 4.8 Composition of two networks for problem 4.1. a) The output y of the first network becomes the input to the second. b) The first network computes this function with output values $y \in [-1, 1]$. c) The second network computes this function on the input range $y \in [-1, 1]$.



3. Non-negative homogeneity of ReLU:

$\forall$ ReLU$(\alpha z) = \alpha$ ReLU(z)

$$\text{ReLU}[\beta_1 + \lambda_1 \cdot \Omega_1 \text{ReLU}[\beta_0 + \lambda_0 \cdot \Omega_0 x]]$$

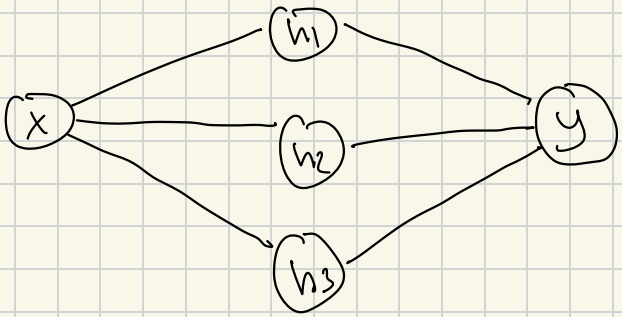
$$= \text{ReLU}\left[\lambda_0\lambda_1\left(\frac{\beta_1}{\lambda_0\lambda_1} + \frac{\lambda_1\Omega_1}{\lambda_0\lambda_1}\text{ReLU}\left[\lambda_0\left(\frac{\beta_0}{\lambda_0} + \frac{\lambda_0}{\lambda_0}\Omega_0 x\right)\right]\right)\right]$$

$$= \lambda_0\lambda_1 \text{ReLU}\left[\frac{\beta_1}{\lambda_0\lambda_1} + \frac{\Omega_1}{\lambda_0}\lambda_0 \text{ReLU}\left[\frac{\beta_0}{\lambda_0} + \Omega_0 x\right]\right]$$

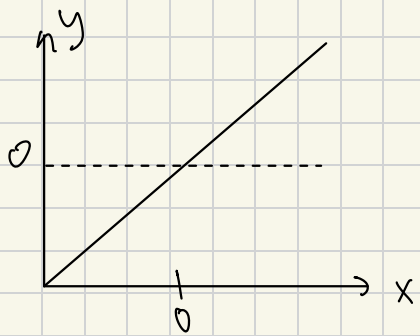
$$= \lambda_0\lambda_1 \text{ReLU}\left[\frac{\beta_1}{\lambda_0\lambda_1} + \Omega_1 \text{ReLU}\left[\frac{\beta_0}{\lambda_0} + \Omega_0 x\right]\right]$$

5. Depth → No. layers = 20
width → No. units / layer = 30

7.
$$\begin{aligned} y &= f[x, \phi] \\ &= \phi_0 + \phi_1 a[\theta_{10} + \theta_{11} x] + \phi_2 a[\theta_{20} + \theta_{21} x] + \phi_3 a[\theta_{30} + \theta_{31} x]. \end{aligned} \tag{3.1}$$

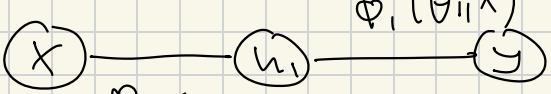


We want:



$-x \to -x$
$x \to x$

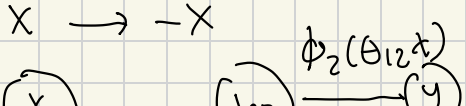
(1) $x \to x$



$\theta_{11} = 1$
$\phi_1 = 1$

$a(\theta_{11} x)$ ($-x$ gets clipped)

(2) $-x \to -x$



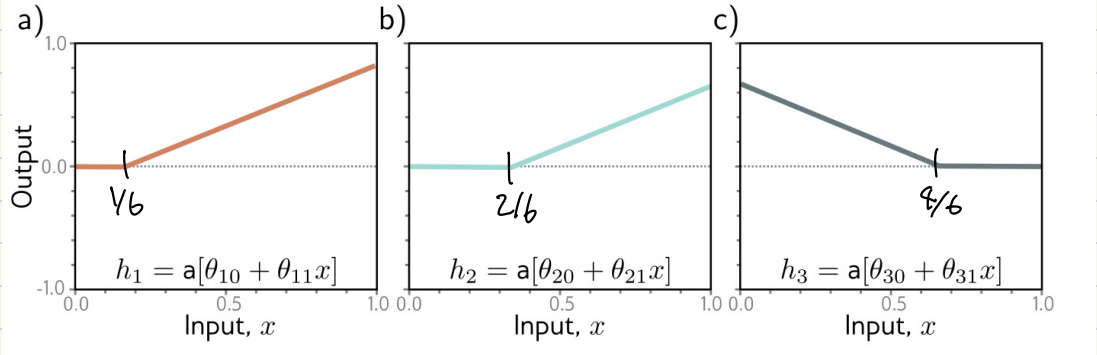
$\theta_{12} = -1$
$\phi_2 = -1$

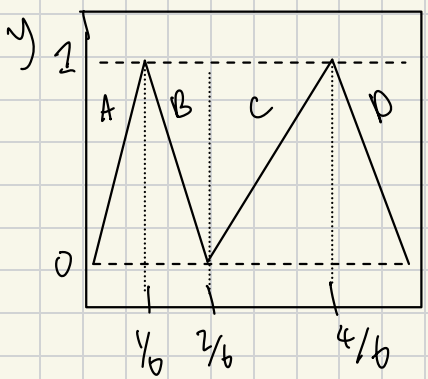
[other params can be set to zero]

(x gets clipped)

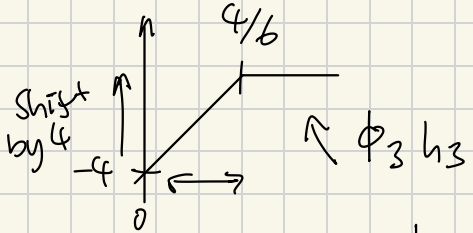
$y = \text{ReLU}[x] + \text{ReLU}[-x]$

8.





$y = \phi_0 + \phi_1 h_1 + \phi_2 h_2 + \phi_3 h_3$



$\rightarrow \phi_0 = 4$

A: only h_3 is active. $x=0, y=0, x=1/6, y=1 \rightarrow \phi_3 = -6$

B: h_1, h_3 active. Slope $\phi_1 - \phi_3$

we need a slope of $-6 = \phi_1 + 6 \rightarrow \phi_1 = -12$

C: All active. Slope $\phi_1 + \phi_2 - \phi_3$

we need a slope of $3 = -12 + \phi_2 + 6$

$3 = -6 + \phi_2$

$\rightarrow \phi_2 = 9$

$\therefore \phi_0 = 4, \phi_1 = -12, \phi_2 = 9, \phi_3 = -6$

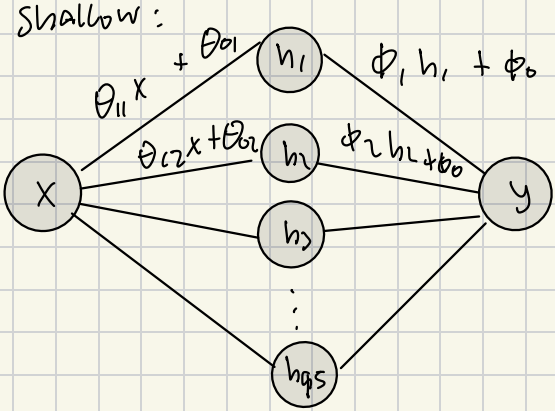
- Composing the network on itself creates 4 new regions in each region $\rightarrow$ $4 \times 4 = 16$ new regions. $= 4^2$
- Composing this K times is 4^{K+1} new regions.

4.

Imagine activations of two hidden units

- No matter how you scale and add them, you cannot create three linear regions that oscillate between 0 and 1.
- But this is possible to create 5 linear regions oscillating between 0 and 2 with 4 units (not shown).

11. Shallow:

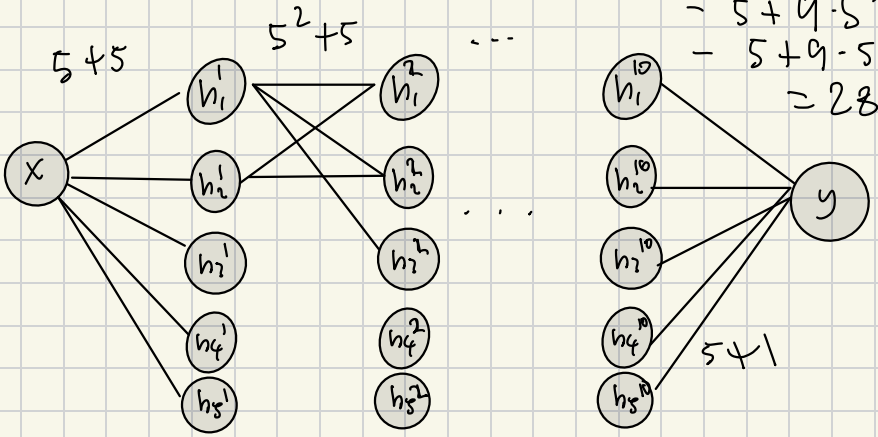


Contains

- 95 x 2 weights
- 95 + 1 biases

= 286 parameters.

Deep:



Contains

- $5+9\cdot5^2+5$ weights
- $5+9\cdot5+1$ biases

= 286 parameters

Shallow network has 96 linear regions

- Deep network has

$$N_r = \left(\frac{D}{D_i}+1\right)^{D_i(k-1)} \sum_{j=0}^{D_i} \binom{D}{j}$$

$D_i = 1$, $D = 5$, $k = 16$

$$N_r = (5+1)^4 = 6^4$$

Typically, the shallow network will run faster because the operations are parallelizable, whereas in the deep network, the outputs of each layer must be inputted into the following layers (sequential).