Multilayer Perceptrons: Multiclassification and Universal Approximation

Martín Hernández
joint work with E. Zuazua.
FAU, Department of Mathematics.
martin.hernandez@fau.de

08/2024

Friedrich-Alexander-Universität
Naturwissenschaftliche Fakultät
Friedrich-Alexander-Universität
DYNAMICS, CONTROL, MACHINE LEARNING AND NUMERICS
DAAD
Deutscher Akademischer Austauschdienst
German Academic Exchange Service
Agencia Nacional de Investigación y Desarrollo
TRR 154

Overview

1. Deep neural network architecture

2. Simultaneous controllability

3. Universal approximation theorem

Deep neural network architecture

Multilayer perceptron

We consider the neural network architecture

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

where $L \geq 1$, $\{W_k, b_k\}_{k=1}^{L} \subset \mathbb{R}^{d_{k+1} \times d_k} \times \mathbb{R}^{d_{k+1}}$ with $d_k \geq 1$.

Multilayer perceptron

We consider the neural network architecture

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

where $L \geq 1$, $\{W_k, b_k\}_{k=1}^{L} \subset \mathbb{R}^{d_{k+1} \times d_k} \times \mathbb{R}^{d_{k+1}}$ with $d_k \geq 1$.

Here σ is the ReLu function $\sigma(x) = \max\{0, x\}$ for $x \in \mathbb{R}$.

Multilayer perceptron

We consider the neural network architecture

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

where $L \geq 1$, $\{W_k, b_k\}_{k=1}^{L} \subset \mathbb{R}^{d_{k+1} \times d_k} \times \mathbb{R}^{d_{k+1}}$ with $d_k \geq 1$.

Here σ is the ReLu function $\sigma(x) = \max\{0, x\}$ for $x \in \mathbb{R}$. If $\mathbf{x} \in \mathbb{R}^d$, then

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

Multilayer perceptron

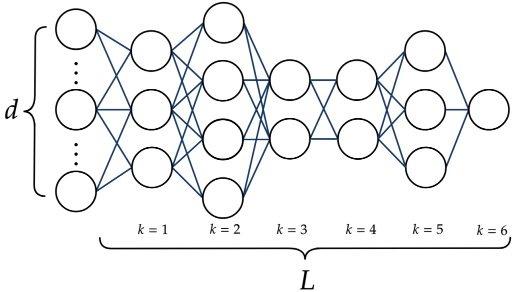
We consider the neural network architecture

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

where $L \geq 1$, $\{W_k, b_k\}_{k=1}^{L} \subset \mathbb{R}^{d_{k+1} \times d_k} \times \mathbb{R}^{d_{k+1}}$ with $d_k \geq 1$.

Here σ is the ReLu function $\sigma(x) = \max\{0, x\}$ for $x \in \mathbb{R}$. If $\mathbf{x} \in \mathbb{R}^d$, then

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



We denote by $N(\mathcal{W}) = \max_{k \in \{1,\ldots,L\}} \{d_k\}$ the neuronal network width.

Finite sample memorization

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

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

Finite sample memorization

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

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

Let $\mathcal{W}^L = \{W_k\}_{k=1}^L$ and $\mathcal{B}^L = \{b_k\}_{k=1}^L$.

Finite sample memorization

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

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

Let $\mathcal{W}^L = \{W_k\}_{k=1}^L$ and $\mathcal{B}^L = \{b_k\}_{k=1}^L$.

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

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

This is *simultaneous controllability* or *finite sample memorization*.

Simultaneous controllability

Simultaneous controllability theorem

Theorem 1: Simultaneous controllability

Consider the integers d, N, $M \geq 1$ and a dataset $\{x_i, y_i\}_{i=1}^N \subset \mathbb{R}^d \times \{1, \ldots, M\}$. Then, 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

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

Simultaneous controllability theorem

Theorem 1: Simultaneous controllability

Consider the integers d, N, $M \geq 1$ and a dataset $\{x_i, y_i\}_{i=1}^{N} \subset \mathbb{R}^d \times \{1, \ldots, M\}$. Then, 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

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

Moreover, this result cannot be achieved with width 1.

Simultaneous controllability theorem

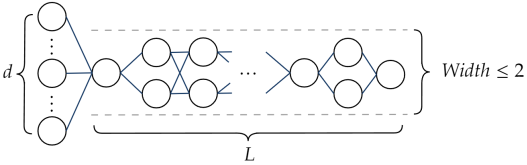
Theorem 1: Simultaneous controllability

Consider the integers d, N, $M \geq 1$ and a dataset $\{x_i, y_i\}_{i=1}^N \subset \mathbb{R}^d \times \{1, \dots, M\}$. Then, 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 width 1.

The neural network of the theorem corresponds to the following



- $N(\mathcal{W}) = \max_{k \in \{0, \dots, L-1\}} \{d_k\} = 2.$
- Depth $L = 2N + 4M - 1$

Geometric analysis of dynamics

Let us analyze $\sigma(Wx + b)$.

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

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

define a hyperplane.

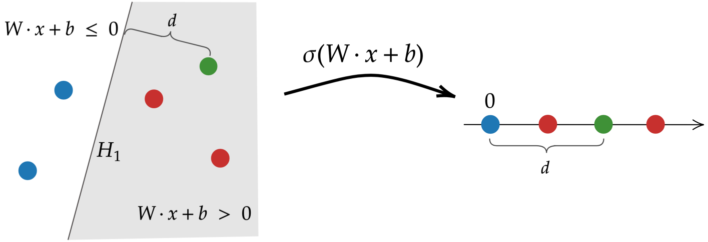
Geometric analysis of dynamics

Let us analyze $\sigma(Wx + b)$.

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

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

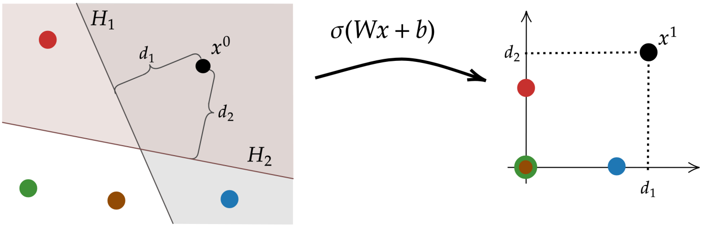
define a hyperplane.



All points to the left of the hyperplane H_1 collapse to zero.

Geometric analysis of dynamics

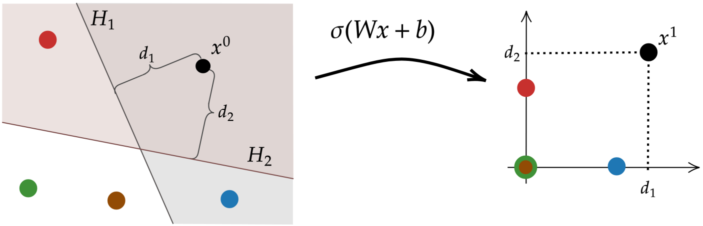
In the case that $(w_1, w_2)^T = W \in \mathbb{R}^{2\times 2}$ and $(b_1, b_2)^T = b \in \mathbb{R}^2$ they define two hyperplanes $H_1(w_1, b_1)$ and $H_2(w_2, b_2)$.



Different regions are mapped to different locations.

Geometric analysis of dynamics

In the case that $(w_1, w_2)^T = W \in \mathbb{R}^{2\times 2}$ 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.

Key idea: Construct the parameters such that in each iteration, points of the same color collapse in the same point.

Sketch of the construction of the parameter

We construct the parameters in four steps.

Sketch of the construction of the parameter

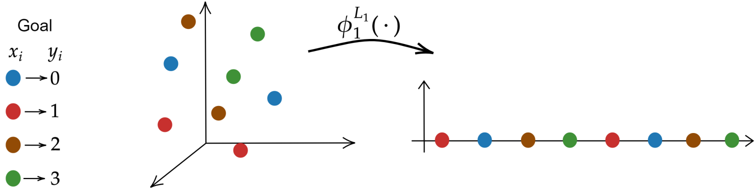
We construct the parameters in four steps.

(1) **Data preconditioning:** Projection from d dimensions into a single dimension.

Sketch of the construction of the parameter

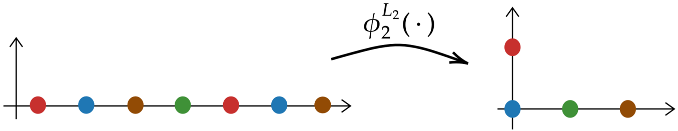
We construct the parameters in four steps.

(1) **Data preconditioning:** Projection from d dimensions into a single dimension.



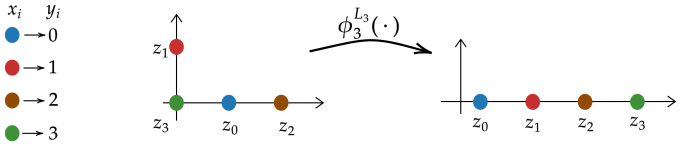
Sketch of the construction of the parameter

(2) **Compression process:** We drive the data from the same class into single points. Defining the map $\phi_2^{L_2}$.



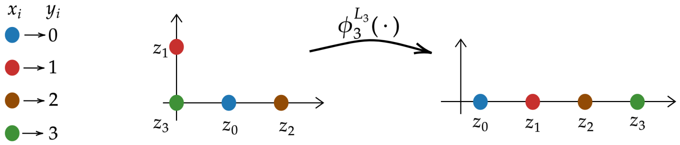
Sketch of the construction of the parameter

(3) **Data sorting:** We sort the data with a map $\phi_3^{L_3}$.

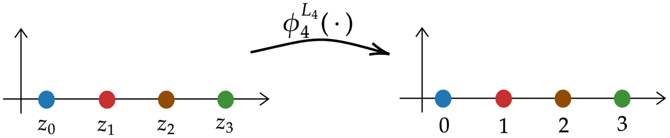


Sketch of the construction of the parameter

(3) **Data sorting:** We sort the data with a map $\phi_3^{L_3}$.

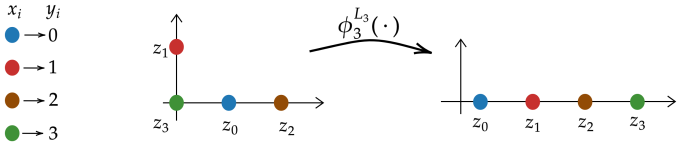


(4) **Mapping to the respective labels:** With a map $\phi_4^{L_4}$ we drive the data to their respective labels.

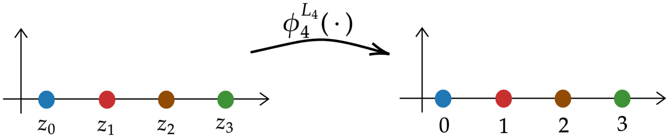


Sketch of the construction of the parameter

(3) **Data sorting:** We sort the data with a map $\phi_3^{L_3}$.



(4) **Mapping to the respective labels:** With a map $\phi_4^{L_4}$ we drive the data to their respective labels.



Finally, the map $\phi^L = (\phi_4^{L_4} \circ \phi_3^{L_3} \circ \phi_2^{L_2} \circ \phi_1^{L_1})$ can memorizes the dataset.

Universal approximation theorem

Universal Approximation Theorem for $L^p(\Omega; \mathbb{R}_+)$

Let $1 \leq p < \infty$, $d \geq 1$ 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 ϕ^L with $N(\mathcal{W}) = d + 1$ satisfies

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

Universal approximation theorem

Universal Approximation Theorem for $L^p(\Omega; \mathbb{R}_+)$

Let $1 \le p < \infty$, $d \ge 1$ 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 $\mathcal{L} = \mathcal{L}(\varepsilon) \ge 1$ and parameters $\mathcal{W}^{\mathcal{L}}$ and $\mathcal{B}^{\mathcal{L}}$ such that the input-output map ϕ^L with $N(\mathcal{W}) = d + 1$ satisfies

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

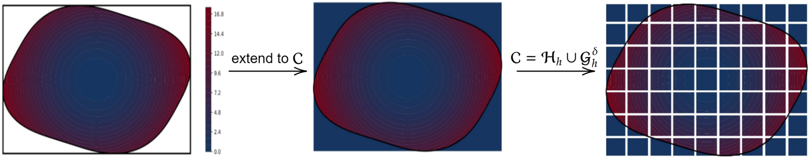
Additionally, for all $f(\cdot) \in W^{1,p}(\Omega; \mathbb{R}_+)$, we have

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

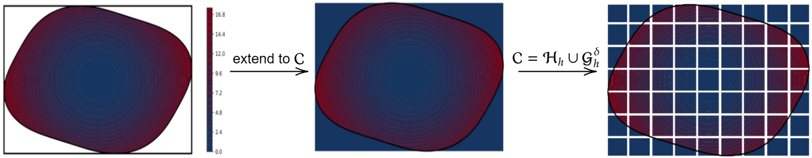
where C is a positive constant independent of f and ε.

Proof: Two step approximation.

Sketch of the proof (Step 1)



Sketch of the proof (Step 1)



Let $\quad f_h(x) = \displaystyle\sum_{H \in \mathcal{H}_h} f_H \chi_H(x),$

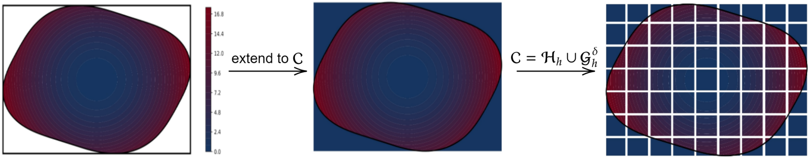
where

$$f_H := \frac{1}{m_d(H)} \int_H f(x)\, dx, \quad \text{for } H \in \mathcal{H}_h.$$

Then,

$$\|f - f_h\|_{L^p(\mathcal{C};\mathbb{R}_+)} \leq \varepsilon/2$$

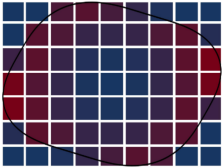
Sketch of the proof (Step 1)



Let $\quad f_h(x) = \sum_{H \in \mathcal{H}_h} f_H \chi_H(x),$

where

$$f_H := \frac{1}{m_d(H)} \int_H f(x)\,dx, \quad \text{for } H \in \mathcal{H}_h.$$

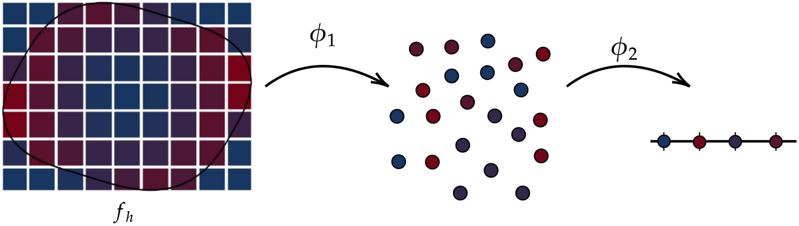


f_h

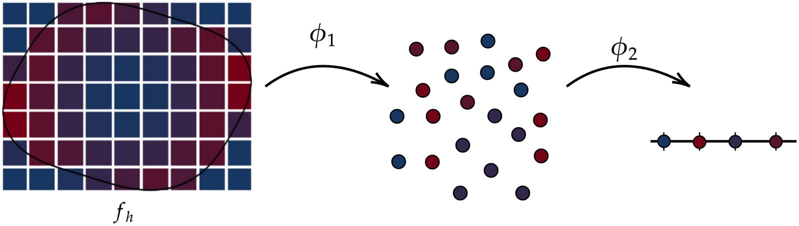
Then,

$$\|f - f_h\|_{L^p(\mathcal{C};\mathbb{R}_+)} \leq \varepsilon/2$$

Sketch of the proof (Step 2)



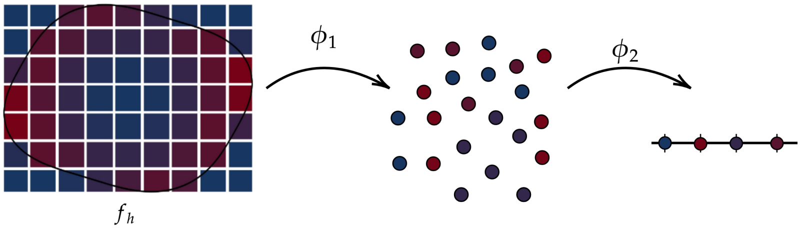
Sketch of the proof (Step 2)



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

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

Sketch of the proof (Step 2)



We define $\phi^{\mathcal{L}} = \phi_2 \circ \phi_1$ and we 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

$$\|f - \phi^{\mathcal{L}}\|_{L^2(\Omega;\mathbb{R}_+)} \leq \|f - f_h\|_{L^p(\mathcal{C};\mathbb{R}_+)} + \|f_h - \phi^{\mathcal{L}}\|_{L^p(\mathcal{C};\mathbb{R}_+)} < \varepsilon.$$

Conclusions

Some novelties of this work are:

Conclusions

Some novelties of this work are:

1. We have proven that any multilayer perceptron with a depth $L = O(N)$ and a width greater than or equal to two satisfies the simultaneous controllability property.

Conclusions

Some novelties of this work are:

1. We have proven that any multilayer perceptron with a depth $L = O(N)$ and a width greater than or equal to two satisfies the simultaneous controllability property.

2. Our geometric analysis employed in our proofs departs from existing techniques.

Conclusions

Some novelties of this work are:

1. We have proven that any multilayer perceptron with a depth $L = O(N)$ and a width greater than or equal to two satisfies the simultaneous controllability property.

2. Our geometric analysis employed in our proofs departs from existing techniques.

3. This explicit construction in the UAT allows us to estimate the number of layers $\mathcal{L}(\varepsilon)$ to approximate a given function $f \in W^{1,p}(\Omega)$.

Conclusions

Some novelties of this work are:

1. We have proven that any multilayer perceptron with a depth $L = O(N)$ and a width greater than or equal to two satisfies the simultaneous controllability property.

2. Our geometric analysis employed in our proofs departs from existing techniques.

3. This explicit construction in the UAT allows us to estimate the number of layers $\mathcal{L}(\varepsilon)$ to approximate a given function $f \in W^{1,p}(\Omega)$.

Thanks for your attention.

References

D. Ruiz-Balet and E. Zuazua
Neural ode control for classification, approximation, and transport
arXiv:2104.05278, 2021.

S. Park, C. Yun, J. Lee, and J. Shin.
Minimum width for universal approximation
arXiv:2006.08859, 2020.

Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang.
The expressive power of neural networks: A view from the width.
arXiv:1709.02540, 2017.

G. Vardi, G. Yehudai, and O. Shamir.
On the optimal memorization power of ReLU neural networks.
arXiv:2110.03187, 2021.