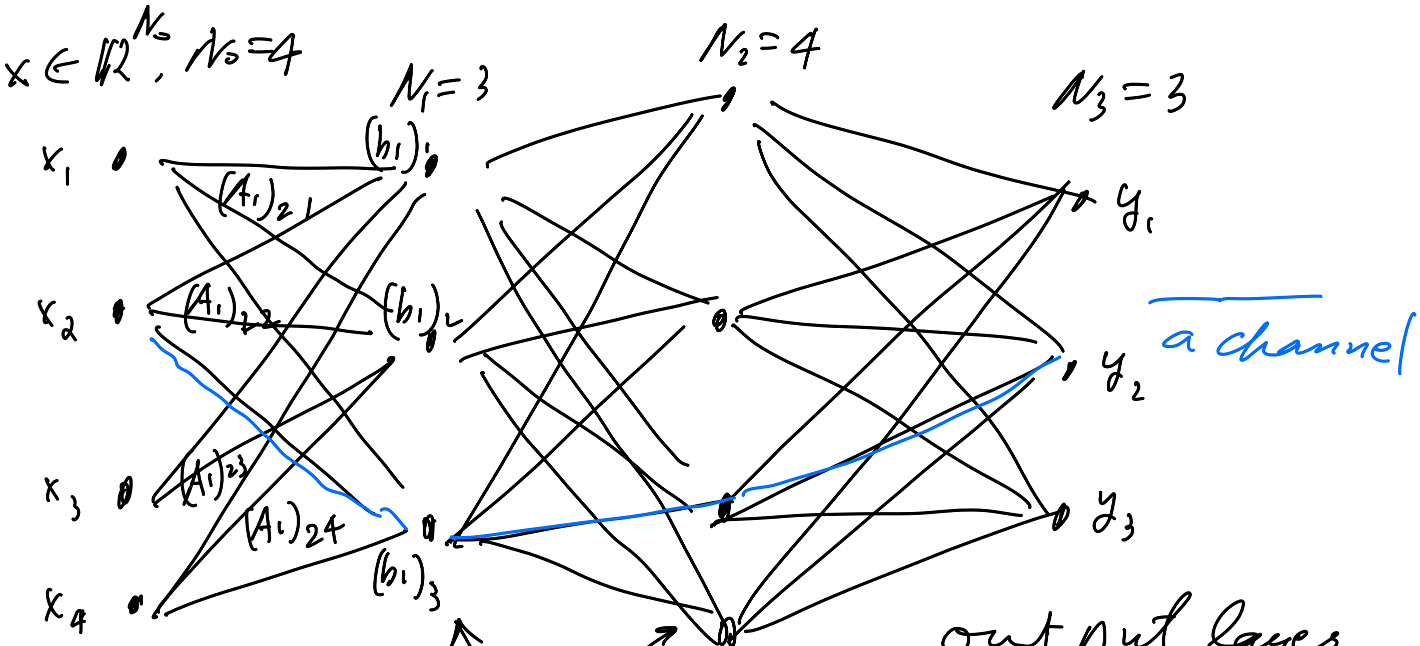
Lecture 2, Friday, April 5, 2024

Plan ① Some remarks on the definition of NNs
② Simple operations / constructions of NNs

Definition of a NN $\Phi : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$: [MLP (multilayer perceptron), or feedforward NN]

$$y = \Phi(x) = (W_L \circ \sigma \circ W_{L-1} \circ \cdots \circ W_2 \circ \sigma \circ W_1)(x)$$

$W_k : \mathbb{R}^{N_{k-1}} \to \mathbb{R}^{N_k}$, $W_k(x) = A_k x + b_k$, $A_k \in \mathbb{R}^{N_k \times N_{k-1}}$, $b_k \in \mathbb{R}^{N_k}$.



input layer $N_0 = 4$ hidden layers $N_1 = 3$, $N_2 = 4$ output layer $N_3 = 3$

$(A_k)_{ij}$ represents the weight associated with the edge between jth node in the $(k-1)$-th layer and the ith node in the k-th layer.

$\mathcal{L}(\Phi) = L$: depth (#layers: $L+1$, #hidden layers $L-1$).

$\mathcal{W}(\Phi) = \max_{0 \le k \le L} N_k$: width

$\mathcal{M}(\Phi) =$ # nonzero weights (entries of A_k, b_k for all k)
connectivity

$\mathcal{B}(\Phi) = \max_{1 \le k \le L} \max(\|A_k\|_\infty, \|b_k\|_\infty)$: weight magnitude

The NN architecture of Φ : $(N_0, \cdots, N_L)$.

Some remarks

- Definition using iteration

$\Phi(x) = y \quad x \in \mathbb{R}^{N_0},\ y \in \mathbb{R}^{N_L}.$

$$\begin{cases} X_0 = x \\ X_k = \sigma(A_k X_{k-1} + b_k), \quad k = 1, \dots, L-1 \\ X_L = A_L X_{L-1} + b_L \\ y = X_L \end{cases}$$

In general, we can allow σ to depend on k, i.e., we can replace σ by σ_k in the definition of X_k.

These NNs are called ~~feed-forward~~ NNs. More generally, they can be defined by <u>directed acyclic graphs</u>. Also called multilayer perceptron (MLP).

Add the def. of NNs using directed acyclic graphs

Also: $M(\Phi) \le \mathcal{L}(\Phi) W(\Phi)(W(\Phi)+1)$

Why: # of entries of $A_k \le W(\Phi)^2$.

# of entries of $b_k \le W(\Phi)$.

$k = 1, 2, \dots, L = \mathcal{L}(\Phi)$.

So, $M(\Phi) \le \mathcal{L}(\Phi) W(\Phi)^2 + \mathcal{L}(\Phi) W(\Phi)$

$= \mathcal{L}(\Phi) W(\Phi)(W(\Phi)+1)$.

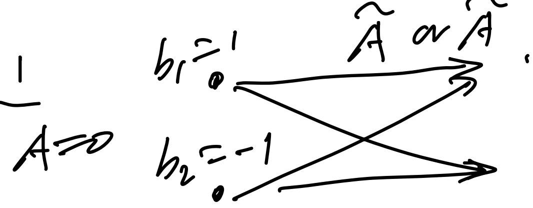
- We write $\Phi = NN(\sigma, N_0, \dots, N_L, A_1, \dots, A_L, b_1, \dots, b_L)$.

Φ is not uniquely determined by its weights!

$NN(\sigma, N_0, \dots, N_L, A_1, \dots, A_L, b_1, \dots, b_L) = NN(\sigma, N_0, \dots, N_L, \tilde{A}_1, \dots, \tilde{A}_L, \tilde{b}_1, \dots, \tilde{b}_L)$

$\not\Rightarrow$ all $A_k = \tilde{A}_k,\ b_k = \tilde{b}_k \quad (1 \le k \le L)$.

<u>Example 1</u>



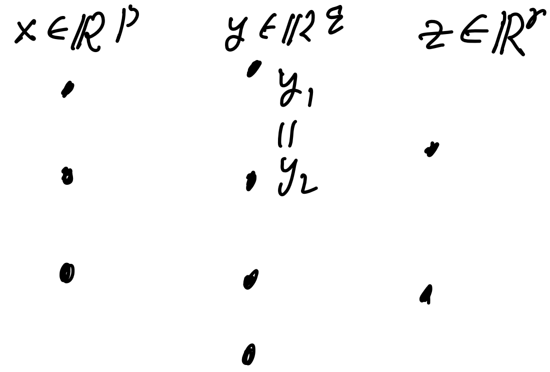
$\tilde{A} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$

$\tilde{\tilde{A}} = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix}$

$\tilde{A} b = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. $\tilde{\tilde{A}} b = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

$\tilde{A} \ne \tilde{\tilde{A}}$ but the output is always the same.

Example 2



$y = Ax + b$, $z = By + c$. choose row 1 of A = row 2 of A, $b = \begin{bmatrix} b_1 \\ \vdots \\ b_q \end{bmatrix}$, $b_1 = b_2$. Then $y_1 = y_2$.

$$z = By + c = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1q} \\ B_{21} & B_{22} & \cdots & B_{2q} \\ \cdots & \cdots & \cdots & \cdots \\ B_{r1} & B_{r2} & \cdots & B_{rq} \end{bmatrix} \begin{bmatrix} y_1 \\ y_1 \\ y_3 \\ \vdots \\ y_q \end{bmatrix} + \begin{bmatrix} c_1 \\ \vdots \\ c_q \end{bmatrix}$$

$$= \begin{bmatrix} (B_{11}+B_{12})\,y_1 + \cdots \\ (B_{21}+B_{22})\,y_1 + \cdots \\ (B_{31}+B_{32})\,y_3 + \cdots \\ \cdots \\ (B_{r1}+B_{r2})\,y_q + \cdots \end{bmatrix} + \begin{bmatrix} c_1 \\ \vdots \\ c_q \end{bmatrix}$$

where all "$\cdots$" are indep. of B_{i1}, B_{i2} $(1 \le i \le r)$ and y_1, y_2.

So, if B and $\tilde{B}$ have the same first 2 columns thus $((A,b),(B,c))$ and $((A,b),(\tilde{B},c))$ define 2 NNs that are the same map. But $B \neq \tilde{B}$ possible.

This example is from C. Fefferman (1994) where the concept and conditions of isomorphic sets of parameters are given, and where it is proven that under generic conditions, sets of parameters defining the same NN (as function) are isomorphic. More recent works (Berner et al. NeurIPS 2019 & Petersen et al. Found. Comp. Math 2021) study the related issues, e.g. the inverse stability $\Phi(\theta) \to \Phi(\tilde{\theta}) \Rightarrow \theta \to \tilde{\theta}$, and the structures of NN functions generated by parameters of same size.

Some simple operations / constructions of NNs

Concatenation / Composition. Let Φ^1 and Φ^2 be two NNs with the same activation function σ:

$$\Phi^1 = NN(\sigma, N_0^{(1)}, \dots, N_{L_1}^{(1)}, (A_1^{(1)}, b_1^{(1)}), \dots (A_{L_1}^{(1)}, b_{L_1}^{(1)})),$$
$$\Phi^2 = NN(\sigma, N_0^{(2)}, \dots, N_{L_2}^{(2)}, (A_1^{(2)}, b_1^{(2)}), \dots, (A_{L_2}^{(2)}, b_{L_2}^{(2)})).$$

Assume $N_0^{(2)} = N_{L_1}^{(1)}$, i.e.,

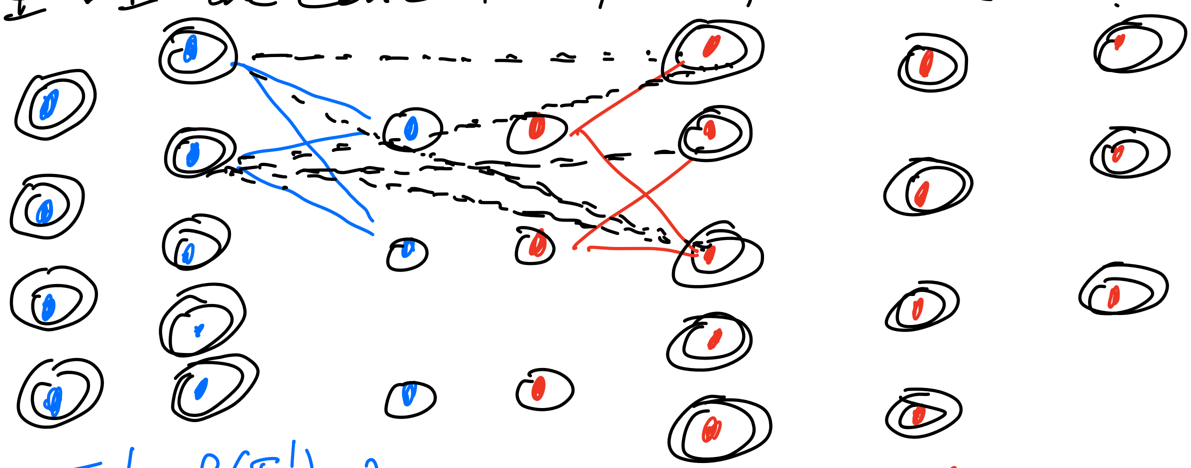
the input dim. of Φ^2 = the output dim. of Φ^1.

Define

$$\Phi^2 \odot \Phi^1 = NN(\sigma, N_0^{(1)}, \dots, N_{L_1-1}^{(1)}, N_1^{(2)}, \dots, N_{L_2}^{(2)}, (A_1^{(1)}, b_1^{(1)}), \dots (A_{L_1-1}^{(1)}, b_{L_1-1}^{(1)}),$$
$$(A_1^{(2)} A_{L_1}^{(1)}, A_1^{(2)} b_{L_1}^{(1)} + b_1^{(2)}), (A_2^{(2)}, b_2^{(2)}), \dots (A_{L_2}^{(2)}, b_{L_2}^{(2)})).$$

Then, this is a NN with the same activation function σ and with the depth $\mathcal{L}(\Phi^2 \bullet \Phi^1) = \mathcal{L}(\Phi^1) + \mathcal{L}(\Phi^2) - 1$.

Call $\Phi^2 \bullet \Phi^1$ the concatenation of Φ^1 and Φ^2.



Φ^1 $\mathcal{L}(\Phi^1) = 2$

Φ^2, $\mathcal{L}(\Phi^2) = 3$

- $\Phi^2 \odot \Phi^1$ $\mathcal{L}(\Phi^{(2)} \odot \Phi^{(1)}) = \mathcal{L}(\Phi^{(1)}) + \mathcal{L}(\Phi^{(2)}) - 1$
- Remove the output layer of Φ^1 and the input layer of Φ^2 since they are not activated (no application of σ)

We have

$$\Phi^2 \odot \Phi^1 = \Phi^2 \circ \Phi^1,$$

i.e., concatenation = composition. This follows from the fact that for any $x^{(L_1-1)} \in \mathbb{R}^{N^{(1)}_{L_1-1}}$,

$$A_1^{(2)} A_{L_1}^{(1)} x^{(L_1-1)} + A_1^{(2)} b_{L_1}^{(1)} + b_1^{(2)} = A_1^{(2)} \underbrace{(A_L^{(1)} x^{(L_1-1)} + b_{L_1}^{(1)})}_{\in \mathbb{R}^{N^{(1)}_{L_1}} = \mathbb{R}^{N_0^{(2)}}} + b_1^{(2)} \in \mathbb{R}^{N_0^{(2)}}.$$

Parallelization

Given two σ-NNs

$$\Phi^1 = NN(\sigma, N_0^{(1)}, \dots, N_{L_1}^{(1)}, (A_1^{(1)}, b_1^{(1)}), \dots (A_{L_1}^{(1)}, b_{L_1}^{(1)})),$$

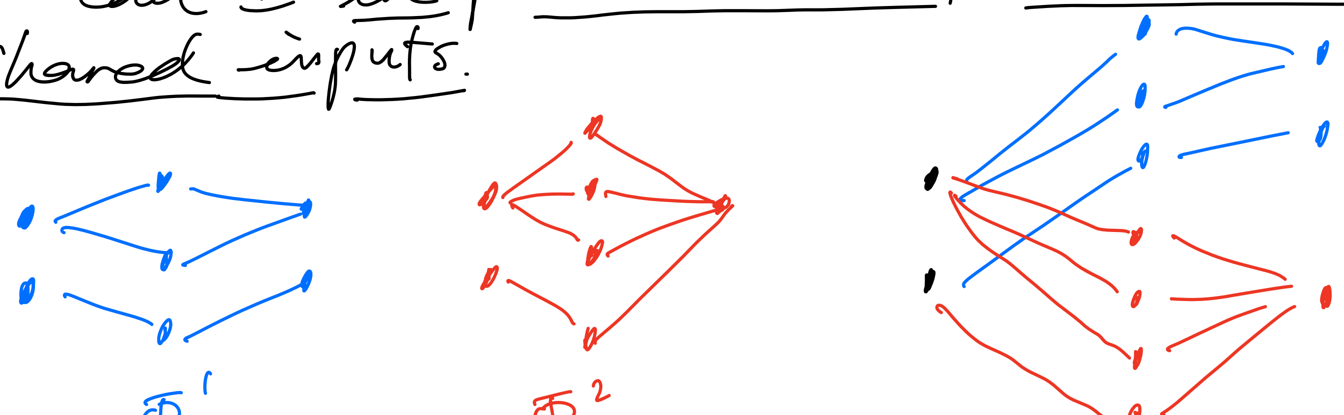
$$\Phi^2 = NN(\sigma, N_0^{(2)}, \dots, N_{L_2}^{(2)}, (A_1^{(2)}, b_1^{(2)}), \dots, (A_{L_2}^{(2)}, b_{L_2}^{(2)})).$$

Assume $L_1 = L_2 = L$ and $N_0^{(1)} = N_0^{(2)} = N_0$ (i.e., same depth and same input dim.). Define

$$\Phi = P(\Phi^1, \Phi^2) = NN\Big(\sigma, N_0, N_1^{(1)} + N_1^{(2)}, \dots, N_1^{(L)} + N_2^{(L)},$$
$$\left(\begin{bmatrix} A_1^{(1)} \\ A_1^{(2)} \end{bmatrix}, \begin{bmatrix} b_1^{(1)} \\ b_1^{(2)} \end{bmatrix}\right), \left(\begin{bmatrix} A_2^{(1)} & 0 \\ 0 & A_2^{(2)} \end{bmatrix}, \begin{bmatrix} b_2^{(1)} \\ b_2^{(2)} \end{bmatrix}\right), \dots \left(\begin{bmatrix} A_L^{(1)} & 0 \\ 0 & A_L^{(2)} \end{bmatrix}, \begin{bmatrix} b_L^{(1)} \\ b_L^{(2)} \end{bmatrix}\right)\Big).$$

Then Φ is a σ-NN with $\mathcal{L}(\Phi) = L$. (check it!)

Call Φ the parallelization of Φ^1 and Φ^2 with shared inputs.



Φ^1 $\quad$ Φ^2 $\quad$ $P(\Phi^1, \Phi^2)$.

We have

$$(P(\Phi^1, \Phi^2))(x) = \begin{bmatrix} \Phi^1(x) \\ \Phi^2(x) \end{bmatrix} \quad \forall x \in \mathbb{R}^{N_0}.$$

Given two σ-NNs Φ^1 and Φ^2 as above.
Assume $L_1 = L_2 =: L$. Define

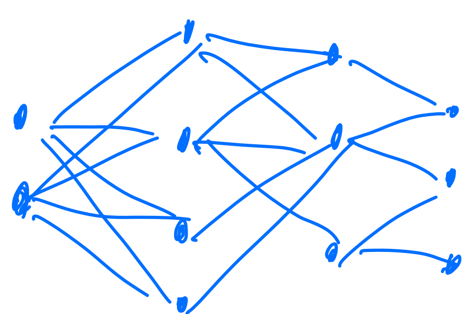
$$\Phi = FP(\Phi^1, \Phi^2) = NN\Big(\sigma, N_0^{(1)}+N_0^{(2)}, \dots, N_{L-1}^{(1)}+N_{L-1}^{(2)}, N_L^{(1)}+N_L^{(2)};$$
$$\left(\begin{bmatrix} A_1^{(1)} & 0 \\ 0 & A_2^{(1)} \end{bmatrix}, \begin{bmatrix} b_1^{(1)} \\ b_1^{(2)} \end{bmatrix}\right), \dots, \left(\begin{bmatrix} A_L^{(1)} & 0 \\ 0 & A_L^{(2)} \end{bmatrix}, \begin{bmatrix} b_L^{(1)} \\ b_L^{(2)} \end{bmatrix}\right)\Big).$$

Then Φ is a σ-NN with $\mathcal{L}(\Phi) = L$, the input dimension $N_0^{(1)}+N_0^{(2)}$, and the output dimension $N_L^{(1)}+N_L^{(2)}$. <u>We call Φ the parallelization of Φ^1 and Φ^2 without shared inputs</u>. We have

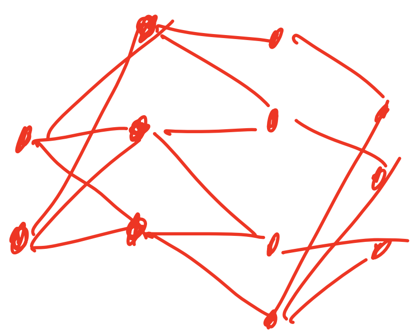
$$FP(\Phi^1(x), \Phi^2(y)) = (\Phi^1(x), \Phi^2(y)) \quad \forall (x,y) \in \mathbb{R}^{N_0^{(1)}} \times \mathbb{R}^{N_0^{(2)}}.$$

We denote by $NN(\sigma, L, N_0, N_L)$ the class of all NNs with the activation function σ, the depth L, the input dimension N_0, and the output dimension N_L.

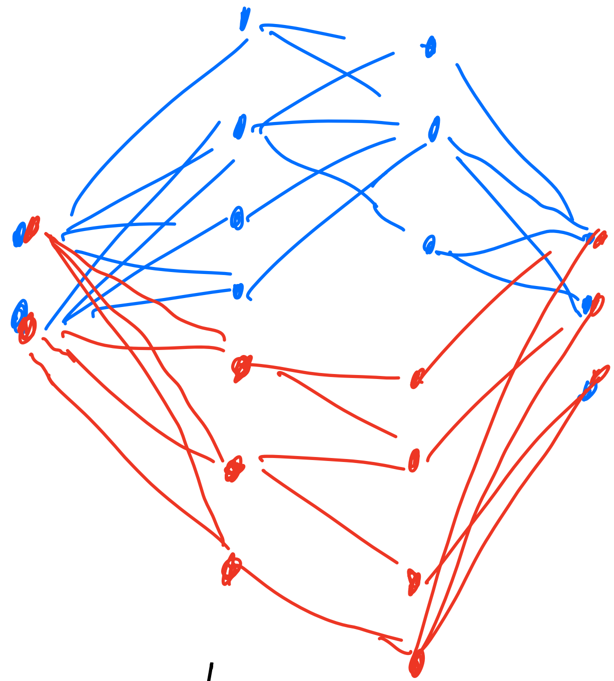
<u>Parallelization again</u>. Let $\Phi^i \in NN(\sigma, L, N_0, N_L)$ be given by $\Phi^i = NN(\sigma, N_0^{(i)}, N_1^{(i)}, \dots, N_L^{(i)}, (A_1^{(i)}, b_1^{(i)}), \dots, (A_L^{(i)}, b_L^{(i)}))$, $i=1,2$, with $N_0^{(1)} = N_0^{(2)} = N_0$, $N_L^{(1)} = N_L^{(2)} = N_L$. We construct $\Phi^1 \oplus \Phi^2 \in NN(\sigma, L, N_0, N_L)$ so that $(\Phi^1 \oplus \Phi^2)(x) = \Phi^1(x) + \Phi^2(x)$ $\forall x \in \mathbb{R}^{N_0}$.



Φ^1



Φ^2



$$\Phi^1 \oplus \Phi^2 = \Phi^1 + \Phi^2.$$

We define

$$\Phi^1 \oplus \Phi^2 = NN\big(\sigma, N_0, N_1^{(1)} + N_1^{(2)}, \dots, N_{L-1}^{(1)} + N_{L-1}^{(1)}, N_L, (\hat{A}_1, \hat{b}_1), \dots, (\hat{A}_L, \hat{b}_L)\big),$$

where

$$\hat{A}_1 = \begin{bmatrix} A_1^{(1)} \\ A_2^{(1)} \end{bmatrix}, \quad \hat{b}_1 = \begin{bmatrix} b_1^{(1)} \\ b_1^{(2)} \end{bmatrix}$$

$$\hat{A}_j = \begin{bmatrix} A_j^{(1)} & 0 \\ 0 & A_j^{(2)} \end{bmatrix}, \quad \hat{b}_j = \begin{bmatrix} b_j^{(1)} \\ b_j^{(2)} \end{bmatrix}, \quad j = 2, \dots, L-1 \quad (\text{if } L \geq 3)$$

$$\hat{A}_L = [A_L^{(1)} \; A_L^{(2)}] \quad \hat{b}_L = b_L^{(1)} + b_L^{(2)}.$$

One verifies that $(\Phi^1 \oplus \Phi^2)(x) = \Phi^1(x) + \Phi^2(x) \;\; \forall x \in \mathbb{R}^{N_0}$.

We denote by $NN(\sigma, N_0, \dots, N_L)$ the class of σ-NNs with depth L and dimension of the layers $N_0, \dots, N_L$.

Shifted dilates Given a NN $\Phi \in NN(\sigma, N_0, \dots, N_L)$. Also, given $\alpha \in \mathbb{R}$ and $\xi \in \mathbb{R}^{N_0}$. Define $T(x) = S(\alpha x + \xi)$ $\forall x \in \mathbb{R}^{N_0}$. Then T is also a NN. To see this, assume

$\Phi = NN(\sigma, N_0, \cdots, N_L, (A_1, b_1), \cdots, (A_L, b_L))$. Define

$\hat{A}_1 = \alpha A_1$, $\hat{b}_1 = A_1 \xi + b_1$, and $\hat{A}_k = A_k$, $\hat{b}_k = b_k$, $k = 2, \cdots, L$.

Then $T = NN(\sigma, N_0, \cdots, N_L, (\hat{A}_1, \hat{b}_1), \cdots (\hat{A}_L, \hat{b}_L))$.

From the above constructions, we have:

Proposition $NN(\sigma, L, N_0, N_L)$ is a vector space, a subspace of the vector space of all maps from $\mathbb{R}^{N_0}$ to $\mathbb{R}^{N_L}$.

QED.

Note: If we fix the architecture, $NN(\sigma, L, N_0, \cdots N_L)$ is not a vector space. as in general the addition is not closed. (Find an example!)

Question:

$$NN(\sigma, L, N_0 = d, N_L = m) \subseteq NN(\sigma, L+1, N_0 = d, N_{L+1} = m)?$$

Question: Let Φ^1, Φ^2 be two NNs with the same activation function σ, the same input dimension N_0, and the same 1-dimensional output $N_L = 1$. Then, is there a NN with σ, N_0, $N_L = 1$ such that it is the same as $\Phi^1(x) \cdot \Phi^2(x)$ $\forall x \in \mathbb{R}^{N_0}$?

Note: If $\sigma = \mathrm{ReLU}$, then we have more operations. In particular, Augmentation!