Deep Learning theory - Lecture 2

$[0,1]$

for a given $P, \ell, \mathcal{F}_\Theta, \mathcal{A}$ we study

$$\boxed{\Delta(\mathcal{D}_n) = \mathcal{R}(\mathcal{A}(\mathcal{D}_n)) - \inf_{f: X \to Y} \mathcal{R}(f)}$$

Population Risk decomposition:

$$\underbrace{\mathcal{R}(f_{\hat\theta}) - \inf_{f: X \to Y} \mathcal{R}(f)}_{\text{population risk.}} = \underbrace{\mathcal{R}(f_{\hat\theta}) - \inf_{\theta\in\Theta} \mathcal{R}(f_\theta)}_{\text{estimation err.}} + \underbrace{\inf_{\theta'\in\Theta} \mathcal{R}(f_{\theta'}) - \inf_{f: X \to Y} \mathcal{R}(f)}_{\text{approximation err.}}$$

ANALYSIS $(\mathcal{F})$

Question: What is an (artificial) neural network?
What is the corresponding $\mathcal{F}_\Theta$? how big is it?

"Definition" of neural network:

Neural networks are a class of functions combining (usually pointwise) nonlinear operations - the neurons - through linear maps / transformations / connections - the network

Different connection structures / nonlinearities give different NN.

1) Single (hidden) layer neural networks (shallow)

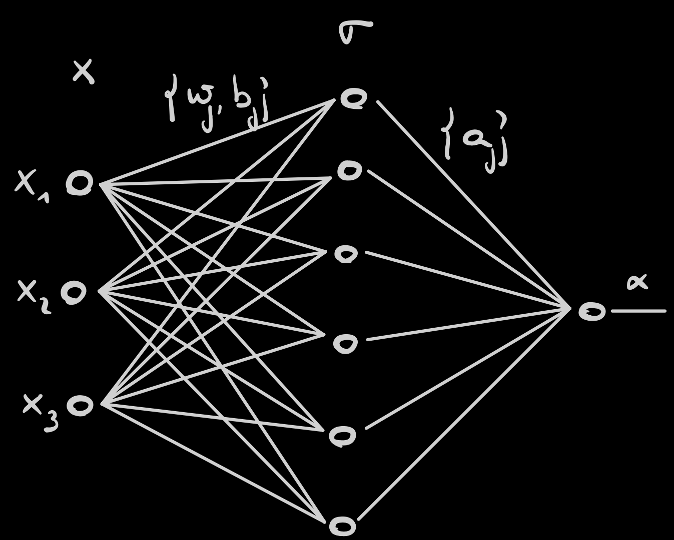
Def: Let $\mathcal{X} = \mathbb{R}^d$. A single layer neural network of width $n \in \mathbb{N}$ is a function of the form:

$$\hat{f}_\theta(x) = \alpha \cdot \sum_{j=1}^{n} a_j \, \sigma(w_j \cdot x + b_j)$$

where
- $\alpha \in \mathbb{R}$ is the scaling of the network
- $a_j \in \mathbb{R}^{d'}$ are the output weights
- $w_j \in \mathbb{R}^d$ are the input weights
- $b_j \in \mathbb{R}$ are the biases.
- $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function.

we will define $\theta = (a_j, w_j, b_j)_j \in \mathbb{R}^{(1+d+1)n} =: \Theta$

Note:
- Most of the time we will set $d' = 1$.
- The above network can be represented as a graph:



here $d = 3$, $d' = 1$, $n = 6$

$$f(x) = \alpha \sum_j a_j \, \sigma(w_j \cdot x + b_j)$$

- Often (e.g. today) $\alpha = 1$, but we will use it when $n \to \infty$.
- The network can also be written using matrix notation, defining the action of σ on $\mathbb{R}^n$ to be componentwise:

$$\hat{f}_\theta(x) = \alpha\, \vec{a} \cdot \sigma(W\vec{x} + \vec{b})$$

where $\vec{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \in \mathbb{R}^n$, $b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \in \mathbb{R}^n$, $W = \begin{pmatrix} w_{1,1} & \cdots & w_{1,d} \\ \vdots & & \\ w_{n,1} & \cdots & w_{n,d} \end{pmatrix} \in \mathbb{R}^{n \times d}$

- There are many possible choices for activation σ:

a) Rectified Linear Unit (ReLU):

$$\sigma(z) = z_+ := \max(0, z) = z \cdot \mathbb{1}(z \geq 0)$$

b) Sigmoid

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

c) Hyperbolic tangent

$$\sigma(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

Less common activations include:

d) step function: $\sigma(z) = \mathbb{1}(z \geq 0) = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{else.} \end{cases}$

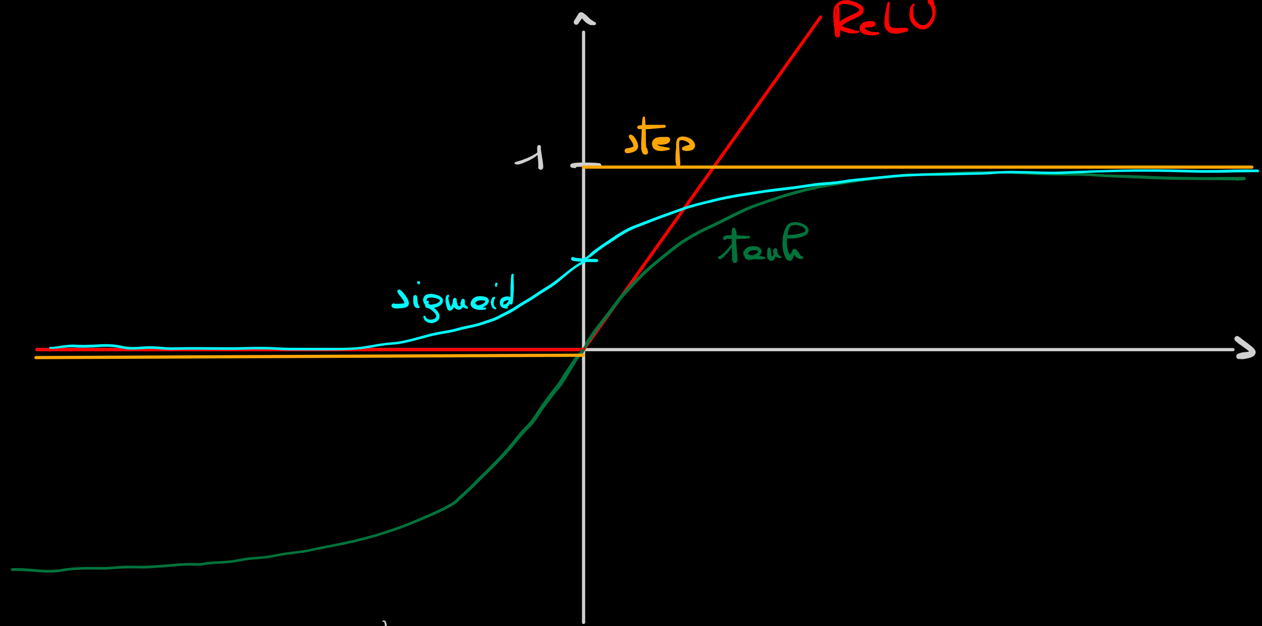
e) identity: $\sigma(z) = z$

f) smooth ReLUs: $\sigma(z) = x \cdot \Phi(x)$, $\sigma(z) = x \cdot (1 + e^{-z})^{-1}$

g) gaussian density : $\sigma(z) = e^{-x^2}$

h) squared : $\sigma(z) = z^2$

d) (... anything ...)



For each nonlinearity σ, we denote by

$$\mathcal{F}_{\sigma,n} := \left\{ x \mapsto \sum_{j=1}^{n} a_j \cdot \sigma(w_j \cdot x + b_j) : (a_j, w_j, b_j) \in \mathbb{R}^{1+d+1} \ \forall j \right\}$$

the space of single layer NN of width n with activation σ,

$$\mathcal{F}_\sigma = \bigcup_{n=1}^{\infty} \mathcal{F}_{\sigma,n}$$

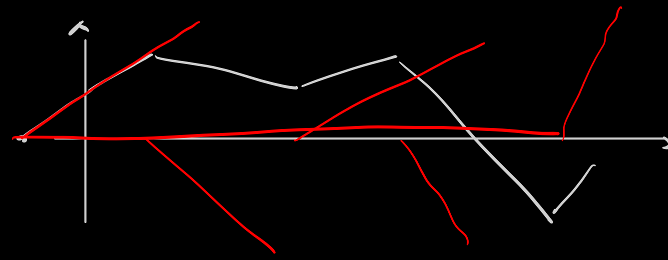
open/closed

<u>Example</u>: Let $\mathcal{X} \subseteq \mathbb{R}$

Let $PL_n(\mathcal{X}) = \left\{ f \in C(\mathcal{X}) : f(x) = (\tilde{a}_\ell \cdot x + \tilde{c}_\ell) \mathbb{1}(x \in [\tilde{b}_\ell, \tilde{b}_{\ell+1}]), \ \tilde{b}_\ell \in \mathcal{X}, \ \tilde{a}_\ell, \tilde{c}_\ell \in \mathbb{R}, \ \ell \in \{1, \dots, n\} \right\}$

<u>Lemma 2.2</u>: For every $h \in PL_n(\mathcal{X})$ there exists a single layer NN f of width n with ReLU nonlinearity such that $f(x) = h(x)$. In other words, $PL_n(\mathcal{X}) \subseteq \mathcal{F}_{ReLU,n}$

Proof: Set $w_j = 1 \ \forall j \in \{1, \dots, n\}$.



Then we can write

$$f_\theta(x) = \sum_{j=1}^{n} a_j (x + b_j) \mathbb{1}(x + b_j \geq 0) = \sum_{j=1}^{n} (a_j x + a_j b_j) \mathbb{1}(x \geq -b_j)$$

So, for $k = 1$: $-b_1 = \tilde{b}_1$, $a_1 = \tilde{a}_1$

$k \to k+1$: solve. $\displaystyle\sum_{j=1}^{k+1} (a_j x + a_j b_j) = \tilde{a}_{k+1} x + \tilde{c}_{k+1} \implies \begin{cases} a_{k+1} = \tilde{a}_{k+1} - \sum_{j=1}^{k} a_j \\ b_{k+1} = \frac{1}{a_{k+1}} \left(\tilde{c}_{k+1} - \sum_{j=1}^{k} a_j b_j \right) \end{cases}$

$\square$

- Sometimes a function is applied to the output of a network:

$$\hat{f}_\theta(x) = \left(1 + \exp\left(-\alpha \sum_{j=1}^{n} a_j \sigma(w_j x + b_j)\right)\right)^{-1}$$

Approximation error for single layer NNs

Which functions can we approximate arbitrarily well?

Def: a class of functions $\mathcal{F}$ is an universal approximator over a compact set $\mathcal{X}$ if for every $g \in C(\mathcal{X})$ there exists $f \in \mathcal{F}$ with $\|f - g\|_\infty \leq \varepsilon$.

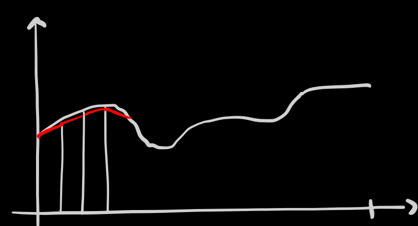
Consider first the univariate case. Throughout, we set wlog. $\mathcal{X} = [0,1]$

We start with intuitive, constructive pf.

Prop 2.1: Let $g: \mathbb{R} \to \mathbb{R}$ be ρ-Lipschitz. Then for every $\varepsilon > 0$ there is a single-layer NN f_θ with $\sigma(z) = z_+$ of width $n = \lceil \rho/\varepsilon \rceil$ such that $\|g - f_\theta\|_\infty \leq \varepsilon$

Proof: By the above lemma we only need to show that $\exists h \in PL_n(\mathcal{X})$ such that $\|g - h\|_\infty \leq \varepsilon$

We choose h as the linear interpolator of $\{g(n\varepsilon \wedge 1)\}_{n=0}^{\lceil 1/\varepsilon \rceil}$



Let $w_j =$, $\tilde{b}_j = j\frac{\varepsilon}{\rho}$ -

$a_0 = g(0) \qquad \tilde{a}_j = \frac{g(b_j) - g(b_{j-1})}{\varepsilon}$

so that $f_\theta(x) = \sum_{j=0}^{n-1} a_j \mathbb{1}(x \geq b_j) = \sum_{j=0}^{n-1} g(b_j) \mathbb{1}(x \in [b_j, b_{j+1}))$

$\Rightarrow |g(x) - f_\theta(x)| \leq |g(x) - g(b_j)| + |g(b_j) - f_\theta(b_j)| + |f_\theta(b_j) - f_\theta(x)|$

$x \in (b_j, b_{j+1})$ $\leq \rho(x - b_j) + 0 + 0 \leq \rho\frac{\varepsilon}{\rho} = \varepsilon.$

Alternatively we could have used the result

Lemma 2.2 Let $\mathcal{X}$ be a compact subset of $\mathbb{R}$. Then PL is dense in $C(\mathcal{X})$ in the $\|\cdot\|_\infty$ topology.

Proof: see course of real analysis. (e.g. Rudin)

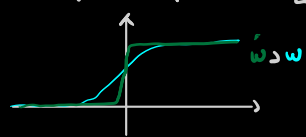
Cor. 2.3: For every compact set $\mathcal{X} \subseteq \mathbb{R}$, single layer neural networks with ReLU nonlinearity are universal approximators

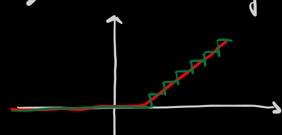
Remark: While the above result is more general (approximation of $C(\mathcal{X})$ instead of Lipschitz) Prop 2.1 gives bounds on the size of the network   low complexity $\leftrightarrow$ low width

Exercise: Let $\mathcal{X}$ be compact, then $\mathcal{F}_{\text{sigmoid}}$ is univ. approx in $\mathcal{X}$.

Idea: By Cor 2.3 $\exists\, h_n \in \mathcal{F}_{\text{ReLU},n}$ with $\|h_n - g\|_\infty \leq \frac{\varepsilon}{2}$

Then, approximate each neuron of h_n (with error $\frac{\varepsilon}{2n}$) with $\sum a_j \sigma(w_j x + b_j)$ for $w_j \gg 1$, $a_j \ll 1$





Lemma 2.4 Let $\mathcal{X} \subseteq \mathbb{R}$ be compact. $\mathcal{F}_{\mathbb{1}_{\geq 0}}$ is a universal approximator over $\mathcal{X}$

Pf: This follows from the fact that

$$\mathbb{1}(x \in [b_1, b_2]) = \mathbb{1}(x \geq b_1) \cdot \mathbb{1}(x \leq b_2) = \mathbb{1}(x - b_1 \geq 0)\, \mathbb{1}(-x + b_2 \geq 0)$$

together with the density of simple functions in $C(\mathcal{X})$

Remark: Polynomials of bounded degree are closed under linear transf.

$$p \in P(n) \implies a\, p(a'x + b') + b \in P(n)$$

and are <u>not</u> dense in $C(\mathcal{X})$ for fixed $n \implies$ no univ. approx

Remark: The compactness of $\mathcal{X}$ is necessary: for any finite relu network f, if $\mathcal{X} = \mathbb{R}$ $\quad \|\cos(x) - f(x)\|_\infty \geq 1$.

<u>Remark</u>: Why the $\|\cdot\|_\infty$ topology? Because we want to bound $\mathcal{R}(\theta)$ without knowing P.

Q: how about higher dimensions? $d > 1$

<u>Idea</u>: We would like to approximate with sums.

$$\prod_e \mathbb{1}(x \in A_e) \qquad A_e = \bigtimes_{j=1}^{d} [b_{e,j}, b'_{e,j}).$$

However, this cannot be done (at least directly) with shallow nets.

Or can it?

Consider $\sigma(z) = \cos(z)$. Then

$$\begin{aligned}\sigma(z)\cdot\sigma(y) &= \cos(z)\cos(y)\\ &= \cos(z+y) + \cos(z-y).\\ &= \sigma(z+y) + \sigma(z-y).\end{aligned}$$

So, at least in principle, using the above property of cosines and the Fourier approximation of step functions:

$$g(x) \approx \sum_j \alpha_j \prod_e^d \mathbb{1}(x_e \in [b_{j,e}, b'_{j,e})) \approx \sum_j \alpha_j \prod_e \sum_k \beta_k \cos(\gamma_k x_e)$$

$$\xrightarrow{\text{algebra}} \approx \sum_j \alpha_j \sum_k \beta_k \prod \cos(\gamma_k x_e) \approx \sum_j \alpha_j \sum_k \beta \sum_m \delta_m \cos\left(\sum \gamma'_{ek} x_e\right)$$

$\overrightarrow{\text{expressivity}} \approx \sum_j \alpha_j \sum_e \beta_e \sum \delta_m \sum \eta_n \mathbb{1}[\gamma_n' \cdot x \geq b_n]$

In practice, this is painful, but the heavy lifting was done for us:

<u>Thm</u> 2.5 (Stone-Weierstrass): Let $\mathcal{F} \subseteq C(\mathcal{X})$ for compact $\mathcal{X} \subseteq \mathbb{R}^d$ satisfy.

a) for every $x \in \mathcal{X}$, there exists $f \in \mathcal{F}$ such that $f(x) \neq 0$

b) for every pair $x, x' \in \mathcal{X}$ with $x \neq x'$ there exists $f \in \mathcal{F}$ with

$$f(x) \neq f(x') \qquad (\mathcal{F} \text{ separates points})$$

c) $\mathcal{F}$ is closed under pointwise multiplication ($\mathcal{F}$ is an algebra)

then $\mathcal{F}$ is an universal approximator.

<u>Lemma</u> 2.6 $\mathcal{F}_{\cos}$ is universal

<u>Pf</u>: a) each $f \in \mathcal{F}_{\cos}$ is continuous (finite sum of cont. functions)

b) $\cos(0 \cdot x) = 1 \quad \forall x \in \mathcal{X}$

c) $x \neq x' \Rightarrow f(z) = \cos\left(\frac{(z-x')\cdot(x-x')}{\|x-x'\|^2}\right)$ satisfies $\begin{cases} f(x') = 1 \\ \neq \\ f(x) = 0 \end{cases}$

d) already checked. $\square$

<u>Thm</u> 2.7 Suppose $\sigma \in C(\mathbb{R})$ is sigmoidal: $\begin{cases} \lim_{z \to -\infty} \sigma(z) = 0 \\ \lim_{z \to \infty} \sigma(z) = 1 \end{cases}$

then $\mathcal{F}_\sigma$ is universal.

Also, $\mathcal{F}_{\mathrm{ReLU}}$ is universal

Pf (sketch): By Lemma 2.7 we have there exists $n \in \mathbb{N}$,

$$h_n(x) = \sum_{j=1}^{n} \tilde{a}_j \cos(\tilde{w}_j \cdot x + \tilde{b}_j) \in \mathcal{F}_{\cos}$$

with $\|h - g\|_\infty \leq \frac{\varepsilon}{2}$

Then, since $h_{n,j}(x) = \hat{a}_j \cos(\hat{w}_j \cdot x + \hat{b}_j) \in C(\mathcal{X})$, by exercise we have $\exists f_{n,j} \in \mathcal{F}_{\text{sigmoid}} : \|f_j - h_{n,j}\| \leq \frac{\varepsilon}{2n}$

$\implies$ for $f(x) = \sum_j f_{n,j}(x) \in \mathcal{F}_{\text{sigmoid}}$

$$\|f_n - g\|_\infty \leq \|f_n - h_n\| + \|h_n - g\| \leq \sum_{j=1}^{n} \frac{\varepsilon}{2n} + \frac{\varepsilon}{2} \leq \varepsilon. \qquad \square$$

Note: the algebra condition does not hold for polynomials of bounded degree. In fact.

Thm (Leshno, 1993): $\mathcal{F}_\sigma$ is universal iff $\sigma \in C(\mathcal{X})$ is not a polynomial