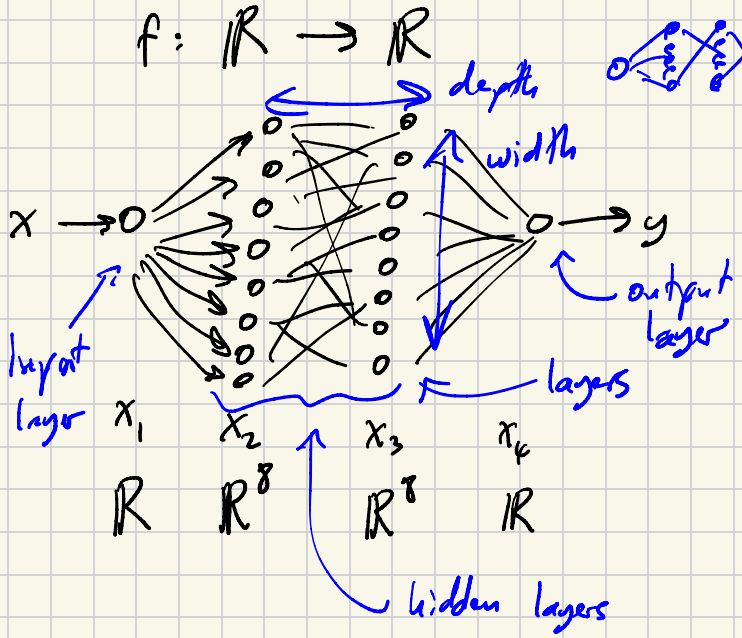
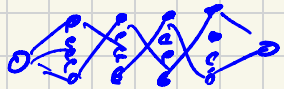
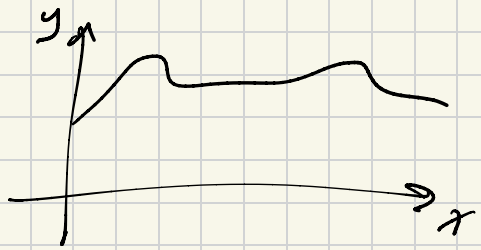
Function approximation with neural networks







$x_1 = x$ $\quad x_1 \in \mathbb{R}$

$x_2 = \sigma(W_1 x_1 + b_1)$ $\quad \mathbb{R} \to \mathbb{R}^8$

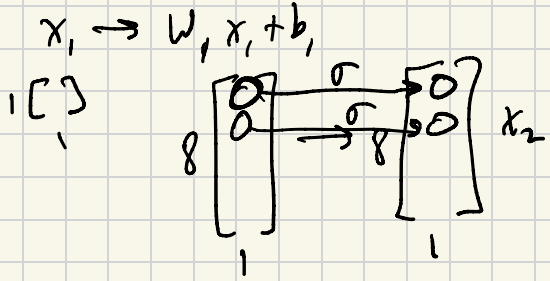
$x_3 = \sigma(W_2 x_2 + b_2)$

$x_4 = W_3 x_3 + b_3$

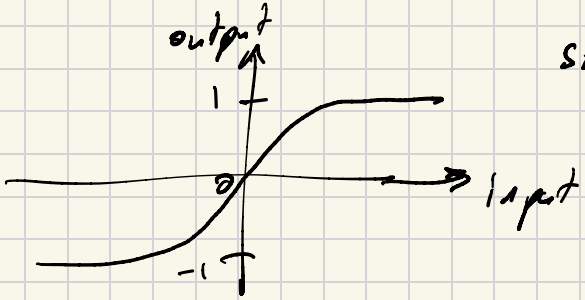
$y = x_4$

W_1	8×1
b_1	8
W_2	8×8
b_2	8
W_3	1×8
b_3	1

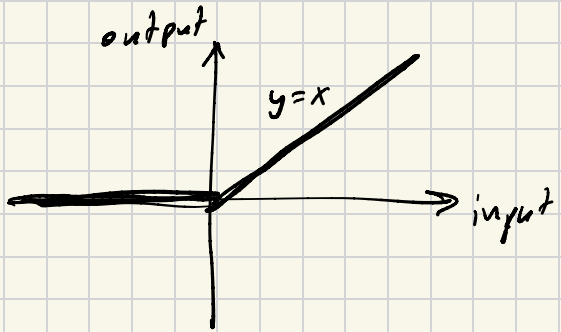
parameters



σ = activation function



sigmoid
tanh



ReLU

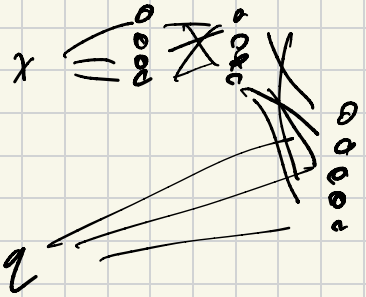
$x \in \mathbb{R}$ → (NN) → $y = f(x)$

$\theta \in \mathbb{R}^{97}$ →

θ ← parameters $= (W_1, b_1, W_2, b_2, W_3, b_3)$

$\nabla_\theta f$

ex.



Universal approximation theorems

https://en.wikipedia.org/wiki/Universal_approximation_theorem

Single hidden layer, arbitrary width (1990s)

$\mathbb{R}^n \to \mathbb{R}^m$

Universal approximation theorem — Let $C(X, \mathbb{R}^m)$ denote the set of continuous functions from a subset X of a Euclidean $\mathbb{R}^n$ space to a Euclidean space $\mathbb{R}^m$. Let $\sigma \in C(\mathbb{R}, \mathbb{R})$. Note that $(\sigma \circ x)_i = \sigma(x_i)$, so $\sigma \circ x$ denotes σ applied to each component of x.

Then σ is not polynomial if and only if for every $n \in \mathbb{N}$, $m \in \mathbb{N}$, compact $K \subseteq \mathbb{R}^n$, $f \in C(K, \mathbb{R}^m), \varepsilon > 0$ there exist $k \in \mathbb{N}$, $A \in \mathbb{R}^{k \times n}$, $b \in \mathbb{R}^k$, $C \in \mathbb{R}^{m \times k}$ such that

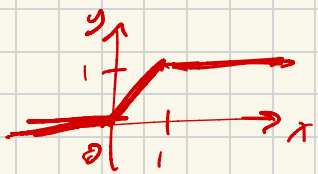
$$\sup_{x \in K} \|f(x) - g(x)\| < \varepsilon$$

where $g(x) = C \cdot (\sigma \circ (A \cdot x + b))$ ← NN

Cybenko, G. (1989). "Approximation by superpositions of a sigmoidal function". Mathematics of Control, Signals, and Systems. 2 (4): 303–314. doi:10.1007/BF02551274

Hornik, Kurt (1991). "Approximation capabilities of multilayer feedforward networks". Neural Networks. 4 (2): 251–257. doi:10.1016/0893-6080(91)90009-T

Pinkus, Allan (1999). "Approximation theory of the MLP model in neural networks". Acta Numerica. 8: 143–195. doi:10.1017/S0962492900002919



Fixed width, arbitrary depth (2010s)

→ deep learning revolution

Universal approximation theorem *(L1 distance, ReLU activation, arbitrary depth, minimal width).* For any Bochner–Lebesgue p-integrable function $f : \mathbb{R}^n \to \mathbb{R}^m$ and any $\epsilon > 0$, there exists a fully-connected ReLU network F of width exactly $d_m = \max\{n+1, m\}$, satisfying

$$\int_{\mathbb{R}^n} \|f(x) - F(x)\|^p \mathrm{d}x < \epsilon.$$

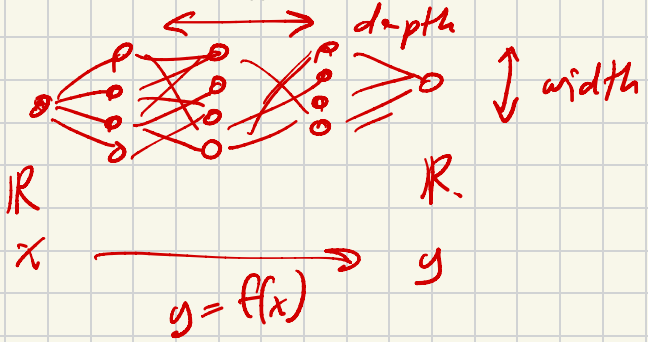
← NN

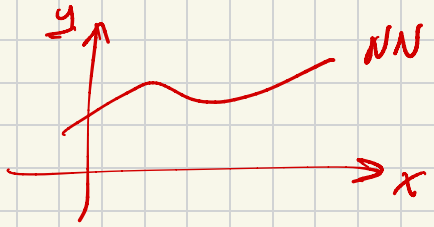
Moreover, there exists a function $f \in L^p(\mathbb{R}^n, \mathbb{R}^m)$ and some $\epsilon > 0$, for which there is no fully-connected ReLU network of width less than $d_m = \max\{n+1, m\}$ satisfying the above approximation bound.

Lu, Zhou; Pu, Hongming; Wang, Feicheng; Hu, Zhiqiang; Wang, Liwei. "The Expressive Power of Neural Networks: A View from the Width". NeurIPS 2017. arXiv:1709.02540

Park, Yun, Lee, Shin, Sejun, Chulhee, Jaeho, Jinwoo. "Minimum Width for Universal Approximation". ICLR 2021. arXiv:2006.08859

R. DeVore, B. Hanin, and G. Petrova, "Neural network approximation", Acta Numerica 30, 327-444, 2021. doi:10.1017/S0962492921000052





$\frac{\partial f}{\partial x} \rightarrow$ x.grad

$\frac{\partial f}{\partial x_1}$

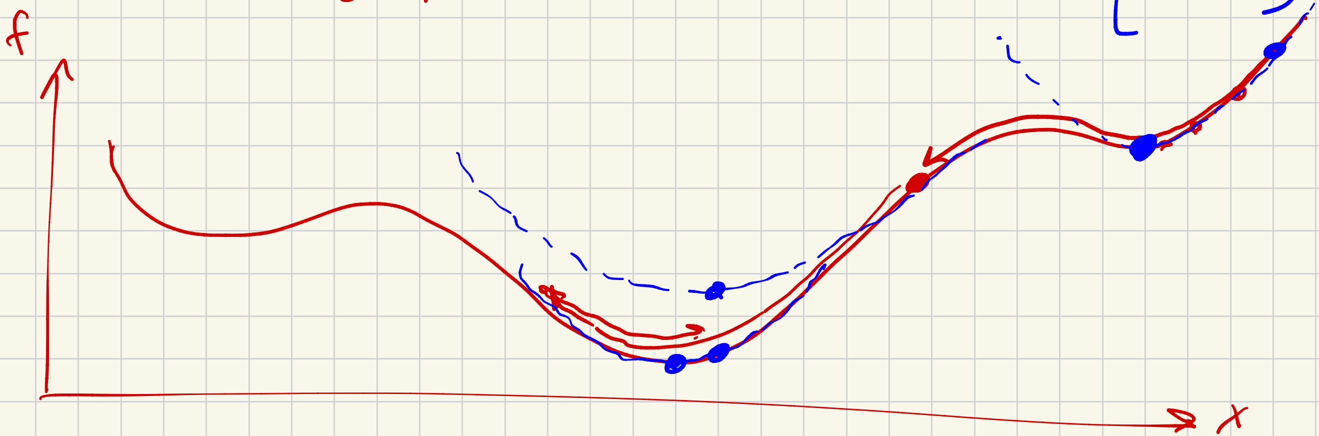
$\frac{\partial f}{\partial x_2}$

$\rightarrow f \in \mathbb{R}$

$\nabla_x f$

$\nabla(\nabla f)$

$w_i \in \mathbb{R}^n$

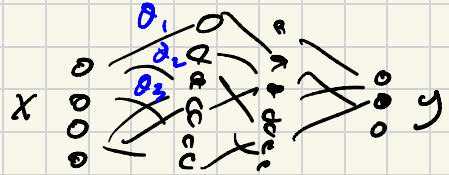


$$f : \mathbb{R}^n \to \mathbb{R}^m$$

$\hat{f}_\theta(x)$ is a NN

θ ← params

$$x_2 = \sigma(w_1 x_1 + b_1)$$



$x \in \mathbb{R}^k$ $\quad$ $y \in \mathbb{R}^s$

sample set $\{x_i\}$ compute true $y_i = f(x_i)$

training data $\{(x_i, y_i)\}$

Loss function $\quad L_\theta = \sum_i \| f_\theta(x_i) - y_i \|^2$

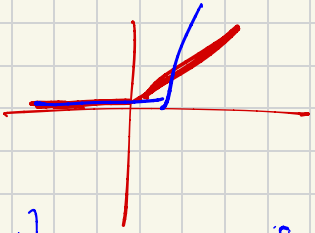
$\theta \longmapsto L_\theta \qquad \mathbb{R}^{96} \to \mathbb{R}$

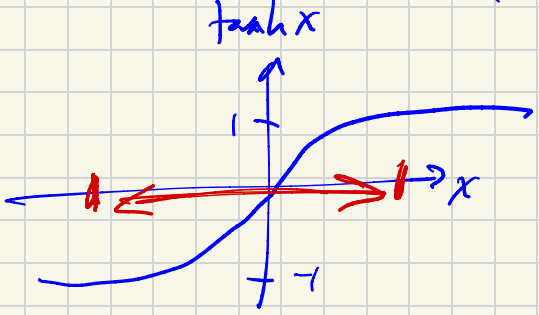
$$\theta^{k+1} = \theta^k - \alpha \nabla_\theta L(\theta^k)$$

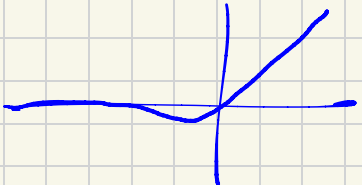
$$y_3 = \mathrm{ReLU}(w_2 x_2 + b_2)$$

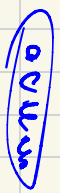
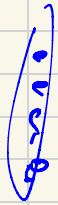
weight bias

$\mathrm{ReLU}(2x - 0.5)$







$\circ$  $\circ$  $\circ$