Universal Joint Approximation of Manifolds and Densities by Simple Injective Flows

Michael Puthawala [1] **Matti Lassas** [2] **Ivan Dokmanić** [3] **Maarten de Hoop** [1]

Abstract

We study approximation of probability measures supported on n-dimensional manifolds embedded in $\mathbb{R}^m$ by injective flows—neural networks composed of invertible flows and injective layers. We show that in general, injective flows between $\mathbb{R}^n$ and $\mathbb{R}^m$ universally approximate measures supported on images of extendable embeddings, which are a subset of standard embeddings: when the embedding dimension m is small, topological obstructions may preclude certain manifolds as admissible targets. When the embedding dimension is sufficiently large, $m \geq 3n + 1$, we use an argument from algebraic topology known as the clean trick to prove that the topological obstructions vanish and injective flows universally approximate any differentiable embedding. Along the way we show that the studied injective flows admit efficient projections on the range, and that their optimality can be established "in reverse," resolving a conjecture made in (Brehmer & Cranmer, 2020)

1. Introduction

Invertible flow networks emerged as powerful deep learning models to learn maps between distributions (Durkan et al., 2019a; Grathwohl et al., 2018; Huang et al., 2018; Jaini et al., 2019; Kingma et al., 2016; Kingma & Dhariwal, 2018; Kobyzev et al., 2020; Kruse et al., 2019; Papamakarios et al., 2019). They generate high-quality samples (Kingma & Dhariwal, 2018) and facilitate solving scientific inference problems (Brehmer & Cranmer, 2020; Kruse et al., 2021).

[1]Department of Computational and Applied Math, Rice University, Houston, TX, USA [2]Department of Mathematics and Statistics, University of Helsinki, Finland [3]Department of Mathematics and Computer Science, University of Basel, Basel, Switzerland. Correspondence to: Michael Puthawala <map19@rice.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

By design, however, invertible flows are bijective and may not be a natural choice when the target distribution has low-dimensional support. This problem can be overcome by combining bijective flows with expansive, injective layers, which map to higher dimensions (Brehmer & Cranmer, 2020; Cunningham et al., 2020; Kothari et al., 2021). Despite their empirical success, the theoretical aspects of such globally injective architectures are not well understood.

In this work, we address approximation-theoretic properties of injective flows. We prove that under mild conditions these networks universally approximate probability measures supported on low-dimensional manifolds and describe how their design enables applications to inference and inverse problems.

1.1. Prior Work

The idea to combine invertible (coupling) layers with expansive layers has been explored by (Brehmer & Cranmer, 2020) and (Kothari et al., 2021). Brehmer & Cranmer (2020) combine two flow networks with a simple expansive element (in the sense made precise in Section 2.1) and obtain a network that parameterizes probability distributions supported on manifolds.[1]

Kothari et al. (2021) propose expansive coupling layers and build networks similar to that of Brehmer & Cranmer (2020) but with an arbitrary number of expressive and expansive elements. They observe that the resulting network trains very fast with a small memory footprint, while producing high-quality samples on a variety of benchmark datasets.

While (to the best of our knowledge) there are no approximation-theoretic results for injective flows, there exists a body of work on universality of invertible flows; see Kobyzev et al. (2020) for an overview. Several works show that certain bijective flow architectures are distributionally universal. This was proved for autoregressive flows with sigmoidal activations by Huang et al. (2018) and for sum-of-squares polynomial flows (Jaini et al.,

[1]More precisely, distributions on manifolds are parameterized by the pushforward (via their network) of a simple probability measure in the latent space.

2019). Teshima et al. (2020) show that several flow networks including those from Huang et al. (2018) and Jaini et al. (2019) are also universal approximators of diffeomorphisms.

The injective flows considered here have key applications in inference and inverse problems; for an overview of deep learning approaches to inverse problems, see (Arridge et al., 2019). Bora et al. (2017) proposed to regularize compressed sensing problems by constraining the recovery to the range of (pre-trained) generative models. Injective flows with efficient inverses as generative models give an efficient algorithmic projection[2] on the range, which facilitates implementation of reconstruction algorithms. An alternative approach is Bayesian, where flows are used to obtain tractable variational approximations of posterior distributions over parameters of interest, via supervised training on labeled input-output data pairs. Ardizzone et al. (2018) encode the dimension-reducing forward process by an invertible neural network (INN), with additional outputs used to encode posterior variability. Invertibility guarantees that a model of the inverse process is learned implicitly. For a given measurement, the inverse pass of the INN approximates the posterior over parameters. Sun & Bouman (2020) propose variational approximations of the posterior using an untrained deep generative model. They train a normalizing flow which produces samples from the posterior, with the prior and the noise model given implicitly by the regularized misfit functional. In Kothari et al. (2021) this procedure is adapted to priors specified by injective flows which yields significant improvements in computational efficiency.

1.2. Our Contribution

We derive new approximation results for neural networks composed of bijective flows and injective expansive layers, including those introduced by (Brehmer & Cranmer, 2020) and (Kothari et al., 2021). We show that these networks universally jointly approximate a large class of manifolds and densities supported on them.

We build on the results of Teshima et al. (2020) and develop a new theoretical device which we refer to as the *embedding gap*. This gap is a measure of how nearly a mapping from $\mathbb{R}^o \to \mathbb{R}^m$ embeds an n-dimensional manifold in $\mathbb{R}^m$, where $n \leq o$. We find a natural relationship between the embedding gap and the problem of approximating probability measures with low-dimensional support.

We then relate the embedding gap to a relaxation of universality we call the *manifold embedding property*. We show that this property captures the essential geometric aspects of universality and uncover important topological restrictions on the approximation power of these networks, to our knowledge, heretofore unknown in the literature. We give an example of an absolutely continuous measure μ and embedding $f\colon \mathbb{R}^2 \to \mathbb{R}^3$ such that $f_{\#}\mu$ can not be approximated with combinations of flow layers and linear expansive layers. This may be surprising since it was previously conjectured that networks such as those of Brehmer & Cranmer (2020) can approximate any "nice" density supported on a "nice" manifold. We establish universality for manifolds with suitable topology, described in terms of *extendable embeddings*. We find that the set of extendable embeddings is a proper subset of all embeddings, but when $m \geq 3n+1$, via an application of the *clean trick* from algebraic topology, we show that all diffeomorphisms are extendable and thus injective flows approximate distributions on arbitrary manifolds. Our universality proof also implies that optimality of the approximating network can be established in reverse: optimality of a given layer can be established without optimality of preceding layers. This settles a (generalization of a) conjecture posed for a three-part network (composed of two flow networks and zero padding) in (Brehmer & Cranmer, 2020). Finally, we show that these universal architectures are also practical and admit exact layer-wise projections, as well as other properties discussed in Section 3.5.

2. Architectures Considered

Let $C(X, Y)$ denote the space of continuous functions $X \to Y$. Our goal is to make statements about networks in $\mathcal{F} \subset C(X, Y)$ that are of the form:

$$\mathcal{F} = \mathcal{T}_L^{n_L} \circ \mathcal{R}_L^{n_{L-1}, n_L} \circ \cdots \circ \mathcal{T}_1^{n_1} \circ \mathcal{R}_1^{n_0, n_1} \circ \mathcal{T}_0^{n_0} \quad (1)$$

where $\mathcal{R}_\ell^{n_{\ell-1}, n_\ell} \subset C(\mathbb{R}^{n_{\ell-1}}, \mathbb{R}^{n_\ell})$, $\mathcal{T}_\ell^{n_\ell} \subset C(\mathbb{R}^{n_\ell}, \mathbb{R}^{n_\ell})$, $L \in \mathbb{N}$, $n_0 = n$, $n_L = m$, and $n_\ell \geq n_{\ell-1}$ for $\ell = 1, \ldots, L$. We introduce a well-tuned shorthand notation and write $\mathcal{H} \circ \mathcal{G} := \{h \circ g\colon h \in \mathcal{H}, g \in \mathcal{G}\}$ throughout the paper.

We identify $\mathcal{R}$ with the expansive layers and $\mathcal{T}$ with the bijective flows. Loosely speaking, the purpose of the expansive layers is to allow the network to parameterize high-dimensional functions by low-dimensional coordinates in an injective way. The flow networks give the network the expressivity necessary for universal approximation of manifold-supported distributions.

2.1. Expansive Layers

The expansive elements transform an n-dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^{n_{\ell-1}}$, and embed it in a higher dimensional space $\mathbb{R}^{n_\ell}$. To preserve the topology of the manifold they are injective. We thus make the following assumptions about the expansive elements:

Definition 2.1 (Expansive Element)**.** A family of functions $\mathcal{R} \subset C(\mathbb{R}^n, \mathbb{R}^m)$ is called an family of expansive elements if $m > n$, and each $R \in \mathcal{R}$ is both injective and Lipschitz.

[2] Idempotent but in general not orthogonal.

Examples of expansive elements include

(R1) Zero padding: $R(x) = \left[x^T, \mathbf{0}^{(m-n)}\right]^T$ where $\mathbf{0}^{(m-n)}$ is the zero vector (Brehmer & Cranmer, 2020).

(R2) Multiplication by an arbitrary full-rank matrix, or one-by-one convolution:

$$R(x) = Wx, \quad \text{or} \quad R(x) = w \star x \tag{2}$$

where $W \in \mathbb{R}^{m\times n}$ and $\mathrm{rank}(W) = n$ (Cunningham et al., 2020), and w is a convolution kernel $\star$ denotes convolution (Kingma & Dhariwal, 2018).

(R3) Injective ReLU layers: $R(x) = \mathrm{ReLU}(Wx)$, $W = \left[B^T, -DB^T, M^T\right]^T$, or $R(x) = \mathrm{ReLU}\left(\left[w^T, -w^T\right] \star x\right)$ for matrix $B \in \mathrm{GL}_n(\mathbb{R})$, positive diagonal matrix $D \in \mathbb{R}^{n\times n}$, and arbitrary matrix $M \in \mathbb{R}^{(m-2n)\times n}$ (Puthawala et al., 2020).

(R4) Injective ReLU networks (Puthawala et al., 2020, Theorem 5). These are functions $R : \mathbb{R}^n \to \mathbb{R}^m$ of the form $R(x) = W_{L+1}\,\mathrm{ReLU}(\ldots \mathrm{ReLU}(W_1 x + b_1)\ldots) + b_L$ where W_ℓ are $n_{\ell+1} \times n_\ell$ matrices and b_ℓ are the bias vectors in $\mathbb{R}^{n_{\ell+1}}$. The weight matrices W_L satisfy the Directed Spanning Set (DSS) condition for $\ell \leq L$ (that make all layers injective) and W_{L+1} is a generic matrix which makes the map $R : \mathbb{R}^n \to \mathbb{R}^m$ injective where $m \geq 2n+1$. Note that the DSS condition requires that $n_\ell \geq 2n_{\ell-1}$ for $\ell \leq L$ and we have $n_1 = n$ and $n_{L+1} = m$.

Continuous piecewise-differentiable functions with bounded gradients are always Lipschitz. Thus, the Lipschitzness assumption is automatically satisfied by feed-forward networks with piecewise-differentiable activation functions with bounded gradients. This includes compositions of ReLU and sigmoid layers.

2.2. Bijective Flow Networks

The bulk of our theoretical analysis is devoted to the bijective flow networks, which bend the range of the expansive elements into the correct shape. We make the following assumptions about the expressive elements:

Definition 2.2 (Bijective Flow Network)**.** Let $\mathcal{T} \subset C(\mathbb{R}^n, \mathbb{R}^n)$ for $n \in \mathbb{N}$. We call $\mathcal{T}$ a family of bijective flow networks if every $T \in \mathcal{T}$ is Lipschitz and bijective.

Examples of bijective flow networks include

(T1) *Coupling flows*, introduced by (Dinh et al., 2014) consider $R(\mathbf{x}) = H_k \circ \cdots \circ H_1(\mathbf{x})$ where

$$H_i(\mathbf{x}) = \begin{bmatrix} h_i\left([\mathbf{x}]_{1:d}, g_i\left([\mathbf{x}]_{d+1:n}\right)\right) \\ [\mathbf{x}]_{d+1:n} \end{bmatrix}. \tag{3}$$

In Eqn. 3, $h_i : \mathbb{R}^d \times \mathbb{R}^e \to \mathbb{R}^d$ is invertible w.r.t. the first argument given the second, and $g_i : \mathbb{R}^{n-d} \to \mathbb{R}^e$ is arbitrary. Typically in practice the operation in Eqn. 3 is combined with additional invertible operations such as permutations, masking or convolutions (Dinh et al., 2014; 2016; Kingma & Dhariwal, 2018).

(T2) *Autoregressive flows*, introduced by Kingma et al. (2016) are generalizations of triangular flows $A : \mathbb{R}^n \to \mathbb{R}^n$ where for $i = 1, \ldots, n$ the i'th value of A is given by of the form

$$[A]_i(\mathbf{x}) = h_i\left([\mathbf{x}]_i, g_i\left([\mathbf{x}]_{1:i-1}\right)\right) \tag{4}$$

In Eqn. 4, $h_i : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}$ where again h_i is invertible w.r.t. the first argument given the second, and $g_i : \mathbb{R}^{i-1} \to \mathbb{R}^m$ is arbitrary except for $g_1 = \mathbf{0}$. In Huang et al. (2018), the authors choose $h_i(x, \mathbf{y})$, where $\mathbf{y} \in \mathbb{R}^m$, to be a multi-layer perceptron (MLP) of the form

$$h_i(x, \mathbf{y}) = \phi \circ W_{p,\mathbf{y}} \circ \cdots \circ \phi \circ W_{1,\mathbf{y}}(x) \tag{5}$$

where ϕ is a sigmoidal increasing non-linear activation function.

3. Main Results

3.1. Embedding Gap

We call a function f an embedding and denote it by $f \in \mathrm{emb}(X, Y)$ if $f : X \to Y$ is continuous, injective, and $f^{-1} : f(X) \to X$ is continuous[3]. Also we denote by $\mathrm{emb}^k(\mathbb{R}^n, \mathbb{R}^m)$ the set of maps $f \in \mathrm{emb}(\mathbb{R}^n, \mathbb{R}^m) \cap C^k(\mathbb{R}^n, \mathbb{R}^m)$ which differential $df|_x : \mathbb{R}^n \to \mathbb{R}^m$ is injective at all points $x \in \mathbb{R}^n$. We now introduce the *embedding gap*, a non-symmetric notion of distance between f and g. This quantifies the degree to which a mapping $g \in \mathrm{emb}(\mathbb{R}^o, \mathbb{R}^m)$ fails to embed a manifold $\mathcal{M} = f(K)$ for compact $K \subset \mathbb{R}^n$ where $f \in \mathrm{emb}(K, \mathbb{R}^m)$. Later in the paper, f will be the function to be approximated, and g an approximating flow-network.

Definition 3.1 (Embedding Gap)**.** Let $n \leq p \leq o \leq m$, $K \subset \mathbb{R}^n$ be compact and non-empty, $W \subset \mathbb{R}^o$ be compact and contain the closure of set U which is open in the subspace topology of some vector subspace V of dimension p, where $f \in \mathrm{emb}(K, \mathbb{R}^m)$ and $g \in \mathrm{emb}(W, \mathbb{R}^m)$. The Embedding Gap between f and g on K and W is

$$B_{K,W}(f, g) = \inf_{r \in \mathrm{emb}(f(K), g(W))} \|I - r\|_{L^\infty(f(K))} \tag{6}$$

[3]Note that if X is a compact set, then continuity of the of $f^{-1} : f(X) \to X$ is automatic, and need not be assumed (Sutherland, 2009, Cor. 13.27). Moreover, if $f : \mathbb{R}^n \to \mathbb{R}^m$ is a continuous injective map that satisfies $|f(x)| \to \infty$ as $|x| \to \infty$, then by (Mukherjee, 2015, Cor. 2.1.23) the map $f^{-1} : f(\mathbb{R}^n) \to \mathbb{R}^n$ is continuous.

where $I\colon f(K) \to f(K)$ is the identity function and $\|h\|_{L^\infty(X)} = \operatorname{ess\,sup}_{x\in X} \|h(x)\|_2$ for $h\colon X \to Y$, where Y is some L^∞ space. We refer to the embedding gap between f and g without specifying K and W when it is clear from context.

Remark 3.2. As $W \subset \mathbb{R}^o$ contains U, an open set in V, there is an affine map $A : \mathbb{R}^n \to V$ such that $A(K) \subset W$. Thus, the map $r_0 = g \circ A \circ f^{-1} : f(K) \to g(W)$ is an injective continuous map from a compact set to its range and hence $r_0 \in \operatorname{emb}(f(K), g(W))$. This proves that the infimum in 6 is non-empty.

Before giving properties of $B_{K,W}(f, g)$, we briefly describe its interpretation and meaning. We denote by $\mathcal{P}(X)$ the set of probability measures over X. If the embedding gap between f and g is small, then $g^{-1}\circ r$ embeds the range of f for an r that is nearly the identity. Hence g^{-1} nearly embeds the range of f into $\mathbb{R}^o$. $B_{K,W}(f, g)$ also serves as an upper bound

$$\inf_{\mu_o \in \mathcal{P}(W)} \mathrm{W}_2\left(f_{\#}\mu_n, g_{\#}\mu_o\right) \leq B_{K,W}(f, g)$$

where $\mu_n \in \mathcal{P}(K)$ is given, and $\mathrm{W}_2\left(\nu_1, \nu_2\right)$ denotes the Wasserstein-2 distance with ℓ^2 ground metric (Villani, 2008). This is proven in Lemma C.1 part 9. The above result has a simple meaning in the context of machine learning. Suppose we want to learn a generative model g to (approximately) sample from a probability measure ν with low-dimensional support, by applying g to samples from a *base* distribution μ_o. Suppose further that ν is a push-forward of some (known or unknown) distribution μ_n via f. The embedding gap $B_{K,W}(f, g)$ then upper bounds the 2-Wasserstein distance between ν and $g_{\#}\mu_0$ for the best possible choice of μ_o.[4]

In the context of optimal transport, the embedding r can be interpreted as a candidate transport map from any measure pushed forward by f, that can be pulled back through g. Loosely speaking, for $\mu_o' = g^{-1} \circ r \circ f_{\#}\mu_n$, r transports $f_{\#}\mu_n$ to $g_{\#}\mu_o'$ with cost no more than $\|I - r\|_{L^\infty(f(K))}$. See Fig. 1 for a visualization of the embedding gap between two toy functions. The embedding gap satisfies inequalities useful for studying networks of the form of Eqn. 1, see Lemma C.1.

In the remainder of this section we use the embedding gap to prove universality of neural networks. The set $f(K)$ will be a target manifold to approximate, and g will be a neural network of the form Eq. 1. The embedding gap requires g to be a proper embedding and so, in particular, injective. This is why we require injectivity of both the expansive and bijective flow layers.

[4]The choice of p-Wasserstein distance is suitable for measures with mismatched low-dimensional support; this has been widely exploited in training generative models (Arjovsky et al., 2017).

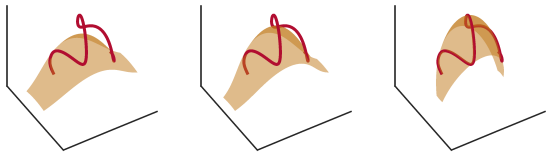


Figure 1: A visualization of the embedding gap. In all three figures we plot $f(K)$ and $g_i(W)$ for Left: $i = 1$, Center: $i = 2$ and Right: $i = 3$. Visually, we see that $g_i(W)$ approaches $f(W)$ as i increases, and we compute $B_{K,W}(f, g_1) > B_{K,W}(f, g_2) > B_{K,W}(f, g_3) = 0$.

3.2. Manifold Embedding Property

We now introduce a central concept, the manifold embedding property (MEP). A family of networks has the MEP if it can, as measured by the embedding gap, nearly embed a large class of manifolds of certain dimension and regularity. The MEP is a property of a family of functions $\mathcal{E} \subset \operatorname{emb}(W, \mathbb{R}^m)$ where $W \subset \mathbb{R}^o$. In this manuscript, $\mathcal{E}$ will always be formed by taking $\mathcal{E} \coloneqq \mathcal{T} \circ \mathcal{R}$, where $\mathcal{R}$ and $\mathcal{T}$ are the expansive layers and bijective flow networks described in sections 2.1 and 2.2 respectively.

We note here that $\mathcal{E}$ having the MEP is closely related to the question of whether or not a given n-dimensional manifold $\mathcal{M} = f(K)$ for $f \in \operatorname{emb}(K, \mathbb{R}^m)$, $K \subset \mathbb{R}^n$, can be approximated by an $E \in \mathcal{E}$. This choice of first applying (possibly non-universal) expansive layers, and then universal layers puts some topological restrictions on the expressivity, which we discuss in great detail in Section 3.3.

In anticipation of these topological difficulties, when we refer to the MEP, we consider it with respect to a class of functions $\mathcal{F} \subset \operatorname{emb}(\mathbb{R}^n, \mathbb{R}^m)$. The MEP can be interpreted as a density statement, saying that our networks $\mathcal{E}$ are dense in some set $\mathcal{F} \subset \operatorname{emb}(\mathbb{R}^n, \mathbb{R}^m)$ in the topology induced by the '$B_{K,W}$ distance.' Two examples of $\mathcal{F}$ that we are particularly interested in are the following. When $\mathcal{F} = \operatorname{emb}(\mathbb{R}^n, \mathbb{R}^m)$, and also when each $f \in \mathcal{F}$ can be written as $f = D \circ L$ where $L : \mathbb{R}^{m\times n}$ is a linear map of rank n and $D : \mathbb{R}^m \to \mathbb{R}^m$ is a C^k diffeomorphism with $k \geq 1$.

Definition 3.3 (Manifold Embedding Property)**.** Let $\mathcal{E} \subset \operatorname{emb}(\mathbb{R}^o, \mathbb{R}^m)$ and $\mathcal{F} \subset \operatorname{emb}(\mathbb{R}^n, \mathbb{R}^m)$ be two families of functions. We say that $\mathcal{E}$ has the m, n, o Manifold Embedding Property (MEP) w.r.t. $\mathcal{F}$ if for every compact non-empty set $K \subset \mathbb{R}^n$, $f \in \mathcal{F}$, and $\epsilon > 0$, there is an $E \in \mathcal{E}$ and a compact set $W \subset \mathbb{R}^o$ such that the restriction of f to K and the restriction of E to W satisfies

$$B_{K,W}(f, E) < \epsilon. \tag{7}$$

When it is clear from the context, we abbreviate the m, n, o

MEP w.r.t. $\mathcal{F}$ simply by the m, n, o MEP, or simply the MEP.

We also present the following two lemmas which relate to the algebra of the MEP.

Lemma 3.4. *Let $\mathcal{E}_1^{p,o} \subset \mathrm{emb}(\mathbb{R}^p, \mathbb{R}^o)$ have the o, n, p MEP w.r.t. $\mathcal{F}_1^{n,o} \subset \mathrm{emb}(\mathbb{R}^n, \mathbb{R}^o)$, and likewise let $\mathcal{E}_2^{o,m} \subset \mathrm{emb}(\mathbb{R}^o, \mathbb{R}^m)$ have the m, o, o MEP w.r.t. $\mathcal{F}_2^{o,m} \subset \mathrm{emb}(\mathbb{R}^o, \mathbb{R}^m)$. If each $E_2^{o,m} \in \mathcal{E}_2^{o,m}$ is locally Lipschitz, then $\mathcal{E}_2^{o,m} \circ \mathcal{E}_1^{p,o}$ has the m, n, p MEP w.r.t. $\mathcal{F}^{o,m} \circ \mathcal{F}^{n,o}$.*

The proof of Lemma 3.4 is in Appendix C.2.1.

We note that when the elements of $\mathcal{E}_2^{o,m}$ are differentiable, local Lipschitzness is automatic, and need not be assumed, see e.g. (Tao, 2009, Ex. 10.2.6). We also record the following lemma, proved in C.2.2, which is a weak-converse of Lemma 3.4. It states that if $\mathcal{E}_2^{o,m} \circ \mathcal{E}_1^{p,o}$ has the m, n, p MEP, then $\mathcal{E}_2^{o,m}$ has the m, n, o MEP.

Lemma 3.5. *Let $\mathcal{E}_1^{p,o} \subset \mathrm{emb}(\mathbb{R}^p, \mathbb{R}^o)$ and $\mathcal{E}_2^{o,m} \subset \mathrm{emb}(\mathbb{R}^o, \mathbb{R}^m)$ be such that $\mathcal{E}_2^{o,m} \circ \mathcal{E}_1^{p,o}$ has the m, n, p MEP with respect to family $\mathcal{F} \subset \mathrm{emb}(\mathbb{R}^n, \mathbb{R}^m)$. Then $\mathcal{E}_2^{o,m}$ has the m, n, o MEP with respect to family $\mathcal{F}$.*

Definition 3.6 (Uniform Universal Approximator)**.** For a non-empty subset $\mathcal{F}^{n,m} \subset C(\mathbb{R}^n, \mathbb{R}^m)$, a family $\mathcal{E}^{n,m} \subset C(\mathbb{R}^n, \mathbb{R}^m)$ is said to be a uniform universal approximator of $\mathcal{F}^{n,m}$ if for every $f \in \mathcal{F}^{n,m}$, every non-empty compact $K \subset \mathbb{R}^n$, and each $\epsilon > 0$, there is an $E \in \mathcal{E}^{n,m}$ satisfying:

$$\sup_{x \in K} \|f(x) - E(x)\|_2 < \epsilon. \tag{8}$$

If $\mathcal{E} \subset \mathrm{emb}(\mathbb{R}^o, \mathbb{R}^m)$ is a uniform universal approximator of $\mathcal{F}^{o,m} = C^0(\mathbb{R}^o, \mathbb{R}^m)$ on compact sets, then it has the m, n, o MEP w.r.t $C^0(\mathbb{R}^n, \mathbb{R}^m)$ for any $n \leq o$, see Lemma 3.9. As an example, when $m \geq 2o+1$ injective ReLU networks $E : \mathbb{R}^o \to \mathbb{R}^m$ (i.e., mappings of the form (R4)) are uniform universal approximator of $C^0(\mathbb{R}^o, \mathbb{R}^m)$ on compact sets, see e.g. (Puthawala et al., 2020) and (Yarotsky, 2017; 2018). Thus, networks that are uniform universal approximators automatically possess the MEP. Generalizations of this are considered in Lemma 3.9.

With the definition of the MEP and uniform universal approximator established, we now discuss in detail the nature of the topological obstructions to approximating all one-chart manifolds.

3.3. Topological Obstructions to Manifold Learning with Neural Networks

We show that using non-universal expansive layers and flow layers imposes some topological restrictions on what can be approximated. Let $n = 2$, $m = 3$, and $K = S^1 \subset \mathbb{R}^2$ be the circle, and let

$$\mathcal{E} = \left\{ T \circ R \in C(\mathbb{R}^2, \mathbb{R}^3) \colon R \in \mathbb{R}^{3\times 2}, T \in \mathrm{hom}(\mathbb{R}^3, \mathbb{R}^3) \right\}.$$

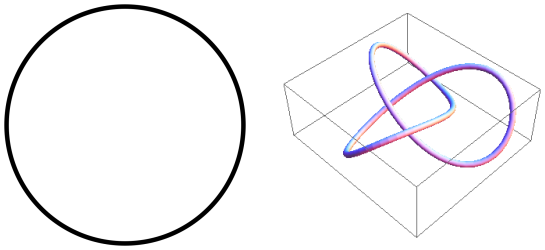


Figure 2: An illustration of the case when $n = 2$, $m = 3$, and $K = S^1$ is the circle. Here $f : S^1 \to \mathbb{R}^3$ is an embedding such that the curve $\mathcal{M} = f(S^1)$ is a trefoil knot. Due to knot theoretical reasons, there are no map $E = T \circ R : \mathbb{R}^2 \to \mathbb{R}^3$ such that $E(S^1) = \mathcal{M}$, where $R : \mathbb{R}^2 \to \mathbb{R}^3$ is a full rank linear map and $T : \mathbb{R}^3 \to \mathbb{R}^3$ is a homeomorphism. This shows that a combination of linear maps and coupling flow maps can not represent all embedded manifolds. For this reason, we define the class $\mathcal{I}(\mathbb{R}^n, \mathbb{R}^m)$ of extendable embeddings f in Definition 3.7. A similar 2-dimensional example can be obtained to a knotted ribbon, see Sec. C.3.1.

That is, $\mathcal{E}$ is the set of maps that can be written as compositions of linear maps from $\mathbb{R}^2$ to $\mathbb{R}^3$ and homeomorphisms on all of $\mathbb{R}^3$. Let $f \in \mathrm{emb}(K, \mathbb{R}^3)$ be an embedding that maps K to a trefoil knot $\mathcal{M} = f(S^1)$, see Fig. 2. Such a function f can not be written as a restriction of an $E \in \mathcal{E}$ to S^1. In Sec. C.3.1 we prove this fact and build a related example where a measure, $\mu \in \mathcal{P}(\mathbb{R}^2)$, supported on an annulus is pushed forward to a measure supported on a knotted ribbon in $\mathbb{R}^3$ by an embedding $g \colon \mathbb{R}^2 \to \mathbb{R}^3$. For this measure, there are no $E \in \mathcal{E}$ such that $g_{\#}\mu = E_{\#}\mu$. We note that the counterexample is still valid if $\mathcal{E}$ is replaced with $\hat{\mathcal{E}} = \mathcal{T} \circ \mathcal{D}$ where $\mathcal{T} = \mathrm{hom}(\mathbb{R}^3, \mathbb{R}^3)$ and $\mathcal{D} = \mathrm{hom}(\mathbb{R}^3, \mathbb{R}^3) \circ \mathbb{R}^{3\times 2}$. See C.3.2 for a proof. The point here is not that R is linear, but rather that it embeds all of $\mathbb{R}^2$ into $\mathbb{R}^3$, rather than only S^1 into $\mathbb{R}^3$.

With this difficulty in mind, we define the MEP property with respect to a certain subclass of manifolds $\{f(K) : f \in \mathcal{F}\}$. Additionally, when considering flow networks which are universal approximators of C^2 diffeomorphisms, we restrict the class of manifolds to be approximated even further. This is necessary because manifolds that are homeomorphic are not necessarily diffeomorphic[5]. Moreover, it is known that C^2-smooth diffeomorphisms can not approximate general homeomorphisms in the C^0 topology, see (Müller, 2014) for a precise statement. All C^1-smooth diffeomorphisms $f : \mathbb{R}^m \to \mathbb{R}^m$, however, can be approximated in the strong topology of C^1 by C^2-smooth

[5] A classic example are the exotic spheres. These are topological structures that are homeomorphic, but not diffeomorphic, to the sphere (Milnor, 1956).

diffeomorphism $\tilde{f} : \mathbb{R}^m \to \mathbb{R}^m$, $\ell \geq k$, see (Hirsch, 2012, Ch. 2, Theorem 2.7). Because of this, we have to pay attention to the smoothness of the maps in the subset $\mathcal{F} \subset \mathrm{emb}(K, \mathbb{R}^m)$.

Definition 3.7 (Extendable Embeddings)**.** We define the set of Extendable Embeddings as

$$\begin{aligned}\mathcal{I}(\mathbb{R}^n, \mathbb{R}^m) &:= \mathcal{D} \circ \mathcal{L} \\ \mathcal{D} &= \mathrm{Diff}^1(\mathbb{R}^m, \mathbb{R}^m) \\ \mathcal{L} &= \left\{ L \in \mathbb{R}^{m \times n} \colon \mathrm{rank}(L) = n \right\},\end{aligned}$$

where $\mathrm{Diff}^k(\mathbb{R}^m, \mathbb{R}^m)$ is the set of C^k-smooth diffeomorphisms from $\mathbb{R}^m$ to itself. Note that $\mathcal{I}(\mathbb{R}^n, \mathbb{R}^m) \subset \mathrm{emb}(\mathbb{R}^n, \mathbb{R}^m)$.

The word extendable in the name extendable embeddings refers to the fact that the family $\mathcal{D}$ in Definition 3.7 is a proper subset of $\mathrm{emb}(L(K), \mathbb{R}^m)$ for some compact $K \subset \mathbb{R}^n$ and linear $L \in \mathbb{R}^{m \times n}$. Mappings in the set $\mathcal{D}$ are embeddings $D : L(K) \to \mathbb{R}^m$ that extend to diffeomorphisms from all of $\mathbb{R}^m$ to itself. Said differently, a $D \in \mathcal{D}$ is a map in $\mathrm{emb}^1(L(K), \mathbb{R}^m)$ that can be extended to a map $\tilde{D} \in \mathrm{Diff}^1(\mathbb{R}^m, \mathbb{R}^m)$ such that $\tilde{D}\Big|_{L(K)} = D$. This distinction is important, as there are maps in $\mathrm{emb}^1(L(K), \mathbb{R}^m)$ that can not be extended to diffeomorphisms on all of $\mathbb{R}^m$, as can be seen from the counterexample developed at the beginning of this section.

We also present here a theorem that states that when m is more than three times larger than n, any differentiable embedding from compact $K \subset \mathbb{R}^n$ to $\mathbb{R}^m$ is necessarily extendable.

Theorem 3.8. *When $m \geq 3n + 1$ and $k \geq 1$, for any C^k embedding $f \in \mathrm{emb}^k(\mathbb{R}^n, \mathbb{R}^m)$ and compact set $K \subset \mathbb{R}^n$, there is a map $E \in \mathcal{I}^k(\mathbb{R}^n, \mathbb{R}^m)$ (that is, E is in the closure of the set of flow type neural networks) such that $E(K) = f(K)$. Moreover,*

$$\mathcal{I}^k(K, \mathbb{R}^m) = \mathrm{emb}^k(K, \mathbb{R}^m) \tag{9}$$

The proof of Theorem 3.8 in Appendix C.3.3. We also remark here that the proof of the above theorem relies on the so called 'clean trick' from differential topology. This trick is related to fact that in $\mathbb{R}^4$, all knots can be reduced to the simple knot continuously.

3.4. Universality

We now combine the notions of universality and extendable embeddings to produce a result stating that many commonly used networks of the form studied in Section 2 have the MEP.

Lemma 3.9. *(i) If $\mathcal{R} \subset \mathrm{emb}(\mathbb{R}^n, \mathbb{R}^m)$ is a uniform universal approximator of $C(\mathbb{R}^n, \mathbb{R}^m)$ and $I \in \mathcal{T}$ where I is the identity map, then $\mathcal{E} := \mathcal{T} \circ \mathcal{R}$ has the MEP w.r.t.* $\mathrm{emb}(\mathbb{R}^n, \mathbb{R}^m)$.

(ii) If $\mathcal{R}$ is such that there is an injective $R \in \mathcal{R}$ and open set $U \subset \mathbb{R}^o$ such that $R\big|_{\overline{U}}$ is linear, and $\mathcal{T}$ is a sup *universal approximator in the space of $\mathrm{Diff}^2(\mathbb{R}^m, \mathbb{R}^m)$, in the sense of (Teshima et al., 2020), of the C^2-smooth diffeomorphisms, then $\mathcal{E} := \mathcal{T} \circ \mathcal{R}$ has the MEP w.r.t. $\mathcal{I}(\mathbb{R}^n, \mathbb{R}^m)$.*

For uniform universal approximators that satisfy the assumptions of (i), see e.g. (Puthawala et al., 2020). The proof of Lemma 3.9 is in Appendix C.4.1. It has the following implications for the architectures studied in Section 2.

Example 1. *Let $\mathcal{E} := \mathcal{T} \circ \mathcal{R}$ and (T1), (T2), (R1), ..., (R4) be as described in Section 2. Then*

(i) If $\mathcal{T}$ is either (T1) or (T2) and $\mathcal{R}$ is (R4), then $\mathcal{E}$ has the m, n, o MEP w.r.t. $\mathrm{emb}(\mathbb{R}^n, \mathbb{R}^m)$.

(ii) If $\mathcal{T}$ is (T2) with sigmoidal activations (Huang et al., 2018), then if $\mathcal{R}$ is any of (R1), ..., (R4), then $\mathcal{E}$ has the m, n, o MEP w.r.t. $\mathcal{I}(\mathbb{R}^n, \mathbb{R}^m)$.

The proof of Example 1 is in Appendix C.4.2.

We now present our universal approximation result for networks given in Eqn. 1 and a decoupling property. Below, we say that a measure μ in $\mathbb{R}^n$ is absolutely continuous if it is absolutely continuous w.r.t. the Lebesgue measure.

Theorem 3.10. *Let $n_0 = n, n_L = m$ $K \subset \mathbb{R}^n$ be compact, $\mu \in \mathcal{P}(K)$ be an absolutely continuous measure. Further let, for each $\ell = 1, \ldots, L$, $\mathcal{E}_\ell^{n_{\ell-1}, n_\ell} := \mathcal{T}_\ell^{n_\ell} \circ \mathcal{R}_\ell^{n_{\ell-1}, n_\ell}$ where $\mathcal{R}_\ell^{n_{\ell-1}, n_\ell}$ is a family of injective expansive elements that contains a linear map, and $\mathcal{T}_\ell^{n_\ell}$ is a family of bijective family networks. Finally let $\mathcal{T}_0^n$ be distributionally universal, i.e. for any absolutely continuous $\mu \in \mathcal{P}(\mathbb{R}^n)$ and $\nu \in \mathcal{P}(\mathbb{R}^n)$, there is a $\{T_i\}_{i=1,2,\ldots}$ such that $T_{i\#}\mu \to \nu$ in distribution. Let one of the following two cases hold:*

(i) $f \in \mathcal{F}_L^{n_{L-1}, m} \circ \cdots \circ \mathcal{F}_1^{n, n_1}$ and $\mathcal{E}_\ell^{n_{\ell-1}, n_\ell}$ have the the n_ℓ, $n_{\ell-1}$,$n_{\ell-1}$ MEP for $\ell = 1, \ldots, L$ with respect to $\mathcal{F}_\ell^{n_{\ell-1}, n_\ell}$.

(ii) $f \in \mathrm{emb}^1(\mathbb{R}^n, \mathbb{R}^m)$ be a C^1-smooth embedding, for $\ell = 1, \ldots, L$ $n_\ell \geq 3n_{\ell-1} + 1$ and the families $\mathcal{T}_\ell^{n_\ell}$ are dense in $\mathrm{Diff}^2(\mathbb{R}^{n_\ell})$.

Then, there is a sequence of $\{E_i\}_{i=1,2,\ldots} \subset \mathcal{E}_L^{n_{L-1}, m} \circ \cdots \circ \mathcal{E}_1^{n_1, n} \circ \mathcal{T}_0^n$ such that

$$\lim_{i \to \infty} \mathrm{W}_2\left(f_{\#}\mu, E_{i\#}\mu\right) = 0. \tag{10}$$

The proof of Theorem 3.10 is in Appendix C.4.3. The results of Theorems 3.8 and 3.10 have a simple interpretation,

omitting some technical details. Densities on 'nice' manifolds embedded in high-dimensional spaces can always be approximated by neural networks of the form Eq. 1. Here, 'nice' manifolds are smooth and homeomorphic to $\mathbb{R}^n$. This proves that networks like Eqn. 1 are 'up to task' of solving generation problems.

As discussed in the above and in Figure 2, there are topological obstructions to obtaining the results of Theorem 3.10 with a general embedding $f : \mathbb{R}^n \to \mathbb{R}^m$. When $n = 2$, $m = 3$, $L = 1$, and μ is the uniform measure on an annulus $K \subset \mathbb{R}^2$ target measure $F_{\#}\mu$ is the uniform measure on a knotted ribbon $\mathcal{M} = f(K) \subset \mathbb{R}^3$. There are no injective linear maps $R : \mathbb{R}^2 \to \mathbb{R}^3$ and diffeomorphisms $T : \mathbb{R}^3 \to \mathbb{R}^3$ such that $E = T \circ R$ would satisfy $\mathcal{M} = E(K)$ and $E_{\#}\mu = F_{\#}\mu$.

We note that our networks are designed expressly to approximate manifolds, and hence injectivity is key. This separates our results from, e.g. (Lee et al., 2017, Theorem 3.1) or (Lu & Lu, 2020, Theorem 2.1), where universality results of ReLU networks are also obtained.

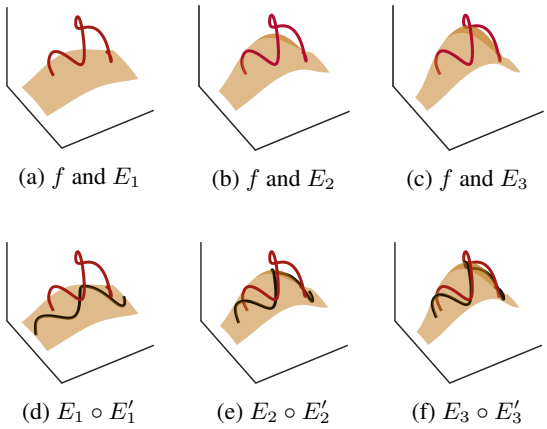
The previous theorem states that the entire network is universal if it can be broken into pieces that have the MEP. The following lemma, proved in Appendix C.4.4, shows that if $\mathcal{E}^{n,m} = \mathcal{H}^{o,m} \circ \mathcal{G}^{n,o}$, then $\mathcal{H}^{o,m}$ must have the m, n, o MEP if $\mathcal{E}^{n,m}$ is universal.

Lemma 3.11. *Suppose that $\mathcal{E}^{n,m} = \mathcal{H}^{o,m} \circ \mathcal{G}^{n,o}$ where $\mathcal{E}^{n,m} \subset \mathrm{emb}(\mathbb{R}^n, \mathbb{R}^m)$, $\mathcal{H}^{o,m} \subset \mathrm{emb}(\mathbb{R}^o, \mathbb{R}^m)$, and $\mathcal{G}^{n,o} \subset \mathrm{emb}(\mathbb{R}^n, \mathbb{R}^o)$. If $\mathcal{H}^{o,m}$ does not have the m, n, o MEP w.r.t. $\mathcal{F}$, then there exists a $f \in \mathcal{F}$, compact $K \subset \mathbb{R}^n$ and $\epsilon > 0$ such that for all $E \in \mathcal{E}^{n,m}$, and $r \in \mathrm{emb}(f(K), E(W))$*

$$\|I - r\|_{L^\infty(K)} \geq \epsilon. \tag{11}$$

Lemma 3.11 has a simple takeaway: If a bijective neural network is universal, then the last layer, last two layers, etc., must have the MEP. In other words, a network is only as universal as its last layer. Earlier layers, on the other hand, need not satisfy the MEP. 'Strong' layers close to the output can compensate for 'weak' layers closer to the input, but not the other way around.

There is a gap between the negation of Theorem 3.10 and Lemma 3.11. That is, it is possible for a family of functions $\mathcal{E}$ to satisfy Lemma 3.11 but nevertheless satisfy the conclusion of Theorem 3.10; these functions approximate measures without matching manifolds. Theorem 3.10 considers approximating measures, whereas Lemma 3.11 refers to matching manifolds exactly. As discussed in Section 3.3, there are no extendable embeddings that map S^1 to the trefoil knot in $\mathbb{R}^3$. Nevertheless, it is possible to construct a sequence of functions $(E_i)_{i=1,\dots}$ so that $\mathrm{W}_2(\nu, E_{i\#}\mu) = 0$ where μ and ν are the uniform distributions on S^1 and trefoil knot respectively. Such a construction is given in C.4.6.



(a) f and E_1 (b) f and E_2 (c) f and E_3

(d) $E_1 \circ E_1'$ (e) $E_2 \circ E_2'$ (f) $E_3 \circ E_3'$

Figure 3: A visualization of the construction described in Corollary 3.12 applied to a toy example when $m = 3, o = 2$ and $n = 1$. In all figures, $f(K)$ is the red curve, $E_i(W)$ are the orange surfaces, $E_i \circ E_i'(W')$ are the black curves, T_i and μ are not pictured. (a) - (c) The orange surfaces approach the red curves. This means that the sequence of E_1, E_2 and E_3 send $B_{K,W}(f, E_i)$ to zero as i increases. (d) - (f) The black curves, a subset of the orange surfaces, approach the red curves. This means that given E_1, E_2 and E_3 we can always find another sequence E_1', E_2' and E_3' that sends $B_{K,W}(f, E_i \circ E_i')$ to zero as i increases too. This as a consequence, sends $\mathrm{W}_2(f_{\#}\mu, E_i \circ E_i' \circ T_{i\#}\mu)$ to zero as i increases too for some choice of T_1, T_2 and T_3.

Although there are sequences of functions that approximate measure without matching manifolds, these sequences are never uniformly Lipschitz. This is proven in C.4.6. Under an idealization of training, we may consider a network undergoing training as successively better and better approximators of a target mapping. If the target mapping does not match the topology, then training necessarily leads to gradient blowup.

The proof of Theorem 3.10 also implies the following result which, loosely speaking, says that optimality of later layers can be determined without requiring optimality of earlier layers, while still having a network that is end-to-end optimal. The conditions and result of this is visualized on a toy example in Figure 3.

Corollary 3.12. *Let $\mathcal{F}^{n,o} \subset \mathrm{emb}(\mathbb{R}^n, \mathbb{R}^o)$, $\mathcal{F}^{o,m} \subset \mathrm{emb}(\mathbb{R}^o, \mathbb{R}^m)$, and let $\mathcal{E}^{o,m} \subset \mathrm{emb}(\mathbb{R}^o, \mathbb{R}^m)$ have the m, n, o MEP w.r.t. $\mathcal{F}^{o,m} \circ \mathcal{F}^{n,o}$. Then for every $f \in \mathcal{F}^{o,m} \circ \mathcal{F}^{n,o}$ and compact sets $K \subset \mathbb{R}^n$ and $W \subset \mathbb{R}^o$*

there is a sequence $\{E_i\}_{i=1,2,\ldots} \subset \mathcal{E}^{o,m}$ *such that*

$$\lim_{i\to\infty} B_{K,W}(f, E_i) = 0. \tag{12}$$

Further,if there is a compact $W' \subset \mathbb{R}^n$ *and* $\mathcal{E}^{n,o} \subset \mathrm{emb}(W', \mathbb{R}^o)$ *has the* o, n, n *MEP w.r.t.* $\mathcal{F}^{n,o}$*, and a* $\mathcal{T}^n$ *is a universal approximator for distributions, then for any absolutely continuous* $\mu \in \mathcal{P}(K)$ *where* $K \subset \mathbb{R}^n$ *is compact, there is a sequence* $\{E'_i\}_{i=1,2,\ldots} \subset \mathcal{E}^{n,o}$ *and* $\{T_i\}_{i=1,2,\ldots,} \subset \mathcal{T}^n$ *so that*

$$\lim_{i\to\infty} \mathrm{W}_2\left(f_{\#}\mu, E_i \circ E'_i \circ T_{i\#}\mu\right) = 0. \tag{13}$$

The proof of Corollary 3.12 is in Appendix C.4.5. Approximation results for neural networks are typically given in terms of the network end-to-end. Corollary 3.12 shows that the layers of approximating networks can in fact be built one at a time. This is related to an observation made in (Brehmer & Cranmer, 2020, Section B) about training strategies, where the authors remark that they 'expect faster and more robust training of a network' of the form in Eqn. 1 when $L = 1$, that is $\mathcal{F} = \mathcal{T}_1^m \circ \mathcal{R}_1^{n,m} \circ \mathcal{T}_0^n$. Corollary 3.12 shows that there exists a minimizing sequence in $\mathcal{T}_1^m$ that need only minimize Eqn. 12; the $\mathcal{T}_0^n$ layers can be minimized after. We can further combine Lemma 3.11 and Cor. 3.12 to prove that not only can the network from (Brehmer & Cranmer, 2020) be trained layerwise, but that *any* universal network can *necessarily* be trained layerwise, provided that it can be written as a composition of two smaller layers.

3.5. Layer-wise Inversion and Recovery of Weights

In this subsection, we describe how our network can be augmented with more useful properties if the architecture satisfies a few more assumptions without affecting universal approximation. We focus on a new layerwise projection result, with a further discussion of black-box recovery of our network's weights in Appendix C.5.2.

Given a point $y \in \mathbb{R}^m$ that does not lie in the range of the network, projecting y onto the range of the network is a practical problem without an obvious answer. The crux of the problem is inverting the injective (but non-invertible) $\mathcal{R}$ layers when $\mathcal{R}$ contains only full-rank matrices as in (R1) or (R2) then we can compute a least-squares solution. If, however, $\mathcal{R}$ contains layers which are only piecewise linear, as in (R3), then the problem of computing a least squares solution is more difficult, see Fig. 4. Nevertheless, we find that if $\mathcal{R}$ is (R3) we can still compute a least-squares solution.

Assumption 3.13. Let $\mathcal{R}$ be given by one of (R1) or (R2), or else (R3) when $m = 2n$.

If $\mathcal{R}$ only contains linear operators, then the least-squares problem can be computed by solving the normal equations

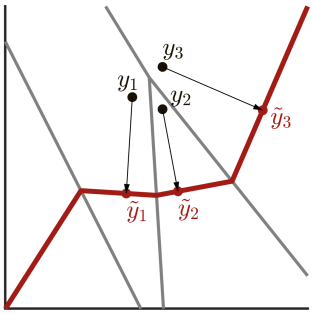


Figure 4: A schematic showing that, for a toy problem, the least-squares projection to a piecewise affine range can be discontinuous. Left: A partitioning of $\mathbb{R}^2$ into classes with gray boundaries. Two points y, y' are in the same class if they are both closest to the same affine piece of $R(\mathbb{R})$, the range of R. The three points y_1, y_2 and y_3 are each projected to the closest three points on $R(\mathbb{R})$ yielding $\tilde{y}_1$, $\tilde{y}_2$ and $\tilde{y}_3$. Note that the projection operation is continuous within each section, but discontinuous across gray boundaries between section.

(see (Golub, 1996, Section 5.3).) This includes cases (R1) or (R2). For (R3) we have the following result when $D = I^{n\times n}$ and $M \in \mathbb{R}^{0\times n}$.

Definition 3.14. Let $W = \begin{bmatrix} B^t & -DB^t \end{bmatrix}^t \in \mathbb{R}^{2n\times n}$ and $y \in \mathbb{R}^{2n}$ be given, and let $R(x) = \mathrm{ReLU}(Wx)$. Then define $c(y) \in \mathbb{R}^{2n}, \Delta_y \in \mathbb{R}^{n\times n}, M_y \in \mathbb{R}^{n\times 2n}$ where

$$c(y) \coloneqq \max\left(\begin{bmatrix} I^{n\times n} & -I^{n\times n} \\ -I^{n\times n} & I^{n\times n} \end{bmatrix} y, 0\right) \tag{14}$$

$$[\Delta_y]_{i,j} \coloneqq \begin{cases} 0 & \text{if } i \neq j \\ 0 & \text{if } [c(y)]_{i+n} = 0 \\ 1 & \text{if } [c(y)]_{i+n} > 0 \end{cases} \tag{15}$$

$$M_y \coloneqq \begin{bmatrix} (I^{n\times n} - \Delta_y) & \Delta_y \end{bmatrix} \tag{16}$$

where the max in Eqn. 14 is taken element-wise.

Theorem 3.15. *Let* $y \in \mathbb{R}^{2n}$*. If for* $i = 1, \ldots, n$*,* $[y]_i \neq [y]_{i+n}$ *then*

$$R^\dagger(y) \coloneqq (M_y W)^{-1} M_y y = \operatorname*{argmin}_{x\in\mathbb{R}^n} \|y - R(x)\|_2 . \tag{17}$$

Further, if there is a $i \in \{1, \ldots, n\}$ *such that* $[y]_i = [y]_{i+n}$*, then there are multiple minimizers of* $\|y - R(x)\|_2$*, one of which is* $R^\dagger(y)$*.*

The proof of Theorem 3.15 is given in Appendix C.5.1.

Remark 3.16. We note that Theorem 3.15 is different from many of the existing work on inverting expansive layers, e.g. (Aberdam et al., 2020; Bora et al., 2017; Lei et al., 2019), our result gives a direct inversion algorithm that is

provably the least-squares minimizer. Further, if each expansive layer is any combination of (R1), (R2), or (R3) then the entire network can be inverted end-to-end by using either the above result or solving the normal equations directly.

4. Conclusion

Bijective flow networks are a powerful tool for learning push-forward mappings in a space of fixed dimension. Increasingly, these flow networks have been used in combination with networks that increase dimension in order to produce networks which are purportedly universal.

In this work, we have studied the theory underpinning these flow and expansive networks by introducing two new notions, the embedding gap and the manifold embedding property. We show that these notions are both necessary and sufficient for proving universality, but require important topological and geometrical considerations which are, heretofore, under-explored in the literature. We also find that optimality of the studied networks can be established 'in reverse,' by minimizing the embedding gap, which we expect opens the door to convergence of layer-wise training schemes. Without compromising universality, we can also use specific expansive layers with a new layerwise projection result.

5. Acknowledgements

We would like to thank Anastasis Kratsios for his editorial input and mathematical discussions that helped us refine and trim our presentation; Pekka Pankka for his suggestion of the 'clean trick,' which was crucial to the development of the proof of Lemma 3.8; and Reviewer for supplying 5 and suggesting the addition of C.4.6.

I.D. was supported by the European Research Council Starting Grant 852821—SWING. M.L. was supported by Academy of Finland, grants 284715, 312110. M.V.dH. gratefully acknowledges support from the Department of Energy under grant DE-SC0020345, the Simons Foundation under the MATH + X program, and the corporate members of the Geo-Mathematical Imaging Group at Rice University.

References

Aberdam, A., Simon, D., and Elad, M. When and how can deep generative models be inverted? arXiv preprint arXiv:2006.15555, 2020.

Ambrosio, L. and Gigli, N. A user's guide to optimal transport. In Modelling and optimisation of flows on networks, volume 2062 of Lecture Notes in Math., pp. 1–155. Springer, Heidelberg, 2013. doi: 10.1007/978-3-642-32160-3_1. URL https://doi.org/10.1007/978-3-642-32160-3_1.

Ambrosio, L., Gigli, N., and Savaré, G. Gradient flows: in metric spaces and in the space of probability measures. Springer Science & Business Media, 2008.

Ardizzone, L., Kruse, J., Wirkert, S., Rahner, D., Pellegrini, E. W., Klessen, R. S., Maier-Hein, L., Rother, C., and Köthe, U. Analyzing inverse problems with invertible neural networks. arXiv preprint arXiv:1808.04730, 2018.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan. arXiv preprint arXiv:1701.07875, 2017.

Arridge, S., Maass, P., Öktem, O., and Schönlieb, C.-B. Solving inverse problems using data-driven models. Acta Numerica, 28:1–174, 2019.

Billingsley, P. Convergence of probability measures. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, second edition, 1999. ISBN 0-471-19745-9. doi: 10.1002/9780470316962. URL https://doi.org/10.1002/9780470316962. A Wiley-Interscience Publication.

Bora, A., Jalal, A., Price, E., and Dimakis, A. G. Compressed sensing using generative models. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 537–546. JMLR. org, 2017.

Brehmer, J. and Cranmer, K. Flows for simultaneous manifold learning and density estimation. arXiv preprint arXiv:2003.13913, 2020.

Bui Thi Mai, P. and Lampert, C. Functional vs. parametric equivalence of relu networks. In 8th International Conference on Learning Representations, 2020.

Cunningham, E., Zabounidis, R., Agrawal, A., Fiterau, I., and Sheldon, D. Normalizing flows across dimensions. arXiv preprint arXiv:2006.13070, 2020.

Dinh, L., Krueger, D., and Bengio, Y. Nice: Non-linear independent components estimation. arXiv preprint arXiv:1410.8516, 2014.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. arXiv preprint arXiv:1605.08803, 2016.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. Cubic-spline flows. arXiv preprint arXiv:1906.02145, 2019a.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. Neural spline flows. Advances in Neural Information Processing Systems, 32:7511–7522, 2019b.

Golub, G. H. Matrix computations. Johns Hopkins University Press, 1996.

Gomez, A. N., Ren, M., Urtasun, R., and Grosse, R. B. The reversible residual network: Backpropagation without storing activations. In Advances in neural information processing systems, pp. 2214–2224, 2017.

Grathwohl, W., Chen, R. T., Bettencourt, J., Sutskever, I., and Duvenaud, D. Ffjord: Free-form continuous dynamics for scalable reversible generative models. arXiv preprint arXiv:1810.01367, 2018.

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

Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. Neural autoregressive flows. In International Conference on Machine Learning, pp. 2078–2087. PMLR, 2018.

Jacobsen, J.-H., Smeulders, A., and Oyallon, E. i-revnet: Deep invertible networks. arXiv preprint arXiv:1802.07088, 2018.

Jaini, P., Selby, K. A., and Yu, Y. Sum-of-squares polynomial flow. In International Conference on Machine Learning, pp. 3009–3018. PMLR, 2019.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In Advances in Neural Information Processing Systems, pp. 10215–10224, 2018.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improving variational inference with inverse autoregressive flow. arXiv preprint arXiv:1606.04934, 2016.

Kobyzev, I., Prince, S., and Brubaker, M. Normalizing flows: An introduction and review of current methods. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

Kothari, K., Khorashadizadeh, A., de Hoop, M., and Dokmanić, I. Trumpets: Injective flows for inference and inverse problems. arXiv preprint arXiv:2102.10461, 2021.

Kruse, J., Detommaso, G., Scheichl, R., and Köthe, U. Hint: Hierarchical invertible neural transport for density estimation and bayesian inference. arXiv preprint arXiv:1905.10687, 2019.

Kruse, J., Ardizzone, L., Rother, C., and Köthe, U. Benchmarking invertible architectures on inverse problems. arXiv preprint arXiv:2101.10763, 2021.

Lee, H., Ge, R., Ma, T., Risteski, A., and Arora, S. On the ability of neural nets to express distributions. In Conference on Learning Theory, pp. 1271–1296. PMLR, 2017.

Lei, Q., Jalal, A., Dhillon, I. S., and Dimakis, A. G. Inverting deep generative models, one layer at a time. In Advances in Neural Information Processing Systems, pp. 13910–13919, 2019.

Lu, Y. and Lu, J. A universal approximation theorem of deep neural networks for expressing distributions. arXiv preprint arXiv:2004.08867, 2020.

Madsen, I. H., Tornehave, J., et al. From calculus to cohomology: de Rham cohomology and characteristic classes. Cambridge university press, 1997.

Milnor, J. On manifolds homeomorphic to the 7-sphere. Annals of Mathematics, pp. 399–405, 1956.

Mukherjee, A. Differential topology. Hindustan Book Agency, New Delhi; Birkhäuser/Springer, Cham, second edition, 2015. ISBN 978-3-319-19044-0; 978-3-319-19045-7. doi: 10.1007/978-3-319-19045-7. URL https://doi.org/10.1007/978-3-319-19045-7.

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

Murasugi, K. Knot theory & its applications. Modern Birkhäuser Classics. Birkhäuser Boston, Inc., Boston, MA, 2008. ISBN 978-0-8176-4718-6. doi: 10.1007/978-0-8176-4719-3. URL https://doi.org/10.1007/978-0-8176-4719-3. Translated from the 1993 Japanese original by Bohdan Kurpita, Reprint of the 1996 translation [MR1391727].

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. arXiv preprint arXiv:1912.02762, 2019.

Puthawala, M., Kothari, K., Lassas, M., Dokmanić, I., and de Hoop, M. Globally injective relu networks. arXiv preprint arXiv:2006.08464, 2020.

Rolnick, D. and Körding, K. Reverse-engineering deep relu networks. In International Conference on Machine Learning, pp. 8178–8187. PMLR, 2020.

Séquin, C. H. Tori story. In Sarhangi, R. and Séquin, C. H. (eds.), Proceedings of Bridges 2011: Mathematics, Music, Art, Architecture, Culture, pp. 121–130. Tessellations Publishing, 2011. ISBN 978-0-9846042-6-5.

Sun, H. and Bouman, K. L. Deep probabilistic imaging: Uncertainty quantification and multi-modal solution characterization for computational imaging. arXiv preprint arXiv:2010.14462, 2020.

Sutherland, W. A. Introduction to metric and topological spaces. Oxford University Press, Oxford, 2009. ISBN 978-0-19-956308-1. Second edition [of MR0442869], Companion web site: www.oup.com/uk/companion/metric.

Tao, T. Analysis, volume 185. Springer, 2009.

Teshima, T., Ishikawa, I., Tojo, K., Oono, K., Ikeda, M., and Sugiyama, M. Coupling-based invertible neural networks are universal diffeomorphism approximators. arXiv preprint arXiv:2006.11469, 2020.

Villani, C. Optimal transport: old and new, volume 338. Springer Science & Business Media, 2008.

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

Yarotsky, D. Optimal approximation of continuous functions by very deep relu networks. In Conference on Learning Theory, pp. 639–649. PMLR, 2018.

A. Summary of Notation

Throughout the paper we make heavy use of the following notation.

1. Unless otherwise stated, X and Y always refer to subsets of Euclidean space, and K and W always refer to compact subsets of Euclidean space.
2. $f \in C(X, Y)$ means that $f\colon X \to Y$ is continuous.
3. For families of functions $\mathcal{F}$ and $\mathcal{G}$ where each $\mathcal{F} \ni f\colon X \to Y$ and $\mathcal{G} \ni g\colon Y \to Z$, then we define $\mathcal{G} \circ \mathcal{F} = \{g \circ f\colon X \to Z\colon f \in \mathcal{F},\ g \in \mathcal{G}\}$.
4. $f \in \mathrm{emb}(X, Y)$ means that $f \in C(X, Y)$ is continuous and injective on the range of f, i.e. an embedding, and furthermore that $f^{-1}\colon f(X) \to X$ is continuous.
5. $\mu \in \mathcal{P}(X)$ means that μ is a probability measure over X.
6. $\mathrm{W}_2(\mu, \nu)$ for $\mu, \nu \in \mathcal{P}(X)$ refers to the Wasserstein-2 distance, always with ℓ_2 ground metric.
7. $\|\cdot\|_{L^p(X)}$ refers to the L^p norm of functions, from X to $\mathbb{R}$.
8. For vector-valued $f\colon X \to Y$, $\|f\|_{L^\infty(X)} = \operatorname{ess\,sup}_{x \in \mathcal{X}} \|f\|_2$. Note that Y is always finite dimensional, and so all discrete $1 \leq q \leq \infty$ norms are equivalent.
9. $\mathrm{Lip}(g)$ refers to the Lipschitz constant of f.
10. For $x \in \mathbb{R}^n$, $[x]_i \in \mathbb{R}$ is the i'th component of x. Similarly, for matrix $A \in \mathbb{R}^{m \times n}$, $[A]_{ij}$ refers to the j'th element in the i'th column.

B. Detailed Comparison to Prior work

B.1. Connection to Brehmer & Cranmer (2020)

In (Brehmer & Cranmer, 2020), the authors introduce manifold-learning flows as an invertible method for learning probability density supported on a low-dimensional manifold. Their model can be written as

$$\mathcal{F} = \mathcal{T}_1^m \circ \mathcal{R}^{n,m} \circ \mathcal{T}_0^n \tag{18}$$

where $\mathcal{T}_1^m \subset C(\mathbb{R}^m, \mathbb{R}^m)$, $\mathcal{T}_0^m \subset C(\mathbb{R}^n, \mathbb{R}^n)$, and $\mathcal{R} = \left\{ \begin{bmatrix} I^{n \times n} \\ \mathbf{0}^{(m-n) \times n} \end{bmatrix} \right\}$ is a zero-padding (R1). They invert $f \in \mathcal{F}$ in two different ways. For manifold-learning flows ($\mathcal{M}$-flows) they restrict $\mathcal{T}_1^m$ to be an invertible flow, and for manifold-learning flows with separate encoder ($\mathcal{M}_e$-flows) they place no such restrictions on $\mathcal{T}_1^m$ and instead train a separate neural network e to invert elements of $\mathcal{T}_1^m$.

Our results apply out-of-the-box to the architectures used in Experiment A of (Brehmer & Cranmer, 2020). The architecture described in Eqn. 18 is of the form of Eqn. 1 where $L = 1$. Further, although they are not studied here, our analysis can also be applied to quadratic flows.

The network used in (Brehmer & Cranmer, 2020, Experiment 4.A) uses coupling networks, (T1), where $\mathcal{T}_1^m$ and $\mathcal{T}_0^n$ are both 5 layers deep. For (Brehmer & Cranmer, 2020, Experiments 4.B and 4.C) the authors choose expressive elements $\mathcal{T}$ that are rational quadratic flows (Durkan et al., 2019b) for both $\mathcal{T}_1^m$ and $\mathcal{T}_0^n$. In Experiment 4.B they let T_1 and T_0 again be 5 layers deep, and in 4.C they again let T_1 by 20 layers deep and T_0 15 layers. For the final experiment, 4.D, the choose more complicated expressive elements that combine Glow (Kingma & Dhariwal, 2018) and Real NVP (Dinh et al., 2016) architectures. These elements include the actnorm, 1×1 convolutions and rational-quadratic coupling transformations along with a multi-scale transformation.

The authors mention universality of their network without our proof, but our universality results in Theorem 3.10 apply to their networks from Experiment A wholesale. Further in their work the authors describe how training can be split into a manifold phase and density phase, wherein the manifold phase $\mathcal{T}_1^m$ is trained to learn the manifold, and in the density phase $\mathcal{T}_1^m$ if fixed and $\mathcal{T}_0^n$ is trained to learn the density thereupon. This statement is made formal and proven by our Cor. 3.12.

B.2. Connection to Kothari et al. (2021)

In (Kothari et al., 2021), the authors introduce the 'Trumpet' architecture, for its architecture, which has many alternating flow networks & expansive layers with many flow-networks in the low-dimensional early stages of the network, which gives the architecture a shape similar to the titular instrument.

The architecture studied in (Kothari et al., 2021) is precisely of the form of Eqn. 1, where the bijective flow networks are revnets (Gomez et al., 2017; Jacobsen et al., 2018) architecture, and the expansive elements are 1×1 convolutions, as in (R2). To out knowledge, there are no results that show that the revnets used are universal approximators, but if they revnets are substituted with either (T1) or (T2), then the, we could apply Theorem 3.10 to the resulting architecture.

C. Proofs

C.1. Main Results

C.1.1. Embedding Gap

To aid all of our subsequent proofs, we first present the following lemma which present inequalities and identities for the embedding gap.

Lemma C.1. *For all of the following results, $f \in \mathrm{emb}(K, \mathbb{R}^m)$ and $g \in \mathrm{emb}(W, \mathbb{R}^m)$ and $n \leq o \leq m$.*

1. $$B_{K,W}(f,g) \geq \sup_{x_n \in K} \inf_{x_o \in W} \|g(x_o) - f(x_n)\|_2 . \tag{19}$$

2. *Let $X, Y \subset W$, let g be Lipschitz on W, and $r \in \mathrm{emb}(X, Y)$. Then, there is a $r' \in \mathrm{emb}(g(X), g(Y))$ such that $g \circ r = r' \circ g$ and $\|I - r'\|_{L^\infty(g(X))} \leq \|I - r\|_{L^\infty(X)} \mathrm{Lip}(g)$.*

3. $$\|I - r\|_{L^\infty(K)} = \left\|I - r^{-1}\right\|_{L^\infty(r(K))} \tag{20}$$

4. *Let $K \subset \mathbb{R}^n$, $X \subset \mathbb{R}^p$ and $W \subset \mathbb{R}^o$ be compact sets. Also, let $f \in \mathrm{emb}(K, W)$ and $h \in \mathrm{emb}(X, W)$, and let $g \in \mathrm{emb}(W, \mathbb{R}^m)$ be a Lipschitz map. Then*
$$B_{K,X}(g \circ f, g \circ h) \leq \mathrm{Lip}(g) B_{K,X}(f,h). \tag{21}$$

5. *$B_{K,W}(f,g) \leq \sup_{x \in K} \|g \circ h(x) - f(x)\|_2$ where $h \in \mathrm{emb}(K, \mathbb{R}^o)$ is a map satisfying $h(K) \subset W$.*

6. *For any X that is the closure of an open set , if $h \in \mathrm{emb}(X, W)$ then*
$$B_{K,W}(f,g) \leq B_{K,X}(f, g \circ h) \tag{22}$$

7. *For any $r \in \mathrm{emb}(f(K), \mathbb{R}^m)$,*
$$B_{K,W}(f,g) \leq \|I - r\|_{L^\infty(f(K))} + B_{K,W}(r \circ f, g). \tag{23}$$

8. *For any $r \in \mathrm{emb}(f(K), g(W))$ and $h \in \mathrm{emb}(X, W)$ where $X \subset \mathbb{R}^p$ is the closure of a set U which is open in the subspace topology of some vector space of dimension p, where $n \leq p \leq o$ we have that*
$$B_{K,X}(f, g \circ h) \leq \|I - r\|_{L^\infty(f(K))} + \mathrm{Lip}(g) B_{K,X}(g^{-1} \circ r \circ f, h) \tag{24}$$
where $\mathrm{Lip}(g)$ denotes the Lipschitz constant of g.

9. *For any $\mu_n \in \mathcal{P}(K)$ there is a $\mu_o \in \mathcal{P}(W)$ such that*
$$\mathrm{W}_2\left(f_{\#}\mu_n, g_{\#}\mu_o\right) \leq B_{K,W}(f,g) \tag{25}$$

Proof. 1. Let $r \in C(f(K), g(W))$, then

$$\begin{aligned}\|I - r\|_{L^\infty(f(K))} &= \sup_{x_n \in K} \|(I-r)f(x_n)\|_2 = \sup_{x_n \in K} \|f(x_n) - r \circ f(x_n)\|_2 \\ &= \sup_{x_n \in K} \|f(x_n) - g(x_o)\|_2 \text{ where } x_o = g^{-1} \circ r \circ f(x_n) \\ &\geq \sup_{x_n \in K} \inf_{x_o \in W} \|f(x_n) - g(x_o)\|_2 .\end{aligned}$$

2. g is injective on X, hence we can define r' such that $r' = g \circ r \circ g^{-1} : g(X) \to g(r(X)) \subset g(Y)$ such that $r' \in \mathrm{emb}(g(X), g(Y))$, and thus $\forall x \in X$,

$$\|(I - r') \circ g(x)\|_2 = \|g(x) - g \circ r(x)\|_2 \leq \mathrm{Lip}(g) \, \|I - r\|_{L^\infty(X)} \tag{26}$$

where we have used $\|r(x) - x\|_2 \leq \|I - r\|_{L^\infty(X)}$.

3. For every $x \in r(K)$, we have a $y \in K$ such that $x = r(y)$, thus $\forall x \in r(K)$,

$$\left\| \left(I - r^{-1}\right)(x) \right\|_2 = \|(r - I)(y)\|_2 . \tag{27}$$

But r is clearly surjective onto it's range, hence taking the supremum over all $x \in X$ yields

$$\left\| I - r^{-1} \right\|_{L^\infty(r(K))} = \|I - r\|_{L^\infty(K)} \tag{28}$$

4. As $g \in \mathrm{emb}(W, \mathbb{R}^m)$, the map $g : W \to g(W)$ is a homeomorphism and there is $g^{-1} \in \mathrm{emb}(g(W), W)$. For a map $r \in \mathrm{emb}(g \circ f(K), g \circ h(X))$, we see that $\hat{r} = g^{-1} \circ r \circ g \in \mathrm{emb}(f(K), h(X))$. Also, the opposite is valid as if $\hat{r} \in \mathrm{emb}(f(K), h(X))$ then $r = g \circ \hat{r} \circ g^{-1} \in \mathrm{emb}(g \circ f(K), g \circ h(X))$. Thus

$$\begin{aligned}B_{K,X}(g \circ f, g \circ h) &= \inf_{r \in \mathrm{emb}(g \circ f(K), g \circ h(X))} \|I - r\|_{L^\infty(g \circ f(K))} \\ &= \inf_{r = g \circ \hat{r} \circ g^{-1} \in \mathrm{emb}(g \circ f(K), g \circ h(X))} \|I - g \circ \hat{r} \circ g^{-1}\|_{L^\infty(g \circ f(K))} \\ &= \inf_{\hat{r} \in \mathrm{emb}(f(K), h(X))} \|g \circ (I - \hat{r}) \circ g^{-1}\|_{L^\infty(g \circ f(K))} \\ &\leq \mathrm{Lip}(g) \inf_{\hat{r} \in \mathrm{emb}(f(K), h(X))} \|(I - \hat{r}) \circ g^{-1}\|_{L^\infty(g \circ f(K))} \\ &\leq \mathrm{Lip}(g) \inf_{\hat{r} \in \mathrm{emb}(f(K), h(X))} \|I - \hat{r}\|_{L^\infty(f(K))} \\ &\leq \mathrm{Lip}(g) \, B_{K,X}(f, h)\end{aligned}$$

5. If we let $r \coloneqq g \circ h \circ f^{-1}$, then $r \in \mathrm{emb}(f(K), g(W))$, and

$$B_{K,W}(f, g) \leq \| \|(I - r) \circ f(x)\|_2 \|_{L^\infty(K)} \tag{29}$$

$$= \left\| \left\| f(x) - g \circ h \circ f^{-1} \circ f(x) \right\|_2 \right\|_{L^\infty(K)} \leq \sup_{x \in K} \|f(x) - g \circ h(x)\|_2 . \tag{30}$$

6. Given that $g \circ h(X) \subset g(W)$, we have that $\mathrm{emb}(f(K), g \circ h(X)) \subset \mathrm{emb}(f(K), g(W))$, thus the infimum in Eqn. 6 is taken over a smaller set, thus $B_{K,W}(f, g) \leq B_{K,X}(f, g \circ h)$.

7. Note that for any $r' \in \mathrm{emb}(r \circ f(K), g(W))$, $r' \circ r \in \mathrm{emb}(f(K), g(W))$, and so we have

$$B_{K,W}(f, g) \leq \|I - r' \circ r\|_{L^\infty(f(K))} \leq \|I - r\|_{L^\infty(f(K))} + \|r - r' \circ r\|_{L^\infty(f(K))} \tag{31}$$

$$= \|I - r\|_{L^\infty(f(K))} + \|I - r'\|_{L^\infty(r \circ f(K))} \tag{32}$$

where we have used that r is injective for the final equality. This holds for all possible r', hence we have the result.

8. Recall that $f \in \mathrm{emb}(K, W)$, $g \in \mathrm{emb}(W, \mathbb{R}^m)$, $h \in \mathrm{emb}(X, W)$ and $r \in \mathrm{emb}(f(K), g(W))$. Then $g^{-1} \in \mathrm{emb}(g(W), W)$. As $r \circ f(K) \subset g(W)$, we see that

$$r \circ f = g \circ g^{-1} \circ r \circ f.$$

Thus Lemma C.1 points 4 and 8 yield that

$$\begin{aligned} B_{K,X}(f, g \circ h) &\leq \|I - r\|_{L^\infty(f(K))} + B_{K,X}(r \circ f, g \circ h) \\ &\leq \|I - r\|_{L^\infty(f(K))} + B_{K,X}(g \circ g^{-1} \circ r \circ f, g \circ h) \\ &\leq \|I - r\|_{L^\infty(f(K))} + \mathrm{Lip}(g)\, B_{K,X}(g^{-1} \circ r \circ f, h), \end{aligned}$$

which proves the claim.

9. Let $r_\epsilon \in \mathrm{emb}(f(K), g(W))$ be such that $\|I - r_\epsilon\|_{L^\infty(\mathrm{Range}(f))} \leq B_{K,W}(f, g) + \epsilon$, then for every $x \in K$, there exists $y \in W$ such that $g(y) = r \circ f(x)$. From injectivity of g, we have that $y = g^{-1} \circ r_\epsilon \circ f(x)$. Note that $g^{-1} \circ r_\epsilon \circ f \in \mathrm{emb}(K, W)$, hence $K' \coloneqq g^{-1} \circ r_\epsilon \circ f(K) \subset W$ is compact. Define $\mu'_\epsilon \in \mathcal{P}(K')$ where $\mu'_\epsilon \coloneqq (g^{-1} \circ r_\epsilon \circ f)_{\#}\mu$. Clearly $g_{\#}\mu'_\epsilon = r_\epsilon \circ f_{\#}\mu$, and thus

$$\mathrm{W}_2\left(f_{\#}\mu, g_{\#}\mu'_\epsilon\right) = \mathrm{W}_2\left(f_{\#}\mu, r_\epsilon \circ f_{\#}\mu\right) \tag{33}$$

and so

$$\mathrm{W}_2\left(f_{\#}\mu, g_{\#}\mu'_\epsilon\right) \leq \left(\int_K \|I - r_\epsilon\|_2^2 \, df_{\#}\mu\right)^{1/2} \leq B_{K,W}(f, g) + \epsilon. \tag{34}$$

As the set W is compact, by Prokhoros's theorem, see (Billingsley, 1999, Theorem 5.1), the set of probability measures $P(W)$ is a compact set in the topology of weak convergence. Thus there is a sequence $\epsilon_i \to 0$ such that the measures μ'_{ϵ_i} converge weakly to a probability measure μ_o. As $g : W \to K$ is a continuous function, the push-forward operation $\mu \to g_{\#}\mu$ is continuous $g_{\#} : P(W) \to P(K)$ and thus $g_{\#}\mu'_{\epsilon_i}$ converge weakly to $g_{\#}\mu_o$. Finally, as $g_{\#}\mu'_{\epsilon_i}$ are supported in a compact set K, their second moments converge to those of $g_{\#}\mu_o$ as $i \to \infty$. By (Ambrosio & Gigli, 2013), Theorem 2.7, see also Remark 28, the weak convergence and the convergence of the second moments imply the convergence in the Wasserstein-2 metric. Hence, $g_{\#}\mu'_{\epsilon_i}$ converge to $g_{\#}\mu_o$ in Wasserstein-2 metric and we see that

$$\mathrm{W}_2\left(f_{\#}\mu, g_{\#}\mu_o\right) \leq B_{K,W}(f, g). \tag{35}$$

□

C.2. Manifold Embedding Property

C.2.1. The Proof of Lemma 3.4

The proof of Lemma 3.4. Let $f = F_2 \circ F_1$ where $F_2 \in \mathcal{F}^{o,m}$ and $F_1 \in \mathcal{F}^{n,o}$ and $\epsilon > 0$ be given, and let $E^{o,m}$. Clearly, $B_{K,W}(f, E) \leq B_{K,W}(F_2, E)$ and so by the m, o, o MEP of $\mathcal{E}^{o,m}$ with respect to $\mathcal{F}^{o,m}$, we have the existence of an $r_m \in \mathrm{emb}(f(K), E^{o,m})$ such that $\|I - r\|_{L^\infty(f(K))} < \epsilon$. $K_o \coloneqq \left(E^{o,m}\right)^{-1} \circ r \circ f(K)$ is compact, hence $E^{o,m}$ is Lipschitz on K_o, so we can apply Lemma C.1 point 8, so

$$B_{K,W}(f, E^{o,m} \circ E^{p,o}) \leq \|I - r\|_{L^\infty(f(K))} + \mathrm{Lip}(E^{o,m}) B_{K,W}(\left(E^{o,m}\right)^{-1} \circ r \circ f, E^{p,o}). \tag{36}$$

But, because $f \in \mathcal{F}^{o,m} \circ \mathcal{F}^{n,o}$, we can choose a $E^{p,o} \in \mathcal{E}_1^{p,o}$ so that $B_{K,W}(\left(E^{o,m}\right)^{-1} \circ r \circ f, E^{p,o}) \leq \frac{\epsilon}{2} \mathrm{Lip}(E^{o,m})$ which, combined with Eqn. 36, proves the result. □

C.2.2. The Proof of Lemma 3.5

The proof of Lemma 3.5. Recall that $\mathcal{F} \subset \mathrm{emb}(\mathbb{R}^n, \mathbb{R}^m)$. Suppose that $\mathcal{E}_2^{o,m}$ does not have the m, n, o MEP with respect to $\mathcal{F}$, then there are some $\epsilon > 0$ and $f \in \mathcal{F}$ so that

$$\forall E^{o,m} \in \mathcal{E}_2^{o,m} \; \forall W_1 \subset\subset \mathbb{R}^o, \quad B_{K,W_1}(f, E_2^{o,m}) \geq \epsilon. \tag{37}$$

From Lemma C.1 point 6, we have that

$$\epsilon \leq B_{K,W_1}(f, E_2^{o,m}) \leq B_{K,W}(f, E_2^{o,m} \circ E_1^{p,o}) \tag{38}$$

for all $E_1^{p,o} \in \mathcal{E}_1^{p,o}$ and for all compact sets $W \subset \mathbb{R}^p$ that satisfy $E_1^{p,o}(W_1) \subset W$. We observe that if $W' \subset \mathbb{R}^p$ is a compact set such that $W' \subset W$, we have

$$B_{K,W}(f, E_2^{o,m} \circ E_1^{p,o}) \leq B_{K,W'}(f, E_2^{o,m} \circ E_1^{p,o})$$

Thus, inequality Eq. 38 holds for all $E_1^{p,o} \in \mathcal{E}_1^{p,o}$ and for all compact sets $W \subset \mathbb{R}^p$. Summarising, we have seen that there are $f \in \mathcal{F}$ and $\epsilon > 0$ such that for all $E_1^{p,o} \in \mathcal{E}_1^{p,o}$ and for all compact sets $W \subset \mathbb{R}^p$ we have $\epsilon \leq B_{K,W}(f, E_2^{o,m} \circ E_1^{p,o})$. Hence $\mathcal{E}_2^{o,m}$ does not have the m, n, o MEP with respect to $\mathcal{F}$, and we have obtained a contradiction, which proves the result. □

C.3. Topological Obstructions to Manifold Learning with Neural Networks

C.3.1. S^1 CAN NOT BE MAPPED EXTENDABLY TO THE TREFOIL KNOT

We first show that there are no maps $E \coloneqq T \circ R$ where $R\colon \mathbb{R}^2 \to \mathbb{R}^3$ such that T is a homeomorphism and $E(S^1)$ is a trefoil knot. We use the fact that the trivial knot S^1 and the trefoil knot $\mathcal{M} = f(S^1)$ are not equivalent, that is, there are no homeomorphisms in $\mathbb{R}^3$ that map S^1 to $\mathcal{M}$. Indeed, by (Murasugi, 2008, Section 3.2), the trefoil knot $\mathcal{M}$ and its mirror image are not equivalent, whereas the trivial knot S^1 and its mirror image are equivalent. Hence, $\mathcal{M}$ and $R(S^1)$ are not equivalent knots in $\mathbb{R}^3$. Thus by (Murasugi, 2008, Definition 1.3.1 and Theorem 1.3.1), we see that there is no orientation preserving homeomorphism $T : \mathbb{R}^3 \to \mathbb{R}^3$ such that $T(\mathbb{R}^3 \setminus R(S^1)) = \mathbb{R}^3 \setminus \mathcal{M}$. As the orientation of the map T can be changed by composing T with the reflection $J : \mathbb{R}^3 \to \mathbb{R}^3$ across the plane $\mathrm{Range}(R)$ that defines a homeomorphism $J : \mathbb{R}^3 \setminus R(S^1) \to \mathbb{R}^3 \setminus R(S^1)$, we see that there is no homeomorphism $T : \mathbb{R}^3 \to \mathbb{R}^3$ such that $T(\mathbb{R}^3 \setminus R(S^1)) = \mathbb{R}^3 \setminus \mathcal{M}$.

This example shows that the composition $E = T \circ R$ of a linear map R and a coupling flow T cannot have the property that $E(S^1) = f(S^1)$ for this embedding f. Moreover, the complement $\mathbb{R}^3 \setminus E(S^1)$ is never homeomorphic to $\mathbb{R}^3 \setminus f(S^1)$ for any such map E.

We now construct another example, similar to Figure 2, where an annulus that is mapped to a knotted ribbon in $\mathbb{R}^3$. To do this, replace the circle S^1 by an annulus $K = \{x \in \mathbb{R}^2 : \ 1/2 \leq |x| \leq 3/2\}$, that in the polar coordinates is $\{(r, \theta) : \ 1/2 \leq r \leq 3/2\}$ and define a map $F : K \to \mathbb{R}^3$ by defining in the polar coordinates

$$F(r, \theta) = f(\theta) + a(r - 1)v(\theta)$$

where $f : S^1 \to \Sigma_1 \subset \mathbb{R}^3$ is an smooth embedding of S^1 to a trefoil knot Σ_1 and $v(\theta) \in \mathbb{R}^3$ is a unit vector normal to Σ_1 at the point $f(\theta)$ such that $v(\theta)$ is a smooth function of θ, and $a > 0$ is a small number. In this case, $M_1 = F(K)$ is a 2-dimensional submanifold of $\mathbb{R}^3$ with boundary, which can visualizes M_1 as a knotted ribbon.

We now show that there are no maps $E = T \circ R$ such that $E(K) = F(K)$ where $T\colon \mathbb{R}^3 \to \mathbb{R}^3$ is an embedding, and $R\colon \mathbb{R}^2 \to \mathbb{R}^3$ injective and linear. The key insight is that if such a T existed, then this implies that the trefoil knot is equivalent to S^1 in $\mathbb{R}^3$, which is known to be false.

Let $U_\rho(A)$ denote the ρ-neighborhood of the set A in $\mathbb{R}^3$. It is easy to see that $\mathbb{R}^2 \setminus (\{0\} \times [-1, 1])$ is homeomorphic to $\mathbb{R}^2 \setminus \overline{B}_{\mathbb{R}^2}(0, 1)$, which is further homeomorphic to $\mathbb{R}^2 \setminus \{0\}$. Thus, using tubular coordinates near Σ_1 and a sufficiently small $\rho > 0$, we see that $\mathbb{R}^3 \setminus M_1$ is homeomorphic to $\mathbb{R}^3 \setminus U_\rho(\Sigma_1)$, which is further homeomorphic to $\mathbb{R}^3 \setminus \Sigma_1$. Also, when $R : \mathbb{R}^2 \to \mathbb{R}^3$ is an injective linear map, we see that $M_2 = R(K)$ is a un-knotted band in $\mathbb{R}^3$ and $\mathbb{R}^3 \setminus M_2$ is homeomorphic to $\mathbb{R}^3 \setminus \Sigma_2$. If $\mathbb{R}^3 \setminus M_1$ and $\mathbb{R}^3 \setminus M_2$ would be homeomorphic, then also $\mathbb{R}^3 \setminus \Sigma_1$ and $\mathbb{R}^3 \setminus \Sigma_2$ would be homeomorphic that is not possible by knot theory, see (Murasugi, 2008, Definition 1.3.1 and Theorem 1.3.1). This shows that there are no injective linear maps $R : \mathbb{R}^2 \to \mathbb{R}^3$ and homeomorphisms $\Phi : \mathbb{R}^3 \to \mathbb{R}^3$ such that $(\Phi \circ R)(K) = M_1$.

Similar examples can be obtained in a higher dimensional case by using a knotted torus (Séquin, 2011)[6] and their Cartesian products.

[6]On the knotted torus, see `http://gallery.bridgesmathart.org/exhibitions/2011-bridges-conference/sequin`.

C.3.2. LINEAR HOMEOMORPHISM COMPOSITION

In this subsection we prove that the topological obstructions to universality presented in Section 3.3 still apply when the expansive elements are allowed to be $\hom(\mathbb{R}^3, \mathbb{R}^3) \circ \mathbb{R}^{3\times 2}$. This fact follows from the observation that $\hom(\mathbb{R}^3, \mathbb{R}^3) \circ \hom(\mathbb{R}^3, \mathbb{R}^3) = \hom(\mathbb{R}^3, \mathbb{R}^3)$, which yields that $\hat{\mathcal{E}} = \mathcal{E}$.

C.3.3. THE PROOF OF THEOREM 3.8

Given an $f \in \mathrm{emb}^k(K, \mathbb{R}^m)$, for $k \geq 1$, we first show that for $m \geq 2n+1$ there is always a diffeomorphism $\Psi \colon \mathbb{R}^m \to \mathbb{R}^m$ so that $\Psi \circ f \colon \mathbb{R}^n \to \{0\}^n \times \mathbb{R}^{m-n}$. The existence of such a Ψ borrows ideas from Whitney's embedding theorem (Hirsch, 2012, Theorems 3.4 & 3.5) and is constructed by iteratively constructing an injective projection.

Next if $m - n \geq 2n + 1$, then we can apply (Madsen et al., 1997, Lemma 7.6), a result analogous to the Tietze extension theorem, to show that $\Psi \colon \mathcal{M} \to \{0\}^n \times \mathbb{R}^{m-n}$ can be extended to a diffeomorphism on the entire space, $h \colon \mathbb{R}^m \to \mathbb{R}^m$. Hence $f(x) = \Psi^{-1} \circ h \circ R(x)$ for diffeomorphism $\Psi^{-1} \circ h \colon \mathbb{R}^m \to \mathbb{R}^m$ and zero-padding operator $R \colon \mathbb{R}^n \to \mathbb{R}^m$, and thus $f \in \mathcal{I}^k(K, \mathbb{R}^m)$. This fact that for m sufficiently large compared to n such a diffeomorphism can always be extended is related to the fact that in 4-dimensions, all knots can be opened. This can be contrasted with the case in Figure 2.

We now present our proof.

Proof. Let us next prove Eq. 9 when $m \geq 3n + 1$. Let

$$f \in \mathrm{emb}^k(\mathbb{R}^n, \mathbb{R}^m) \tag{39}$$

be a C^k map and $\mathcal{M} = f(\mathbb{R}^n)$ be an embedded submanifold of $\mathbb{R}^m$.

We have that $m \geq 3n + 1 > 2n + 1$. Let S^{m-1} be the unit sphere of $\mathbb{R}^m$ and let

$$S\mathbb{R}^m = \{(x, v) \in \mathbb{R}^m \times \mathbb{R}^m : \ \|v\| = 1\}$$

be the sphere bundle of $\mathbb{R}^m$ that is a manifold of dimension $2m - 1$. By the proof's of Whitney's embedding theorem, by Hirsch, (Hirsch, 2012, Chapter 1, Theorems 3.4 and 3.5), there is a set of 'problem points' $H_1 \subset S^{m-1}$ of Hausdorff dimension $2n$ such that for all $w \in \mathbb{R}^m \setminus H_1$ the orthogonal projection

$$P_w : \mathbb{R}^m \to \{w\}^\perp = \{y \in \mathbb{R}^m : \ y \perp w\}$$

has a restriction $P_w|_{\mathcal{M}}$ on $\mathcal{M}$ defines an injective map

$$P_w|_{\mathcal{M}} : \mathcal{M} \to \{w\}^\perp.$$

Moreover, let $T_x\mathcal{M}$ be the tangent space of manifold $\mathcal{M}$ at the point x and let us define another set of 'problem points' as

$$H_2 = \{v \in S^{m-1} : \ \exists x \in \mathcal{M}, v \in T_x\mathcal{M}\}.$$

For $w \in S^{m-1} \setminus H_2$ the map

$$P_w|_{\mathcal{M}} : \mathcal{M} \to \{w\}^\perp \subset \mathbb{R}^m$$

is an immersion, that is, it has an injective differential. The sphere tangent bundle $S\mathcal{M}$ of $\mathcal{M}$ has dimension $2n - 1$, and the set H_2 has the Hausdorff dimension at most $2n-1$. Thus $H = H_1 \cup H_2$ has Hausdorff dimension at most $2n < m-1$ and hence the set $S^{m-1} \setminus H$ is non-empty. For $w \in S^{m-1} \setminus H$ the map $P_w|_{\mathcal{M}} : \mathcal{M} \to \{w\}^\perp$ is a C^k injective immersion and thus

$$\tilde{N} = P_w(\mathcal{M}) \subset \{w\}^\perp$$

is a C^k submanifold.

Let $Z : P_w(\mathcal{M}) \to \mathcal{M}$ be the C^k function defined by

$$Z(y) \in \mathcal{M}, \quad P_w(Z(y)) = y,$$

that is it is the inverse of $P_w|_{\mathcal{M}} : \mathcal{M} \to P_w(\mathcal{M})$, where $P_w(\mathcal{M}) \subset \{w\}^\perp$. Let $g : \tilde{N} = P_w(\mathcal{M}) \to \mathbb{R}$ be the function

$$g(y) = (Z(y) - y) \cdot w, \quad y \in P_w(\mathcal{M}).$$

Then $\tilde{N}$ is a n-dimensional C^k submanifold of $(m-1)$-dimensional Euclidean space $H = \{w\}^\perp$ and g is a C^k function defined on it. By definition of a C^k submanifold of H, any point $x \in \tilde{N}$ has a neighborhood $U \subset H$ with local C^k coordinates $\psi : U \to \mathbb{R}^m$ such that $\psi(\tilde{N} \cap U) = (\{0\}^{m-1-n} \times \mathbb{R}^n) \cap \psi(U)$. Using these coordinates, we see that g can be extended to a C^k function in U. Using a suitable partition of unity, we see that there is a C^k map $G : \{w\}^\perp \to \mathbb{R}$ that a C^k extension of g that is, $G|_{\tilde{N}} = g$.

Then the map

$$\Phi_1 : \mathbb{R}^m \to \mathbb{R}^m, \quad \Phi_1(x) = x - G(P_w(x))w$$

is a C^k diffeomorphism of $\mathbb{R}^m$ that maps $\mathcal{M}$ to $m-1$ dimensional space $\{w\}^\perp$, that is

$$\Phi_1(\mathcal{M}) \subset \{w\}^\perp.$$

In the case when $m \geq 3n+1$, we can repeat this construction n times. This is possible as $m - n \geq 2n + 1$. Then we obtain C^k diffeomorphisms $\Phi_j : \mathbb{R}^m \to \mathbb{R}^m$, $j = 1, \ldots, n$ such that their composition $\Phi_n \circ \cdots \circ \Phi_1 : \mathbb{R}^m \to \mathbb{R}^m$ is a C^k-diffeomorphism such that which

$$\mathcal{M}' = \Phi_n \circ \cdots \circ \Phi_1(\mathcal{M}) \subset Y',$$

where $Y' \subset \mathbb{R}^m$ is a $m-n$ dimensional linear space. By letting $\Psi = Q \circ \Phi_n \circ \cdots \circ \Phi_1$ for rotation matrix $Q \in \mathbb{R}^{m \times m}$, we have that $Y \coloneqq Q(Y') = \{0\}^n \times \mathbb{R}^{m-n}$. Also, let $X = \mathbb{R}^n \times \{0\}^{m-n}$, $A = Q(\mathcal{M}') \subset X$ and $\phi : X \to \mathbb{R}^m$ be the map

$$\phi(x, 0) = \Psi(f(x)) \in Y,$$

where f is the function given in Eq. 39 and $B = \Psi(f(A)) \subset Y$. Then A is a C^k-submanifold X, B is a C^k-submanifold Y and $\phi : A \to B$ is a C^k-diffeomorphism. We observe that $m - n \geq 2n + 1$ and so we can apply (Madsen et al., 1997, Lemma 7.6) to extend ϕ to a C^k-diffeomorphism

$$h : \mathbb{R}^m \to \mathbb{R}^m$$

such that $h|_A = \phi$. Note that (Madsen et al., 1997, Lemma 7.6) concerns an extension of a homeomorphism, but as the extension h is given by an explicit formula which is locally a finite sum of C^k functions, the same proof gives a C^k-diffeomorphic extension h to a diffeomorphism ϕ. Indeed, let $A' \subset \mathbb{R}^n$ and $B' \subset \mathbb{R}^{m-n}$ be such sets that $A = A' \times \{0\}^{m-n}$, and $B = \{0\}^n \times B'$. Moreover, let $\tilde{\phi} : A' \to \mathbb{R}^{n-m}$ and $\tilde{\psi} : B' \to \mathbb{R}^n$ be such C^k-smooth maps that $\phi(x, 0) = (0, \tilde{\phi}(x))$ for $(x, 0) \in A$ and $\phi^{-1}(0, y) = (\tilde{\psi}(y))$ for $(0, y) \in B$. As A' and B' are C^k-submanifolds, the map $\tilde{\phi}$ has a C^k-smooth extension $f_1 : \mathbb{R}^n \to \mathbb{R}^{n-m}$ and the map $\tilde{\psi}$ has a C^k-smooth extension $f_2 : \mathbb{R}^{n-m} \to \mathbb{R}^n$, that is, $f_1|_{A'} = \tilde{\phi}$ and $f_2|_{B'} = \tilde{\psi}$. Following (Madsen et al., 1997, Lemma 7.6), we define the maps $h_1 : \mathbb{R}^n \times \mathbb{R}^{m-n} \to \mathbb{R}^n \times \mathbb{R}^{m-n}$,

$$h_1(x, y) = (x, y + f_1(x))$$

and $h_2 : \mathbb{R}^n \times \mathbb{R}^{m-n} \to \mathbb{R}^n \times \mathbb{R}^{m-n}$,

$$h_2(x, y) = (x + f_2(y), y).$$

Observe that h_2 has the inverse map $h_2^{-1}(x, y) = (x - f_2(y), y)$. Then the map

$$h = h_2^{-1} \circ h_1 : \mathbb{R}^n \times \mathbb{R}^{m-n} \to \mathbb{R}^n \times \mathbb{R}^{m-n}$$

is a C^k-diffeomorphism that satisfies $h|_A = \phi$. This technique is called the 'clean trick'.

Finally, to obtain the claim, we observe that when $R : \mathbb{R}^n \to \mathbb{R}^m$, $R(x) = (x, 0) \in \{0\}^n \times \mathbb{R}^{m-n}$ is the zero padding operator, we have

$$f(x) = \Psi^{-1}(\phi(R(x))), \quad x \in \mathbb{R}^n.$$

As $h|_X = \phi$ and $R(x) \in X$, this yields

$$f(x) = \Psi^{-1}(h(R(x))), \quad x \in \mathbb{R}^n,$$

that is,

$$f = E \circ R$$

where $E = \Psi^{-1} \circ h : \mathbb{R}^m \to \mathbb{R}^m$ is a C^k diffeomorphism. Thus $f \in \mathcal{I}^k(\mathbb{R}^n, \mathbb{R}^m)$. This proves Eq. 9 when $m \geq 3n + 1$. □

C.4. Universality

C.4.1. The Proof of Lemma 3.9

The proof of Lemma 3.9. (i) Let us consider $\epsilon > 0$, a compact set $K \subset \mathbb{R}^n$ and $f \in \mathrm{emb}(\mathbb{R}^n, \mathbb{R}^m)$. Let $W = K \times \{0\}^{o-n}$ and $F : \mathbb{R}^o \to \mathbb{R}^m$ be the map given by $F(x, y) = f(x)$, $(x, y) \in \mathbb{R}^n \times \mathbb{R}^{o-n}$. Because $\mathcal{R}^{o,m} \subset \mathrm{emb}(\mathbb{R}^o, \mathbb{R}^m)$ is a uniform universal approximator of $C(\mathbb{R}^n, \mathbb{R}^m)$, there is an $R \in \mathcal{R}^{o,m}$ such that $\|F - R\|_{L^\infty(W)} < \epsilon$. Then for the map $E = I \circ R$ we have that $B_{K,W}(f, E) < \epsilon$. This is true for every $\epsilon > 0$, and so $\mathcal{E}^{o,m}$ has the MEP property w.r.t. the family $\mathrm{emb}(\mathbb{R}^n, \mathbb{R}^m)$.

(ii) Recall that $f \coloneqq \Phi_0 \circ R_0$ for $\Phi_0 \in \mathrm{Diff}^1(\mathbb{R}^m, \mathbb{R}^m)$ and linear $R_0 \colon \mathbb{R}^n \to \mathbb{R}^m$, and that $R \in \mathcal{R}$ is such that $R\big|_{\overline{U}}$ is linear for open U. We present the proof in the case when $n = o$, and we make the assumption that $R\big|_K$ is linear. In this case, we have the existence of an affine map $A \colon \mathbb{R}^m \to \mathbb{R}^m$ so that $R_0 = A \circ R$ so that $\tilde{K} \coloneqq R_0(K) = A(R(K))$. Let $\epsilon > 0$ be given. By (Hirsch, 2012, Chapter 2, Theorem 2.7), the space $\mathrm{Diff}^2(\mathbb{R}^m, \mathbb{R}^m)$ is dense in the space $\mathrm{Diff}^1(\mathbb{R}^m, \mathbb{R}^m)$, and so there is some $\Phi_1 \in \mathrm{Diff}^2(\mathbb{R}^m, \mathbb{R}^m)$ such that

$$\|\Phi_1|_{\tilde{K}} - \Phi_0|_{\tilde{K}}\|_{L^\infty(\tilde{K};\mathbb{R}^m)} < \frac{\epsilon}{2}.$$

Then, let $T \in \mathcal{T}^m$ be such that $\|T - \Phi_1 \circ A\|_{L^\infty(R(K);\mathbb{R}^m)} < \frac{\epsilon}{2}$. Then we have that

$$\begin{aligned}
\|T \circ R - f\|_{L^\infty(K)} &= \|T \circ R - \Phi_0 \circ R_0\|_{L^\infty(K)} \\
&\leq \|T \circ R - \Phi_1 \circ A \circ R\|_{L^\infty(K)} + \|\Phi_1 \circ A \circ R - \Phi_0 \circ R_0\|_{L^\infty(K)} \\
&\leq \|T - \Phi_1 \circ A\|_{L^\infty(R(K))} + \|\Phi_1 \circ A \circ R - \Phi_0 \circ A \circ R\|_{L^\infty(K)} \\
&< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
\end{aligned}$$

Hence, if we let $r = T \circ R \circ f^{-1} \in \mathrm{emb}(f(K), T \circ R(K))$ then we obtain that $B_{K,K}(f, T \circ R) < \epsilon$. This holds for any ϵ, and hence we have that $\mathcal{T} \circ \mathcal{R}$ has the MEP for $\mathcal{I}(\mathbb{R}^n, \mathbb{R}^m)$.

The proof in the case that $o \geq n$ follows with minor modification, and applying Lemma C.1 point 5.

□

C.4.2. The Proof of Example 1

Proof. (i) From (Puthawala et al., 2020, Theorem 15) we have that $\mathcal{R}^{o,m}$ can approximate any continuous function $f \in \mathrm{emb}(\mathbb{R}^n, \mathbb{R}^m)$. Further, clearly (T1) and (T2) both contain the identity map, thus Lemma 3.9 (i) applies.

(ii) Let $\mathcal{T}^m$ be the family autoregressive flows with sigmoidal activations defined in (Huang et al., 2018). By (Teshima et al., 2020, App. G, Theorem 1 and Proposition 7), $\mathcal{T}^m$ are sup-universal approximators in the space $\mathrm{Diff}^2(\mathbb{R}^m, \mathbb{R}^m)$ of C^2-smooth diffeomorphisms $\Phi : \mathbb{R}^m \to \mathbb{R}^m$. When $\mathcal{R}^{o,m}$ is one of (R1) or (R2) the network is always linear, hence the conditions are satisfied. If $\mathcal{R}^{o,m}$ is (R4), then $\mathcal{R}^{o,m}$ contains linear mappings, and if (R3), then we can shift the origin, so that $R(x)$ is linear on K. In all cases, Lemma 3.9 part (ii) applies.

□

C.4.3. The Proof of Theorem 3.10

The proof of Theorem 3.10. First we prove the claim under the assumptions (i).

First we prove the claim under assumption (i).

Let $W \subset \mathbb{R}^n$ be an open relatively compact set. From Lemma 3.4 we have that

$$\mathcal{E}^{n,m} \coloneqq \mathcal{E}_L^{n_{L-1},m} \circ \cdots \circ \mathcal{E}_1^{n,n_1} \tag{40}$$

has the m, n, n MEP w.r.t. $\mathcal{F} \coloneqq \mathcal{F}_L^{n_{L-1},m} \circ \cdots \circ \mathcal{F}_1^{n,n_1}$. Thus for any $\epsilon_1 > 0$, we have an $\tilde{E} \in \mathcal{E}^{n,m}$ s.t. $B_{K,W}(f, \tilde{E}) < \epsilon_1$.

From Lemma C.1 point 9, we have the existence of a $\mu' \in \mathcal{P}(W)$ so that $\mathrm{W}_2\left(f_{\#}\mu, \tilde{E}_{\#}\mu'\right) < \epsilon_1$. By convolving μ' with a suitable mollifier ϕ, we can obtain a measure $\mu'' = \mu' * \phi \in \mathcal{P}(W)$ that is absolutely continuous with respect to the Lebesgue measure so that

$$\mathrm{W}_2\left(\mu', \mu''\right) < \frac{\epsilon_1}{1 + \mathrm{Lip}(\tilde{E})},$$

see (Ambrosio et al., 2008, Lemma 7.1.10.), and so $\mathrm{W}_2\left(\tilde{E}_{\#}\mu', \tilde{E}_{\#}\mu''\right) < \epsilon_1$. Hence,

$$\mathrm{W}_2\left(f_{\#}\mu, \tilde{E}_{\#}\mu''\right) < 2\epsilon_1. \tag{41}$$

Next, from universality of $\mathcal{T}_0^n$ for any $\epsilon_2 > 0$, we have the existence of a $T_0 \in \mathcal{T}_0^n$ so that $\mathrm{W}_2\left(\mu'', T_{0\#}\mu\right) < \epsilon_2$. From Lemma C.1 points 7 and 8 we have that

$$\mathrm{W}_2\left(f_{\#}\mu, \tilde{E} \circ T_{0\#}\mu\right) \leq 2\epsilon_1 + \epsilon_2 \,\mathrm{Lip}(\tilde{E}). \tag{42}$$

For a given $\epsilon > 0$, choosing $\epsilon_1 < \frac{\epsilon}{4}$ and $\epsilon_2 < \frac{\epsilon}{2(1+\mathrm{Lip}(\tilde{E}))}$ yields that the map $E = \tilde{E} \circ T_0 \in \mathcal{E}$ is such that $\mathrm{W}_2\left(f_{\#}\mu, E_{\#}\mu\right) < \epsilon$. This yields the result.

Next we prove the claim under the assumptions (ii). By our assumptions, in the weak topology of the space $C^2(\mathbb{R}^{n_j}, \mathbb{R}^{n_j})$, the closure of the set $\mathcal{T}^{n_j} \subset C^2(\mathbb{R}^{n_j}, \mathbb{R}^{n_j})$ contains the space of $\mathrm{Diff}^2(\mathbb{R}^{n_j}, \mathbb{R}^{n_j})$. Moreover, by our assumptions $\mathcal{R}^{n_{j-1}, n_j}$ contains a linear map R. We observe that as $\mathcal{R}^{n_{j-1}, n_j}$ is a space of expansive elements, the map R is injective. and hence by Lemma 3.9, the family

$$\mathcal{E}_j^{n_{j-1}, n_j} = \mathcal{T}^{n_j} \circ \mathcal{R}^{n_{j-1}, n_j}$$

has the MEP w.r.t. $\mathcal{F} = \mathcal{I}^1(\mathbb{R}^n, \mathbb{R}^m)$. By Theorem 3.8, we have that $\mathcal{I}^1(\mathbb{R}^n, \mathbb{R}^m)$ coincides with the space $\mathrm{emb}^1(\mathbb{R}^n, \mathbb{R}^m)$. Finally, by the assumption that $\mathcal{T}_0^{n_0}$ is dense in the space of C^2-diffeomorphism $\mathrm{Diff}^2(\mathbb{R}^{n_\ell})$ implies that $\mathcal{T}_0^{n_0}$ is a L^p-universal approximator for the set of C^∞-smooth triangular maps for all $p < \infty$. Hence by Lemma 3 in Appendix A of (Teshima et al., 2020), $\mathcal{T}_0^{n_0}$ is a distributionally universal. From these the claim in the case (ii) follows in the same way as the case (i) using the family $\mathcal{F} = \mathrm{emb}^1(\mathbb{R}^n, \mathbb{R}^m)$. □

C.4.4. The Proof of Lemma 3.11

The proof of Lemma 3.11. The proof follows from taking the logical negation of the MEP for $\mathcal{F}$. If the MEP is not satisfied, then there is some $f \in \mathcal{F}$ so that $B_{K,W}(f, E)$ is never smaller than $\epsilon > 0$ for all $E \in \mathcal{E}$. Applying the definition of $B_{K,W}(f, E)$ from Eqn. 6 yields the result. □

C.4.5. The Proof of Cor. 3.12

The proof of Cor. 3.12. The proof of Eqn 12 follows from the definition of the MEP.

From Eqn. 12 for $i = 1, \ldots$ we have the existence of a $\epsilon_i := B_{K,W}(f, E_i)$, where $\lim_{i\to\infty} \epsilon_i = 0$, and a $r_i \in \mathrm{emb}(f(K), E_i(W))$ such that $\|I - r_i\|_{L^\infty(f(K))} \leq 2\epsilon_i$. Applying Lemma C.1 point 8, we have that for any $E' \in \mathcal{E}^{n,o}(X, W)$

$$B_{K,X}(f, E_i \circ E') \leq 2\epsilon_i + \mathrm{Lip}(E_i) B_{K,X}(E_i^{-1} \circ r_i \circ f, E'). \tag{43}$$

Because $\mathcal{E}^{n,o}(X, W)$ has the o, n, n MEP, for each $i = 1, \ldots$, we can find a $E_i' \in \mathcal{E}^{n,o}(X, W)$ such that $B_{K,X}(E_i^{-1} \circ r_i \circ f, E_i') \leq \frac{1}{1+\mathrm{Lip}(E_i)}\epsilon_i$, and so $B_{K,X}(f, E_i \circ E_i') \leq 3\epsilon_i$. For this choice of E_i', we have that $\lim_{i\to\infty} B_{K,X}(f, E_i \circ E_i') = 0$.

From Lemma C.1 point 9, we have that for any absolutely continuous $\mu \in \mathcal{P}(K)$, there is a absolutely continuous $\mu' \in \mathcal{P}(X)$ such that $\mathrm{W}_2\left(f_{\#}\mu, E_i \circ E_{i\#}'\mu'\right) \leq 3\epsilon$. By the universality of $\mathcal{T}^n$, continuity of $E_i \circ E_i'$, and absolute continuity of μ and μ', we have the existence of $T_i \in \mathcal{T}^n$ so that

$$\mathrm{W}_2\left(f_{\#}\mu, E_i \circ E_i' \circ T_{i\#}\mu\right) \leq 4\epsilon_i \tag{44}$$

for each $i = 1, \ldots$. This proves the claim. □

(a) The unknot (b) The trefoil knot

Figure 5: An example showing how the unknot (left) can be deformed to approximate the trefoil knot (right). The black part of both knots are identical, and the red section can be made arbitrarily skinny by bringing the black points together. This can be done while sending the measure of the red sections to zero, if the starting measure have no atoms. In this way, we can construct a sequence of diffeomorphisms $(E_i)_{i=1,\dots}$ so that $\mathrm{W}_2\left(E_{i\#}\mu, \nu\right) \to 0$ where μ is the uniform measure on S^1, and ν the uniform measure on the trefoil knot. We would like to thank Reviewer 4 for suggesting this discussion and providing the figure (in tikz code!).

C.4.6. Further Discussion on Matching Topology Exactly vs Approximately

In this section we discuss a theoretical gap between the positive approximation results of Theorem 3.10 and the negative exact mapping results of Lemma 3.11. We show two main results.

First we construct sequences of maps of the form $\mathcal{E} = \mathcal{T} \circ \mathcal{R}$ that map the uniform measure on S^1 to the uniform measure on the trefoil knot. As discussed in Section 3.3, there are no mappings of this form which map S^1 to the trefoil knot exactly, but there are approximate mappings. This shows that there is some overlap between the two results, and extendable mappings may be approximated by non-extendable mappings.

Second we prove that sequences of functions that approximate non-extendable embeddings with extendable ones necessarily have unbounded gradients. This result shows that, when restricted to approximation by sequences with bounded gradients, either Theorem 3.10 or Lemma 3.11 can apply, but never both.

Example 2. *There is a sequence of extendable embeddings $(E_i)_{i=1,\dots}$ that map the uniform measure on S^1, denoted μ, to the uniform measure on the trefoil knot, denoted ν, so that*

$$\lim_{i\to\infty} \mathrm{W}_2\left(E_{i\#}\mu, \nu\right) = 0.$$

Proof. The key idea of the construction is shown in Figure 5. In that figure the unknot is bent so that it overlaps the trefoil knot, outside of an exceptional set (shown in red in Figure 5) which can be made as small as desired. The result follows by constructing a sequence of functions which 'squeeze' this red section as small as possible.

Let μ be the uniform probability measure on $S^1 \subset \mathbb{R}^2$, and ν the uniform probability measure on the trefoil knot, $\mathcal{M}$. Let $R\colon \mathbb{R}^2 \to \mathbb{R}^3$ be a fixed linear map of the form $R(x) = (x, 0)$.

We define a sequence $(X_i)_{i=1,\dots}$ of unknots in the following way. For any choice of two points on the top of the trefoil knot as shown in black in Figure 5a, we can replace the straight-line red section with a U-shaped section as shown in Figure 5a so that the resulting knot is the unknot. We obtain X_1 by letting the black points be a distance 1 apart, X_2 by letting them be a distance $\frac{1}{2}$ apart and so on, so that for X_i the two points are a distance $\frac{1}{i}$ apart. Further, for each X_i, we define A_i and B_i where A_i is the U-shaped piece of X_i (in red), and $B_i = X_i \setminus A_i$. Observe that $B_i \subset \mathcal{M}$.

Let $(T_i')_{i=1,\dots}$ be a family of diffeomorphisms so that $E_i\colon \mathbb{R}^3 \to \mathbb{R}^3$ maps $S^1 \times \{0\}$ to X_i. Further, let $(T_i'')_{i=1,\dots}$ be such that $T_i''\colon X_i \to X_i$ so that $\chi_{B_i}\left(T_i'' \circ T_i' \circ R\right)_{\#}\mu = \chi_{B_i}\nu$ when χ_{B_i} is the characteristic function of the set B_i.

Then we define $E_i \coloneqq T_i'' \circ T_i' \circ R$ and compute

$$\begin{aligned}\mathrm{W}_2\left(E_{i\#}\mu, \nu\right) &\leq \mathrm{W}_2\left(\chi_{A_i}E_{i\#}\mu, \chi_{\mathcal{M}\setminus B_i}\nu\right) + \mathrm{W}_2\left(\chi_{B_i}E_{i\#}\mu, \chi_{B_i}\nu\right) \\ &= \mathrm{W}_2\left(\chi_{A_i}E_{i\#}\mu, \chi_{\mathcal{M}\setminus B_i}\nu\right).\end{aligned}$$

As i increases, the length of $\mathcal{M} \setminus B_i$ goes to zero, thus $\nu(\mathcal{M} \setminus B_i) = \mu(A_i)$ converges to zero. Hence taking limits yields

$$\lim_{i\to\infty} \mathrm{W}_2\left(E_{i\#}\mu, \nu\right) \leq \lim_{i\to\infty} \mathrm{W}_2\left(\chi_{A_i} E_{i\#}\mu, \chi_{\mathcal{M}\setminus B_i}\nu\right) = 0.$$

Finally, E_i is certainly an extendable embedding, as R is linear, and $T_i'' \circ T_i'$ are diffeomorphisms. □

The above proof also applies when ν or μ have finitely many atoms. The same construction works if A_i is chosen so that it contains no atoms for sufficiently large i.

Next, we show that all function sequence for which implication of Theorem 3.10 and conditions of Lemma 3.11 apply are not uniformly Lipschitz. This implies that if they are differentiable they have unbounded gradients.

Lemma C.2. *Let f be continuous and E_i be a sequence of continuous functions that are uniformly Lipschitz with constant L. Let E_i be such that for all compact K and W subsets of $\mathbb{R}^n$, there is an $\epsilon > 0$, so $\forall i$ and $r \in \mathrm{emb}(f(K), E(W))$, $\|I - r\|_{L^\infty(K)} \geq \epsilon$. If μ is the indicator function of d, then $\lim_{i\to\infty} \mathrm{W}_2\left(f_{\#}\mu, E_{i\#}\mu\right) > 0$.*

Proof. Let E_i be uniformly Lipschitz with constant L. Consider a $\frac{\epsilon}{2}$ tubular neighborhood of $f(K)$. From the fact that $\|I - r\|_{L^\infty(K)} \geq \epsilon$, we have that there is a point $x \in E(W)$ so that x lies outside of this neighborhood. From uniform Lipschitzness of E_i, for each i there is a ball B of radius $\frac{\epsilon}{4L}$ around x so that all points in $E_i \cap B$ are more than $\frac{\epsilon}{4}$ away from $f(K)$. We also have that $\mu(E_i \cap B) > c$ where c is the volume of the n dimensional ball. Thus, $\mathrm{W}_2\left(f_{\#}\mu, E_{i\#}\mu\right) > \frac{c\epsilon}{4L}$ for each i, and so $\lim_{i\to\infty} \mathrm{W}_2\left(f_{\#}\mu, E_{i\#}\mu\right) > 0$. □

C.5. Layerwise Inversion and Recovery of Weights

C.5.1. Layer-wise Projection

Here we provide the details of our closed-form layerwise projection algorithm The flow layers are injective, and are often implemented to be numerically easy to invert. Thus, the crux of the algorithm comes from inverting the injective expansive layers, R. The range of the ReLU layer is piece-wise affine, hence the inversion follows a two-step program. First, identify which affine piece (described algebraically, onto which sign pattern) to project. Second, project to this point using a standard least-squares solver.

The second step is always straight-forward to analyze, but the first is more complicated.

The key step in our algorithm is the fact that for the specific choice of weight matrix $W = \begin{bmatrix} B \\ -DB \end{bmatrix}$, given any $y \in \mathbb{R}^{2n}$, we can always solve the least-squares inversion problem exactly.

We prove this result in several parts given below.

1. For any $y \in \mathbb{R}^{2n}$, $M_y W \in \mathbb{R}^{n\times n}$ is full-rank.
2. If $[y]_i \neq [y]_{i+n}$ for each $i = 1, \ldots, n$, then the argmin in Eqn. 17 is well defined, i.e. that there is a unique minimizer. Otherwise there are 2^I minimizers, where I is the number of distinct i such that $[y]_i = [y]_{i+n}$.
3. If $\tilde{M}_y = \begin{bmatrix} \Delta_y & (I^{n\times n} - \Delta_y) \end{bmatrix}$, then
$$\min_{x\in\mathbb{R}^n} \|y - R(x)\|_2^2 = \min_{x\in\mathbb{R}^n} \|M_y\left(y - Wx\right)\|_2^2 + \left\|\tilde{M}_y y\right\|_2^2. \tag{45}$$
4. We verify Eqn. 17.

The proof of Theorem 3.15. 1. Using the definition of M_y, we have,

$$M_y \begin{bmatrix} B \\ -DB \end{bmatrix} = \left(I^{n\times n} - \Delta_y\right) B - \Delta_y DB = \left(I^{n\times n} - \Delta_y - \Delta_y D\right) B. \tag{46}$$

But, $(I^{n\times n} - \Delta_y - \Delta_y D)$ is a full-rank diagonal matrix (with entries either 1 or $[D]_{i,i}$), and B is full rank by assumption, hence $M_y \begin{bmatrix} B \\ -DB \end{bmatrix}$ is too.

2. Because B is square and full rank there exists a basis[7] $\left\{\hat{b}_i\right\}_{i=1,\ldots,n}$ of $\mathbb{R}^n$ such that

$$\left\langle \hat{b}_j, b_i \right\rangle = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}. \tag{47}$$

For an $x \in \mathbb{R}^n$, let $\alpha_i = \langle x, b_i \rangle$ for $i = 1, \ldots, n$ be the expansion of x in the $\hat{b}_i$ basis.

$$\min_{x \in \mathbb{R}^n} \|y - R(x)\|_2^2 = \min_{x \in \mathbb{R}^n} \sum_{i=1}^{2n} [y - R(x)]_i^2 \tag{48}$$

$$= \sum_{i=1}^{n} \min_{x_i \in \mathbb{R}} \left([y]_i - \max(\langle x, b_i \rangle, 0)\right)^2 + \left([y]_{i+n} - \max(\langle x, -[D]_{ii} b_i \rangle, 0)\right)^2 \tag{49}$$

We now consider minizing Eqn. 49 by minimizing the basis expansion in terms of α_i,

$$\sum_{i=1}^{n} \min_{\alpha_i \in \mathbb{R}} \left([y]_i - \max(\alpha_i, 0)\right)^2 + \left([y]_{i+n} - \max(-[D]_{ii} \alpha_i, 0)\right)^2 \tag{50}$$

Eqn. 50 is clearly minimized by minizing each term in the sum, hence we search for a minimizer of the i'th term

$$\min_{\alpha_i \in \mathbb{R}} \left([y]_i - \max(\alpha_i, 0)\right)^2 + \left([y]_{i+n} - \max(-[D]_{ii} \alpha_i, 0)\right)^2 \tag{51}$$

Noting $f(\alpha_i)$ as the quantity inside the minimum of Eqn. 51, we consider the positive, negative and zero α_i cases of Eqn. 51 separately and we get

$$\min_{\alpha_i \in \mathbb{R}^+} f(\alpha_i) = \min_{\alpha_i \in \mathbb{R}^+} \left([y]_i - \alpha_i\right)^2 + [y]_{i+n}^2 = [y]_{i+n}^2 \tag{52}$$

$$\min_{\alpha_i \in \mathbb{R}^-} f(\alpha_i) = \min_{\alpha_i \in \mathbb{R}^+} [y]_i^2 + \left([y]_{i+n} + [D]_{ii} \alpha_i\right)^2 = [y]_i^2 \tag{53}$$

$$f(0) = [y]_i^2 + [y]_{i+n}^2 . \tag{54}$$

If $[y]_{i+n} > [y]_i$, then the minimizer of Eqn. 51 is $\alpha_i = -\frac{[y]_{i+n}^2}{[D]_{ii}} < 0$. Conversely if $[y]_{i+n} < [y]_i$ then the minimizer of Eqn. 51 is $\alpha_i = [y]_i > 0$. This argument applies all $i = 1, \ldots, n$, and hence if $[y]_i \neq [y]_{i+1}$ for all $i = 1, \ldots, n$ then the minimizing x is unique.

If $[y]_i = [y]_{i+1}$ then there are exactly two minimizers of $f(\alpha_i)$, $-\frac{[y]_{i+n}^2}{[D]_{ii}}$ and $[y]_i$, for both of which $f(\alpha_i) = [y]_i^2 = [y]_{i+n}^2$.

3. If we suppose that $[y]_{i+n} - [y]_i > 0$, then $[c(y)]_i = 0$ and $[c(y)]_{i+n} > 0$, thus $[\Delta_y]_{ii} = 1$, hence if we let $x_{\min}$ be the minimizing x from part 1, then

$$\left([y]_i - \max(\langle x_{\min}, b_i \rangle, 0)\right)^2 + \left([y]_{i+n} - \max(\langle x_{\min}, -[D]_{ii} b_i \rangle, 0)\right)^2 \tag{55}$$

$$= [y]_i^2 + \left([y]_{i+n} - \max(\langle x_{\min}, -[D]_{ii} b_i \rangle, 0)\right)^2 \tag{56}$$

$$= \left[\tilde{M}_y y\right]_i^2 + \left[M_y (y - W x_{\min})\right]_i^2 \tag{57}$$

If $[y]_{i+n} - [y]_i \leq 0$ then we have

$$\left([y]_i - \max(\langle x_{\min}, b_i \rangle, 0)\right)^2 + \left([y]_{i+n} - \max(\langle x_{\min}, -[D]_{ii} b_i \rangle, 0)\right)^2 \tag{58}$$

$$= \left([y]_i - \max(\langle x_{\min}, b_i \rangle, 0)\right)^2 + [y]_{i+n}^2 \tag{59}$$

$$= \left[M_y (y - W x_{\min})\right]_i^2 + \left[\tilde{M}_y y\right]_i^2 . \tag{60}$$

Thus combining Eqn.s 48, 49, 57 and 60 for each $i = 1, \ldots, n$, we have that

$$\min_{x \in \mathbb{R}^n} \|y - R(x)\|_2^2 = \min_{x \in \mathbb{R}^n} \|M_y (y - W x)\|_2^2 + \|M_y y\|_2^2 . \tag{61}$$

[7] Namely the columns of the matrix B^{-1}

4. For the final point, combining all of the above points we have

$$\min_{x\in\mathbb{R}^n} \|y - R(x)\|_2^2 = \min_{x\in\mathbb{R}^n} \|M_y\,(y - Wx)\|_2^2\,. \tag{62}$$

Further we have from Point 1 that $M_y W$ is full rank, hence $(M_y W)^{-1} M_y y = R^\dagger(y)$ is a minimizer of Eqn. 62. If $[y]_i \neq [y]_{i+n}$ for all $i = 1, \ldots, n$ then Part 2 applies, and $R^\dagger(y)$ is the unique minimizer of $\|y - R(x)\|_2^2$. In either case, we have that $R^\dagger(y)$ is a minimizer.

□

C.5.2. BLACK-BOX RECOVERY

We now discuss assumptions that enable black-box recovery of the weights of our entire network post-training.

Assumption C.3. For each $\ell = 1, \ldots, L$, $\mathcal{R}_\ell$ is an affine ReLU layer. Each $\mathcal{T}_\ell$ and $\mathcal{T}_0$ is constructed from a finite number of affine ReLU layers.

Remark C.4. If a network $\mathcal{F}$ of the form of Eqn. 1 satisfies Assumption C.3, then given the range of the network, the range of the network can be recovered exactly.

Further, if the linear region assumption from (Rolnick & Körding, 2020) is satisfied, then the exact weights are recovered, subject to two natural isometries discussed below.

Remark C.5. The ReLU part of Assumption C.3 is for all examples in Sec. 2.1. Further it is also satisfied by both flows considered in Sec. 2.2, provided that the various g_i are given by layers of affine ReLU's.

In (Rolnick & Körding, 2020), the authors show that, although ReLU networks depend on the value of their weight matrix in non-linear ways, it is still possible to recover the exact weights of a given ReLU network in a black-box way, subject to natural isometrics. The authors show that this is possible not only in theory, but in numerical applications as well.

The works of (Rolnick & Körding, 2020; Bui Thi Mai & Lampert, 2020) imply that provided the activation functions of the expressive elements are ReLU then the entire network can be recovered in a black-box way. Further, provided that either the 'linear region assumption' from (Rolnick & Körding, 2020) or the generality assumption from (Bui Thi Mai & Lampert, 2020) is satisfied, then the entire network can be recovered uniquely modulo the natural isometries of rescaling and permutation of weight matrices.

First we describe the two natural isometries of scaling and permutation. Consider the following function

$$f(x) = W_2\phi(W_1 x) \tag{63}$$

where ϕ is coordinate-wise homogeneous degree 1 (such as ReLU) and $W_1 \in \mathbb{R}^{n_1\times n_2}$ and $W_2 \in \mathbb{R}^{n_2\times n_3}$. If we let $P \in \mathbb{R}^{n_2\times n_2}$ be any permutation matrix, and D_+ be a diagonal matrix with strictly positive elements, then we can write

$$f(x) = W_2 P' D_+^{-1}\phi(D_+ P W_1 x) \tag{64}$$

as well. Thus ReLU networks can only ever be uniquely given subject to these two isometries. When describe unique recovery in the rest of this section, we mean modulo these two isometries.

In (Rolnick & Körding, 2020), the authors describe how all parameters of a ReLU network can be recovered uniquely (called reverse engineered in (Rolnick & Körding, 2020)), subject to the so called 'linear[8] region assumption', LRA.

The input space $\mathbb{R}^n$ can be partitioned into a finite number of open $\{\mathcal{S}_i\}_{i=1}^{n_i}$, where for each k, $f(x) = \mathbf{W}_k i + \mathbf{b}_i$, i.e. the network corresponds to an affine polyhedron in the output space. The algorithms (Rolnick & Körding, 2020, Alg.s 1 & 2) are roughly described below.

First, identify at least one point within each affine polyhedra $\{\mathcal{H}_j\}_{j=1}^{n_j}$. Then identify the boundaries between polyhedra. The boundaries between sections are always one affine 'piece' of piecewise hyperplanes $\{\mathcal{H}_j\}_{j=1}^{n_j}$. These $\{\mathcal{H}_j\}_{j=1}^{n_j}$ are the central objects which indicate the (de)activation of an element of a ReLU somewhere in the network. If the $\mathcal{H}_j$ are full hyperplanes, then the ReLU that is (de)activates occurs in the first layer of the network. If $\mathcal{H}_j$ is not a full hyperplane, then

[8]The use of 'linear' in this context is somewhat non-standard, and instead means affine. In this section we use the term 'linear region assumption', but use 'affine' where (Rolnick & Körding, 2020) would use 'linear' to preserve mathematical meaning.

it necessarily has a bend where it intersects another hyperplane $\mathcal{H}_{j'}$. Further, except for a Lebesgue measure 0 set, when $\mathcal{H}_j$ intersects $\mathcal{H}_{j'}$ the latter does not have a bend. If this is the case, then $\mathcal{H}_{j'}$ corresponds to a ReLU (de)activation in an earlier layer than $\mathcal{H}_j$. In this way the activation functions of every layer can be deduced. Once this is done, the normals of the hyperplanes can be used to infer the row-vectors of the various weight matrices, letting one recover the entire network.

The above algorithm recovers all of the weights exactly provided that the LRA is satisfied. The LRA is satisfied if for every distinct $\mathcal{S}_i$ and $\mathcal{S}_{i'}$, either $\mathbf{W}_i \neq \mathbf{W}_{i'}$ or $\mathbf{b}_i \neq \mathbf{b}_{i'}$. That is, different sign patterns produce different affine sections in the output space. This is a natural assumption, as the algorithm as described above reconstruction works by first detecting the boundaries between adjacent affine polyhedra, which is only possible if the LRA holds.

Given the weights of a network there is currently no simple way to detect if the LRA is satisfied, to our knowledge. Nevertheless the authors of (Rolnick & Körding, 2020) show that if it is satisfied, then unique recovery follows. Nevertheless recovery of the range of the entire network is possible, but this recovery may not be unique.

In (Bui Thi Mai & Lampert, 2020) the authors also consider the problem of recovering weights of a ReLU neural network, however the authors therein study the question of when there exist isometries beyond the two natural ones described above. In particular the main result (Bui Thi Mai & Lampert, 2020, Theorem 1) shows the following. Let $\mathcal{E}^{n_0,n_L}$ be a ReLU network that is L layers deep and non-increasing. Suppose that $E_1, E_2 \in \mathcal{E}^{n_0,n_L}$, E_1 and E_2 are general[9] and for all $x \in \mathbb{R}^{n_0}$ $E_1(x) = E_2(x)$, then E_1 is parametrically identical to E_2 subject to the two natural isometries.

This work provides the stronger result, however does not apply to the networks that we consider out of the box. It does apply to our expressive elements (provided that they use ReLU activation functions, and are non-increasing), but not necessarily apply to the network on the whole.

[9] A set is general in the topological sense if its complement is closed and nowhere dense