Incremental Generation is Necessity and Sufficient for Universality in Flow-Based Modelling

Hossein Rouhvarzi ROUHVARZI@GMAIL.COM
McMaster University and the Vector Institute
Department of Mathematics
1280 Main Street West, Hamilton, Ontario, L8S 4K1, Canada

Anastasis Kratsios* KRATSIOA@MCMASTER.CA
McMaster University and the Vector Institute
Department of Mathematics
1280 Main Street West, Hamilton, Ontario, L8S 4K1, Canada

Abstract

Incremental flow-based denoising models have reshaped generative modelling, but their empirical advantage still lacks a rigorous approximation-theoretic foundation. We show that incremental generation is necessary and sufficient for universal flow-based generation on the largest natural class of self-maps of $[0,1]^d$ compatible with denoising pipelines, namely the orientation-preserving homeomorphisms of $[0,1]^d$. All our guarantees are uniform on the underlying maps and hence imply approximation both samplewise and in distribution.

Using a new topological-dynamical argument, we first prove an impossibility theorem: the class of all single-step autonomous flows, independently of the architecture, width, depth, or Lipschitz activation of the underlying neural network, is meagre and therefore not universal in the space of orientation-preserving homeomorphisms of $[0,1]^d$. By exploiting algebraic properties of autonomous flows, we conversely show that every orientation-preserving Lipschitz homeomorphism on $[0,1]^d$ can be approximated at rate $\mathcal{O}(n^{-1/d})$ by a composition of at most K_d such flows, where K_d depends only on the dimension. Under additional smoothness assumptions, the approximation rate can be made dimension-free, and K_d can be chosen uniformly over the class being approximated. Finally, by linearly lifting the domain into one higher dimension, we obtain structured universal approximation results for continuous functions and for probability measures on $[0,1]^d$, the latter realized as pushforwards of empirical measures with vanishing 1-Wasserstein error.

Keywords: Incremental Generation; Normalizing Flows; Neural ODEs; Universal Approximation; Dynamical Systems; Homeomorphism Groups; Flowability; Optimal Transport, Structure-Aware AI.

MSC (2020): 41A30, 41A65, 58D05, 58D15, 37C10.

1 Introduction

Flow-based diffusion models are changing our understanding of generative modelling, combining exact likelihoods with high-quality samples Rezende and Mohamed (2015); Kingma and Dhariwal (2018), and the effectiveness of the modern *incremental generation* paradigm has undeniably revolutionized a range of areas from image synthesis to molecular design and time-series modelling Papamakarios et al. (2021); Ho et al. (2020); Kobyzev et al. (2020). Nevertheless, our fundamental question remains: why and how is incremental generation advantageous over (non-incremental) predecessors such as GANs Goodfellow et al. (2014). We answer this question in the foundations of AI, using an approximation-theoretic lens.

Here, a *generative model* refers to any function that transforms a source of noise, given by a random variable Z on $\mathbb{R}^d$, into another "generated" random variable

$$X = \varphi(Z), \tag{1}$$

where $\varphi : \mathbb{R}^d \to \mathbb{R}^D$. Generation is performed by sampling Z, which induces a sample of X. The transformation φ is required to be *continuous*, often Lipschitz, to ensure numerical stability during generation Gulrajani

∗. Corresponding author.

et al. (2017); Cisse et al. (2017); Miyato et al. (2018). We consider *denoising*-based generative models, such as normalizing flows Rezende and Mohamed (2015); Kingma and Dhariwal (2018) which conceptually build on their variational autoencoder (VAE) predecessors Kingma and Welling (2013). These models require the transformation φ to be *continuously invertible* and leverage the inverse map $\varphi^{-1} : \varphi(\mathbb{R}^d) \to \mathbb{R}^d$ to train the model: given i.i.d. samples from X, the model learns to map them back to the noise source Z. The invertibility, thus injectivity of φ necessitates that $D \geqslant d$; we focus on the difficult *critical case* where $D = d$ and φ is surjective (other cases are simple consequence which we discussed below). As with log-likelihood based generative models[1], cf. Dinh et al. (2014, 2017), prohibits φ from flipping inputs during generation; i.e. φ is *orientation preserving*. Lastly, keeping with the approximation-theoretic tradition in deep learning Mhaskar and Poggio (2016); Yarotsky (2017); Petersen and Voigtlaender (2018); Elbrächter et al. (2021); Kratsios and Zamanlooy (2022); Zhang et al. (2022); Siegel (2023), we will focus only on maps φ that are supported† (in the sense of differential topology, cf. (Banyaga, 1997a, Chapter 2)) on the unit hypercube $[0,1]^d$; that is, $\varphi(x) = x$ for every $x \notin [0,1]^d$. This means that, the largest *concept class* of denoising-type generative models consists of all random variables X, as in (1), for which the transformation φ belongs to the class $\mathcal{H}_d([0,1]^d)$ of *orientation-preserving homeomorphism supported*† *on* $[0,1]^d$. In this paper, we adopt a strong viewpoint by fixing the noise variable Z itself, rather than just fixing its distribution. This allows our results to hold *sample-wise*, not just in expectation or in distribution (e.g. a weaker Wasserstein sense), and it reduces the analysis of X in (1) to the properties of the map φ. Thus, when the worst-case gap between any two generated random variables $X = \varphi(Z)$ and $\tilde{X} = \psi(Z)$ for any sample of Z (samples/ω-wise), with $\varphi, \psi \in \mathcal{H}_d([0,1]^d)$, is quantified exactly by the uniform distance between φ and ψ

$$\sup_{\omega} \|\varphi(Z(\omega)) - \psi(Z(\omega))\| \leqslant d_\infty(\phi, \psi) \stackrel{\text{def.}}{=} \sup_{x \in \mathbb{R}^d} \|\varphi(x) - \psi(x)\| \tag{2}$$

where the supremum on the left-hand side of (2) is taken over all outcomes ω in the probability space on which Z is defined. We emphasize that $d_\infty(\phi, \psi)$ must be finite since both ϕ and ψ coincide outside of $[0,1]^d$. In other words, the relevant *concept class* in this paper is the function space $\mathcal{H}_d([0,1]^d)$ with distance quantified by d_∞, which will serve as the focus of all our approximation-theoretic analysis.

Unfortunately, $\mathcal{H}_d([0,1]^d)$ is not a vector space which places it outside the scope of the available constructive approximation toolbox Pinkus (2012); Lorentz et al. (1996); Cohen et al. (2022); Petrova and Wojtaszczyk (2023), which assumes a Banach space structure—our concept class $\mathcal{H}_d([0,1]^d)$ possesses a rich and well-studied infinite-dimensional topological group structure Mather (1974); Thurston (1974); Fathi (1980); Mather (1984); Banyaga (1997b); Haller et al. (2013); Fukui et al. (2019) that we will exploit extensively. Moreover, unlike signature-based machine learning methods, cf. Gyurkó et al. (2013); Chevyrev and Kormilitzin (2016); Cuchiero et al. (2023); Andrès et al. (2024); Cass and Salvi (2024); Bayer et al. (2025), rooted in rough path theory Lyons (1998); Gubinelli (2004); Hambly and Lyons (2010), we do not have access to a global exponential map (in the sense of infinite-dimensional Lie groups; cf. Kriegl and Michor (1997a); Michor et al. (2023)) to transfer results from linear spaces onto $\mathcal{H}_d([0,1]^d)$, as in Kratsios and Bilokopytov (2020); Kratsios and Papon (2022). In short, there are currently no available tool to apply classical approximation results from linear spaces onto $\mathcal{H}_d([0,1]^d)$ using global exponential map, we are required to develop new approximation-theoretic techniques suited to its non-vectorial structure, where our approximators/hypothesis class consists only of (neural network-based) homeomorphisms in $\mathcal{H}_d([0,1]^d)$. Our approximators preserve the structure of the maps in (1), whereas replacing φ with an arbitrary, e.g. non-invertible neural network as in the standard GAN framework Radford et al. (2016); Arjovsky et al. (2017); Peyré et al. (2019); Rout et al. (2022); Korotin et al. (2021, 2023); Kolesov et al. (2024) would prevent the use of denoising-based training.

Semi-Discrete Dynamics We identify and study the approximation-theoretic gap between *incremental* and *non-incremental* flow-based generative models. In this paper, a flow-based generative model means a

1. Assuming the necessary absolute continuity conditions, let p denote the Radon-Nikodym density of the law of X with respect to the law of Z. If p is sufficiently differentiable, the log-likelihood under a flow-based model takes the form $\log p_X(x) = \log p_Z(z) - \log(\det J_\varphi(z))$,where $J_\varphi(z)$ is the Jacobian matrix of φ at z. If φ is not orientation-preserving, $\det J_\varphi(z)$ can become negative, making the logarithm undefined. Thus, log-likelihood-based generative models must use orientation-preserving transformations.

map $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ decomposable as a "semi-discrete dynamical system"; by which we mean φ can be expressed as the composition of finitely many (discrete part) flows of continuous-time ODEs (continuous part). That is

$$\varphi = \varphi_G \circ \cdots \circ \varphi_1, \tag{3}$$

for some $G \in \mathbb{N}_+$, where for each $l \in \{1, \ldots, G\}$, the map $\varphi_l : \mathbb{R}^d \to \mathbb{R}^d$ is a time-1 flow of an autonomous ODE; that is, φ_l sends any $x \in \mathbb{R}^d$ to the time-1 solution x_1^x of the autonomous ODE with initial condition x

$$\frac{dx_t^x}{dt} = V_l(x_t^x), \quad x_0^x = x \tag{4}$$

where $V_l : \mathbb{R}^d \to \mathbb{R}^d$ is a Lipschitz vector field, parameterized by a ReLU multilayer perceptron (MLP)whose weights matrices has finite operator norm. Under this formalism, a flow-based generative model φ, as in (3), is called *incremental* if $G > 1$ and it is called *non-incremental* if $G = 1$. We note some L^p-version, for finite p, of these results in this direction are known via controllability arguments Ruiz-Balet and Zuazua (2024), however these results come at a cost as they require the user to have active control of the vector field of the Neural ODE (i.e. non-autonomous or controlled NeuralODEs) which is effectively the case where $G = \infty$.

Figure 1: Visualizing Incremental Flow-Based Generation (3)**:** In a denoising-type, flow-based incremental generator, an image x (left) is mapped to noise by the invertible sequence $\varphi_G^{-1} \circ \cdots \circ \varphi_1^{-1}$. The model is trained to undo this via $\varphi_1 \circ \cdots \circ \varphi_G$. At inference, a noise sample Z is injected and propagated through $\varphi_1, \ldots, \varphi_G$ to synthesize an image [2].
Here, $G = 2$. Each arrow depicts the vector field—visualized in the panel directly below—that induces one step of the incremental flow. By contrast, a non-incremental generator attempts a single-shot mapping from left to right, while non-denoising pipelines (e.g., GANs Goodfellow et al. (2014)) do not enforce invertibility.

In contrasts, incremental flow-based generative models can be approximately represented as a single *non-autonomous flow*, i.e., where the vector field in (4) is allowed to be time-dependent but undergoes only finitely many changes in direction; which directly allows the conversion of our theory of incremental flow-based generation to time-dependent generation. Such time-dependent models differ from *controlled neural ODEs* Kidger et al. (2020); Cuchiero et al. (2020); Cirone et al. (2023), which mimic controlled rough paths Morrill et al. (2021); Walker et al. (2024) and can potentially adapt their vector fields arbitrarily often over time resulting in a higher parametric complexity. The following question drives our manuscript and an affirmative answer to it would imply a concrete mathematical support for the advantage of incrimental generation over classical non-incremental generation to support the undeniable success of incremental generation

Question 1

Is there any approximation theoretic advantage of incremental generation?

2. Image credit: *Sans Soleil* (1983), written and directed by Chris Marker.

1.1 Main Construction

Our first result (Theorem 1) shows that incremental generation, i.e. when $G > 1$, is *necessary* for universality of flow-based generative models in $\mathcal{H}_d([0,1]^d)$ for $d \in \mathbb{N}_+$ and $d > 1$. Specifically, there exists a $\varphi \in \mathcal{H}_d([0,1]^d)$ that cannot be approximated by the flow of any autonomous neural ODE, regardless of network depth, width, or choice of activation function. We deduce this from a more general result of ours showing that: the class of *all* flows of autonomous ODEs is meagre (in particular, not dense) in $\mathcal{H}_d([0,1]^d)$. This result extends the differential topology results of Wescley Bonomo (2020) to the non-compact case and the classical embedding results of Fort (1955); Jones (1972); Palis (1974) to the non-smooth, non-compact setting.

Our second result (Theorem 7) shows that incremental generation with ReLU vector fields is *sufficient*. Moreover, the result is quantitative, and, surprisingly, we find that the added homeomorphism structure allows our approximation rates to match the minimax-optimal rates achieved by ReLU neural networks in the broader classes of uniformly continuous Yarotsky (2017); Shen (2020) and C^k-functions Petersen and Voigtlaender (2018); Yarotsky and Zhevnerchuk (2020); Lu et al. (2021) from $[0,1]^d$ to $\mathbb{R}^d$, respectively.

Our final main result (Theorem 9) shows, somewhat surprisingly, that G never needs to become arbitrarily large when approximating diffeomorphisms. Even more strikingly, we prove that there exists a dimensional constant $K_d > 0$ such that one can constrain $G \leqslant K_d$, *independantly of the target diffeomorphism being approximated* and maintain universality in $\mathcal{H}_d([0,1]^d)$.

1.2 Secondary Contributions

Several other consequences are considered in section 4 namely Universal Approximation of arbitrary Lipschitz functions between arbitrary dimensions and Universal Approximation of measures satisfying Caffarelli conditions.

2 Preliminaries

2.1 Background

2.1.1 Homeomorphisms, Flows and Flowability

This section contains all necessary terminology and background needed to formulate our main results. Additional background required only for proofs is contained in Appendix C. We say a function is **supported** on a set S if it is zero out of that set and a function is **supported**† if it is the identity out of that set. For practical reasons said in the introduction section, we need to work with **compactly supported/supported**† functions and vector fields; meaning that S is contained in a compact set.

We recall that, a **homeomorphism** is a continuous bijective function with continuous inverse. Furthermore, if the function and its inverse are C^k-differentiable, it is called **C^k-diffeomorphism**. By convention, C^∞-diffeomorphism is called *diffeomorphism*. Two continuous functions $f, g : X \to Y$ between the same topological spaces X and Y are thought of as being "topologically similar" if one can be continuously deformed into another; such as continuous deformation is called a homotopy; i.e. $H(t,x) : [0,1] \times X \to Y$ such that $H(0,\cdot) = f$ and $H(1,\cdot) = g$. We will say that two homeomorphisms are **isotopic** if one can continuously transform one into the other with homeomorphisms.

Example 1 *If $d, D \in \mathbb{N}_+$ and $f, g : \mathbb{R}^d \to \mathbb{R}^D$ are continuous then the so-called linear homotopy $H : [0,1] \times \mathbb{R}^d \to \mathbb{R}^D$ sending any pair (t,x) to $(1-t)f(x) + tg(x)$ is indeed a homotopy.*

A homeomorphism/C^k-diffeomorphism is **orientation preserving** if it is homotopic to identity[3].

Example 2 (A $1d$ Example of (Non-)Orientation Preserving Homeomorphism) *A transparent example of an orientation preserving homeomorphism in one dimension is the identity $\mathbb{R} \ni x \to x \in \mathbb{R}$. A*

3. By Alexander's trick, compactly supported† homeomorphisms in $\mathbb{R}^n$ are isotopic to identity so orientation preserving. This is etymologically more natural definition in terms of orientations; however this is the most mathematically convenient and concise for our analysis

homeomorphism from $\mathbb{R}$ to itself which is not is the mirroring map $\mathbb{R} \ni x \mapsto -x \in \mathbb{R}$. Indeed, both of these maps are homeomorphisms since they are their own inverses.

Example 3 (Orientation-Preserving Homeomorphisms From Computer Vision (Rotations)) *Standard Multidimensional examples arising in rotation invariances in computer vision, e.g. Lui (2012); Cohen and Welling (2016); Thomas et al. (2018), include linear transformerization $\varphi : \mathbb{R}^d \ni x \mapsto Ox \in \mathbb{R}^d$ where O is an $d \times d$ orthogonal matrix; in which case φ is orientation preserving if and only if $\det(O) = 0$ and it fails to be precisely when $\det(O) = -1$.*

If O is orientation-preserving; i.e. if $\det(O) = 1$, then any such homeomorphism can be expressed as the solution to an ordinary differential equation (ODE) at time 1; namely, $\varphi(x) = Ox = x_1^x$ where

$$\begin{aligned} \tfrac{d}{dt}\, x_t^x &= \mathfrak{o} x_t^x \\ x_0^x &= x \end{aligned} \tag{5}$$

where $O = \exp(\mathfrak{o})$ for some $d \times d$-skew-symmetric matrix $\mathfrak{o}$; where $\exp$ is the matrix exponential. This can be noted upon observing that the ODE solution to the ODE (5) is given by the curve $x_\cdot^x = (x_t^x)_{t \geqslant 0}$ where

$$x_t^x = \exp(t\mathfrak{o})\, x.$$

This connection is the starting-point of the theory of finite-dimensional Lie groups; cf. Helgason (1979). Indeed, finite-dimensional Lie groups induce the prototypical and simplest classes of "well-behaved" homeomorphisms on $\mathbb{R}^d$ in this way; however, only the identity is compactly supported in any such construction where the vector field in (5) is "constant" multiplication against a single matrix.

Generalizing Example 3 we bring our attention to autonomous ODEs of the form (4), where vector-field $V : \mathbb{R}^d \to \mathbb{R}^d$ is allowed to be any Lipschitz or C^k-differentiable vector field compactly supported on interval domains (i.e. $[a, b]^d$ for $a, b \in \mathbb{R}$). By the Picard-Lindelöf Theorem (Hartman, 2002, *Chapter II* Theorem 1.1) this system has a unique global solution x_t^x and the map sending the point x to the time-1 value x_1^x of the solution to this problem (often referred to as Cauchy problem) will be called the **flow** induced by the vector field V, and is denoted by Flow$\,(V)$. The flow of any Lipschitz vector field supported on a compact set S defines a compactly supported† homeomorphism, supported† on S. If, moreover, V is C^k-differentiable then its flow is actually a C^k-diffeomorphism. We say that an orientation-preserving homeomorphism in d dimensional euclidean space, $\varphi \in \mathcal{H}_d([0, 1]^d)$ is **flowable** if there exists a Lipschitz Vector field V supported within $[0, 1]^d$ such that $\varphi = \text{Flow}\,(V)$. An interesting connection, worth noting here in analogy with Example 3—is that the spaces of flows and diffeomorphisms Banyaga (1997a) constitute the prototypical infinite-dimensional Lie groups, a fact that sparked the foundational work of Kriegl and Michor (1997c,b); Teichmann (2001); Neeb and Pianzola (2007); Omori (2017); Michor et al. (2023).

2.1.2 From Neural ODEs to Incremental Flow-Based Generators

Let $\Delta \in \mathbb{N}_+$ and consider a multi-index $\mathbf{d} \overset{\text{def.}}{=} [d_1, \dots, d_{\Delta+1}] \in \mathbb{N}_+^{\Delta+1}$. The class $\mathcal{NN}(\mathbf{d})$ consists of all multilayer perceptrons with $\sigma \in C(\mathbb{R})$ Lipschitz-activation function (σ-MLPs)

$$\Phi :\ \mathbb{R}^{d_1} \to \mathbb{R}^{d_{\Delta+1}}$$

admitting the following iterative representation

$$\begin{aligned} \Phi(\mathbf{x}) &= \mathbf{W}^{(\Delta)}\mathbf{x}^{(\Delta)} + \mathbf{b}^{(\Delta)}, \\ \mathbf{x}^{(l+1)} &\overset{\text{def.}}{=} \sigma \bullet \big(\mathbf{W}^{(l)}\mathbf{x}^{(l)} + \mathbf{b}^{(l)}\big), \qquad \text{for } l = 1, \dots, \Delta - 1, \\ \mathbf{x}^{(1)} &\overset{\text{def.}}{=} \mathbf{x}\,. \end{aligned} \tag{3.2}$$

Here, for $l = 1, \dots, \Delta$, $\mathbf{W}^{(l)}$ is a $d_{l+1} \times d_l$ matrix and $\mathbf{b}^{(l)} \in \mathbb{R}^{d_{l+1}}$, and $\sigma\,\bullet$ denotes componentwise application of the σ function.

Fix a Lipschitz activation function $\sigma \in C(\mathbb{R})$ and real numbers $a \leqslant b$. A σ-neural ODE of respective depth and width $\Delta, W \in \mathbb{N}_+$ is a flowable homeomorphism $\varphi \in \mathcal{H}_d([0,1]^d)$ for which there exists a σ-MLP $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ of depth Δ and width W supported in $[a,b]^d$ such that φ is time one solution of (4) with vector field Φ. The class of all σ-neural ODEs supported on $[a,b]^d$ is denoted as $\mathrm{NODE}_\sigma([a,b]^d)$. The class of incrimental flow-based generators with activation function σ denoted as $\mathrm{IFG}_\sigma([a,b]^d)$ is defined as all functions of $\mathcal{H}_d([a,b]^d)$ that can be written as composition of at least two or more but finitely many σ-neural ODEs

2.2 Notation

Before moving on, we now collect a list of notation used in our manuscript.

- Let $\mathbb{N} \overset{\text{def.}}{=} \{0,1,2,\dots,\}$, $\mathbb{N}_+ \overset{\text{def.}}{=} \{n \in \mathbb{N} : n > 0\}$ and for $N \in \mathbb{N}_+$ denote $[N] \overset{\text{def.}}{=} \{1,\dots,N\}$
- For a Lipschitz function f we denote its Lipschitz constant as L^f
- Given a function $f : \mathbb{R}^d \to \mathbb{R}^D$, we denote its support by $\mathrm{supp}(f) \overset{\text{def.}}{=} \overline{\{x \in \mathbb{R}^d : f(x) \neq 0\}}$ and its support† by $\mathrm{supp}^\dagger(f) \overset{\text{def.}}{=} \overline{\{x \in \mathbb{R}^d : f(x) \neq x\}} = \overline{\mathrm{supp}(f - \mathrm{id})}$
- For a set P denote its interior by $int(P)$
- Denote (open Euclidean) ball with radious r around point p as $B_r(p) = \{x \in \mathbb{R}^d | \ \|x - p\| < r\}$ where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$.
- Let $\mathcal{H}_d$ be the set of orientation preserving homeomorphisms of $\mathbb{R}^d$. Given subset $P \subset \mathbb{R}^d$, let $\mathcal{H}_d(P)$ be the set of orientation preserving homeomorphisms of $\mathbb{R}^d$ compactly supported† on P
- For $s \in \mathbb{N} \cup \{\infty\}$ let $\mathcal{X}^s$ be the set of C^s vector fields on $\mathbb{R}^d$ supported† on $[0,1]^d$. when $s = 0$, we only consider *Lipschitz* continuous vector fields for $\mathcal{X}^0$.
- For a given vector field $V \in \mathcal{X}^0$ let $\mathrm{Flow}(V)$ be the time one solution of:

$$\begin{cases} \dot{x}_t^x &= V(x_t^x) \\ x_0^x &= x. \end{cases} \tag{6}$$

  and $\mathrm{Flow}([a,b]^d)$ be the set of time one solutions above with vector fields supported on $[a,b]^d$ and

$$\mathrm{Flow} \overset{\text{def.}}{=} \bigcup_{\forall k \in \mathbb{N} \cup \{\infty\}} \{\mathrm{Flow}(V) | V \in \mathcal{X}^k\}$$

  also $\mathrm{Flow}_\sigma([a,b]^d)$ be the set of time one solutions with σ neural vector fields supported on $[a,b]^d$
- Fix a granularity $G \in \mathbb{N}_+$ and a smoothness $s \in \mathbb{N}$. A homeomorphism $\varphi \in \mathcal{H}_d([0,1]^d)$ belongs to $\mathcal{H}_d^{G,s}([0,1]^d)$ if and only if:
  - (i) **Representation:** There exist vector fields $V_1,\dots,V_G \in \mathcal{X}^s$ such that

$$\varphi = \bigcirc_{g=1}^{G} \mathrm{Flow}\,(V_g). \tag{7}$$

  - (ii) **Minimality:** There is no integer $1 \leqslant \tilde{G} < G$ and vector fields $\tilde{V}_1,\dots,\tilde{V}_G \in \mathcal{X}^s$ such that

$$\varphi = \bigcirc_{g=1}^{\tilde{G}} \mathrm{Flow}\,(\tilde{V}_g).$$

- Let $\mathrm{Diff}_0(P)$ be the set of orientation preserving diffeomorphisms of $\mathbb{R}^d$ compactly supported on the set P .

- Given G functions $\{f_1, \ldots, f_G\}$, the iterated composition operator $\bigcirc$ maps any finite set of composable functions $f_1, \ldots, f_G$ to their composition $\bigcirc_{g=1}^G f_i \stackrel{\text{def.}}{=} f_G \circ \cdots \circ f_1$.
- Given $E \subseteq \mathbb{R}^d$, d-tuple $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_d]^T \in \mathbb{N}^d$ and functions $f(x) : \mathbb{R}^d \to \mathbb{R}^D$ and $g : \mathbb{R}^d \to \mathbb{R}$ let :
  - $\|f(x_0)\|_{l^\infty} = \|(f_1(x_0), \ldots, f_D(x_0))\|_{l^\infty} = \max_{i\in[D]} f_i(x_0)$
  - $\|g\|_{L^\infty(E)} = \text{ess sup}_{x\in E} |g(x)|$
  - $\|f\|_{L^\infty(E)l^\infty} = \|\text{ess sup}_{x\in E} |f_1(x)|, \ldots, \text{ess sup}_{x\in E} |f_1(x)| \,\|_{l^\infty} = \max_{i\in[D]} \text{ess sup}_{x\in E} f_i(x_0)$
  - $\|\alpha\|_1 = |\alpha_1| + |\alpha_2| + \cdots + |\alpha_d|$
  - $\partial^\alpha = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \frac{\partial^{\alpha_2}}{\partial x_2^{\alpha_2}} \cdots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}}$
  - $\|g\|_{C^s(E)} := \max \left\{ \|\partial^\alpha g\|_{L^\infty(E)} : \alpha \in \mathbb{N}^d \text{ with } \|\alpha\|_1 \leqslant s \right\}$[4]
- Compositional Notation: During the course of our analysis, it will be convenient to describe ReLU MLPs via the role of each of their (sets of) layers. Specifically, the structure of a ReLU MLP Φ is represented in the following way: suppose $\Phi = \mathcal{L}_m \circ (\sigma \circ \mathcal{L}_{m-1}) \circ \cdots \circ (\sigma \circ \mathcal{L}_2) \circ (\sigma \circ \mathcal{L}_1)$ where the $\mathcal{L}_i$'s are affine transformations. Following Hong and Kratsios (2024), we express this notationally as

$$\begin{aligned}\mathbf{x} &\Longrightarrow (\sigma\circ\mathcal{L}_1)(\mathbf{x}) \Longrightarrow (\sigma\circ\mathcal{L}_2) \circ (\sigma\circ\mathcal{L}_1)(\mathbf{x}) \\ &\Longrightarrow (\sigma\circ\mathcal{L}_{m-1}) \circ \cdots \circ (\sigma\circ\mathcal{L}_2) \circ (\sigma\circ\mathcal{L}_1)(\mathbf{x}) \\ &\Longrightarrow \mathcal{L}_m \circ (\sigma\circ\mathcal{L}_{m-1}) \circ \cdots \circ (\sigma\circ\mathcal{L}_2) \circ (\sigma\circ\mathcal{L}_1)(\mathbf{x}) \\ &= \Phi(\mathbf{x}) \,.\end{aligned}$$

  In other words, if $x_1, x_2, \ldots, x_{m-1}$ are the $1, 2, \ldots, (m-1)$-th hidden layers of Φ and x_m is the output layer, then the structure of Φ is expressed as

$$x \Longrightarrow x_1 \Longrightarrow x_2 \Longrightarrow \cdots \Longrightarrow x_{m-1} \Longrightarrow x_m = \Phi(x) \,.$$

3 Main Results

3.1 Negative Results: Incremental Generation is Necessary

Our first main results shows that the collection of *all* flowable homeomorphisms compactly supported† on $[\delta, 1-\delta]^d$ for any given $0 < \delta < \frac{1}{2}$. (a subset of $\mathcal{H}_d([\delta, 1-\delta]^d)$) are a small subset of homeomorphisms compactly supported† on $[0,1]^d$. This means there is not only one but *many* functions in $\mathcal{H}_d([0,1]^d)$ which cannot be approximated by flowable functions in $\mathcal{H}_d([\delta, 1-\delta]^d)$ for any $0 < \delta < \frac{1}{2}$.

Theorem 1 (Non-Incremental Generation is Not Universal) *Let $d \in \mathbb{N}_+$ with $d > 1$ and consider the class of all non-incremental generators, i.e. autonomous neural ODEs, for any Lipschitz activation function*

$$\text{NODE}_\sigma((0,1)^d) \stackrel{\text{def.}}{=} \bigcup_{\sigma\in\text{Lip}(\mathbb{R}^d,\mathbb{R}^d)} \bigcup_{0<\delta<\frac{1}{2}} \text{Flow}_\sigma([\delta, 1-\delta]^d).$$

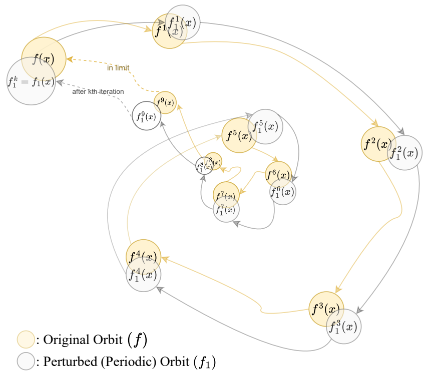
Then, $\text{NODE}_\sigma((0,1)^d)$ *is nowhere dense in* $\mathcal{H}_d([0,1]^d)$.

Now, Theorem 1 is implied by our more general result in topological dynamics, illustrated in Figure 2,

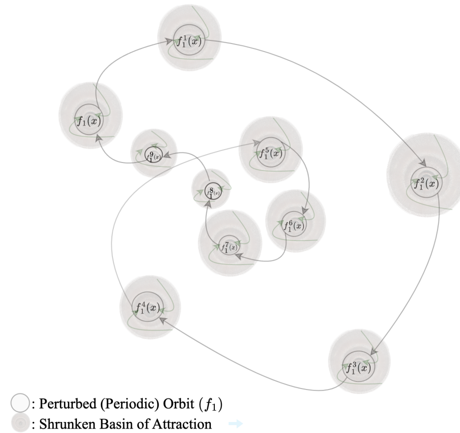
Theorem 2 (Few C^0 Homeomorphisms are Flowable on $[0,1]^d$ for $d > 1$) *Let $d \in \mathbb{N}_+$ with $d > 1$. The set of flowable homeomorphisms in $\mathcal{H}_d([0,1]^d)$ is meager in $\mathcal{H}_d(P)$ for every $d > 1$ and compact set P s.t. $[0,1]^d \subset int(P)$.*

Remark 3 *For simplicity and convention $[0,1]^d$ is considered but this theorem is also true for any compact set K, P with $K \subset int(P)$. To understand and see the proof refer to Appendix A.*

4. Note that $\|g\|_{C^0} = \|g\|_{L^\infty}$



(a) The orbits of any (topologically) generic dynamical system (yellow) are perturbed to become periodic, by perturbing the underlying system (gray) – C^0-Closing lemma.



(b) The dynamical system is then further perturbed so that each periodic point becomes a (non-fixed) periodic attractor; which cannot happen for any flow.

Figure 2: Why Non-Incremental Generation is Not Universal: The reason why non-incremental generators (Theorem 1) fail to be universal is that most homeomorphisms cannot be represented as flows (Theorem 2). The idea is that there is a dense open set of orientation-preserving homeomorphisms supported on the hypercube $[0,1]^d$, which can be approximated/perturbed so that any given orbit becomes periodic (SubFigure 2a). Then, these perturbations can be further perturbed so that the a small neighbourhood around the given orbit becomes a basin of attraction (SubFigure 2b), which cannot happen for any flow. Consequently, the complement of any such map, which contains the set $\mathrm{NODE}^\star_\sigma([0,1]^d)$ cannot be dense, implying that all non-incremental generators/autonomous Neural ODEs fail to be universal approximators of orientation-preserving homeomorphisms supported on the hypercube $[0,1]^d$.

3.2 Positive Results: Incremental Generation with ReLU MLP Vector Fields is Sufficient

3.2.1 Qualitative Universal Approximation Guarantee

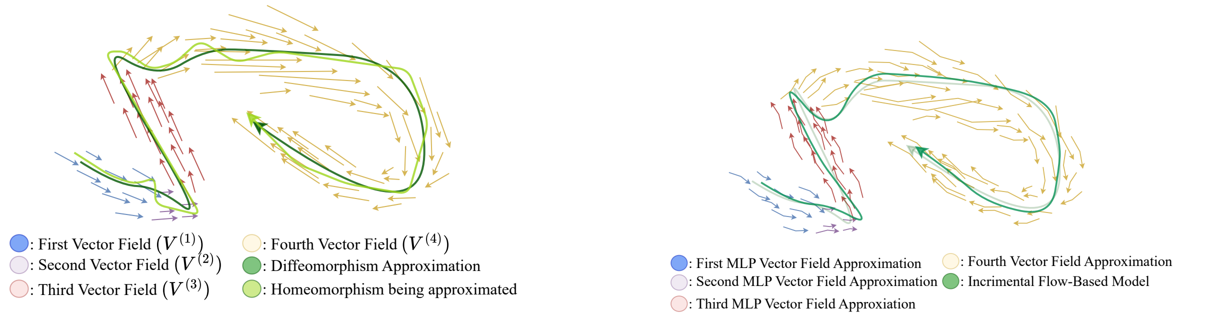
We now contrast our previous result, which shows the limitations of non-incremental generation with our *universal* approximation guarantee, showing that the set of incremental flow-based generators $\mathrm{IFG}_\sigma([0,1]^d)$ is universal in $\mathcal{H}_d([0,1]^d)$; *quantitatively.* We emphasize our critical *structural* point that the approximation is "within" the class $\mathcal{H}_d([0,1]^d)$ not from "outside"; by which we mean that our approximating class only consists of homeomorphisms not standard continuous functions.

We now state the streamlined qualitative version of our main result, in the high-dimensional setting where $d \geqslant 5$, before entering into a more technical analysis involving approximation rates and exactly parameter estimates, shortly. The main intuition behind this result is summarized in Figure 3, whose proof technique largely combines both algebraic and approximation theoretic tools.

Theorem 4 (Universal Approximation of Orientation-Preserving Homeomorphisms) *Let $d \in \mathbb{N}_+$ and $d \geqslant 5$. There exists a constant $K_d \in \mathbb{N}_+$ such that: for every $\varphi \in \mathcal{H}_d([0,1]^d)$ and every $\varepsilon > 0$ there exists some $G \leqslant K_d$ and ReLU neural ODEs $\{\Psi^{(g)} = \mathrm{Flow}(\Phi^{(g)})\}_{g=1}^G \subset \mathcal{X}^0$ such that the diffeomorphism $\Psi \overset{\text{def.}}{=} \bigcirc_{g=1}^G \Psi^{(g)}$ is Lipschitz and compactly supported and satisfies the approximation guarantee:*

$$\left\|\varphi - \Psi\right\|_{L^\infty(\mathbb{R}^d) l^\infty} \leqslant \varepsilon. \tag{8}$$

Remark 5 *For simplicity and convention $[0,1]^d$ is considered but all of the theorems in this section are also true for any compact set P.(Zhang et al. (2024),Yarotsky (2018),Petersen and Voigtlaender (2018)) To understand and see the proof refer to Appendix B.*



(a) Approximation of homeomorphisms using decomposition of diffeomorphisms into composition of flows.

(b) Approximation of each flow by ReLU MLP approximation of its vector field.

Figure 3: By (a uniform version) of Thursten's Theorem Thurston (1974) we $\mathcal{H}_d([0,1]^d)$ is a simple group and since the group generated by flows is normal, then every diffeomorphism (green curve) must be the composition of *finitely many* flows of vector fields $V^{(1)}, \ldots, V^{(G)}$) (here $G = 4$) – Sub-Figure 3a. Each vector field is then approximated by a ReLU MLP at an optimal rate (Sub-Figure 3b) with maximal Lipschitz regularity using Hong and Kratsios (2024); the approximation of the original homeomorphism (Theorem 7) is concluded using Grönwall's inequality.
Furthermore, in dimension $d \geqslant 5$, every orientation preserving homeomorphism is isotopic to the identity, then Müller (2014) implies that it can be uniformly approximated by diffeomorphisms; reducing (Theorem 4) to the smooth case.

3.2.2 Quantitative Formulations

We start this section by remarking that, the class $\mathcal{H}_d^{G,0}([0,1]^d)$ for some $G > 1$ is non-empty. Thus the theorem provides a non-vacuous statement for $G > 1$.

Proposition 6 (Non-triviality) *there exists some $G \in \mathbb{N}_+$ and $G > 1$ that $\mathcal{H}_d^{G,0}([0,1]^d)$ is non-empty,*

We now provide a detailed quantitative analysis of the general qualitative result of the previous section.

Theorem 7 (Universal Approximation by Deep Neural ODEs) *Let $n, d \in \mathbb{N}$, and $\varphi \in \mathcal{H}_d^{G,0}([0,1]^d)$ then, there exists ReLU neural ODEs $\{\Psi^{(g)} = \mathrm{Flow}(\Phi^{(g)})\}_{g=1}^G \subset \mathcal{X}^0$ such that $\Psi \stackrel{\text{def.}}{=} \bigcirc_{g=1}^G \Psi^{(g)}$ satisfies the approximation guarantee:*

$$\|\varphi - \Psi\|_{L^\infty(\mathbb{R}^d)l^\infty} \leqslant \sum_{g=1}^G \left(2\|\omega^{(g)}(\frac{d}{2n})\|_{l^\infty} \prod_{j=g}^G e^{L^{V^{(j)}}}\right) \tag{9}$$

The right hand side converges to zero as $n \to \infty$.
Moreover, Ψ is a compactly-supported† *homeomorphism on $\mathbb{R}^d$ supported*† *on $(0,1)^d$, with Lipschitz constant at-most $\prod_{g=1}^G e^{L^{\Phi^{(g)}}}$ and $\Phi^{(1)}, \ldots, \Phi^{(G)}$ are ReLU MLP vector fields of depth $\lceil log_2 d\rceil + 6$, width $8d(n+1)^d + 9$, and at-most $16d(n+1)^d + 9$ non-zero parameters. In particular, these vector fields and L^Ψ do not depend on the parameter n.*
Furthermore, $\omega^{(g)}$ is the modulus of regularity of $\varphi^{(g)}$ which is equal to the modulus of regularity of $\Phi^{(g)}$.

Naturally, one may wonder if improved rates are achievable under additional smoothness of the target homeomorphism. Indeed, we confirm that this is the case, mirroring the classical approximation theory of smooth functions by ReLU MLPs.

Proposition 8 (Universal Approximation by Deep Neural ODEs (Differentiable Case)) *Let $\varphi \in \mathcal{H}_d^{G,s}([0,1]^d)$ and N, L, $s \in \mathbb{N}_+$ then, there exists ReLU neural ODEs $\{\Psi^{(g)} = \mathrm{Flow}(\Phi^{(g)})\}_{g=1}^G \subset \mathcal{X}^0$ such that $\Psi \stackrel{\text{def.}}{=} \bigcirc_{g=1}^G \Psi^{(g)}$ satisfies the approximation guarantee:*

$$\|\varphi - \Psi\|_{L^\infty(\mathbb{R}^d)l^\infty} \leqslant \sum_{g=1}^G \left(2\|\omega^{(g)}(N,L)\|_{l^\infty} \prod_{j=g}^G e^{L^{V^{(j)}}}\right) \tag{10}$$

The right hand side goes to zero as $L \to \infty$ and $N \to \infty$.
Moreover, Ψ is a compactly-supported† C^s-diffeomorphism on $\mathbb{R}^d$ supported† on $(0,1)^d$, with Lipschitz constant at-most $\prod_{g=1}^G e^{L^{\Psi(g)}}$ and $\Phi^{(1)}, \ldots, \Phi^{(G)}$ are ReLU MLPs of width less than or equal to $17s^{d+1}3^d d^2(N+2)\log_2(8N)$ and depth $18s^2(L+2)\log_2(4L)+2(d+1)$ where $N, L \in \mathbb{N}_+$. In particular, L^Ψ depends on the parameter N.
Furthermore, $(\omega^{(g)})_j = 85(s+1)^d 8^s \|V_j^{(g)}\|_{C^s([0,1]^d)}(NL)^{-2s/d}$.

Complexity of MLP Vector fields	**Lipschitz Case**	**C^s-Differentiable Case ($s \geqslant 1$)**
Depth	$\lceil \log_2(d) \rceil + 4$	$18s^2(L+2)\log_2(4L) + 2(d+1)$
Width	$8d(n+1)^d$	$17sd^{d+13}d^2(N_j+2)\log_2(8N_j)$
Nonzero parameters	$16d(n+1)^d$	

Table 1: Parametric Complexity of the Flow-Based Generative Model of Theorem 7 and of Proposition 8.

Now using the theorem bellow we give an order of approximation for any diffeomorphism in Proposition 10.

Theorem 9 (Finite Composition of Flows for Diffeomorphisms) *There exists a positive number $K_d \in \mathbb{N}_+$ such that any diffeomorphism $\varphi \in \operatorname{Diff}_0([0,1]^d)$ can be written as at most K_d flows.*

Proposition 10 (Universal Approximation by Deep Neural ODEs (Smooth Case)) *Let $\varphi \in \operatorname{Diff}_0([0,1]^d)$ and $N, L \in N_+$ and take L fixed; then, for $\forall s \in \mathbb{N}_+$ there exists ReLU neural ODEs $\{\Psi^{(g)} = \operatorname{Flow}(\Phi^{(g)})\}_{g=1}^G \subset \mathcal{X}^0$ such that $\Psi \overset{\text{def.}}{=} \bigcirc_{g=1}^G \Psi^{(g)}$ satisfies the approximation guarantee:*

$$\left\|\varphi - \Psi\right\|_{L^\infty(\mathbb{R}^d)l^\infty} \in O\left(N^{-2s/d}\right) \tag{11}$$

Moreover, Ψ is a compactly-supported† C^s-diffeomorphism on $\mathbb{R}^d$ supported† on $(0,1)^d$, with Lipschitz constant at-most $\prod_{g=1}^G e^{L^{\Psi(g)}}$ and $\Phi^{(1)}, \ldots, \Phi^{(G)}$ are ReLU MLPs of width less than or equal to $17s^{d+1}3^d d^2(N+2)\log_2(8N)$ and depth $18s^2(L+2)\log_2(4L)+2(d+1)$ where $N, L \in \mathbb{N}^+$. In particular, L^Ψ depends on the parameter n.

4 Implications: Lifted Flow-Based Generation Imply Universal Approximation

We now present a sequence of corollaries that further illustrate the scope of our results. In particular, the simplest form of our main positive result—our approximation theorem—yields *structured* (in the sense of homeomorphisms "lying over," as explained below) versions of state-of-the-art universal approximations for ReLU MLPs, implemented via a "tweaked" incremental flow-based model. These results also imply universal generation in the classical sense of Wasserstein GANs Arjovsky et al. (2017).

4.1 Universal Approximation of arbitrary Lipschitz functions between arbitrary dimensions

At first glance, the homeomorphisms and dimensional constraints $d = D$, constraint defining (1), may seem to suggest that our universal flow-based generative models are overly restrictive and thus cannot approximate continuous functions between Euclidean spaces, locally on compact subsets as standard multilayer perceptrons do; cf. Hornik et al. (1989); Cybenko (1989); Funahashi (1989). However, this is not the case, and in fact the full-power of incremental generative models are not needed if one is prepared to sacrifice injectivity. The idea is similar to Klee's trick, cf. Klee (1955), where we encode transport the graph of an arbitrary continuous function into a homeomorphism. Rather, we can do better, by embedding a function into a *single* very simple time-1 flow up to conjuration be simple linear maps; our simple approach is summarized in Figure 4.

Fix $d, D \in \mathbb{N}_+$ and $L \geqslant 0$. Then, for every L-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}^D$ induces a $\bar{L} \overset{\text{def.}}{=} \max\{1, L\}$-Lipschitz vector field $V_f : \mathbb{R}^{d+D} \to \mathbb{R}^{d+D}$ defined for each $(x, y) \in \mathbb{R}^{d+D}$ by

$$V_f(x, y) \overset{\text{def.}}{=} \big(0, f(x)\big). \tag{12}$$

The solution $\Phi : [0, \infty) \times \mathbb{R}^{d+D} \to \mathbb{R}^{d+D}$ to the *autonomous* ODE induced by V_f defined by

$$\frac{dx_t^{(x,y)}}{dt} = 0, \qquad \frac{dy_t^{(x,y)}}{dt} = f(x_t^{(x,y)}), \text{ where } (x_0, y_0) = (x, 0)$$

is easily explicitly solved:

$$\frac{dx_t^{(x,y)}}{dt} = 0 \Rightarrow x_t^{(x,y)} = x \Rightarrow y_t^{(x,y)} = y_0 + \int_0^t f(x)dt \Rightarrow y_t^{(x,y)} = 0 + t\, f(x)$$

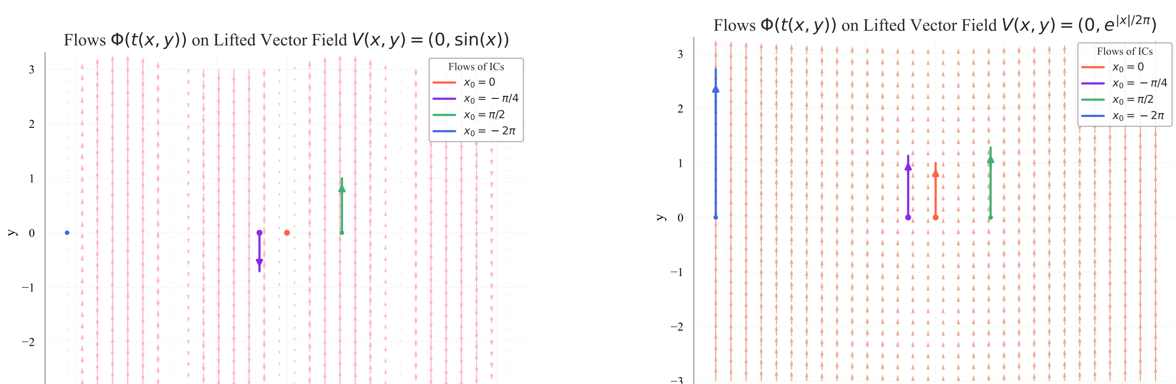
and yields the $\bar{L}$-Lipschitz map

$$\Phi(t, (x, y)) = \big(x, t\, f(x)\big). \tag{13}$$

Therefore, the time-1 flow of V_f, induced by f, is $\operatorname{Flow}(V_f)(x, y) = (x, y + f(x))$. Restricting this flow to the "lifted hypercube" $[0,1]^{d\uparrow} \overset{\text{def.}}{=} \{(x, y) \in \mathbb{R}^{d+D} : x \in [0,1]^d, \text{ and } y = 0\}$, which is a compact subset of $[0,1]^{d+D}$, we find that for any $(x, 0) \in [0,1]^{d\uparrow}$ we have $\operatorname{Flow}(V_f)(x, 0) = (x, f(x))$. Now, let $\pi_D^{d+D} : \mathbb{R}^{d+D} \to \mathbb{R}^D$ denote the canonical (linear) projection sending any $(x, y) \in \mathbb{R}^{d+D}$ to $y \in \mathbb{R}^D$ and let $\iota_d^{d+D} : \mathbb{R}^d \to \mathbb{R}^{d+D}$ denotes the (linear) embedding sending any $x \in \mathbb{R}^d$ to $(x, 0) \in \mathbb{R}^{d+D}$; in particular, note that $\iota_d^{d+D}([0,1]^d) = [0,1]^{d\uparrow}$. Putting it all together, we have that: for all $x \in [0,1]^d$

$$\pi_D^{d+D} \circ \operatorname{Flow}(V_f) \circ \iota_d^{d+D}(x) = f(x). \tag{14}$$

Applying Theorem 4 to f, we conclude that by post- and pre-composing our incremental flow-based models, we can uniformly approximate any continuous function on the cube $[0,1]^d$.



(a) Lift of: $f(x) = \sin(x)$ to $V(x, y) = (0, \sin(x))$.

(b) Lift of: $f(x) = e^{\frac{|x|}{2\pi}}$ to $V(x, y) = (0, e^{\frac{|x|}{2\pi}})$.

Figure 4: Any continuous function $f : \mathbb{R}^d \to \mathbb{R}$ can be realized as a time-1 flow $\Phi(1, (x, y))$ for the $(d+1)$-dimensional vector field $V(x, y) \overset{\text{def.}}{=} (0, f(x))$ (illustrated by the pink vector fields), which acts trivially in its first "dummy" coordinates and acts as the target function $f(x)$ in the $(d+1)^{\text{st}}$ coordinate. By mapping any given input $x \in \mathbb{R}^d$ to the initial condition $(x, 0) \in \mathbb{R}^{d+1}$ of the flow Φ "lifting" f, we then simply flow linearly relative to the $(d+1)^{\text{st}}$ (here y) axis until it arrives at $(x, f(x))$ at time 1, at which point the final value can be linearly projected-off and the value $f(x)$ is recovered. In this way, every real-valued continuous (resp. Lipschitz, resp. smooth) function can be realized as a time-1 flow of the same regularity in a space of only one more dimension.

However, we note that the model in (14) does not scale well when D is large as one would obtain an approximation rate of $\mathcal{O}(n^{d+D})$ which is significantly slower than the "unconstrained" optimal rate achievable by ReLU MLPs; cf. Yarotsky (2018). Instead, a rate nearly equal to the "unconstrained" ReLU MLP rate is possible if we alternatively approximate each component of the target function f independently using this strategy, and then concatenate the resulting lifted flow-based approximations. Denoting $f = (f_1, \dots, f_d)$, we replace (14) with the model

$$\bigoplus_{i=1}^{D} \Big(\pi_1^{d+1} \circ \operatorname{Flow}(V_{f_i}) \circ \iota_d^{d+1}(x)\Big) = \big(f_1(x), \dots, f_D(x)\big) = f(x). \tag{15}$$

The advantage of the representation in (15), with its greater width, over the more "naive narrow" version in (14), is that each $\operatorname{Flow}(V_{f_i})$ performs its approximation in only one extra dimension beyond the physical dimension d. This design minimizes the approximation-theoretic difficulties that typically arise from high dimensionality. Importantly, doing so achieves the minimax optimal approximation rates (cf. (Shen et al., 2022, Theorem 2.4) and Yarotsky (2017)) for the *lifted, and thus higher-dimensional* space $[0,1]^{d+1}$ and nearly achieves the optimal rate on the original low-dimensional domain $[0,1]^d$, up to an extra factor of 1; which is possibly inevitable due to the extra invertibility structure of the incremental models studied herein.

Corollary 11 (Approximation of Arbitrary Lipschitz Functions by Linear Lifting) *Let $d, D \in \mathbb{N}_+$, $L \geqslant 0$, and $f : \mathbb{R}^d \to \mathbb{R}^D$ be Lipschitz. For every $n \in \mathbb{N}_+$ and each $i \in [D]$ there exists ReLU Neural ODEs $\Psi_i : \mathbb{R}^{d+1} \to \mathbb{R}^{d+1}$ such that the Latent Neural ODE $\Psi \stackrel{\text{def.}}{=} \bigoplus_{i=1}^{D} (\pi_1^{d+1} \circ \Psi_i \circ \iota_d^{d+1})$ satisfies*

$$\big\| f(x) - \Psi \big\|_{L^\infty(\mathbb{R}^d) l^\infty} \in O\big(\tfrac{1}{n}\big).$$

Moreover, Ψ is Lipschitz and the ReLU MLPs parameterizing the vector fields defining each Ψ_i, have width $\mathcal{O}(dn^{d+1})$ and depth $\mathcal{O}(\log_2(d))$ and $\mathcal{O}(dn^{d+1})$ non-zero parameters for each $i \in [D]$.

Thus, Corollary 11 provides quantitative and, more topologically more explicit, version of the very recent qualitative result of De Marinis et al. (2025) as a direct consequence of our main theorem.

Remark 12 (Optimality of the $\mathcal{O}(n^{d+1})$ rate) *The question of the optimality of the rate $\mathcal{O}(n^{d+1})$ is not known, as there are no available approximation theoretic lower bounds, nor tools for establishing lower bounds, in our non-vector space setting it is currently unknown. However, is likely that a rate of $\mathcal{O}(n^d)$ cannot be achieved while requiring that the core of the model is the conjugation of a homeomophism by linear maps, due to the need to lift in order to approximate general continuous functions.*

Corollary 13 (Universal Approximation) *Let $d, D \in \mathbb{N}_+$. The set $\mathcal{F} \subseteq C([0,1]^d, \mathbb{R}^D)$ of all maps of the form (15) where $f : \mathbb{R}^d \to \mathbb{R}^D$ is a ReLU MLP of arbitrary depth and width is dense in $C([0,1]^d, \mathbb{R}^D)$.*

Discussion: Only one flow is enough when lifting If f were smooth, then the topological embedding $x \mapsto (x, f(x))$ defines a *differentiable d-cell*, in the sense of (Palais, 1960, page 274). By (Palais, 1960, Theorem C), this embedding extends to a compactly supported†, orientation-preserving homeomorphism $\Psi : \mathbb{R}^{d+D} \to \mathbb{R}^{d+D}$, which can then be approximated using our main positive result (Theorem 4). Projecting away the second coordinate, as in (15), yields an approximation of f. Moreover, the smoothness assumption is not restrictive: Compactly supported† homeomorphisms are identity out of their support, this means that there exists a r such that they are all identity on the boundary of the disc D^r. By Alexander's trick, this means they are all isotopic to identity. As a compactly supported† homeomorphism of $\mathbb{R}^d$ for $d \geqslant 5$ can be approximated uniformly by compactly supported† diffeomorphisms if and only if it is isotopic to a diffeomorphism by Müller (2014) this shows in our case working on diffeomorphisms in high dimension is not restrictive. A similar approach was recently explored in Puthawala et al. (2022).
We highlight that this line of reasoning misses the central insight of the flow-based construction in (15). Namely, it does not specify how *many flows* are required to represent Ψ. In contrast, our construction

shows that only one flow suffices if we allow for linear *lifting/projecting*, thereby *breaking the homeomorphism structure* in (1).

We examine the implications our results to classical generative modelling, in the sense of Wasserstein GANs Arjovsky et al. (2017); Korotin et al. (2021), and highlighting the differences and similarities.

4.2 Universal Approximation of measures satisfying Caffarelli Conditions

We first recall that for any hyperparameter $1 \leqslant p < \infty$, we denote by $\mathcal{P}_p([0,1]^d)$ the subset of probabilities that finitely integrate $x \mapsto \|x\|^p$. We equip $\mathcal{P}_p([0,1]^d)$ with the Wasserstein p-distance $\mathcal{W}_p$, that is, for $\mu, \nu \in \mathcal{P}_p([0,1]^d)$, the metric defined by

$$\mathcal{W}_p(\mu,\nu)^p \stackrel{\text{def.}}{=} \inf_{\pi \in \mathrm{Cpl}(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\pi}\big[\|X-Y\|^p\big],$$

where $\mathrm{Cpl}(\mu,\nu) \stackrel{\text{def.}}{=} \{\pi \in \mathcal{P}([0,1]^d \times [0,1]^d) \colon \pi \text{ has first marginal } \mu, \text{ second marginal } \nu\}$. Given any $n, m \in \mathbb{N}_+$, any Borel map $g : [0,1]^n \to [0,1]^m$, and any probability measure $\nu \in \mathcal{P}([0,1]^n)$ we recall that the *pushforward measure* is $g_\sharp \nu \stackrel{\text{def.}}{=} \nu(g^{-1}[\cdot])$ belongs to $\mathcal{P}([0,1]^m)$. If g is Lipschitz and ν belongs to the Wasserstein space $\mathcal{P}_p([0,1]^n)$ then so does $g_\sharp \nu \in \mathcal{P}_p([0,1]^m)$. We require the following standard regularity condition.

Assumption 4.1 (Caffarelli Conditions; cf. Caffarelli (1996)) *Let μ and ν be measures which are absolutely continuous with respect to the uniform measure U on $(0,1)^d$; and whose Radon-Nikodym densities $\frac{d\mu}{dU}$ and $\frac{d\nu}{dU}$ are α-Hölder for some $\alpha \in (0,1)$, and are bounded (above and below on $[0,1]^d$).*

Armed with these definitions we are ready to show that incremental flow-based generation, augmented by lifts, are universal generative models in the more general but less structured context where GANs are typically studied; Lu and Lu (2020) with related guarantees in Biau et al. (2020).

Corollary 14 *Let $2 < d \in \mathbb{N}_+$ and μ, ν be probability measures on $[0,1]^d$ satisfying Assumption 4.1. For every $\varepsilon, \delta > 0$, there exists a large enough $N \in \mathbb{N}_+$ and a Latent Neural ODE $\Psi : \mathbb{R}^{d+D} \to \mathbb{R}^{d+D}$ such that for all i.i.d. random variables $Z_1 \sim \cdots \sim Z_N$, with law ν (defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$)*

$$\mathbb{P}\Big(\mathcal{W}_1(\mu, \Psi_\sharp \nu^{(N)}) \leqslant L^\Psi \frac{\sqrt{d} C_{[0,1]^d}}{\sqrt[d]{N}} + \delta + \varepsilon\Big) \geqslant 1 - 2\, e^{\frac{-2N\delta^2}{d(L^\Psi)^2}}$$

where $\nu^{(N)} \stackrel{\text{def.}}{=} \frac{1}{N}\sum_{n=1}^N \delta_{Z_n}$ is the empirical distribution induced from the random sample and $C_{[0,1]^d}$ is a constant.

5 Conclusion

In this article, we have shown that incremental generation is both necessary and sufficient for universal flow-based approximation, while non-incremental flows form a meagre, non-universal subset. By coupling dynamical, algebraic, and approximation-theoretic tools, we established quantitative rates and demonstrated that lifted incremental flows yield structured universality for both functions and probability measures.

Acknowledgments and Disclosure of Funding

The authors would like to thank Paulo Varandas for his encouraging suggestion for proving the meagerness of the set of flows and to Giulia Livieri for her incredibly helpful feedback during the final stages of the manuscript. A. Kratsios and H. Rouhvarzi would like to acknowledge financial support from an NSERC Discovery Grant No. RGPIN-2023-04482 and No. DGECR-2023-00230. They also acknowledge that resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute[5].

5. https://vectorinstitute.ai/partnerships/current-partners/

Appendix A. Proof of the Negative Result

Proposition 15 (Existence of a recurrent point) *For every compact set M and any $f \in \mathcal{H}_d(M)$ there exists a recurrent point in M.*

Proof : Take $f \in \mathcal{H}_d(M)$. Since M is compact the set

$$\mathcal{A} = \{S \subseteq M : S \neq \emptyset, S \text{ closed}, f(S) = S\}$$

is non-empty (e.g. $M \in \mathcal{A}$). Order $\mathcal{A}$ by inclusion. Any totally ordered chain $\{S_i\}$ in $\mathcal{A}$ has

$$\bigcap_i S_i \neq \emptyset$$

by compactness, and $\bigcap_i S_i$ is closed and f-invariant. By Zorn's Lemma there is a minimal element $S_{\min} \in \mathcal{A}$, i.e. a nonempty closed invariant set containing no proper nonempty closed invariant subset. Take any $x \in S_{\min}$. If x was not recurrent then there would exist a neighborhood $U \ni x$ and an integer N such that

$$f^n(x) \notin U \quad \forall n \geqslant N.$$

Set

$$S' = \overline{\{ f^n(x) : n \geqslant N\}}.$$

Then S' is nonempty, closed, and f-invariant, but $x \notin S'$, so $S' \subsetneq S_{\min}$, contradicting minimality. Thus x must be recurrent. Since x was arbitrary in $S_{\min}$, every point of $S_{\min}$ is recurrent. ■

In order to prove Lemma 18 We restate two lemmas for the convenience of the reader. For the proof refer to Lemma 9 and 13 of Nitecki and Shub (1975) :

Lemma 16 *Given $\varepsilon > 0$ and a flow ϕ. Suppose γ is a C^1 curve in M (an embedded closed interval or circle) such that at each point x in the image of γ one of the following conditions holds:*

(i) $\|\dot{\phi}(x)\| < \varepsilon/2$, *or*

(ii) $x \notin \mathcal{Z}(\phi)$, *and γ has inclination* $\sigma < \varepsilon/\|\dot{\phi}\|$ *at* x.

Then, given any neighborhood U of the image of γ, there exists a flow ψ on M satisfying:

(a) $\dot{\psi} = \dot{\phi}$ *off* U,

(b) $\|\dot{\psi} - \dot{\phi}\| < \varepsilon$ *on* M,

(c) γ *is (a segment of an) integral curve of* ψ.

Lemma 17 *Let M be a manifold of dimension $\geqslant 2$ with distance d coming from a Riemannian metric. Suppose a finite collection $\{(p_i, q_i) \in M \times M : i = 1, \ldots, k\}$ of pairs of points of M is specified, together with a small positive constant $\delta > 0$ such that:*

(a) *For each* i, $d(p_i, q_i) < \delta$.

(b) *If* $i \neq j$, *then* $p_i \neq p_j$ *and* $q_i \neq q_j$.

Then there exists $f \in \mathrm{Diff}(M)$ *such that*

(i) $d\big(f(x), x\big) < 2\pi\delta$ *for every* $x \in M$.

(ii) $f(p_i) = q_i$ *for* $i = 1, \ldots, k$.

Lemma 18 (C^0-closing lemma for compactly supported functions) *Let $f\colon M \to M$ be a compactly supported† homeomorphism on the compact set $K \subset M$, where M is a $\mathcal{C}^\infty$ manifold of dimension $\geqslant 2$ with distance d coming from a Riemannian metric, and let $x_0 \in M$ be a non-wandering point. Then for every $\varepsilon > 0$ and compact set P s.t. $K \subset int(P)$ (K is compactly embedded in P) there exists a compactly supported† homeomorphism $g\colon M \to M$ on P such that x_0 is a periodic point of g and*

$$d_1(f,g) < \varepsilon$$

also, if M is orientable and f is orientation-preserving, then g is also orientation-preserving.

Proof : The Lemma is trivial when x_0 is a fixed point or a periodic point. Consider not.
If M is compact and $K = M$ by (Kwiecińska, 1996, Theorem 4) the result follows. Suppose not.

As f is compactly supported†, $\mathcal{O}_f(x_0)$ is a subset of K. Note f is uniformly continuous on P and Lemma 16 and Lemma 17 are true on any Riemannian manifold like M so one can apply the same proof of (Kwiecińska, 1996, Theorem 4) which results the desirable function $g = f \circ h$ which h is isotopic to the identity (considering the flow time as the isotopy parameter) which means h is orientation-preserving. As a result, if f is orientation-preserving then so is g.
Also, Lemma 16 property (a) makes the perturbed function in Lemma 17 to have support on a finite number of neighborhoods of the curves joining p_i to q_i. As h is the result of Lemma 17 take η and those neighborhoods small enough such that these neighborhoods stay inside P. This makes the function h to be compactly supported† on P. As a result, g must be compactly supported† on P. ■

We are now in a position to prove our main result in *topological dynamics* which will imply our main negative result (namely Theorem 1).
Proof of Theorem 2:

Choose arbitrary $f \in \mathcal{H}_d([0,1]^d)$. By Proposition 15 (for $M = [0,1]^d$) there exists $x \in R(f)$. (cf. (Wescley Bonomo, 2020, Theorem A)) First consider $x \in R(f)\backslash \mathrm{Fix}(f)$. By Lemma 18, there exists an $\frac{\varepsilon}{2}$-C^0-purterbation $f_1 = f \circ h$ such that $f_1^k(x) = x$ for $k > 1$ and $f_1 \in \mathcal{H}_d(P)$. Reduce G so that $\mathcal{B} = \{B(f_1^i(x), G) : 1 \leqslant i \leqslant k\}$ become pairwise disjoint collection of balls inside P. Using bump-function, one can attain homeomorphism f_2 s.t. it is C^1-smooth around a neighborhood of $\mathcal{O}_{f_2}(x)$ and $f_2^i(x)$ is a periodic attractor for f_2^k for $i \in \{1,\ldots,k\}$ (one can also use mollifiers to do so) and one gets $d_{C^0}(f_2, f_1) = \|f_2 - f_1\|_{C^0} < \frac{\varepsilon}{2}$. $\{f_2\}$ is a dense set of homeomorphisms compactly supported† on P with (non-fixed) periodic points of period $k > 1$ s.t. $\mathcal{O}_f(p)$ is not contained in a continuum in $Per_k(f)$ so f_2 can't be written as a time-1 map of a flow[6] with the derivative of the same order.

Now consider $x \in R(f) = \mathrm{Fix}(f)$. Take G small enough $0 < G < \varepsilon$ s.t. the exponential map $\exp_x : \overline{D}(G) \to B(z, G)$, (which is defined as $\exp_x(v) = x + v$) [7] has its range inside P where $\overline{D}(G) := \{v \in \mathbb{R}^d : \|v\| < G\}$. Rewrite $f(z) = \exp_x \circ F \circ \exp_x^{-1}(z)$ where $F(v) = \exp_x^{-1} \circ f \circ \exp_x(v)$ where $v \in \overline{D}(G)$. Define C^1-bump function $\rho|_{B(x,\frac{G}{2})} = 1$, $\rho|_{\mathbb{R}^d \backslash B(x,G)} = 0$. Define:

$$f_1(z) \stackrel{\text{def.}}{=} \exp_x \circ [\rho \cdot R + (1-\rho) \cdot F] \circ \exp_x^{-1}(z)$$

on $z \in B(x,G)$ and $f_1 = f$ for $z \in M \backslash \overline{D}(G)$, where $R \neq Id$ is an involution of $\overline{D}(G)$. f_1 is an orientation-preserving homeomorphism as it is a convex combination of two functions in the same connected component of invertible group $\mathrm{GL}^+(n)$.
As $D\exp_x(0) = Id$ there exists $L > 1$ tending to 1 as $G \to 0$ s.t. :

$$d(f_1(z), f(z)) \leqslant L\|(F - R) \circ \exp_x^{-1}(z)\| \leqslant L\|F - R\|G \leqslant L\|F - R\|\varepsilon \tag{16}$$

can do the same in case one for f_1 as it has period-two periodic point because of R. This proves the density. For each homeomorphism f_2 in case one there exists small enough neighborhood $U \ni x$ s.t. $f^k_{2\,|U}$ has no fixed point on ∂U (note : $f_2^k(x) = x$ and $x \in \mathrm{R}(f)\backslash \mathrm{Fix}(f)$).

6. Here the flow is the usual group action of the additive group of real numbers on the manifold
7. For a Riemannian manifold M it is defined as $\exp : M \times T_xM \to M$ for more read (Gallot et al., 2004, Section 2.C)

For every C^0-Perturbation g of f_2 in $\mathcal{H}_d(P)$ one has: $0 \neq \deg(f_2^k(x) - x, 0) = \deg(g^k(x) - x, 0)$. There exists a point z s.t. $g^k(z) - z = 0$ so g^k has a fixed point. Thus, g has a periodic point of period $k > 1$ that is not contained in a continuum subset of the set of k-periodic points. It means f_2 has an open neighborhood. This proves the openness. ■

Proof of Theorem 1: Since every neural ODE is, by definition, a flowable homeomorphisms supported† on any such $[\delta, 1-\delta]^d$ then our next results necessitates that the set of neural ODEs is itself meager in $\mathcal{H}_d([0,1]^d)$ as well. ■

Appendix B. Proof of the Positive Result

First we handle "Quantitative Formulations" and so we can easily prove the theorem in "Qualitative Universal Approximation Guarantee" section.

B.1 Quantitative Formulations

B.1.1 General Case

First we prove Proposition 6 to show it is reasonable to think of incremental flow-based generation.

Proof of Proposition 6: Consider vector fields $V'(x,y) = (-\pi(y - \frac{1}{2}), \pi(x - \frac{1}{2}))$ and $W'(x,y) = (-x + \frac{1}{2}, 0)$ which are π rotation around the point $p = (\frac{1}{2}, \frac{1}{2})$ and squeeze to the line $x = \frac{1}{2}$ (call this line l). Consider the bump function ρ which is 1 on the circle with radius $r = \frac{1}{8}$ around p and zero out of the circle with radius $R = \frac{1}{4}$. $V = V' \circ \rho, W = W' \circ \rho$ are compactly supported on $[0,1]^d$. Denote time one map of V, W as φ_V, φ_W. Then $\varphi_V \circ \varphi_W \notin \text{Flow}$. suppose not, i.e. there exists a vector field F which its time one flow φ_F is equal to $\varphi_V \circ \varphi_W$.
Note that for the composition $\varphi_V \circ \varphi_W$ any point $a \in B_r(p) \backslash l$ converges to zero on iteration but any $b \in l$ has periodic point two. Consider the trajectory γ of the flow of $\varphi_V \circ \varphi_W$. For any point $c = r_c e^{i\theta}$ (considering polar coordinate) on a neighborhood S of the point $\varphi_V \circ \varphi_W(b)$ on γ, radius of $\varphi_V \circ \varphi_W(c)$ will be smaller than r_c as a result radius of any point other than b on $(\varphi_V \circ \varphi_W)^2(S)$ will be smaller but $(\varphi_V \circ \varphi_W)^2(b) = b$ so there will be a discontinuity at the point b but all periodic orbits of a continuous flow form a continuum (i.e. a compact and connected metric space that contains at least two points) which is a contradiction. ■

Now, we first show how one can approximate a flow, then we consider the composition of flows and prove Theorem 7.

Lemma 19 (Transfer: Universal Approximation of Vector Field to Flow) *Let $n \in \mathbb{N}$ and $V \in \mathcal{X}^0$, then for every $\varphi = \text{Flow}(V)$ there exists a ReLU MLP $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ compactly supported on $(0,1)^d$ with width $8d(n+1)^d$, depth $\lceil log_2 d \rceil + 6$, and at-most $16d(n+1)^d + 9$ non-zero parameters such that the flow (Neural ODE) given for each $x \in \mathbb{R}^d$ by*

$$\begin{aligned} \Psi(x) &= z_1^x \\ z_t^x &= x + \int_0^t \Phi(z_s^x)\, ds \end{aligned} \tag{17}$$

for $0 \leqslant t \leqslant 1$; satisfies the uniform estimates

$$\|\Psi(x) - \varphi(x)\|_{L^\infty(\mathbb{R}^d)l^\infty} \leqslant 2\|\omega(\frac{d}{2n})\|_{l^\infty} e^{L^V} \tag{18}$$

Where ω is the modulus of regularity of φ which is equal to the modulus of regularity of Φ. Note that e^{L^V} does not depend on n. One gets convergence as $n \to \infty$

Proof : If we define $(V)_j := V_j$ then as V_j is Lipschitz and nonzero on a compact subset from (Hong and Kratsios, 2024) there exists a ReLU MLP $(\phi(z_s^x))_j := \phi_j(z_s^x)$ and modulus of regularity $(\omega)_j := \omega_j$ such that:

$$\|V_j(z_s^x) - \phi_j(z_s^x)\|_{L^\infty([0,1]^d)} \leqslant \omega_j(\frac{d}{2n})$$

To make it compactly supported on $[\frac{\delta}{4}, 1 - \frac{\delta}{4}]^d$ for a $\delta \in (0, 2)$ we compose it componentwise with ReLU bump function below:

$$\begin{aligned} b(x) &= \sigma_{ReLU}\big(2\sigma_{ReLU}(x - \frac{\delta}{4}) - \sigma_{ReLU}(x - \frac{\delta}{2}) - \frac{1}{\delta}\sigma_{ReLU}(x - (1 - \frac{\delta}{2}))\big) \\ &= \begin{cases} 0, & \text{if } x < \frac{\delta}{4} \\ 2x - \frac{\delta}{2} & \text{if } x \in [\frac{\delta}{4}, \frac{\delta}{2}] \\ x, & \text{if } x \in [\frac{\delta}{2}, 1 - \frac{\delta}{2}] \\ (1 - \frac{1}{\delta})x + (\frac{2-\delta}{2\delta}), & \text{if } x \in [1 - \frac{\delta}{2}, 1 - \frac{\delta}{4}] \\ 0, & \text{if } x > 1 - \frac{\delta}{4} \end{cases} \end{aligned}$$

implemented by the following network:

$$\begin{aligned} \mathbf{x} &\Rightarrow \begin{bmatrix} \sigma_{ReLU}(x - \frac{\delta}{4}) \\ \sigma_{ReLU}(x - \frac{2\delta}{4}) \\ \sigma_{ReLU}(x - (1 - \frac{2\delta}{4})) \end{bmatrix} \\ &\Rightarrow \big[\sigma_{ReLU}\big(2\sigma_{ReLU}(x - \frac{\delta}{4}) - \sigma_{ReLU}(x - \frac{2\delta}{4}) - \frac{1}{\delta}\sigma_{ReLU}(x - (1 - \frac{2\delta}{4}))\big)\big] \\ &\Rightarrow \big[\phi_j\big(\sigma_{ReLU}\big(2\sigma_{ReLU}(x - \frac{\delta}{4}) - \sigma_{ReLU}(x - \frac{2\delta}{4}) - \frac{1}{\delta}\sigma_{ReLU}(x - (1 - \frac{2\delta}{4}))\big)\big)\big] = \phi_j\big(b(x)\big) \\ &\Rightarrow \Phi_j(x) \end{aligned}$$

which is 2-Lipschitz and supported on $[0, 1]$. The result $\Phi_j = \phi_j \circ b$ has width:

$$\begin{aligned} \text{width}(\Phi_j) &= \max\{3d, d, \text{width}(\phi_j)\} \\ &= \max\{3d, d, 8d(n+1)^d\} \\ &= 8d(n+1)^d \\ &= \text{width}(\phi_j) \end{aligned}$$

And depth:

$$\text{depth}(\phi_j \circ b) = \text{depth}(\phi_j) + 2 = \lceil log_2 d \rceil + 6$$

then note that $\Phi_j = \phi_j$ on $[\frac{\delta}{2}, 1 - \frac{\delta}{2}]^d$ so:

$$\|\Phi_j - V_j\|_{L^\infty([\frac{\delta}{2}, 1-\frac{\delta}{2}]^d)} \leqslant \|\phi_j - V_j\|_{L^\infty([0,1]^d)} \leqslant \omega_j(\frac{d}{2n})$$

And:

$$\|\Phi_j - V_j\|_{L^\infty([0,\frac{\delta}{2}]^d)} \leqslant \frac{\delta}{2}(\|V_j\|_{L^\infty([0,1]^d)} + \|\phi_j\|_{L^\infty([0,1]^d)})$$

Then:

$$\begin{aligned} \|\Phi_j - V_j\|_{L^\infty([0,1]^d)} &\leqslant \|\Phi_j - V_j\|_{L^\infty([0,\frac{\delta}{2}]^d)} + \|\Phi_j - V_j\|_{L^\infty([\frac{\delta}{2},1-\frac{\delta}{2}]^d)} + \|\Phi_j - V_j\|_{L^\infty([1-\frac{\delta}{2},1]^d)} \\ &\leqslant \omega_j(\frac{d}{2n}) + \delta(\|V_j\|_{L^\infty([0,1]^d)} + \|\phi_j\|_{L^\infty([0,1]^d)}) \end{aligned}$$

Note that $\text{supp}(\Phi) = \text{supp}(V) = [0,1]^d$ so we can consider the the inequalities above on $\mathbb{R}^d$ then we'll have:

$$\begin{aligned}\|V(x_s^x) - \Phi(z_s^x)\|_{L^\infty(\mathbb{R}^d)l^\infty} &\leqslant \|V(x_s^x) - V(z_s^x)\|_{L^\infty(\mathbb{R}^d)l^\infty} + \|V(z_s^x) - \Phi(z_s^x)\|_{L^\infty(\mathbb{R}^d)l^\infty} \\ &\leqslant L^V \|x_s^x - z_s^x\|_{L^\infty(\mathbb{R}^d)l^\infty} + \|\omega(\frac{d}{2n})\|_{l^\infty} + \delta(\|V\|_{L^\infty(\mathbb{R}^d)l^\infty} + \|\phi\|_{L^\infty(\mathbb{R}^d)l^\infty})\end{aligned}$$

Note that as ϕ and V are continuous on a compact set, they attain their supremum. This means there exists a constant C such that:

$$\|V\|_{L^\infty(\mathbb{R}^d)l^\infty} + \|\phi\|_{L^\infty(\mathbb{R}^d)l^\infty} \leqslant C$$

Observe that, $C \geqslant 0$ not only depends on V but also on n and on d, as ϕ depends only on V and n and on d; hence by Jensen's inequality:

$$\begin{aligned}\|z_t^x - x_t^x\|_{L^\infty(\mathbb{R}^d)l^\infty} &= \|\int_0^t V(x_s^x) - \Phi(z_s^x) ds\|_{L^\infty(\mathbb{R}^d)l^\infty} \\ &\leqslant \int_0^t \|V(x_s^x) - \Phi(z_s^x)\|_{L^\infty(\mathbb{R}^d)l^\infty} ds \\ &\leqslant L^V \int_0^t \|x_s^x - z_s^x\|_{L^\infty(\mathbb{R}^d)l^\infty} ds + \left(\|\omega(\frac{d}{2n})\|_{l^\infty} + \delta C\right) t\end{aligned}$$

On the other hand, observe that δ does not depend on any of V, n, d and it was arbitrary; thus, we may take $\delta C \leqslant \|\omega(\frac{d}{2n})\|_{l^\infty}$ thus:

$$\|z_t^x - x_t^x\|_{L^\infty(\mathbb{R}^d)l^\infty} \leqslant L^V \int_0^t \|x_s^x - z_s^x\|_{L^\infty(\mathbb{R}^d)l^\infty} ds + 2\|\omega(\frac{d}{2n})\|_{l^\infty} t$$

By applying Grönwall's inequality (Pachpatte (1998) Theorem 1.3.1)

$$\|z_t^x - x_t^x\|_{L^\infty(\mathbb{R}^d)l^\infty} \leqslant 2\|\omega(\frac{d}{2n})\|_{l^\infty} t e^{L^V t}$$

let $t = 1$:

$$\|\Psi(x) - \varphi(x)\|_{L^\infty(\mathbb{R}^d)l^\infty} = \|z_1^x - x_1^x\|_{L^\infty(\mathbb{R}^d)l^\infty} \leqslant 2\|\omega(\frac{d}{2n})\|_{l^\infty} e^{L^V} \tag{19}$$

The right hand side goes to zero as $n \to \infty$. Finally, observe that e^{L^V} depends only on V and *not* on the approximation parameter $n \in \mathbb{N}_+$. ■

Now we are ready to prove the Theorem 7.

Proof of Theorem 7: As $\varphi \in \mathcal{H}_d^{G,s}([0,1]^d)$ in (7) define $\varphi_g \overset{\text{def.}}{=} \text{Flow}(V_g)$ for every $g \in [G]$. So one can write:

$$\varphi = \bigcirc_{g=1}^G \varphi^{(g)}(x) \tag{20}$$

We use Lemma 19 for each $\varphi^{(g)}$ to find Neural ODEs $\{\Psi^{(g)}\}_{g=1}^G$. One can write $\Psi^{(g)}(x) = z_1^x$ as in (17), then as $\Phi^{(g)}$ is Lipschitz we can write:

$$\|z_t^x - z_t^y\|_{l^\infty} \leqslant \|x - y\|_{l^\infty} + \int_0^t \|\Phi^{(g)}(z_s^x) - \Phi^{(g)}(z_s^y)\|_{l^\infty} ds \leqslant \|x - y\|_{l^\infty} + L^{\Phi^{(g)}} \int_0^t \|z_s^x - z_s^y\|_{l^\infty} ds$$

By applying Grönwall's inequality (Pachpatte (1998) Theorem 1.2.2)

$$\|z_t^x - z_t^y\|_{l^\infty} \leqslant (\|x - y\|_{l^\infty}) e^{L^{\Phi^{(g)}} t}$$

Let $t = 1$:

$$\|\Psi^{(g)}(x) - \Psi^{(g)}(y)\|_{l^\infty} = \|z_1^x - z_1^y\|_{l^\infty} \leqslant \|x - y\|_{l^\infty} e^{L^{\Phi^{(g)}}} \tag{21}$$

Define $\Psi \overset{\text{def.}}{=} \bigcirc_{g=1}^{G} \Psi^{(g)}$. We try to show that the Lipschitz constant L^{Ψ} is at-most $\prod_{g=1}^{G} e^{L^{\Phi^{(g)}}}$ using induction. The base case is already proved in inequality (21); for the step of the induction, consider the induction assumption bellow :

$$\| \bigcirc_{g=1}^{G-1} \Psi^{(g)}(x) - \bigcirc_{g=1}^{G-1} \Psi^{(g)}(y)\| \leqslant \prod_{g=1}^{G-1} e^{L^{\Phi^{(g)}}} \|x-y\|$$

Using (21) we can write:

$$\| \bigcirc_{g=1}^{G} \Psi^{(g)}(x) - \bigcirc_{g=1}^{G} \Psi^{(g)}(y)\| \leqslant e^{L^{\Phi^{(G)}}} \| \bigcirc_{g=1}^{G-1} \Psi^{(g)}(x) - \bigcirc_{g=1}^{G-1} \Psi^{(g)}(y)\| \leqslant \prod_{g=1}^{G} e^{L^{\Phi^{(g)}}} \|x-y\| \tag{22}$$

So one can deduce the Lipschitz constant is at-most $\prod_{g=1}^{G} e^{L^{\Phi^{(g)}}}$.

We again use induction to show inequality (9). The base case is already proved in Lemma 19. For the step of the induction, consider the induction assumption bellow:

$$\| \bigcirc_{g=1}^{G-1} \varphi^{(g)}(x) - \bigcirc_{g=1}^{G-1} \Psi^{(g)}(x)\|_{L^\infty(\mathbb{R}^d)l^\infty} \leqslant \sum_{g=1}^{G-1} \left(2\|\omega^{(g)}(\frac{d}{2n})\|_{l^\infty} \prod_{j=g}^{G-1} e^{L^{V^{(j)}}} \right)$$

Using (29) we can write:

$$\begin{aligned}
\|\varphi(x) - \bigcirc_{g=1}^{G} \Psi^{(g)}(x)\|_{L^\infty(\mathbb{R}^d)l^\infty} &= \| \bigcirc_{g=1}^{G} \varphi^{(g)}(x) - \bigcirc_{g=1}^{G} \Psi^{(g)}(x)\|_{L^\infty(\mathbb{R}^d)l^\infty} \\
&\leqslant \| \bigcirc_{g=1}^{G} \varphi^{(g)}(x) - \varphi^{(G)} \circ \bigcirc_{g=1}^{G-1} \Psi^{(g)}(x)\|_{L^\infty(\mathbb{R}^d)l^\infty} \\
&\quad + \|\varphi^{(G)} \circ \bigcirc_{g=1}^{G-1} \Psi^{(g)}(x) - \bigcirc_{g=1}^{G} \Psi^{(g)}(x)\|_{L^\infty(\mathbb{R}^d)l^\infty} \\
&\leqslant \| \bigcirc_{g=1}^{G-1} \varphi^{(g)}(x) - \bigcirc_{g=1}^{G-1} \Psi^{(g)}(x)\|_{L^\infty(\mathbb{R}^d)l^\infty} e^{L^{V^{(G)}}} + 2\|\omega^{(G)}(\frac{d}{2n})\|_{l^\infty} e^{L^{V^{(G)}}}
\end{aligned}$$

Where for the last inequality we used (18) and (21).
Then by the induction assumption:

$$\|\varphi(x) - \bigcirc_{g=1}^{G} \Psi^{(g)}(x)\|_{L^\infty(\mathbb{R}^d)l^\infty} \leqslant \sum_{g=1}^{G} \left(2\|\omega^{(g)}(\frac{d}{2n})\|_{l^\infty} \prod_{j=g}^{G} e^{L^{V^{(j)}}} \right)$$

By Lemma 19 the right hand side goes to zero as $n \to \infty$. We have thus concluded the proof. ■

B.1.2 Differentiable Case

We do the same steps for the differentiable case but with another ReLU MLP which gives faster approximation rates. Also we first prove a proposition which Proposition 8 is a special case of it.

Lemma 20 (Transfer: Universal Approximation of Vector Field to Flow (Differentiable Case)) *Let $s \in \mathbb{N}_+$, and $V = (V_1, \dots, V_d) \in \mathcal{X}^s$, and $\omega_j \overset{\text{def.}}{=} 85(s+1)^d 8^s \|V_j\|_{C^s([0,1]^d)} (N_j L)^{-2s/d}$ for $1 \leqslant j \leqslant d$ and $\omega \overset{\text{def.}}{=} (\omega_1, \dots, \omega_d)$ then for every $\varphi = \text{Flow}(V)$ and for any $N_j, L \in \mathbb{N}^+$, $1 \leqslant j \leqslant d$, There exists a ReLU MLP*

$$\Phi : \mathbb{R}^d \to \mathbb{R}^d$$

compactly supported in $(0,1)^d$ with width less than or equal $\sum_{j=1}^{d} 17 s d^{d+13} d(N_j+2)\log_2(8N_j)$ and depth $18s^2(L+2)\log_2(4L) + 2(d+1)$ such that the flow (Neural ODE) given for each $x \in \mathbb{R}^d$ by

$$\begin{aligned}
\Psi(x) &= z_1^x \\
z_t^x &= x + \int_0^t \Phi(z_s^x)\, ds
\end{aligned} \tag{23}$$

for $0 \leqslant t \leqslant 1$*; satisfies the uniform estimates*

$$\|\Psi(x) - \varphi(x)\|_{L^\infty(\mathbb{R}^d)l^\infty} \leqslant 2\|\omega(N_1, \ldots, N_d, L)\|_{l^\infty} e^{L^V} \tag{24}$$

where Ψ *is also compactly supported*† *in* $(0,1)^d$.

Proof : If we define $(V)_j := V_j$ for $1 \leqslant j \leqslant d$ then by (Lu et al., 2021, Theorem 1.1) for any $N_j, L \in \mathbb{N}^+$, there exists a function ϕ_j implemented by a ReLU FNN with width $C_1(N_j+2)\log_2(8N_j)$ and depth $C_2(L+2)\log_2(4L) + 2d$ such that

$$\|\phi_j - V_j\|_{L^\infty([0,1]^d)} \leqslant C_3\|V_j\|_{C^s([0,1]^d)}(N_j L)^{-2s/d}, \tag{25}$$

where $C_1 = 17s^{d+1}3^d d$, $C_2 = 18s^2$, and $C_3 = 85(s+1)^d 8^s$.
Let $\delta \in (0,2)$, one can make ϕ compactly supported on $[\frac{\delta}{4}, 1-\frac{\delta}{4}]^d$ by composing it componentwise with ReLU bump function below:

$$\begin{aligned} b(x) &= \sigma_{ReLU}\big(2\sigma_{ReLU}(x - \frac{\delta}{4}) - \sigma_{ReLU}(x - \frac{\delta}{2}) - \frac{1}{\delta}\sigma_{ReLU}(x - (1 - \frac{\delta}{2}))\big) \\ &= \begin{cases} 0, & \text{if } x < \frac{\delta}{4} \\ 2x - \frac{\delta}{2} & \text{if } x \in [\frac{\delta}{4}, \frac{\delta}{2}] \\ x, & \text{if } x \in [\frac{\delta}{2}, 1 - \frac{\delta}{2}] \\ (1 - \frac{1}{\delta})x + (\frac{2-\delta}{2\delta}), & \text{if } x \in [1 - \frac{\delta}{2}, 1 - \frac{\delta}{4}] \\ 0, & \text{if } x > 1 - \frac{\delta}{4} \end{cases} \end{aligned}$$

implemented by the following network:

$$\begin{aligned} \mathbf{x} &\Rightarrow \begin{bmatrix} \sigma_{ReLU}(x - \frac{\delta}{4}) \\ \sigma_{ReLU}(x - \frac{2\delta}{4}) \\ \sigma_{ReLU}(x - (1 - \frac{2\delta}{4})) \end{bmatrix} \\ &\Rightarrow \big[\sigma_{ReLU}\big(2\sigma_{ReLU}(x - \frac{\delta}{4}) - \sigma_{ReLU}(x - \frac{2\delta}{4}) - \frac{1}{\delta}\sigma_{ReLU}(x - (1 - \frac{2\delta}{4}))\big)\big] \\ &\Rightarrow \big[\phi_j\big(\sigma_{ReLU}\big(2\sigma_{ReLU}(x - \frac{\delta}{4}) - \sigma_{ReLU}(x - \frac{2\delta}{4}) - \frac{1}{\delta}\sigma_{ReLU}(x - (1 - \frac{2\delta}{4}))\big)\big)\big] = \phi_j\big(b(x)\big) \\ &\Rightarrow \Phi_j(x) \end{aligned}$$

which is 2-Lipschitz and supported on $[0,1]$. The result $\Phi_j = \phi_j \circ b$ has width:

$$\begin{aligned} \text{width}(\Phi_j) &= \max\{3d, d, \text{width}(\phi_j)\} \\ &= \max\{3d, d, 17sd^{d+13}d(N_j+2)\log_2(8N_j)\} \\ &= 17sd^{d+13}d(N_j+2)\log_2(8N_j) \\ &= \text{width}(\phi_j) \end{aligned}$$

And depth:

$$\text{depth}(\phi_j \circ b) = \text{depth}(\phi_j) + 2 = 18s^2(L+2)\log_2(4L) + 2(d+1)$$

and as $\text{depth}(\phi_1) = \text{depth}(\phi_2) = \cdots = \text{depth}(\phi_d)$ then $\text{depth}(\Phi_1) = \text{depth}(\Phi_2) = \cdots = \text{depth}(\Phi_d)$, by parallelization (Petersen and Zech, 2024, Proposition 2.3.) the neural network

$$\Phi(x) = (\Phi_1(x), \ldots, \Phi_d(x)) : \mathbb{R}^d \to \mathbb{R}^d$$

has the same depth. The width of it is at least :

$$\text{width}(\Phi) \leqslant \sum_{j=1}^{d} \text{width}(\Phi_j)$$

Set $\omega_j \stackrel{\text{def.}}{=} C_3 \|V_j\|_{C^s([0,1]^d)} (N_j L)^{-2s/d}$, $\phi \stackrel{\text{def.}}{=} (\phi_1, \ldots, \phi_d)$ and $\omega \stackrel{\text{def.}}{=} (\omega_1, \ldots, \omega_d)$ then note that $\Phi_j = \phi_j$ on $[\frac{\delta}{2}, 1 - \frac{\delta}{2}]^d$ so:

$$\|\Phi_j - V_j\|_{L^\infty([\frac{\delta}{2}, 1-\frac{\delta}{2}]^d)} \leqslant \|\phi_j - V_j\|_{L^\infty([0,1]^d)} \leqslant C_3 \|V_j\|_{C^s([0,1]^d)} (N_j L)^{-2s/d},$$

Also on the interval $[0, \frac{\delta}{2}]^d$ one can write:

$$\|\Phi_j - V_j\|_{L^\infty([0,\frac{\delta}{2}]^d)} \leqslant \frac{\delta}{2} (\|V_j\|_{L^\infty([0,1]^d)} + \|\phi_j\|_{L^\infty([0,1]^d)})$$

Then:

$$\begin{aligned}
\|\Phi_j - V_j\|_{L^\infty([0,1]^d)} &\leqslant \|\Phi_j - V_j\|_{L^\infty([0,\frac{\delta}{2}]^d)} + \|\Phi_j - V_j\|_{L^\infty([\frac{\delta}{2},1-\frac{\delta}{2}]^d)} + \|\Phi_j - V_j\|_{L^\infty([1-\frac{\delta}{2},1]^d)} \\
&\leqslant C_3 \|V_j\|_{C^s([0,1]^d)} (N_j L)^{-2s/d} + \delta (\|V_j\|_{L^\infty([0,1]^d)} + \|\phi_j\|_{L^\infty([0,1]^d)})
\end{aligned}$$

Note that $\operatorname{supp}(\Phi) = \operatorname{supp}(V) = [0,1]^d$ so we can consider the the inequalities above on $\mathbb{R}^d$ then we'll have:

$$\begin{aligned}
\|V(x_s^x) - \Phi(z_s^x)\|_{L^\infty(\mathbb{R}^d) l^\infty} &\leqslant \|V(x_s^x) - V(z_s^x)\|_{L^\infty(\mathbb{R}^d) l^\infty} + \|V(z_s^x) - \Phi(z_s^x)\|_{L^\infty(\mathbb{R}^d) l^\infty} \\
&\leqslant L^V \|x_s^x - z_s^x\|_{L^\infty(\mathbb{R}^d) l^\infty} + \|\omega(N_1, \ldots, N_d, L)\|_{l^\infty} \\
&\quad + \delta (\|V\|_{L^\infty(\mathbb{R}^d) l^\infty} + \|\phi\|_{L^\infty(\mathbb{R}^d) l^\infty})
\end{aligned}$$

Note that as ϕ and V are continuous on a compact set, they attain their supremum. This means there exists a constant C such that:

$$\|V\|_{L^\infty(\mathbb{R}^d) l^\infty} + \|\phi\|_{L^\infty(\mathbb{R}^d) l^\infty} \leqslant C.$$

Observe that, $C \geqslant 0$ not only depends on V but also on $N_1, \ldots, N_d$ and on L, as ϕ depends only on V and $N_1, \ldots, N_d$ and on L; hence by Jensen's inequality:

$$\begin{aligned}
\|x_t^x - z_t^x\|_{L^\infty(\mathbb{R}^d) l^\infty} &= \| \int_0^t V(x_s^x) - \Phi(z_s^x) ds \|_{L^\infty(\mathbb{R}^d) l^\infty} \\
&\leqslant \int_0^t \|V(x_s^x) - \Phi(z_s^x)\|_{L^\infty(\mathbb{R}^d) l^\infty} ds \\
&\leqslant L^V \int_0^t \|x_s^x - z_s^x\|_{L^\infty(\mathbb{R}^d) l^\infty} ds + (\|\omega(N_1, \ldots, N_d, L)\|_{l^\infty} + \delta C)\, t
\end{aligned} \tag{26}$$

On the other hand, observe that δ does not depend on any of $V, N_1, \ldots, N_d, L$ and it was arbitrary; thus, we may take $\delta C \leqslant \|\omega(N_1, \ldots, N_d, L)\|_{l^\infty}$ thus:

$$\|x_t^x - z_t^x\|_{L^\infty(\mathbb{R}^d) l^\infty} \leqslant L^V \int_0^t \|x_s^x - z_s^x\|_{L^\infty(\mathbb{R}^d) l^\infty} ds + 2\|\omega(N_1, \ldots, N_d, L)\|_{l^\infty} t \tag{27}$$

By applying Grönwall's inequality (Pachpatte (1998) Theorem 1.3.1)

$$\|z_t^x - x_t^x\|_{L^\infty(\mathbb{R}^d) l^\infty} \leqslant 2\|\omega(N_1, \ldots, N_d, L)\|_{l^\infty} t e^{L^V t}$$

Let $t = 1$:

$$\|\Psi(x) - \varphi(y)\|_{L^\infty(\mathbb{R}^d) l^\infty} = \|z_1^x - x_1^y\|_{L^\infty(\mathbb{R}^d) l^\infty} \leqslant 2\|\omega(N_1, \ldots, N_d, L)\|_{l^\infty} e^{L^V}.$$

■

Proposition 21 (Universal Approximation by Deep Neural ODEs (Differentiable Case)) *Let* $\varphi \in \mathcal{H}_d^{G,s}([0,1]^d)$ *and* $1 \leqslant s$*; then, there exists ReLU neural ODEs* $\{\Psi^{(g)}\}_{g=1}^G \subset \mathcal{X}^0$ *such that* $\Psi \overset{\text{def.}}{=} \bigcirc_{g=1}^G \Psi^{(g)}$ *satisfies the approximation guarantee:*

$$\|\varphi - \Psi\|_{L^\infty(\mathbb{R}^d)l^\infty} \leqslant \sum_{g=1}^G \left(2\|\omega^{(g)}(N_1, \ldots, N_d, L)\|_{l^\infty} \prod_{j=g}^G e^{L^{V^{(j)}}} \right) \tag{28}$$

Where $(\omega^{(g)})_j = 85(s+1)^d 8^s \|V_j^{(g)}\|_{C^s([0,1]^d)} (N_j L)^{-2s/d}$ *each* $\omega^{(g)}$*. The right hand side goes to zero as* $L \to \infty$ *and* $N_j \to \infty$ *for each* $j \in [G]$*.*
Moreover, Ψ *is a compactly-supported*† C^s*-diffeomorphism* t *on* $\mathbb{R}^d$ *supported*† *on* $(0,1)^d$*, with Lipschitz constant at-most* $\prod_{g=1}^G e^{L^{\Psi^{(g)}}}$ *and* $\Psi^{(g)}$ *is a ReLU MLP of width less than or equal to* $\sum_{j=1}^d 17 s^{d+1} 3^d d (N_j^g + 2) \log_2(8N_j^g)$ *and depth* $18s^2(L^g + 2)\log_2(4L^g) + 2(d+1)$ *where* $N_j^g, L^g \in \mathbb{N}^+$ *for each* $1 \leqslant j \leqslant d$*, and* $1 \leqslant g \leqslant G$*.*
In particular, L^Ψ *depends on the parameters* $N_1, \ldots, N_d$*.*

Proof : As $\varphi \in \mathcal{H}_d^{G,s}([0,1]^d)$ in (7) define $\varphi_g \overset{\text{def.}}{=} \text{Flow}\,(V_g)$ for every $g \in [G]$. So one can write:

$$\varphi = \bigcirc_{g=1}^G \varphi^{(g)}(x) \tag{29}$$

We use Lemma 20 for each $\varphi^{(g)}$ to find Neural ODEs $\{\Psi^{(g)}\}_{g=1}^G$. One can write $\Psi^{(g)}(x) = z_1^x$ as in (23), then as $\Phi^{(g)}$ is Lipschitz we can write:

$$\|z_t^x - z_t^y\|_{l^\infty} \leqslant \|x-y\|_{l^\infty} + \int_0^t \|\Phi^{(g)}(z_s^x) - \Phi^{(g)}(z_s^y)\|_{l^\infty} ds \leqslant \|x-y\|_{l^\infty} + L^\Phi \int_0^t \|z_s^x - z_s^y\|_{l^\infty} ds$$

By applying Grönwall's inequality (Pachpatte (1998) Theorem 1.2.2)

$$\|z_t^x - z_t^y\|_{l^\infty} \leqslant (\|x-y\|_{l^\infty})\, e^{L^{\Phi^{(g)}} t}$$

set $t = 1$:

$$\|\Psi^{(g)}(x) - \Psi^{(g)}(y)\|_{l^\infty} = \|z_1^x - z_1^y\|_{l^\infty} \leqslant (\|x-y\|_{l^\infty})\, e^{L^{\Phi^{(g)}}} \tag{30}$$

Define $\Psi \overset{\text{def.}}{=} \bigcirc_{g=1}^G \Psi^{(g)}$. We try to show that the Lipschitz constant L^Ψ is at-most $\prod_{g=1}^G e^{L^{\Phi^{(g)}}}$ using induction. The base case is already proved in inequality (30); for the step of the induction, consider the induction assumption bellow :

$$\|\bigcirc_{g=1}^{G-1} \Psi^{(g)}(x) - \bigcirc_{g=1}^{G-1} \Psi^{(g)}(y)\|_{l^\infty} \leqslant \prod_{g=1}^{G-1} e^{L^{\Phi^{(g)}}} \|x-y\|_{l^\infty}$$

Using (30) we can write:

$$\|\bigcirc_{g=1}^{G} \Psi^{(g)}(x) - \bigcirc_{g=1}^{G} \Psi^{(g)}(y)\|_{l^\infty} \leqslant e^{L^{\Phi^{(G)}}} \|\bigcirc_{g=1}^{G-1} \Psi^{(g)}(x) - \bigcirc_{g=1}^{G-1} \Psi^{(g)}(y)\|_{l^\infty} \leqslant \prod_{g=1}^{G} e^{L^{\Phi^{(g)}}} \|x-y\|_{l^\infty} \tag{31}$$

So one can deduce the Lipschitz constant is at-most $\prod_{g=1}^G e^{L^{\Phi^{(g)}}}$.
Note that one can apply the same procedure of (30) for $\varphi^{(g)}$ and conclude

$$L^{\varphi^{(g)}} = e^{L^{V^{(g)}}} \tag{32}$$

We again use induction to show inequality (34). The base case is already proved in Lemma 20. For the step of the induction consider the induction assumption bellow:

$$\| \bigcirc_{g=1}^{G-1} \varphi^{(g)}(x) - \bigcirc_{g=1}^{G-1} \Psi^{(g)}(x) \|_{L^\infty(\mathbb{R}^d)l^\infty} \leqslant \sum_{g=1}^{G-1} \left(2\|\omega^{(g)}(N_1, \ldots, N_d, L)\|_{l^\infty} \prod_{j=g}^{G-1} e^{L^{V^{(j)}}} \right)$$

Using (29) we can write:

$$\begin{aligned}
\left\|\varphi(x) - \bigcirc_{g=1}^{G} \Psi^{(g)}(x)\right\|_{L^\infty(\mathbb{R}^d)l^\infty} &= \left\| \bigcirc_{g=1}^{G} \varphi^{(g)}(x) - \bigcirc_{g=1}^{G} \Psi^{(g)}(x)\right\|_{L^\infty(\mathbb{R}^d)l^\infty} \\
&\leqslant \left\| \bigcirc_{g=1}^{G} \varphi^{(g)}(x) - \varphi^{(G)} \circ \bigcirc_{g=1}^{G-1} \Psi^{(g)}(x)\right\|_{L^\infty(\mathbb{R}^d)l^\infty} \\
&\quad + \left\|\varphi^{(G)} \circ \bigcirc_{g=1}^{G-1} \Psi^{(g)}(x) - \bigcirc_{g=1}^{G} \Psi^{(g)}(x)\right\|_{L^\infty(\mathbb{R}^d)l^\infty} \\
&\leqslant \left\| \bigcirc_{g=1}^{G-1} \varphi^{(g)}(x) - \bigcirc_{g=1}^{G-1} \Psi^{(g)}(x)\right\|_{L^\infty(\mathbb{R}^d)l^\infty} e^{L^{V^{(G)}}} \\
&\quad + 2\|\omega^{(G)}(N_1, \ldots, N_d, L)\|_{l^\infty} e^{L^{V^{(G)}}}
\end{aligned}$$

Where for the last inequality we used (24) and (32).
Then by the induction assumption:

$$\left\|\varphi(x) - \bigcirc_{g=1}^{G} \Psi^{(g)}(x)\right\|_{L^\infty(\mathbb{R}^d)l^\infty} \leqslant \sum_{g=1}^{G} \left(2\|\omega^{(g)}(N_1, \ldots, N_d, L)\|_{l^\infty} \prod_{j=g}^{G} e^{L^{V^{(j)}}} \right)$$

By Lemma 20 the right hand side goes to zero as $L \to \infty$ and $N_j \to \infty$ for each $j \in [d]$. We have thus concluded the proof. ■

Now we are in a position to prove Proposition 8.
Proof of Proposition 8: Take free parameters $N = N_1 = N_2 = \cdots = N_d$ then the result follows from Proposition 21 ■

B.1.3 Smooth Case

First we prove Theorem 9 and then Proposition 10.

Lemma 22 (Normality) *The set* Flow *generates a normal subgroup (denoted as)* $\langle \text{Flow} \rangle$ *of* $\text{Diff}_0([0,1]^d)$.

Proof : If $\varphi = \text{Flow}(V)$ then using Corollary 9.14 (Lee, 2013) in our case it shows that for every $F \in \text{Diff}_0([0,1]^d)$ the Flow of the (pushforward) vector field $F_* V$ is $\eta = F \circ \varphi \circ F^{-1}$, in other words $\eta = \text{Flow}(F_* D) \Rightarrow \eta \in \text{Flow}$ so for φ in the generated subgroup $\langle \text{Flow} \rangle$ and the diffeomorphism $F \in \text{Diff}_0([0,1]^d)$ we can write :

$$\varphi \in \langle \text{Flow} \rangle \Rightarrow \exists n \in \mathbb{N}_+ \text{ s.t. } \forall i \in \{1, \ldots, n\},\ \exists \varphi^{(i)} \in \text{Flow}(V_i) \text{ s.t. } \varphi = \varphi^{(n)} \circ \cdots \circ \varphi^{(1)}$$

then one can write:

$$F \circ \varphi \circ F^{-1} = (F \circ \varphi^{(n)} \circ F^{-1}) \circ (F \circ \varphi^{(n-1)} \circ F^{-1}) \circ \cdots \circ (F \circ \varphi^{(1)} \circ F^{-1}) = \eta^{(n)} \circ \cdots \circ \eta^{(1)} \in \langle \text{Flow} \rangle$$

■

Lemma 23 (Large Granularity Triviality for Smooth Regularity) *There exists a* $K_d \in \mathbb{N}_+$ *s.t. for any integer* $G > K_d$, $\mathcal{H}_d^{G,\infty}([0,1]^d)$ *is empty.*

Proof : Note $\mathcal{H}_d^{G,\infty}\left([0,1]^d\right) = \mathcal{H}_d^{G,\infty}\left((0,1)^d\right)$ because these set of functions are zero on the boundary of $[0,1]^d$ by continuity. Define the autonomous norm

$$\|f\|_{\text{frag}} := \min\{m \in \mathbb{N} \mid f = h_1 \cdots h_m \text{ where } h_i = \text{Flow}(V_i) \text{ for some } V_i \in \mathcal{X}\}$$

one can check that this norm is conjugate invariant. By Burago et al. (2013) [Theorem 1.17] the autonomous norm on the group $\text{Diff}_0\left((0,1)^d\right)$ is bounded by a constant K_d as $(0,1)^d$ is portable.[8] ■

Lemma 24 (Decomposition) *For every diffeomorphism $\varphi \in \text{Diff}_0([0,1]^d)$, there exist some $G \in \mathbb{N}$ and flows $\varphi^{(1)}, \ldots, \varphi^{(G)} \in \text{Flow}$ such that*

$$\varphi = \bigcirc_{g=1}^{G} \varphi^{(i)} \overset{\text{def.}}{=} \varphi^{(G)} \circ \cdots \circ \varphi^{(1)}. \tag{33}$$

Proof : By Thurston's Theorem, see e.g. (Banyaga, 1997a, Theorem 2.1.1), the group $\text{Diff}_0([0,1]^d)$ is simple; i.e. it has no *proper* normal subgroups besides the trivial group consisting only of the identity (diffeomorphism) on $\mathbb{R}^d$. Since Flow does not contain only the identity diffeomorphism on $\mathbb{R}^d$ and, by Lemma (22), since Flow generates a normal subgroup of $\text{Diff}_0([0,1]^d)$ then Flow generates the entire group $\text{Diff}_0([0,1]^d)$. Consequentially, each $\varphi \in \text{Diff}_0([0,1]^d)$ admits a representation of the form (33). ■

Now we are in a position to prove Theorem 9.
Proof of Theorem 9: By Lemma 24 there exists a G such that $\varphi \in \mathcal{H}_d^{G,\infty}([0,1]^d)$; as $\mathcal{H}_d^{G,\infty}([0,1]^d)$ is non-empty, by Lemma 23 $G \leqslant K_d$. This means one can write any diffeomorphism as a composition of at most K_d flows. ■

Using Theorem 9 we can prove Proposition 10.
Proof of Proposition 10: As $\varphi \in \text{Diff}_0([0,1]^d)$ it is also s differentiable so by Theorem 9 there exists $G \leqslant K_d$ such that $\varphi \in \mathcal{H}_d^{G,s}([0,1]^d)$. By Proposition 8 there exists Ψ such that:

$$\begin{aligned}
\|\varphi - \Psi\|_{L^\infty(\mathbb{R}^d)l^\infty} &\leqslant \sum_{g=1}^{G} \left(2\|\omega^{(g)}(N,L)\|_{l^\infty} \prod_{j=g}^{G} e^{L^{V^{(j)}}} \right) \\
&\leqslant 170(s+1)^d 8^s \max_{g,j}\{\|V_j^{(g)}\|_{C^s([0,1]^d)}\}(NL)^{-2s/d} e^{\max_g\{L^{V^{(g)}}\}} K_d \left(\prod_{j=2}^{G} L^{V^{(j)}} \right) \\
&\in O\big(N^{-2s/d}\big)
\end{aligned} \tag{34}$$

as $(\omega^{(g)})_j = 85(s+1)^d 8^s \|V_j^{(g)}\|_{C^s([0,1]^d)}(NL)^{-2s/d}$ ■

We are now able to deduce our main qualitative universal approximation guarantee within the class $\mathcal{H}_d([0,1]^d)$.
Proof of Theorem 4: Now, since φ is supported on $B(0,R)$ for large enough R then there exists some $M > 0$ such that φ is supported on $[-M,M]^d \supset B(0,R)$. Recalling by a contraction/expansion diffeomorphism $\varphi_M : \mathbb{R}^d \to \mathbb{R}^d$ of the form $\varphi_M(x) \overset{\text{def.}}{=} Ax$ for some $d \times d$ scalar matrix $A = kI_d$, for some $k > 0$, satisfying $\varphi_M([-M,M]) \subseteq B(0,1/8)$. Moreover φ_M can be written as the time 1 flow whose integral curve with initial condition $x \in \mathbb{R}^d$ is

$$x_t^x = k^t x \text{ for all } t \geqslant 0$$

We may without loss of generality consider $\sup(\varphi \circ \varphi_M^{-1}) = B(0,\frac{1}{8})$ thus $\varphi \circ \varphi_M^{-1}$ fixes all points on a neighborhood of the boundary of $B(0,1/4)$. Consequently, we may apply the Munkres-Connel-Bing Theorem, as formulated in (Müller, 2014, Lemma 2) to deduce that: for every $\varepsilon > 0$ there exists a diffeomorphism $\varphi_\varepsilon : \mathbb{R}^d \to \mathbb{R}^d$ supported on $B(0,\frac{1}{4})$ satisfying the uniform approximation guarantees

$$\|\varphi_\varepsilon(x) - \varphi \circ \varphi_M^{-1}(x)\|_{L^\infty(\mathbb{R}^d)l^\infty} \leqslant \tfrac{\varepsilon}{3}. \tag{35}$$

8. Take vector field X as the vector field that points toward the point $(\frac{1}{2},\ldots,\frac{1}{2})$ at every point and θ a translation.

Which by substituting $x = \varphi_M(x)$ we'll still have:

$$\|\varphi_\varepsilon \circ \varphi_M - \varphi\|_{L^\infty(\mathbb{R}^d)l^\infty} \leqslant \tfrac{\varepsilon}{3}. \tag{36}$$

Now, since φ_ε is supported on $B(0, \frac{1}{4}) \subset [0,1]^d$ and since φ_ε is a diffeomorphism then Theorem 9 implies that there exists some $\tilde{K}_d$ (not depending on φ_ε nor on ε) and some $1 \leqslant G \leqslant \tilde{K}_d$ such that $\varphi_\varepsilon \in \mathcal{H}_d^{G,1}([0,1]^d)$. Applying Theorem 7 implies that there exists ReLU neural ODEs $\{\Psi^{(g)} = \operatorname{Flow}(\Phi^{(g)})\}_{g=1}^G \subset \mathcal{X}^0$ such that the diffeomorphism $\Psi_\varepsilon \stackrel{\text{def.}}{=} \bigcirc_{g=1}^G \Psi^{(g)}$ is Lipschitz and compactly supported and satisfies the approximation guarantee:

$$\left\|\varphi_\epsilon - \Psi_\varepsilon\right\|_{L^\infty(\mathbb{R}^d)l^\infty} \leqslant \tfrac{\varepsilon}{3}. \tag{37}$$

As $\varphi_M \in \mathcal{H}_d([-M,M]^d)$ by the same theorem:

$$\left\|\varphi_M - \Psi_M\right\|_{L^\infty(\mathbb{R}^d)l^\infty} \leqslant \tfrac{\varepsilon}{3L^{\Psi_\varepsilon}}. \tag{38}$$

Combining (36), (37), (38):

$$\begin{aligned}
\left\|\Psi_\varepsilon \circ \Psi_M - \varphi\right\| &\leqslant \left\|\Psi_\varepsilon \circ \Psi_M - \varphi_\epsilon \circ \varphi_M\right\| + \left\|\varphi_\epsilon \circ \varphi_M - \varphi\right\| \\
&\leqslant \left\|\Psi_\varepsilon \circ \Psi_M - \Psi_\varepsilon \circ \varphi_M\right\| + \left\|\Psi_\varepsilon \circ \varphi_M - \varphi_\epsilon \circ \varphi_M\right\| + \left\|\varphi_\epsilon \circ \varphi_M - \varphi\right\| \\
&\leqslant L^{\Psi_\varepsilon}\left\|\Psi_M - \varphi_M\right\| + \left\|\Psi_\varepsilon \circ \varphi_M - \varphi_\epsilon \circ \varphi_M\right\| + \left\|\varphi_\epsilon \circ \varphi_M - \varphi\right\| \\
&\leqslant L^{\Psi_\varepsilon} \tfrac{\varepsilon}{3L^{\Psi_\varepsilon}} + \tfrac{\varepsilon}{3} + \tfrac{\varepsilon}{3} = \varepsilon
\end{aligned} \tag{39}$$

and setting $\Psi = \Psi_\varepsilon \circ \Psi_M$ and $K_d \stackrel{\text{def.}}{=} \tilde{K}_d + 1$ yields the conclusion.

■

Appendix C. Additional Geometric and Topological Background

1. continuum: a compact and connected metric space that contains at least two points.

2. Tangent vector, tangent bundle and vector field: Let M be a smooth n-manifold.

   (a) A *tangent vector* at a point $p \in M$ can be defined as the velocity $\gamma'(0)$ of a smooth curve $\gamma : (-\varepsilon, \varepsilon) \to M$ with $\gamma(0) = p$, or equivalently as a derivation at p, i.e. a linear map $v : C^\infty(M) \to \mathbb{R}$ satisfying the Leibniz rule

   $$v(fg) = v(f)g(p) + f(p)v(g), \qquad f, g \in C^\infty(M).$$

   The set of all tangent vectors at p forms an n-dimensional vector space called the *tangent space* at p, denoted T_pM.

   (b) The *tangent bundle* of M is the disjoint union

   $$TM = \bigsqcup_{p \in M} T_pM,$$

   together with the natural projection $\pi : TM \to M$ given by $\pi(v) = p$ for $v \in T_pM$. It is itself a smooth $2n$-dimensional manifold.

   (c) A *vector field* on M is a smooth map

   $$X : M \to TM$$

such that $X(p) \in T_pM$ for every $p \in M$. Equivalently, a vector field is a smooth section of the tangent bundle $\pi : TM \to M$. In local coordinates $(x^1, \dots, x^n)$, any vector field has the form

$$X = \sum_{i=1}^{n} f_i(x) \frac{\partial}{\partial x^i},$$

where $f_i \in C^\infty(M)$.

3. Homotopic and C^k-diffeotopy: A homotopy between two continuous functions f and g from a topological space X to a topological space Y is defined to be a continuous function

$$H : X \times [0,1] \to Y$$

such that $H(x,0) = f(x)$ and $H(x,1) = g(x)$ for all $x \in X$.
Let M and N be smooth manifolds, and let $f_0, f_1 : M \to N$ be C^k diffeomorphisms ($1 \leqslant k \leqslant \infty$).

We say that f_0 and f_1 are *C^k-diffeotopic* if there exists a map

$$F : M \times [0,1] \to N$$

such that:

(a) For each $t \in [0,1]$, the map
$$f_t(x) := F(x,t) : M \to N$$
is a C^k diffeomorphism.

(b) $f_0 = F(\cdot, 0)$ and $f_1 = F(\cdot, 1)$.

(c) The map F is C^k in x and continuous (sometimes C^k) in (x,t).

4. Conjugation invariant and autonomous norm: **conjugation-invariant norm** $\nu : G \to [0; +\infty)$ on a group G is a function which satisfies the following axioms:

(i) $\nu(1) = 0$;

(ii) $\nu(f) = \nu(f^{-1}) \quad \forall f \in G$;

(iii) $\nu(fg) \leqslant \nu(f) + \nu(g) \quad \forall f, g \in G$;

(iv) $\nu(f) = \nu(gfg^{-1}) \quad \forall f, g \in G$;

(v) $\nu(f) > 0$ for all $f \neq 1$

for a smooth function f define $\|f\|_{\text{frag}} := \min\{m \in \mathbb{N} \mid f = h_1 \cdots h_m \text{ where } h_i = \text{Flow}(V_i) \text{ for some } V_i \in \mathcal{X}\}$ where $\text{Flow}(V_i)$ is the time one solution of Cauchy Problem for smooth vector field V_i

5. Nowhere dense and meagerness: Let $T = (S, \tau)$ be a topological space and $A \subseteq S$. A is *nowhere dense* in T if and only if
$$\left(\overline{A}\right)^\circ = \varnothing,$$
where $\overline{A}$ denotes the closure of A and A° its interior.
A is *meager* in T if and only if it can be written as a countable union

$$A = \bigcup_{n=1}^{\infty} A_n,$$

where each $A_n \subseteq S$ is nowhere dense in T. Equivalently, complement of A is open and dense.

6. Complete vector field: Let M be a smooth manifold and let X be a smooth vector field. Denote by φ_t the local flow of X, that is, the solution to the ODE

$$\frac{d}{dt}\varphi_t(p) = X(\varphi_t(p)), \qquad \varphi_0(p) = p.$$

We say that X is *complete* if for every $p \in M$, the integral curve $\gamma_p(t) := \varphi_t(p)$ is defined for all $t \in \mathbb{R}$. Equivalently, X is complete if its local flow extends to a global flow

$$\varphi : \mathbb{R} \times M \to M.$$

7. Portable manifold: We say that a smooth connected open manifold M is **portable** if it admits a complete vector field X and a compact subset M_0 with the following properties:
   - M_0 is an attractor of the flow X^t generated by X: for every compact subset $K \subset M$ there exists $\tau > 0$ so that $X^\tau(K) \subset M_0$.
   - There exists a diffeomorphism $\theta \in \mathrm{Diff}_0(M)$ so that $\theta(M_0) \cap M_0 = \varnothing$.

8. Recurrent point: Let X be a topological space and let $f : X \to X$ be a continuous map. A point $x \in X$ is called a *recurrent point* of f if there exists a sequence of integers (n_k) with $n_k \to \infty$ such that

$$f^{n_k}(x) \longrightarrow x \quad \text{as } k \to \infty.$$

Equivalently, x is recurrent if it belongs to its own ω-limit set,

$$x \in \omega(x) := \{y \in X : f^{n_k}(x) \to y \text{ for some sequence } n_k \to \infty\}.$$

9. Orbit, fixed point and periodic point: Let X be a set and $f : X \to X$ a map. For $x \in X$, the *forward orbit* of x under f is the set

$$\mathcal{O}^+(x) := \{f^n(x) : n \in \mathbb{N}_0\},$$

where $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ and $f^0 = \mathrm{id}_X$.

If f is invertible, the *(full) orbit* of x is

$$\mathcal{O}(x) := \{f^n(x) : n \in \mathbb{Z}\}.$$

A point $x \in X$ is called a fixed point of f if

$$f(x) = x.$$

A point $x \in X$ is called a periodic point of period $k \geqslant 1$ if

$$f^k(x) = x,$$

and k is the smallest positive integer with this property.

The set of all periodic points of period k is denoted

$$\mathrm{Per}_k(f) := \{\, x \in X : f^k(x) = x \text{ and } f^j(x) \neq x \text{ for all } 0 < j < k \,\}.$$

Appendix D. Proofs of Corollaries

Proof of Corollary 11: By Theorem 7 and using (15), there exists Lipschtiz ReLU neural ODEs $\{\Psi_i = \mathrm{Flow}\,(\Phi_i)\}_{i=1}^D$ such that for each $i \in [D]$ we have:

$$\|\,\mathrm{Flow}\,(V_{f_i}) - \Psi_i\|_{L^\infty(\mathbb{R}^d)l^\infty} < \epsilon$$

Which by Theorem 7, $\varepsilon \in O\left(\frac{1}{n}\right)$ so restricting the domain and only considering the last coordinate of the output we'll have:

$$\|\pi_1^{d+1} \circ \operatorname{Flow}(V_{f_i}) \circ \iota_d^{d+1} - \pi_1^{d+1} \circ \Psi_i \circ \iota_d^{d+1}\|_{L^\infty(\mathbb{R}^d)} < \epsilon$$

but by (14) we know $f_i = \pi_1^{d+1} \circ \operatorname{Flow}(V_{f_i}) \circ \iota_d^{d+1}$ so define $\Psi = \bigoplus_{i=1}^{D}(\pi_1^{d+1} \circ \Psi_i \circ \iota_d^{d+1})$ then:

$$\|f - \Psi\|_{L^\infty(\mathbb{R}^d)l^\infty} < \epsilon.$$

also width and depth are given in table 1. Also note that by lifting, projecting and concatination the function remains Lipschitz. This completes our proof. ■

Proof of Corollary 13: The Benyamini-Lindenstrauss theorem; see e.g. (Benyamini and Lindenstrauss, 2000, Theorem 1.12), implies that for every non-empty subset $A \subseteq [0,1]^d$, each $L \geqslant 0$, and every L-Lipschitz function $g : A \to \mathbb{R}^D$, there exists a L-Lipschitz extension $G : [0,1]^d \to \mathbb{R}^D$; i.e. $g|_A = G$. Therefore, (Miculescu, 2002/03, Theorem 1) implies that the set of Lipschitz functions is dense in $C([0,1]^d, \mathbb{R}^D)$ for the uniform topology. The result now follows from Corollary 11. ■

Proof of Corollary 14:
Applying the Benamou-Brenier Theorem (Villani, 2003, Theorem 2.12 (ii) and (iii)), we deduce that there exists a convex function $\varphi : \mathbb{R}^d \to \mathbb{R}$ inducing the unique optimal transport map between μ and ν, namely $\mu = \nabla\varphi_\sharp\nu$ i.e. φ is a Kantorovich potential. note that, at this state $\nabla\varphi$ may only be defined μ-a.s. therefore we verify its regularity before being able to continue further with any uniform approximation thereof. Under Assumption 4.1 the Caffarelli's regularity theorem, as formulated (Villani, 2009, Theorem 12.50 (ii)) applies and we deduce that $\varphi \in C^{2,\alpha}((0,1)^d)$; in particular, $\nabla\varphi$ is defined on all of $(0,1)^d$ and it's Lipschitz thereon by the mean-valued theorem.

Since $\nabla\varphi : (0,1)^d \to (0,1)^d$ is L'-Lipschitz, for some $L' \geqslant 0$, then it admits an L'-Lipschitz extension $\tilde{\varphi} : [0,1]^d \to [0,1]^d$, again by the Benyamini-Lindenstrauss Theorem, see e.g. (Benyamini and Lindenstrauss, 2000, Theorem 1.12) (it is easy to see that it must map the close cube to itself, since $\nabla\varphi$ maps $(0,1)^d$ to itself and the latter is continuous).

Fix $\varepsilon > 0$, applying Corollary 11 we find that there exists a Lipschitz Latent Neural ODE $\Psi : \mathbb{R}^{d+D} \to \mathbb{R}^{d+D}$ induced by ReLU MLP of depth $\mathcal{O}(\log_2(d))$, width $\mathcal{O}(dn^{d+1})$, with $\mathcal{O}(dn^{d+1})$ non-zero parameters, such that

$$\left(\max_{x\in[0,1]^d} \left\|\tilde{\varphi}(x) - \Psi(x)\right\|^2\right)^{1/2} < \varepsilon \tag{40}$$

Since μ and ν are both supported on $[0,1]^d$ then they belong to $\mathcal{P}_p([0,1]^d)$ for every $1 \leqslant p < \infty$; in particular, μ and $\Psi_\sharp\nu$ belong to $\mathcal{P}_2(\mathbb{R}^d)$. Therefore, $\mathcal{W}_2(\mu, \Psi_\sharp\nu)$ is finite and we have

$$\mathcal{W}_1\big(\mu, \Psi_\sharp\nu^{(N)}\big) \leqslant \mathcal{W}_1\big(\Psi_\sharp\nu, \Psi_\sharp\nu^{(N)}\big) + \mathcal{W}_1\big(\mu, \Psi_\sharp\nu\big) \leqslant \underbrace{\mathcal{W}_1\big(\Psi_\sharp\nu, \Psi_\sharp\nu^{(N)}\big)}_{\text{(I)}} + \underbrace{\mathcal{W}_2\big(\mu, \Psi_\sharp\nu\big)}_{\text{(II)}}. \tag{41}$$

We begin by controlling term (I) . Indeed, since Ψ is L^Ψ-Lipschitz, then we have

$$\text{(I)} = \mathcal{W}_1\big(\Psi_\sharp\nu, \Psi_\sharp\nu^{(N)}\big) \leqslant \operatorname{Lip}(\Psi)\, \mathcal{W}_1\big(\nu, \nu^{(N)}\big). \tag{42}$$

Now, applying the concentration of measure result in (Hou et al., 2023, Lemma 18) we have that: for every $\frac{\delta}{L^\Psi} > 0$

$$\mathbb{P}\bigg(\Big|\mathcal{W}_1\big(\nu, \nu^{(N)}\big) - \mathbb{E}\big[\mathcal{W}_1\big(\nu, \nu^{(N)}\big)\big]\Big| \geqslant \frac{\delta}{L^\Psi}\bigg) \leqslant 2\, e^{\frac{-2N\delta^2}{d(L^\Psi)^2}} \quad \text{and} \quad \mathbb{E}\big[\mathcal{W}_1\big(\nu, \nu^{(N)}\big)\big] \leqslant \frac{\sqrt{d}C_{[0,1]^d}}{\sqrt[d]{N}}$$

for some constant $C_{[0,1]^d} > 0$. Consequently, we may control (I) probabilistically: for every $\delta > 0$

$$\mathbb{P}\Big(L^\Psi \mathcal{W}_1\big(\nu, \nu^{(N)}\big) \leqslant L^\Psi \frac{\sqrt{d}C_{[0,1]^d}}{\sqrt[d]{N}} + \delta\Big) \geqslant 1 - 2\, e^{\frac{-2N\delta^2}{d(L^\Psi)^2}} \tag{43}$$

So

$$\mathbb{P}\Big((\mathrm{I}) \leqslant L^{\Psi} \frac{\sqrt{d} C_{[0,1]^d}}{\sqrt[d]{N}} + \delta \Big) \geqslant 1 - 2\, e^{\frac{-2N\delta^2}{d(L^{\Psi})^2}} \tag{44}$$

It remains to control term (II). In particular,

$$(\mathrm{II}) = \mathcal{W}_2\big(\mu, \Psi_\sharp \nu\big) = \mathcal{W}_2\big(\nabla\varphi_\sharp \nu, \Psi_\sharp \nu\big), \tag{45}$$

Since $[0,1]^d \backslash (0,1)^d$ is of ν-measure zero, we may again rephrase (45) as

$$(\mathrm{II}) = \mathcal{W}_2\big(\nabla\varphi_\sharp \nu, \Psi_\sharp \nu\big) \leqslant \mathbb{E}_{X\sim\nu}\big[\|(\nabla\varphi)_\sharp(X) - (\Psi)_\sharp(X)\|^2\big]^{1/2} = \mathbb{E}_{X\sim\nu}\big[\|\tilde{\varphi}(X) - (\Psi)_\sharp(X)\|^2\big]^{1/2} \overset{\text{def.}}{=} (\mathrm{III}). \tag{46}$$

But

$$(\mathrm{II}) \leqslant (\mathrm{III}) \leqslant \Big(\max_{x\in[0,1]^d} \big\|\tilde{\varphi}(x) - \Psi(x)\big\|^2 \Big)^{1/2} \tag{47}$$

Consequently, combining (40), (44) ,and (47) yields the conclusion. ■

References

Hervé Andrès, Alexandre Boumezoued, and Benjamin Jourdain. Signature-based validation of real-world economic scenarios. In *Astin Bulletin*, 2024.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, 2017.

Augustin Banyaga. *The structure of classical diffeomorphism groups*, volume 400 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1997a. ISBN 0-7923-4475-8. doi: 10.1007/978-1-4757-6800-8. URL https://doi.org/10.1007/978-1-4757-6800-8.

Augustin Banyaga. *The Structure of Classical Diffeomorphism Groups*, volume 400 of *Mathematics and Its Applications*. Springer, Dordrecht, 1997b. ISBN 978-0-7923-4832-3.

Christian Bayer, Luca Pelizzari, and Jan Schoenmakers. Primal and dual optimal stopping with signatures. *Finance and Stochastics*, 2025.

Yoav Benyamini and Joram Lindenstrauss. *Geometric nonlinear functional analysis. Vol. 1*, volume 48 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2000. ISBN 0-8218-0835-4. doi: 10.1090/coll/048.

Gérard Biau, Benoît Cadre, Maxime Sangnier, and Ugo Tanielian. Some theoretical properties of GANs. *Annals of Statistics*, 48(3):1539–1562, 2020.

Dmitri Burago, Sergei Ivanov, and Leonid Polterovich. Conjugation-invariant norms on groups of geometric origin. *In "Groups of Diffeomorphisms", Adv. Stud. Pure Math., vol. 52, 2008, pp. 221-250*, 2013. Available at http://arxiv.org/abs/0710.1412.

Luis A. Caffarelli. Boundary regularity of maps with convex potentials–ii. *Annals of Mathematics*, 144(3):453–496, 1996. doi: 10.2307/2118610. URL https://www.jstor.org/stable/2118564.

Thomas Cass and Cristopher Salvi. Lecture notes on rough paths and applications to machine learning. *arXiv preprint arXiv:2404.06583*, 2024.

Ilya Chevyrev and Andrey Kormilitzin. A primer on the signature method in machine learning. *arXiv preprint arXiv:1603.03788*, 2016.

Nicola Muca Cirone, Maud Lemercier, and Cristopher Salvi. Neural signature kernels as infinite-width-depth-limits of controlled resnets. In *International Conference on Machine Learning*, pages 25358–25425. PMLR, 2023.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 70:854–863, 2017.

Albert Cohen, Ronald DeVore, Guergana Petrova, and Przemyslaw Wojtaszczyk. Optimal stable nonlinear approximation. *Foundations of Computational Mathematics*, 22(3):607–648, 2022.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016.

Christa Cuchiero, Martin Larsson, and Josef Teichmann. Deep neural networks, generic universal interpolation, and controlled odes. *SIAM Journal on Mathematics of Data Science*, 2(3):901–919, 2020.

Christa Cuchiero, Wahid Khosrawi, and Josef Teichmann. Signature-based models: Theory and calibration. *SIAM Journal on Financial Mathematics*, 14(3), 2023.

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

Arturo De Marinis, Davide Murari, Elena Celledoni, Nicola Guglielmi, Brynjulf Owren, and Francesco Tudisco. Approximation properties of neural odes. *arXiv preprint arXiv:2503.15696*, 2025.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. In *arXiv preprint arXiv:1410.8516*, 2014.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *International Conference on Learning Representations (ICLR)*, 2017.

Dennis Elbrächter, Dmytro Perekrestenko, Philipp Grohs, and Helmut Bölcskei. Deep neural network approximation theory. *IEEE Transactions on Information Theory*, 67(5):2581–2623, 2021.

Albert Fathi. Structure of the group of homeomorphisms preserving a good measure on a compact manifold. *Annales Scientifiques de l'École Normale Supérieure*, 13(1):45–93, 1980.

Marion Kirkland Fort. The embedding of homeomorphisms in flows. *Proceedings of the American Mathematical Society*, 6(6):960–967, 1955.

Kazuhiko Fukui, Tomasz Rybicki, and Tatsuhiko Yagasaki. The uniform perfectness of diffeomorphism groups of open manifolds. *arXiv preprint arXiv:1905.07664*, 2019.

Ken-Ichi Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural networks*, 2(3):183–192, 1989.

Sylvestre Gallot, Dominique Hulin, and Jacques Lafontaine. *Riemannian Geometry*. Springer Berlin Heidelberg, 3 edition, 2004. ISBN 978-3-642-18854-1. doi: 10.1007/978-3-642-18855-8.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Massimiliano Gubinelli. Controlling rough paths. *Journal of Functional Analysis*, 216(1):86–140, 2004.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5767–5777, 2017.

Lajos G. Gyurkó, Terry Lyons, Mark Kontkowski, and Jonathan Field. Extracting information from the signature of a financial data stream. *arXiv preprint arXiv:1307.7244*, 2013.

Stefan Haller, Tomasz Rybicki, and Josef Teichmann. Smooth perfectness for the group of diffeomorphisms. *Journal of Geometric Mechanics*, 5(3):281–294, 2013.

Ben Hambly and Terry Lyons. Uniqueness for the signature of a path of bounded variation and the reduced path group. *Annals of Mathematics*, 171(1):109–167, 2010. doi: 10.4007/annals.2010.171.109. URL https://annals.math.princeton.edu/wp-content/uploads/annals-v171-n1-p02-p.pdf.

Philip Hartman. *Ordinary Differential Equations*, volume 38 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, 2 edition, 2002.

Sigurdur Helgason. *Differential Geometry, Lie Groups, and Symmetric Spaces*, volume 80 of *Pure and Applied Mathematics*. Academic Press, New York, 1979. ISBN 978-0-12-338460-1. A foundational reference on the geometry and analysis of Lie groups and symmetric spaces.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Songyan Hou, Parnian Kassraie, Anastasis Kratsios, Andreas Krause, and Jonas Rothfuss. Instance-dependent generalization bounds via optimal transport. *Journal of Machine Learning Research*, 24(349): 1–51, 2023. URL https://www.jmlr.org/papers/v24/22-1293.html.

Gary Jones. The embedding of homeomorphisms of the plane in continuous flows. *Pacific Journal of Mathematics*, 41(2):421–436, 1972.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. *Advances in neural information processing systems*, 33:6696–6707, 2020.

Durk P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10215–10224, 2018.

Durk P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Victor L. Klee. Some topological properties of convex sets. *Transactions of the American Mathematical Society*, 78(1):30–45, 1955. doi: 10.2307/1992353. URL https://www.jstor.org/stable/1992353.

Ivan Kobyzev, Simon Prince, and Marcus Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, 2020.

Alexander Kolesov, Petr Mokrov, Igor Udovichenko, Milena Gazdieva, Gudmund Pammer, Anastasis Kratsios, Evgeny Burnaev, and Aleksandr Korotin. Energy-guided continuous entropic barycenter estimation for general costs. *Advances in Neural Information Processing Systems*, 37:107513–107546, 2024.

Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Burnaev. Wasserstein-2 generative networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=bEoxzW_EXsa.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=d8CBRlWNkqH.

Anastasis Kratsios and Ievgen Bilokopytov. Non-euclidean universal approximation. In *Advances in Neural Information Processing Systems*, volume 33, pages 10635–10646, 2020.

Anastasis Kratsios and Léonie Papon. Universal approximation theorems for differentiable geometric deep learning. *Journal of Machine Learning Research*, 23(196):1–73, 2022.

Anastasis Kratsios and Behnoosh Zamanlooy. Do ReLU networks have an edge when approximating compactly-supported functions? *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=sNxNi54B8b.

Andreas Kriegl and Peter W. Michor. Regular infinite dimensional lie groups. *Journal of Lie Theory*, 7(1): 61–99, 1997a. URL https://www.heldermann-verlag.de/jlt/jlt07/MICHPL.PDF.

Andreas Kriegl and Peter W Michor. *The convenient setting of global analysis*, volume 53. American Mathematical Soc., 1997b.

Andreas Kriegl and Peter W. Michor. Regular infinite dimensional lie groups. *Journal of Lie Theory*, 7(1): 61–99, 1997c.

Anna A. Kwiecińska. On the c0-closing lemma. *Annales Polonici Mathematici*, 64(2):131–138, 1996. URL http://eudml.org/doc/269953.

John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, second edition, 2013. ISBN 978-1-4419-9981-8. doi: 10.1007/978-1-4419-9982-5. URL https://doi.org/10.1007/978-1-4419-9982-5.

George G. Lorentz, Manfred v. Golitschek, and Yuly Makovoz. *Constructive approximation*, volume 304 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1996. ISBN 3-540-57028-4. doi: 10.1007/978-3-642-60932-9. URL https://doi.org/10.1007/978-3-642-60932-9. Advanced problems.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

Yulong Lu and Jianfeng Lu. A universal approximation theorem of deep neural networks for expressing probability distributions. *Advances in neural information processing systems*, 33:3094–3105, 2020.

Yui Man Lui. Human gesture recognition on product manifolds. *The Journal of Machine Learning Research*, 13(1):3297–3321, 2012.

Terry J Lyons. Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2): 215–310, 1998.

John N. Mather. Commutators of diffeomorphisms. *Commentarii Mathematici Helvetici*, 49(1):512–528, 1974.

John N. Mather. A curious remark concerning the geometric transfer map. *Commentarii Mathematici Helvetici*, 59:86–110, 1984.

Hrushikesh N Mhaskar and Tomaso Poggio. Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, 14(06):829–848, 2016.

Peter W Michor, Philipp Harms, and Martin Bauer. Regularity and completeness of half lie groups. *Journal of the European Mathematical Society*, 2023.

Radu Miculescu. Approximations by Lipschitz functions generated by extensions. *Real Anal. Exchange*, 28(1):33–40, 2002/03. ISSN 0147-1937,1930-1219. doi: 10.14321/realanalexch.28.1.0033. URL https://doi.org/10.14321/realanalexch.28.1.0033.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

James Morrill, Cristopher Salvi, Patrick Kidger, and James Foster. Neural rough differential equations for long time series. In *International Conference on Machine Learning*, pages 7829–7838. PMLR, 2021.

Stefan Müller. Uniform approximation of homeomorphisms by diffeomorphisms. *Topology and its Applications*, 178:315–319, 2014.

Stefan Müller. Uniform approximation of homeomorphisms by diffeomorphisms. *Topology and its Applications*, 178:315–319, 2014. ISSN 0166-8641. doi: https://doi.org/10.1016/j.topol.2014.10.003. URL https://www.sciencedirect.com/science/article/pii/S0166864114003915.

Karl-Hermann Neeb and Arturo Pianzola. Infinite-dimensional lie theory. *Oberwolfach Reports*, 3(4):3259–3320, 2007.

Zbigniew Nitecki and Michael Shub. Filtrations, decompositions, and explosions. *American Journal of Mathematics*, 97(4):1029–1047, 1975.

Hideki Omori. *Infinite-dimensional Lie groups*, volume 158. American Mathematical Soc., 2017.

B. G. Pachpatte. *Inequalities for Differential and Integral Equations*, volume 197 of *Mathematics in Science and Engineering*. Academic Press, San Diego, 1998. ISBN 0-12-543430-8.

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

Jacob Palis. Vector fields generate few diffeomorphisms. 1974.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330, 2018.

Philipp Petersen and Jakob Zech. Mathematical theory of deep learning. *arXiv preprint arXiv:2407.18384*, 2024.

Guergana Petrova and Przemysław Wojtaszczyk. Lipschitz widths. *Constructive Approximation*, 57(2): 759–805, 2023.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Allan Pinkus. *N-widths in Approximation Theory*, volume 7 of *Springer Series in Computational Mathematics*. Springer Science & Business Media, Berlin, Heidelberg, 2012. ISBN 978-3-642-23260-0.

Michael Puthawala, Matti Lassas, Ivan Dokmanic, and Maarten De Hoop. Universal joint approximation of manifolds and densities by simple injective flows. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17959–17983. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/puthawala22a.html.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2016.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1530–1538, 2015.

Litu Rout, Alexander Korotin, and Evgeny Burnaev. Generative modeling with optimal transport maps. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=5JdLZg346Lw.

Domenec Ruiz-Balet and Enrique Zuazua. Control of neural transport for normalising flows. *Journal de Mathématiques Pures et Appliquées*, 181:58–90, 2024.

Zuowei Shen. Deep network approximation characterized by number of neurons. *Communications in Computational Physics*, 28(5):1768–1811, 2020.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Optimal approximation rate of relu networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, 157:101–135, 2022. ISSN 0021-7824. doi: https://doi.org/10.1016/j.matpur.2021.07.009. URL https://www.sciencedirect.com/science/article/pii/S0021782421001124.

Jonathan W Siegel. Optimal approximation rates for deep relu neural networks on Sobolev and Besov spaces. *Journal of Machine Learning Research*, 24(357):1–52, 2023.

Josef Teichmann. Regularity of infinite-dimensional lie groups by metric space methods. *Tokyo Journal of mathematics*, 24(1):39–58, 2001.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

William Thurston. Foliations and groups of diffeomorphisms. *Bulletin of the American Mathematical Society*, 80:304–307, 1974.

Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003. ISBN 0-8218-3312-X. doi: 10.1090/gsm/058. URL https://doi.org/10.1090/gsm/058.

Cédric Villani. *Optimal transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. ISBN 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9. URL https://doi.org/10.1007/978-3-540-71050-9. Old and new.

Benjamin Walker, Andrew Donald Mcleod, Tiexin Qin, Yichuan Cheng, Haoliang Li, and Terry Lyons. Log neural controlled differential equations: The lie brackets make a difference. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 49822–49844. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/walker24a.html.

Paulo Varandas Wescley Bonomo. Continuous flows generate few homeomorphisms. *Proceedings of the Edinburgh Mathematical Society*, 2020.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural networks*, 94:103–114, 2017.

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In *Conference on learning theory*, pages 639–649. PMLR, 2018.

Dmitry Yarotsky and Anton Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. *Advances in neural information processing systems*, 33:13005–13015, 2020.

Shijun Zhang, Zuowei Shen, and Haizhao Yang. Deep network approximation: Achieving arbitrary accuracy with fixed number of neurons. *Journal of Machine Learning Research*, 23(276):1–60, 2022.

Shijun Zhang, Jianfeng Lu, and Hongkai Zhao. Deep network approximation: Beyond relu to diverse activation functions. *Journal of Machine Learning Research*, 25(35):1–39, 2024. URL http://jmlr.org/papers/v25/23-0912.html.