Friedrich-Alexander-Universität
Naturwissenschaftliche Fakultät

Friedrich-Alexander-Universität
DYNAMICS, CONTROL, MACHINE LEARNING AND NUMERICS

# MULTILAYER PERCEPTRONS: MULTICLASSIFICATION AND UNIVERSAL APPROXIMATION


**Martín Hernández** Chair for Dynamics, Control, Machine Learning, and Numerics – AvH Professorship. FAU Erlangen-Nürnberg

**Enrique Zuazua** Chair for Dynamics, Control, Machine Learning, and Numerics – AvH Professorship. FAU Erlangen-Nürnberg


## Introduction

This poster presents two main results from [1]:

- We construct ReLU neural networks with fixed width and explicit parameters that achieve simultaneous controllability, ensuring the classification of any dataset with $N$ points and $M$ classes.
- We establish a universal approximation result for $L^p$ functions using neural networks with a fixed width, providing explicit estimates for the required depth (number of layers) for the approximation.

In both cases, the network parameters are explicitly constructed.

## Multilayer Perceptron

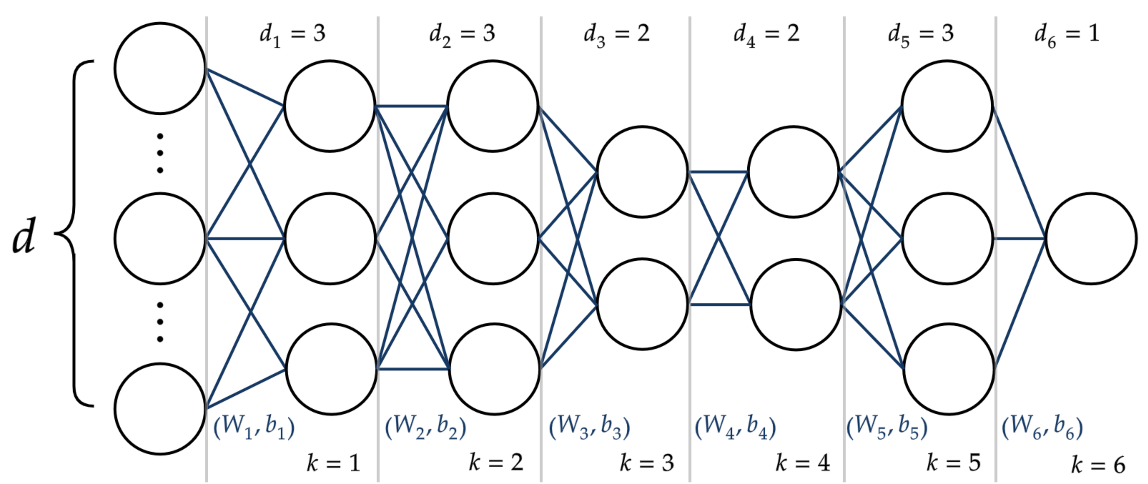
We consider the following neural network architecture:

$$\mathbf{x}^k = \sigma(W_k \cdot \mathbf{x}^{k-1} + b_k), \quad k \in \{1, \dots, L\},$$

where $L \geq 1$, and $\{W_k, b_k\}_{k=1}^L \subset \mathbb{R}^{d_{k+1} \times d_k} \times \mathbb{R}^{d_{k+1}}$, with $d_k \geq 1$. Here, $\sigma$ is the ReLU function $\sigma(x) = \max\{0, x\}$ for $x \in \mathbb{R}$. If $\mathbf{x} \in \mathbb{R}^d$, then:

$$\sigma(\mathbf{x}) = \sigma\left(x_1, \dots, x_d\right)^\top = \left(\sigma(x_1), \dots, \sigma(x_d)\right)^\top.$$

The following diagram illustrates this discrete dynamical system:




Denote by $h^k(x) = W_k \cdot x + b_k$, and consider the input-output map:

$$\phi^L(\mathbf{x}) = \phi^L(\{W_k, b_k\}_{k=1}^L, \mathbf{x}) = (\sigma \circ h^L \circ \cdots \circ \sigma \circ h^1)(\mathbf{x}).$$

Let $\mathcal{W}^L = \{W_k\}_{k=1}^L$ and $\mathcal{B}^L = \{b_k\}_{k=1}^L$, and denote by:

$$N(\mathcal{W}) = \max_{k \in \{1, \dots, L\}} \{d_k\}$$

the neural network width.

**Main question:** Let $d, N, M \geq 1$, and let $\{x_i, y_i\}_{i=1}^N \subset \mathbb{R}^d \times \{1, \dots, M\}$ be a given dataset. Does there exist $L > 0$ and $(\mathcal{W}^L, \mathcal{B}^L)$ such that:

$$\phi^L(x_i) = y_i \quad \text{for every } i \in \{1, \dots, N\}?$$

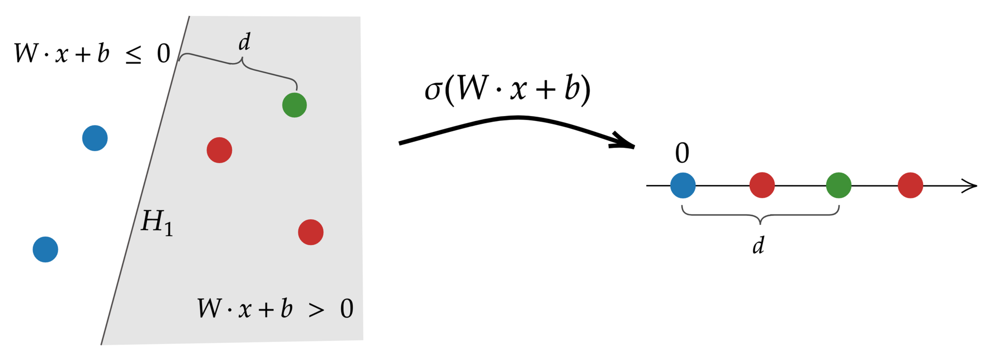
This is referred to as *simultaneous controllability* or *finite sample memorization*.

## Dynamics Interpretation

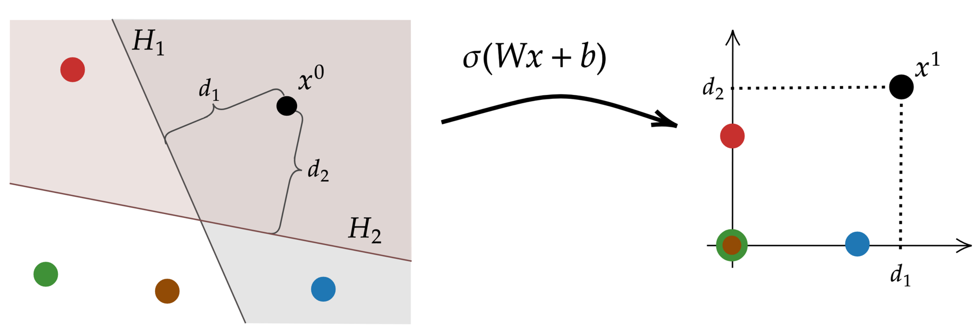
If $W \in \mathbb{R}^{1 \times 2}$ and $b \in \mathbb{R}$, then

$$H(W, b) = \{x \in \mathbb{R}^2 : W \cdot x + b = 0\},$$

defines a hyperplane.




In the case where $(w_1, w_2)^T = W \in \mathbb{R}^{2 \times 2}$ and $(b_1, b_2)^T = b \in \mathbb{R}^2$, they define two hyperplanes $H_1(w_1, b_1)$ and $H_2(w_2, b_2)$.




**Different regions are mapped to different locations, and one region collapses to a single point.**

## Main Results

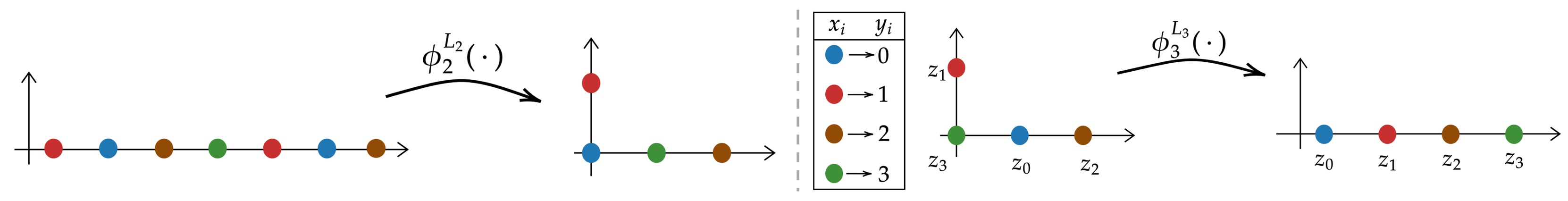
**Theorem 1** (Simultaneous Controllability): Consider integers $d, N, M \geq 1$ and a dataset $\{x_i, y_i\}_{i=1}^N \subset \mathbb{R}^d \times \{1, \dots, M\}$. For $L = 2N + 4M - 1$ and $N(\mathcal{W}) = 2$, there exist parameters $\mathcal{W}^L$ and $\mathcal{B}^L$ such that the input-output map satisfies:

$$\phi^L(\mathcal{W}^L, \mathcal{B}^L, x_i) = y_i, \qquad \text{for every } i \in \{1, \dots, N\}.$$

Moreover, this result cannot be achieved with a width of $1$.

**Proof:** The proof consists of 4 steps:

Step 1 We define $\phi_1^{L_1}$ that projects $d$-dimensional points into $1$-dimensional points.
Step 2 We define $\phi_2^{L_2}$ that collapses points of the same class into a single point.
Step 3 We define $\phi_3^{L_3}$ that sorts the data based on the labels.
Step 4 We define $\phi_4^{L_4}$ that maps the sorted data to their respective labels.




Finally, $\phi^L = (\phi_4^{L_4} \circ \phi_3^{L_3} \circ \phi_2^{L_2} \circ \phi_1^{L_1})$ satisfies simultaneous controllability. □

**Theorem 2** (Universal Approximation Theorem for $L^p$): Let $1 \leq p < \infty$, $d \geq 1$ be an integer, and $\Omega \subset \mathbb{R}^d$ a bounded domain. For any $f \in L^p(\Omega; \mathbb{R}_+)$ and $\varepsilon > 0$, there exist a depth $\mathcal{L} = \mathcal{L}(\varepsilon) \geq 1$ and parameters $\mathcal{W}^\mathcal{L}$ and $\mathcal{B}^\mathcal{L}$ such that the input-output map $\phi^\mathcal{L}$ with $N(\mathcal{W}) = d + 1$ satisfies:

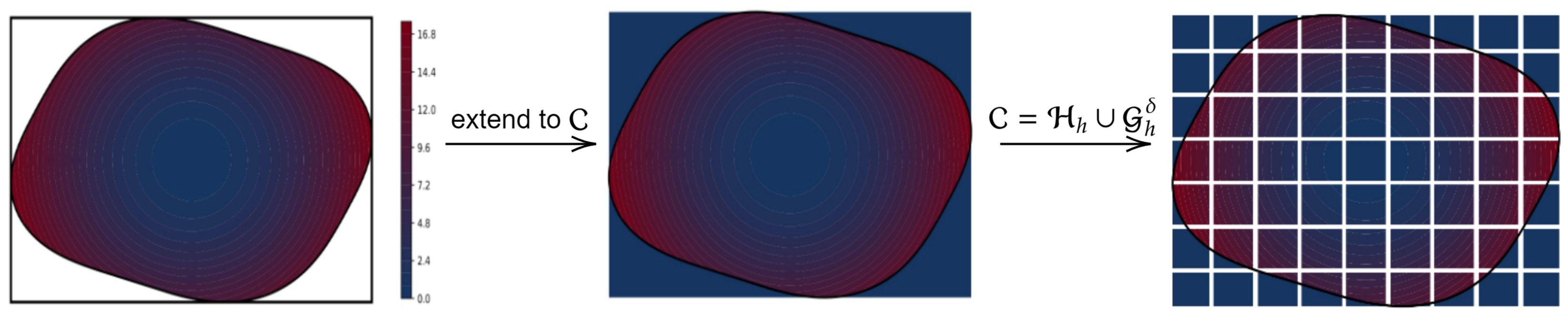
$$\|\phi^\mathcal{L}(\mathcal{W}^\mathcal{L}, \mathcal{B}^\mathcal{L}, \cdot) - f(\cdot)\|_{L^p(\Omega;\mathbb{R}_+)} < \varepsilon.$$

Additionally, for all $f(\cdot) \in W^{1,p}(\Omega; \mathbb{R}_+)$, we have:

$$\mathcal{L}(\varepsilon) \leq C \|f(\cdot)\|_{W^{1,p}(\Omega;\mathbb{R}_+)}^{dp} \varepsilon^{-dp}, \tag{1}$$

where $C$ is a positive constant independent of $f$ and $\varepsilon$.

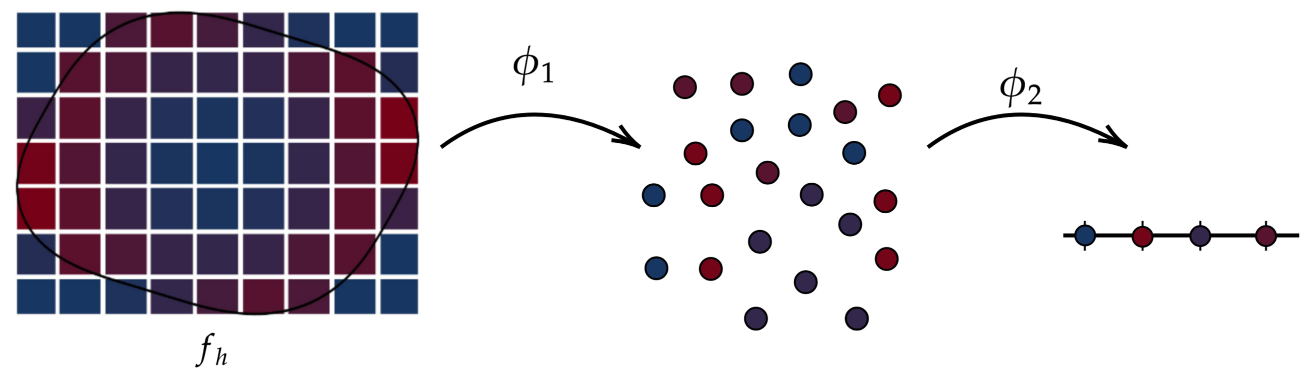
**Proof:** Two-step approximation:




Let

$$f_h(x) = \sum_{H \in \mathcal{H}_h} f_H \chi_H(x), \quad \text{where} \quad f_H := \frac{1}{m_d(H)} \int_H f(x)\, dx,$$

for each $H \in \mathcal{H}_h$. Then, there exists $h_1 > 0$ such that for all $h < h_1$, we have $\|f - f_h\|_{L^p(\mathcal{C};\mathbb{R}_+)} < \varepsilon/2$. Next, we construct two neural networks such that:




We define $\phi^\mathcal{L} = \phi_2 \circ \phi_1$ and show that:

$$\|f_h - \phi^\mathcal{L}\|_{L^p(\mathcal{H};\mathbb{R}_+)} = 0 \quad \text{and} \quad \|f_h - \phi^\mathcal{L}\|_{L^p(\mathcal{G}_h^\delta;\mathbb{R}_+)} < \varepsilon/2.$$

Finally, we deduce:

$$\|f - \phi^\mathcal{L}\|_{L^p(\Omega;\mathbb{R}_+)} \leq \|f - f_h\|_{L^p(\mathcal{C};\mathbb{R}_+)} + \|f_h - \phi^\mathcal{L}\|_{L^p(\mathcal{C};\mathbb{R}_+)} < \varepsilon. \qquad \square$$

## Remarks

- Our work is motivated by [2], where simultaneous controllability results and the UAT were proven using a geometrical interpretation of NODEs.
- In [1], simultaneous controllability is also proven when labels are in $\mathbb{R}^m$, as well as the universal approximation for functions in $L^p(\Omega; \mathbb{R}_+^m)$. In both cases, the parameters are explicitly characterized.
- The explicit parameters can be used for classification problems; **see** ⇒
- The neural network width in Theorem 2 is near optimal. In [3], it was proven that the UAT does not hold for networks with a width less than $d$.

## Selected publications


[1] Hernández, M., Zuazua, E. (2024). **Deep Neural Networks: Multi-Classification and Universal Approximation.** arXiv:2409.06555v1

[2] Ruiz-Balet, D., Zuazua, E. (2023). **Neural ODE Control for Classification, Approximation, and Transport.** SIAM Rev.,65(3):735–773.

[3] Lu, Z., Pu, H., Wang, F., Hu, Z., Wang, L. (2017). **The expressive power of neural networks: A view from the width .**


Department MATHEMATIK


Unterstützt von / Supported by
Alexander von Humboldt Stiftung / Foundation