UNIVERSALITY THEOREMS FOR GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite the fact that generative models are extremely successful in practice, the theory underlying this phenomenon is only starting to catch up with practice. In this work we address the question of the *universality* of generative models: is it true that neural networks can approximate any data manifold arbitrarily well? We provide a positive answer to this question and show that under mild assumptions on the activation function one can always find a feedforward neural network that maps the latent space onto a set located within the specified Hausdorff distance from the desired data manifold. We also prove similar theorems for the case of multiclass generative models and cycle generative models, trained to map samples from one manifold to another and vice versa.

1 INTRODUCTION

Generative models such as Generative Adversarial Networks (GANs) are widely used for tasks such as image synthesis, semi-supervised learning, and domain adaptation (Brock et al., 2018; Radford et al., 2015; Zhang et al., 2017; Isola et al., 2017). Such generative models are trained to perform a mapping from a latent space of a small dimension to some specified data manifold, typically represented by a dataset of natural images. Despite their success and excellent performance, the theory behind such models is not yet well understood. A recent survey of open questions about generative models (Odena, 2019) among others presents the following question: what sorts of distributions can GANs model? In particular, what does it even mean for a GAN to *model a distribution*?

To answer these questions we adopt the following geometric approach, very amenable to precise mathematical analysis. Under the assumption of the *Manifold Hypothesis* (Goodfellow et al., 2016), data comes from a certain data manifold. Then the goal of a generator network is to reproduce this data manifold as closely as possible by mapping the latent space into the ambient space of the data manifold. This intuitive understanding can be written more concretely as follows. Suppose that we are given the latent space $\mathcal{M}_z$, feedforward neural network f_θ as a generator, and some target data manifold $\mathcal{M}$. In order for the manifold $\mathcal{M}$ to be generated by f_θ we require that the image of $\mathcal{M}_z$ under f_θ is sufficiently close to $\mathcal{M}$, more specifically that the Hausdorff distance between $f_\theta(\mathcal{M}_z)$ and $\mathcal{M}$ is less than the given parameter ε. Hausdorff distance is a well-defined metric on the space of all compact subsets of Euclidean space and hence is equal to zero if and only if $f_\theta(\mathcal{M}_z) = \mathcal{M}$ — the case of precise replication of the data manifold. Thus, the question at hand can be formulated as follows: is it possible to approximate in the sense of the Hausdorff distance an arbitrary compact (connected) manifold using standard feedforward neural networks? By combining techniques from Riemannian geometry with well–known properties of neural networks we provide a positive answer to this question. We also show that the condition of being *smooth* is not necessary and the results are also valid for just topological manifolds.

We further extend the discussed geometric approach for the theoretical analysis of many practical situations, for instance, to the case of data manifolds, which consist of multiple disjoint manifolds and correspond to multiclass datasets, and cycle generative models (Zhu et al., 2017; Isola et al., 2017), which for two manifolds learn an approximately invertible mapping from one manifold to another. For the latter case we prove a somewhat surprising result that for *any* given pair of data manifolds of the same dimension, one can always train a pair of neural networks which are approximately inverses of one another, and map the first manifold *almost* onto the second one, and vice versa. In this work,

we ignore specifics of the training algorithm (for instance, what loss function is used) and merely focus on understanding the generative capabilities of neural networks.

2 Related work

A large body of papers is devoted to analyzing the universality of neural networks. Classical works on universality (Cybenko, 1989; Hornik, 1991; Haykin, 1994; Hassoun et al., 1995) prove that neural networks with one hidden layer are *universal approximators* and can approximate arbitrary continuous functions on compact sets. Similar results also stand for deep wide networks with ReLU nonlinearities (Lu et al., 2017), convolutional neural networks (Cohen & Shashua, 2016) and recurrent neural networks (Khrulkov et al., 2019).

GANs were mostly studied from point of view of convergence properties (Feizi et al., 2017; Balduzzi et al., 2018; Lucic et al., 2018). Several works focus on the relationship between geometric properties of datasets and the behavior of GANs. To analyze what characteristics of datasets lead to better convergence, synthetic datasets were studied in (Lucic et al., 2018). A case of disconnected data manifold (similar in spirit to our analysis in Section 5) was analyzed in (Khayatkhoei et al., 2018). A metric for analyzing the quality of GANs based on comparing geometric properties of the original and generated datasets was proposed in (Khrulkov & Oseledets, 2018).

3 Notation and assumptions

We will denote the d-cube $[-1, 1]^d$ by I_d. We will often use an approximation of a continuous function by a neural network, in that case, the "network version" of the function will be indicated by a subscript θ or ϕ indicating a collection of trainable parameters, e.g., f_θ or g_ϕ.

In this work, we deal with data manifolds. We assume that all these manifolds are smooth, orientable, compact and connected unless stated explicitly. We also assume that all the manifolds are embedded into a Euclidean space $\mathbb{R}^n$, and inherit the Riemannian metric tensor g. By *smooth* we will mean infinitely differentiable manifolds (functions), i.e, of class C^∞; all the results, however, will stay true if we consider class C^r for some finite r. As a norm of a function f defined on some compact set D we will use the C-norm: $\|f\|_D = \max_{x \in D} |f(x)|$, and for vectors we use the 2-norm.

We will often make use of a natural geometric measure μ on a manifold, which can be constructed by integrating the *volume form* associated with the Riemannian metric tensor over the corresponding set.

4 Background

Let us first present some background material necessary for understanding the proofs. We will freely use the term *manifold* in the precise mathematical sense. Due to limited space, we do not provide the definition and refer the reader to thorough introductions such as (Lee, 2013; Sakai, 1996).

First important construction in the proof is the *exponential map*.

4.1 Exponential map

Let $\mathcal{M}$ be a Riemannian manifold endowed with a metric tensor g. Recall that *geodesics* are locally length minimizing curves, defined as a solution of a certain second-order differential equation. An important property of geodesics is that the length of the velocity vector is preserved along the curve, i.e., for a geodesic $\gamma(t)$ we have

$$\frac{d}{dt}\|\dot{\gamma}(t)\| = 0. \tag{1}$$

.

The *exponential map* is defined in the following manner. Let $q \in \mathcal{M}$ and $v \in T_p\mathcal{M}$, and suppose that there exists a geodesic $\gamma : [0, 1] \to \mathcal{M}$ with

$$\gamma(0) = q, \quad \dot{\gamma}(0) = v.$$

Then the point $\gamma(1) \in \mathcal{M}$ is denoted by $\exp_q(v)$ and called the exponential of the tangent vector v. The geodesic γ can then be written as $\gamma(t) = \exp_q(vt)$. While apriori the exponential map is defined only if $\|v\|$ is small enough, for certain class of manifolds it is globally defined. Namely, if a manifold is *geodesically complete*, then $\exp_q(v)$ is defined for all q and $v \in T_q\mathcal{M}$. Our proof is based on the following classical result.

Theorem 4.1 (Hopf-Rinow)**.** *Let $(\mathcal{M}, g)$ be a connected Riemannian manifold. Then the following statements are equivalent.*

- *The closed and bounded subsets of $\mathcal{M}$ are compact;*
- *$\mathcal{M}$ is a complete metric space;*
- *$\mathcal{M}$ is geodesically complete.*

Furthermore, any of the above implies that any points p and q in $\mathcal{M}$ can be connected by a minimal (length–minimizing) geodesic.

In particular, this implies that any compact connected manifold $\mathcal{M}$ is geodesically complete.

4.2 Hausdorff distance

The *Hausdorff distance* between two sets $X, Y \subset \mathbb{R}^n$ is defined as follows.

$$d_H(X, Y) = \inf \{\varepsilon \geq 0; X \subseteq [Y]_\varepsilon \text{ and } Y \subseteq [X]_\varepsilon\}, \tag{2}$$

where

$$[X]_\varepsilon := \bigcup_{x \in X} \{z \in \mathbb{R}^n; d(z, x) \leq \varepsilon\}. \tag{3}$$

It is well–known that the set of all compact subsets of $\mathbb{R}^n$ endowed with the Hausdorff distance becomes a complete metric space (Henrikson, 1999).

4.3 Universal Approximation Property of Neural Networks

In this paper we heavily rely on the ability of neural networks to approximate *functions*. In particular, we use the following classical results on neural networks (Cybenko, 1989; Hornik, 1991).

Theorem 4.2 (Universal Approximation Theorem)**.** *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a nonconstant, bounded and continuous function. Then for any continuous function $f : I_n \to \mathbb{R}$ and $\varepsilon > 0$ there exists a fully connected neural network f_θ with the activation function ϕ and one hidden layer, such that*

$$\max_{x \in I_n} |f(x) - f_\theta(x)| < \varepsilon.$$

Similarly, our analysis is also valid for deep networks with Rectified Linear Unit (ReLU) nonlinearities by means of the following result (Arora et al., 2018, Theorem 2.2).

Theorem 4.3 (Arora et al. (2018))**.** *Every piecewise linear function $\mathbb{R}^n \to \mathbb{R}$ can be represented by a ReLU DNN with at most $\lfloor \log_2(n+1) \rfloor + 1$ depth.*

Since piecewise linear functions are dense in the space of continuous function, we obtain a simple corollary.

Corollary 4.1. *For any continuous function $f : I_n \to \mathbb{R}$ and $\varepsilon > 0$ there exists a fully connected neural network f_θ with ReLU nonlinearity, such that*

$$\max_{x \in I_n} |f(x) - f_\theta(x)| < \varepsilon.$$

For conciseness, we will call nonlinearities satisfying Theorem 4.2 or Theorem 4.3 simply *universal nonlinearities*, explicitly specifying other required properties when necessary.
Similarly, our results can also be extended to the case of non-compact latent space $\mathbb{R}^d$. In Appendix B we provide relevant theorems and show how results of next sections generalize.

5 GEOMETRIC UNIVERSALITY THEOREM

In this section we prove that for an arbitrary manifold it is possible to construct a neural network, mapping the cube I_d approximately onto this manifold. Our analysis is based on the following lemma. In fact, this is a particular case of a much stronger theorem valid even for *topological* manifolds (without smooth structure), for which we provide a discussion and reference further in the text. We, however, believe that this particular case is instructive and provides an intuition on how the generative mappings may look like.

Lemma 5.1. *Let $\mathcal{M} \subset \mathbb{R}^n$ be a compact connected d-dimensional manifold. Then there exists a smooth map*

$$f : I_d \to \mathbb{R}^n,$$

such that $f(I_d) = \mathcal{M}$.

Proof. We will construct this map explicitly. Choose an arbitrary point $q \in \mathcal{M}$, and consider

$$\exp_q : T_q\mathcal{M} \to \mathcal{M}.$$

Since $\mathcal{M}$ is compact and connected, it is geodesically complete and the Hopf-Rinow theorem applies. Thus, this map is defined on $T_q\mathcal{M} \cong \mathbb{R}^d$ and surjective.

We now need to show that we can choose a compact subset of $T_q\mathcal{M}$ such that the restriction of $\exp_q$ to this subset is also surjective. To do this observe that since $\mathcal{M}$ is compact it has finite *diameter*, namely $\forall p, q : d(p, q) \leq R_0$ for some finite constant R_0. Here d is the Riemannian distance, defined as the arc length of a minimizing geodesic. From Eq. (1) it instantly follows that for the (Euclidean) ball $B_{R_0} = \{v \in T_q\mathcal{M} : \|v\| \leq R_0\}$ we have $\exp_q(B_{R_0}) = \mathcal{M}$. Indeed, since any point on $\mathcal{M}$ is within distance R_0 from q, there exists a minimal geodesic connecting these points with length bounded by R_0. But for any vector $v \in T_q\mathcal{M}$ from Eq. (1) we obtain that the length of the corresponding geodesic connecting q and $\exp_q(v)$ is exactly $\|v\|$, which proves the claim. Statement of the lemma then follows after selecting an arbitrary cube containing B_{R_0} and appropriate rescaling. □

Recall from Section 4.3 that *universal nonlinearities* include ReLU and nonconstant bounded continuos functions.

Theorem 5.1 (Geometric Universality of Generative Models)**.** *Let $\mathcal{M}$ be a compact connected d-dimensional manifold. For every universal nonlinearity σ there exists a fully connected neural network $f_\theta(z) : I_d \to \mathbb{R}^n$ with the activation function σ such that $d_H(\mathcal{M}, \mathcal{M}_\theta) < \varepsilon$. Here $\mathcal{M}_\theta = f_\theta(I_d)$.*

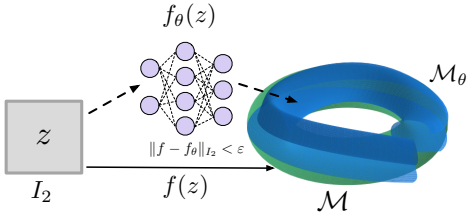

Figure 1: Visualization of the construction in the proof of Theorem 5.1. The latent space I_2 is mapped onto the manifold $\mathcal{M}$ via the function f. This mapping is then approximated via neural network f_θ, which in turn maps I_2 onto the compact set $\mathcal{M}_\theta$. If f_θ is sufficiently close to f then so are $\mathcal{M}$ and $\mathcal{M}_\theta$.

Proof. Choose an arbitrary f as in Lemma 5.1.

By universality of σ we can find such a neural network that $\|f - f_\theta\|_{I_d} < \varepsilon$. Statement of the theorem then follows from the definition of the Hausdorff distance. Indeed, by surjectivity of f we find that every point $x_0 = f(z_0) \in \mathcal{M}$ is within distance ε from the point $f_\theta(z_0) \in \mathcal{M}_\theta$, and thus $\mathcal{M} \subset [\mathcal{M}_\theta]_\varepsilon$ as in Eq. (3), and conversely $\mathcal{M}_\theta \subset [\mathcal{M}]_\varepsilon$. See Fig. 1 for illustration of the proof. □

Previously we have noted that our Lemma 5.1 is a particular case of a much stronger result (Brown, 1962). Namely, it can be stated as follows.

Lemma 5.2 (Brown's mapping theorem)**.** *Let $\mathcal{M}$ be a compact connected d-dimensional* topological *manifold (with or without boundary). Then there exists a continuous map*

$$f : I_d \to \mathbb{R}^n,$$

such that $f(I_d) = \mathcal{M}$.

Based on this lemma Theorem 5.1 can be generalized to include the more general case of topological data manifolds (as well as of manifolds with boundary).

Corollary 5.1 (Strong Geometric Universality)**.** *Theorem 5.1 holds true for $\mathcal{M}$ being an arbitrary **compact connected topological manifold with or without boundary**.*

This class of manifolds is extremely general, and it seems plausible that manifolds of natural images satisfy these conditions, which may partially explain the success of generative models. Indeed, spaces of natural images of shape $H \times W \times C$ are closed subsets of I_{HWC}, and thus are compact. We hypothesize that the property of being connected holds for manifolds representing single–class datasets. We will now address the case of multiclass manifolds.

Multiclass case The previous theorem considers only the case of a single data manifold. However, commonly in practice, single datasets contain samples from multiple data manifolds (e.g, MNIST digits, ImageNet classes). Since we can assume that these manifolds do not intersect, it is impossible to map a connected latent space surjectively onto this disconnected joint data manifold. To counteract this effect we can allow small pieces of latent space to map into thin "tunnels" connecting those manifolds. This can be made precise by the following statement.

Theorem 5.2 (Geometric Universality for Multiclass Manifolds)**.** *Let $\mathcal{M} = \sqcup_{i=1}^{c} \mathcal{M}_i$ be a "multiclass" data manifold, with each $\mathcal{M}_i$ being a compact connected d-dimensional topological manifold (with or withour boundary). Then for every $\varepsilon > 0$ and $\delta > 0$ and every universal nonlinearity σ there exists a fully connected neural network $f_\theta(z) : I_d \to \mathbb{R}^n$ with the activation function σ such that the following properties hold.*

- *There exists a collection $\{D_i\}_{i=1}^{c}$ of disjoint compact subsets of I_d such that*
$$\forall i \; d_H(f_\theta(D_i), \mathcal{M}_i) < \varepsilon. \tag{4}$$
- $\mu(I_d \setminus \sqcup_{i=1}^{c} D_i) \leq \delta$.

We refer the reader to Appendix A for the proof.

6 Invariance property of deep expanding networks

Our previous results state that it is possible to approximate any given manifold $\mathcal{M}$ up to some accuracy. However, neural networks used in the proof are shallow (they have one hidden layer) and are not practical. In this section, we study how the set $\mathcal{M}_\theta$ looks like for more practical networks consisting of a series of fully connected and convolutional layers. We will show a somewhat surprising result that under certain mild conditions such networks cannot significantly transform the latent space, more precisely the generated set $\mathcal{M}_\theta$ will be diffeomorphic to the open unit cube $(-1, 1)^d$. In fact, our results will be more general and will demonstrate that this property holds for arbitrary latent spaces, that is if z is sampled from some manifold $\mathcal{M}_z$, then $\mathcal{M}_\theta$ will be diffeomorphic to $\mathcal{M}_z$.

6.1 Reminder on embeddings

Recall the following definition.

Definition 6.1 (Smooth embedding)**.** *Let $\mathcal{M}$ and $\mathcal{N}$ be smooth manifolds and $f : \mathcal{M} \to \mathcal{N}$ be a smooth map. Then f is called an embedding is the following conditions hold.*

- *Derivative of f is everywhere injective;*
- *f is an injective, continuous and open map (i.e, maps opens sets to open sets).*

The main property of a smooth embedding is the following (Lee, 2013).

Proposition 6.1. *The domain of an embedding is diffeomorphic to its image.*

We will show that certain neural networks commonly used for generative models are in fact smooth embeddings, and thus their image is diffeomorphic to the domain (latent space). We analyze the two most commonly used layers in such models: fully connected and convolutional layers (both standard

and transposed). For the sake of simplicity we assume that convolutions are *circularly* padded, i.e., the input presents a two-dimensional torus; in this case, when the offset calls for a pixel that is off the left end of the image, the layer "wraps around" to take it from the opposite end. We consider arbitrary stride, in order to allow for a layer to increase the spatial size of a feature tensor, as commonly done.

Let us fix the nonlinearity $\sigma(z)$ to be an arbitrary smooth monotonous function without saddle points ($\sigma'(z) \neq 0$). Then the following two lemmas hold. Let us first assume that the latent space is the Euclidean space $\mathbb{R}^d$ (or equivalently, an open unit cube $(-1,1)^d$).

Lemma 6.1. *Let $f(z) = \sigma(Az + b)$ with $A \in \mathbb{R}^{n\times m}$ be a fully connected layer. If $n \geq m$ then $f(z)$ is a smooth embedding for all A except for a set of measure zero. We will call such a layer an **expanding fully connected layer**.*

Proof. Indeed, such a map is injective. It is open as a composition of a linear map (which is trivially open), and of $\sigma(z)$ which is open since it is a continuous monotonous function. Then for all matrices A of full rank (which form a set of full measure in the space of matrices of size $n \times m$) the derivative is injective by a simple application of the chain rule and the fact that $\sigma'(z) \neq 0$. □

Let us now deal with the convolutional layers.

Lemma 6.2. *Let z be a 3rd–order tensor tensor representing a feature tensor of size $m \times m$ with k channels. Suppose that $f(z) = \sigma(\mathrm{Conv}(z) + b)$ is a standard convolutional or transposed convolutional layer with an arbitrary stride. Suppose that* Conv *is parameterized via a kernel parameter $C \in \mathbb{R}^{l\times k\times s\times s}$, such that $f(z)$ is a feature tensor of size $n \times n$ with l channels. If $n^2 l \geq m^2 k$ then $f(z)$ is a smooth embedding for all C except for a set of measure zero. We will call such a layer an **expanding convolutional layer**.*

Proof. The only non-trivial part of the proof is showing injectivity of this layer for all C but measure zero. Note that if $n^2 l \geq m^2 k$ then the matrix representing the linear map performing the Conv operation is vertical, hence it is sufficient to show that generically it is of full rank. In the case of the transposed convolution, we can transpose this matrix and analyze the corresponding convolutional layer.

Stride one Let us start with the most important case of stride being one, in which case $m = n$. Denote the matrix of the linear map underlying Conv by $\widehat{C} \in \mathbb{R}^{n^2 l\times n^2 k}$, that is $\mathrm{vec}(\mathrm{Conv}(x)) = \widehat{C}\mathrm{vec}(x)$, where vec denotes the *vectorization* operator. We need to show that for all C but measure zero this matrix is of full rank.

To prove the lemma we use the following simple argument coming from algebraic geometry. The condition of matrix $\widehat{C}$ *not* being a full rank is *algebraic* (i.e., is given by polynomial equations) in the space of parameters C. Indeed, the operation of constructing $\widehat{C}$ based on C is linear with respect to C, and the condition of not being a full rank in the space of *all* matrices is specified by a set of polynomial equations (namely, determinants of all maximal square submatrices should be zero). Thus, we have shown that set $C_{singular} = \{C \in \mathbb{R}^{l\times k\times s\times s} \mid \widehat{C} \text{ is not of full rank}\}$ is algebraic; and by the well-known property of algebraic sets there are two options: either $\mu(C_{singular}) = 0$ or $C_{singular} = \mathbb{R}^{l\times k\times s\times s}$ (with μ being the standard Lebesgue measure). To show that the latter does not hold, we provide a concrete example of a weight C not in $C_{singular}$. Namely, consider the following C.

$$C[i,j,p,q] = \begin{cases} \delta_{ij}, & p = q = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Here δ_{ij} denotes the Kronecker delta symbol:

$$\delta_{ij} = \begin{cases} 1, i = j, \\ 0, i \neq j. \end{cases}$$

We observe that the corresponding matrix $\widehat{C}$ is of particularly simple structure:

$$\widehat{C}[i,j] = \delta_{ij},$$

which trivially is of full rank.

Arbitrary stride The same argument as before applies. Notice that selection of a bigger stride corresponds to selecting specific rows from the matrix $\widehat{C}$ obtained for stride one. By using the same weight tensor C as in the case of stride one, we find that the obtained matrix $\widehat{C}$ contains $\min(m^2k, n^2l)$ distinct rows of the identity matrix, followed by possible zero rows and thus also has full rank.

□

After these preliminary results, we are ready to extend them to the case of arbitrary latent space. Namely, suppose that z is sampled from an arbitrary manifold $\mathcal{M}_z \subset \mathbb{R}^d$. We use the following simple lemma and refer the reader to Appendix A for the proof.

Lemma 6.3. *Let $f : \mathcal{M} \to \mathcal{N}$ be an arbitrary smooth embedding. Let $\mathcal{S} \subset \mathcal{M}$ be a smooth embedded submanifold. Then $f|_{\mathcal{S}}$ is also a smooth embedding.*

By combining Lemmas 6.1 to 6.3 and Proposition 6.1 we obtain the following result.

Theorem 6.1. *Let $f_\theta(z)$ be an arbitrary neural network consisting of expanding fully connected layers and expanding convolutions, and $z \in \mathcal{M}_z \subset \mathbb{R}^d$. Denote $\mathcal{M}_\theta = f_\theta(\mathcal{M}_z)$. Then for all parameters θ but measure zero the following properties hold:*

- *$\mathcal{M}_\theta$ is a smooth embedded manifold;*
- *$\mathcal{M}_\theta \simeq \mathcal{M}_z$.*

Proof. Theorem follows from Lemmas 6.1 to 6.3 and Proposition 6.1 and the fact that a composition of embeddings is also an embedding. □

For many datasets used in practice, it seems very unlikely that the data comes from manifolds with very simple topological properties, as even basic visual patterns may possess quite non-trivial topological structure (Ghrist, 2008). Thus on the first sight, it seems that Theorem 6.1 suggests that using only expanding architectures, it is impossible to approximate an arbitrary data manifold with latent space being $\mathbb{R}^d$ (or an open unit cube). Such models are, however, extremely successful in practice. While we do not provide a precise theorem for this case, based on the discussion in Section 7, we hypothesize that it may possible to approximate an arbitrary compact data manifold using expanding networks up to a subset of *arbitrary small measure*, and thus limitations imposed by Theorem 6.1 are negligible in practice.

7 Cycle generative models

Another popular class of models used for instance for the unsupervised image to image translation (Zhu et al., 2017; Isola et al., 2017) learn a mapping along with its inverse from one data manifold to another. We specify this task as follows. Given two data manifolds $\mathcal{M}$ and $\mathcal{N}$ of the same dimension, the goal is two train two neural networks $f_\theta(x)$ and $g_\phi(y)$ such that $f_\theta(x)$ is a diffeomorphism of $\mathcal{M}$ and $\mathcal{N}$ with g being inverse of f.

First of all, let us notice that we do not expect for such f and g to exist for two general manifolds since two manifolds of different topological properties cannot be diffeomorphic. However, based on Theorem 6.1 we expect that the desired properties may hold *approximately*. Let us start with lemmas ensuring existence of functions f and g which map $\mathcal{M}$ *approximately* to $\mathcal{N}$ and $\mathcal{N}$ *approximately* to $\mathcal{M}$ correspondingly. In this section, we again consider only the case of smooth data manifolds.

First of all, we recall the following result (Sakai, 1996), proved in a very similar manner to Lemma 5.1.

Lemma 7.1. *Every compact connected d-dimensional manifold $\mathcal{M}$ contains an open dense set diffeomorphic to $\mathbb{R}^d$. Moreover, complement of this set has measure zero in $\mathcal{M}$.*

We use this result to obtain the following lemma (see Appendix A for the proof).

Lemma 7.2. *Let $\mathcal{M}$ and $\mathcal{N}$ be two manifolds of the same dimension. For every $\delta > 0$ there exist compact subsets $\mathcal{M}_\delta \subset \mathcal{M}$ and $\mathcal{N}_\delta \subset \mathcal{N}$ such that $\mu(\mathcal{M} \setminus \mathcal{M}_\delta) < \delta$ and $\mu(\mathcal{N} \setminus \mathcal{N}_\delta) < \delta$ and $\mathcal{M}_\delta$ is diffeomorphic to $\mathcal{N}_\delta$.*

We are now ready to provide our main result on cycle generative models. As before, recall from Section 4.3 that universal nonlinearities include nonconstant bounded continuous functions as well as ReLU.

Theorem 7.1 (Geometric Universality for Cycle Models)**.** *Fix any two compact connected manifolds $\mathcal{M}$ and $\mathcal{N}$ of the same dimension and a universal nonlinearity σ. Then for every $\delta > 0$ and $\varepsilon > 0$ there exist compact subsets $\mathcal{M}_\delta \subset \mathcal{M}$ and $\mathcal{N}_\delta \subset \mathcal{N}$ and a pair of feedforward neural networks $f_\theta(x)$, $g_\phi(y)$ with the activation function $\sigma(x)$ satisfying the following conditions:*

- $\mu(\mathcal{M} \setminus \mathcal{M}_\delta) < \delta$ *and* $\mu(\mathcal{N} \setminus \mathcal{N}_\delta) < \delta$*;*
- $d_H(f_\theta(\mathcal{M}_\delta), \mathcal{N}_\delta) < \varepsilon$ *and* $d_H(g_\phi(\mathcal{N}_\delta), \mathcal{M}_\delta) < \varepsilon$*;*
- $\|g_\phi \circ f_\theta - id\|_{\mathcal{M}_\delta} < C\varepsilon$ *and* $\|f_\theta \circ g_\phi - id\|_{\mathcal{N}_\delta} < C\varepsilon$ *with constant C depending only on manifolds $\mathcal{M}$ and $\mathcal{N}$.*

Proof. Let us start by selecting subsets $\mathcal{M}_\delta$ and $\mathcal{N}_\delta$ and a diffeomorphism $f : \mathcal{M}_\delta \to \mathcal{N}_\delta$ along with its inverse g as specified by Lemma 7.2. For simplicity let us also assume that $\mathcal{M} \subset I_n$ and $\mathcal{N} \subset I_n$. By means of the Whitney extension theorem (Whitney, 1934) we can smoothly extend f and g to the entire cube I_n, and by universality of σ we construct two feedforward neural networks $f_\theta(x)$ and $g_\phi(y)$ such that

$$\|f_\theta - f\|_{I_n} < \varepsilon, \tag{6}$$

and

$$\|g_\phi - g\|_{I_n} < \varepsilon, \tag{7}$$

with all the functions defined on the unit cube I_n. This proves first two points in the theorem. To show the last property we find that $\forall x \in \mathcal{M}_\delta$ the following estimate holds.

$$\begin{aligned} \|g_\phi \circ f_\theta(x) - x\| &= \|g_\phi \circ f_\theta(x) - g \circ f_\theta(x) + g \circ f_\theta(x) - x\| \\ &\leq \|g_\phi \circ f_\theta(x) - g \circ f_\theta(x)\| + \|g \circ f_\theta(x) - x\| \\ &\leq \varepsilon + \|g \circ f_\theta(x) - g \circ f(x)\| \\ &\leq \varepsilon + \max_{\mathcal{M}_\delta} \|Dg\| \|f_\theta(x) - f(x)\| \\ &\leq (1 + \max_{\mathcal{M}_\delta} \|Dg\|)\varepsilon, \end{aligned} \tag{8}$$

where he have used the fact that $g \circ f(x) = x$ for $x \in \mathcal{M}_\delta$ and properties (6) and (7). The second part of the claim is proved similarly. □

Neural networks f_θ and g_ϕ constructed in the proof perform *translation* from data sampled from $\mathcal{M}_\delta$ to data coming from approximately $\mathcal{N}_\delta$, and existence of such networks for arbitrary manifolds may partially explain huge empirical success of cyclic models. Even though the theorem is valid for an arbitrary pair of manifolds, we hypothesize that for datasets containing visually similar images such a map may be much easier to model, than for two arbitrary manifolds without such a connection.

Latent space $\mathbb{R}^d$ Results we provided in previous sections are valid for the compact latent space I_d. However, we can generalize them to include the other popular case of latent space being the entire space $\mathbb{R}^d$ (even though for a smaller set of activation functions). See Appendix B for the discussion of this case.

8 Conclusion and Future Work

In this work we have attempted to partially explain huge empirical success of generative models. Our results show only existence of neural networks approximating arbitrary manifolds, and do not specify how one can estimate the size of a network required for any given manifold. We hypothesize, however, that there might exist a connection between certain geometrical properties of a manifold (curvature, various topological properties), and the width/depth of a neural network required. One interesting direction of research left for a future work is analyzing this relation for datasets popular in computer vision, such as MNIST or CelebA, or toy datasets sampled from simple small dimensional manifolds (tori, circles), where one can easily vary the topological properties.

REFERENCES

Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. In *International Conference on Learning Representations*, 2018.

David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The mechanics of n-player differentiable games. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 354–363. PMLR, 2018.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Morton Brown. A mapping theorem for untriangulated manifolds. *Topology of*, 3:92–94, 1962.

Nadav Cohen and Amnon Shashua. Convolutional rectifier networks as generalized tensor decompositions. In *International Conference on Machine Learning*, pp. 955–963, 2016.

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

Soheil Feizi, Farzan Farnia, Tony Ginart, and David Tse. Understanding gans: the lqg setting. *arXiv preprint arXiv:1710.10793*, 2017.

Robert Ghrist. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Mohamad H Hassoun et al. *Fundamentals of artificial neural networks*. MIT press, 1995.

Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

Jeff Henrikson. Completeness and total boundedness of the hausdorff metric. *MIT Undergraduate Journal of Mathematics*, 1:69–80, 1999.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2): 251–257, 1991.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Yoshifusa Ito. Approximation of continuous functions on rd by linear combinations of shifted rotations of a sigmoid function with and without scaling. *Neural Networks*, 5(1):105–115, 1992.

Mahyar Khayatkhoei, Maneesh K Singh, and Ahmed Elgammal. Disconnected manifold learning for generative adversarial networks. In *Advances in Neural Information Processing Systems*, pp. 7343–7353, 2018.

Valentin Khrulkov and Ivan Oseledets. Geometry score: A method for comparing generative adversarial networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2621–2629. PMLR, 2018.

Valentin Khrulkov, Oleksii Hrinchuk, and Ivan Oseledets. Generalized tensor models for recurrent neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1gNni0qtm.

John M Lee. Smooth manifolds. Springer, 2013.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In *Advances in Neural Information Processing Systems*, pp. 6231–6239, 2017.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. In *Advances in neural information processing systems*, pp. 700–709, 2018.

Augustus Odena. Open questions about generative adversarial networks. *Distill*, 2019. doi: 10.23915/distill.00018. https://distill.pub/2019/gan-open-problems.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Takashi Sakai. *Riemannian geometry*, volume 149. American Mathematical Soc., 1996.

Hassler Whitney. Analytic extensions of differentiable functions defined in closed sets. *Transactions of the American Mathematical Society*, 36(1):63–89, 1934.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5907–5915, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

A PROOFS

Theorem 5.2 (Geometric Universality for Multiclass Manifolds). *Let $\mathcal{M} = \sqcup_{i=1}^{c} \mathcal{M}_i$ be a "multiclass" data manifold, with each $\mathcal{M}_i$ being a compact connected d-dimensional topological manifold (with or withour boundary). Then for every $\varepsilon > 0$ and $\delta > 0$ and every universal nonlinearity σ there exists a fully connected neural network $f_\theta(z) : I_d \to \mathbb{R}^n$ with the activation function σ such that the following properties hold.*

- *There exists a collection $\{D_i\}_{i=1}^{c}$ of disjoint compact subsets of I_d such that*
$$\forall i \; d_H(f_\theta(D_i), \mathcal{M}_i) < \varepsilon. \tag{4}$$
- $\mu(I_d \setminus \sqcup_{i=1}^{c} D_i) \leq \delta$.

Proof. Similar to the proof of Theorem 5.1 we will apply the universal approximation theorem to a certain function constructed with the help of Lemma 5.2. To construct such function let us select sets D_i in the following way. We divide the interval $[-1, 1]$ uniformly into c intervals, namely $[x_0, x_1], [x_1, x_2], \ldots, [x_{c-1}, x_c]$ with length of each interval being $\frac{1}{c}$ and $x_0 = -1, x_c = 1$. We propose to use the following D_i, satisfying conditions of the corollary. Denote $h = \frac{\delta}{2(c-1)}$,

$$D_i = \begin{cases} [x_i, x_{i+1} - h] \times [-1, 1]^{d-1}, \; i = 0, \\ [x_i + h, x_{i+1} - h] \times [-1, 1]^{d-1}, 0 < i < c - 1, \\ [x_i + h, x_{i+1}] \times [-1, 1]^{d-1}, i = c - 1. \end{cases} \tag{9}$$

Intuition is very simple: we chop down the cube D on the first axis into smaller boxes, and remove some space between them. On each of the chunks D_i we can now apply Lemma 5.2 for the corresponding manifold $\mathcal{M}_i$, obtaining a collection of maps $\{f_i\}_{i=1}^{c}$. To construct a global continuous map f we can now simply linearly interpolate each of the maps f_i from the right boundary $[x_{i+1} - h] \times [-1, 1]^{d-1}$ of one box to the left boundary $[x_{i+1} + h] \times [-1, 1]^{d-1}$ of the neighboring one. By applying the universal approximation theorem to this function f, we finalize the proof. □

Lemma 6.3. *Let $f : \mathcal{M} \to \mathcal{N}$ be an arbitrary smooth embedding. Let $\mathcal{S} \subset \mathcal{M}$ be a smooth embedded submanifold. Then $f|_{\mathcal{S}}$ is also a smooth embedding.*

Proof. The proof follows from the definition. Indeed, for every point $x \in \mathcal{S} \subset \mathcal{M}$ we have $T_x\mathcal{S} \subset T_x\mathcal{M}$ and restriction of the derivative of f onto this subspace is also injective. Note that $f|_{\mathcal{S}}$ is also injective and open map. □

Lemma 7.2. *Let $\mathcal{M}$ and $\mathcal{N}$ be two manifolds of the same dimension. For every $\delta > 0$ there exist compact subsets $\mathcal{M}_\delta \subset \mathcal{M}$ and $\mathcal{N}_\delta \subset \mathcal{N}$ such that $\mu(\mathcal{M} \setminus \mathcal{M}_\delta) < \delta$ and $\mu(\mathcal{N} \setminus \mathcal{N}_\delta) < \delta$ and $\mathcal{M}_\delta$ is diffeomorphic to $\mathcal{N}_\delta$.*

Proof. For each of the manifolds $\mathcal{M}$ and $\mathcal{N}$ select the open dense set of full measure as in Lemma 7.1. Each of these subsets is diffeomorphic to an open unit ball in $\mathbb{R}^d$ via maps $h_{\mathcal{M}}$ and $h_{\mathcal{N}}$. In order to construct $\mathcal{M}_\delta$ and $\mathcal{N}_\delta$ it sufficient to take preimages under $h_{\mathcal{M}}$ and $h_{\mathcal{N}}$ correspondingly of a sufficiently large closed ball B_r (as with $r \to 1$ we have $\mu(h_{\mathcal{M}}^{-1}(B_r)) \to \mu(\mathcal{M})$ and $\mu(h_{\mathcal{N}}^{-1}(B_r)) \to \mu(\mathcal{N})$). □

B LATENT SPACE $\mathbb{R}^d$

In this section we discuss how the results in the main text can be generatlized to the case of the latent variable z sampled from $\mathbb{R}^d$ rather than I_d. The only principal difference is that we now have to deal with approximating functions defined on the noncompact space, which is less trivial. We make use of the following result (Ito (1992)), where we for simplicity provide concrete formulas for the activation functions for which the results hold.

Theorem B.1. *Let σ belong to one of the three families:*

- *Gaussian distribution family:* $\sigma(t) = (2\pi)^{-1/2} \int_{-\infty}^{t} e^{-u^2/2} du$

- *Sigmoid function:* $\sigma(t) = \frac{1}{1+e^{-t}}$
- *Inverse tangent:* $\sigma(t) = \arctan(t)$

Then for any function $f \in \mathcal{C}(\bar{\mathbb{R}}^d)$ *and every* $\varepsilon > 0$ *there exists a neural network* f_θ *with activation* σ *such that*

$$\|f - f_\theta\|_{\mathbb{R}^d} < \varepsilon.$$

Here $\bar{\mathbb{R}}^d$ denotes the one point compactification of $\mathbb{R}^d$, so this class, for instance, includes functions with compact support. Note that the same result holds for ReLU networks due to Theorem 4.3. Let us denote the class of activations considered in Theorem B.1 along with ReLU by *weak universal functions*. By composing the function constructed in Lemma 5.2 with any surjection from $\mathbb{R}^d$ to I_d with compact support, we obtain the following result.

Corollary B.1. *Let* $\mathcal{M}$ *be a compact connected topological manifold with or without boundary. For every weak universal nonlinearity* σ *there exists a fully connected neural network* $f_\theta(z) : \mathbb{R}^d \to \mathbb{R}^n$ *with activation* σ *such that* $d_H(\mathcal{M}, \mathcal{M}_\theta) < \varepsilon$. *Here* $\mathcal{M}_\theta = f_\theta(\mathbb{R}^d)$.