Optimal Minimum Width for the Universal Approximation of Continuously Differentiable Functions by Deep Narrow MLPs

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

Abstract

In this paper, we investigate the universal approximation property of deep, narrow multilayer perceptrons (MLPs) for C^1 functions under the Sobolev norm, specifically the $W^{1,\infty}$ norm. Although the optimal width of deep, narrow MLPs for approximating continuous functions has been extensively studied, significantly less is known about the corresponding optimal width for C^1 functions. We demonstrate that *the optimal width* can be determined in a wide range of cases within the C^1 setting. Our approach consists of two main steps. First, leveraging control theory, we show that any diffeomorphism can be approximated by deep, narrow MLPs. Second, using the Borsuk-Ulam theorem and various results from differential geometry, we prove that the optimal width for approximating arbitrary C^1 functions via diffeomorphisms is $\min(n+m, \max(2n+1, m))$ in certain cases, including $(n, m) = (8, 8)$ and $(16, 8)$, where n and m denote the input and output dimensions, respectively. Our results apply to a broad class of activation functions.

1 Introduction

$$\min(n+m, \max(2n+1, m)) \tag{1}$$

$$\|f-g\|_{W^{k,p}(U)} := \sum_{|\alpha|\le k} \|D^\alpha(f-g))\|_{L^p(U)} \tag{2}$$

The choice of neural network architecture plays a crucial role in determining performance. However, in practice, architectural decisions are often made through trial and error. It is therefore important to provide theoretical guidance on what should be avoided and how to select appropriate width and depth based on the input space, target function, and specific tasks. The *universal approximation property* (UAP) refers to the ability of deep learning models to approximate a given class of functions. Since deep networks must approximate general functions to perform specific tasks, the UAP has received considerable attention as a theoretical foundation. While various forms of universal approximation theorems exist depending on the network type and its characteristics, one actively studied setting is the universal approximation property of *deep, narrow multilayer perceptrons* (deep, narrow MLPs), which reflects the practical scenario where networks are deep but relatively narrow in width.

MLPs with fixed width and arbitrarily large depth exhibit different universal approximation behavior depending on whether their width exceeds a critical threshold (Johnson, 2018; Kidger & Lyons, 2020). This threshold is called the *minimum width*, and numerous studies have investigated upper and lower bounds for this threshold based on the input dimension n, output dimension m, and the choice of activation function.

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

The most intensively studied case involves the approximation of continuous functions under the uniform norm. Notable results include the upper bound $n + m + c(\sigma)$, where $c(\sigma)$ is a constant depending on the activation function, shown by Hanin & Sellke (2017); Kidger & Lyons (2020). More recently, Hwang (2023) improved this upper bound to $\max(2n + 1, n)$.

For lower bounds, Johnson (2018); Cai (2022); Kim et al. (2023) proved that the minimum width must be at least $n + 1$ or $m + \mathbf{1}_{d<m\leq 2n}$, depending on the setting. However, few studies have succeeded in narrowing the gap between known lower and upper bounds. Among the few, Park et al. (2020); Hwang (2023) proved optimality in specific cases: the minimum width is 3 for $(n, m) = (1, 2)$ and 4 for $(2, 2)$.

Beyond the uniform norm, there has also been research under other norms. Park et al. (2020) established the optimal minimum width of deep, narrow MLPs with ReLU activation in the L_p norm. However, research on general norms beyond the L_p and uniform norms remains limited.

However, there has been a scarce number of papers that study norms involving derivatives of functions in the setting of deep narrow MLPs. Many deep learning techniques directly penalize the difference between the derivative of the target function and that of the network. These include Sobolev Training (Czarnecki et al., 2017), Physics-Informed Neural Networks (PINNs) (Raissi et al., 2019), and Generative Adversarial Networks with gradient penalty (Gulrajani et al., 2017; Arbel et al., 2018).

In this work, we determine the minimum width required to approximate continuously differentiable functions in Sobolev spaces. Specifically, we focus on approximation with respect to the $W^{1,\infty}$ norm. Compared to the uniform norm, the topology of the Sobolev norm $W^{1,\infty}$ is finer, enabling tighter lower bound estimates. For the upper bound, tools from control theory allow us to match the upper bounds known in the uniform norm setting. Using these ideas, we compute both upper and lower bounds for approximation in Sobolev spaces. In some cases, the lower bound coincides with the upper bound, thus identifying the optimal minimum width. This includes interesting cases such as $(8, 8)$ and $(16, 8)$. The exact pairs to which our result applies can be found in Theorem 5.9.

Our contributions are as follows:

- We show that deep, narrow MLPs can approximate arbitrary diffeomorphisms with respect to the Sobolev norm $W^{1,\infty}$. (Theorem 4.1)
- We precisely characterize the additional width required to approximate arbitrary continuously differentiable functions as compositions of diffeomorphisms and linear transformations. (Definition 4.3 and Theorem 4.6)
- Using these results, we prove that the known upper bounds $n + m$ and $\max(2n + 1, m)$ under the uniform norm also hold under the Sobolev norm $W^{1,\infty}$. (Theorem 5.1)
- We prove that these upper bounds are also lower bounds for infinitely many combinations of n and m, and therefore, these values represent the optimal minimum width in those cases. (Theorem 5.9)

2 Related Words

In this section, we review previous studies on the universal approximation property (UAP). Cybenko (1989) proved that a two-layer MLP possesses the UAP in the space of continuous functions. This result was extended by Leshno et al. (1993) to more general activation functions.

While these early results focus on two-layer networks, subsequent research has investigated the UAP of deep, narrow MLPs. Hanin & Sellke (2017) established a universal approximation theorem for deep, narrow MLPs with ReLU activation, providing both lower and upper bounds on the minimum width. Johnson (2018) showed that a width of at least $n + 1$ is required for networks with monotonic activation functions. Kidger & Lyons (2020) proved that a width of $n + m + 1$ suffices for general non-polynomial activation functions, while $n + m + 2$ is sufficient for polynomial activations. Park et al. (2020) demonstrated that the optimal minimum width is three when $n = 1$ and $m = 2$ with ReLU. Cai (2022) showed that a width of at least $\max(n, m)$ is necessary for general activation functions. Kim et al. (2023) proved a lower bound of $m + \mathbf{1}_{n<m\leq 2m}$. Hwang (2023) established an upper bound of $\max(2n + 1, m)$ for networks using the Leaky-ReLU activation and showed that the optimal minimum width is four when $n = m = 2$.

There have also been investigations of the UAP under norms other than the uniform norm. Park et al. (2020) showed that the optimal minimum width is $\max(n+1, m)$ in the $L_p(\mathbb{R}^n, \mathbb{R}^m)$ space for ReLU networks. Additionally, Kim et al. (2024) demonstrated that in the $L_p([0,1]^n, \mathbb{R}^m)$ setting, the optimal minimum width becomes $\min(n, m, 2)$.

In addition to studies on MLPs, there has been significant progress in understanding the universal approximation property of residual networks (ResNets). Lin & Jegelka (2018) demonstrated that even ResNets with one-neuron hidden layers can serve as universal approximators, highlighting the expressive power that arises from their residual structure. Aizawa et al. (2020) extended this line of research by analyzing both ResNets and ODENets, providing rigorous mathematical results along with supporting numerical experiments. More recently, Tabuada & Gharesifard (2022) investigated ResNets from a control-theoretic perspective.

Beyond function value approximation, some universal approximation theorems also consider derivatives. Li (1996) proved that a two-layer MLP can approximate arbitrary derivatives of a function, provided the activation function is sufficiently smooth.

However, these results do not cover the Sobolev norm in the context of deep narrow MLPs. In this paper, we provide a partial answer to this open question by establishing results under the $W^{1,\infty}$ norm.

3 Notation and Definition

In this section, we introduce the notations and definitions used throughout this paper: $\mathbb{N}$ denotes the set of natural numbers, and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. $B_n(r)$ denotes the open ball in $\mathbb{R}^n$ centered at the origin with radius r. For a set $A \subset \mathbb{R}^d$, $\overline{A}$ denotes the closure of A with respect to the Euclidean norm.

For two open sets $V \subset U \subset \mathbb{R}^d$, we say that V is a precompact subset of U if $\overline{V} \subset \mathbb{R}^d$ is compact and $\overline{V} \subset U$. We denote this as $V \Subset U$. For sets $A, B \subset \mathbb{R}^d$, the Minkowski sum is defined as $A + B = \{x + y \in \mathbb{R}^d \mid x \in A,\, y \in B\}$.

For a d-dimensional vector $x \in \mathbb{R}^d$, we denote by x_i the i-th component of x; in other words, $x = (x_1, x_2, \ldots, x_d)$. Similarly, for a function $f : X \to \mathbb{R}^n$, we write f_i to denote the i-th component function, so that $f(x) = (f_1(x), \ldots, f_n(x))$. We use $x_{i:j}$ to represent the $(j-i+1)$-dimensional subvector $(x_i, x_{i+1}, \ldots, x_j)$. For vectors $x, y \in \mathbb{R}^d$, the dot product is denoted by $x \cdot y \in \mathbb{R}$ and defined as $x \cdot y := \sum_{i=1}^{d} x_i y_i$. For vectors $x = (x_1, \ldots, x_{d_1}) \in \mathbb{R}^{d_1}$ and $y = (y_1, \ldots, y_{d_2}) \in \mathbb{R}^{d_2}$, we define the operation $\oplus$ as $x \oplus y := (x_1, \ldots, x_{d_1}, y_1, \ldots, y_{d_2}) \in \mathbb{R}^{d_1+d_2}$. Similarly, for functions $f : X \to \mathbb{R}^{d_1}$ and $g : X \to \mathbb{R}^{d_2}$, we define $f \oplus g : X \to \mathbb{R}^{d_1+d_2}$ by $(f \oplus g)(x) := f(x) \oplus g(x)$.

Let $\mathrm{Aff}_{n,m}$ denote the set of affine transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$. For a function $f : X \to Y$ and a subset $X' \subset X$, we write $f|_{X'}$ to denote the restriction of f to the domain X'. For $r \in \mathbb{N}_0$, the space $C^r(X; Y)$ denotes the set of functions that are r-times continuously differentiable. For $U \subset \mathbb{R}^n$ and $\mathbf{r} = (r_1, \ldots, r_n) \in \mathbb{N}_0^n$, the space $C^{\mathbf{r}}(U; \mathbb{R}^m)$ consists of functions f such that the mixed partial derivative $\frac{\partial^{r_1+\cdots+r_n} f}{\partial x_1^{r_1} \ldots \partial x_n^{r_n}}$ exists and is continuous. For $k \in \mathbb{N}_0$ and $r \in [0,1]$, the space $C^{k,r}(U; \mathbb{R}^m)$ consists of functions whose k-th order partial derivatives are Hölder continuous with exponent r. In particular, $C^{0,1}(U; \mathbb{R}^m)$ denotes the space of Lipschitz continuous functions. We define $C^{0,1}_{\mathrm{loc}}(U; \mathbb{R}^m)$ as the space of locally Lipschitz continuous functions: that is, $f \in C^{0,1}_{\mathrm{loc}}(U; \mathbb{R}^m)$ if for every precompact set $V \Subset U$, there exists a constant L_V such that $\|f(x) - f(y)\| \le L_{\overline{V}} \|x - y\|$ for all $x, y \in \overline{V}$. We denote the Lipschitz constant of f on $\overline{V}$ by $\mathcal{L}_V(f)$.

3.1 Sobolev Space

We define the Sobolev space as follows: We denote the weak derivative of u by Du.

Definition 3.1 (Sobolev Space). *Let $n, k \in \mathbb{N}$, $p \in \mathbb{N} \cup \{\infty\}$, and let $U \subset \mathbb{R}^n$ be an open set. The Sobolev space $W^{k,p}(U)$ is defined by*

$$W^{k,p}(U) := \{u \in L^p(U) \mid D^\alpha u \in L^p(U) \text{ for all multi-indices } \alpha \text{ with } |\alpha| \le k\}, \tag{3}$$

equipped with the norm

$$\|u\|_{W^{k,p}(U)} := \sum_{|\alpha| \le k} \|D^\alpha u\|_{L^p(U)}. \tag{4}$$

The vector-valued Sobolev space $W^{k,p}(U;\mathbb{R}^m)$ for $m,k \in \mathbb{N}$ is defined as

$$W^{k,p}(U;\mathbb{R}^m) := \left\{ u = (u_1, \dots, u_m) \;\middle|\; u_i \in W^{k,p}(U) \right\}, \tag{5}$$

with the norm

$$\|u\|_{W^{k,p}(U;\mathbb{R}^m)} := \sum_{i=1}^{m} \|u_i\|_{W^{k,p}(U)}. \tag{6}$$

More specifically, we focus on the following local Sobolev space, considering compact domains:

Definition 3.2 (Local Sobolev Space)**.** *Let $U \subset \mathbb{R}^m$, $r \in \mathbb{N}_0$, and $p \in [1,\infty]$. The local Sobolev space $W^{r,p}_{\mathrm{loc}}(U;\mathbb{R}^n)$ is defined as the projective limit:*

$$W^{r,p}_{\mathrm{loc}}(U;\mathbb{R}^n) := \varprojlim_{V \Subset U} W^{r,p}(V;\mathbb{R}^n), \tag{7}$$

where the right-hand side is given explicitly as

$$\left\{ (f_V)_V \in \prod_{V \Subset U} W^{r,p}(V;\mathbb{R}^n) \;\middle|\; f_{V_1}|_{V_2} = f_{V_2} \text{ for all } V_2 \subset V_1 \right\}. \tag{8}$$

The local Sobolev space is equipped with the relative topology inherited from the product topology of the spaces $W^{r,p}(V;\mathbb{R}^n)$.

In this paper, we focus on the Sobolev norm $W^{1,\infty}$. It is well known that $W^{1,\infty}_{\mathrm{loc}} = C^{0,1}_{\mathrm{loc}}$. See Theorem 4.5, p.155 in Evans (2018) for details. It is also known that for convex domains, the Sobolev and Lipschitz spaces coincide (Theorem 4.1 in Heinonen (2005)): if $V \subset \mathbb{R}^d$ is convex, then $W^{1,\infty}(V) = C^{0,1}(V)$. Moreover, there exist constants $C_1, C_2 > 0$ depending only on d and n such that (see Theorem 4, p.279 and Theorem 6, p.281 in Evans (2022)):

$$C_1 \|f\|_{W^{1,\infty}(V;\mathbb{R}^n)} \leq \mathcal{L}_V(f) + \|f\|_{L^\infty(V)} \leq C_2 \|f\|_{W^{1,\infty}(V;\mathbb{R}^n)}. \tag{9}$$

For convenience, we will always take the continuous representative among functions that differ only on a set of Lebesgue measure zero.

For a set of functions $\mathcal{A} \subset W^{1,p}(U;\mathbb{R}^m)$, we denote by $\overline{\mathcal{A}}^{W^{1,\infty}} = \overline{\mathcal{A}}$ the closure of $\mathcal{A}$ with respect to the norm $\|\cdot\|_{W^{1,p}(U;\mathbb{R}^m)}$. Similarly, for a set of functions $\mathcal{A} \subset W^{1,p}_{\mathrm{loc}}(U;\mathbb{R}^m)$, we denote the closure in the local Sobolev topology by $\overline{\mathcal{A}}^{W^{1,\infty}_{\mathrm{loc}}} = \overline{\mathcal{A}}^{\mathrm{loc}}$.

3.2 Activation Function

We adopt the commonly used condition on activation functions, as proposed by Kidger & Lyons (2020).

Condition 1. *There exist constants $\alpha \in \mathbb{R}$ and $\epsilon \in \mathbb{R}_+$ such that a nonlinear activation function σ is a C^1 function on the interval $(\alpha - \epsilon, \alpha + \epsilon)$, and $\sigma'(\alpha) \neq 0$.*

The ReLU activation function is defined as

$$\mathrm{ReLU}(x) := \begin{cases} x & \text{if } x \geq 0, \\ 0 & \text{if } x < 0 \end{cases} \tag{10}$$

and the Leaky-ReLU activation function is defined as

$$\mathrm{LR}_\beta(x) := \begin{cases} x & \text{if } x \geq 0, \\ \beta x & \text{if } x < 0 \end{cases}, \tag{11}$$

We consider MLPs with sets of activation functions. For example, MLPs with the Leaky-ReLU activation function select an activation function from the following set at each layer:

$$\mathrm{LR} := \{\mathrm{LR}_\beta \mid \beta \in \mathbb{R}_+\}. \tag{12}$$

We use the symbols σ and Σ to denote an activation function and a set of activation functions, respectively. We define Leaky-ReLU-like activation functions as follows:

Definition 3.3 (Leaky-ReLU-like)**.** *A set of activation functions Σ is called* Leaky-ReLU-like *if and only if for each $\beta \in \mathbb{R}_+$, there exists a C^1 activation function $\sigma_\beta \in \Sigma$ such that*

$$\lim_{x\to\infty} \frac{\sigma_\beta(x)}{x} = 1, \quad \lim_{x\to-\infty} \frac{\sigma_\beta(x)}{x} = \beta, \tag{13}$$

and

$$\sup_{x\in\mathbb{R}} |D\sigma_\beta(x) - 1| \xrightarrow[\beta\to 1]{} 0. \tag{14}$$

We also denote the identity function by id:

$$\mathrm{id}(x) := x. \tag{15}$$

Activation functions applied to vectors operate componentwise. For a set of activation functions Σ, define Σ^d as

$$\Sigma^d := \left\{ f : \mathbb{R}^d \to \mathbb{R}^d \,\middle|\, f_i \in \Sigma \right\}. \tag{16}$$

3.3 Deep Narrow MLP

We define the set of deep, narrow MLPs with a set of activation functions Σ, arbitrary depth, input dimension n, output dimension m, and at most w intermediate dimensions as $\Delta^\Sigma_{n,m,w}$. (The exact definition is provided in Appendix A.1.) For a singleton activation function σ, we define:

$$\Delta^\sigma_{n,m,w} := \Delta^{\{\sigma\}}_{n,m,w}. \tag{17}$$

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 zero-padding inclusion $q_{m,n} : \mathbb{R}^m \to \mathbb{R}^n$ as:

$$p_{n,m}(x_1, \dots, x_n) := (x_1, \dots, x_m), \tag{18}$$

$$q_{m,n}(x_1, \dots, x_m) := (x_1, \dots, x_m, 0, \dots, 0). \tag{19}$$

Any function $f \in \Delta^\Sigma_{n,m,w}$ can be decomposed as:

$$f = p_{w,n} \circ g \circ q_{n,w}, \tag{20}$$

where $g \in \Delta^\Sigma_{w,w,w}$. Note that if $g_1, g_2 \in \Delta^\Sigma_{w,w,w}$, then their composition $g_1 \circ g_2$ also belongs to $\Delta^\Sigma_{w,w,w}$.

From this point on, we will use the notation σ to refer to either a single activation function or a set of activation functions Σ, depending on the context.

3.4 Subsets of Diffeomorphisms

We define the sets of diffeomorphisms. For definitions of concepts from differential geometry, see Appendix A.2.

Definition 3.4 (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. Then $\mathcal{D}^r(U)$ denotes the set of C^r-diffeomorphisms from U to $\mathbb{R}^d$.*

4 Universal Approximation

4.1 Problem Formulation

Our primary goal is to identify the minimum width $w^{W^{1,\infty}}_{\min} \in \mathbb{N}$ such that any continuously differentiable function $f \in C^1(\mathbb{R}^n; \mathbb{R}^m)$ can be approximated by elements of $\Delta^\sigma_{n,m,w^{W^{1,\infty}}_{\min}}$ in the topology of $W^{1,\infty}_{\mathrm{loc}}(\mathbb{R}^n; \mathbb{R}^m)$. In other words, our aim is to determine the value of $w^{W^{1,\infty}}_{\min}(n, m, \sigma)$ such that

$$w^{W^{1,\infty}}_{\min}(n, m, \sigma) := \min \left\{ l \in \mathbb{N} \,\middle|\, C^1(\mathbb{R}^n; \mathbb{R}^m) \subset \overline{\Delta^\sigma_{n,m,l}}^{W^{1,\infty}_{\mathrm{loc}}} \right\}. \tag{21}$$

$w^{W^{1,\infty}}_{\min}(n, m, \sigma)$ denotes the minimum width for which MLPs of this width and arbitrary depth can approximate C^1 functions to any accuracy in the $W^{1,\infty}$ norm.

4.2 Diffeomorphisms and Continuously Differentiable Functions

Our proof strategy is divided into two parts. First, we approximate a diffeomorphism using deep narrow MLPs with a small additional width. Next, we show that any continuously differentiable function can be approximated by a composition of affine transformations and diffeomorphisms, and we rigorously estimate the required width. In this subsection, we aim to prove the following theorem, which asserts that a diffeomorphism can be approximated by deep narrow MLPs.

Theorem 4.1. *Let σ be one of a non-polynomial $C^{1,1}$-function, LR, ReLU, or Leaky-ReLU-like activation functions. Then, for any natural number $d \in \mathbb{N}$, the following relation holds:*

$$\mathcal{D}^1(\mathbb{R}^d) \subset \overline{\Delta^{\sigma}_{d,d,d+\alpha(\sigma)}}^{\mathrm{loc}}, \tag{22}$$

where

$$\alpha(\sigma) = \begin{cases} 0 & \text{if } \sigma = \text{LR or } \sigma \text{ is Leaky-ReLU-like} \\ 1 & \text{if } \sigma = \text{ReLU or } \sigma \text{ is a non-polynomial } C^{1,1}\text{-function} \end{cases}. \tag{23}$$

The theorem states that deep, narrow MLPs with a small additional width can approximate arbitrary diffeomorphisms. The proof relies on techniques from control theory. Approximating an entire diffeomorphism directly using a neural network is challenging. To address this, we interpret a diffeomorphism as the solution of an ordinary differential equation evolving over time. In other words, the existence of a diffeomorphism implies the existence of a continuous flow connecting the identity map to the diffeomorphism. The direction and magnitude of this flow are determined by a vector field. The problem then reduces to approximating this continuous flow step by step, which is equivalent to approximating a vector field. Deep, narrow MLPs can approximate such flows over sufficiently small time intervals by leveraging the universal approximation property. Then, by approximating the flow generated by this two-layer MLP using a deep narrow MLP, we complete the argument. The full proof is provided in Appendix C.1.

Now, we introduce a quantity $\Omega(n,m)$ such that any continuously differentiable function from $[0,1]^n$ to $\mathbb{R}^m$ can be approximated by a composition of affine transformations and $\Omega(n,m)$-dimensional diffeomorphisms. We further show that this width is optimal. To this end, we begin with the following lemma.

Lemma 4.2 (Theorem C of Palais (1960))**.** *Let $n, m \in \mathbb{N}$ with $n \leq m$, and let $f : K = [0,1]^n \to \mathbb{R}^m$ be a smooth embedding. Then, there exists a smooth diffeomorphism $F : \mathbb{R}^m \to \mathbb{R}^m$ such that the following equation holds:*

$$F \circ q_{n,m} = f. \tag{24}$$

The lemma implies that any smooth embedding can be decomposed into an affine transformation followed by a diffeomorphism. Now, let $\mathrm{Emb}(X,Y)$ denote the set of smooth embeddings from X to Y. We define the quantity $\Omega(n,m)$ as follows:

Definition 4.3 ($\Omega(n,m)$)**.**

$$\Omega(n,m) := \min \left\{ l \in \mathbb{N}_0 \,\middle|\, p_{l,m}\left(\overline{\mathrm{Emb}([0,1]^n, \mathbb{R}^l)}\right) = C^1([0,1]^n; \mathbb{R}^m) \right\}, \tag{25}$$

where the closure is taken with respect to the C^1-norm.

Using the lemma above and the definition of $\Omega(n,m)$, we state the following theorem:

Theorem 4.4. *Let σ be one of a non-polynomial $C^{1,1}$-function, LR, ReLU, or Leaky-ReLU-like activation functions. Then, for any natural numbers n and m, the following relation holds:*

$$C^1(\mathbb{R}^n; \mathbb{R}^m) \subset \overline{\Delta^{\sigma}_{n,m,\Omega(n,m)+\alpha(\sigma)}}^{\mathrm{loc}}, \tag{26}$$

where

$$\alpha(\sigma) = \begin{cases} 0 & \text{if } \sigma = \text{LR or } \sigma \text{ is Leaky-ReLU-like} \\ 1 & \text{if } \sigma = \text{ReLU or } \sigma \text{ is a non-polynomial } C^{1,1}\text{-function} \end{cases}. \tag{27}$$

The proof of the theorem is provided in Appendix D.1. The preceding theorem shows that $\Omega(n,m)$ is a sufficient width for approximating functions with n-dimensional input and m-dimensional output. Conversely, the following proposition demonstrates that $\Omega(n,m)$ is also a necessary width for such approximation.

Proposition 4.5. *Let σ be a set of non-decreasing, C^1 activation functions. Then, for natural numbers n and m, the following relation holds:*

$$C^1(\mathbb{R}^n;\mathbb{R}^m) \not\subset \overline{\Delta^{\sigma}_{n,m,\Omega(n,m)-1}}^{\text{loc}}. \tag{28}$$

The proof of this proposition is provided in Appendix D.2. By combining the previous theorems with this proposition, we derive the following result. This theorem demonstrates that the purely geometrically defined quantity $\Omega(n,m)$ has a fundamental connection to the minimum width of deep narrow MLPs.

Theorem 4.6. *The following relation holds:*

$$w_{\min}^{W^{1,\infty}}(n,m,\sigma) = \Omega(n,m) \tag{29}$$

for a Leaky-ReLU-like σ in which every element is increasing, and

$$\Omega(n,m) \leq w_{\min}^{W^{1,\infty}}(n,m,\sigma) \leq \Omega(n,m)+1, \tag{30}$$

for a set of $C^{1,1}$ increasing activation functions σ.

5 Calculation of $\Omega(n,m)$

In the previous section, we showed that $\Omega(n,m)$ nearly determines the minimum width required for universal approximation. In this section, we provide general bounds for $\Omega(n,m)$ and compute exact values for specific cases.

5.1 Upper Bound of $\Omega(n,m)$

We begin by establishing the following general upper bound.

Theorem 5.1. *The following relation holds:*

$$\Omega(n,m) \leq \min(n+m, \max(2n+1,m)). \tag{31}$$

Proof. The inequality $\Omega(n,m) \leq n+m$ follows directly from the definition of $\Omega(n,m)$. $\Omega(n,m) \leq \max(2n+1,m)$ is by Lemma 5.2. □

Lemma 5.2. *Consider natural numbers n and m where $m > 2n$. Let $f \in C^1(\mathbb{R}^n;\mathbb{R}^m)$ be a continuously differentiable function. Then, for a bounded open set $U \subset \mathbb{R}^n$ and a positive number $\epsilon \in \mathbb{R}_+$, there exists a smooth embedding $g : \overline{U} \to \mathbb{R}^m$ such that*

$$\|f-g\|_{W^{1,\infty}(U,\mathbb{R}^m)} < \epsilon. \tag{32}$$

Proof. This is a direct consequence of the transversality theorem. (See Chapter 3, Theorem 2.1 of Hirsch (2012) for details) □

In some cases, we can improve the general bound established above.

Lemma 5.3. *For even k, the following eqaution holds:*

$$\Omega(k,2k-1) = 2k. \tag{33}$$

Proof. By Kim et al. (2023), we have $\Omega(k,2k-1) \geq 2k$. Thus, it suffices to prove that $\Omega(k,2k-1) \leq 2k$. As immersions are dense in $C^1(\mathbb{R}^k,\mathbb{R}^{2k-1})$, it is enough to approximate an immersion f. By Corollary 3.2 of Lashof & Smale (1959), there exists a smooth embedding g such that $\|p_{2k,2k-1} \circ g - f\|_{W^{1,\infty}(U;\mathbb{R}^m)} < \epsilon$. Note that while the original result is stated for the uniform norm, the same proof applies directly in the C^1 norm setting. □

5.2 Lower Bound of $\Omega(n, m)$

In this subsection, we present a lower bound for certain cases, which coincides with the upper bound established in the previous section, thereby yielding the optimal minimum width. To prove the lower bound, we require an argument of the following form: Given a function $f : \mathbb{R}^n \to \mathbb{R}^m$, there exists $\epsilon > 0$ such that if the codomain dimension of another function g is small, then the concatenation $f \oplus g$ cannot be an embedding. To this end, we construct a function f whose self-intersection is transversal and has the structure of a sphere S^r. If all antipodal points on the sphere are mapped to the same value by f, then we can apply the Borsuk–Ulam theorem.

Lemma 5.4 (Borsuk–Ulam Theorem)**.** *Let $h : S^n \to \mathbb{R}^n$ be a continuous function. Then there exists a point $x \in S^n$ such that*

$$h(x) = h(-x). \tag{34}$$

The Borsuk–Ulam theorem states that every continuous map from an n-dimensional sphere to $\mathbb{R}^n$ maps some pair of antipodal points to the same point. Now, suppose we have an embedding $S^r \hookrightarrow \mathbb{R}^n$ and a map f such that $f(x) = f(-x)$ for all antipodal points $x \in S^r$. Then, for any function $g : \mathbb{R}^r \to \mathbb{R}^r$, there exists a pair of antipodal points on S^r that are mapped to the same value by g. Therefore, the map $G = f \oplus g$ cannot be injective, and hence cannot be an embedding. This leads to the conclusion $\Omega(n, m) \geq m + r + 1$.

The difficulty, however, lies in the fact that we must consider a map G such that $\|p_{m+r,n} \circ G - f\|$ is small, rather than requiring exact equality $p_{m+r,n} \circ G = f$. The following lemma guarantees that the diffeomorphic structure of the self-intersection is preserved under small perturbations in the C^1 norm.

Lemma 5.5 (Ehresmann's Lemma for Intersection)**.** *For $n, m \in \mathbb{N}$ with $2n > m$, consider a precompact set $U \subset \mathbb{R}^n$ and a C^1 function $f : U \to \mathbb{R}^m$ in transversal position. Then there exists $\epsilon > 0$ such that the following holds: Consider arbitrary $g \in C^1(U; \mathbb{R}^m)$ satisfying*

$$\|f - g\|_{W^{1,\infty}(U;\mathbb{R}^m)} < \epsilon. \tag{35}$$

Define the diagonal Δ of $U \times U$ as

$$\Delta := \{(x, x) \in U \times U \mid x \in U\}. \tag{36}$$

Define $\widetilde{f} : U \times U - \Delta \to \mathbb{R}^m$ as

$$\widetilde{f}(x, y) := f(x) - f(y). \tag{37}$$

Similarly, define $\widetilde{g}$ as

$$\widetilde{g}(x, y) := g(x) - g(y). \tag{38}$$

Then, there exists a C^1 diffeomorphism $\Phi : \widetilde{f}^{-1}(0) \to \widetilde{g}^{-1}(0)$ such that

$$\Phi(x, y) = T(\Phi(y, x)), \tag{39}$$

where T denotes the involution $(x, y) \mapsto (y, x)$.

The proof of Lemma 5.5 is provided in Appendix E.1. Now, the only remaining task is to construct a function with such a self-intersection structure. The following two lemmas provide results for specific cases.

Lemma 5.6. *Assume that there exists a submersion $f : \mathbb{RP}^{n-1} \times (-1, 1) \to \mathbb{R}^m$. Then, the following relation holds:*

$$\Omega(n, m) = n + m. \tag{40}$$

Proof. Let $f : \mathbb{RP}^{n-1} \times (-1, 1) \to \mathbb{R}^m$ be a submersion. Then, there exists a lifting $\widetilde{f} : S^{n-1} \times (-1, 1) \to \mathbb{R}^m$ such that for a canonical two-to-one covering $p : S^{n-1} \times (-1, 1) \to \mathbb{RP}^{n-1} \times (-1, 1)$, we have $p \circ \widetilde{f} = f$. Then, as all antipodal points have the same $\widetilde{f}$ values and the intersection is transversal, it follows from the previous arguments that $\Omega(n, m) \geq n + m$. This completes the proof. $\square$

Lemma 5.7 (Projective Space Submersion Lemma)**.** *For $n \in \mathbb{N}$, consider $a, b, c \in \mathbb{N}_0$ such that $n + 1 = 2^{4a+b} \times c$, where $0 \leq b \leq 3$ and c is an odd number. Then, for any natural number $m \in \mathbb{N}$ satisfying $m \leq 8a + 2^b$, $\mathbb{RP}^n \times (-1, 1)$ can be submerged into $\mathbb{R}^m$.*

Proof. By Theorem B of Phillips (1967), there exists a submersion $M \to \mathbb{R}^m$ if and only if there exists a section in $F_m(M)$ where $F_m(M)$ is the bundle of m-frames tangent to M. This condition is equivalent to the existence of m linearly independent vector fields. By Theorem 1.1 of Davis (2012), the maximum number of linearly independent vector fields on $\mathbb{RP}^n$ equals $8a + 2^b - 1$ where $n + 1 = 2^{4a+b} \times c$ for $0 \leq b \leq 3$ and an odd number $c \in \mathbb{N}$. Therefore, $\mathbb{RP}^n \times (-\epsilon, \epsilon)$ has $8a + 2^b$ independent vector fields, and thus can be submerged into $\mathbb{R}^{8a+2^b}$. $\square$

We can also verify that an immersion with an $\mathbb{RP}^n$-structure intersection exists when $3n+1 < 2m \leq 4n$, which implies that $\Omega(n, m) = 2n + 1$.

Lemma 5.8 (Theorem 3 of Miller (1969))**.** *Given n and $m, 3n + 1 < 2m \leq 4n - 2$. Then if $n + 1 \cong 0 \,(\mathrm{mod} c_{2n-m})$ there exists a transversal immersion $S^n \to \mathbb{R}^{m+1}$ with self-intersection $\mathbb{RP}^{2n-m-1}$. Here, c_m is defined as*

$$c_m = \begin{cases} 2^{4r} & \text{if } m = 8r, \\ 2^{4r+1} & \text{if } m = 8r + 1, \\ 2^{4r+2} & \text{if } m = 8r + 2 \text{ or } 8r + 3, \\ 2^{4r+3} & \text{if } m = 8r + 4,\ 8r + 5,\ 8r + 6,\ \text{or } 8r + 7. \end{cases} \tag{41}$$

By combining all the results, we obtain the following theorem.

Theorem 5.9. *If $2^{4a+b} | n$ for $a, b \in \mathbb{N}_0$ satisfying $0 \leq b \leq 3$ and $m \leq 8a + 2^b$,*

$$\Omega(n, m) = m + n. \tag{42}$$

If $\frac{3n+3}{2} < m \leq 2n$ and $n + 1 \cong 0 \,(\mathrm{mod} c_{2n-m+1})$,

$$\Omega(n, m) = 2n + 1. \tag{43}$$

If $2n + 1 \leq m$,

$$\Omega(n, m) = m. \tag{44}$$

Remark 5.10. *At first glance, the dependence of the optimal minimum width on the parity of the input and output dimensions may appear somewhat artificial. However, Lemma 5.3 and Theorem 5.9 together yield the relation*

$$\Omega(k, 2k - 1) = \begin{cases} 2k, & \text{if } k \text{ is even}, \\ 2k + 1, & \text{if } k \text{ is odd}, \end{cases} \tag{45}$$

which provides strong evidence that this parity dependence may be a fundamental property.

6 Limitation

Although our results yield optimal values in many cases—such as $\Omega(8, 8) = 16$ and $\Omega(16, 8) = 24$—they do not apply to all combinations of n and m. Our analysis is asymptotically valid primarily when m is either much smaller than n or significantly larger, specifically in the regime where $3n < 2m$. Developing a theoretical framework that addresses the intermediate regime not covered by our theory would be a compelling direction for future research. Furthermore, determining the exact lower bound in cases where $n + 1$ is not divisible by a power of 2 remains an open and intriguing problem.

Also, our analysis is non-constructive and asymptotic in nature. In particular, we do not provide explicit rates of approximation or quantitative bounds on the depth required for a network to achieve a given precision. As a result, our results establish existence guarantees but leave open the practical question of how deep a network must be to approximate a target function within a prescribed accuracy. This limitation stands in contrast to constructive approximation results that do provide explicit dependence on approximation error.

Furthermore, our work does not characterize the role of the smoothness of the target function in the approximation behavior. It is natural to expect that smoother functions should be easier to approximate, and indeed prior studies have demonstrated this by analyzing approximation rates in terms of Sobolev smoothness classes (Schmidt-Hieber, 2020). Incorporating such smoothness-dependent considerations into the analysis of deep, narrow networks remains an important direction for future work.

7 Conclusion

In this study, we investigated the minimum width of deep narrow MLPs required to approximate continuously differentiable functions under the Sobolev norm. Our analysis established optimality in a broad range of cases. However, our proof techniques rely on the robustness of the topological structure under small perturbations in the derivatives of the target functions and therefore do not directly extend to the uniform norm. Nonetheless, the structure of the proofs suggests that similar bounds may still hold under the uniform norm. Developing more refined algebraic topological tools to rigorously bridge this gap presents an interesting direction for future research.

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

Aizawa, Y., Kimura, M., and Matsui, K. Universal approximation properties for an odenet and a resnet: Mathematical analysis and numerical experiments. *arXiv preprint arXiv:2101.10229*, 2020.

Arbel, M., Sutherland, D. J., Bińkowski, M., and Gretton, A. On gradient regularizers for mmd gans. *Advances in neural information processing systems*, 31, 2018.

Biagi, S. and Bonfiglioli, A. *An Introduction to the Geometrical Analysis of Vector Fields: with Applications to Maximum Principles and Lie Groups*. World Scientific, 2019.

Cai, Y. Achieve the minimum width of neural networks for universal approximation. In *The Eleventh International Conference on Learning Representations*, 2022.

Caponigro, M. Orientation preserving diffeomorphisms and flows of control-affine systems. *IFAC Proceedings Volumes*, 44(1):8016–8021, 2011.

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

Czarnecki, W. M., Osindero, S., Jaderberg, M., Swirszcz, G., and Pascanu, R. Sobolev training for neural networks. *Advances in neural information processing systems*, 30, 2017.

Davis, D. Vector fields on rp^n $\times$rp^m. *Proceedings of the American Mathematical Society*, 140(12): 4381–4388, 2012.

Evans, L. *Measure theory and fine properties of functions*. Routledge, 2018.

Evans, L. C. *Partial differential equations*, volume 19. American Mathematical Society, 2022.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Hanin, B. and Sellke, M. Approximating continuous functions by relu nets of minimal width. *arXiv preprint arXiv:1710.11278*, 2017.

Heinonen, J. *Lectures on Lipschitz analysis*. Number 100. University of Jyväskylä, 2005.

Hirsch, M. W. *Differential topology*, volume 33. Springer Science & Business Media, 2012.

Hwang, G. Minimum width for deep, narrow mlp: A diffeomorphism and the whitney embedding theorem approach. *arXiv preprint arXiv:2308.15873*, 2023.

Johnson, J. Deep, skinny neural networks are not universal approximators. *arXiv preprint arXiv:1810.00393*, 2018.

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. *arXiv preprint arXiv:2309.10402*, 2023.

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, X. Simultaneous approximations of multivariate functions and their derivatives by neural networks with one hidden layer. *Neurocomputing*, 12(4):327–343, 1996.

Lin, H. and Jegelka, S. Resnet with one-neuron hidden layers is a universal approximator. *Advances in neural information processing systems*, 31, 2018.

Miller, J. G. Self-intersections of some immersed manifolds. *Transactions of the American Mathematical Society*, 136:329–338, 1969.

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. *arXiv preprint arXiv:2006.08859*, 2020.

Phillips, A. Submersions of open manifolds. *Topology*, 6(2):171–206, 1967.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Schmidt-Hieber, J. Nonparametric regression using deep neural networks with relu activation function. 2020.

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

A Definitions and Notations

A.1 Sets of Neural Networks

For a set of activation functions Σ, the set of MLPs denoted by $\mathcal{N}^{\Sigma}_{d_0,d_1,\dots,d_N}$ is defined as:

$$\mathcal{N}^{\Sigma}_{d_0,d_1,\dots,d_N} := \left\{ f : \mathbb{R}^{d_0} \to \mathbb{R}^{d_N} \;\middle|\; W_i \in \mathrm{Aff}_{d_{i-1},d_i},\; g_i \in \Sigma^{d_i},\; f = W_N \circ g_{N-1} \circ \cdots \circ g_1 \circ W_1 \right\}.$$

Note that, in general, an MLP can have different activation functions in each layer. If the set Σ is a singleton, i.e., $\Sigma = \{\sigma\}$, we omit the set notation and simply write:

$$\mathcal{N}^{\sigma}_{d_0,d_1,\dots,d_N} := \mathcal{N}^{\{\sigma\}}_{d_0,d_1,\dots,d_N}. \tag{46}$$

We define the set of deep, narrow MLPs with input dimension n, output dimension m, and at most w intermediate dimensions as:

$$\Delta^{\Sigma}_{n,m,w} := \bigcup_{N\in\mathbb{N}_0} \bigcup_{1\le d_1,\dots,d_N\le w} \mathcal{N}^{\Sigma}_{n,d_1,\dots,d_N,m}. \tag{47}$$

A.2 Some Definitions from Differential Geometry

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

Definition A.2 (Immersion)**.** *Let M and N be smooth manifolds, and let $f : M \to N$ be a C^1-map. The map f is called an* immersion *if for every point $p \in M$, the differential*

$$df_p : T_pM \to T_{f(p)}N \tag{48}$$

is injective.

Definition A.3 (Submersion)**.** *Let M and N be smooth manifolds, and let $f : M \to N$ be a C^1-map. The map f is called a* submersion *if, for every point $p \in M$, the differential*

$$df_p : T_pM \to T_{f(p)}N \tag{49}$$

is surjective.

Definition A.4 (Embedding)**.** *Let M and N be smooth manifolds, and let $f : M \to N$ be a C^1-map. The map f is called an embedding if it is an immersion and a homeomorphism onto its image $f(M)$, where $f(M)$ is equipped with the subspace topology from N.*

Definition A.5 (Transversality)**.** *Let M, N, P be smooth manifolds and let $f : M \to P$, $g : N \to P$ be smooth maps. We say that f and g are* transverse *(written $f \pitchfork g$) if for every pair of points $p \in M$, $q \in N$ with $f(p) = g(q)$, we have*

$$df_p(T_pM) + dg_q(T_qN) = T_{f(p)}P. \tag{50}$$

That is, the images of the differentials at p and q together span the tangent space of P at $f(p) = g(q)$.

B Practical Lemmas

In this section, we present several useful lemmas that are employed throughout the paper. The composition of functions is addressed by the following lemma.

Lemma B.1. *Let $f_i \to f$ in the $W^{1,\infty}_{\mathrm{loc}}(\mathbb{R}^m;\mathbb{R}^l)$ topology and $g_i \to g$ in the $W^{1,\infty}_{\mathrm{loc}}(\mathbb{R}^n;\mathbb{R}^m)$ topology. Then $f_i \circ g_i \to f \circ g$ in the $W^{1,\infty}_{\mathrm{loc}}(\mathbb{R}^n;\mathbb{R}^l)$ topology.*

Proof. It is sufficient to prove that, for each $V \Subset \mathbb{R}^n$, the Lipschitz constant of $f \circ g - f_i \circ g_i$ on V converges to zero as i increases. Choose a sufficiently large number i_0 such that for any $i \ge i_0$, we have $\|g - g_i\|_{L^\infty(V;\mathbb{R}^m)} < 1$. Then,

$$\begin{aligned}\mathcal{L}_V(f\circ g - f_i\circ g_i) &\le \mathcal{L}_V(f\circ g - f\circ g_i) + \mathcal{L}_V(f\circ g_i - f_i\circ g_i)\\ &\le \mathcal{L}_{g(V)+B_m(1)}(f)\mathcal{L}_V(g-g_i) + \mathcal{L}_{g(V)+B_m(1)}(f-f_i) \xrightarrow{i\to\infty} 0,\end{aligned} \tag{51}$$

where $g(V) + B_n(1)$ is the Minkowski sums. □

This lemma implies that if each function can be approximated by neural networks in the local Sobolev topology, then their composition can also be approximated in the same topology.

We can apply a partial activation function using the following lemma.

Lemma B.2. *For natural numbers $n, m, w \in \mathbb{N}$, and an activation function σ satisfying Condition 1, the following relation holds:*

$$\overline{\Delta^{\sigma}_{n,m,w}}^{\text{loc}} \supset \Delta^{\{\sigma,\text{id}\}}_{n,m,w}. \tag{52}$$

Proof. For each $f \in \Delta^{\{\sigma,\text{id}\}}_{n,m,w}$, f can be represented as:

$$f = p_{w,m} \circ g \circ q_{n,w}, \tag{53}$$

where $g \in \Delta^{\{\sigma,\text{id}\}}_{w,w,w}$. By the definition of $\Delta^{\{\sigma,\text{id}\}}_{w,w,w}$, there exists a natural number $N \in \mathbb{N}$ such that the following equation holds:

$$g = W_N \circ g_{N-1} \circ \cdots \circ g_1 \circ W_1, \tag{54}$$

where $W_i \in \text{Aff}_{w,w}, g_i \in \{\sigma, \text{id}\}^w$ for each $i \in [1, N]_{\mathbb{N}}$. By Lemma B.1, if $W_i, g_i \in \overline{\Delta^{\sigma}_{w,w,w}}^{\text{loc}}$ for each $i \in [1, N]_{\mathbb{N}}$, the composition g is also in $\overline{\Delta^{\sigma}_{w,w,w}}^{\text{loc}}$, again, leading to $f \in \overline{\Delta^{\sigma}_{n,m,w}}^{\text{loc}}$. Obviously, $W_i \in \overline{\Delta^{\sigma}_{w,w,w}}^{\text{loc}}$, and it is sufficient to prove that $\overline{\Delta^{\sigma}_{w,w,w}}^{\text{loc}} \supset \{\sigma, \text{id}\}^w$. For $g \in \{\sigma, \text{id}\}^w$, consider $I \subset [1, w]_{\mathbb{N}}$ such that $g_i(x) = \sigma(x)$ if $i \in I$ and $g_i(x) = x$ if $i \notin I$. By Condition 1, there exists $\alpha \in \mathbb{R}$ and $\epsilon \in \mathbb{R}_+$ such that $\sigma'(\alpha) \neq 0$ and σ is C^1 function in $(\alpha - \epsilon, \alpha + \epsilon)$. For an arbitrary precompact set $V \Subset \mathbb{R}$ and sufficiently large M so that $\alpha + \frac{x}{M} \subset (\alpha - \epsilon, \alpha + \epsilon)$ for any $x \in V$, the following relation holds:

$$\left\| \frac{M\left(\sigma\left(\alpha + \frac{x}{M}\right) - \sigma(\alpha)\right)}{\sigma'(\alpha)} - x \right\|_{W^{1,\infty}(V)} \xrightarrow{M \to \infty} 0. \tag{55}$$

Because $\frac{M\left(\sigma\left(\alpha + \frac{x}{M}\right) - \sigma(\alpha)\right)}{\sigma'(\alpha)} \in \mathcal{N}^{\sigma}_{1,1,1}$, the identity function $x \mapsto x \in \overline{\mathcal{N}^{\sigma}_{1,1,1}}^{\text{loc}}$. Define $f_i \in \mathcal{N}^{\sigma}_{1,1,1}$ as

$$f_i(x) = \begin{cases} \sigma(x) & \text{if } x \in I \\ \frac{M\left(\sigma\left(\alpha + \frac{x}{M}\right) - \sigma(\alpha)\right)}{\sigma'(\alpha)} & \text{if } x \notin I \end{cases}, \tag{56}$$

and concatenation $f^M \in \mathcal{N}^{\sigma}_{w,w,w}$ as $f^M(x) := (f_1(x_1), \ldots, f_w(x_w))$. Then, for arbitrary precompact set $V \Subset \mathbb{R}^w$,

$$\|g - f^M\|_{W^{1,\infty}(V;\mathbb{R}^w)} \xrightarrow{M \to \infty} 0. \tag{57}$$

Therefore, $g \in \overline{\Delta^{\sigma}_{w,w,w}}^{\text{loc}}$, and this completes the proof. □

C Proof of Theorem 4.1

C.1 Main Proof of Theorem 4.1

The theorem is proved using the following lemma, which states that any continuously differentiable function can be approximated by a two-layer neural network.

Lemma C.1 (Theorem 2.1. of Li (1996)). *Let K be a compact subset of $\mathbb{R}^s, s \geq 1$, and $f \in C^{\mathbf{m}_1}(K) \cap \cdots \cap C^{\mathbf{m}_q}(K)$, where $\mathbf{m}_i \in \mathbb{N}_0^s$ for $1 \leq i \leq q$. Also, let σ be any non-polynomial function in $C^n(\mathbb{R})$, where $n = \max\{|\mathbf{m}_i| : 1 \leq i \leq q\}$. Then for any $\epsilon > 0$, there is a network*

$$N(\mathbf{x}) = \sum_{i=0}^{v} c_i \sigma\left(\mathbf{w}_i \cdot \mathbf{x} + \theta_i\right), \quad \mathbf{x} \in \mathbb{R}^s, \tag{58}$$

where $c_i \in \mathbb{R}, \mathbf{w}_i \in \mathbb{R}^s$, and $\theta_i \in \mathbb{R}, 0 \leq i \leq v$, such that

$$\left\| D^{\mathbf{k}} f - D^{\mathbf{k}} N \right\|_{L^\infty(K)} < \epsilon, \quad \mathbf{k} \in \mathbb{N}_0^s, \mathbf{k} \leq \mathbf{m}_i, \textit{ for some } i, 1 \leq i \leq q \tag{59}$$

The following two lemmas state that any arbitrary increasing function can be approximated using Leaky-ReLU and Leaky-ReLU-like activation functions, respectively.

Lemma C.2 (Increasing Functions to Leaky-ReLU)**.** *For any increasing C^1 function f,*

$$f \in \overline{\Delta_{1,1,1}^{LR}}^{\text{loc}}. \tag{60}$$

The proof of Lemma C.2 is provided in Appendix C.2.

Lemma C.3. *Let $\Sigma = \{\sigma_\beta \mid \beta \in \mathbb{R}_+\}$ be a set of Leaky-ReLU-like activation functions. Then, for any increasing C^1 function f,*

$$f \in \overline{\Delta_{1,1,1}^{\Sigma}}^{\text{loc}}. \tag{61}$$

The proof of Lemma C.3 is provided in Appendix C.3. The two lemmas yield the following corollary.

Corollary C.4 (Generalization of Activation)**.** *For a natural number $d \in \mathbb{N}$ and any increasing, C^1 activation function ρ, the following relation holds:*

$$\Delta_{d,d,d}^{\rho} \subset \overline{\Delta_{d,d,d}^{\sigma}}^{\text{loc}}, \tag{62}$$

where σ is the Leaky-ReLU or a set of Leaky-ReLU-like activation functions.

The following lemma is a technical result used to approximate a vector field with deep narrow MLPs.

Lemma C.5. *For $t, b \in \mathbb{R}$, and $w \in \mathbb{R}^d$, define $f_t : \mathbb{R}^d \to \mathbb{R}^d$ as:*

$$f_t : x = (x_1, \dots, x_d) \mapsto (x_1, \dots, x_{d-1}, x_d + t \tanh(w \cdot x + b)), \tag{63}$$

Let σ be the Leaky-ReLU or a set of Leaky-ReLU-like activation functions. Then, there exists a positive real number $\delta \in \mathbb{R}_+$ such that, for $|t| < \delta$, the following relation holds:

$$f_t \in \overline{\Delta_{d,d,d}^{\sigma}}^{\text{loc}}. \tag{64}$$

The proof of Lemma C.5 is provided in Appendix C.4. The following lemma states that any smooth diffeomorphism can be approximated by flows generated by (time-dependent) vector fields. The definition of a vector field is as follows:

Definition C.6 (Flow of a Vector Field)**.** *Let $f : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ be a function that is Lipschitz continuous with respect to x and a piecewise continuous with respect to t. For each $f \in \mathcal{A}$, consider a ODE system*

$$\dot{x}(t) = f(x(t), t), \tag{65}$$

where $x : \mathbb{R} \to \mathbb{R}^d$. We define a flow map $\phi_f^{t,s} : \mathbb{R}^d \to \mathbb{R}^d$, corresponding to f as follows:

$$\phi_f^{t,s} : x(t) \mapsto x(t+s). \tag{66}$$

For $t = 0$, we omit t and just denote it as ϕ_f^s:

$$\phi_f^s = \phi_f^{0,s} \tag{67}$$

We define the maximal domain $\mathcal{M}_f \subset \mathbb{R}^d \times \mathbb{R}$ as the set which satisfies $(x, t) \in \mathcal{M}_f$ if and only if the solution $\phi_f^t(x)$ is well-defined. It is well known that $\mathcal{M}_f$ is an open set. It is also well known that if f is a C^k-function with respect to x and t, then $\phi_f^t(x)$ is also C^k-function with respect to x and t. (See Theorem B.41 of Biagi & Bonfiglioli (2019) for example.)

When we consider Df, we only consider a Jacobian with respect to x:

$$Df(x, t) := D_x f(x, t). \tag{68}$$

Lemma C.7 (Theorem 5 of Caponigro (2011))**.** *Any orientation preserving diffeomorphism can be represented by a flow map: For any diffeomorphism $f \in \mathcal{D}^\infty(\mathbb{R}^d)$ with $det(Df) > 0$, there exists a flow map ϕ_F^t generated by a ODE system $\dot{x} = F(x, t)$ with a smooth vector field $F : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ such that the following equation holds:*

$$f = \phi_F^1. \tag{69}$$

If two vector fields are close, then the flows they generate are also close.

Lemma C.8. *Consider $C^{1,1}$ functions $f_1, f_2 : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$. Define two ODE systems $\dot{x} = f_i(x,t)$ for $i = 1, 2$ and let $\phi_i^t := \phi_{f_i}^t$ be a flow map defined by each f_i. Assume that $\overline{V} \times [0,\tau] \subset \mathcal{M}_{f_1}$ for a precompact set $V \Subset \mathbb{R}^d$ and $\tau \in \mathbb{R}_+$. Define $\widetilde{V}$ as*

$$\widetilde{V} := \left\{ \phi_1^t(x) \in \mathbb{R}^d \mid t \in [0,\tau], x \in V \right\} + B_d(1). \tag{70}$$

Then, for any $\epsilon \in \mathbb{R}_+$, there exists a positive number $\delta \in (0,1)$ such that if

$$\|f_1(\cdot,t) - f_2(\cdot,t)\|_{W^{1,\infty}(\widetilde{V};\mathbb{R}^d)} < \delta, \tag{71}$$

for all $t \in [0,\tau]$, then,

$$\|\phi_1^\tau - \phi_2^\tau\|_{W^{1,\infty}(V;\mathbb{R}^d)} < \epsilon. \tag{72}$$

The proof of Lemma C.8 is provided in Appendix C.5.

Now, we approximate a two-layered-MLP-like vector field using deep narrow MLPs.

Lemma C.9. *For $i \in [1,N]_{\mathbb{N}}$ and $C^{1,1}$-functions $v_i : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$, let*

$$v(x,t) := \sum_{i=1}^{N} v_i(x,t). \tag{73}$$

For a real number $\tau \in \mathbb{R}_+$ and a precompact set $U \Subset \mathbb{R}^d$, assume that $\overline{U} \times [0,\tau] \subset \mathcal{M}_f$.

Consider $n \in \mathbb{N}$, $t_k := \frac{k\tau}{n}$, $\Delta t := \frac{\tau}{n}$,

$$f_{i,k} : x \mapsto x + \Delta t v_i(x, t_{k-1}), \tag{74}$$

$$T_k := f_{N,k} \circ f_{N-1,k} \circ \cdots \circ f_{1,k}, \tag{75}$$

and

$$S_k := T_k \circ \cdots \circ T_1. \tag{76}$$

Then, there exists a natural number $n_0 \in \mathbb{N}$ such that if $n \geq n_0$, the following relation holds:

$$\|\phi_v^\tau - S_n\|_{W^{1,\infty}(V;\mathbb{R}^d)} < \epsilon. \tag{77}$$

The proof of Lemma C.9 is provided in Appendix C.6.

By combining all the lemmas, we prove the theorem.

Proof of Theorem 4.1. By Theorem 2.7, p.50 in Hirsch (2012), we only have to consider $\mathcal{D}^\infty(\mathbb{R}^d)$. If f is an orientation reversing diffeomorphism, $g \circ f$ is orientation preserving where $g \in \Delta_{d,d,d}^{\mathrm{LR}}$ is defined as:

$$g : (x_1, \ldots, x_d) \mapsto (-x_1, x_2, x_3, \ldots, x_d). \tag{78}$$

Therefore, we only consider an orientation preserving diffeomorphism $f \in \mathcal{D}^\infty(\mathbb{R}^d)$. Consider an arbitrary precompact set $V \Subset \mathbb{R}^d$ and $\epsilon \in \mathbb{R}_+$. By Lemma C.7, there exists an ODE flow induced by $F \in C^\infty(\mathbb{R}^d \times \mathbb{R}; \mathbb{R}^d)$ such that

$$f = \phi_F^1. \tag{79}$$

By Lemma C.8, there exists $\delta \in \mathbb{R}_+$ and $\widetilde{V} \Subset \mathbb{R}^d$ such that if $\|F(\cdot,t) - F_2(\cdot,t)\|_{W^{1,\infty}(\widetilde{V};\mathbb{R}^d)} < \delta$ for all $t \in [0,1]$, then, $\|\phi_F^1 - \phi_{F_2}^1\|_{W^{1,\infty}(V;\mathbb{R}^d)} < \epsilon$. Consider a compact set K such that $\widetilde{V} \subset K$. By Lemma C.1, there exists a $F_2 : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ such that

$$\|F - F_2\|_{L^\infty(K\times[0,\tau])} + \|DF - DF_2\|_{L^\infty(K\times[0,\tau])} < \delta. \tag{80}$$

where F_2 is represented as

$$F_2 := \sum_{i=1}^{N} c_i \rho(w_i \cdot x + a_i t + b_i), \tag{81}$$

where $a_i, b_i, c_i \in \mathbb{R}$, and $w_i \in \mathbb{R}^d$. Here, $\rho = tanh$ if σ is Leaky-ReLU or Leaky-ReLU-like, and ρ equals to σ if σ is an activation function satisfying Condition 1. As both F and F_2 are $C^{1,1}$ functions,

$$\|F(\cdot,t) - F_2(\cdot,t)\|_{W^{1,\infty}(\widetilde{V};\mathbb{R}^d)} < \delta, \tag{82}$$

for all $t \in [0, \tau]$. Thus, $\|\phi_F^1 - \phi_{F_2}^1\|_{W^{1,\infty}(V;\mathbb{R}^d)} < \frac{\epsilon}{2}$. Then, by Lemma C.9, there exists a natural number $n_0 \in \mathbb{N}$ so that if $n \geq n_0$, then, for $t_k := \frac{k\tau}{n}$, $\Delta t = \frac{\tau}{n}$,

$$f_{i,k} : x \mapsto x + \Delta t c_i \rho(w_i \cdot x + a_i t_k + b_i), \tag{83}$$

$$T_k := f_{n,k} \circ f_{n-1,k} \circ \cdots \circ f_{1,k}, \tag{84}$$

and

$$S_k := T_k \circ \cdots \circ T_1, \tag{85}$$

the following inequality holds:

$$\left\|\phi_{F_2}^1 - S_n\right\|_{W^{1,\infty}(V;\mathbb{R}^d)} < \frac{\epsilon}{2}. \tag{86}$$

For the Leaky-ReLU or Leaky-ReLU like σ, by Lemma C.5, there exists $i \in [1, N]_{\mathbb{N}}$, $k \in [1, n]_{\mathbb{N}}$, and $\delta_{i,k} \in \mathbb{R}_+$ such that if $|t| < \delta_i$, then, $f_{i,k} \in \overline{\Delta^{\sigma}_{d,d,d}}^{\text{loc}}$. Choose sufficiently large n so that $|\Delta t c_i| < \delta_i$ for all , each $f_{i,k} \in \overline{\Delta^{\sigma}_{d,d,d}}^{\text{loc}}$. Then, $S_n \in \overline{\Delta^{\sigma}_{d,d,d}}^{\text{loc}}$. For σ satisfying Condition 1, $f_{i,k} \in \overline{\Delta^{\sigma}_{d,d,d+1}}^{\text{loc}}$, thus, $S_n \in \overline{\Delta^{\sigma}_{d,d,d+1}}^{\text{loc}}$. Thus, $\|\phi_F^1 - S_n\|_{W^{1,\infty}(V;\mathbb{R}^d)} < \epsilon$ for $S_n \in \overline{\Delta^{\sigma}_{d,d,d+\alpha(\sigma)}}^{\text{loc}}$. This completes the proof. □

C.2 Proof of Lemma C.2

Proof. $\Delta^{\text{LR}}_{1,1,1}$ is the set of strictly increasing piecewise linear functions with finite segments. Consider any increasing C^1 function $\sigma : \mathbb{R} \to \mathbb{R}$, a compact set $K \subset \mathbb{R}$, and a positive real number $\epsilon \in \mathbb{R}_+$. We will construct a function $f \in U$ such that $\|\sigma - f\|_{W^{1,\infty}} < \epsilon$. Consider a closed interval $[a, b] \supset K$. Then, there exists a natural number $n \in \mathbb{N}$ such that $\|f(x) - f(y)\| < \frac{\epsilon}{4}$ and $\|Df(x) - Df(y)\| < \frac{\epsilon}{4}$ for $\|x - y\| < \frac{1}{n}$. Define $f \in U$ as a piecewise linear function with breaking points $x = a + (b-a)i/n$ for $0 \leq i \leq n$, which has the same values with f in each breaking point. For all $x \in K$ and the closest breaking point $y \in [a, b]$, $|x - y| < \epsilon$. Then,

$$|\sigma(x) - f(x)| < |\sigma(x) - \sigma(y)| + |\sigma(y) - f(y)| + |f(y) - f(x)| < \frac{\epsilon}{2}. \tag{87}$$

And for two adjacent breaking points y_0, y_1 such that $x \in [y_0, y_1]$,

$$|Df(x) - D\sigma(x)| = \left|\frac{f(y_1) - f(y_0)}{y_1 - y_0} - D\sigma(x)\right| = |Df(c) - Df(x)| < \frac{\epsilon}{4} \tag{88}$$

for a $c \in (y_0, y_1)$ by mean value theorem, almost everywhere. Therefore, $\|\sigma(x) - f(x)\|_{W^{1,\infty}(K)} < \epsilon$. Because the selection of a compact set $K \subset \mathbb{R}$ is arbitrary, $\sigma \in \overline{\Delta^{\text{LR}}_{1,1,1}}^{\text{loc}}$, and this completes the proof. □

C.3 Proof of Lemma C.3

Proof. As strictly increasing C^1 functions are dense in the set of increasing C^1 functions in C^1 topology, we only need to approximate a strictly increasing C^1 function f. Consider an arbitrarily small error $\epsilon \in \mathbb{R}_+$ and an open interval (a, b). It is sufficient to prove that there exists $g \in \overline{\Delta^{\sigma}_{1,1,1}}^{\text{loc}}$ such that

$$\|f - g\|_{W^{1,\infty}((a,b);\mathbb{R})} < \epsilon. \tag{89}$$

Define $L_1, L_2 \in \mathbb{R}_+$ as uniquely determined value as follows:

$$[L_1, L_2] = \{Df(x) \in \mathbb{R}_+ \mid x \in [a, b]\}. \tag{90}$$

Define $b : \mathbb{R}_+ \to \mathbb{R}$ as

$$b(\beta) := \sup_x \|D\sigma_\beta(x) - 1\| \xrightarrow{\beta \to 1} 0 \tag{91}$$

Choose a sufficiently small $\epsilon' \in \mathbb{R}_+$ so that $(6L_2 + 4)b(1+\epsilon') + 2\epsilon' < \epsilon$. There exists a natural number $N \in \mathbb{N}$ such that if $\|x - y\| < \frac{1}{N}$, then, $\|f(x) - f(y)\| < \frac{\epsilon}{4}$ and $\|Df(x) - Df(y)\| < \min\left(\frac{\epsilon}{4}, \epsilon'\right)$. Define h as a piecewise linear function with breaking points $\alpha_i = a + (b-a)i/N$ for $0 \leq i \leq N$, which has the same values with σ in each breaking point. Then,

$$\|f - h\|_{W^{1,\infty}((a,b);\mathbb{R})} < \epsilon. \tag{92}$$

Now, it is sufficient to prove that there exists a function $h' \in \Delta^{\sigma}_{1,1,1}$ such that

$$\|h - h'\|_{W^{1,\infty}((a,b);\mathbb{R})} < \epsilon. \tag{93}$$

Define γ_i as

$$\gamma_i := \frac{f(\alpha_{i+1}) - f(\alpha_i)}{\alpha_{i+1} - \alpha_i}, \tag{94}$$

so that γ_i be the slope of h in (α_i, α_{i+1}). We use mathematical induction on n to prove the following: There exists a $f_{n,m} \in \Delta^{\sigma}_{1,1,1}$ such that

1. $\|h - f_{n,m}\|_{L^\infty((a,\alpha_{n+1}))} \xrightarrow{m\to\infty} 0$,
2. there exists a natural number M such that if $m \geq M$, then, $\|Dh - Df_{n,m}\|_{L^\infty((a,\alpha_{n+1}))} < \frac{\epsilon}{2}$,
3. and, for an arbitrary $\delta \in \left(0, \frac{1}{N}\right)$, $\|f_{n,m} - (\gamma_n(x - \alpha_{n+1}) + f(\alpha_{n+1}))\|_{W^{1,\infty}((\alpha_n+\delta,b);\mathbb{R})} \xrightarrow{m\to\infty} 0$

For $n = 0$, there is nothing to prove. Assume that the induction hypothesis is satisfied for n. Define $f_{n+1,m} \in \Delta^{\sigma}_{1,1,1}$ as

$$f_{n+1,m}(x) := \frac{\gamma_{n+1}}{\gamma_n}\frac{\sigma_{\frac{\gamma_n}{\gamma_{n+1}}}\left(m(f_{n,m}(x) - f(\alpha_{n+1}))\right)}{m} + f(\alpha_{n+1}). \tag{95}$$

As $\frac{\sigma_\beta(mx)}{m} \xrightarrow{m\to\infty} \mathrm{LR}_\beta(x)$ in C^0-topology,

$$\begin{aligned} f_{n+1,m} &\xrightarrow{m\to\infty} \frac{\gamma_{n+1}}{\gamma_n}\frac{\mathrm{LR}_{\frac{\gamma_n}{\gamma_{n+1}}}\left(h - f(\alpha_{n+1}))\right)}{m} + f(\alpha_{n+1}) \\ &= \begin{cases} h & \text{in } (a, \alpha_{n+1}) \\ \gamma_{n+1}(x - \alpha_{n+1}) + f(\alpha_{n+1}) & \text{in } (\alpha_{n+1}, b) = \gamma_{n+1}(x - \alpha_{n+2}) + f(\alpha_{n+2}) \end{cases}, \end{aligned} \tag{96}$$

with C^0-topology. Now, it is sufficient to prove that the derivate-related assumptions. $Df_{n+1,m}$ can be calculated as

$$Df_{n+1,m}(x) = \frac{\gamma_{n+1}}{\gamma_n} D\sigma_{\frac{\gamma_n}{\gamma_{n+1}}}\left(m(f_{n,m}(x) - f(\alpha_{n+1}))\right)Df_{n,m}(x) \tag{97}$$

Then, for any $\delta \in \mathbb{R}_+$ and $x \in [a, \alpha_{n+1} - \delta]$,

$$\begin{aligned} &\sup_{x\in[a,\alpha_{n+1}-\delta]} \|Df_{n+1,m}(x) - Df_{n,m}(x)\| \\ &\leq \sup_{x\in[a,\alpha_{n+1}-\delta]} \|Df_{n,m}(x)\| \left\|1 - \frac{\gamma_{n+1}}{\gamma_n}\sigma_{\frac{\gamma_n}{\gamma_{n+1}}}(m(f_{n,m}(x) - f(\alpha_{n+1})))\right\| \xrightarrow{m\to\infty} 0. \end{aligned} \tag{98}$$

And, for $x \in [\alpha_{n+1} + \delta, b]$, $\lim_{m\to\infty} \sup_{x\in[\alpha_{n+1}+\delta,b]} \|Df_{n,m}(x) - \gamma_n\| = 0$, and therefore,

$$\begin{aligned} &\lim_{m\to\infty} \sup_{x\in[\alpha_{n+1}+\delta,b]} \|Df_{n+1,m}(x) - \gamma_{n+1}\| \\ &= \lim_{m\to\infty} \sup_{x\in[\alpha_{n+1}+\delta,b]} \gamma_{n+1} \left\|D\sigma_{\frac{\gamma_n}{\gamma_{n+1}}}\left(m(f_{n,m}(x) - f(\alpha_{n+1}))\right)\right\| = \gamma_{n+1}. \end{aligned} \tag{99}$$

Therefore, the induction hypothesis 3 is satisfied. As $h(x) = \gamma_{n+1}(x - \alpha_{n+2}) + f(\alpha_{n+2})$ for $x \in [\alpha_{n+1}, \alpha_{n+2}]$,

$$\|Dh - Df_{n+1,m}\|_{L^\infty((\alpha_{n+1}+\delta,\alpha_{n+2}))} \xrightarrow{m\to\infty} 0. \tag{100}$$

Now, it remains to prove that there exists a natural number M' such that if $m \geq M'$, then $\|Dh - Df_{n+1,m}\|_{L^\infty((\alpha_{n+1}-\delta,\alpha_{n+1}+\delta))} < \epsilon$. Choose sufficiently large M' so that if $m \geq M'$, then,

$$\sup_{x\in(\alpha_{n+1}-\delta,\alpha_{n+1}+\delta)} \|Df_{n,m}(x) - \gamma_i\| \leq \min(\epsilon, 1). \tag{101}$$

Then, for $x \in (\alpha_{n+1} - \delta, \alpha_{n+1} + \delta)$,

$$\begin{aligned}
&\|\gamma_{i+1} - Df_{n+1,m}(x)\| = \left\| \gamma_{i+1} - \frac{\gamma_{i+1}}{\gamma_i} D\sigma_{\frac{\gamma_i}{\gamma_{i+1}}} \left(m(f_{n,m}(x) - f(\alpha_{n+1})) \right) Df_{n,m}(x) \right\| \\
&\leq \frac{\gamma_{i+1}}{\gamma_i} \|Df_{n,m}(x)\| \left\| D\sigma_{\frac{\gamma_i}{\gamma_{i+1}}} \left(m(f_{n,m}(x) - f(\alpha_{n+1})) \right) - 1 \right\| + \left\| \gamma_{i+1} - \frac{\gamma_{i+1}}{\gamma_i} Df_{n,m}(x) \right\| \\
&\leq \frac{\gamma_{i+1}}{\gamma_i} \|Df_{n,m}(x)\| b\left(\frac{\gamma_i}{\gamma_{i+1}} \right) + \left\| \gamma_{i+1} - \frac{\gamma_{i+1}}{\gamma_i} Df_{n,m}(x) \right\| \\
&\leq (1+1)\left(L_2+1\right) b(1+\epsilon') + L_2\epsilon' \leq (3L_2+2)b(1+\epsilon') < \frac{\epsilon}{2},
\end{aligned} \tag{102}$$

and

$$\|\gamma_i - Df_{n+1,m}(x)\| \leq \|\gamma_{i+1} - Df_{n+1,m}(x)\| + \|\gamma_{i+1} - \gamma_i\| \leq (3L_2+2)b(1+\epsilon') + \epsilon' < \frac{\epsilon}{2}. \tag{103}$$

Therefore, for $x \in (\alpha_{n+1} - \delta, \alpha_{n+1} + \delta)$,

$$\|Dh(x) - Df_{n+1,m}(x)\| < \frac{\epsilon}{2}. \tag{104}$$

By mathematical induction, we conclude that there exists $f_{N,m} \in \Delta^{\sigma}_{1,1,1}$ and $M \in \mathbb{N}$ such that if $m \geq M$, then,

$$\|h - f_{N,m}\|_{W^{1,\infty}((a,b);\mathbb{R})} < \epsilon. \tag{105}$$

This completes the proof. □

C.4 Proof of Lemma C.5

Proof. For $w = (w_1, \dots, w_d)$, if $w_i = 0$ for $i \in [1, d-1]_{\mathbb{N}}$, the last term of f_t can be calculated as:

$$x_d + t\tanh(w_d x_d + b). \tag{106}$$

For sufficiently small $\delta \in \mathbb{R}_+$ and $|t| < \delta$, this function is increasing. Thus, by Corollary C.4, the following relations holds:

$$\Delta^{\{x \mapsto x + t\tanh(wx+b), \mathrm{id}\}}_{d,d,d} \subset \overline{\Delta^{\sigma}_{d,d,d}}^{\mathrm{loc}}, \tag{107}$$

Also, the following relation holds:

$$\Delta^{\{\tanh, \mathrm{id}\}}_{d,d,d}, \Delta^{\{\tanh^{-1}, \mathrm{id}\}}_{d,d,d} \subset \overline{\Delta^{\sigma}_{d,d,d}}^{\mathrm{loc}}. \tag{108}$$

Now assume that there exists $i \in [1, d-1]_{\mathbb{N}}$ such that $w_i \neq 0$. Further, without loss of generality, assume that $w_1 \neq 0$. Then, the following functions are elements of $\overline{\Delta^{\sigma}_{d,d,d}}^{\mathrm{loc}}$

$$f_1 : (x_1, \dots, x_d) \mapsto (w \cdot x + b, x_2, \dots, x_d), \tag{109}$$

$$f_2 : (x_1, \dots, x_d) \mapsto (\tanh(x_1), x_2, \dots, x_d), \tag{110}$$

$$f_3 : (x_1, \dots, x_d) \mapsto (x_1, \dots, x_{d-1}, x_d + tx_1), \tag{111}$$

$$f_4 : (x_1, \dots, x_d) \mapsto (\tanh^{-1}(x_1), x_2, \dots, x_d), \tag{112}$$

$$f_5 : (x_1, \dots, x_d) \mapsto (x_1 + w_d t\tanh(x_1), x_2, \dots, x_d), \tag{113}$$

and

$$f_6 : (x_1, \dots, x_d) \mapsto \left(\frac{x_1 - w_d x_d - w_{2:d-1} \cdot x_{2:d-1} - b}{w_1}, x_2, \dots, x_d \right). \tag{114}$$

Then, the composition $f_6 \circ f_5 \circ f_4 \circ f_3 \circ f_2 \circ f_1$ becomes

$$f_6 \circ f_5 \circ f_4 \circ f_3 \circ f_2 \circ f_1 : x \mapsto (x_1, \dots, x_{d-1}, x_d + t\tanh(w \cdot x + b)). \tag{115}$$

□

C.5 Proof of Lemma C.8

Proof. Let L and L' be a Lipschitz constant of f_1 and Df_1 with respect to x in $\widetilde{V}$, respectively. Restrict δ to

$$\delta < \min\left(1, \frac{1}{2\tau e^{L\tau}}\right). \tag{116}$$

We first prove that, for all $x \in V$ and $t \in [0, \tau]$, $\phi_2^t \in \widetilde{V}$. Define T as

$$T := \left\{t \in [0, \tau] \mid \overline{V} \times \{t\} \subset \mathcal{M}_{f_2} \text{ and } \phi_2^t(x) \in \widetilde{V} \text{ for all } x \in \overline{V}\right\}. \tag{117}$$

1. Obviously, $0 \in T$.
2. And T is an open set relative to $[0, \tau]$: Assume that $t \in T$. Then, as $\phi_2^t(x) \in \widetilde{V}$ for all $x \in \overline{V}$, and $\mathcal{M}_{f_2}$ and $\widetilde{V}$ is open, there exist $\epsilon_{1,x}, \epsilon_{2,x} \in \mathbb{R}_+$ such that if $\|y - x\| < \epsilon_{1,x}$ and $|t' - t| \leq \epsilon_{2,x}$, then, $\phi_2^{t'}(y) \in \widetilde{V}$. Because $\overline{V}$ is compact, we can choose a finite cover $\{\{x\} + B_d(\epsilon_{1,x})\}_{x \in S}$ of $\overline{V}$. Then, $[t, t + \min_{x \in S} \epsilon_{2,x}) \subset T$, and T becomes an open set.
3. T is closed relative to $[0, \tau]$: Assume that $T = [0, t)$ for $t \in (0, \tau]$. It is sufficient to prove that

   $$\phi_2^t(x) = x + \int_0^t f_2(\phi_2^s(x), s) ds \tag{118}$$

   is finite and in $\widetilde{V}$. Define $e(x, t)$ as

   $$e(x, t) := \phi_1^t(x) - \phi_2^t(x). \tag{119}$$

   Then, the following equation holds:

   $$e(x, t) = \int_0^t f_1(\phi_1^s(x), s) - f_2(\phi_2^s(x), s) ds. \tag{120}$$

   Then, as $\phi_1^s(x), \phi_1^s(x) \in \widetilde{V}$ for $s \in [0, t)$, the following inequalities hold:

   $$\begin{aligned} \|e(x,t)\| &\leq \int_0^t \|f_1(\phi_1^s(x), s) - f_1(\phi_2^s(x), s)\| ds + \int_0^t \|f_1(\phi_2^s(x), s) - f_2(\phi_2^s(x), s)\| ds \\ &\leq \int_0^t \|Le(x, s)\| + \delta ds \leq \delta t + L \int_0^t \|e(x, s)\| ds \leq \delta t e^{Lt}, \end{aligned} \tag{121}$$

   where the last inequality is by Gronwall's inequality. As $\delta t e^{Lt} < \delta \tau e^{L\tau} < 1$,

   $$\phi_2^t(x) = \phi_1^t(x) + e(x, t) \in \widetilde{V}, \tag{122}$$

   for all $x \in V$, which leads to $t \in T$.
4. We conclude that $T = [0, \tau]$.

Next, we prove that $\|e(x, t)\|$ can be bounded. It is already proven by setting $\delta < \frac{\epsilon}{2\tau e^{L\tau}}$. Then, $\|e(x, t)\| < \frac{\epsilon}{2}$.

Finally, we will prove that $\|De(x, t)\|$ can be bounded.

$$De(x, t) = \int_0^t Df_1(\phi_1^s(x), s) D\phi_1^s(x) - Df_2(\phi_2^s(x), s) D\phi_2^s(x) ds. \tag{123}$$

Then,

$$
\begin{aligned}
\|De(x,t)\| \le \int_0^t \|Df_1(\phi_1^s(x),s)D\phi_1^s(x) - Df_1(\phi_2^s(x),s)D\phi_1^s(x)\|\,ds \\
+ \int_0^t \|Df_1(\phi_2^s(x),s)D\phi_1^s(x) - Df_2(\phi_2^s(x),s)D\phi_1^s(x)\|\,ds \\
+ \int_0^t \|Df_2(\phi_2^s(x),s)D\phi_1^s(x) - Df_2(\phi_2^s(x),s)D\phi_2^s(x)\|\,ds \\
\le \int_0^t L^2\|e(x,s)\| + \delta L + L\|De(x,t)\|ds \le \int_0^t LL'\delta s e^{Ls} + \delta L + LL'\|De(x,t)\|ds \\
\le \delta e^{Lt}(Lt-1) + \delta + \delta Lt + LL'\int_0^t \|De(x,t)\|ds \le \delta\left(e^{Lt}(Lt-1)+1+Lt\right)e^{LL't}, \quad (124)
\end{aligned}
$$

where the last inequality is by Gronwall's inequality again. By setting sufficiently small δ, we get

$$\|e(x,t)\| + \|De(x,t)\| < \epsilon, \tag{125}$$

for all $x \in V$ and $t \in [0,\tau]$. This completes the proof. □

C.6 Proof of Lemma C.9

Proof. Define $V^0 \subset \mathbb{R}^d \times \mathbb{R}$ and $V_t^0 \subset \mathbb{R}^d$ as

$$V_t^0 := \{\phi_v^t(x) \mid x \in U\}. \tag{126}$$

and

$$V^0 := \left\{(x,t) \in \mathbb{R}^d \times \mathbb{R} \mid x \in V_t^0\right\}. \tag{127}$$

As $\overline{V^0}$ is compact, there exists a positive number $\delta \in \mathbb{R}_+$ such that

$$V_t := \left(V_t^0 + B_d(\delta)\right) \times [0, \tau - t] \Subset \mathcal{M}_f, \tag{128}$$

for all $t \in [0,\tau]$. Define $V \subset \mathbb{R}^d \times \mathbb{R}$ as

$$V := \left\{(x,t) \in \mathbb{R}^d \times \mathbb{R} \mid x \in V_t\right\}. \tag{129}$$

We will conduct all our discussions on V where ϕ_v^t is well-defined. Denote the supremum and the Lipschitz constant of v in $\overline{V}$ as C and L, respectively. Also, denote the supremum and the Lipschitz constant (as operator norm) of Dv in $\overline{V}$ as C' and L'.

In this proof, we will use a big-O notation with respect to Δt; that is, a function $f : \mathbb{R} \to \mathbb{R}$ is denoted as

$$f = O\left(\Delta t^i\right), \tag{130}$$

if and only if

$$|f(\Delta t)| < c\Delta t^i, \tag{131}$$

where c is a constant independent of Δt and polynomially dependent on N, L, C, L', C'.

We will check that

$$\left\|\phi_v^{t_k,t_{k+1}} - T_{k+1}\right\|_{L^\infty\left(V_{t_k};\mathbb{R}^d\right)} = O\left(\Delta t^2\right). \tag{132}$$

We define $U_{l,k} \in C^1(\mathbb{R}^d;\mathbb{R}^d)$ as

$$U_{l,k} : x \mapsto x + \Delta t \sum_{i=1}^{l} v_i(x, t_{k-1}). \tag{133}$$

And define U_k as

$$U_k := U_{N,k}. \tag{134}$$

Then, it is sufficient to bound two terms:

$$\left\|\phi_v^{t_k,t_{k+1}} - U_{k+1}\right\|_{L^\infty\left(V_{t_k};\mathbb{R}^d\right)} \quad \text{and} \quad \|T_{k+1} - U_{k+1}\|_{L^\infty\left(V_{t_k};\mathbb{R}^d\right)}. \tag{135}$$

The first term can be calculated as

$$\left\|\phi_v^{t_k,t_{k+1}} - U_{k+1}\right\|_{L^\infty\left(V_{t_k};\mathbb{R}^d\right)} = \left\|\int_{t_k}^{t_{k+1}} v\left(\phi_v^{t_k,s}, s\right) ds - \Delta t v(\cdot, t_k)\right\|_{L^\infty\left(V_{t_k};\mathbb{R}^d\right)}$$
$$= \left\|\int_{t_k}^{t_{k+1}} v\left(\phi_v^{t_k,s}, s\right) - v(\cdot, t_k) ds\right\|_{L^\infty\left(V_{t_k};\mathbb{R}^d\right)} \leq \sup_{x\in V_{t_k}} \int_{t_k}^{t_{k+1}} \left\|v\left(\phi_v^{t_k,s}(x), s\right) - v(x, t_k)\right\| ds$$
$$\leq \Delta t \sup_{x\in V_{t_k}} \sup_{s\in[t_k,t_{k+1}]} \left\|v\left(\phi_v^{t_k,s}, s\right) - v(x, t_k)\right\| \leq C\left(\Delta t\right)^2 \left(e^{L\Delta t} + L\right), \quad (136)$$

where the last equality is by the following arguments: for any k and $x \in V_k$,

$$\|\phi_v^{t_k,s}(x) - x\| = \left\|\int_{t_k}^s v(\phi_v^{t_k,r}(x), r) dr\right\| = \left\|\int_{t_k}^s v(\phi_v^{t_k,r}(x), r) - v(x, r) + v(x, r) dr\right\|$$
$$\leq L \int_{t_k}^s \|\phi_v^{t_k,r}(x) - x\| dr + C(s - t_k) \leq C(s - t_k)e^{L(s-t_k)} \leq C\Delta t e^{L\Delta t}, \quad (137)$$

where the second last inequality is by Gronwall's inequality. Therefore,

$$\left\|v\left(\phi_v^{t_k,s}, s\right) - v(x, t_k)\right\| \leq CL\Delta t e^{L\Delta t} + L\Delta t, \quad (138)$$

and the bound is independent of k. To calculate the second term $\|T_{k+1} - U_{k+1}\|_{W^{1,\infty}(V_{t_k};\mathbb{R}^d)}$, for $l \in [1, N]_{\mathbb{N}}$, define $T_{l,k} \in C^1(\mathbb{R}^d; \mathbb{R}^d)$ as

$$T_{l,k} := f_{l,k} \circ f_{l-1,k} \circ \cdots \circ f_{1,k}. \quad (139)$$

Then, $T_{N,k} = T_k$. We inductively bound

$$\|T_{l,k} - U_{l,k}\|_{L^\infty(V_{t_{k-1}};\mathbb{R}^d)}. \quad (140)$$

When $l = 1$, $T_{1,k} = U_{1,k} = f_{1,k}$, and there is nothing to prove. Assume that the above induction hypothesis is satisfied for l. Then,

$$T_{l+1,k}(x) = f_{l+1,k} \circ T_{l,k}(x) = T_{l,k}(x) + \Delta t v_{l+1}(T_{l,k}(x), t_{k-1}). \quad (141)$$

Therefore,

$$\|T_{l+1,k} - U_{l+1,k}\|_{L^\infty(V_{t_{k-1}};\mathbb{R}^d)} \leq \sup_{x\in V_{t_{k-1}}} \|T_{l+1,k}(x) - U_{l+1,k}(x)\|$$
$$\leq \sup_{x\in V_{t_{k-1}}} \|T_{l,k}(x) + \Delta t v_{l+1}(T_{l,k}(x), t_{k-1}) - (U_{l,k}(x) + \Delta t v_{l+1}(x, t_{k-1}))\|$$
$$\leq \sup_{x\in V_{t_{k-1}}} \|T_{l,k}(x) - U_{l,k}(x)\| + \Delta t \|v_{l+1}(T_{l,k}(x), t_{k-1}) - v_{l+1}(x, t_{k-1})\|$$
$$\leq (\Delta t)^2 CNL. \quad (142)$$

Therefore,

$$\|T_k - U_k\|_{L^\infty(V_{t_{k-1}};\mathbb{R}^d)} \leq (\Delta t)^2 CN^2L. \quad (143)$$

And thus,

$$\left\|\phi_v^{t_k,t_{k+1}} - T_{k+1}\right\|_{L^\infty\left(V_{t_k};\mathbb{R}^d\right)} \leq (\Delta t)^2 \left(CN^2L + L + e^{L\Delta t}\right) =: c_1 (\Delta t)^2. \quad (144)$$

Now, define $e_k : \mathbb{R}^d \to \mathbb{R}^d$ as

$$e_k := \phi_v^{t_k} - S_k. \quad (145)$$

We restrict Δt sufficiently small so that

$$\frac{e^{L\tau} - 1}{L} c_1(\Delta t) < \min\left(\frac{\epsilon}{2}, \delta\right). \quad (146)$$

Under this assumption, we use the mathematical induction on k to prove that $S_k(x) \in V_{t_k}$ for an arbitrary $k \in [1,n]_{\mathbb{N}}$. It is obvious when $k=0$. Assume that the induction hypothesis is satisfied for $k=k_0$. For $x \in U$ and $k \le k_0$,

$$\begin{aligned}
\|e_{k+1}(x)\| &= \|\phi_v^{t_{k+1}}(x) - S_{k+1}(x)\| = \|\phi_v^{t_k,t_{k+1}} \circ \phi_v^{t_k}(x) - T_{k+1} \circ S_k(x)\| \\
&\le \|\phi_v^{t_k,t_{k+1}} \circ \phi_v^{t_k}(x) - \phi_v^{t_k,t_{k+1}} \circ S_k(x)\| + \|\phi_v^{t_k,t_{k+1}} \circ S_k(x) - T_{k+1} \circ S_k(x)\| \\
&\le \mathcal{L}_{V_{t_k}}(\phi_v^{t_k,t_{k+1}})\|e_k(x)\| + \|\phi_v^{t_k,t_{k+1}} - T_{k+1}\|_{L^\infty(V_{t_k};\mathbb{R}^d)} < (1+L\Delta t)\|e_k(x)\| + \Delta t c_1(\Delta t).
\end{aligned} \tag{147}$$

Then, for any $k \le \frac{\tau}{\Delta t}$,

$$\|e_k(x)\| = (1+L\Delta t)^k \|e_0(x)\| + \frac{(1+L\Delta t)^k - 1}{L\Delta t}\Delta t c_1(\Delta t) \le \frac{e^{L\tau}-1}{L} c_1(\Delta t) < \frac{\epsilon}{2}. \tag{148}$$

As $S_{k+1}(x) = \phi_v^{t_{k+1}}(x) + e_{k+1}(x) \in V_{t_{k+1}}$, the induction hypothesis is satisfied. Also, $\|e_k\|_{L^\infty(U;\mathbb{R}^d)} < \frac{\epsilon}{2}$.

Now, we bound $D\phi_v^t(x)$. First, we bound a derivative $D\left(\phi_v^{t_k,t_{k+1}} - I_d\right)$. For arbitrary $s,t \in [0,\tau]$ and $x \in V_s$, consider the following equation.

$$\phi_v^{s,t}(x) - x = \int_s^t v(\phi_v^{s,r}(x), r)dr. \tag{149}$$

Apply derivative to both sides, and we get

$$\begin{aligned}
\|D\phi_v^{s,t}(x) - I_d\| &= \left\|\int_s^t Dv(\phi_v^{s,r}(x),r)D\phi_v^{s,r}(x)dr\right\| \le \int_s^t L'\left\|D\phi_v^{s,r}(x)\right\|dr \\
&\le \int_s^t L'\left\|D\phi_v^{s,r}(x) - I_d\right\| + dL'dr \le dL'(t-s)e^{L't} \le dL'(t-s)e^{L'\tau},
\end{aligned} \tag{150}$$

where the last inequality is by the Gronwall's inequality. Denote the last constant as $L'_1 := dL'e^{L'\tau}$; that is,

$$\|D\phi_v^{s,t}(x) - I_d\| \le L_1'(t-s). \tag{151}$$

Calculate the Lipschitz constant of $D\phi_v^{t_k,t_{k+1}}$. For $s,t \in [t_k,t_{k+1}]$, and $x,y \in V_{t_k}$.

$$\begin{aligned}
\left\|D\phi_v^{s,t}(x) - D\phi_v^{s,t}(y)\right\| &\le \left\|\int_s^t Dv(\phi_v^{s,r}(x),r)D\phi_v^{s,r}(x)dr - \int_s^t Dv(\phi_v^{s,r}(y),r)D\phi_v^{s,r}(y)dr\right\| \\
&\le \left\|\int_s^t Dv(\phi_v^{s,r}(x),r)D\phi_v^{s,r}(x)dr - Dv(\phi_v^{s,r}(x),r)D\phi_v^{s,r}(y)dr\right\| \\
&\quad + \left\|\int_s^t Dv(\phi_v^{s,r}(x),r)D\phi_v^{s,r}(y)dr - Dv(\phi_v^{s,r}(y),r)D\phi_v^{s,r}(y)dr\right\| \\
&\le (1+L_1'(t-s))\int_s^t \left\|D\phi_v^{s,r}(x) - D\phi_v^{s,r}(y)\right\|dr + 2L'\int_s^t \|\phi_v^{s,r}(x) - \phi_v^{s,r}(y)\|dr \\
&\le (1+L'(t-s))\int_s^t \left\|D\phi_v^{s,r}(x) - D\phi_v^{s,r}(y)\right\|dr + 2L'C'\Delta t\|x-y\| \\
&\le 2L'\Delta t\|x-y\|e^{1+L_1'(t-s)} \le 4L'C'\Delta t\|x-y\|e^2,
\end{aligned} \tag{152}$$

where the second last inequality is by Gronwall's inequality. Denote the last constant as $L'_2 := 4L'C'e^2$; that is,

$$\left\|D\phi_v^{s,t}(x) - D\phi_v^{s,t}(y)\right\| \le L_2'\Delta t\|x-y\|. \tag{153}$$

Now we calculate the followings:

$$\left\|D\phi_v^{t_k,t_{k+1}} - DT_{k+1}\right\|_{L^\infty\left(V_{t_k};\mathbb{R}^d\right)} = O\left(\Delta t\right)^2. \tag{154}$$

$$
\begin{aligned}
&\left\|D\phi_v^{t_k,t_{k+1}}(x) - DU_{k+1}(x)\right\| = \left\|\int_{t_k}^{t_{k+1}} Dv\left(\phi_v^{t_k,s}(x), s\right) D\phi_v^{t_k,s}(x) - Dv(x,t_k)ds\right\| \\
&\leq \int_{t_k}^{t_{k+1}} \left\|Dv\left(\phi_v^{t_k,s}(x), s\right) - Dv(x,t_k)\right\| ds + \int_{t_k}^{t_{k+1}} \left\|Dv\left(\phi_v^{t_k,s}(x), s\right)\left(D\phi_v^{t_k,s}(x) - I_d\right)\right\| ds \\
&= O\left(\Delta t^2\right). \qquad (155)
\end{aligned}
$$

We use the mathematical induction on l to prove that

$$
\|DT_{l,k} - DU_{l,k}\|_{L^\infty(V_{t_{k-1}};\mathbb{R}^d)} = O\left(\Delta t^2\right). \qquad (156)
$$

When $l = 1$, $T_{1,k} = U_{1,k} = f_{1,k}$, and there is nothing to prove. Assume that the above induction hypothesis is satisfied for l. Then,

$$
DT_{l+1,k}(x) = Df_{l+1,k}(T_{l,k}(x))DT_{l,k}(x) = (I_d + \Delta t Dv_{l+1}(T_{l,k}(x), t_{k-1}))DT_{l,k}(x) \qquad (157)
$$

Therefore,

$$
\begin{aligned}
&\|DT_{l+1,k} - DU_{l+1,k}\|_{L^\infty(V_{t_{k-1}};\mathbb{R}^d)} \leq \sup_{x\in V_{t_{k-1}}} \|DT_{l+1,k}(x) - DU_{l+1,k}(x)\| \\
&\leq \sup_{x\in V_{t_{k-1}}} \|(I_d + \Delta t Dv_{l+1}(T_{l,k}(x), t_{k-1}))DT_{l,k}(x) - (DU_{l,k}(x) + \Delta t Dv_{l+1}(x, t_{k-1})))\| \\
&\leq \sup_{x\in V_{t_{k-1}}} \|DT_{l,k} - DU_{l,k}\| + \Delta t\|Dv_{l+1}(T_{l,k}(x), t_{k-1})DT_{l,k}(x) - Dv_{l+1}(x, t_{k-1}))\| \\
&= O\left(\Delta t^2\right). \qquad (158)
\end{aligned}
$$

Therefore, the induction hypothesis is satisfied.

For any $x \in U$ and k, we have

$$
\begin{aligned}
\|De_{k+1}(x)\| = \|D\phi_v^{t_{k+1}}(x) - DS_{k+1}(x)\| &= \left\|D\phi_v^{t_k,t_{k+1}}(\phi_v^{t_k}(x))D\phi_v^{t_k}(x) - DT_{k+1}(S_k(x))DS_k(x)\right\| \\
&\leq \left\|D\phi_v^{t_k,t_{k+1}}(\phi_v^{t_k}(x))D\phi_v^{t_k}(x) - D\phi_v^{t_k,t_{k+1}}(\phi_v^{t_k}(x))DS_k(x)\right\| \\
&+ \left\|D\phi_v^{t_k,t_{k+1}}(\phi_v^{t_k}(x))DS_k(x) - D\phi_v^{t_k,t_{k+1}}(S_k(x))DS_k(x)\right\| \\
&+ \left\|D\phi_v^{t_k,t_{k+1}}(S_k(x))DS_k(x) - DT_{k+1}(S_k(x))DS_k(x)\right\|. \qquad (159)
\end{aligned}
$$

For the first term, we have

$$
\begin{aligned}
&\left\|D\phi_v^{t_k,t_{k+1}}(\phi_v^{t_k}(x))D\phi_v^{t_k}(x) - D\phi_v^{t_k,t_{k+1}}(\phi_v^{t_k}(x))DS_k(x)\right\| \\
&\leq \|D\phi_v^{t_k}(x) - DS_k(x)\| + \left\|\left(D\phi_v^{t_k,t_{k+1}}(\phi_v^{t_k}(x)) - I_d\right)\left(D\phi_v^{t_k}(x) - DS_k(x)\right)\right\| \\
&\leq (1 + L_1'\Delta t)\|De_k(x)\|. \qquad (160)
\end{aligned}
$$

For the second term, there exists a constant $c_2 \in \mathbb{R}_+$ satisfying

$$
\begin{aligned}
&\left\|D\phi_v^{t_k,t_{k+1}}(\phi_v^{t_k}(x))DS_k(x) - D\phi_v^{t_k,t_{k+1}}(S_k(x))DS_k(x)\right\| \\
&\leq \mathcal{L}_{V_{t_k}}\left(D\phi_v^{t_k,t_{k+1}}\right)\|DS_k(x)\|\|e_k(x)\| \leq L_2'\Delta t\|e_k(x)\|\|DS_k(x)\| \leq c_2\Delta t\|e_k(x)\|. \qquad (161)
\end{aligned}
$$

For the last term, we have

$$
\left\|D\phi_v^{t_k,t_{k+1}}(S_k(x))DS_k(x) - DT_{k+1}(S_k(x))DS_k(x)\right\| = O\left(\Delta t^2\right). \qquad (162)
$$

Then, by selecting a sufficiently small Δt, we have

$$
\begin{aligned}
&\|De_{k+1}(x)\| \leq (1 + L_1'\Delta t)\|De_k(x)\| + c_2\Delta t\|e_k(x)\| + O\left(\Delta t^2\right) \\
&\leq (1 + L_1'\Delta t)\|De_k(x)\| + c_3\Delta t^2 \leq \frac{e^{L_1'\tau - 1}}{L_1'}c_3(\Delta t) < \frac{\epsilon}{2}, \qquad (163)
\end{aligned}
$$

for a constant $c_3 \in \mathbb{R}_+$. We conclude that

$$
\|e_{k+1}(x)\| + \|De_{k+1}(x)\| < \epsilon, \qquad (164)
$$

and this completes the proof. □

D Proofs of Approximation Lemmas

D.1 Proof of Theorem 4.4

Proof. Consider a function $f \in C^1(\mathbb{R}^n;\mathbb{R}^m)$ and a precompact set $V \Subset \mathbb{R}^n$. It is sufficient to prove that, for any $\epsilon \in \mathbb{R}_+$, there exists a function $\widetilde{f} \in \Delta^\sigma_{n,m,\Omega(n,m)+\alpha(\sigma)}$ such that $\|f-\widetilde{f}\|_{W^{1,\infty}(V;\mathbb{R}^m)} < \epsilon$. Because $\Delta^\sigma_{n,m,\Omega(n,m)+\alpha(\sigma)}$ is closed under affine transformation composition, we only need to consider V satisfying $V \Subset (0,1)^n$. By the definition of $\Omega(n,m)$, for any $\epsilon \in \mathbb{R}_+$, there exists an embedding $g \in \mathrm{Emb}([0,1]^n, \mathbb{R}^{\Omega(n,m)})$ such that

$$\|f - p_{\Omega(n,m),n} \circ g\|_{C^1([0,1]^n;\mathbb{R}^m)} < \frac{\epsilon}{2}. \tag{165}$$

Because $\Omega(n,m) \geq n$, by Lemma 4.2, for $q_{n,\Omega(n,m)} : (x_1,\dots,x_n) \mapsto (x_1,\dots,x_n,0,\dots,0)$, there exists a smooth diffeomorphism G such that $g = G \circ q_{n,\Omega(n,m)}$. By Theorem 4.1, there exists an MLP $H \in \Delta^\sigma_{\Omega(n,m),\Omega(n,m),\Omega(n,m)+\alpha(\sigma)}$ such that

$$\|G - H\|_{W^{1,\infty}(V\times(0,1)^{\Omega(n,m)-n};\mathbb{R}^{\Omega(n,m)})} < \frac{\epsilon}{2}. \tag{166}$$

Then,

$$\|p_{\Omega(n,m),m} \circ H \circ q_{n,\Omega(n,m)} - p_{\Omega(n,m),m} \circ G \circ q_{n,\Omega(n,m)}\|_{W^{1,\infty}(V;\mathbb{R}^m)} < \frac{\epsilon}{2}. \tag{167}$$

Therefore,

$$\|f - p_{\Omega(n,m),m} \circ H \circ q_{n,\Omega(n,m)}\|_{W^{1,\infty}(V;\mathbb{R}^m)} < \epsilon. \tag{168}$$

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

D.2 Proof of Proposition 4.5

Proof. For a non-decreasing C^1 activation function σ, there exist smooth, strictly increasing activation functions σ_n that converge to σ in $W^{1,\infty}_{\mathrm{loc}}$ topology. Therefore, $\overline{\Delta^\sigma_{d,d,d}}^{\mathrm{loc}} \subset \overline{\Delta^{\{\sigma_n|n\in\mathbb{N}\}}_{d,d,d}}^{\mathrm{loc}}$, making it sufficient to consider only a smooth, strictly increasing activation function σ.

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

$$f = p_{\Omega(n,m)-1,m} \circ g \circ q_{n,\Omega(n,m)-1}, \tag{169}$$

where $g \in \Delta^\sigma_{\Omega(n,m)-1,\Omega(n,m)-1,\Omega(n,m)-1}$. As $\Delta^\sigma_{\Omega(n,m)-1,\Omega(n,m)-1,\Omega(n,m)-1} \subset \overline{\mathcal{D}^\infty(\mathbb{R}^{\Omega(n,m)-1})}^{\mathrm{loc}}$, $g \circ q_{n,\Omega(n,m)-1}\big|_{(0,1)^n} \in \overline{\mathrm{Emb}((0,1)^n,\mathbb{R}^{\Omega(n,m)-1})}^{\mathrm{loc}}$. Therefore, we have:

$$f|_{(0,1)^n} \in p_{\Omega(n,m)-1,m}\left(\overline{\mathrm{Emb}((0,1)^n,\mathbb{R}^{\Omega(n,m)-1})}^{\mathrm{loc}}\right), \tag{170}$$

and as the selection of $f \in \Delta^\sigma_{n,m,\Omega(n,m)-1}$ is arbitrary, we get the following:

$$\Delta^\sigma_{n,m,\Omega(n,m)-1}\Big|_{(0,1)^n} \subset p_{\Omega(n,m)-1,m}\left(\overline{\mathrm{Emb}((0,1)^n,\mathbb{R}^{\Omega(n,m)-1})}^{\mathrm{loc}}\right). \tag{171}$$

As $\Omega(n,m)-1 < \Omega(n,m)$, by the definition of $\Omega(n,m)$:

$$p_{\Omega(n,m)-1,m}\left(\overline{\mathrm{Emb}((0,1)^n,\mathbb{R}^{\Omega(n,m)-1})}\right) \not\supseteq C^1((0,1)^n,\mathbb{R}^m), \tag{172}$$

and thus,

$$\Delta^\sigma_{n,m,\Omega(n,m)-1}\Big|_{(0,1)^n} \not\supseteq C^1((0,1)^n,\mathbb{R}^m). \tag{173}$$

Therefore, we have $C^1(\mathbb{R}^n,\mathbb{R}^m) \not\subseteq \overline{\Delta^\sigma_{n,m,\Omega(n,m)-1}}^{\mathrm{loc}}$. This completes the proof. □

E Proofs of Topological Lemmas

E.1 Proof of Lemma 5.5

Proof. Define $F : (-\delta, 1+\delta) \times U \times U \to \mathbb{R}^{m+1}$ as

$$F(\alpha, x, y) = (\alpha, \alpha\,(f(x) - f(y)) + (1-\alpha)\,(g(x) - g(y))) \,. \tag{174}$$

Then F is a proper submersion for sufficiently small ϵ. Then, $DF(\alpha, x, y)$ can be calculated as

$$DF = \begin{bmatrix} 1 & 0 & 0 \\ f(x) - f(y) - (g(x) - g(y)) & \alpha Df(x) + (1-\alpha)Dg(x)) & -\alpha Df(y) - (1-\alpha)Dg(y)) \end{bmatrix}. \tag{175}$$

Consider a vector field X_i in $(-\delta, 1+\delta) \times U \times U$ for $i \in [1, m+1]_{\mathbb{N}}$ which satisfy the following:

$$(DF)X_i = e_i, \tag{176}$$

where e_i is the i-th coordinate vector. Then, define $G : F^{-1}(\mathbb{R}^{m+1}) \to \mathbb{R}^{m+1} \times F^{-1}(0)$ as

$$G(z) := (F(z), \phi_{X_{m+1}}^{-F(z)_{m+1}} \circ \cdots \circ \phi_{X_1}^{-F(z)_1}). \tag{177}$$

Then, G has a inverse $G^{-1} : \mathbb{R}^{m+1} \times F^{-1}(0) \to F^{-1}(\mathbb{R}^{m+1})$ which can be calculated as

$$G^{-1}(t_1, t_2, \ldots, t_{m+1}, x) = \phi_{X_1}^{t_1} \circ \cdots \circ \phi_{X_{m+1}}^{t_{m+1}}(x), \tag{178}$$

for $x \in \mathbb{R}$ and $x \in F^{-1}(0)$. Then, for the projection $p : \mathbb{R}^{m+1} \times F^{-1}(0) \to \mathbb{R}^{m+1}$, the following equation holds:

$$p = F \circ G^{-1}. \tag{179}$$

Therefore, $F^{-1}(c_1)$ is diffeomorphic to $F^{-1}(c_2)$ for $c_1, c_2 \in \mathbb{R}^{m+1}$.

Note that the above diffeomorphism G can be defined for all X_i that satisfy Equation (176).

We set X_1 as

$$X_1 := (DF)^T \left(DF(DF)^T\right)^{-1} e_1 \tag{180}$$

Then, $\phi_{X_1}^1$ is the diffeomorphism between $F^{-1}(0,0) = \{0\} \times \widetilde{f}^{-1}(0)$ and $F^{-1}(1,0) = \{1\} \times \widetilde{g}^{-1}(0)$. Let X_1 be represented as

$$X_1(\alpha, x, y) = \begin{bmatrix} 1 \\ M_1(\alpha, x, y) \\ M_2(\alpha, x, y) \end{bmatrix}. \tag{181}$$

It is enough to prove that $M_1(\alpha, y, x) = M_2(\alpha, x, y)$. Let

$$A(x, y) := f(x) - f(y) - (g(x) - g(y)), \tag{182}$$

and

$$B(\alpha, x) = B(x) := \alpha Df(x) + (1-\alpha)Dg(x)). \tag{183}$$

Then, $A(y, x) = -A(x, y)$.

$(DF)^T DF$ can be represented as

$$DF(DF)^T = \begin{bmatrix} 1 & A^T \\ A & AA^T + B(x)B(x)^T + B(y)B(y)^T \end{bmatrix}. \tag{184}$$

Then,

$$\left(DF(DF)^T\right)^{-1} = \begin{bmatrix} 1 + A^T(B(x)B(x)^T + B(y)B(y)^T)^{-1}A & -A^T(B(x)B(x)^T + B(y)B(y)^T)^{-1} \\ -(B(x)B(x)^T + B(y)B(y)^T)^{-1}A & (B(x)B(x)^T + B(y)B(y)^T)^{-1} \end{bmatrix}. \tag{185}$$

$$X_1 = \begin{bmatrix} 1 \\ M_1(\alpha, x, y) \\ M_2(\alpha, x, y) \end{bmatrix} = (DF)^T \left(DF(DF)^T\right)^{-1} e_1 = \begin{bmatrix} 1 \\ -B(x)(B(x)B(x)^T + B(y)B(y)^T)^{-1}A \\ B(y)(B(x)B(x)^T + B(y)B(y)^T)^{-1}A \end{bmatrix}. \tag{186}$$

$M_1(\alpha, y, x) = M_2(\alpha, x, y)$. And this completes the proof. □

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.