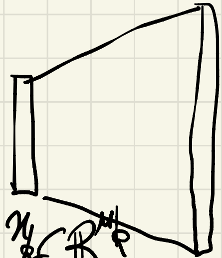
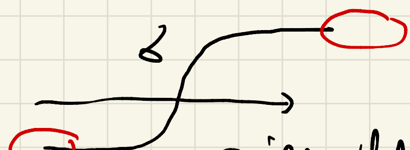
Recap : MLP



$x_k \in \mathbb{R}^{N_k}$

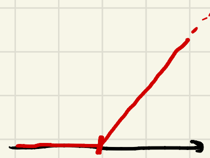
$$x_{k+1} = \sigma(W_k x_k + b_k)$$



sigmoids

$$\Theta_k = (W_k, b_k)$$

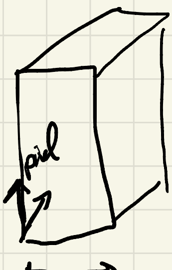
$$\sigma(x) = \begin{cases} \mathrm{Arctan}(x) \\ \dfrac{e^x}{1+e^x} \end{cases} \qquad \sigma(x) = \mathrm{Relu}(x)$$



Design choices : "Weight sharing" → convolution

audio / image / video

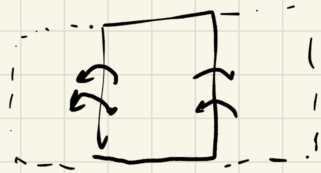
$x_k =$



$$x_k = \left(x_k[\cdot,\cdot,s]\right)_{s=1}^{S}$$

"Translat° inv. lin operator ⟺ convolution"

$$T_\tau : x[\underbrace{\cdot,\cdot}_{\text{pixel}},\overset{\text{channel}}{\cdot}] \rightsquigarrow x[\cdot-\tau_1, \cdot-\tau_2, \cdot]$$



<u>Thm</u>: $W: x\in\mathbb{R}^{M\times M\times S} \to \mathbb{R}^{M\times M\times T}$

that commutes with transl°

$$W\circ T_\tau = T_\tau \circ W$$

$$\begin{tikzcd} x \arrow[r, "W"] \arrow[d, "T_\tau"] & Wx \\ {} \arrow[r, "W"] & {} \arrow[u, "T_{-\tau}"] \end{tikzcd}$$

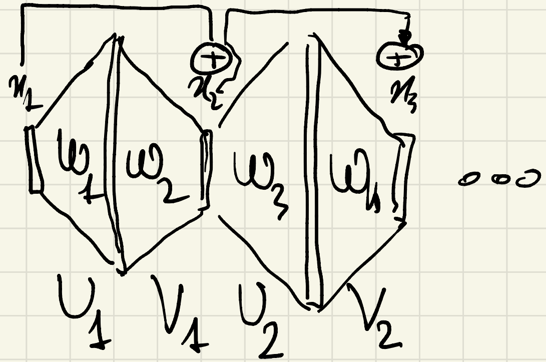
if and only if $\exists$ $\{\psi_{st}\}_{\substack{s=1\dots S\\ t=1\dots T}}$ so that

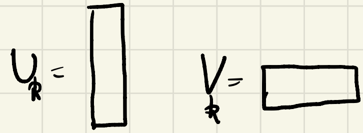
$$Wx = \Big(\sum_s \psi_{st} * x[\cdot,\cdot,s]\Big)_t$$

<u>Example</u>: 1.1 conv°

2nd design choice: skip connex°

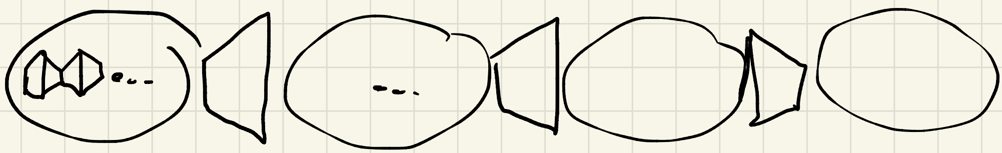
Res-Net





$$x_{k+1} = x_k + \sigma\left[V_k\, \sigma\left[U_k x_k\right]\right]$$

removed / batch norm.



[Generative net

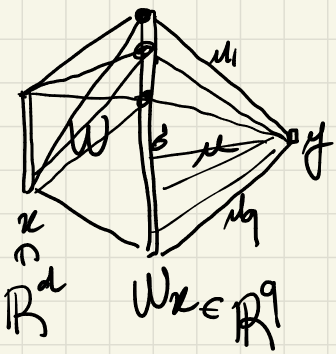
[U-Net → sept. generative model
diffu°

Transformers : convolut° + nonlinearity → Attent°

"Theory": Universality / Expressivity

2 layers NN

"linear" single layer $\qquad q = \#\text{neurons}$



$$y = \langle \underset{\mathbb{R}^q}{\mu}, \underbrace{\sigma(Wx)}_{\in \mathbb{R}^q} \rangle$$

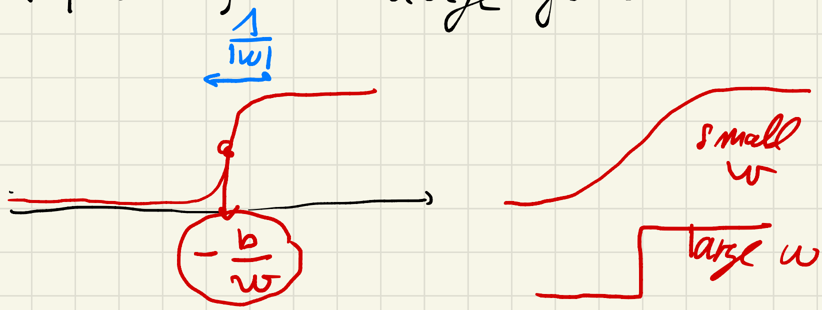
$$W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ \\ \end{bmatrix}$$

$$y = \underbrace{\sum_k \mu_k \sigma(\langle x, w_k \rangle + \overset{\mathbb{R}}{b_k})}$$

$$\varphi(x) = \sum_k \mu_k \varphi_{w_k, b_k}(x)$$

$$\varphi_{w,b}(x) = \sigma(\langle x, w \rangle + b)$$ "Ridge funct°"

1-D



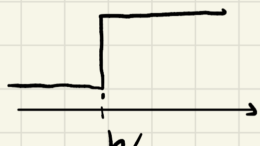
$\rightsquigarrow \langle w, x \rangle + b = 0$

Bounded σ

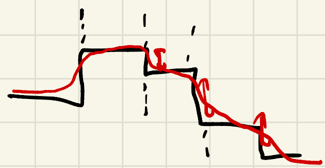
<u>Thm</u>: [Cibenko] if $f(x)$ continuous, $\varepsilon > 0$ precision, $R > 0$, $\exists\, q = \#$ neurons, $\exists\, w, b, u$,

such that, $\forall \|x\| \le R$ $\quad \Big| f(x) - \underbrace{\sum u_k \varphi_{w_k b_k}(x) + c}_{\text{NN. } q \text{ neurons}} \Big| \le \varepsilon$

$\varphi_{\frac{w}{\varepsilon}, \frac{b}{\varepsilon}}(x) \xrightarrow{\varepsilon \to 0}$



$\sum u_k \varphi_{\frac{w_k}{\varepsilon}, \frac{b}{\varepsilon}}(x)$



Speed of approx^{n} : · No free lunch

Barron's function : $\hat{f}(\omega) = \int f(x) \cdot e^{-i\langle \omega, x\rangle} dx$

Frequency

Barron's norm : $\|f\|_B \triangleq \int \|\omega\| \cdot |\hat{f}(\omega)| \, d\omega$

Sobolev norm : $\|f\|_{W^1}^2 = \int \|\omega\|^2 \cdot \|\hat{f}(\omega)\|^2 \, d\omega$

$\widehat{f'}(\omega) = i\omega \cdot \hat{f}(\omega)$

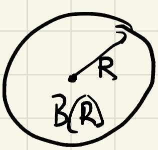
$= \int \|\widehat{f'}(\omega)|^2 \, d\omega$

$= \int |f'(x)|^2 dx = \|f'\|_{L^2}^2$

R>0 radius

Thm : If $\|f\|_B < +\infty$ then $\forall q>0$, $\exists$ a neural net

$f_q(x) = \sum u_k \varphi_{\omega_k, b_k}(x)$. such that



$$\sqrt{\frac{1}{|B(R)|}\int_{B(R)} |f(x) - f_q(x)|^2 dx} \le \frac{\|f\|_B}{\sqrt{q}} = \varepsilon$$

$\boxed{\text{Loss}}$

With pol.

$$\frac{\|f\|_{Barron}}{q^{1/d}} = \varepsilon \quad \leftarrow \text{curse of the dim}^c.$$

⊕ No curse of dim, explicit

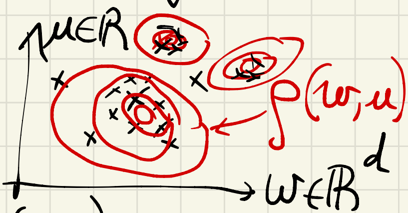
⊖ 2 layers.

⊖ existence, how to find f_q

<u>Proof</u>: mean field analysis.

$$f_q(x) = \frac{1}{q}\sum_{k=1}^{q} u_k \, \sigma(\langle x, w_k \rangle)$$

$$\Big\downarrow \; q \to +\infty \qquad \theta_k = (u_k, w_k) \in \mathbb{R}^{d+1}$$



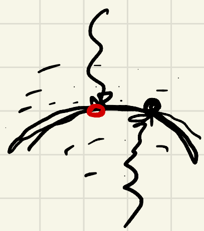
$$f^\rho(x) = \int u \, \sigma(\langle x, w \rangle) \, d\rho(w,u)$$

Difficult: $\|f\|_B < +\infty \iff \exists \rho, \; f \overset{!}{=} f^\rho$

Q°: optimiz°. Chizat / Bach.

if $q \to +\infty$, GD on the loss

Q°: How many neurons ($q = ??$)

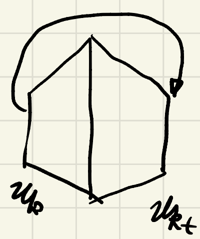


Q°: Generaliz° $\to$ Implicit bias.

Deep & Wide ResNet.

$$x_{k+1} \overset{(*)}{=} x_k + \frac{1}{K} V_k \cdot \sigma(\underbrace{U_k x_k}_{\in \mathbb{R}^q})$$

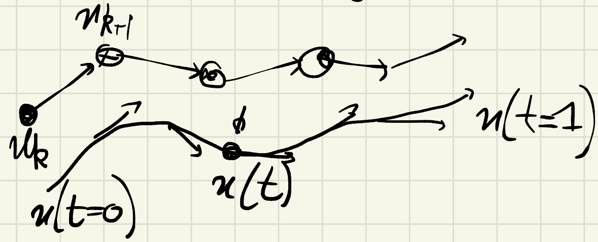
$k = 1 \dots K$ Depth



$$\begin{cases} K = \text{depth} \\ q = \text{width} \end{cases}$$ "$\to +\infty$" mean field

$$(*) \qquad \frac{x_{k+1}-x_k}{1/K} = v_{\theta_k}(x_k) \qquad \theta_k = (U_k, V_k)$$

$$v_\theta(x) = U\sigma(Vx)$$



"$K \to +\infty$"

$$\dot{x}(t) = v_{\underline{\underline{\theta(t)}}}(x(t)).$$

out : $x(1)$.

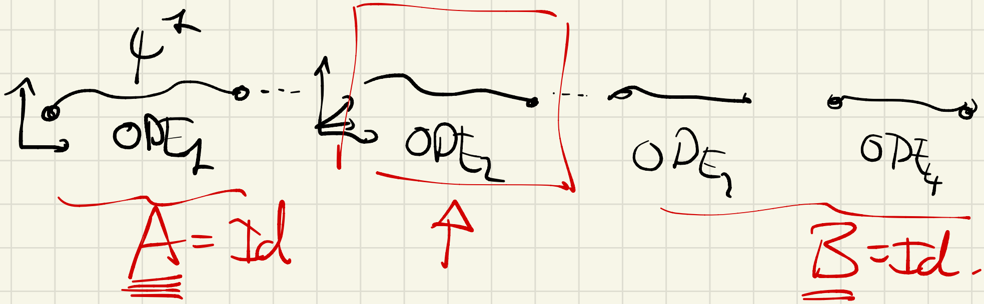
$$x(t=0) = x_0$$

inf. depth limit : $\Psi_\theta : x(0) \longrightarrow x(1)$.

?? learning ??

$$(x_i)_{i=1}^N \longrightarrow (y_i)_{i=1}^N$$

<u>Learn</u> : $$\min_{\{\theta(t)\}_{t=0}^1} E(\theta) = \frac{1}{N}\sum_{i=1}^N \ell(\Psi_\theta(x_i), y_i).$$

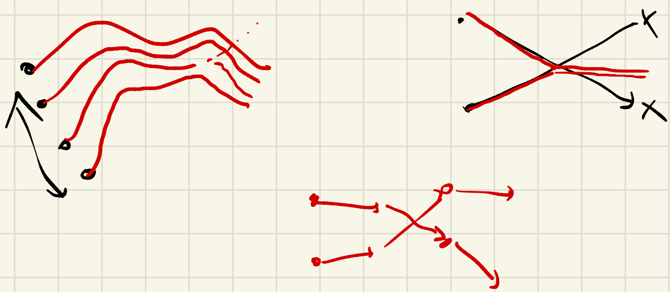


Thm: Barboni Neurips 2022

$$C'(\theta)\cdot\|\nabla E(\theta)\|^2 \overset{(2)}{\leq} E(\theta)-E(\theta^*) \overset{(1)}{\leq} C(\|\theta\|)\times\|\nabla E(\theta)\|^2$$

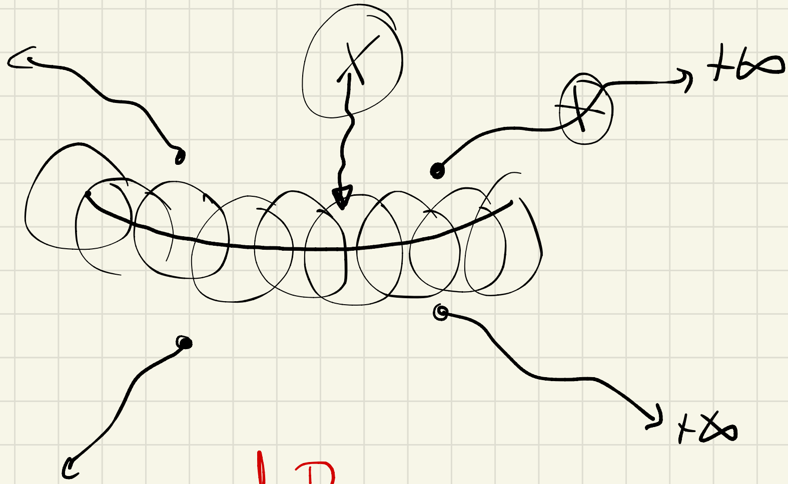
P-L

Pollak-Łojajevitch.



(1) $\Rightarrow$ No loc. min. but $\omega \to +\infty$

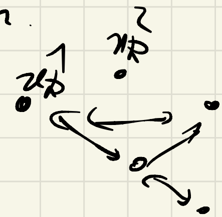
(2) $\Rightarrow$ if θ "close" to θ^* $\omega \not\to +\infty$



$A \in \mathbb{R}^{d \times D}$ $\quad$ $\underline{\underline{D}}$ large enough

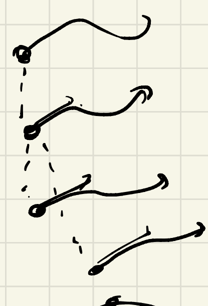
$$\dot{x}_{k+1} = v_{\theta_k}(x_k)$$

$$X_k = (x_k^1, x_k^2 \; - \; - \; x_k^{\overset{\text{\#token}}{s}})$$



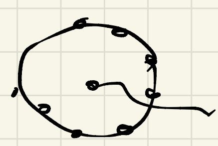
$$X_{k+1} = X_k + \frac{1}{k} A_{\theta_k}(X_k)$$

$$\dot{X}(t) = A_{\theta(t)}(X(t))$$



$$\dot{x}^i(t) = v_{\theta(t)}(x^i(t); X(t))$$

$$x_i \to \frac{x_i - m_i}{\| \quad \|}$$



$$v_{\theta^{(t)}}\left(x^i ; (x^j)\right) = \sum_j A_{ij} \cdot (V x_j)$$

$\theta^{(t)} \leftarrow (K, Q, V)$

$\sum_j A_{ij}$) Hard

≥ 0

$$A_{ij} = e^{\frac{\langle K x_i, Q x_j \rangle}{\sqrt{d}}}$$