Neural Networks as Universal Approximators

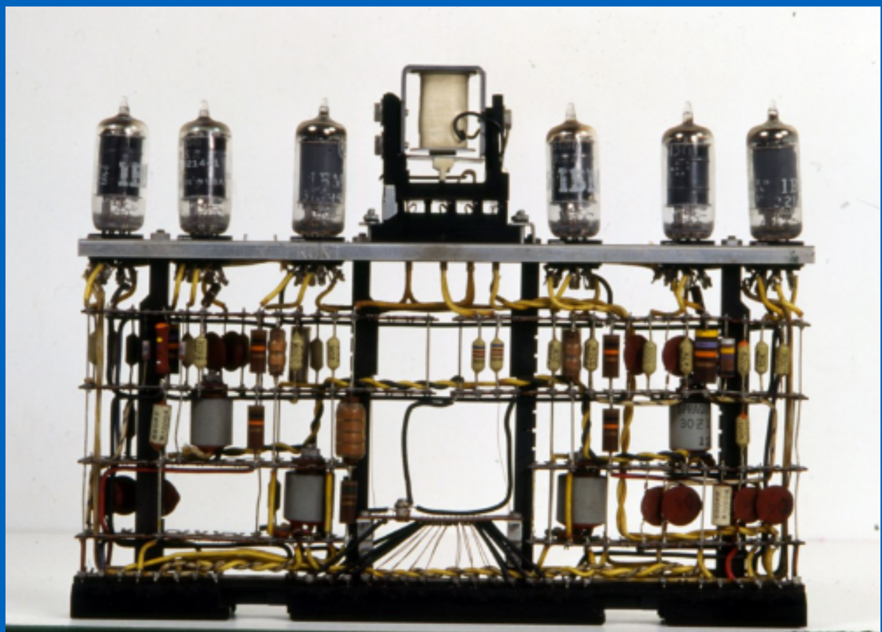
Pratik Shah
Jan 20, 2023

1943 – McCulloh and Pitts "A Logical Calculus of the Ideas Immanent in Nervous Activity."

In 1943, neurophysiologist Warren McCulloch and mathematician Walter Pitts wrote a paper on how neurons might work. In order to describe how neurons in the brain might work, they modeled a simple neural network using electrical circuits.

1949 – "Organization of the Behaviour" Donald Hebb

1957 – Frank Rosenblatt – first paper – "perceptron"



IBM 704

1957-62 Perceptron — HYPE AI

1969 Perceptron, Minsky and Papert

(limitations and critique to AI)

1982-85 John Hopfied (Caltech)

Today...

- What is learning?
- What is the role of data in learning?

Data $\{(x_i, y_i)\}_{i\in I}$ $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}^m$

$\left\{\begin{matrix}\text{training}\\ \text{instances}\end{matrix}\right\}$ I/p O/p

$$P(Y_i \mid X_i) \Rightarrow P(Y_i \mid X_i, \Theta)$$

parameters

$$P(Y_i \mid X_i, S_T, \Theta)$$

w, b

training set

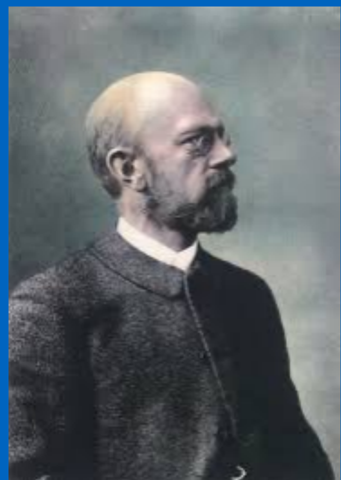
$$P(Y_i \mid X_i, r, S_T, \Theta)$$

seed

Further back in 1900

8th August

David Hilbert



Sorbonne, Paris

ICM

The 13^{th} Hilbert's problem:

$$x^7 + ax^3 + bx^2 + cx + 1 = 0$$

- Is it possible to write its solution x, as a function of a, b and c $\longrightarrow$ as a COMPOSITION of a finite number of two variable functions?

$$x = f(a, b, c). \qquad \phi_2(\phi_1(a, b), c)$$

- $f = \sum a_i \phi_i$ — linear combination

$f = \phi_n \circ \phi_{n-1} \circ \dots \circ \phi_2 \circ \phi_1$ — composition of functions.

1957 Arnold (Kolmogorov)

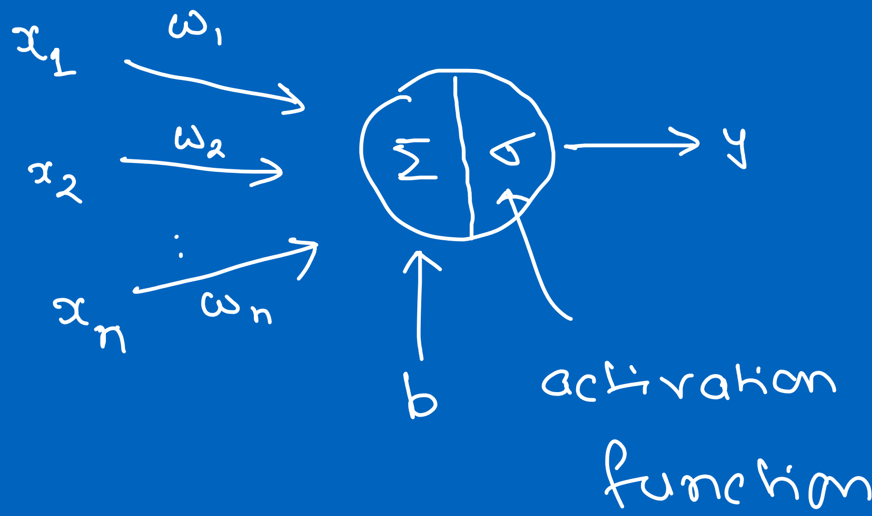
"Can every continuous function of 3 variables be expressed as a composition of finitely many continuous functions of two variables?"

$$f(\underbrace{x, y, z}) = x^3 + xy^2z + zx = 0$$

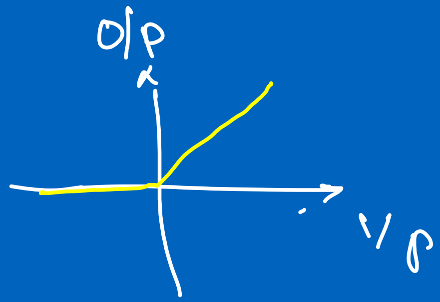
$$f(x, y, z) = \phi_n(\overbrace{\phi_{n-1}(-,)}, \overbrace{\quad - \quad})$$

Neuron :

$\mathbb{R}^n \rightarrow \mathbb{R}$



- ReLU



- tanh , Sigmoid

$[-1, +1]$ $[0, 1]$

$$y = \sigma\left(\sum_{i=1}^{n} \omega_i x_i + b\right)$$

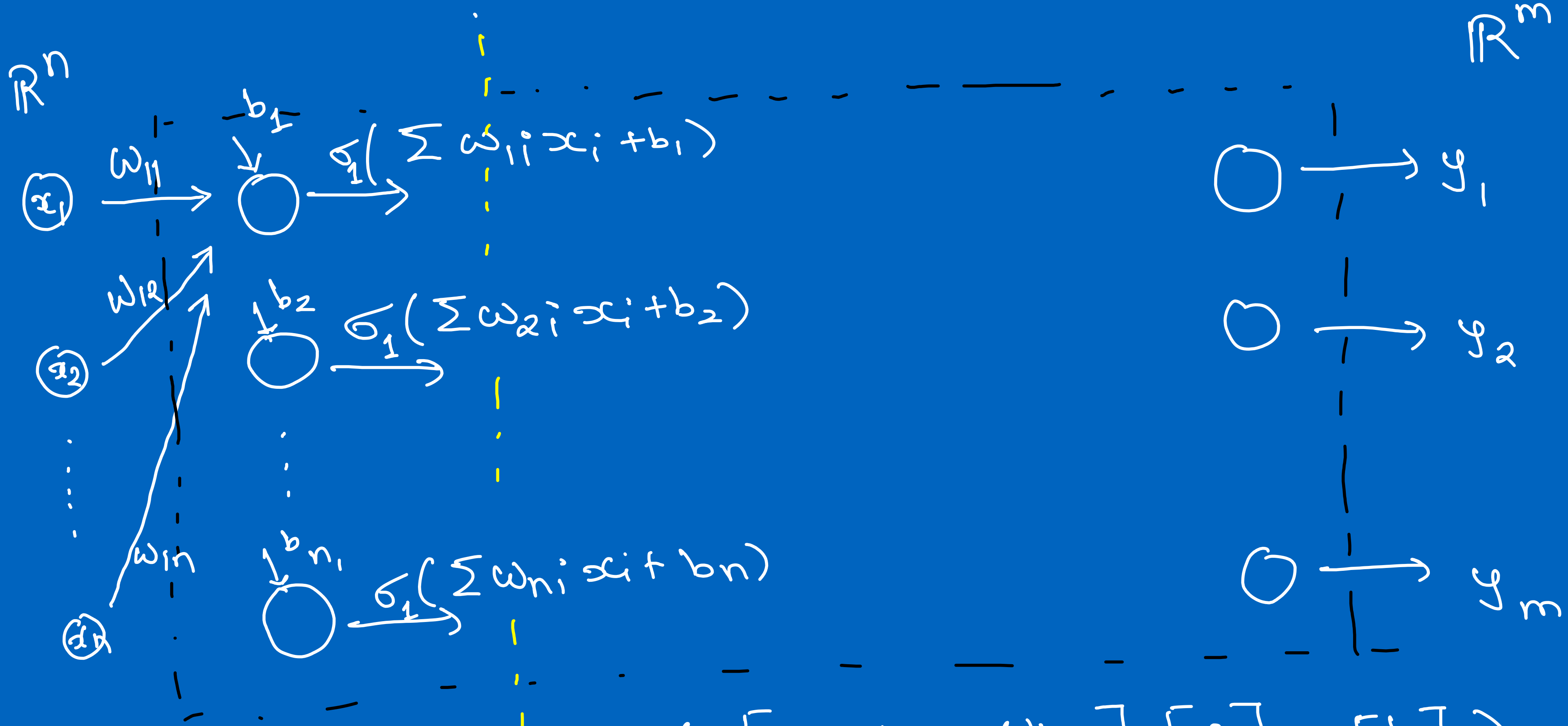
$\uparrow$ affine function

$$\sigma(\omega^T x + b)$$

- Nonlinearity!

$$x = [x_1 \; x_2 \ldots x_n]^T \qquad x \in \mathbb{R}^n$$

$$\omega = [\omega_1 \; \omega_2 \ldots \omega_n]^T \qquad \omega \in \mathbb{R}^n$$



$\# n_1$
Layer-1

$$\sigma_1\left(\underbrace{\begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ w_{n_1 1} & & \dots & w_{n_1 n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n_1} \end{bmatrix}}_{t_1 : \mathbb{R}^n \to \mathbb{R}^{n_1}}\right)$$

Layer 1

Output $\rightarrow$ $\sigma_1 \circ t_1(x)$

$$\sigma\left(\underbrace{\begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} & \cdots & \omega_{1n} \\ \omega_{21} & \omega_{22} & \cdots & & \omega_{2n} \\ \vdots & & & & \\ \omega_{n_1 1} & \omega_{n_1 2} & \cdots & & \omega_{n_1 n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n_1} \end{bmatrix}}_{\text{Affine transform}}\right)$$

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

$\downarrow n_1$

$\sigma_1(Wx+b) \in \mathbb{R}^{n_1}$

$\vdots$

$Y \in \mathbb{R}^m$

$$f(x) = t_N \circ \cdots \sigma_3 \circ t_3 \circ \sigma_2 \circ t_2 \circ \sigma_1 \circ t_1(X)$$

$\boxed{\sigma\text{-net}} \rightarrow f$ $\quad \sigma = \{\sigma_1, \ldots, \sigma_K\}$

$$\theta = (W^i, b^i)$$

$$f(x^{(i)}, \theta) \sim y^{(i)}$$

$$\boxed{\min_{\theta} \sum_{i=1}^{\#} \left(f(x^{(i)}, \theta) - y^{(i)}\right)^2}$$

n		n_1		n_2		n_3		n_k		m
I/p	b_{n_1}, $W_{n_1 \times n}$	L1	b_{n_2}, $W_{n_2 \times n_1}$	L2	b_{n_3}	L3	....	Lk	b_{n_k}, $W_{m \times n_k}$	O/p
X		$\sigma_1 \circ t_1$		$\sigma_2 \circ t_2$		$\sigma_3 \circ t_3$		$\sigma_k \circ t_k$		Y

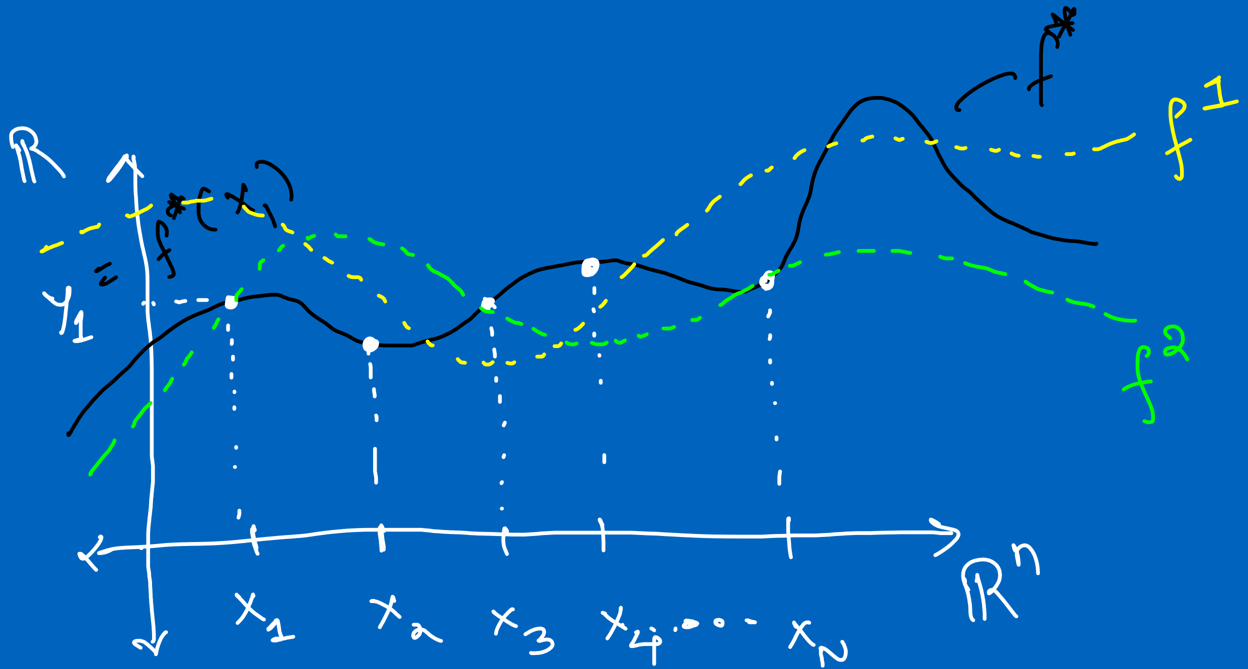
$$Y = \underbrace{\sigma_k \circ t_k \circ \sigma_{k-1} \circ t_{k-1} \circ \dots \circ \sigma_2 \circ t_2 \circ \sigma_1 \circ t_1(X)}.$$

σ – set of activation functions $\qquad \sigma_i \in \sigma$

<u>σ-net</u> $\qquad Y := f(x, \overbrace{W, b}^{\theta}) \qquad f : \mathbb{R}^n \to \mathbb{R}^m$

$$C(\mathbb{R}^n, \mathbb{R}^m) := \left\{ \begin{array}{c} \text{set of all continuous functions} \\ \text{from } \mathbb{R}^n \to \mathbb{R}^m \end{array} \right\}$$

Universal Approximation:



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

$D = \{(x_i, y_i)\}_{i \in \{1, \dots, N\}}$

σ-net $f^1: \mathbb{R}^n \to \mathbb{R}$

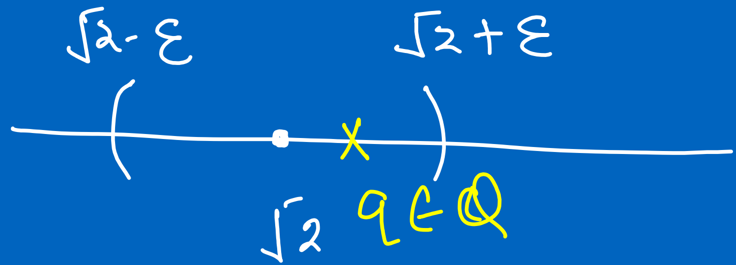
$\mathbb{Q} \subset \mathbb{R}$

rational $\sqrt{2}$

irrationals

- Floating point arithmetic

$\mathbb{Q}$ is dense in $\mathbb{R}$.



$\forall \varepsilon > 0$

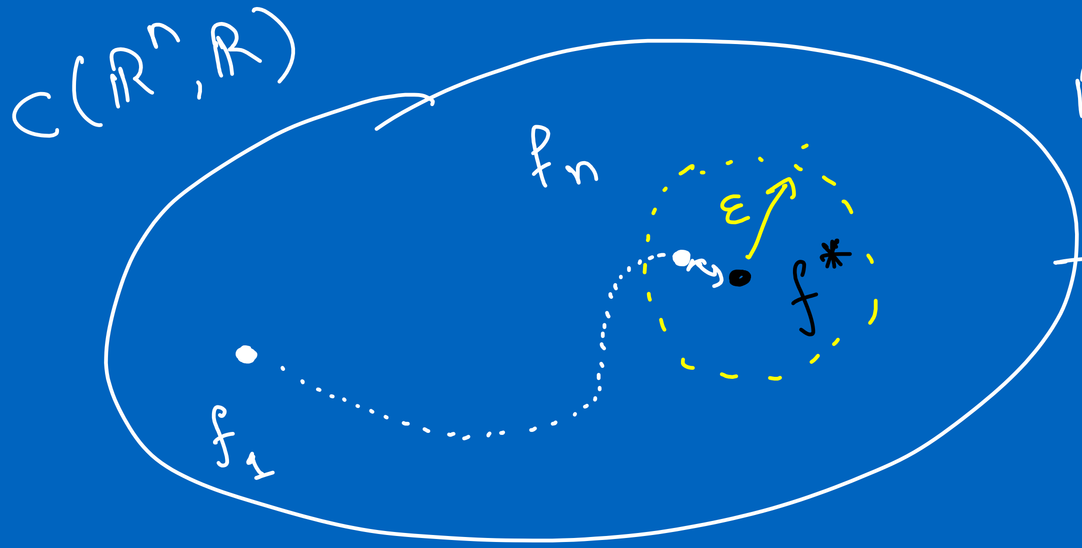
$d(a,b) = |a-b|$

$$d_D(f^1, f^*) = \sum_{i=1}^{N} \left(f^1(x_i) - f^*(x_i)\right)^2$$

$$\underset{w,b}{\text{minimize}} \quad e_f(w,b) = \sum_{i=1}^{N} \left(f(x_i, w, b) - y_i \right)^2$$

$$W_{i+1} \leftarrow W_i - \Gamma \, \nabla_{w,b} \, e_f(w,b)$$

Γ ← Learning rate



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

set of all cont. fun from $\mathbb{R}^n \to \mathbb{R}$

$f^* \in C(\mathbb{R}^n, \mathbb{R})$

σ-net is dense in $C(\mathbb{R}^n, \mathbb{R})$

Reference	Function class	Activation ρ	Upper / lower bounds
Lu et al. (2017)	$L^1(\mathbb{R}^{d_x}, \mathbb{R})$	ReLU	$d_x + 1 \leq w_{\min} \leq d_x + 4$
	$L^1(\mathcal{K}, \mathbb{R})$	ReLU	$w_{\min} \geq d_x$
Hanin and Sellke (2017)	$C(\mathcal{K}, \mathbb{R}^{d_y})$	ReLU	$d_x + 1 \leq w_{\min} \leq d_x + d_y$
Johnson (2019)	$C(\mathcal{K}, \mathbb{R})$	uniformly conti.†	$w_{\min} \geq d_x + 1$
Kidger and Lyons (2020)	$C(\mathcal{K}, \mathbb{R}^{d_y})$	conti. nonpoly‡	$w_{\min} \leq d_x + d_y + 1$
	$C(\mathcal{K}, \mathbb{R}^{d_y})$	nonaffine poly	$w_{\min} \leq d_x + d_y + 2$
	$L^p(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$	ReLU	$w_{\min} \leq d_x + d_y + 1$
Ours (Theorem 1)	$L^p(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$	ReLU	$w_{\min} = \max\{d_x + 1, d_y\}$
Ours (Theorem 2)	$C([0,1], \mathbb{R}^2)$	ReLU	$w_{\min} = 3 > \max\{d_x + 1, d_y\}$
Ours (Theorem 3)	$C(\mathcal{K}, \mathbb{R}^{d_y})$	ReLU+Step	$w_{\min} = \max\{d_x + 1, d_y\}$
Ours (Theorem 4)	$L^p(\mathcal{K}, \mathbb{R}^{d_y})$	conti. nonpoly‡	$w_{\min} \leq \max\{d_x + 2, d_y + 1\}$

ICLR 2020, Min width for Universal approximation

Park, Lee

- Fully connected NN, alternating composition of affi transformations and activation functions.

§ A set of activation functions Σ

§ L-layer NN 'f' of dimension $d_x \to d_y$ and hidden layer dimensions $d_1, \dots, d_{L-1}$

$$f := t_L \circ \sigma_{L-1} \circ \dots \circ t_2 \circ \sigma_1 \circ t_1$$

§ $t_\ell : \mathbb{R}^{d_{\ell-1}} \to \mathbb{R}^{d_\ell}$ is an affine transformation

§ σ_ℓ is a vector activation, $\varsigma_i^\ell \in \Sigma$.

$$\sigma_\ell(x_1, \dots, x_{d_\ell}) = (\varsigma_1^\ell(x_1), \dots, \varsigma_{d_\ell}^\ell(x_{d_\ell}))$$

§ width of NN $\omega := \max\{d_1, \dots, d_{L-1}\}$

escription of Universal Approximation:

$L^P(X,Y)$

ς-nets of width ω are dense in $C(X,Y)$, if for any

$L^P(X,Y)$

$f^* \in C(X,Y)$ and $\varepsilon > 0$, $\exists$ a ς-net f of width ω s.t.

$\|f^* - f\|_p \leq \varepsilon$

$$\|f^* - f\|_\infty \leq \varepsilon.$$