Lecture 1. Wed., 4/3/2024

Math 217: Topics in Applied Mathematics

Mathematics of Artificial Neural Networks / with possible extension to general machine learning

About the course: goals, style, expectation, etc.

- This is a math graduate topic course. It is a combination of research oriented and subject learning course. The selection of the topics to be covered is biased by ~~the~~ the instructor's own research interest. (See below for a tentative list of topics to be covered.)
- My plan is to look into the mathematical aspects of artificial neural networks (NNW) and also to discuss some related research (mainly my own research) on NNW dynamical systems and partial differential equations with application to physical and biological sciences (molecular and cellular biology, charged systems, materials, soft matter, etc.).
- I will not use any textbooks (as there are no such books meeting my interest and the goal of this course). Instead, I will mainly give lectures on some published papers. There will be a lot of proofs, and some of them can be technical. I encourage discussions.
- The prerequisites of this course include some measure theory and graduate level numerical analysis. Functional analysis, probability theory, dynamical systems, and partial differential equations are useful but not required.
- There will be no exams but there will be possibly some reading assignments. There will be a course

project. It should be brief and with some minimal work. Guidelines and some suggestions of project topics will be provided later.

Students can design their own projects. If so, discussions with the instructor are encouraged.

No previously completed project or project irrelavant to the course will be accepted.

The course project is due 3:00 pm, Wednesday, June 12, 2024.

A tentative outline of topics

1. Basics of neural networks
   - definition, terminologies, etc. an example: a 3-layer net
   - activation functions
   - Nonuniqueness of NN representations
   - operations/constructions of NNs
2. Approximation Theory
   - Universal approximation theorems
   - Shallow NN approximations, error bounds
   - Deep NN approximations, error bounds
3. Training NNs:
   - Loss functions
   - back propagation. differentiations
   - Optimization algorithm. Stoch. grad. descent, etc. Convergence analysis
4. Neural networks for dynamical systems and PDE
   - Recurrent NNs for dynamical systems
   - Parameter-solution networks for learning steady states and their stability
   - Deep Ritz / PINN, their variations
   - Fixed-point interations with NNW

Basics of Neural Networks (NNs)

- Definition of NNs
- Activation functions
- Loss functions, and their minimization

<u>Definition</u> A NN is a map $\Phi: \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ given by

$$\Phi = \begin{cases} W_1 & \text{if } L = 1, \\ W_L \circ \sigma \circ W_{L-1} \circ \cdots \circ W_2 \circ \sigma \circ W_1 & \text{if } L \geq 2. \end{cases}$$

where

- $\sigma: \mathbb{R} \to \mathbb{R}$ is given;
- all $L \in \mathbb{N}$, and $N_0, N_1, \dots, N_L \in \mathbb{N}$;
- For each $k \in \{1, \dots, L\}$, $W_k: \mathbb{R}^{N_{k-1}} \to \mathbb{R}^{N_k}$ is an affine map, given by

$$W_k(x) = A_k x + b_k, \quad \forall x \in \mathbb{R}^{N_{k-1}}.$$

All $A_k \in \mathbb{R}^{N_k \times N_{k-1}}$ are matrices and all $b_k \in \mathbb{R}^{N_k}$ are vectors.

Here and below,

$$\sigma(y) = (\sigma(y_1), \dots, \sigma(y_m)) \quad \text{if } y = (y_1, \dots, y_m) \in \mathbb{R}^m.$$

<u>Notations and terminologies</u>

$\sigma: \mathbb{R} \to \mathbb{R}$ an activation function

N_0: input dimension

N_L: output dimension

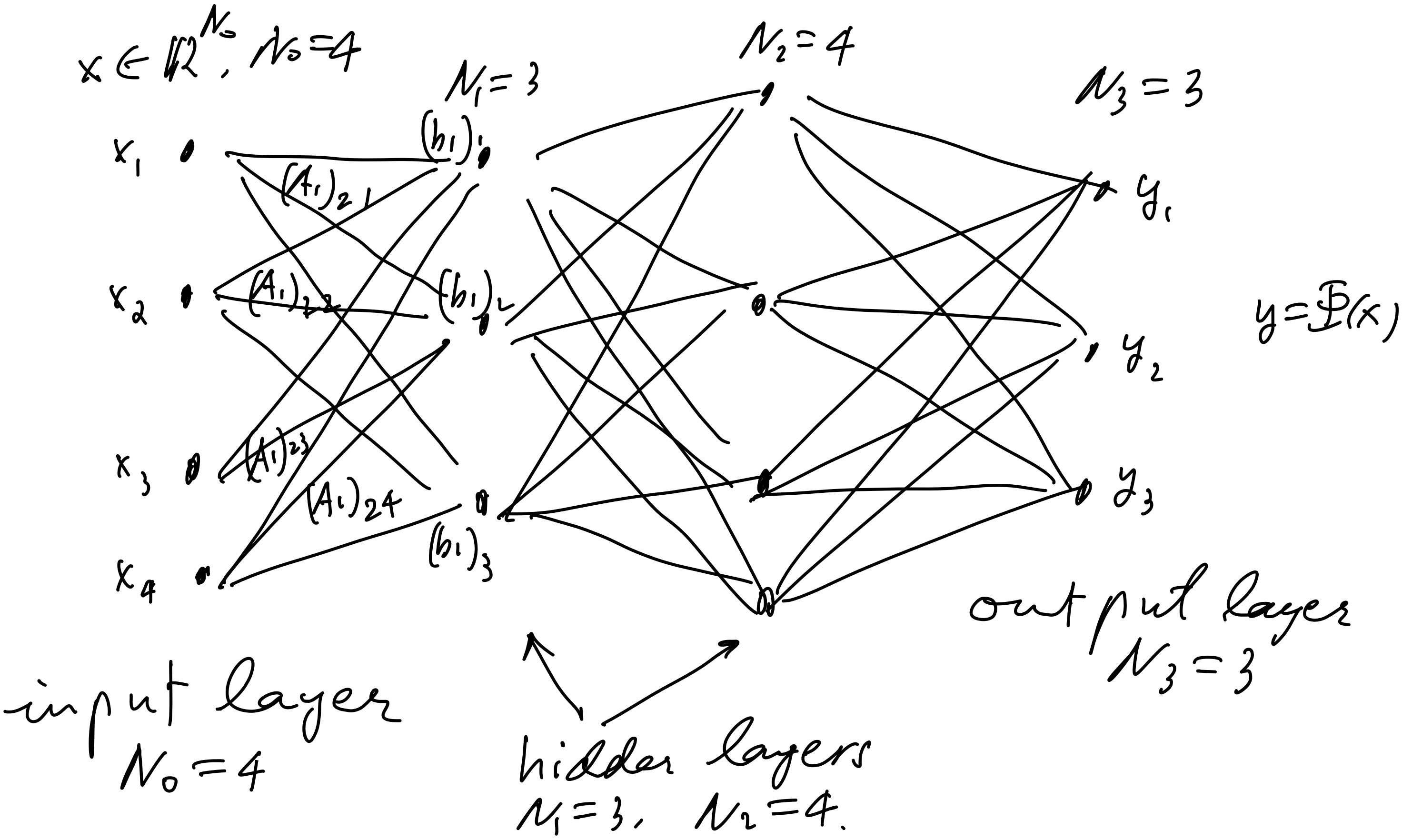
$L-1$: number of hidden layers

b_k: bias (vector), or threshold (vector)

$\mathcal{L}(\Phi) := L$: the depth of the NN
$= \#$ affine transforms

$\mathcal{W}(\Phi) = \max_{k=0,1,\cdots,L} N_k$: the width of the NN

$\mathcal{M}(\Phi)$: the connectivity $\mathcal{M}(\Phi) :=$ total number of nonzero entries in A_k (for all k) and nonzero components in b_k (for all k).



$$\left[A_1 \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + b_1 \right]_2 = \begin{bmatrix} (A_1)_{11} x_1 + (A_1)_{12} x_2 + (A_1)_{13} x_3 + (A_1)_{14} x_4 + (b_1)_1 \\ (A_1)_{21} x_1 + (A_1)_{22} x_2 + (A_1)_{23} x_3 + (A_1)_{24} x_4 + (b_1)_2 \\ \cdots \\ \cdots \end{bmatrix}_2$$

$$= (A_1)_{21} x_1 + (A_1)_{22} x_2 + (A_1)_{23} x_3 + (A_1)_{24} x_4 + (b_1)_2$$

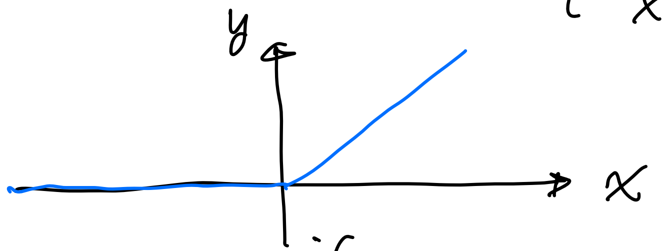
$(A_k)_{ij}$ represents the weight associated with the edge betweed jth node in the $(k-1)$-th layer and the ith node in the kth layer.

(Same remarks on the def. of NNs are in lecture 2.)

The activation function σ (Or response function)

ReLU (= Rectified Linear Unit)

$$\sigma(x) = \max(x,0) = \begin{cases} 0 & \text{if } x<0 \\ x & \text{if } x \geq 0 \end{cases}$$



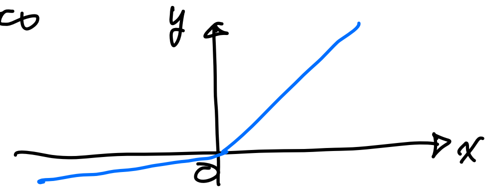
$$\sigma'(x) = \begin{cases} 0 & \text{if } x<0 \\ 1 & \text{if } x>0 \end{cases} = H(x) \quad \left(\begin{array}{c}\text{Heaviside}\\ \text{function}\end{array}\right)$$

Weakly differentiable:

$$\int_{-\infty}^{\infty} \sigma(x)\,\varphi'(x)\,dx = -\int_{-\infty}^{\infty} H(x)\,\varphi(x)\,dx \quad \forall \varphi \in C_c^1(\mathbb{R})$$

(Exercise: Prove it!)

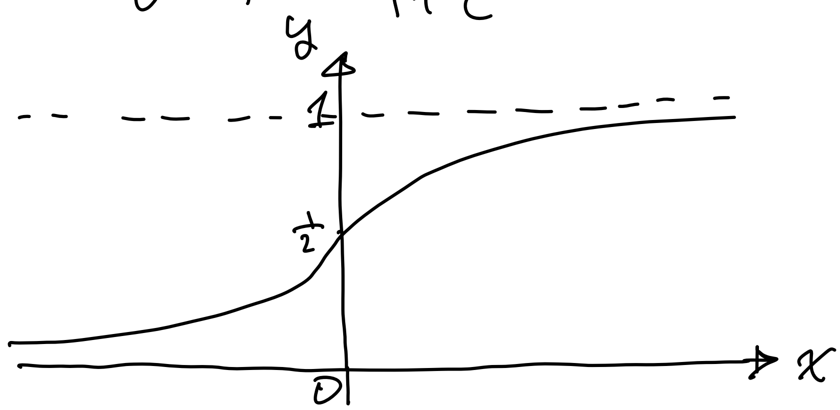
The leaky ReLU



The standard sigmoid function (or the logistic function)

$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}$$

$\sigma' > 0$, $\sigma \in C^\infty(\mathbb{R})$



$\sigma(+\infty) = 1$

$\sigma(-\infty) = 0$

$\sigma'(x) = \sigma(x)(1-\sigma(x))$

A general sigmoid function ($c>0$)

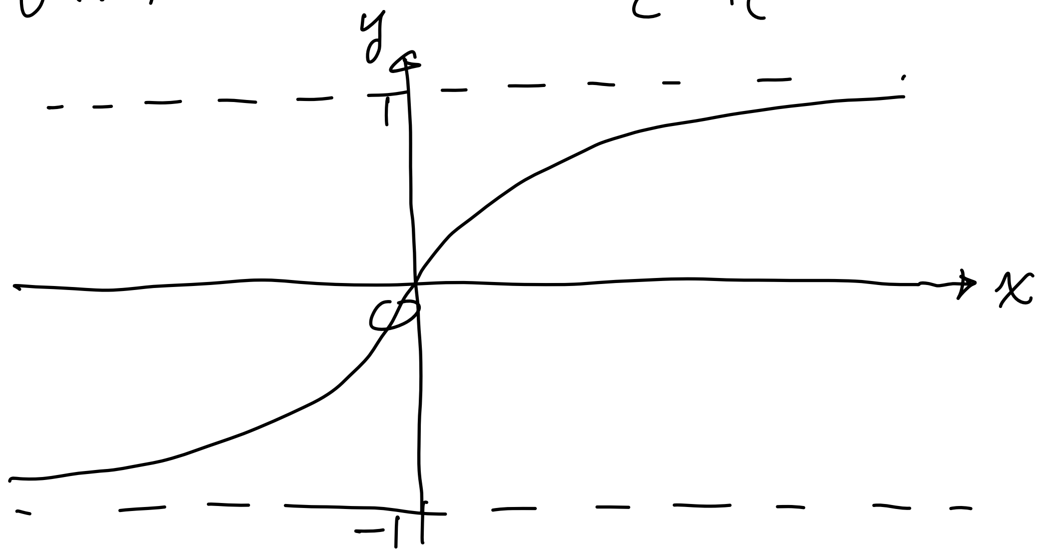
$$\sigma_c(x) = \frac{1}{1+e^{-cx}}$$

$$\lim_{c\to+\infty} \sigma_c(x) = H(x) = \begin{cases} 1 & \text{if } x>0, \\ 0 & \text{if } x<0, \end{cases} \quad x \neq 0.$$

$$\sigma_c(0) = \frac{1}{2}$$

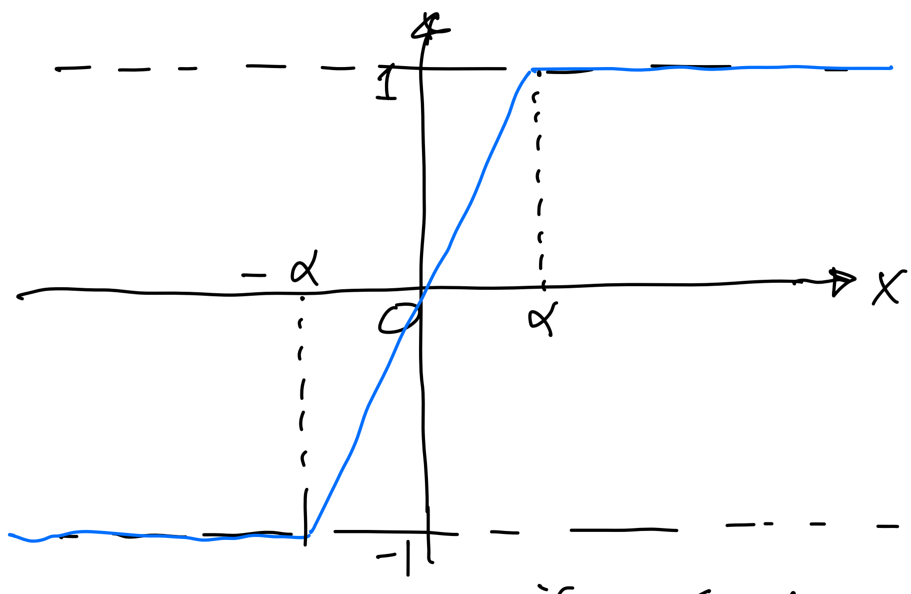
The hyperbolic tangent

$$\sigma(x) = \tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}} = 2\sigma_2(x) - 1$$



$$\sigma'(x) = 1 - (\sigma(x))^2$$

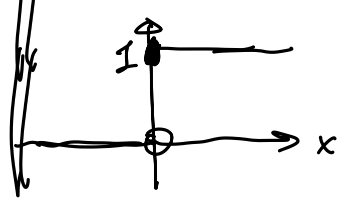
A piecewise linear function $(\alpha > 0)$



$$\sigma_\alpha(x) = \begin{cases} -1 & \text{if } x \le -\alpha \\ \frac{x}{\alpha} & \text{if } -\alpha < x < \alpha \\ 1 & \text{if } x \ge \alpha \end{cases}$$

$$\lim_{\alpha \to 0} \sigma_\alpha(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \end{cases} \qquad x \neq 0$$

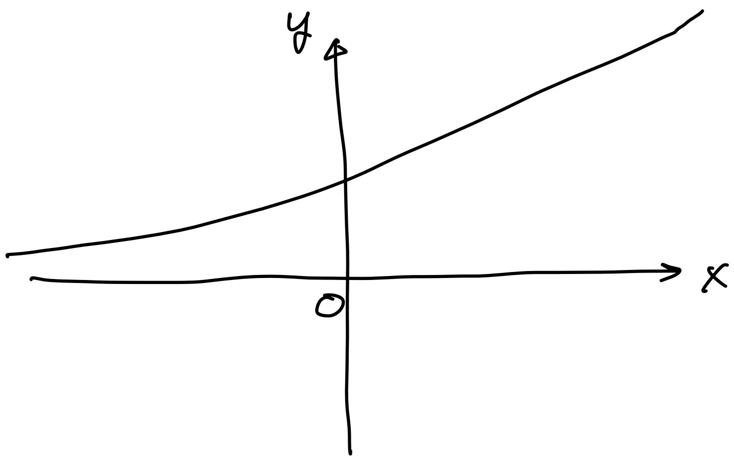
An important and originally used one: Heaviside function



$\sigma_\alpha(0)=0$.

The soft plus function

$$\sigma(x)=sp(x)=\ln(1+e^x)$$



$sp(x)\approx x$ for $x>>1$.

$sp(x)\approx 0$ for $x<<-1$

$sp(x)-sp(-x)=x$.

$sp'(x)=\dfrac{1}{1+e^{-x}}$

$sp^{-1}(x)=\log(e^x-1)$

Related: The Dirac δ-function $\delta=\delta(x)$.

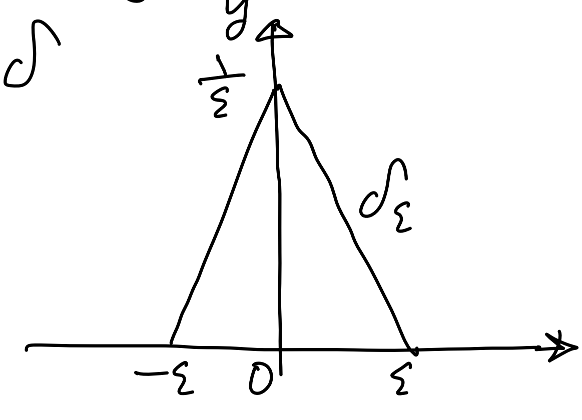
A physicists' version: $\delta(x)=0\ \forall x\neq 0$, $\delta(0)=+\infty$, and $\int_{-\infty}^{\infty}\delta(x)\,dx=1$.

A mathematician's version. δ is a distribution defined by $\langle\delta,\phi\rangle=\phi(0)$ for any $\phi\in C_c(\mathbb{R})$

Approximation of δ: $\lim_{\varepsilon\to 0}\delta_\varepsilon=\delta$

i.e., $\displaystyle\lim_{\varepsilon\to 0^+}\int_{-\infty}^{\infty}\delta_\varepsilon(x)\phi(x)\,dx=\phi(0)$

$=\langle\delta,\phi\rangle\quad\forall\phi\in C_c(\mathbb{R})$.



Note: $H'=\delta$.

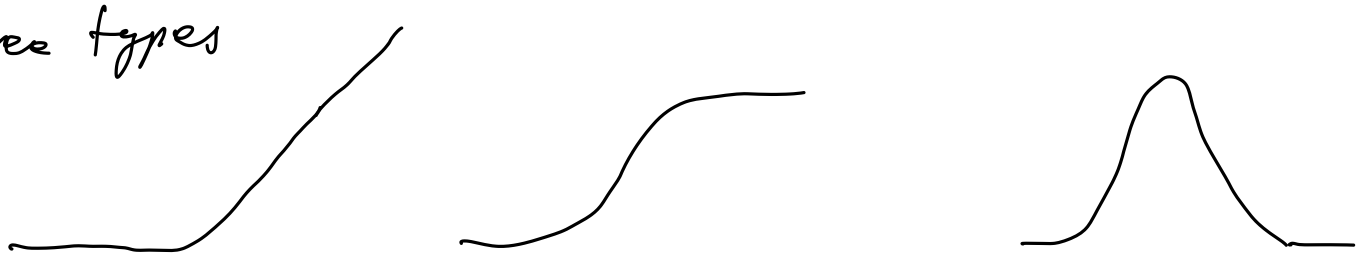
i.e. $\displaystyle\int_{-\infty}^{\infty}H'(x)\,\phi(x)\,dx=\langle\delta,\phi\rangle=\phi(0)\quad\forall\phi\in C_c^1(\mathbb{R})$

(Exercise: Prove it.)

A Gaussian activation function

$$\sigma(x) = a e^{-b(x-c)^2} \qquad a, b > 0, \; c \in \mathbb{R}.$$

Three types



ReLU type Sigmoidal Bumps

sharp versions, smoothened versions, etc

Loss functions (Or: Cost Functions)

(More on this topic in the part of training NNs.)

Let $\Omega \subseteq \mathbb{R}^{N_0}$. Given $f: \Omega \to \mathbb{R}^{N_L}$, we approximate f by the previously defined NNW $\Phi: \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$. We denote by Θ the collection (in certain order) of all the weights (i.e., all the entries of A_k and b_k, $k = 1, \cdots, L$) of Φ and write $\Phi = \Phi_\Theta$

The supremum error function

$$E(\Theta, f) = \sup_{x \in \Omega} | \Phi_\Theta(x) - f(x) | .$$

Note that Ω can be a finite subset of $\mathbb{R}^{N_0}$.

The L^2-error function (or the least-squares error function)

$$E(\theta, F) = \|\Phi_\theta - F\|_{L^2}^2$$

If $\Omega = \{x_1, \dots, x_M\}$ then

$$E(\theta, F) = \sum_{k=1}^{M} \|\Phi_\theta(x_k) - F(x_k)\|^2$$

where $\|\cdot\|$ is the ℓ^2-norm (Euclidean norm) of $\mathbb{R}^{N_2}$.

The Kullback-Leibler divergence

$$D_{KL}(p\|q) = -\int_{\mathbb{R}^n} p(x) \ln \frac{q(x)}{p(x)} dx$$

where p, q are (positive) PDFs on $\mathbb{R}^n$

i.e., $p, q \geq 0$, $\int_{\mathbb{R}^n} p(x) dx = 1$, $\int_{\mathbb{R}^n} q(x) dx = 1$.

Exercise: show that $D_{KL}(p\|q) \geq 0$.