Minimum Width for Universal Approximation using ReLU Networks on Compact Domain

Namjun Kim[1], Chanho Min[2], Sejun Park[1]

[1]Korea University, [2]Ajou University

TL;DR. We find the exact minimum width and the lower bound for universal approximation using ReLU networks on compact domain.

Motivation

Universal Approximation (UA). For any continuous function f^* and error $\varepsilon > 0$, we want to find a neural network f such that

$$\text{distance}(f^*, f) \leq \varepsilon.$$

- Two popular choices for distance:

$$L^p \text{ distance} : \|f^* - f\|_p, \quad \text{Uniform distance} : \sup_x \|f^*(x) - f(x)\|_\infty.$$

Classical Results. Mainly focus on shallow and wide networks.

Theorem ([Hornik+89; Cybenko89; Leshno+93; Pinkus99]). Two-layer neural networks with a non-polynomial activation function are universal approximators in both L^p and uniform distance.

- Namely, the minimum depth for universal approximation is exactly two.
- The universal approximation property of deep and narrow networks has been studied as a dual problem.

Problem. The minimum width enabling universal approximation?

Summary of Bounds on Minimum Width

Ref.	Distance	Function class	Activation σ	Exact minimum width
Park+21	L^p	$C(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$	ReLU	$w_{\min} = \max\{d_x+1, d_y\}$
Cai23	L^p	$C([0,1]^{d_x}, \mathbb{R}^{d_y})$	Leaky-ReLU	$w_{\min} = \max\{d_x, d_y, 2\}$
Thm. 1.	L^p	$C([0,1]^{d_x}, \mathbb{R}^{d_y})$	ReLU	$w_{\min} = \max\{d_x, d_y, 2\}$
Thm. 2.	L^p	$C([0,1]^{d_x}, \mathbb{R}^{d_y})$	ReLU-Like†	$w_{\min} = \max\{d_x, d_y, 2\}$

† is an activation function similar to ReLU such as Leaky-ReLU, GELU, and Mish.

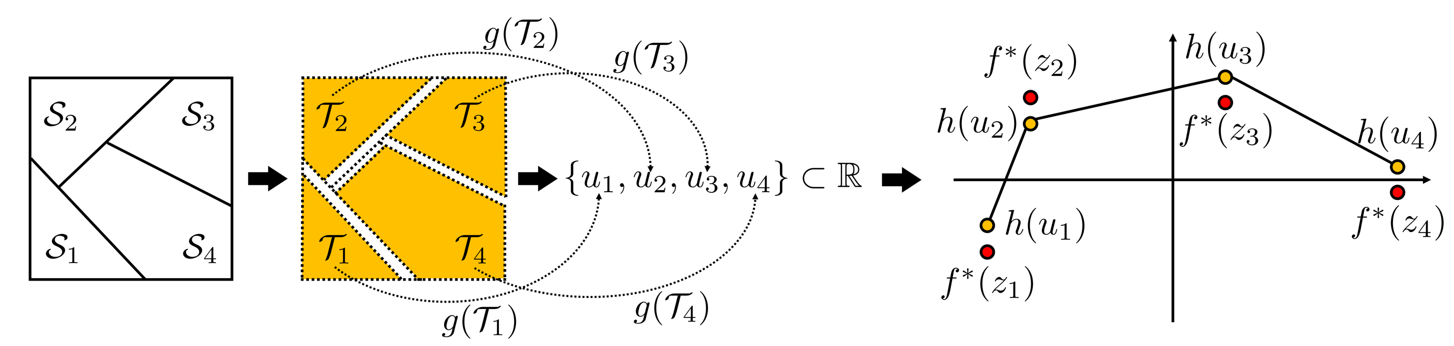
Ref.	Distance	Function class	Activation σ	Lower bound
Park+21	Uniform	$C([0,1], \mathbb{R}^2)$	ReLU	$w_{\min} > 2 = \max\{d_x+1, d_y\}$
Cai23	Uniform	$C([0,1], \mathbb{R}^2)$	Leaky-ReLU	$w_{\min} > 2 = \max\{d_x+1, d_y\}$
Thm. 3.	Uniform	$C([0,1]^{d_x}, \mathbb{R}^{d_y})$	Conti. monotone	$w_{\min} \geq d_y + 1_{d_x < d_y \leq 2d_x}$

Contributions.

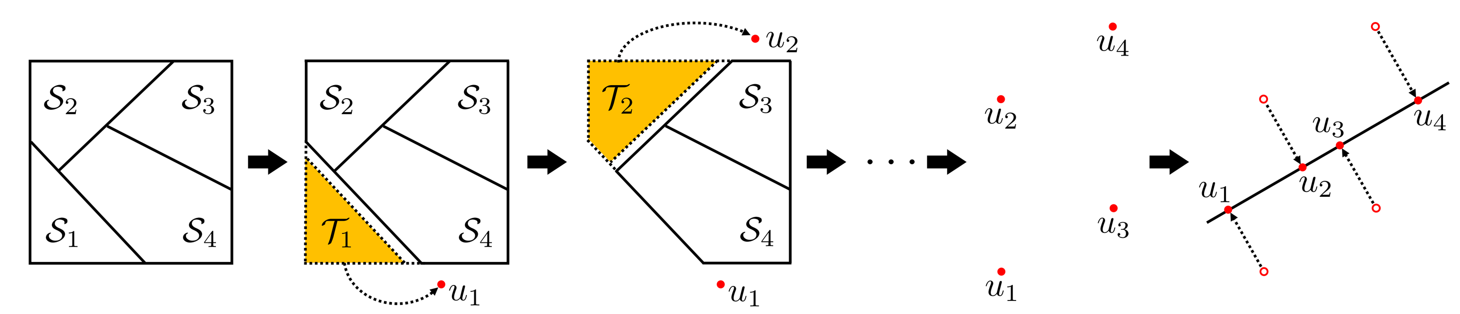
- **Theorem 1.** For L^p distance and ReLU networks on compact domain, $w_{\min} = \max\{d_x, d_y, 2\}$ for UA.
  - This shows a dichotomy between bounded and unbounded domains: $w_{\min} = \max\{d_x+1, d_y\}$ when the domain is unbounded.
- **Theorem 2.** $w_{\min} = \max\{d_x, d_y, 2\}$ for the networks using any of ReLU-Like activation functions, which generalizes the previous result for Leaky-ReLU networks.
- **Theorem 3.** For uniform distance and networks using continuous monotone activation function (e.g., ReLU, Leaky-ReLU) $w_{\min} \geq d_y + 1$ if $d_x < d_y \leq 2d_x$.
  - This generalizes the previous result: $w_{\min} \geq d_y + 1$ for ReLU networks if $d_x = 1$ and $d_y = 2$.

Proof Sketch: Achieving Exact Minimum Widths

Idea. Given a partition $\{\mathcal{S}_1, \dots, \mathcal{S}_k\}$ of the domain with $\text{diam}(\mathcal{S}_i)$ is small, map almost all of each $\mathcal{S}_i$ (i.e., $\mathcal{T}_i$ below) to an approximate target vector.



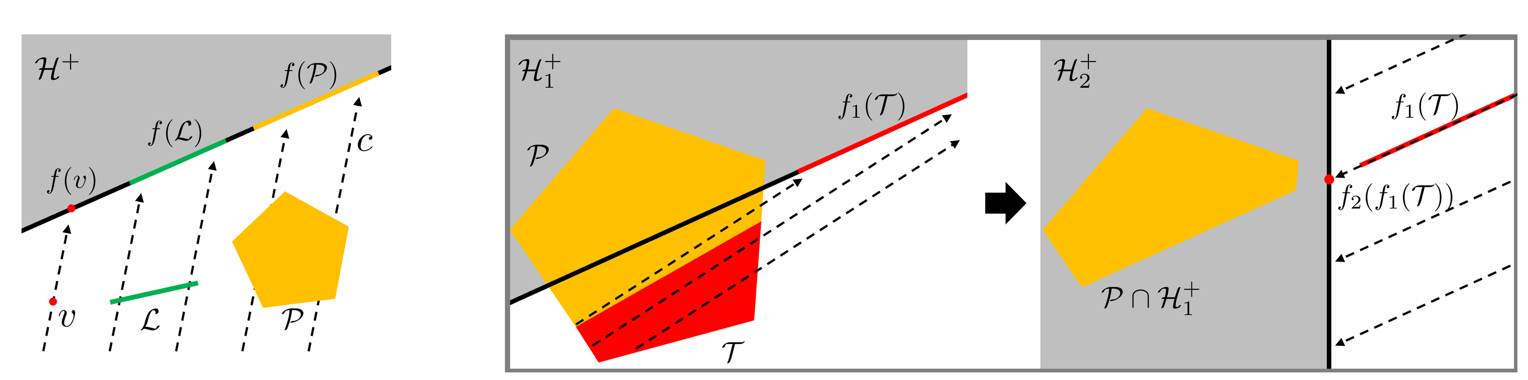
Encoder. Our encoder iteratively maps $\mathcal{T}_1, \dots, \mathcal{T}_k$ to distinct scalar values using width $\max\{d_x, 2\}$ ReLU networks, using Lemma 1.



Lemma 1. For any $d_x \in \mathbb{N}$, a compact set $\mathcal{K} \subset \mathbb{R}^{d_x}$, $a, c \in \mathbb{R}^{d_x}$ such that $a^\top c > 0$, and $b \in \mathbb{R}$, there exists a two-layer ReLU network $f : \mathcal{K} \to \mathbb{R}^d_x$ of width d_x such that

$$f(x) = \begin{cases} x & \text{if } a^\top x + b \geq 0 \\ x - \frac{a^\top x + b}{a^\top c} \times c & \text{if } a^\top x + b < 0 \end{cases}.$$

- Using a ReLU network of width d_x, preserve the points in the half-space $\mathcal{H}^+ = \{x \in \mathbb{R}^{d_x} : a^\top x + b \geq 0\}$ and project points not in $\mathcal{H}^+$ to the boundary along the direction c.



Decoder. Our decoder maps each scaler value generated by the encoder to an approximate target vector, which can be implemented by a ReLU network of width $\max\{d_y, 2\}$.

- Overall, for L^p distance and ReLU networks on compact domain, $w_{\min} \leq \max\{d_x, d_y, 2\}$ for UA.

Matching lower bound. $w_{min} \geq \max\{d_x, d_y, 2\}$ is rather straightforward.

- Width either $d_x - 1$ or $d_y - 1$: lose input/output information.
- Width 1: cannot approximate non-monotonic function.

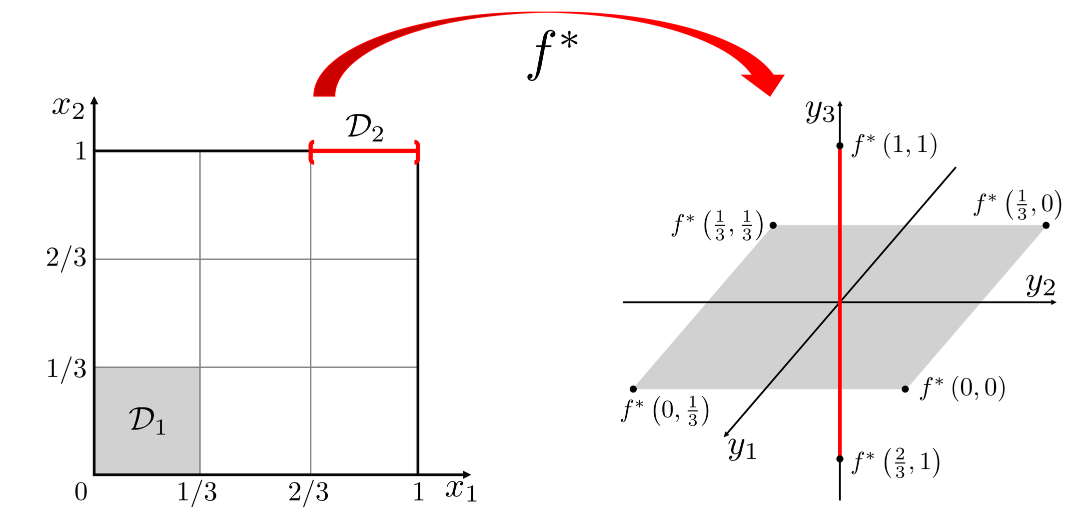
ReLU-Like activation functions. ReLU can be approximated by a width 1 network using any of ReLU-Like activation functions.

- Thus, our proof techniques can be generalized to networks using any of ReLU-Like activation functions; $w_{\min} = \max\{d_x, d_y, 2\}$ for UA.

Proof Sketch: Lower Bound on Minimum Width

We assume our activation function σ is a continuous injection; this easily generalizes to continuous monotone functions.

Proof by Contradiction. Our counterexample $f^* : [0,1]^{d_x} \to \mathbb{R}^{d_y}$ is defined as follows:



For $r = d_y - d_x$, $x = (x_1, \dots, x_{d_x}) \in [0,1]^{d_x}$, $\mathcal{D}_1 = [0, 1/3]^{d_x}$, and $\mathcal{D}_2 = [2/3, 1]^r \times \{1\}^{d_x - r}$,

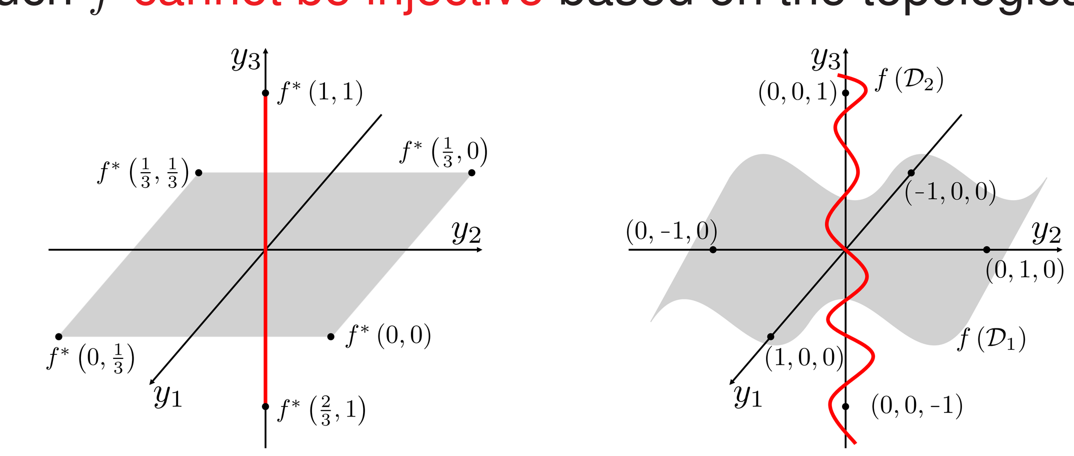
$$f^*(x) = \begin{cases} (1 - 6x_1, 1 - 6x_2, \dots, 1 - 6x_{d_x}, 0, \dots, 0) & \text{if } x \in \mathcal{D}_1 \\ (0, \dots, 0, 6x_1 - 5, 6x_2 - 5, \dots, 6x_r - 5) & \text{if } x \in \mathcal{D}_2 \\ g^*(x) & \text{otherwise} \end{cases},$$

where g^* is some continuous function that makes f^* continuous.

Network f Approximating f^*. Suppose for a contradiction that there is a σ network f of width d_y such that $\|f^* - f\|_\infty$ is small enough.

- Since φ is injective, f is also an injection, i.e., $f(\mathcal{D}_1) \cap f(\mathcal{D}_2) = \emptyset$.

However, such f cannot be injective based on the topological argument.



Minimum Width for Recurrent Neural Networks

Our encoder & decoder can also be applied to recurrent neural networks and bidirectional recurrent neural networks.

Networks	Distance	Function class	Activation σ	Upper / lower bounds
RNN	L^p	$C([0,1]^{d_x \times T}, \mathbb{R}^{d_y \times T})^{\dagger,\ddagger}$	ReLU ReLU-Like	$w_{\min} = \max\{d_x, d_y, 2\}$ $w_{\min} = \max\{d_x, d_y, 2\}$
BRNN	L^p	$C([0,1]^{d_x \times T}, \mathbb{R}^{d_y \times T})^{\dagger}$	ReLU ReLU-Like	$w_{\min} \leq \max\{d_x, d_y, 2\}$ $w_{\min} \leq \max\{d_x, d_y, 2\}$

† consists of all continuous functions with length T from $[0,1]^{d_x}$ to $\mathbb{R}^{d_y}$.
‡ are *past-dependent*; the t-th output is a function of the first to t-th inputs.

- **RNNs.** For L^p distance, $w_{\min} = \max\{d_x, d_y, 2\}$ for UA.
- **BRNNs.** For L^p distance, $w_{\min} \leq \max\{d_x, d_y, 2\}$ for UA.