Minimum Width for Deep, Narrow MLP: A Diffeomorphism Approach

Geonho Hwang
Department of Mathematical Sciences
Gwangju Institute for Science and Technology
Gwangju, Buk-gu 61005
hgh2134@gist.ac.kr

Abstract

Recently, there has been a growing focus on determining the minimum width requirements for achieving the universal approximation property in deep, narrow Multi-Layer Perceptrons (MLPs). Among these challenges, one particularly challenging task is approximating a continuous function under the uniform norm, as indicated by the significant disparity between its lower and upper bounds. To address this problem, we propose a framework that simplifies finding the minimum width for deep, narrow MLPs into determining a purely geometrical function denoted as $w(d_x, d_y)$. This function relies solely on the input and output dimensions, represented as d_x and d_y, respectively. To achieve this, we first demonstrate that deep, narrow MLPs, when provided with a small additional width, can approximate any C^2-diffeomorphism. Subsequently, using this result, we prove that $w(d_x, d_y)$ equates to the optimal minimum width required for deep, narrow MLPs to achieve universality. By employing the aforementioned framework and the Whitney embedding theorem, we provide an upper bound for the minimum width, given by $\max(2d_x + 1, d_y) + \alpha(\sigma)$, where $0 \leq \alpha(\sigma) \leq 2$ represents a constant depending explicitly on the activation function. Furthermore, we provide novel optimal values for the minimum width in several settings, including $w(2, 2) = w(2, 3) = 4$.

1 Introduction

The *universal approximation property* (UAP) denotes the capability of neural networks to approximate a specific class of functions, forming the foundation for their efficacy and garnering significant interest in the research community. Among the extensively explored subjects is the investigation of the UAP of *deep, narrow multilayer perceptrons* (MLPs) characterized by restricted width and an arbitrary number of layers. Given the practical application of MLPs involving modest widths and more than two layers, comprehending the UAP of such networks has emerged as a pivotal focus.

In this context, a series of papers has aimed to determine the *minimum width*, representing the necessary and sufficient width for achieving the UAP. The minimum width depends on input dimension d_x, output dimension d_y, activation functions, and the employed norm. Out of various norms, one particularly important and challenging task is to determine the minimum width under the uniform norm. In this paper, our focus centers on the universal approximation of continuous functions under the uniform norm. The earlier findings concerning uniform approximation are summarized in Table 1. In general, research on determining the minimum width for approximations under the uniform norm, using continuous activation functions, suggests a range between $\max(d_x + 1, d_y)$ and $d_x + d_y$. This discrepancy highlights a significant gap.

In this context, we present novel upper and lower bounds for the minimum width required by deep, narrow MLPs to possess the UAP. Our support for these bounds unfolds in two steps. Initially, we

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

approximate diffeomorphisms using deep, narrow MLPs, building upon the concept of the UAP of invertible neural networks. Subsequently, we approximate arbitrary continuous functions through a composition of linear transformations and a diffeomorphism. We characterize this process by revealing that the minimum width of Leaky-ReLU MLPs corresponds to a geometrical function denoted as $w(d_x, d_y)$. This function represents the required dimension of diffeomorphisms necessary to approximate arbitrary continuous functions given an input dimension of d_x and an output dimension of d_y.

Building upon these statements, we establish upper and lower bounds. By applying results from geometric topology, we prove that MLPs with width $\max(2d_x + 1, d_y)$ can approximate any continuous function. Additionally, by utilizing more refined results of topological algebra, we can improve this bound in certain specific cases. For example, when the input dimension is two and the output dimension is three, the optimal minimum width is **exactly** four, and when both the input and output dimensions are two, the optimal minimum width remains four. These findings highlight the practical utility of our framework.

Our contributions can be summarized as follows:

- We prove that deep, narrow MLPs with a width of d, employing the Leaky-ReLU activation function, can approximate any C^2-diffeomorphisms on $\mathbb{R}^d$. Furthermore, for more general activation functions, we demonstrate that deep, narrow MLPs with width $d + 1$ using ReLU and $d + 2$ employing a general activation function, respectively, can approximate any C^2-diffeomorphisms uniformly on $\mathbb{R}^d$.
- We propose the purely topological quantity $w(d_x, d_y)$, representing the optimal minimum width for achieving the UAP of deep, narrow MLPs employing the Leaky-ReLU activation function.
- Building upon the aforementioned results, we prove that deep, narrow MLPs with a width of $\max(2d_x + 1, d_y) + \alpha(\sigma)$ can approximate any continuous function in $C(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$ uniformly on a compact domain, where $0 \leq \alpha(\sigma) \leq 2$ is a constant depending on the activation function.
- We prove that when the input dimension is $2k$ and the output dimension is $4k - 1$, the optimal minimum width is $4k$.
- We demonstrate that a width of 4 is the optimal minimum width for deep, narrow MLPs to approximate arbitrary continuous function mapping $[0, 1]^2$ to $\mathbb{R}^2$.

2 Related Works

In this section, we explore previous research concerning the universal approximation property of neural networks. Initial studies primarily focused on two-layered MLPs. Cybenko (1989) demonstrated that two-layered MLPs with sigmoidal activation functions possess the UAP for approximating continuous functions. Later, Leshno et al. (1993) expanded the scope of activation functions to more general ones.

Beyond two-layered MLPs, extensive investigations have explored the UAP of deep, narrow MLPs. For instance, Lu et al. (2017) showed that deep, narrow MLPs with a width of $d_x + 4$ and ReLU activation functions possess the UAP in $L_1([0, 1]^{d_x}, \mathbb{R})$, while they do not have the UAP with a width of d_x. This early research paved the way for further studies that narrowed down the minimum width range and expanded the application scope of theories. Hanin (2019) extended the study to encompass arbitrary output dimensions d_y: The minimum width lies between $d_x + 1$ and $d_x + d_y$ for the ReLU activation function. Johnson (2019) demonstrated that a width of d_x is insufficient to achieve the UAP in continuous function spaces for an activation function that can be approximated by increasing functions. Kidger & Lyons (2020) proved that a dimension of $d_x + d_y + 1$ is sufficient for activation functions under a weak condition, and $d_x + d_y + 2$ is sufficient for polynomials. Furthermore, Cai (2023) explored the lower bound $\max(d_x, d_y)$ of the minimum width for arbitrary activation functions and proved that the UAP can be achieved with $\max(d_x, d_y)$ using the floor function and an activation function having the universal ordering of extrema property. Tabuada & Gharesifard (2022) addressed the UAP of deep, narrow residual networks. Recently, Li et al. (2023) claimed that the upper bound could be reduced to $\max(d_x + 1, d_y) + \mathbf{1}_{d_x+1=d_y}$. On the other hand, Shen et al. (2022); Hong & Kratsios (2024); Liu & Chen (2024) treat the approximation rates of neural networks with low

Table 1: A summary of known results on minimum width for universal approximation of continuous functions. K denotes a compact domain, and "Conti." is short for continuous.

Reference	Domain	Activation σ	Upper / Lower Bounds
Hanin (2019)	$C(K, \mathbb{R}^{d_y})$	ReLU	$d_x + 1 \leq w_{\min} \leq d_x + d_y$
Johnson (2019)	$C(K, \mathbb{R})$	uniformly conti.†	$w_{\min} \geq d_x + 1$
Kidger & Lyons (2020)	$C(K, \mathbb{R}^{d_y})$	conti. nonpoly‡	$w_{\min} \leq d_x + d_y + 1$
	$C(K, \mathbb{R}^{d_y})$	nonaffine poly	$w_{\min} \leq d_x + d_y + 2$
Park et al. (2021)	$C([0,1], \mathbb{R}^2)$	ReLU	$w_{\min} = 3 > \max(d_x + 1, d_y)$
Cai (2023)	$C(K, \mathbb{R}^{d_y})$	Arbitrary	$w_{\min} \geq \max(d_x, d_y)$
Kim et al. (2024)	$C(K, \mathbb{R}^{d_y})$	uniformly conti.†	$w_{\min} \geq d_y + \mathbf{1}_{d_x < d_y \leq 2d_x}$
Ours	$C(K, \mathbb{R}^{d_y})$	Leaky-ReLU	$w_{\min} \leq \max(2d_x + 1, d_y)$
	$C(K, \mathbb{R}^{d_y})$	ReLU	$w_{\min} \leq \max(2d_x + 1, d_y) + 1$
	$C(K, \mathbb{R}^{d_y})$	conti. nonpoly‡	$w_{\min} \leq \max(2d_x + 1, d_y) + 2$
	$C([0,1]^{2k}, \mathbb{R}^{4k-1})$	Leaky-ReLU	$w_{\min} = 4k$
	$C([0,1]^2, \mathbb{R}^2)$	ReLU	$w_{\min} = 4$
	$C([0,1]^2, \mathbb{R}^2)$	Leaky-ReLU	$w_{\min} = 4$
	$C([0,1]^2, \mathbb{R}^2)$	uniformly conti.†	$w_{\min} \geq 4$

† Requires that σ can be uniformly approximated by a sequence of one-to-one functions.
‡ Requires that σ is continuously differentiable at some point z with $\sigma'(z) \neq 0$.

width. In the recurrent setting (Song et al., 2023), investigations into minimal-width deep RNNs have demonstrated that similar universal approximation properties hold, broadening the scope of width-efficient architectures beyond feedforward models.

In addition to the uniform norm, Park et al. (2021) presented the optimal minimum width as $\max(d_x + 1, d_y)$ for deep, narrow MLPs using ReLU activation functions in L_p space. They demonstrated that this is not applicable for the uniform norm, establishing the optimal width for the UAP in $C([0,1], \mathbb{R}^2)$ as three. Kim et al. (2024) proved that the lower bound equals or exceeds $d_y + 1$ if d_y is less than or equal to $2d_x$, and on a compact domain, the minimum width for L_p space is $\max(d_x, d_y, 2)$ when using the ReLU activation function. Rochau et al. (2024) provided an alternative constructive proof of the same result. HERNÁNDEZ & ZUAZUA proved that width 2 is sufficient in the classification setting.

In this paper, we prove that a width of $\max(2d_x + 1, d_y)$ is sufficient for universal approximation. This provides an advantage in cases where d_y is large compared to previous results, which always required an upper bound of approximately $d_x + d_y$. We also prove that when both the input and output dimensions are two, the minimum required width is 4, demonstrating that the $\max(d_x + 1, d_y)$-type results obtained in the L_p norm setting do not apply to the uniform norm. See Table 1 for a summary.

3 Notation and Definition

In this section, we introduce the notations and definitions utilized throughout this paper: $\mathbb{R}$ represents the set of real numbers. $\mathbb{R}_+$ denotes the set of positive real numbers. $\mathbb{N}$ stands for the set of natural numbers, and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. For $a, b \in \mathbb{R}$, $[a, b]$ and (a, b) represent the closed and open intervals from a to b, respectively. $M_{n,m}$ denotes the set of $n \times m$ real matrices with real inputs. $GL(n) \subset M_{n,n}$ represents the set of invertible $n \times n$-matrices. $\mathrm{Aff}_{n,m}$ and IAff_n stand for the sets of affine transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$ and invertible affine transformations from $\mathbb{R}^n$ to $\mathbb{R}^n$, respectively. For a d-dimensional vector $x \in \mathbb{R}^d$, x_i denote the i-th component of x; in other words, $x = (x_1, x_2, \ldots, x_d)$. Additionally, $x_{i:j}$ represents the $(j - i + 1)$-dimensional vector $(x_i, x_{i+1}, \ldots, x_j)$.

3.1 Compact Approximation

Let X and Y be metric spaces. $C(X, Y)$ represents the set of continuous functions from X to Y. For a function $f \in C(X, Y)$ and a non-empty set $X' \subset X$, $f|_{X'}$ denotes the restriction of the function to the domain X'. For a set of functions $\mathcal{A} \subset C(X,Y)$, $\mathcal{A}|_{X'}$ is defined as $\{f|_{X'} : f \in \mathcal{A}\}$. We focus on the uniform approximation of a continuous function on a compact set, defined as follows:

Definition 3.1. *Given two subsets, $\mathcal{A}, \mathcal{B} \subset C(\mathbb{R}^n, \mathbb{R}^m)$, we say that $\mathcal{A}$ **compactly approximates** $\mathcal{B}$ if, for any $f \in \mathcal{B}$, a compact set $K \subset \mathbb{R}^n$, and $\epsilon > 0$, there exists $g \in \mathcal{A}$ such that*

$$\|f - g\|_{\infty,K} := \sup_{x \in K} \|f(x) - g(x)\|_\infty < \epsilon. \tag{1}$$

This is denoted as $\mathcal{A} \succ \mathcal{B}$ or $\mathcal{B} \prec \mathcal{A}$.

The compact approximation relation is transitive: if $\mathcal{A} \succ \mathcal{B}$, and $\mathcal{B} \succ \mathcal{C}$, then, $\mathcal{A} \succ \mathcal{C}$. Additionally, we use the notation $f \prec \mathcal{A}$ to indicate that $\{f\} \prec \mathcal{A}$. For a set of functions $\mathcal{A} \subset C(X,Y)$, $\overline{\mathcal{A}}$ represents the closure with respect to the uniform norm.

3.2 Activation Function

We follow the commonly used condition for activation functions proposed by Kidger & Lyons (2020). Note that this condition permits a linearization of the activation function, as established in Lemma 4.1 of Kidger & Lyons (2020).

Condition 1. *An activation function σ is a C^1-function near $\alpha \in \mathbb{R}$, with $\sigma'(\alpha) \neq 0$.*

We define several activation functions that satisfy Condition 1:

- ReLU: $\mathrm{ReLU}(x) := \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$.
- Leaky-ReLU : $\mathrm{LR}_\beta(x) := \begin{cases} x & \text{if } x \geq 0 \\ \beta x & \text{if } x < 0 \end{cases}$.

Activation functions applied to vectors function as componentwise operators: For $x \in \mathbb{R}^d$,

$$\sigma(x) := (\sigma(x_1), \ldots, \sigma(x_d)). \tag{2}$$

3.3 Deep, Narrow MLP

A set of MLPs, denoted as $\mathcal{N}^\sigma_{d_0,d_1,\ldots,d_N}$, is defined as follows:

$$\mathcal{N}^\sigma_{d_0,d_1,\ldots,d_N} := \left\{\, f : \mathbb{R}^{d_0} \to \mathbb{R}^{d_N} \,\middle|\, W_i \in \mathrm{Aff}_{d_{i-1},d_i}, f(x) = W_N \circ \sigma \circ \cdots \circ \sigma \circ W_1 \right\}. \tag{3}$$

For Leaky-ReLU, an additional parameter β can vary for each layer, resulting in the set $\mathcal{N}^{\mathrm{LR}}_{d_0,d_1,\ldots,d_N}$:

$$\mathcal{N}^{\mathrm{LR}}_{d_0,d_1,\ldots,d_N} := \left\{ W_N \circ \mathrm{LR}_{\beta_{N-1}} \circ \cdots \circ \mathrm{LR}_{\beta_1} \circ W_1 : \mathbb{R}^{d_0} \to \mathbb{R}^{d_N} \,\middle|\, W_i \in \mathrm{Aff}_{d_{i-1},d_i}, \beta_i \in \mathbb{R}_+, \right\}. \tag{4}$$

We define a set of deep, narrow MLPs with input dimension d_x, output dimension d_y, and at most n intermediate dimensions as follows:

$$\Delta^\sigma_{d_x,d_y,n} := \bigcup_{N \in \mathbb{N}_0} \bigcup_{1 \leq d_1, d_2, \ldots, d_N \leq n} \mathcal{N}^\sigma_{d_x,d_1,d_2\ldots,d_N,d_y}. \tag{5}$$

For natural numbers $n \geq m \in \mathbb{N}$, we define the natural projection $p_{n,m} : \mathbb{R}^n \to \mathbb{R}^m$ and the inclusion $q_{m,n}$ as follows:

$$p_{n,m} : (x_1, \ldots, x_n) \mapsto (x_1, \ldots, x_m), \tag{6}$$

and

$$q_{m,n} : (x_1, \ldots, x_m) \mapsto (x_1, \ldots, x_m, 0, \ldots, 0). \tag{7}$$

For any $n \geq d_x, d_y$ and $f \in \Delta^\sigma_{d_x,d_y,n}$, it can be decomposed as:

$$f = p_{n,d_y} \circ g \circ q_{d_x,n}, \tag{8}$$

where $g \in \Delta^\sigma_{n,n,n}$.

Definitions of several geometric concepts including diffeomorphism and embedding, are provided in Appendix A. We define the set of diffeomorphisms as follows:

Definition 3.2 (Diffeomorphism: $\mathcal{D}^r(U)$)**.** *Let $U \subset \mathbb{R}^d$ be an open subset, and let r be a non-negative integer or infinity. $\mathcal{D}^r(U)$ is the set of C^r-diffeomorphisms from U to $\mathbb{R}^d$.*

4 Main Theorem

4.1 Problem Formulation

Our primary objective is to determine the minimum width $w_{\min} \in \mathbb{N}$ such that for any compact set $K \subset \mathbb{R}^n$, a continuous function $f \in C(K, \mathbb{R}^m)$ can be uniformly approximated by $\Delta^{\sigma}_{n,m,w_{\min}}$. In other words, we aim to determine the value $w_{\min}(n, m, \sigma)$ such that

$$w_{min}(n, m, \sigma) := \min \left\{ l \in \mathbb{N} \,\middle|\, C(\mathbb{R}^n, \mathbb{R}^m) \prec \Delta^{\sigma}_{n,m,l} \right\}. \tag{9}$$

4.2 Approximating Diffeomorphisms and Continuous Functions

In this subsection, we begin by demonstrating the capability of deep, narrow MLPs to approximate diffeomorphisms and aim to prove that any continuous function can be approximated by composing linear transformations and diffeomorphisms.

Lemma 4.1. *Let σ be a continuous function that satisfies Condition 1. Then, for a natural number $d \in \mathbb{N}$, the set $\Delta^{\sigma}_{d,d,d+\alpha(\sigma)}$ compactly approximates $\mathcal{D}^2(\mathbb{R}^d)$, where*

$$\alpha(\sigma) = \begin{cases} 0 & \textit{if } \sigma = \textit{Leaky-ReLU} \\ 1 & \textit{if } \sigma = \textit{ReLU} \\ 2 & \textit{if } \sigma = \textit{Otherwise} \end{cases}. \tag{10}$$

In other words, we have the relation

$$\mathcal{D}^2(\mathbb{R}^d) \prec \Delta^{\sigma}_{d,d,d+\alpha(\sigma)}. \tag{11}$$

In Teshima et al. (2020), it was shown that any diffeomorphism can be approximated by a composition of single coordinate transformations. Therefore, it suffices to prove that deep narrow MLPs can approximate any such single coordinate transformation (see Definition A.5 for the formal definition). This is established in Lemma B.1. With the exception of the Leaky-ReLU case, the proof is relatively direct. For the Leaky-ReLU case, we progressively extend the class of functions that can be approximated. Using Lemma B.3, we show that any increasing scalar function can be approximated by width one Leaky-ReLU networks. This result implies that any width-1 neural network with increasing activation functions can be approximated by a Leaky-ReLU network of the same width (see Corollary B.4). Building on this, we prove in Lemma B.5 that any ACF (see Definition A.6) can be approximated by deep narrow MLPs. Finally, Lemma B.6 shows that any single coordinate transformation can, in turn, be approximated by an ACF.

The detailed proof of Lemma 4.1 is provided in Appendix B.1.

Now, we quantify the required geometric width for approximation as $w(n, m)$. Additionally, we demonstrate that the network-independently defined value $w(n, m)$ equals the minimum width of deep, narrow, Leaky-ReLU MLPs.

Let $\mathrm{Emb}(X, Y)$ denote the set of smooth embeddings from X to Y, and $\mathrm{Emb}_{p.l.}(X, Y)$ represent the set of piecewise linear embeddings from X to Y. For natural numbers $d_1 \geq d_2$, define $p_{d_1,d_2} : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ as the projection to the first d_2 coordinates. And define $w(n, m)$ as:

$$w(n, m) := \min \left\{ l \in \mathbb{N}_0 \middle| p_{l,m}\left(\overline{\mathrm{Emb}([0, 1]^n, \mathbb{R}^l)}\right) = C([0, 1]^n, \mathbb{R}^m) \right\}. \tag{12}$$

Intuitively, $w(n, m)$ represents the minimum width required to approximate any arbitrary continuous function using diffeomorphisms.

Remark 4.2. *We note that the interval $[0, 1]$ can be replaced with the interval $[a, b]$ for $a < b$. Additionally,* $\mathrm{Emb}([0, 1]^{\mathrm{n}}, \mathbb{R}^{\mathrm{l}})$ *can be replaced with any dense subset of* $\overline{\mathrm{Emb}([0, 1]^n, \mathbb{R}^l)}$*, such as* $\mathrm{Emb}_{p.l.}([0, 1]^n, \mathbb{R}^l)$ *(Munkres, 1960).*

We will prove that the difference between $w(n, m)$ and $w(n, m, \sigma)$ is bounded by two, and that $w(n, m)$ has the same value as $w(n, m, \text{Leaky-ReLU})$. The next lemma demonstrates that any smooth embedding can be represented by the composition of an inclusion and a smooth diffeomorphism.

Lemma 4.3 (Theorem C of Palais (1960)). *Consider natural numbers n and m where $n \leq m$, and a smooth embedding $f : K = [0,1]^n \to \mathbb{R}^m$. Then, there exists a smooth diffeomorphism $F : \mathbb{R}^m \to \mathbb{R}^m$ such that the following equation holds:*

$$F \circ q_{n,m} = f. \tag{13}$$

Using the preceding lemma along with the definition of $w(n,m)$, we can present the following theorem:

Theorem 4.4. *Let σ be a continuous function satisfying Condition 1. Then, for natural numbers n and m, $\Delta^\sigma_{n,m,w(n,m)+\alpha(\sigma)}$ compactly approximates $C(\mathbb{R}^n, \mathbb{R}^m)$, where*

$$\alpha(\sigma) = \begin{cases} 0 & \text{if } \sigma = \textit{Leaky-ReLU} \\ 1 & \text{if } \sigma = \textit{ReLU} \\ 2 & \text{if } \sigma = \textit{Otherwise} \end{cases}. \tag{14}$$

In other words,

$$C(\mathbb{R}^n, \mathbb{R}^m) \prec \Delta^\sigma_{n,m,w(n,m)+\alpha(\sigma)}. \tag{15}$$

Proof. Without loss of generality, assume that $K = [0,1]^n$. In cases where K differs, we can achieve this by continuously extending the function to encompass a cube containing K and then rescaling. By the definition of $w(n,m)$, for any $f \in C([0,1]^n, \mathbb{R}^m)$ and $\epsilon > 0$, there exists an embedding $g \in \mathrm{Emb}([0,1]^n, \mathbb{R}^{w(n,m)})$ such that

$$\|f - p_{w(n,m),n} \circ g\|_{\infty,[0,1]^n} < \epsilon. \tag{16}$$

Because $w(n,m) \geq n$, by Lemma 4.3, for $q_{n,w(n,m)} : (x_1, \ldots, x_n) \mapsto (x_1, \ldots, x_n, 0, \ldots, 0)$, there exists a smooth diffeomorphism G such that $g = G \circ q_{n,w(n,m)}$. By Lemma 4.1, there exists an $H \in \Delta^\sigma_{w(n,m),w(n,m),w(n,m)+\alpha(\sigma)}$ such that

$$\|G - H\|_{\infty, K \times [0,1]^{w(n,m)-n}} < \epsilon. \tag{17}$$

Then,

$$\|p_{w(n,m),n} \circ H \circ q_{n,w(n,m)} - p_{w(n,m),n} \circ G \circ q_{n,w(n,m)}\|_{\infty,[0,1]^n} < \epsilon. \tag{18}$$

Therefore,

$$\|f - p_{w(n,m),m} \circ H \circ q_{n,w(n,m)}\|_{\infty,K} < 2\epsilon. \tag{19}$$

$p_{w(n,m),m} \circ H \circ q_{n,w(n,m)} \in \Delta^\sigma_{w(n,m),w(n,m),w(n,m)+\alpha(\sigma)}$. This completes the proof. □

Furthermore, we can give the lower bound for the minimum width required for the UAP.

Theorem 4.5. *Let σ be an increasing, continuous activation function. For natural numbers n and m in $\mathbb{N}$, $\Delta^\sigma_{n,m,w(n,m)-1}$ does not compactly approximate $C(\mathbb{R}^n, \mathbb{R}^m)$. In other words, the following relation holds:*

$$C(\mathbb{R}^n, \mathbb{R}^m) \not\prec \Delta^\sigma_{n,m,w(n,m)-1}. \tag{20}$$

The detailed proof is provided in Appendix C.1.

Combining Theorem 4.4 with Theorem 4.5, we conclude that the minimum width $w_{\min}(n, m, \text{Leaky-ReLU})$ equals $w(n,m)$ for Leaky-ReLU and provides a tight inequality for general increasing activation functions.

Corollary 4.6. *The following equation holds:*

$$w_{\min}(n, m, \textit{Leaky-ReLU}) = w(n,m) \tag{21}$$

For a general increasing activation function σ, which satisfies Condition 1, the following inequality holds:

$$w(n,m) \leq w_{\min}(n,m,\sigma) \leq w(n,m) + \alpha(\sigma), \tag{22}$$

where

$$\alpha(\sigma) = \begin{cases} 1 & \text{if } \sigma = \textit{ReLU} \\ 2 & \text{if } \sigma = \textit{Otherwise} \end{cases}. \tag{23}$$

4.3 Some Observations about the Upper bound of $w(n,m)$

In the previous subsection, we demonstrated the fundamental correlation between the minimum width of the deep, narrow MLP and $w(n,m)$. In this subsection, we will present a sufficient condition for $w(n,m)$ to be equal to m. The following lemma demonstrates that a continuous function can be approximated by a smooth embedding when the output dimension is greater than twice the input dimension:

Lemma 4.7. *Consider natural numbers n and m where $m > 2n$. Let $f : K = [0,1]^n \subset \mathbb{R}^n \to \mathbb{R}^m$ be a continuous function. Then, for $\epsilon \in \mathbb{R}_+$, there exists a smooth embedding $g : K \to \mathbb{R}^m$ such that*

$$\|f - g\|_{\infty,K} < \epsilon. \tag{24}$$

Proof. Consider a connected, open subset U of $\mathbb{R}^n$ such that $K \subset U \subset \mathbb{R}^n$. Since K is compact, there exists a continuous extension f_0 of f such that

$$f_0|_K = f. \tag{25}$$

As U is a manifold, it satisfies the assumptions of Theorems 3.17 and 3.18 in Persson (2014). Therefore, there exists an injective immersion g such that

$$\|f - g\|_{\infty,U} < \epsilon. \tag{26}$$

Consequently, the restriction $g|_K$ defined on the compact set K becomes a smooth embedding. □

Now, we present an upper bound in the following theorem.

Theorem 4.8. *Let σ be a continuous function satisfying Condition 1. Then, for any natural numbers $n, m \in \mathbb{N}$, the set $\Delta^{\sigma}_{n,m,\max(2n+1,m)+\alpha(\sigma)}$ compactly approximates $C(\mathbb{R}^n, \mathbb{R}^m)$, where*

$$\alpha(\sigma) = \begin{cases} 0 & \text{if } \sigma = \textit{Leaky-ReLU} \\ 1 & \text{if } \sigma = \textit{ReLU} \\ 2 & \text{if } \sigma = \textit{Otherwise} \end{cases}. \tag{27}$$

In other words, the following relation holds:

$$C(\mathbb{R}^n, \mathbb{R}^m) \prec \Delta^{\sigma}_{n,m,\max(2n+1,m)+\alpha(\sigma)}. \tag{28}$$

Proof. Lemma 4.7 implies that $w(n,m) \leq \max(2n+1, m)$. By Theorem 4.4, we can immediately get the conclusion. □

Remark 4.9. *As previously demonstrated by Kim et al. (2024), the minimum width $w_{\min}(d_x, d_y, \sigma)$ satisfies the relation $w_{\min} \geq d_y + \mathbf{1}_{d_x < d_y \leq 2d_x}$ for an increasing activation function. It indicates that when the output dimension d_y is twice the input dimension d_x, $w_{\min}(d_x, 2d_x, \sigma)$ is equals or exceeds $d_y + 1$. In the same configuration, following Theorem 4.8, we derive the relation: $w_{\min}(d_x, d_y, \textit{Leaky-ReLU}) \leq 2d_x + 1 = d_y + 1$. By merging these results, we arrive at the optimal minimum width $w_{\min}(d_x, d_y, \textit{Leaky-ReLU}) = d_y + 1 = 2d_x + 1$.*

Furthermore, we know that $w_{\min}(d_x, d_y, \sigma) \geq \max(d_x, d_y)$ for a general activation function σ (Cai, 2023). If $d_y > 2d_x$, then $w_{\min}(d_x, d_y, \textit{Leaky-ReLU}) \leq d_y$ presenting the optimal minimum width as $w_{\min}(d_x, d_y, \textit{Leaky-ReLU}) = d_y$.

In addition to the aforementioned relation, there exists an evident upper bound for $w(n,m)$:

$$w(n,m) \leq n + m, \tag{29}$$

for all $n, m \in \mathbb{N}$. This reaffirms the result presented by Hanin (2019) in the case of Leaky-ReLU.

$$w_{min}(n, m, \text{Leaky-ReLU}) \leq n + m, \tag{30}$$

For ReLU (Hanin, 2019) and other general activation functions (Kidger & Lyons, 2020), the results are slightly less favorable:

$$w_{min}(n, m, \sigma) \leq n + m + \alpha(\sigma), \tag{31}$$

where $\alpha(\sigma) = 1$ for $\sigma =$ ReLU, and $\alpha(\sigma) = 2$ for other activation functions.

4.4 Some Novel Upper Bounds of $w(n, m)$

The framework of diffeomorphism not only gives the upper bound derived from Whitney embedding. By utilizing sophisticated techniques of geometric topology, we can give some non-trivial optimal width including $w(2, 3) = 4$. Note that it is strictly smaller than both of $d_x + d_y = 5$ and $2d_x + 1 = 5$.

Proposition 4.10. *For even k, $w(k, 2k - 1) = 2k$.*

Proof. The lower bound $w(k, 2k - 1) \geq 2k$ is by the result of Kim et al. (2024). For the upper bound, consider an arbitrary continuous function $f : [0, 1]^k \to \mathbb{R}^{2k-1}$. By Theorem 5.10 of Hirsch (1959), for any $\epsilon \in \mathbb{R}_+$, there exists an immersion g such that $\|f - g\|_{\infty,[0,1]^k} < \epsilon$. By Corollary 3.2 of Lashof & Smale (1959), there exists an embedding $h : [0, 1]^k \to \mathbb{R}^{2k}$ such that its first $2k - 1$ components $p_{2k,2k-1} \circ h$ satisfy $\|g - p_{2k,2k-1} \circ h\| < \epsilon$. This completes the proof. ☐

4.5 Lower Bound of $w(n, m)$

In this subsection, we offer a nontrivial example of minimum width using the concept of $w(n, m)$. In particular, we will prove that $w(2, 2) = 4$ employing algebraic topological techniques. To provide a lower bound, we construct a function $f : [0, 1]^2 \to \mathbb{R}^2$ that cannot be made injective by concatenating an additional one-dimensional output function.

Such a function is constructed by designing a function with a winding number of two, meaning that if a circle in the domain wraps around the origin once, its image wraps around the origin twice. Specifically, we use a function that maps (r, θ) to $(r, 2\theta)$ where r and θ are the radius and the angle in the polar coordinate system. This ensures that self-intersections always occur at two antipodal points on a certain circle. Then, the Borsuk-Ulam theorem guarantees that any function from a circle S^1 to $\mathbb{R}$ must have the same value at some pair of antipodal points, which implies that the concatenation also attains the same value at some pair of antipodal points. Therefore, the minimum width is at least 4. Topological algebra helps formalize these abstract discussions mathematically by utilizing relationships between homology, the fundamental group, and related concepts. See Hatcher (2000) for further details.

From now on, all homology will refer to singular homology with $\mathbb{Z}$-coefficients, denoted as $\mathrm{H}_i(X; \mathbb{Z})$. For simplicity, we will write it as $\mathrm{H}_i(X)$. The homology of a topological space provides information about the presence of holes in the space. For example, existence of nonzero H_1 indicates the presence of a one-dimensional hole.

We use the following lemma, which suggests that the homology of a level set of a function is robust under the perturbation.

Lemma 4.11 (Theorem 2 of Bendich et al. (2010))**.** *Let $\mathbb{X}$ be a compact topological space. For a continuous function $f : \mathbb{X} \to \mathbb{R}$ and $f_a : \mathbb{X} \to \mathbb{R}$ defined as $f_a(x) := |f(x) - a|$, we define $\mathbb{X}_r$ as follows:*

$$\mathbb{X}_r (f_a) = f_a^{-1}[0, r] \tag{32}$$

For $h : \mathbb{X} \to \mathbb{R}$, such that $\|f - h\|_{\infty,\mathbb{X}} < r$, $h^{-1}(a)$ is included in $\mathbb{X}_r (f_a)$:

$$h^{-1}(a) \hookrightarrow \mathbb{X}_r (f_a), \tag{33}$$

and this inclusion induces the homomorphism of the homology:

$$\mathrm{j}_h : \mathrm{H}_n \left(h^{-1}(a)\right) \to \mathrm{H}_n (\mathbb{X}_r (f_a)) \tag{34}$$

In addition, as $f^{-1}(a - r)$ and $f^{-1}(a + r)$ are also included in $\mathbb{X}_r (f_a)$, we have inclusions:

$$\iota^0 : f^{-1}(a + r) \hookrightarrow \mathbb{X}_r (f_a), \text{ and } \iota^1 : f^{-1}(a - r) \hookrightarrow \mathbb{X}_r (f_a), \tag{35}$$

which induce the homomorphisms ι^0_* and ι^1_* of the homology:

$$\iota^0_* : \mathrm{H}_n \left(f^{-1}(a + r)\right) \to \mathrm{H}_n (\mathbb{X}_r (f_a)), \text{ and } \iota^1_* : \mathrm{H}_n \left(f^{-1}(a - r)\right) \to \mathrm{H}_n (\mathbb{X}_r (f_a)). \tag{36}$$

Define $\mathrm{B}_{0,r}$ and $\mathrm{B}_{1,r}$ as the images of two homomorphism:

$$\mathrm{B}_{0,r} := \iota^0_* \left(\mathrm{H}_n \left(f^{-1}(a + r)\right)\right), \text{ and } \mathrm{B}_{1,r} := \iota^1_* \left(\mathrm{H}_n \left(f^{-1}(a - r)\right)\right). \tag{37}$$

Define $\mathrm{U}_n(r)$ *as:*

$$\mathrm{U}_n(r) = \bigcap_{\|h-f\|_{\infty,\mathbb{X}} \le r} \mathrm{im}(j_h). \tag{38}$$

Then, the following equation holds:

$$\mathrm{U}_n(r) = \mathrm{B}_{0,r} \cap \mathrm{B}_{1,r}. \tag{39}$$

We employ the well-known theorem as a lemma. It gives a relationship between the homology and the fundamental group of the space.

Lemma 4.12 (Hurewicz Theorem (Theorem 2A.1 of Hatcher (2000)))**.** *By regarding loops as singular* 1*-cycles, we obtain a homomorphism* $h : \pi_1(X, x_0) \to \mathrm{H}_1(X)$*. If* X *is path-connected, then* h *is surjective and has a kernel, which is the commutator subgroup of* $\pi_1(X)$*. Consequently,* h *induces an isomorphism from the abelianization of* $\pi_1(X)$ *onto* $\mathrm{H}_1(X)$*.*

By the preceding lemma, we can ensure that the fundamental group of the self-intersection sets is nontrivial.

Now, we introduce the concept of the winding number in the framework of topological algebra. We can observe that if a curve is given as input to the previously defined function that doubles the angle, the winding number of the output curve also doubles that of the input curve.

Definition 4.13 (Winding Number)**.** *For* $O = (0,0) \in \mathbb{R}^2$ *and a closed curve* $c : [0,1] \to \mathbb{R}^2 - O$*, consider* c *as an element of the fundamental group:*

$$[c] \in \pi_1(\mathbb{R}^2 - O, x_0) = \mathbb{Z}, \tag{40}$$

where the fundamental group $\pi_1(\mathbb{R}^2 - O, x_0)$ *is generated by the curve* $\omega_1 = (\cos(2\pi\theta), \sin(2\pi\theta))$*. Then, the winding number of* c *is the natural number* $[c]$ *as an element of* $\pi_1(\mathbb{R}^2 - O, x_0) = \mathbb{Z}$*.*

As the image curve has a winding number greater than one, it must have a self-intersection by the following lemma.

Lemma 4.14. *For any closed curve* $c : S^1 \to \{(x,y) \in \mathbb{R}^2 | 1 < x^2 + y^2 < 2\}$ *in the annulus with a winding number greater than* 1*,* c *is not injective.*

Proof. Suppose c is an injective curve. By the Jordan Curve Theorem (See Proposition 2B.1 of Hatcher (2000) for details), an injective curve bounds a region homeomorphic to the disk. Therefore, there exists an embedding $C : D^2 \to \mathbb{R}^2$ such that its restriction to the boundary is c:

$$C|_{S^1} = c. \tag{41}$$

Because $S^1 \hookrightarrow D^2 - \{O\}$ induces an isomorphism of the fundamental group, the degree should be 1 or -1. Therefore, a curve with a winding number greater than 1 is not injective. ☐

Using the lemmas, we can rigorously prove the following theorem.

Theorem 4.15. $w(2,2) = 4$*.*

The proof of Theorem 4.15 is provided in Appendix C.2. Then, the theorem yields a direct corollary, stating that the minimum width for universal approximation is 4 when both the input and output dimensions are two.

Corollary 4.16.

$$w_{min}(2,2,\mathit{ReLU}) = w_{min}(2,2,\mathit{Leaky\text{-}ReLU}) = 4. \tag{42}$$

Proof. The lower bound $w_{min}(2,2,\mathrm{ReLU}) \ge w(2,2) \ge 4$ is a direct consequence of Theorem 4.15 and Corollary 4.6. The upper bound is provided by Hanin (2019). This completes the proof. ☐

Remark 4.17. *In Li et al. (2023), authors claimed that the optimal minimum width is given by* $\max(d_x+1, d_y) + \mathbf{1}_{(d_x+1=d_y)}$*. However, as demonstrated in the previous corollary, it is not the case. This may originate from a subtle misconception, specifically the assumption that an arbitrary* d*-dimensional continuous function can be approximated by a* $(d+1)$*-dimensional diffeomorphism, which does not hold in general. In this paper, we carefully control injectivity and embedding properties to rigorously demonstrate that, in some cases, additional width is unavoidable.*

5 Limitations

In Section 4.4, we proved that for even k, the optimal minimum width is $w(k, 2k-1) = 2k$. But what about the case when k is odd? It is highly plausible that for odd k, the minimum width is $w(k, 2k-1) = 2k+1$, since odd k permits the existence of self-intersections with structure $[S^1; \mathbb{RP}^1]$ in the sense of $\mathbb{Z}_2$-bordism. Unfortunately, to the best of our knowledge, there is currently no mathematical tool available for analyzing the self-intersection structure of uniformly perturbed functions in this setting. It would be particularly interesting if one could rigorously prove that the minimum width indeed depends on the parity of the input dimension. Moreover, our current analysis has primarily focused on cases where the output dimension is roughly twice the input dimension, allowing us to exploit properties of homology and the fundamental group. Exploring higher-dimensional intersection settings remains a challenging but important direction for determining the minimum width of MLPs.

6 Conclusion

In this paper, we introduced novel upper and lower bounds for the minimum width of a deep, narrow MLP required to achieve the UAP within continuous function spaces. While our derived bound exhibits optimality only when the output dimension is about twice the input dimension, we propose that the strategy of approximating arbitrary functions through diffeomorphisms could potentially lead to achieving optimality across all cases. Exploring this perspective presents an intriguing avenue for future research. Additionally, we anticipate that analyzing the quantitative approximation capacity of general MLPs from the viewpoint of diffeomorphisms might yield valuable insights.

Acknowledgments and Disclosure of Funding

This work was supported by National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) [RS-2025-00515264, RS-2024-00406127], and Global University Project grant funded by the GIST in 2025.

References

Bendich, P., Edelsbrunner, H., Morozov, D., and Patel, A. The robustness of level sets. In *European symposium on algorithms*, pp. 1–10. Springer, 2010.

Cai, Y. Achieve the minimum width of neural networks for universal approximation. In *The Eleventh International Conference on Learning Representations*, 2023.

Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Hanin, B. Universal function approximation by deep neural nets with bounded width and relu activations. *Mathematics*, 7(10):992, 2019.

Hatcher, A. Algebraic topology, 2000.

HERNÁNDEZ, M. and ZUAZUA, E. Constructive universal approximation and finite sample memorization by narrow deep relu networks.

Hirsch, M. W. Immersions of manifolds. *Transactions of the American Mathematical Society*, 93(2): 242–276, 1959.

Hong, R. and Kratsios, A. Bridging the gap between approximation and learning via optimal approximation by relu mlps of maximal regularity. *arXiv preprint arXiv:2409.12335*, 2024.

Johnson, J. Deep, skinny neural networks are not universal approximators. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Kidger, P. and Lyons, T. Universal approximation with deep narrow networks. In *Conference on learning theory*, pp. 2306–2327. PMLR, 2020.

Kim, N., Min, C., and Park, S. Minimum width for universal approximation using relu networks on compact domain. In *The Twelfth International Conference on Learning Representations*, 2024.

Lashof, R. and Smale, S. Self-intersections of immersed manifolds. *Journal of Mathematics and Mechanics*, pp. 143–157, 1959.

Leshno, M., Lin, V. Y., Pinkus, A., and Schocken, S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

Li, L., Duan, Y., Ji, G., and Cai, Y. Minimum width of leaky-ReLU neural networks for uniform universal approximation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19460–19470. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/li23g.html`.

Liu, C. and Chen, M. Relu network with width d+ o (1) can achieve optimal approximation rate. In *International Conference on Machine Learning*, 2024.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. The expressive power of neural networks: A view from the width. *Advances in neural information processing systems*, 30, 2017.

Munkres, J. Obstructions to the smoothing of piecewise-differentiable homeomorphisms. *Annals of Mathematics*, pp. 521–554, 1960.

Palais, R. S. Extending diffeomorphisms. *Proceedings of the American Mathematical Society*, 11(2): 274–277, 1960.

Park, S., Yun, C., Lee, J., and Shin, J. Minimum width for universal approximation. In *International Conference on Learning Representations*, 2021.

Persson, M. The whitney embedding theorem, 2014.

Rochau, D., Chan, R., and Gottschalk, H. New advances in universal approximation with neural networks of minimal width. *arXiv preprint arXiv:2411.08735*, 2024.

Shen, Z., Yang, H., and Zhang, S. Optimal approximation rate of relu networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, 157:101–135, 2022.

Song, C. h., Hwang, G., Lee, J. h., and Kang, M. Minimal width for universal property of deep rnn. *Journal of Machine Learning Research*, 24(121):1–41, 2023.

Tabuada, P. and Gharesifard, B. Universal approximation power of deep residual neural networks through the lens of control. *IEEE Transactions on Automatic Control*, 2022.

Teshima, T., Ishikawa, I., Tojo, K., Oono, K., Ikeda, M., and Sugiyama, M. Coupling-based invertible neural networks are universal diffeomorphism approximators. *Advances in Neural Information Processing Systems*, 33:3362–3373, 2020.

A Definitions

In this section, we define several subsets of the set of diffeomorphisms.

Definition A.1 (Diffeomorphism). *For $d, r \in \mathbb{N}$ and open sets $U_1, U_2 \subset \mathbb{R}^d$, a function $f : U_1 \to U_2$ is called a C^r-diffeomorphism if and only if f is bijective, r-times continuously differentiable, and its inverse f^{-1} is r-times continuously differentiable.*

Definition A.2 (Embedding). *Let M and N be manifolds. A function $f : M \to N$ is called an embedding if and only if it is an immersion and a homeomorphism onto its image $f(M)$.*

Definition A.3. *For any natural number d, let $\mathcal{G}$ be a subset of invertible functions from $\mathbb{R}^d$ to $\mathbb{R}^d$. Then, $\mathrm{INN}_{\mathcal{G}}$ is defined as:*

$$\mathrm{INN}_{\mathcal{G}} := \left\{ W_1 \circ g_1 \circ \cdots \circ W_n \circ g_n \circ W_{n+1} \middle| n \in \mathbb{N}, g_i \in \mathcal{G}, W_i \in \mathrm{IAff}_d \right\} \tag{43}$$

Note that the approximation capability of $\mathrm{INN}_{\mathcal{G}}$ remains unchanged even if IAff_d, in the definition, is replaced with $\mathrm{Aff}_{d,d}$.

Definition A.4 (Compactly supported diffeomorphism: $\mathrm{Diff}^r_c(\mathbb{R}^d)$)**.** *A diffeomorphism $f : \mathbb{R}^d \to \mathbb{R}^d$ is compactly supported if there exists a compact subset $K \subset \mathbb{R}^d$ such that for any $x \notin K$, $f(x) = x$. $\mathrm{Diff}^r_c(\mathbb{R}^d)$ is the set of all compactly supported C^r-diffeomorphisms from $\mathbb{R}^d$ to $\mathbb{R}^d$.*

Definition A.5 (Single-coordinate transformations: $\mathcal{S}^r_c(\mathbb{R}^d)$)**.** *$\mathcal{S}^r_c(\mathbb{R}^d)$ is the set of all compactly supported C^r-diffeomorphisms defined as follows:*

$$\mathcal{S}^r_c(\mathbb{R}^d) := \left\{ \tau \in \mathrm{Diff}^r_c(\mathbb{R}^d) \middle| \exists i \in \{1, \ldots, d\} \text{ s.t } \tau(x)_j = x_j \text{ for } j \neq i \text{ and } \tau(x)_i = \widetilde{\tau}(x), \widetilde{\tau} \in C(\mathbb{R}^d, \mathbb{R}) \right\}. \quad (44)$$

Definition A.6 (Single-coordinate affine coupling flows)**.** *ACF_d is the set of all single-coordinate affine coupling flows defined as follows:*

$$ACF_d := \left\{ (x_1, \ldots, x_{d-1}, exp(s(x_{1:d-1})x_d + t(x_{1:d-1}))) \middle| s, t \in C(\mathbb{R}^{d-1}, \mathbb{R}) \right\}, \quad (45)$$

B Proofs for Approximation

B.1 Proof of Lemma 4.1

To prove the lemma, we introduce a lemma that suggests we can focus on approximating $\mathcal{S}^\infty_c(\mathbb{R}^d)$ to achieve the approximation of diffeomorphisms.

Lemma B.1. *For a natural number $d \in \mathbb{N}$, the following relation holds:*

$$\mathrm{INN}_{\mathcal{S}^\infty_c(\mathbb{R}^d)} \succ \mathcal{D}^2(\mathbb{R}^d). \quad (46)$$

Proof. This result directly follows from Theorem 1(B) in Teshima et al. (2020). As every element of $\mathcal{S}^\infty_c(\mathbb{R}^d)$ is invertible and locally bounded due to its continuity, it satisfies the conditions outlined in Theorem 1. Given that $\mathrm{INN}_{\mathcal{S}^\infty_c(\mathbb{R}^d)} \succ \mathcal{S}^\infty_c(\mathbb{R}^d)$, we can therefore conclude that $\mathrm{INN}_{\mathcal{S}^\infty_c(\mathbb{R}^d)} \succ \mathcal{D}^2(\mathbb{R}^d)$. □

Proof of Lemma 4.1. According to Lemma B.1, it suffices to prove that the set of neural networks can approximate $\mathcal{S}^\infty_c$: $\Delta^\sigma_{d,d,d+\alpha(\sigma)} \succ \mathcal{S}^\infty_c(\mathbb{R}^d)$.

When $\sigma =$ Leaky-ReLU, we need to prove that $\Delta^\sigma_{d,d,d} \succ \mathcal{S}^\infty_c(\mathbb{R}^d)$. We can accomplish this by employing Lemma B.2.

For $\sigma =$ ReLU, by Theorem 1 in Hanin (2019), for $f(x) = (x_1, \ldots, x_d, \tau(x))$, we have $f \prec \Delta^\sigma_{d,d+1,d+1}$. Therefore, $(x_1, \ldots, x_{d-1}, \tau(x)) \in \Delta^\sigma_{d,d,d+1}$, implying $\mathcal{S}^\infty_c(\mathbb{R}^d) \prec \Delta^\sigma_{d,d,d+1}$.

For other continuous activation functions σ, Proposition 4.2 of Kidger & Lyons (2020) demonstrates that for $f(x) = (x_1, \ldots, x_d, \tau(x))$, we have $f \prec \Delta^\sigma_{d,d+1,d+2}$. Consequently, $(x_1, \ldots, x_{d-1}, \tau(x)) \in \Delta^\sigma_{d,d,d+2}$, concluding that $\mathcal{S}^\infty_c(\mathbb{R}^d) \prec \Delta^\sigma_{d,d,d+2}$.

It is worth noting that while Theorem 1 of Hanin (2019) and Proposition 4.2 of Kidger & Lyons (2020) do not explicitly state the form of approximated function as $(x_1, \ldots, x_d, \tau(x))$, their proofs implicitly rely on this form. □

Now, the remaining task is to prove the following lemma for the Leaky-ReLU case.

Lemma B.2 (Single-Coordinate Transformations to Leaky-ReLU)**.** *For a natural number $d \in \mathbb{N}$, the following relation holds:*

$$\Delta^{LR}_{d,d,d} \succ \mathcal{S}^\infty_c(\mathbb{R}^d). \quad (47)$$

The proof of this lemma involves a series of lemmas and corollaries that gradually extend the scope of functions approximable using Leaky-ReLU. The proof is provided in Appendix B.2.

B.2 Proof of Lemma B.2

The following lemma implies that any increasing function can be approximated by composing Leaky-ReLUs and affine transformations.

Lemma B.3 (Increasing Functions to Leaky-ReLU)**.** *Define the sets as follows:*

$$U_0 := \{ax + b : \mathbb{R} \to \mathbb{R} |\, a \in \mathbb{R}_+, b \in \mathbb{R}\}\,, \tag{48}$$

$$U_{n+1} := \{aLR_\beta(f) + b : \mathbb{R} \to \mathbb{R} |\, a, \beta \in \mathbb{R}_+, b \in \mathbb{R}, f \in U_n\}\,, \tag{49}$$

$$U := \bigcup_{n=0}^{\infty} U_n. \tag{50}$$

Then, for any continuous, increasing activation function $\sigma : \mathbb{R} \to \mathbb{R}$, the following relation holds:

$$\sigma \prec U. \tag{51}$$

The proof of Lemma B.3 is provided in Appendix B.3. This lemma directly implies the subsequent corollary: deep, narrow MLPs using the Leaky-ReLU activation function can approximate a deep, narrow MLP using an increasing activation function and the same width.

Corollary B.4 (Generalization of Activation)**.** *For a natural number $d \in \mathbb{N}$ and any continuous, increasing activation function σ, the following relation holds:*

$$\Delta^{\sigma}_{d,d,d} \prec \Delta^{LR}_{d,d,d}. \tag{52}$$

Utilizing the corollary above, we can demonstrate that Leaky-ReLU deep, narrow MLPs can approximate any ACF.

Lemma B.5 (ACF to Leaky-ReLU)**.** *For a natural number $d \in \mathbb{N}$, the following relation holds:*

$$INN_{ACF_d} \prec \Delta^{LR}_{d,d,d}. \tag{53}$$

The proof of Lemma B.5 is provided in Appendix B.4. Next, we establish a technical lemma serving as the multidimensional counterpart of Lemma B.3. For a multidimensional function from $\mathbb{R}^d$ to $\mathbb{R}$ that increases with a coordinate x_d, we can freely change the value when x_d is large, while it remains unaffected when x_d is small.

Lemma B.6. *Consider a compact set $K = [0, 1]^d \subset \mathbb{R}^d$, two distinct real values $\alpha_1 < \alpha_2$, and a single-coordinate transformation $F = (x_1, \ldots, x_{d-1}, f(x)) \in \mathcal{S}^r_c$, where the function $f(x)$ satisfies the following relation:*

$$f(x) \le 0 \textit{ if } x_d < \alpha_1, \textit{ and } f(x) = 0 \textit{ if } x_d = \alpha_1. \tag{54}$$

Assuming that $F \prec \Delta^{LR}_{d,d,d}\Big|_K$. Then, for a continuous function $b : \mathbb{R}^{d-1} \to \mathbb{R}$ such that $b(x_{1:d-1}) > 0$ for all $x \in K$, there exists a single-coordinate transformation $G = (x_1, \ldots, x_{d-1}, g(x)) \prec \Delta^{LR}_{d,d,d}\Big|_K$ satisfying the following relation:

$$g(x) := \begin{cases} f(x) & \textit{if } x_d \le \alpha_1 \\ f(x)b(x_{1:d-1}) & \textit{if } x_d = \alpha_2 \end{cases}. \tag{55}$$

The proof of Lemma B.6 is provided in Appendix B.5. Utilizing the lemma, we can prove Lemma B.2.

Proof. Consider an arbitrary single-coordinate transformation $F(x) = (x_1, \ldots, x_{d-1}, \tau(x_1, \ldots, x_d))$ and a compact set $K \subset \mathbb{R}^d$. Without loss of generality, assume $K \subset [0, 1]^d$. Additionally, assume τ strictly increases with respect to x_d.

Because τ is a continuous function defined on a compact set, for an $\epsilon > 0$, there exists a natural number $N \in \mathbb{N}$ such that if $\|x - x'\| < \frac{1}{N}$, then $|\tau(x) - \tau(x')| < \epsilon$. Now, define $u_i : \mathbb{R}^{d-1} \to \mathbb{R}$ as follows:

$$u_i(x_{1:d-1}) := F\left(x_{1:d-1}, \frac{i}{N}\right). \tag{56}$$

If there exists a single-coordinate transformation $G = (x_1, \dots, x_{d-1}, g(x_{1:d})) \prec \Delta^{\mathrm{LR}}_{d,d,d}$ such that $u_i(x_{1:d-1}) = g\left(x_{1:d-1}, \frac{i}{N}\right)$ for $x \in K$, then $F \prec \Delta^{\mathrm{LR}}_{d,d,d}$. We will demonstrate the existence of a sequence $\{G_n\}_{n=1}^{\infty} \subset \overline{\Delta^{\mathrm{LR}}_{d,d,d}\Big|_K}$ such that $G_n(x_{1:d-1}, \frac{i}{N}) = u_i(x_{1:d-1})$ for $1 \le i \le n$ via mathematical induction.

By Lemma B.6, there exists a single-coordinate transformation $G_0 = (x_1, \dots, x_{d-1}, g_0(x))$ such that $g_0(x_{1:d-1}, 0) = u_0(x_{1:d-1})$. Assume the induction hypothesis holds for $n = n_0$, implying the existence of a single-coordinate transformation $G_{n_0} = (x_1, \dots, x_{d-1}, g_{n_0}(x))$ such that $g_{n_0}(x_{1:d-1}, \frac{i}{N}) = u_i(x_{1:d-1})$ for $1 \le i \le n_0$. Then, by Lemma B.5, we can construct $G'_{n_0} := (x_1, \dots, x_{d-1}, g_{n_0}(x) - u_{n_0}(x_{1:d-1})) \in \overline{\Delta^{\mathrm{LR}}_{d,d,d}\Big|_K}$. Notably, G'_{n_0} satisfies the assumptions of Lemma B.6 with $\alpha_1 = \frac{n_0}{N}$ and $\alpha_2 = \frac{n_0+1}{N}$. By applying Lemma B.6 with $b(x_{1:d-1}) = \frac{u_{n_0+1}(x_{1:d-1}) - u_{n_0}(x_{1:d-1})}{g_{n_0}(x_{1:d-1}, \frac{n_0+1}{N}) - u_{n_0}(x_{1:d-1})}$, we obtain a single-coordinate transformation $G''(n_0) = (x_1, \dots, x_{d-1}, g''_{n_0}(x))$ such that $g''_{n_0}(x_{1:d-1}, \frac{i}{N}) = u_i(x_{1:d-1}) - u_{n_0}(x_{1:d-1})$ for $i \le n_0 + 1$. Finally, by Lemma B.5, we can get $G_{n_0+1} := (x_1, \dots, x_{d-1}, g''_{n_0}(x) + u_{n_0}(x_{1:d-1})) \in \overline{\Delta^{\mathrm{LR}}_{d,d,d}\Big|_K}$. As a result, the induction hypothesis is satisfied, and this completes the proof. $\square$

B.3 Proof of Lemma B.3

Proof. Because increasing piecewise linear functions are dense in the space of increasing continuous functions defined on a compact interval, it suffices to prove that for any natural number $n \in \mathbb{N}$ and an increasing piecewise linear function f with n breakpoints, we have $f \in U_n$. We will proceed with mathematical induction on n. For the base case, $n = 0$, there is nothing to prove. Assume that the induction hypothesis holds for some $n = n_0$, and consider the case of $n = n_0 + 1$, where we have an increasing piecewise linear function f with $n_0 + 1$ breakpoints, denoted as $\alpha_1 < \alpha_2 < \cdots < \alpha_{n_0+1}$. The function f is affine on each of the intervals $(-\infty, \alpha_1], [\alpha_1, \alpha_2], \dots, [\alpha_{n_0}, \alpha_{n_0+1}], [\alpha_{n_0+1}, \infty)$. Now, let f have values as follows:

$$f(x) = \begin{cases} f(\alpha_{n_0+1}) + \gamma_1(x - \alpha_{n_0+1}) & \text{if } x \in [\alpha_{n_0}, \alpha_{n_0+1}] \\ f(\alpha_{n_0+1}) + \gamma_2(x - \alpha_{n_0+1}) & \text{if } x \in [\alpha_{n_0+1}, \infty) \end{cases}. \tag{57}$$

Consider the function f_0 defined as:

$$f_0(x) := \begin{cases} f(x) & \text{if } x \in (-\infty, \alpha_{n_0+1}] \\ f(\alpha_{n_0+1}) + \gamma_1(x - \alpha_{n_0+1}) & \text{if } x \in [\alpha_{n_0+1}, \infty) \end{cases}. \tag{58}$$

The function f_0 coincides with f on the interval $(-\infty, \alpha_{n_0+1}]$ and is affine on the interval $[\alpha_{n_0}, \infty)$. This means that the affine function on the interval $[\alpha_{n_0}, \alpha_{n_0+1}]$ naturally extends to the interval $[\alpha_{n_0} + 1, \infty)$ with the same slope. Therefore, f_0 has n_0 breakpoints, and by the induction hypothesis, $f_0 \in U_{n_0}$. We can express f in terms of f_0 as follows:

$$f(x) = \frac{\gamma_2}{\gamma_1} \mathrm{LR}_{\frac{\gamma_1}{\gamma_2}} \left(f_0(x) - f(\alpha_{n_0+1})\right) + f(\alpha_{n_0+1}). \tag{59}$$

Thus, $f \in U_{n_0+1}$, and the induction hypothesis is satisfied for $n = n_0 + 1$. This completes the proof. $\square$

B.4 Proof of Lemma B.5

Proof. For $\beta \in \mathbb{R}_+$, $a, c \in \mathbb{R}$ and $b \in \mathbb{R}^{d-1}$, we define the function g as follows:

$$g : (x_1, x_2, \dots, x_d) \mapsto (x_1, x_2, \dots, x_{d-1}, x_d + a\mathrm{LR}_\beta(b \cdot x_{1:d-1} + c)). \tag{60}$$

We will prove that $g \prec \Delta^{\mathrm{LR}}_{d,d,d}$. If b is the zero vector, g is a constant adding function satisfying the statement. If b is not the zero vector and $b = (b_1, \dots, b_{d-1})$, there exists an index $1 \le i \le d-1$ such that $b_i \neq 0$. Let $W \in \mathrm{IAff}_d$ be an invertible affine transformation defined as:

$$W : (x_1, x_2, \dots, x_d) \mapsto (x_1, x_2, \dots, x_{i-1}, b \cdot x_{1:d-1} + c, x_{i+1}, \dots, x_d). \tag{61}$$

Because b_i is nonzero, W is invertible. Applying LR_β to the i-th component gives:

$$(x_1, \ldots, x_{i-1}, \mathrm{LR}_\beta(b \cdot x_{1:d-1} + c), x_{i+1}, \ldots, x_d) \prec \Delta^{\mathrm{LR}}_{d,d,d}. \tag{62}$$

By adding a times the i-th component to the last component, we have:

$$(x_1, \ldots, \mathrm{LR}_\beta(b \cdot x_{1:d-1} + c), \ldots, x_d + a\mathrm{LR}_\beta(b \cdot x_{1:d-1}) + c) \prec \Delta^{\mathrm{LR}}_{d,d,d}. \tag{63}$$

By applying $\mathrm{LR}_{\frac{1}{\beta}}$ to the i-th component and applying W^{-1}, we get:

$$(x_1, \ldots, x_{d-1}, x_d + a\mathrm{LR}_\beta(b \cdot x_{1:d-1} + c)) \prec \Delta^{\mathrm{LR}}_{d,d,d}. \tag{64}$$

Next, we will prove that for the function h defined as:

$$h : (x_1, x_2, \ldots, x_d) \mapsto (x_1, x_2, \ldots, x_{d-1}, x_d + t(x_1, \ldots, x_{d-1})), \tag{65}$$

$h \prec \Delta^{\mathrm{LR}}_{d,d,d}$. By the UAP of two-layered neural networks (Leshno et al., 1993), for any $\epsilon > 0$ and a compact set $K \subset \mathbb{R}^{d-1}$, there exist $\beta \in \mathbb{R}_+$, $a_i, c_i \in \mathbb{R}$, and $b_i \in \mathbb{R}^{d-1}$ such that:

$$\left\| t(x_{1:d-1}) - \sum_{i=1}^{n} a_i \mathrm{LR}_\beta(b_i \cdot x_{1:d-1} + c_i) \right\|_{\infty, K} < \epsilon. \tag{66}$$

Composing Eq (64) for n different a_i, b_i, and c_i yields:

$$\left(x_1, \ldots, x_{d-1}, x_d + \sum_{i=1}^{n} a_i \mathrm{LR}_\beta \left(b_i \cdot x_{1:d-1} + c_i \right) \right) \prec \Delta^{\mathrm{LR}}_{d,d,d}. \tag{67}$$

Thus, $h \prec \Delta^{\mathrm{LR}}_{d,d,d}$.

Finally, by composing the operations described so far, we demonstrate that any ACF can be approximated by $\Delta^{\mathrm{LR}}_{d,d,d}$. It is achieved by combining the following four operations:

- Apply the logarithm to the last component.
- Add $\log(s(x_1, \ldots, x_{d-1}))$ to the last component.
- Apply the exponential function to the last component.
- Add $t(x_1, \ldots, x_{d-1})$ to the last component.

This results in the following transformation:

$$(x_1, x_2, \ldots, x_d) \mapsto (x_1, x_2, \ldots, x_{d-1}, \exp\left(\log(x_d) + \log(s)\right) + t) = (x_1, x_2, \ldots, x_{d-1}, s x_d + t) \prec \Delta^{\mathrm{LR}}_{d,d,d}. \tag{68}$$

This completes the proof. $\square$

B.5 Proof of Lemma B.6

Proof. We begin by observing that it is sufficient to consider functions b satisfying $b(x_{1:d-1}) \geq 1$ for all $x \in K$. Define β as $\beta := \inf_{x \in K} b(x_{1:d-1})$. We introduce the function $\widetilde{F}(x) := (x_1, \ldots, x_{d-1}, \widetilde{f}(x))$, defined as:

$$\widetilde{f}(x) := \beta \mathrm{LR}_{\frac{1}{\beta}} \left(f(x) \right). \tag{69}$$

If $F \in \overline{\Delta^{\mathrm{LR}}_{d,d,d}\Big|_K}$, then $\widetilde{F} \in \overline{\Delta^{\mathrm{LR}}_{d,d,d}\Big|_K}$. The value of $\widetilde{f}(x)$ can be calculated as:

$$\widetilde{f}(x) = \begin{cases} f(x) & \text{if } x_d \leq \alpha_1 \\ \beta f(x) & \text{if } x_d > \alpha_1 \end{cases}. \tag{70}$$

This ratio $\frac{g(x)}{\widetilde{f}(x)} = \frac{b(x_{1:d-1})}{\beta} \geq 1$ for all $x \in K$, and $\widetilde{F}$ also satisfied all the assumptions of the lemma. Therefore, we only need to consider functions b that satisfy $b \geq 1$.

Next, we will inductively construct a sequence $\{G_i = (x_1, \ldots, x_{d-1}, g_i(x))\}_{i=1}^{\infty} \subset \overline{\Delta^{\text{LR}}_{d,d,d}\Big|_K}$ that uniformly converges to G when $x_d = \alpha_2$. We start with $g_0(x) := f(x)$. Define $b_i : \mathbb{R}^{d-1} \to \mathbb{R}$ as :

$$b_i(x_{1:d-1}) := \frac{g_i(x_{1:d-1}, \alpha_2)}{f(x_{1:d-1}, \alpha_2)}, \tag{71}$$

for all $x \in K$. Define $\gamma_i \in \mathbb{R}$ as:

$$\gamma_i := \sup\left\{ \frac{b(x_{1:d-1})}{b_i(x_{1:d-1})} \middle| x \in K \right\}. \tag{72}$$

Next, we define two mutually exclusive sets, denoted as $L_{i,0}$ and $L_{i,1}$:

$$L_{i,0} = \left\{ x_{1:d-1} \in [0,1]^{d-1} \middle| 1 \le \frac{b(x_{1:d-1})}{b_i(x_{1:d-1})} \le \gamma_i^{\frac{1}{3}} \right\}. \tag{73}$$

$$L_{i,1} = \left\{ x_{1:d-1} \in [0,1]^{d-1} \middle| \gamma_i^{\frac{2}{3}} \le \frac{b(x_{1:d-1})}{b_i(x_{1:d-1})} \le \gamma_i \right\}. \tag{74}$$

Define a distance metric D as:

$$D(x, C) := \inf_{y \in C} \|x - y\|_2, \tag{75}$$

And then, we define the function $\phi_i : \mathbb{R}^{d-1} \to \mathbb{R}$ as:

$$\phi_i(x) := \frac{D(x, L_{i,0})}{D(x, L_{i,0}) + D(x, L_{i,1})}. \tag{76}$$

The function ϕ_i satisfies the inequality $0 \le \phi_i(x_{1:d-1}) \le 1$ for all $x \in K$, has a value of zero on $L_{i,0}$, and a value of one on $L_{i,1}$. Define $h_i : \mathbb{R}^{d-1} \to \mathbb{R}$ as follows:

$$h_i(x_{1:d-1}) := (1 - \phi_i(x_{1:d-1})) g_i(x_{1:d-1}, \alpha_2) \tag{77}$$

Then, $0 \le h_i(x_{1:d-1}) \le g_i(x_{1:d-1}, \alpha_2)$ for all $x \in K$, has a value of zero on $L_{i,1}$ and a value of $g_i(x_{1:d-1}, \alpha_2)$ on $L_{i,0}$.

Now, we define $g_{i+1}(x) \in \overline{\Delta^{\text{LR}}_{d,d,d}\Big|_K}$ as follows:

$$g_{i+1}(x) := \gamma_i^{\frac{1}{3}} \text{LR}_{\gamma_i^{-\frac{1}{3}}}(g_i(x) - h_i(x_{1:d-1})) + h_i(x_{1:d-1}). \tag{78}$$

We have

$$g_{i+1}(x) \begin{cases} = g_i(x) = 0 & \text{if } x_d \le \alpha_1 \\ = g_i(x) & \text{if } x_d = \alpha_2 \text{ and } x_{1:d-1} \in L_{i,0} \\ = \gamma_i^{\frac{1}{3}} g_i(x) & \text{if } x_d = \alpha_2 \text{ and } x_{1:d-1} \in L_{i,1} \\ \le \gamma_i^{\frac{1}{3}} g_i(x) & \text{if } x_d = \alpha_2 \text{ and } x_{1:d-1} \notin L_{i,0} \cup L_{i,1} \end{cases}. \tag{79}$$

Thus, for x where $x_d = \alpha_2$ and $x_{1:d-1} \in L_{i,0}$, we have

$$\frac{g(x)}{g_{i+1}(x)} = \frac{g(x)}{g_i(x)} = \frac{b(x_{1:d-1})}{b_i(x_{1:d-1})}. \tag{80}$$

As $1 \le \frac{b(x_{1:d-1})}{b_i(x_{1:d-1})} \le \gamma_i^{\frac{1}{3}}$ for $x_{1:d-1} \in L_{i,0}$, we deduce that $1 \le \frac{g(x)}{g_{i+1}(x)} \le \gamma_i^{\frac{1}{3}}$.

For x where $x_d = \alpha_2$ and $x_{1:d-1} \in L_{i,1}$,

$$\frac{g(x)}{g_{i+1}(x)} = \frac{g(x)}{\gamma_i^{\frac{1}{3}} g_i(x)} = \gamma_i^{-\frac{1}{3}} \frac{b(x_{1:d-1})}{b_i(x_{1:d-1})}. \tag{81}$$

As $\gamma_i^{\frac{2}{3}} \le \frac{b(x_{1:d-1})}{b_i(x_{1:d-1})} \le \gamma_i$ for $x_{1:d-1} \in L_{i,1}$, we get $1 \le \gamma_i^{\frac{1}{3}} \le \frac{g(x)}{g_{i+1}(x)} \le \gamma_i^{\frac{2}{3}}$.

For x where $x_d = \alpha_2$ and $x_{1:d-1} \notin L_{i,0} \cup L_{i,1}$,

$$\gamma_i^{-\frac{1}{3}} \frac{b(x_{1:d-1})}{b_i(x_{1:d-1})} = \frac{g(x)}{\gamma_i^{\frac{1}{3}} g_i(x)} \le \frac{g(x)}{g_{i+1}(x)} \le \frac{g(x)}{g_i(x)} = \frac{b(x_{1:d-1})}{b_i(x_{1:d-1})}. \tag{82}$$

As $\gamma_i^{\frac{1}{3}} \le \frac{b(x_{1:d-1})}{b_i(x_{1:d-1})} \le \gamma_i^{\frac{2}{3}}$ for $x_{1:d-1} \notin L_{i,0} \cup L_{i,1}$, we conclude that $1 \le \frac{g(x)}{g_{i+1}(x)} \le \gamma_i^{\frac{2}{3}}$.

We obtain the following results: for all $x \in K$, where $x_d = \alpha_2$, we have $1 \le \frac{g(x)}{g_{i+1}(x)} = \frac{b(x_{1:d-1})}{b_{i+1}(x_{1:d-1})} \le \gamma_i^{\frac{2}{3}}$. This implies $1 \le \gamma_{i+1} \le \gamma_i^{\frac{2}{3}}$. Consequently, as i tends towards infinity, γ_i converges to one. Therefore, $\frac{g(x_{1:d-1},\alpha_2)}{g_{i+1}(x_{1:d-1},\alpha_2)}$ uniformly converges to one as i increases, implying the convergence of G_i to G. As a result, there exists a function $G = (x_1, \dots, x_{d-1}, g(x_{1:d})) \in \overline{\Delta^{\mathrm{LR}}_{d,d,d}\Big|_K}$ such that $g(x_{1:d-1}, \alpha_2) = b(x_{1:d-1}) f(x_{1:d-1}, \alpha_2)$.

To check that G is a single-coordinate transformation, we observe:

$$g_{i+1}(x_{1:d-1}, x_d) - g_{i+1}(x_{1:d-1}, x'_d) > g_i(x_{1:d-1}, x_d) - g_i(x_{1:d-1}, x'_d), \tag{83}$$

for $x_d > x'_d$, which implies that $g(x_{1:d-1}, x_d) - g(x_{1:d-1}, x'_d) > g_0(x_{1:d-1}, x_d) - g_0(x_{1:d-1}, x'_d) > 0$ for all $x \in K$. Therefore, g satisfies the strictly increasing condition, and G becomes a single-coordinate transformation. □

C Proofs for Topology

C.1 Proof of Theorem 4.5

Proof. For a non-decreasing continuous activation function σ, there exist smooth, strictly increasing activation functions σ_n that uniformly converge to σ. Therefore, $\Delta^{\sigma}_{d,d,d} \prec \bigcup_{n\in\mathbb{N}} \Delta^{\sigma_n}_{d,d,d} \prec \mathcal{D}^{\infty}(\mathbb{R}^d)$, making it sufficient to consider only a set of smooth, strictly increasing activation functions σ.

For $f \in \Delta^{\sigma}_{n,m,w(n,m)-1}$, it can be decomposed as:

$$f = p_{w(n,m)-1,m} \circ g \circ q_{n,w(n,m)-1}, \tag{84}$$

where $g \in \Delta^{\sigma}_{w(n,m)-1,w(n,m)-1,w(n,m)-1}$. Because $\Delta^{\sigma}_{w(n,m)-1,w(n,m)-1,w(n,m)-1} \prec \mathcal{D}^{\infty}(\mathbb{R}^{w(n,m)-1})$, $g \circ q_{n,w(n,m)-1}\big|_{[0,1]^n} \in \overline{\mathrm{Emb}([0,1]^n, \mathbb{R}^{w(n,m)-1})}$. Therefore, we have:

$$f|_{[0,1]^n} \in p_{w(n,m)-1,m}\left(\overline{\mathrm{Emb}([0,1]^n, \mathbb{R}^{w(n,m)-1})}\right), \tag{85}$$

and because $f \in \Delta^{\sigma}_{n,m,w(n,m)-1}$ is arbitrary, we conclude:

$$\Delta^{\sigma}_{n,m,w(n,m)-1}\Big|_{[0,1]^n} \subset p_{w(n,m)-1,m}\left(\overline{\mathrm{Emb}([0,1]^n, \mathbb{R}^{w(n,m)-1})}\right). \tag{86}$$

Because $w(n,m) - 1 < w(n,m)$, by the definition of $w(n,m)$:

$$p_{w(n,m)-1,m}\left(\overline{\mathrm{Emb}([0,1]^n, \mathbb{R}^{w(n,m)-1})}\right) \not\supseteq C([0,1]^n, \mathbb{R}^m), \tag{87}$$

and consequently,

$$\Delta^{\sigma}_{n,m,w(n,m)-1}\Big|_{[0,1]^n} \not\supseteq C([0,1]^n, \mathbb{R}^m). \tag{88}$$

Hence, we conclude that $C(\mathbb{R}^n, \mathbb{R}^m) \not\prec \Delta^{\sigma}_{n,m,w(n,m)-1}$. □

C.2 Proof of Theorem 4.15

Proof. Firstly, it is obvious that $w(2,2) \le 4 = 2 + 2$. Therefore, it is sufficient to prove that $w(2,2) \ge 4$. Assume the opposite, $w(2,2) \le 3$. Then, for an arbitrary continuous function f in $C([-2,2]^2, \mathbb{R}^2)$, f is contained in $p_{3,2} \circ \overline{\mathrm{Emb}([-2,2]^2, \mathbb{R}^3)} = p_{3,2} \circ \overline{\mathrm{Emb}_{p.l.}([-2,2]^2, \mathbb{R}^3)}$. Consider the piecewise linear map $f : [-2,2]^2 \to \mathbb{R}^2$ defined as follows:

$$f(x_1, x_2) := \begin{cases} \begin{pmatrix} 1 & -1 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} & \text{if } 0 \le x_2 \le x_1, \\[2ex] \begin{pmatrix} 1 & -1 \\ 2 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} & \text{if } 0 \le x_1 \le x_2, \\ -f(x_2, -x_1) & \text{if } x_1 \le 0 \text{ and } 0 \le x_2, \\ f(-x_1, -x_2) & \text{if } x_2 \le 0. \end{cases} \tag{89}$$

We can check that f is the piecewise linear double-winding function. By the assumption, there exists a piecewise linear embedding $G \in \mathrm{Emb}_{p.l.}([-2,2]^2, \mathbb{R}^3)$ such that

$$\|f - p_{3,2} \circ G\|_{\infty,[-2,2]^2} < \frac{1}{4}. \tag{90}$$

Let $\Sigma : \mathbb{R}^2 \to \mathbb{R}$ be defined as

$$\Sigma : (x_1, x_2) \mapsto |x_1| + |x_2|. \tag{91}$$

We observe that f conserves the level of Σ: $(\Sigma \circ f)(x) = \Sigma(x)$ for all $x \in \mathbb{R}^2$. Therefore,

$$(\Sigma \circ f)^{-1}(1) = \Sigma^{-1}(1) = \left\{ (x_1, x_2) \in \mathbb{R}^2 \big| \, |x_1| + |x_2| = 1 \right\}, \tag{92}$$

which is homeomorphic to a circle S^1. Similarly,

$$(\Sigma \circ f)^{-1}\left(\frac{1}{2}\right) = \Sigma^{-1}\left(\frac{1}{2}\right) = \left\{ (x_1, x_2) \in \mathbb{R}^2 \big| \, |x_1| + |x_2| = \frac{1}{2} \right\}, \tag{93}$$

And

$$(\Sigma \circ f)^{-1}\left(\frac{3}{2}\right) = \Sigma^{-1}\left(\frac{3}{2}\right) = \left\{ (x_1, x_2) \in \mathbb{R}^2 \big| \, |x_1| + |x_2| = \frac{3}{2} \right\}, \tag{94}$$

are homeomorphic to S^1, and

$$(\Sigma \circ f)^{-1}\left(\left[\frac{1}{2}, \frac{3}{2}\right]\right) = \Sigma^{-1}\left(\left[\frac{1}{2}, \frac{3}{2}\right]\right) = \left\{ (x_1, x_2) \in \mathbb{R}^2 \big| \, \frac{1}{2} \leq |x_1| + |x_2| \leq \frac{3}{2} \right\}, \tag{95}$$

is homeomorphic to a closed annulus $S^1 \times [0,1]$.
Define g as $g := p_{3,2} \circ G$. Because $\|f - g\|_{\infty,[-2,2]^2} < \frac{1}{4}$, we have

$$|\Sigma \circ f - \Sigma \circ g| < \frac{1}{2}. \tag{96}$$

Now, apply Lemma 4.11 to $\Sigma \circ f$. Because $(\Sigma \circ f)^{-1}\left(\frac{1}{2}\right) = \Sigma^{-1}(\frac{1}{2})$ and $(\Sigma \circ f)^{-1}\left(\frac{3}{2}\right) = \Sigma^{-1}(\frac{3}{2})$ are deformation retractions of $(\Sigma \circ f)^{-1}\left([\frac{1}{2}, \frac{3}{2}]\right)$, the following equation holds:

$$\mathrm{U}_1\left(\frac{1}{2}\right) = \mathrm{H}_1\left(\mathrm{B}_{0,\frac{1}{2}}\right) = \mathrm{H}_1\left(\mathrm{B}_{1,\frac{1}{2}}\right) = \mathrm{H}_1\left((\Sigma \circ f)^{-1}\left(\left[\frac{1}{2}, \frac{3}{2}\right]\right)\right) = \mathrm{H}_1\left(\Sigma^{-1}(1)\right) = \mathbb{Z} \tag{97}$$

Thus, $\mathrm{U}_1(\frac{1}{2}) = \mathbb{Z}$. Note that $j_g : \mathrm{H}_1\left((\Sigma \circ g)^{-1}(1)\right) \to \mathrm{U}_1(\frac{1}{2})$ is surjective.

Because g and Σ are piecewise linear, $(\Sigma \circ g)^{-1}(1)$ consists of finite connected components $A_1, \ldots, A_k$. Then, the first homology $\mathrm{H}_1\left((\Sigma \circ g)^{-1}(1)\right)$ is decomposed as:

$$\mathrm{H}_1\left((\Sigma \circ g)^{-1}(1)\right) = \bigoplus_{i=1}^{k} \mathrm{H}_1(A_i), \tag{98}$$

And we can express j_g as a sum of homomorphisms $j_g^i : \mathrm{H}_1(A_i) \to \mathrm{U}_1\left(\frac{1}{2}\right)$:

$$j_g(x) = \sum_{i=1}^{k} j_g^i(x_i), \tag{99}$$

for $x = \bigoplus_{i=1}^{k} x_i$. As j_g is surjective, we can choose an index i_0 such that $j_g^{i_0}$ is a nonzero homomorphism. Set a basepoint $x_0 \in A_{i_0}$. By Lemma 4.12, there exists a surjective Hurewicz homomorphism $h_1 : \pi_1(A_{i_0}, x_0) \to \mathrm{H}_1(A_{i_0})$. Additionally, the Hurewicz homomorphism $h_2 : \pi_1\left((\Sigma \circ f)^{-1}\left([\frac{1}{2}, \frac{3}{2}]\right), x_0\right) \to \mathrm{H}_1\left((\Sigma \circ f)^{-1}\left([\frac{1}{2}, \frac{3}{2}]\right)\right)$ is an isomorphism. By composing homomorphism, we obtain:

$$\pi_1(A_{i_0}, x_0) \xrightarrow{h_1} \mathrm{H}_1(A_{i_0}) \tag{100}$$

$$\xrightarrow{j_g^{i_0}} \mathrm{H}_1\left((\Sigma \circ f)^{-1}\left(\left[\frac{1}{2}, \frac{3}{2}\right]\right)\right) = \mathrm{U}_1\left(\frac{1}{2}\right) \xrightarrow{h_2^{-1}} \pi_1\left((\Sigma \circ f)^{-1}\left(\left[\frac{1}{2}, \frac{3}{2}\right]\right), x_0\right), \tag{101}$$

resulting in the nonzero homomorphism $h_2^{-1} \circ j_g^{i_0} \circ h_1$:

$$h_2^{-1} \circ j_g^{i_0} \circ h_1 : \pi_1 \left(A_{i_0}, x_0\right) \to \pi_1 \left((\Sigma \circ f)^{-1} \left(\left[\frac{1}{2}, \frac{3}{2} \right] \right), x_0 \right). \tag{102}$$

Furthermore, we can observe that

$$h_2^{-1} \circ j_g^{i_0} \circ h_1 = \iota_*, \tag{103}$$

where ι_* represents the homomorphism of the fundamental group induced by the inclusion $\iota : A_{i_0} \hookrightarrow (\Sigma \circ f)^{-1} \left(\left[\frac{1}{2}, \frac{3}{2} \right] \right)$.

Now, we will prove the existence of a simple closed curve $\gamma : S^1 \to (\Sigma \circ g)^{-1}(1)$, homotopic to the cycle $\omega_1 : \theta \mapsto (\cos(2\pi\theta), \sin(2\pi\theta))$. Because g and Σ are piecewise linear, $(\Sigma \circ g)^{-1}(1)$ can be realized by a simplicial complex, and we can assume that $\pi_1 \left((\Sigma \circ g)^{-1}(1), x_0 \right)$ is generated by curves with finite segments, where all self-intersection points are breakpoints of curves. Choose $\gamma_0 \in \pi_1 \left((\Sigma \circ g)^{-1}(1), x_0 \right) = \pi_1 \left(A_{i_0}, x_0\right)$ such that $\iota_* ([\gamma_0])$ is nonzero. We iteratively construct a closed curve γ_i until it has no self-intersection points. Suppose γ_i has a self-intersection point $a \neq b$: $\gamma_i(a) = \gamma_i(b)$. Define γ_i^+ as $\gamma_i|_{[a,b]}$ and γ_i^- be defined as $\gamma_i|_{S^1-(a,b)}$. Then, γ_i^+ and γ_i^- become closed curves again with fewer segments than γ_i. Because the winding number of γ_i equals the sum of those of γ_i^+ and γ_i^-, at least one of γ_i^+ or γ_i^- has a nonzero winding number, and we set γ_{i+1} as the one with a nonzero winding number. Each γ_i has a strictly smaller number of segments as i increases and has a winding number not equal to zero. Because γ_0 has finite segments, this process stops in a finite sequence. Therefore, we can get a non-self-intersecting curve $\gamma := \gamma_n$ with a nonzero winding number. If γ has a winding number with an absolute value greater than one, by Lemma 4.14, it must have a self-intersection point. Thus, γ has a winding number 1 or -1. Reverse reparametrization yields a curve with winding number one.

Because g is homotopic to f through linear interpolation and γ is homotopic to ω_1, their compositions are homotopic. This implies the same winding number between $g \circ h$ and $f \circ \omega_1$. Therefore, the winding number of $g \circ \gamma : S^1 \to S^1 = \Sigma^{-1}(1)$ is two. Now consider G. Because G is an embedding, it is injective. Therefore, $G|_{\gamma(S^1)} : \gamma(I) \to S^1 \times \mathbb{R}$ is injective. As the image $G(\gamma(S^1))$ is compact, the image in $S^1 \times \mathbb{R}$ can be embedded in the annulus $\{(x_1, x_2) \in \mathbb{R} | 1 - \epsilon \leq |x_1| + |x_2| \leq 1 + \epsilon\}$. And the map $G \circ \gamma$ has winding number two. However, by Lemma 4.14, any map with winding number two is not injective, leading to a contradiction.

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: No justification

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: No justification

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: No justification

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: No justification

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.