Abstract

Modifications to a neural network's input and output layers are often required to accommodate the specificities of most practical learning tasks. However, the impact of such changes on architecture's approximation capabilities is largely not understood. We present general conditions describing feature and readout maps that preserve an architecture's ability to approximate any continuous functions uniformly on compacts. As an application, we show that if an architecture is capable of universal approximation, then modifying its final layer to produce binary values creates a new architecture capable of deterministically approximating any classifier. In particular, we obtain guarantees for deep CNNs and deep feed-forward networks. Our results also have consequences within the scope of geometric deep learning. Specifically, when the input and output spaces are Cartan-Hadamard manifolds, we obtain geometrically meaningful feature and readout maps satisfying our criteria. Consequently, commonly used non-Euclidean regression models between spaces of symmetric positive definite matrices are extended to universal DNNs. The same result allows us to show that the hyperbolic feed-forward networks, used for hierarchical learning, are universal. Our result is also used to show that the common practice of randomizing all but the last two layers of a DNN produces a universal family of functions with probability one. We also provide conditions on a DNN's first (resp. last) few layer's connections and activation function which guarantee that these layer's can have a width equal to the input (resp. output) space's dimension while not negatively effecting the architecture's approximation capabilities.

1 Introduction

Modifications made to a neural network's input and output maps to extract features from a data-set or to better suit a learning task is prevalent throughout learning theory. Typically, such changes are made by pre-(resp. post-)composing an architecture with a fixed and untrainable feature (resp. readout) map. Examples prevail classification by neural networks, random feature maps obtained by randomizing all but the last few layers of a feed-forward network, and numerous illustrations throughout geometric deep-learning theory, which we detail below. This motivates the central question of this paper: *"Which modifications to the input and output layers of a neural network architecture preserve its universal approximation capabilities?"*

Specifically, in this paper we obtain a simple sufficient condition on a pair of a feature map $\phi : \mathscr{X} \to \mathbb{R}^m$ and a readout map $\rho : \mathbb{R}^n \to \mathscr{Y}$, where $\mathscr{X}$ and $\mathscr{Y}$ are topological spaces, guaranteeing that if $\mathscr{F}$ is dense in $C(\mathbb{R}^m, \mathbb{R}^n)$ for the uniform convergence on compacts (ucc) topology then

$$\{f \in C(\mathscr{X}, \mathscr{Y}) : \rho \circ f \circ \phi, f \in \mathscr{F}\}, \tag{1}$$

is dense in $C(\mathscr{X}, \mathscr{Y})$ in the uniform convergence on compacts topology when $\mathscr{Y}$ is metric and, more generally, in the compact-open topology when $\mathscr{Y}$ is non-metrizable (such as in the hard classification problem). Simplified conditions are obtained when $\mathscr{Y}$ is a metrizable manifold, and characterization of ρ and ϕ is obtained when both $\mathscr{X}$ and $\mathscr{Y}$ are smooth manifolds.

The set $\mathscr{F}$ represents any expressive neural network architecture. For example, by [36] $\mathscr{F}$ can be taken to be the set of feed-forward networks with one hidden layer and continuous, locally-bounded, and non-polynomial activation function. Or, by [57], $\mathscr{F}$ can be taken to be the set of deep convolution networks with specific sparsity structures and ReLu activation function. Throughout $\mathscr{F}$ is often referred to as an architecture. The results are not limited to neural networks and remain valid when, for example, $\mathscr{F}$ is taken to be the set of posterior means generated by a Gaussian processes universal kernel, as in [42]. The central results are motivated by the following consequences.

Implication: Method for Constructing Non-Euclidean Universal Approximators

A natural hub for our results is in *geometric deep learning*, an emerging field of machine learning, which acknowledges and makes use of the latent non-Euclidean structures present in many types of data. Applications of geometric deep learning are prevalent throughout neuroimaging [17], computer-vision [49], covariance learning [41], and learning from hierarchical structures such as complex social networks [35], undirected graphs [44], and trees [51].

For instance, in [45], it is shown that low-dimensional representations for complex hierarchical structures into hyperbolic space outperform the current state-of-the-art high-dimensional Euclidean embedding methods due to the tree-like geometry of the former. Using the theory of gyro-vector space, introduced in [56], [18] proposed a hyperbolic-space variant of the feed-forward architecture and demonstrated its superior performance in learning hierarchical structure from these hyperbolic representations. A direct application of our main result confirms that this non-Euclidean architecture can indeed approximate any continuous function between hyperbolic spaces.

More generally, we obtain an explicit construction of feed-forward networks between any Cartan-Hadamard manifold and a guarantee that our construction is universal. Cartan-Hadamard manifolds appear throughout applied mathematics from the symmetric positive-definite matrix-valued regression problems of [17, 41], which we extend to universal approximators, to applications in mathematical finance in [24], Gaussian processes in [40], information geometric in [4], and to the geometry of the Wasserstein space [37] commonly used in Generative Adversarial Networks as in [3].

Implication: Universal Approximation Implies Universal Classification

Perhaps the most commonly used readout maps are those used when modifying neural-networks to perform classification tasks. The currently available theoretical results, found in [16], guarantee that for a random vector in $\mathbb{R}^m$ with random labels, the set of feed-forward networks with one hidden layer, step activation function $\sigma(x) = I_{[0,\infty)} - I_{(-\infty,0]}$, and readout map $\rho(x)_i = I_{[\frac{1}{2},\infty)}$ can approximate the Bayes' classifier in probability.

As an application of this paper's main results, we obtain deterministic guarantees of generic hard (n-ary) and soft (fuzzy) classification on $\mathbb{R}^m$ for any given universal approximator in $C(\mathbb{R}^m, \mathbb{R}^n)$ once it's outputs are modified by a continuous surjection ρ to take values in $\{0,1\}^n$ or $(0,1)^n$, respectively. For example, our result applies to feed-forward networks with at-least one hidden layer holds when ρ the component-wise logistic $\rho(x)_i = I_{[\frac{1}{2},1]} \circ \frac{e^{x_i}}{1+e^{x_i}}$ readout map.

Implication: DNNs with Randomly Generated First Layers are Universal

We show that the commonly employed practice of only training the final layers of a deep feed-forward network and randomizing the rest preserves its universal approximation capabilities with probability 1. Though widely used, this practice has only recently begun to be studied in [20, 38]. The link with our results arises from an observation made in [38], stating that the first portion of such a random architecture can be seen as a randomly generated feature map.

Implication: DNNs Can be Narrowed

In [30, 46] the authors provide lower bounds on a DNN layer's width, under which it is no longer a universal approximator. However, there is a wealth of literature which shows that arranging a network's neurons to create depth rather than width yields empirically superior performance. As a final application of our theory, we provide explicit conditions on a DNN's connections and activation functions so additional initial and final few layers may be added to a DNN which do not respect the minimum width requirements of [30, 46] but do not negatively impact the DNN's approximation capabilities. Numerical implementations show that additional depth build using our main results improve predictive performance additional deep layers failing our assumptions reduced the network's predictive performance.

This paper is organized as follows. Section 2 discusses the necessary topological and geometric background needed to formulate the paper's central results. Section 3 contains the paper's main results discussed above. The conclusion follows in section 4. The proofs of the main results are contained within this paper's supplementary material.

2 Background

2.1 General Topology

Before moving on to the main results of the paper, we will require some additional topological terminology. The *interior* of a subset $A \subseteq \mathscr{X}$ of a topological space is the largest open subset contained in A. For example, in the Euclidean space $\mathbb{R}$, the interior of $[0,1)$ is $(0,1)$. The *closure* of A is the smallest closed-set containing A. Therefore, the closure of $[0,1)$ in $\mathbb{R}$ is $[0,1]$. The difference between the closure of A and its interior is called the *boundary* of A and is denoted by ∂A. For example, the boundary of $[0,1)$ in the Euclidean space $\mathbb{R}$ is $\{0,1\}$. A subset $A \subseteq \mathbb{R}^m$ is called a *retract* of $\mathbb{R}^m$ if there is a continuous map $r : \mathbb{R}^m \to A$ such that $r \circ \iota_A = 1_A$, where $\iota_A : A \to \mathbb{R}^m$ is the

inclusion map and 1_A is the identity map on A. Dually, a continuous right-inverse of a continuous surjective map is called a *section*. A *covering projection* $f : \mathbb{R}^n \to \mathscr{Y}$ is a surjective continuous function such that for every $x \in \mathscr{Y}$ there is an open set U_x, containing x, such that $\rho^{-1}[U] = \bigcup_{\alpha \in A} U_x^\alpha$ and $\{U_x^\alpha\}_{\alpha \in A}$ is a disjoint collection of open sets for which $\rho|_{U_x^\alpha} : U_x^\alpha \to U_x$ is a homeomorphism. For example, the map $x \to e^{-i\pi x}$ is a covering projection of $\mathbb{R}$ onto the circle.

Lastly, since continuous maps transfer topological information then a continuous bijection with continuous inverse preserves all topological information. Such a map is called a *homeomorphism* and two topological spaces related by a homeomorphism are said to be *homeomorphic*. For example, the sigmoid (or logistic) function $x \mapsto \frac{e^x}{1+e^x}$ continuously puts $\mathbb{R}$ in bijection with $(0,1)$ and its inverse is the logit function $y \mapsto \ln\left(\frac{y}{1-y}\right)$, which is continuous. Therefore, $\mathbb{R}$ and $(0,1)$ are homeomorphic. A continuous injective map $\phi : \mathscr{X} \to \mathbb{R}^m$ such that $\mathscr{X}$ is homeomorphic to $\phi(\mathscr{X})$ is called an *embedding*.

A topological space $\mathscr{X}$ is said to be *locally-compact* if every $x \in \mathscr{X}$ is contained in an open subset U_x of $\mathscr{X}$ which is in turn contained in an compact subset K_U of $\mathscr{X}$. For example, every point $x \in \mathbb{R}$ is contained in $(x-1, x+1)$ which is in turn contained in the closed-bounded (and therefore compact-set by the Heine-Borel theorem) $[x-1, x+1]$. A topological space is said to be *simply connected*, if every two paths between points can be continuously deformed into one another.

Similarly to [8, 11], we say that a subset A of a topological space $\mathscr{X}$ is *collared* if there exists an open subset $U \subseteq \mathscr{X}$ containing A and a homeomorphism $\phi : U \to A \times [0,1)$ mapping A to $A \times [0,1)$. In this way the open set U is, in a sense, topologically similar to A itself since one can imagine shrinking any point $(a,t) \in A \times [0,1)$ down to $(a,0)$ and then identifying it back with a via ψ.

We denote the set of positive integers by $\mathbb{N}^+$.

2.2 Topology of Function Spaces

Let $C(\mathbb{R}^m, \mathbb{R}^n)$ denote the set of all continuous functions from the Euclidean space $\mathbb{R}^m$ to the Euclidean space $\mathbb{R}^n$. Closeness in $C(\mathbb{R}^m, \mathbb{R}^n)$ can be described in a number of ways but in the context of universal approximation theorems of [13, 28, 36] two functions f and g are thought of as being close if they are uniformly close on compacts, if for a given $\varepsilon > 0$ the following holds:

$$\sum_{k=1}^{\infty} \frac{\sup_{\|x\| \leq k} \sqrt{\sum_{i=1}^{n} \|f(x)_i - g(x)_i\|^2}}{2^k (1 + \sup_{\|x\| \leq k} \sqrt{\sum_{i=1}^{n} \|f(x)_i - g(x)_i\|^2})} < \varepsilon. \tag{2}$$

The topology described by (2) is called the *topology of uniform convergence on compacts*, henceforth ucc topology. If $\mathbb{R}^n$ is replaced by any other topological space $\mathscr{Y}$ whose notion of closeness is defined by a distance function and $\mathbb{R}^m$ is replaced by nearly any other topological space then closeness in the collection of continuous functions from $\mathscr{X}$ to $\mathscr{Y}$, denoted by $C(\mathscr{X}, \mathscr{Y})$, can be described analogously to (2) by replacing the Euclidean distance on $\mathbb{R}^n$ by another distance function d-on $\mathscr{Y}$, the compact subsets $K \subseteq \mathbb{R}^m$ with compact subsets of $\mathscr{X}$, and taking $f, g \in C(\mathscr{X}, \mathscr{Y})$. The topology on $C(\mathscr{X}, \mathscr{Y})$ defined in this way is still called the ucc topology.

If a distance function cannot describe the topology on $\mathscr{Y}$, for example, we will see that this is the case for reasonable topologies on $C(\mathscr{X}, \{0,1\}^n)$, then one cannot define the ucc topology. Instead, consider the smallest topology on $C(\mathscr{X}, \mathscr{Y})$ containing the sets

$$\{V_{K,O} : \emptyset \neq K \subseteq \mathscr{X} \text{ compact and } \emptyset \neq O \subseteq \mathscr{Y} \text{ open}\}, V_{K,O} \triangleq \{f \in C(\mathscr{X}, \mathscr{Y}) : f(K) \subseteq O\}. \tag{3}$$

When the topology on $\mathscr{Y}$ is defined by a distance function and $\mathscr{X}$ is a locally-compact Hausdorff space, then the smallest topology containing (3) coincides with the ucc topology. However, unlike the ucc topology, the smallest topology containing (3) is well-defined on $C(\mathscr{X}, \mathscr{Y})$ for any topological spaces $\mathscr{X}$ and $\mathscr{Y}$. This generalized ucc topology is called the *compact-open topology* (co-topology).

2.3 Manifolds

A (topological) *manifold* is a topological space which "closeup" resembles Euclidean space, whereas a manifold with boundary locally looks like a part of Euclidean space but possibly with a hard edge.

Definition 2.1 (Metrizable Manifold with Boundary; [8]). *A topological space $\mathscr{Y}$ is said to be a metrizable manifold with boundary if*
(i) For every $y \in \mathscr{Y}$, there is an open $U_y \subseteq \mathscr{Y}$ containing y which is homeomorphic to

$$\left\{ (z_1, \ldots, z_n) \in \mathbb{R}^n : \sqrt{\sum_{i=1}^{n} z_i^2} < 1 \text{ and } z_n \geq 0 \right\}, \tag{4}$$

3

(ii) There exists a distance function (metric) $d : \mathscr{Y}^2 \to \mathscr{Y}$ such that the topology on $\mathscr{Y}$ coincides with the smallest topology on $\mathscr{Y}$ containing the open balls $\{B_\varepsilon(y)\}_{\varepsilon>0, y \in \mathscr{Y}}$; where

$$B_\varepsilon(y) \triangleq \{z \in \mathscr{Y} : d(z,y) < \varepsilon\}.$$

We say that d is a metric for $\mathscr{Y}$. The subset of $\mathscr{Y}$ consisting of all points y contained in some open set U_y which is homeomorphic to the interior of (4) *is denoted by* $\mathrm{Int}(\mathscr{Y})$.

A *smooth manifold* without boundary, is a manifold for which there is a well-defined differential calculus admitting arbitrarily many derivatives and which can locally be deformed into Euclidean space via infinitely differentiable maps with infinitely differentiable inverses.

An *m*-dimensional *Riemannian manifold* $\mathscr{M}$ is a manifold without boundary which can be locally smoothly deformed into Euclidean space such that curvature and length can be meaningfully compared, locally, between $\mathbb{R}^m$ and $\mathscr{M}$. Amongst other things, this allows the definition of minimal-length curves connecting points on $\mathscr{M}$, called *geodesics*. If any two points on $\mathscr{M}$ can be connected by such a minimal length curve then $\mathscr{M}$ is said to be *complete*. Moreover, when $\mathscr{M}$ is complete and connected, the function mapping any two points $p, q \in \mathscr{M}$ to the length of a geodesic connecting them defines a metric $d_{\mathscr{M}}$. Thus, $\mathscr{M}$ has a geometrically meaningful metric structure where distance represents the length of maximally efficient trajectories and $C(X, \mathscr{M})$. The existence of $d_{\mathscr{M}}$ also implies that $C(\mathscr{X}, \mathscr{M})$ is equipped with the ucc-topology.

Further, when $\mathscr{M}$ is complete and connected the Hopf-Rinow Theorem, of [27], affirms that for any given $p \in \mathscr{M}$, the map sending any $v \in \mathbb{R}^m$ lying tangent to p to the point on $\mathscr{M}$ arrived at time $t = 1$ by traveling along a the geodesic with initial velocity v defines a surjection from $\mathbb{R}^m$ onto $\mathscr{M}$. This map is called the *Riemannian Exponential map* on $\mathscr{M}$ at p and is denoted by $\mathrm{Exp}_p^{\mathscr{M}}$. In [29], it is shown that, in this case, $\mathrm{Exp}_p^{\mathscr{M}}$ has a smooth inverse outside a low-dimensional subset C_p. This inverse is denoted by $\mathrm{Log}_p^{\mathscr{M}}$ and is known to locally preserve length between $\mathbb{R}^m$ and $\mathscr{M}$ along geodesics emanating from p. This means that $\mathrm{Log}_p^{\mathscr{M}}$ and $\mathrm{Exp}_p^{\mathscr{M}}$ are geometrically meaningful feature and readout maps, respectively.

However, the set ∂C_p can be pathological or difficult to deal with. This issue is overcome by turning to the sub-class of *Cartan-Hadamard manifolds*. A Riemannian manifold $\mathscr{M}$ is Cartan-Hadamard if it is simply connected and has *non-positive curvature*. Non-positive curvature mean that all triangles drawn on $\mathscr{M}$ by geodesics have internal angles adding-up at-most $180°$.

3 Main Results

Let $\phi : \mathscr{X} \to \mathbb{R}^m$ and $\rho : \mathbb{R}^n \to \mathscr{Y}$. Subsets of $\mathbb{R}^n$ (resp $\mathbb{R}^m$) will be equipped with the (relative) Euclidean topology unless explicitly stated otherwise. Equip $C(\mathscr{X}, \mathscr{Y})$ with the co-topology, $C(\mathbb{R}^m, \mathbb{R}^n)$ with the ucc topology, let $\mathscr{F}$ be a *dense* subset of $C(\mathbb{R}^m, \mathbb{R}^n)$ such as the architectures studied in [36, 39, 57] or the posterior means of a Gaussian process with universal kernel as in [42], and define the subset $\mathscr{F}_{\rho,\phi} \subseteq C(\mathscr{X}, \mathscr{Y})$ by

$$\mathscr{F}_{\rho,\phi} \triangleq \{g \in C(\mathscr{X}, \mathscr{Y}) : g = \rho \circ f \circ \phi \text{ where } f \in \mathscr{F}\}. \tag{5}$$

The set $\mathscr{F}_{\rho,\phi}$ is dense in $C(\mathscr{X}, \mathscr{Y})$ under the following assumptions on ϕ and ρ.

Assumption 3.1 (Feature Map Regularity). *The map ϕ is a continuous and injective.*

Assumption 3.2 (Readout Map Regularity). *Suppose that the readout map ρ satisfies the following:*
 (i) Either of the following hold:
   (a) ρ is a continuous and it has a section on $\mathrm{Im}(\rho)$,
   (b) ρ is a covering projection of $\mathbb{R}^m$ onto $\mathrm{Im}(\rho)$ and $\mathscr{X}$ is connected and simply connected,
 (ii) $\mathrm{Im}(\rho)$ is dense in $\mathscr{Y}$,
 (iii) $\partial \mathrm{Im}(\rho)$ is collared.

Theorem 3.3 (General Version). *Suppose that $\mathscr{F}$ is dense in $C(\mathbb{R}^m, \mathbb{R}^n)$. If Assumptions 3.1 and 3.2 hold then $\mathscr{F}_{\rho,\phi}$ is dense in $C(\mathscr{X}, \mathscr{Y})$.*

Just as in the filtering literature of [10], one would hope that the outputs of any learning model should depend continuously on its inputs. Therefore, we only consider feature maps ϕ which are continuous functions. In this case, Assumption 3.1 is *sharp*. We denote the identity map $x \mapsto x$ on $\mathbb{R}^n$ by $1_{\mathbb{R}^n}$.

Theorem 3.4 (Assumption 3.1 is Sharp). *Let $\mathscr{X}$ be a metrizable manifold with boundary, let ϕ be continuous, and $\mathscr{F} \subseteq C(\mathbb{R}^m, \mathbb{R}^n)$. Then $\mathscr{F}_{1_{\mathbb{R}^n}, \phi}$ is dense in $C(\mathscr{X}, \mathbb{R}^n)$ if and only if ϕ is injective.*

Remark 3.5 (Sharpening Assumption 3.2). *Assumption 3.2 is almost sharp and a characterization can be obtained using the $\mathscr{X}$-sets studied in [55, 21]. However, it is unlikely that a non-pathological example can be generated which falls outside the scope of Assumption 3.2.*

Theorem 3.4 shows that it is easy to verify if a feature map preserves the universal approximation property. However, it can be much more challenging to verify if and when the readout map ρ does so.

The following presents a readily applicable case of Theorem 3.3. They highlight the convenient fact that if ρ is surjective then only Assumptions 3.1 and 3.2 (i) need to be verified.

Corollary 3.6. *If ϕ is a continuous injective map, ρ is a surjective covering projection, and $\mathscr{F}$ is dense in $C(\mathbb{R}^d, \mathbb{R}^D)$ then $\mathscr{F}_{\rho,\phi}$ is dense in $C(\mathscr{X}, \mathscr{Y})$. In particular, ϕ and ρ may be homeomorphisms.*

When both ϕ and ρ fully preserve topological structure then Corollary 3.6 sharpens.

Proposition 3.7 (Homeomorphic case is Sharp). *Let ϕ and ρ be homeomorphisms. Then $\mathscr{F}$ is dense in $C(\mathbb{R}^m, \mathbb{R}^n)$ if and only if $\mathscr{F}_{\rho,\phi}$ is dense in $C(\mathscr{X}, \mathscr{Y})$.*

Corollary 3.8. *If ϕ is a continuous injective map, ρ is a continuous surjection with a section, $\mathscr{X}$ is connected and simply connected, and $\mathscr{F}$ is dense in $C(\mathbb{R}^d, \mathbb{R}^D)$ then $\mathscr{F}_{\rho,\phi}$ is dense in $C(\mathscr{X}, \mathscr{Y})$.*

When additional structure is assumed of $\mathscr{Y}$, as is common in most applications, Assumption 3.2 (ii) and (iii) can be omitted and the other assumptions can be simplified. Specifically, the case where $\mathscr{Y}$ is a manifold with boundary is considered. In the case where $\mathscr{X}$ and $\operatorname{Int}(\mathscr{Y})$ are smooth, then Theorem 3.3 can be further streamlined as follows.

Assumption 3.9 (Readout Map Regularity: Geometric Version). *Suppose that ρ satisfies:*
  (i) *ρ satisfies Assumption 3.2 (i) and $\operatorname{Im}(\rho) \subseteq \operatorname{Int}(\mathscr{Y})$,*
  (ii) *$\operatorname{Int}(\mathscr{Y}) - \operatorname{Im}(\rho)$ is a (possibly empty) smooth submanifold of $\operatorname{Int}(\mathscr{Y})$ of dimension strictly less-than $\dim(\operatorname{Int}(\mathscr{Y})) - n$.*

Theorem 3.10 (Geometric Version). *Let $\mathscr{Y}$ be a metrizable manifold with boundary, for which $\operatorname{Int}(\mathscr{Y})$ is a smooth manifold, $\mathscr{X}$ is locally-compact, and $\mathscr{F}$ is dense in $C(\mathbb{R}^m, \mathbb{R}^n)$. If ϕ satisfies Assumption 3.1 and ρ satisfies Assumption 3.9 then $\mathscr{F}_{\rho,\phi}$ is dense in $C(\mathscr{X}, \mathscr{Y})$.*

Consequences of these results in various areas of machine-learning are now considered.

3.1 Dense Families in $C(\mathbb{R}^m, \mathbb{R}^n)$ Induce Universal Classifiers

Let $\mathscr{X}$ be a set, $\phi : \mathscr{X} \to \mathbb{R}^m$ be a bijection, and $\{L_i\}_{i=1}^n$ be a collection of labels describing elements of $\mathscr{X}$. Let $\mathscr{X}_i \triangleq \{x \in \mathscr{X} : x \text{ has label } L_i\}$. For example, $\{\mathscr{X}_i\}_{i=1}^n$ are disjoint and cover $\mathscr{X}$ then we obtain the *n*-ary classification problem, but in general, any $x \in \mathscr{X}$ may simultaneously have distinct multiple labels. Without loss of generality, we may assume that $\mathscr{X}$ is a topological space which is homeomorphic to $\mathbb{R}^m$ since we may equip it with the topology $\{\phi^{-1}[U] : U \text{ open in } \mathbb{R}^m\}$. Assume that the sets $\{\mathscr{X}_i\}_{i=1}^n$ are open subsets of $\mathscr{X}$.

In the stochastic case, the Bayes classifier is the golden standard for classification. In the deterministic case, the standard is clearly the *ideal classifier* $\hat{h} : \mathscr{X} \to \{0, 1\}^n$, introduced here, and defined by

$$\hat{h}(x)_i \triangleq I_{\mathscr{X}_i}(x), \tag{6}$$

where $I_{\mathscr{X}_i}$ is the indicator function of $\mathscr{X}_i$, taking value 1 if $x \in \mathscr{X}_i$ and 0 otherwise.

Since the usual Euclidean topology on $\{0, 1\}^n$ coincides with the discrete topology on $\{0, 1\}^n$ and since a continuous functions to a discrete topological space are constant, see [50], then $\hat{h}$ only belongs to $C(\mathscr{X}, \{0, 1\}^n)$ if it is trivial, i.e.: either $\mathscr{X}_i = \mathscr{X}$ or $\mathscr{X}_i = \emptyset$ for each i. Moreover, a direct computation shows that there are exactly 2^n functions in $C(\mathscr{X}, \{0, 1\}^n)$. Thus, other topologies must be considered on $\{0, 1\}^n$ in order to have a meaningful deterministic classification theory.

When $n = 1$, there are two other choices of topologies on $\{0, 1\}$, up to homeomorphism. These are the trivial topology $\{\emptyset, \{0, 1\}\}$ and the *Sierpiński topology* $\{\emptyset, \{1\}, \{0, 1\}\}$. The trivial topology is uninteresting since a direct computation shows that with it every function in $C(\mathscr{X}, \{0, 1\})$ becomes indistinguishable, i.e.: the co-topology on $C(\mathscr{X}, \{0, 1\})$ becomes trivial and therefore density in $C(\mathscr{X}, \{0, 1\})$ holds trivially for any non-empty subset. In the case of the Sierpiński topology in [54, Chapter 7] it is shown that all indicator functions of any open set $\mathscr{X}$ from any sufficiently regular topological space, such as $\mathscr{X}$, is a continuous function to $\{0, 1\}$ with the Sherpiński topology. This latter property has lead to widespread use of this space in semantics.

The next result shows that $\hat{h}$ can be approximated on two fronts simultaneously. First, by showing that $\hat{h}$ has a natural decomposition as $I_{(\frac{1}{2},1]}$ applied component-wise to continuous *soft (fuzzy) classifier* $\hat{s}$, i.e. $\hat{s} \in C(\mathscr{X}, [0,1]^n)$, satisfying

$$\hat{s}_i^{-1}[(1/2,1]] = \mathscr{X}_i, \qquad (\forall i = 1,\dots,n). \tag{7}$$

Subsequently, the architecture $\mathscr{F}_{\rho,\phi}$ is shown to simultaneously approximate $\hat{s}$ uniformly on compacts in $C(\mathscr{X}, [0,1]^n)$ and $\hat{h}$ in the compact-open topology on $C(\mathscr{X}, \{0,1\}^n)$. Intuitively, (7) represents the philosophy of logistic regression where one approximates on the interval and the thresholds the logistic classifier to obtain a strict decision rule, and thus a hard classifier.

Theorem 3.11 (Universal Classification: General Case). *Let $\{0,1\}^n$ be equipped with the n-fold product of the Sierpiński topology on $\{0,1\}$, ϕ be continuous and injective, $\rho : \mathbb{R}^n \to (0,1)^n$ be a homeomorphism, $\alpha \in (0,1)$, and $\mathscr{F} \subseteq C(\mathbb{R}^m, \mathbb{R}^n)$ be dense. Let $\{\mathscr{X}_i\}_{i=1}^n$ be a set of open subsets of a metric space $\mathscr{X}$ and let $\hat{h}$ be its associated ideal classifier defined by (6). Then the following hold:*
  (i) (Hard-Soft Decomposition) There exist continuous functions $\hat{s}_i \in C(\mathscr{X}, [0,1])$ such that

$$\hat{h} = I_{(\alpha,1]} \bullet (\hat{s}_1, \dots, \hat{s}_n) \qquad \hat{s}_i^{-1}[(\alpha,1]] = \mathscr{X}_i, (\forall i = 1,\dots,n)$$

  (ii) (Universal Classification) There exists a sequence $\{f_k\}_{k \in \mathbb{N}}$ in $\mathscr{F}$ such that:
    (a) (Soft Classification) For each non-empty compact subset $\kappa \subseteq \mathscr{X}$ and every $\varepsilon > 0$, there is some $K \in \mathbb{N}^+$ such that

$$\sup_{x \in \kappa} \max_{i=1,\dots,n} |\rho \circ f_k \circ \phi(x)_i - \hat{s}_i(x_i)| < \varepsilon, \qquad (\forall k \geq K)$$

    (b) (Hard Classification) $I_{(\alpha,1]} \bullet \rho \circ f_k \circ \phi$ converges to $\hat{h}$ in $C(\mathscr{X}, \{0,1\}^n)$ for the co-topology.

Furthermore, $\mathscr{F}_{\rho,\phi}$ is dense in $C(\mathscr{X}, [0,1]^n)$.

As an application, we now show that most feed-forward DNNs and deep CNNs used in practice for classifications, are indeed universal classifiers in the sense of Theorem 3.11.

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous *activation* function, and let $\mathscr{N}\mathscr{N}^\sigma$ denote the set of feed-forward networks from $\mathbb{R}^m$ to $\mathbb{R}^n$, i.e.: continuous functions with representation

$$f(x) = W \circ f^{(J)}, \quad f^{(j)}(x) = \sigma \bullet \left(W^{(j)} \circ f^{(j-1)}(x) \right), f^{(0)}(x) = x, \quad j = 1,\dots,J \tag{8}$$

where W and W^j are affine maps and $\bullet$ denotes component-wise composition. The following results directly follow from Theorem 3.11 and the central result of [36], and validates the principle way neural networks are used for classification.

Corollary 3.12 (Universal Classification: Deep Feed-Forward Networks). *Let $\{\mathscr{X}_i\}_{i=1}^n$ be open subsets of $\mathscr{X}$, and $\hat{h}$ be their associated ideal classifier. Let $\phi : \mathscr{X} \to \mathbb{R}^n$ be a continuous injective feature map. Let σ be a continuous, locally-bounded, and non-constant activation function. Let ρ either be the component-wise logistic function. Then there exists a sequence $\{f_k\}_{k \in \mathbb{N}^+}$ of DNNs satisfying the conclusions of Theorem 3.11.*

Define the set of deep CNNs with ReLu activation and sparsity $2 \leq s \leq m$, denoted by $Conv^s$, to be the collection of all functions from $\mathbb{R}^n$ to $\mathbb{R}$ represented by

$$f(x) = W \circ f^{(J)}, \quad f^{(j)}(x) = \sigma \bullet \left(w^{(j)} \star (f^{(j-1)}(x)) - b^j \right), f^{(0)}(x) = x, \quad j = 1,\dots,J,$$

where W is an affine map from $\mathbb{R}^{d+Js}$ to $\mathbb{R}$, $b^{(j)} \in \mathbb{R}^{d+js}$, $w^{(j)} = \{w_k^{(j)}\}_{k=-\infty}^\infty$ is a *convolutional filter mask* where $w_k \in \mathbb{R}$ and $w_k \neq 0$ only if $0 \leq k \leq s$, and the *convolutional operation* of $w^{(j)}$ with the vectors $\{v_j\}_{j=1}^J$ is the sequence defined by $(w \star v)_i = \sum_{j=0}^{J-1} w_{i-j} v_j$ and $\sigma(x) = \max\{0,x\}$.

Corollary 3.13 (Universal Classification: Deep CNNs). *Let $2 \leq s \leq n$, $\{\mathscr{X}_i\}_{i=1}^n$ be open subsets of $\mathscr{X}$, and $\hat{h}$ be their associated ideal classifier. Let $\phi : \mathscr{X} \to \mathbb{R}^n$ be a continuous injective feature map and let $\rho : \mathbb{R} \to (0,1)$ be the logistic function. Then there is a sequence of deep CNNS $\{f_k\}_{k \in \mathbb{N}^+}$ in $Conv_{\rho,\phi}^s$ satisfying the conclusion of Theorem 3.11.*

3.2 Applications in Geometric Deep Learning

This subsection illustrates the applicability of the main results to geometric deep learning. Our examples focus on two illustrative points, first that many commonly used non-Euclidean regression

models can be extended to non-Euclidean architectures capable of universal approximation and second, we illustrate how our results can be used to validate the approximation capabilities of certain geometric deep learning architectures.

For Cartan-Hadamard manifolds, the Cartan-Hadamard Theorem, [31, Corollary 6.9.1], guarantees that $\partial C_p = \emptyset$ and in particular $\mathrm{Log}_p^{\mathscr{M}}$ is a globally-defined homeomorphism between $\mathscr{M}$ and $\mathbb{R}^m$. Thus, the following result follows from Corollary 3.8.

Corollary 3.14 (Cartan-Hadamard Version). *Let $\mathscr{F}$ be dense in $C(\mathbb{R}^m, \mathbb{R}^n)$, let $\mathscr{M}$ and $\mathscr{N}$ be Cartan-Hadamard manifolds of dimension m and n. Then, $\mathscr{F}_{\mathrm{Log}_p^{\mathscr{M}}, \mathrm{Exp}_q^{\mathscr{N}}}$ is dense in $C(\mathscr{M}, \mathscr{N})$.*

We consider here two consequences of this result.

3.2.1 Symmetric Positive-Definite Matrix Learning

Symmetric positive-definite matrices play a prominent role in many applied sciences, largely due to their relationship to covariance matrices, in areas ranging from computational anatomy in [48], computer vision in [47], and in finance [5]. The space P_d^+ of $d \times d$ symmetric positive-definite matrices is a non-Euclidean subspace of $\mathbb{R}^{d^2}$. In [1], P_d^+ is shown to be a Cartan-Hadamard manifold whose Riemannian exponential and logarithm maps are, respectively, given by

$$\mathrm{Exp}_A(B) = \sqrt{A} \exp\left(\sqrt{A}^{-1} B \sqrt{A}^{-1}\right) \sqrt{A}, \ \mathrm{Log}_A(B) = \sqrt{A} \log\left(\sqrt{A}^{-1} B \sqrt{A}^{-1}\right) \sqrt{A}, \quad (9)$$

where exp and log denote the matrix exponential and logarithms, respectively. Moreover, the distance function on this space is given by

$$d_+(A, B) \triangleq \left\| \sqrt{A} \log\left(\sqrt{A}^{-1} B \sqrt{A}^{-1}\right) \sqrt{A} \right\|_F,$$

where $\|\cdot\|_F$ denotes the Fröbenius norm and $\sqrt{A}$ is well-defined for any matrix in P_d^+. Using this distance, [41] developed non-Euclidean least-squares regression on P_d^+. The parameters involved in these models are typically optimized either using the non-Euclidean line search algorithms of [41] or the non-Euclidean stochastic gradient approach on P_d^+ of [6]. The aforementioned regression models can be extended to form a ucc-dense architecture in $C(P_d^+, P_D^+)$.

Corollary 3.15 (Universal Approximation for Symmetric Positive-Definite Matrices). *Let $d, D \in \mathbb{N}^+$ and $\mathscr{F} \subseteq C(\mathbb{R}^{d(d+1)/2}, \mathbb{R}^{D(D+1)/2})$ be ucc-dense. Then, for any $A \in P_d^+$ and $B \in P_D^+$, $\mathscr{F}_{\mathrm{Log}_A, \mathrm{Exp}_B}$ is ucc-dense in $C(P_d^+, P_D^+)$. In particular, if σ is a continuous, locally-bounded, and non-polynomial activation function then $\mathscr{N}\mathscr{N}^\sigma_{\mathrm{Log}_A, \mathrm{Exp}_B}$ is ucc-dense in $C(P_d^+, P_D^+)$.*

3.2.2 Hyperbolic Feed-Forward Networks

For $c > 0$, the *generalized hyperbolic spaces* $\mathbb{D}_c^n$ of [18] have underlying set $\{x \in \mathbb{R}^n : c\|x\|^2 < 1\}$ and their topology is induced by the following non-Euclidean metric

$$d_c(x, y) \triangleq \frac{2}{\sqrt{c}} \tanh^{-1}\left(\sqrt{c} \left\| \frac{(1 - c\|x\|^2)y - (1 - 2cx^\top y + c\|y\|^2)}{1 - 2cx^\top y + c^2\|x\|^2\|y\|^2} \right\|\right).$$

Though a direct description of hyperbolic feed-forward neural networks would be lengthy, on [18, page 6], it is shown any hyperbolic feed-forward network from $\mathbb{D}_c^m$ to $\mathbb{D}_c^n$ can be represented as

$$\left\{ \mathrm{Exp}_0^{\mathbb{D}_c^k} \circ f \circ \mathrm{Log}_0^{\mathbb{D}_c^k} : f \in \mathscr{N}\mathscr{N}^\sigma \right\}, \quad (10)$$

where $\mathrm{Exp}_0^{\mathbb{D}_c^k}$ is the Riemmanian Exponential map on $\mathbb{D}_c^k$ about 0, as in Corollary 3.14. Closed-form expressions are obtained in [18, Lemma 2] for these feature and readout maps. Since, as discussed in [18], $\mathbb{D}_c^k$ is a complete connected Riemannian manifold of non-positive curvature then the Cartan-Hadamard Theorem implies that $C_0 = \emptyset$. Whence, Corollary 3.14 yields the following.

Corollary 3.16 (Hyperbolic Neural Networks are Universal). *Let σ be a continuous, non-polynomial, locally-bounded activation function and $c > 0$. Then for every $g \in C(\mathbb{D}_c^m, \mathbb{D}_c^n)$, every $\varepsilon > 0$, and every compact subset $K \subseteq \mathbb{D}_c^m$ there exists a hyperbolic neural network $g_{\varepsilon, K, c}$ in (10) satisfying*

$$\sup_{x \in K} d_c(g(x), g_{\varepsilon, K, c}) < \varepsilon.$$

Next, applications of Theorems 3.3 and 3.10 with Euclidean input and output spaces are considered.

3.3 Universality of Deep Networks with First Layers Randomized

Fix $\mathbb{R}$-valued random variables $\{X_i\}_{i=1}^k$ and $\{Z_i\}_{i=1}^k$ defined on a common probability space $(\Omega, \Sigma, \mathbb{P})$. Fix an activation function $\sigma : \mathbb{R} \to [0,1]$, and positive integers $\{d_i\}_{i=1}^k$. Using this data, for each $i = 1, \ldots, k$ define random affine maps $W_i : \mathbb{R}^{d_i} \times \Omega \to \mathbb{R}^{d_{i+1}}$, defined by

$$(x, \omega) \mapsto A_i(\omega)x + b_i(\omega), \tag{11}$$

where the entries of A_i are i.i.d. copies of X_i and the entries of b_i are i.i.d. copies of Z_i.

The random affine maps (11) define the (random) set of deep feed-forward neural networks with first k layers randomized and last 2 layers trainable to be the (random) subset of $C(\mathbb{R}^d, \mathbb{R}^D)$ via

$$\mathcal{NN}_{2,k}^\sigma(\omega) \triangleq \{f \in C(\mathbb{R}^m, \mathbb{R}^n) : (\exists g \in \mathcal{NN}_2^\sigma) f(x) = g \circ [\sigma \bullet W_k(x, \omega) \circ \sigma \bullet \cdots \circ \sigma \bullet W_1(x, \omega)]\},$$

where $\mathcal{NN}_2^\sigma$ is the collection of feed-forward neural networks of the form $W_2 \circ \sigma \bullet W_1$, where $W_1 : \mathbb{R}^m \to \mathbb{R}^d$ and $W_2 : \mathbb{R}^d \to \mathbb{R}^n$ are affine maps and d is a positive integer. Under the following mild assumptions, the random set $\mathcal{NN}_{2,k}^\sigma$ is dense in $C(\mathbb{R}^m, \mathbb{R}^n)$ with probability 1.

Assumption 3.17. *For each $i = 1, \ldots, k$*
  (i) $d_i \leq d_{i+1}$ for each $i = 1, \ldots, k$,
  (ii) σ is a strictly increasing and continuous,
  (iii) $\mathbb{E}[X_i] = \mathbb{E}[Z_i] = 0$, $\mathbb{E}[X_i^2] = \mathbb{E}[Z_i^2] = 1$,
  (iv) For every $C > 1$, $\mathbb{E}[|X_i|^C], \mathbb{E}[|Z_i|^C] < \infty$.

Theorem 3.18. *If Assumption 3.17 holds, then there exists a measurable subset $\Omega' \subseteq \left\{ \omega \in \Omega : \overline{\mathcal{NN}_{2,k}^\sigma(\omega)} = C(\mathbb{R}^d, \mathbb{R}^D) \right\}$ satisfying $\mathbb{P}(\Omega') = 1$.*

Corollary 3.19 (Sub-Gaussian Case with Sigmoid Activation). *Let $X_i = Z_i$ for each $i = 1, \ldots, k$ be independent standardized sub-Gaussian random-variables, $\sigma(x) = \frac{1}{1+e^{-x}}$, and $d_i = d$ for each $i = 1, \ldots, k$. Then the conclusion of Theorem 3.18 holds.*

Corollary 3.20 (Bernoulli Case with PReLU Activation). *Suppose that for every $i, j = 1, \ldots, k$, X_i and Z_j i.i.d. copies of a random variable taking values $\{-1, 1\}$ with probabilities $\{\frac{1}{2}, \frac{1}{2}\}$. Let $d_i = d$ for each $i = 1, \ldots, k$ and σ be the PReLU activation function of [23]. Then Assumptions 3.17 are met; thus the conclusion of Theorem 3.18 holds.*

3.4 Feed-Forward Layers of Sub-Minimal Width

In this section, we use Theorem 3.10 to describe how additional layers can be incorporated into a DNN, which violate the minimum width requirements of $m+1$ in its hidden layers (see [30, 46]) but do not negatively impact the architecture's approximation capabilities. We say that such layers have *sub-minimal width*. We derive specific conditions on the activation functions and structure of the connections between sub-minimal width layer ensuring that Assumptions 3.1 and 3.2 are met.

Proposition 3.21 (Input Layers of Sub-Minimal Width: Continuous Monotone Activations and Invertible Connections). *Let σ be a continuous and strictly increasing activation function, $J \in \mathbb{N}_+$, $A_1, \ldots, A_J$ be $m \times m$ matrices, and $b_1, \ldots, b_J \in \mathbb{R}^d$. Let $\phi(x) \triangleq \phi_J(x)$ where*

$$\phi_j(x) \triangleq \sigma \bullet \left(\exp(A_j)\phi_{j-1}(x) + b_j\right) \qquad \phi_0(x) \triangleq x, \quad j = 1, \ldots, J, \tag{12}$$

where $\exp$ is the matrix exponential. Then ϕ satisfies Assumption 3.1.

Proposition 3.22 (Output Layers of Sub-Minimal Width: Invertible Feed-Forward Layers). *In the setting of Proposition 3.21, if σ is also surjective then ϕ is a homeomorphism, and in particular it satisfies Assumption 3.2.*

Example 3.23 (Sub-Minimal Width via Generalized PReLU Activation). *Fix $\alpha, \beta \in (0, \infty)$, $\alpha \neq \beta$. Then the activation function $\sigma = \beta x I_{x \geq 0} + \alpha x I_{x < 0}$ is monotone increasing and surjective.*

Example 3.24 (ReLU Feed-Forward Layers Cannot Achieve Sub-Minimal Width). *Let ϕ be as in (12) and let $\sigma(x) \triangleq \max\{0, x\}$. By Theorem 3.4, $\mathscr{F}_{1_{\mathbb{R}^n}, \phi}$ is not dense in $C(\mathbb{R}^m, \mathbb{R}^n)$.*

3.4.1 Numerical Illustration

The following numerical illustration will also make use of the following method for constructing feature maps satisfying Assumption 3.1. The method can be interpreted as passing the inputs through an arbitrary function and feeding them into the model's input via a skip connection.

8

Proposition 3.25 (Skip Connections Using Arbitrary Continuous Functions are Good Feature Maps).
Let $d \in \mathbb{N}_+$ and $g \in C(\mathbb{R}^m, \mathbb{R}^d)$. Then $\phi_g(x) \triangleq (x, g(x))$ satisfies Assumption 3.1.

Example 3.26 (Pre-Trained DNN with a Skip Connection are Good Feature Maps). *Let $g \in \mathcal{N}\mathcal{N}^\sigma$. Then $\phi_g(x) \triangleq (x, g(x))$ satisfies Assumption 3.1.*

To illustrate the effect of (in)correctly choosing the networks' input and output layers we implement different DNNs whose initial or final layers are build using the above examples or intentionally fail Assumptions 3.1 or 3.2. Our implementations are on the California housing dataset [32], with the objective of predicting the median housing value given a set of economic and geo-spacial factors as described in [19]. The test-set consists of 30% percent of the total 20k training instances. The implemented networks are of the form $\rho \circ f \circ \phi$, where $f = W_2 \circ \sigma \bullet W_1$ is a shallow feed-forward network with ReLU activation and ρ, ϕ are built using the above examples.

Our reference model (Vanilla) is simply the shallow feed-forward network f. For the first DNN, which we denote (Bad), ρ and ϕ are given by as in Example 3.24 and therefore violate Assumption 3.1. For the second DNN, denoted by (Good), ρ and ϕ are as in Example (3.23) and Assumptions 3.1 and 3.2 are met. The final DNN, denoted by (Rand), ρ is as in Example 3.23 and ϕ is as in Example 3.26 where the pre-trained network is generated randomly following in Corollary 3.20.

Model	Test				Train			
	Good	Rand	Bad	Vanilla	Good	Rand	Bad	Vanilla
MAE	0.318	0.320	0.876	0.320	0.252	0.296	0.887	0.284
MSE	0.247	0.259	1.355	0.257	0.174	0.234	1.409	0.209
MAPE	16.714	17.626	48.051	17.427	12.921	15.668	48.698	14.878

Table 1: Training and test predictive performance.

As anticipated, Table 1 shows that selecting the networks' initial and final layers according to our method either improves performance (Good) when all involved parameters are trainable or does not significantly affect it even if nearly every parameter is random (Rand). However, disregarding Assumptions 3.1 and 3.2 when adding additional deep layers dramatically degrades predictive performance, as is the case for (Bad). Table 1 shows that if a DNN's first and final layers are structured according to Theorem 3.10 then expressibility can be improved, even if these layers violate the minimum width bounds of [30, 46]. Python code for these implementations is available at [34].

4 Conclusion

Modifications to the input and output layers of any neural networks, using carefully chosen feature $\phi : \mathcal{X} \to \mathbb{R}^m$ and readout $\rho : \mathbb{R}^n \to \mathcal{Y}$ maps, are common in practice. Theorems 3.3, 3.10, and Corollary 3.14 provided general conditions on these maps guaranteeing that the new, modified, architecture can approximate any function in the uniform convergence on compacts (or more generally the compact-open) topologies between their new input and output spaces.

As a consequence of our main results, we showed that universal approximation implies universal classification once a component-wise logistic map is applied. This is a deterministic strengthening of the probabilistic results of [16]. We derived a method for constructing universal approximators between a wide class of manifolds. In particular, we extended the symmetric positive-definite matrix-valued regressor of [41] to a universal approximator and we showed that the hyperbolic feed-forward networks of [18] are universal approximators between hyperbolic spaces.

Our main results also described how to structure the first and final layers of a DNN between Euclidean spaces, so as to preserve the approximation capabilities of the network's middle layers. In particular, we provided conditions on a network's activation function and connections so that these layers can be made narrower than the specifications of [30, 46] while maintaining the architecture's expressibility. Lastly, we showed that randomly generated DNNs are good feature maps with probability 1.

Broader Impact

A large portion of available data is non-Euclidean, either in the form of social network data to imaging data relevant in health applications of deep learning. The tools in this paper open up a generic means of translating the currently available deep learning technology to those milieus. The automation of tools in the medical sciences is important to helping reducing waiting times in hospitals and help

make healthcare more accessible to all, so in that way, any automatizing of health science tools helps move society in that direction. Therefore, we hope that the methods presented in paper form a small step towards a greater positive advancement of the social and natural sciences.

Acknowledgments and Disclosure of Funding

Supplement to: Non-Euclidean Universal Approximation

Anastasis Kratsios[*]

Eugene Bilokopytov[†]

This supplement provides theoretical justification of the claims made in this paper. We repeat the assumptions needed for our results and we repeat the statement of each theorem before it's proof for a smoother read.

A Assumptions

Assumption (Feature Map Regularity). *The map ϕ is a continuous and injective.*

Assumption (Readout Map Regularity). *Suppose that the readout map ρ satisfies the following:*
  (i) Either of the following hold:
    (a) ρ is a continuous and it has a section on $\mathrm{Im}(\rho)$,
    (b) ρ is a covering projection of $\mathbb{R}^m$ onto $\mathrm{Im}(\rho)$ and $\mathscr{X}$ is connected and simply connected,
  (ii) $\mathrm{Im}(\rho)$ is dense in $\mathscr{Y}$,
  (iii) $\partial\mathrm{Im}(\rho)$ is collared; i.e.: there exists an open subset $U \subseteq \mathscr{Y}$ containing $\partial\mathrm{Im}(\rho)$ and a homeomorphism $\psi : U \to \partial\mathrm{Im}(\rho) \times [0,1)$ mapping $\partial\mathrm{Im}(\rho)$ to $\partial\mathrm{Im}(\rho) \times \{0\}$.

Assumption (Readout Map Regularity: Geometric Version). *Suppose that ρ satisfies:*
  (i) ρ satisfies Assumption 3.2 (i) and $\mathrm{Im}(\rho) \subseteq \mathrm{Int}(\mathscr{Y})$,
  (ii) $\mathrm{Int}(\mathscr{Y}) - \mathrm{Im}(\rho)$ is a (possibly empty) smooth submanifold of $\mathrm{Int}(\mathscr{Y})$ of dimension strictly less-than $\dim(\mathrm{Int}(\mathscr{Y})) - n$.

Assumption (Regularity of Randomized Deep Networks). *For each $i = 1, \dots, k$*
  (i) $d_i \leq d_{i+1}$ for each $i = 1, \dots, k$,
  (ii) σ is a strictly increasing and continuous,
  (iii) $\mathbb{E}[X_i] = \mathbb{E}[Z_i] = 0$, $\mathbb{E}[X_i^2] = \mathbb{E}[Z_i^2] = 1$,
  (iv) For every $C > 1$, $\mathbb{E}[|X_i|^C], \mathbb{E}[|Z_i|^C] < \infty$.

B Proofs

B.1 Proof of Main Results

In the next lemma, we use the term algebra in the sense of the Stone-Weirestrass theorem, see [12, V.8] for details.

Lemma B.1. *Let $\phi : X \to Z$ be a continuous injection between topological spaces. Let $\mathscr{F} \subset C(Z, \mathbb{R}^n)$ be dense in the compact-open topology. Then $\{f \circ \phi | \ f \in \mathscr{F}$ is dense in $C(X, \mathbb{R}^n)$.*

Proof. Let $\Phi : C(Z, \mathbb{R}^n) \to C(X, \mathbb{R}^n)$ be defined by $[\Phi(f)](x) = f(\phi(x))$, which by [43, Theorem 46.11] is a continuous map. Hence, $\overline{\Phi(\mathscr{F})} = \overline{\Phi(C(Z, \mathbb{R}^n))}$, because $\mathscr{F}$ is dense. Therefore, in order to show that $\Phi(\mathscr{F})$ is dense, it is enough to prove density of $\Phi(C(Z, \mathbb{R}^n))$.

---

[*]Department of Mathematics, Eidgenössische Technische Hochschule Zürich, HG G 32.3, Rämistrasse 101, 8092 ürich, Switzerland. email: *anastasis.kratsios@math.ethz.ch*

[†]Department of Mathematics and Statistical Sciences, University of Alberta, 11324 89 Ave NW, Edmonton, AB T6G 2J5, Canada. email: *bilokopy@ualberta.ca*

Observe that $C(X, \mathbb{R}^n) = C(X) \oplus \ldots \oplus C(X)$, $C(Z, \mathbb{R}^n) = C(Z) \oplus \ldots \oplus C(Z)$ and $\Phi : C(Z) \to C(X)$ on every component independently. Hence, we just need to show that $\Phi(C(Z))$ is dense in $C(X)$.

To that end, observe that since Φ is an algebra homomorphism, $\Phi(C(Z))$ is a subalgebra of $C(X)$. Clearly, $1 \in \Phi(C(Z))$, and since ϕ is an injection, it is easy to see that $\Phi(C(Z))$ separates points of X. Thus, $\Phi(C(Z))$ is dense by the Stone-Weierstrass theorem. $\qquad\square$

Lemma B.2. *Let $\mathscr{X}$ be a topological space, $\rho : \mathbb{R}^n \to \mathscr{Y}$ be a map satisfying Assumption 3.2, and $\mathscr{F}$ be dense in $C(X, \mathbb{R}^n)$ for the compact-open topology. Then the set*

$$\{g \in C(\mathscr{X}, \mathrm{Im}(\rho)) : (\exists f \in \mathscr{F}) \, g = \rho \circ f\}, \tag{13}$$

is dense in $C(\mathscr{X}, \mathrm{Im}(\rho))$. In particular, if $\mathscr{Y}$ is a metric space then the set (13) *is dense in $C(\mathscr{X}, \mathrm{Im}(\rho))$ for the ucc topology.*

Proof. First, suppose that Assumption 3.2 (i.a), (ii), and (iii) hold and that ρ is continuous with continuous section. Define the map $F : C(\phi(\mathscr{X}), \mathbb{R}^n) \to C(\mathscr{X}, \mathrm{Int}(\rho))$ by $f \to \rho \circ f$. Since ρ is continuous then by [43, Theorem 46.11] F is also continuous. Moreover, by Assumption 3.2 (i.a), since there is a section R to ρ, i.e. a continuous map $R : \mathrm{Int}(\rho) \to \mathbb{R}^n$ such that $\rho \circ R = 1_{\mathscr{Y}}$ then by [43, Theorem 46.11] the map $G : C(\mathscr{X}, \mathrm{Int}(\rho)) \to C(\mathscr{X}, \mathbb{R}^n)$ defined by $g \to R \circ g$ is well-defined and continuous. Furthermore, for every $g \in C(\mathscr{X}, \mathrm{Int}(\rho))$ it follows that

$$\begin{aligned}
F \circ G(g) &= F(R \circ g) \\
&= \rho \circ (R \circ g) \\
&= (\rho \circ R) \circ g \\
&= g \\
&= 1_{\mathrm{Int}(\rho)}(g),
\end{aligned}$$

therefore F is a continuous surjection. Since continuous surjections map dense sets to dense sets in their codomain, then since $\mathscr{F}$ is dense in $C(\mathscr{X}, \mathbb{R}^n)$ and the image of $\mathscr{F}$ under F is the set of (13) then the set described in (13) is dense in $C(\mathscr{X}, \mathrm{Int}(\rho))$.

Now, suppose that Assumption 3.2 (i.b), (ii), and (iii) hold. Since $\mathscr{X}$ is simply connected and ρ is a covering map, then the conditions for [52, Chapter 2, Section 2, Theorem 5] are met (since simply connectedness means a trivial fundamental group), therefore, for every $f \in C(\mathscr{X}, \mathrm{Im}(\rho))$ there exists some $\hat{f} \in C(\mathscr{X}, \mathbb{R}^n)$ such that

$$\rho \circ \hat{f} = f. \tag{14}$$

Since $\mathscr{F}$ is dense in $C(\mathscr{X}, \mathbb{R}^n)$ then exists a sequence $\{f_k\}_{k \in \mathbb{N}}$ in $\mathscr{F}$ converging to $\hat{f}$ in the compact-open topology. Since ρ is continuous the continuity of F, following the same notation as the first part of the proof, implies that $\{F(f_k)\}_{k \in \mathbb{N}}$ converges to $F(\hat{f})$ in the compact-open topology. By (14) this implies that $\{F(f_k)\}_{k \in \mathbb{N}} = \{\rho \circ f_k\}_{k \in \mathbb{N}}$ converges to

$$f = F(f) = f$$

in the compact-open topology. Therefore, (13) is dense in $C(\mathscr{X}, \mathrm{Im}(\rho))$.

In particular, if $\mathscr{Y}$ is a metric space, then $\mathrm{Im}(\rho)$ is a metric space. Thus, by [43, Theorem 46.8] the compact-open and ucc topologies coincide on $C(\mathscr{X}, \mathrm{Im}(\rho))$. This gives the final claim. $\qquad\square$

It is convenient to sumamrize the above lemmas into a single lemma.

Lemma B.3. *If Assumptions 3.1 and 3.2 both hold, and if $\mathscr{F}$ is dense in $C(\mathbb{R}^m, \mathbb{R}^n)$ then $\mathscr{F}_{\rho, \phi}$ is dense in $C(\mathscr{X}, \mathrm{Im}(\rho))$ for the compact-open topology. Moreover, if $\mathscr{Y}$ is a metric space then $\mathscr{F}_{\rho, \phi}$ is dense in $C(\mathscr{X}, \mathrm{Im}(\rho))$ for the ucc-topology.*

Proof. The result follows directly from Lemmas B.2 and B.1. $\qquad\square$

Lemma B.4. *If $\mathscr{X}$ is locally-compact and Assumption 3.2 holds then $C(\mathscr{X}, \mathrm{Im}(\rho))$ is dense in $C(\mathscr{X}, \mathscr{Y})$ in the compact-open topology.*

Proof of Lemma B.4. By Assumption 3.2 (iii) there is a homeomorphism $\psi : U \to \partial \mathrm{Im}(\rho) \times [0,1)$ and a (continuous) inclusion map $i : U \to \mathscr{Y}$. Therefore, the [14, Lemma 5.29.1] (existence and universal property of colimits of topological spaces and the fact that any push-out is a colimit) there exists a unique topological space (up to homeomorphism) $\mathscr{Y}'$ containing $\partial \mathrm{Im}(\rho) \times [0,1)$ as a subspace for which there exists a continuous $\Psi : \mathscr{Y} \to \mathscr{Y}'$ satisfying

$$\Psi \circ i = j \circ \psi, \tag{15}$$

where j is the (continuous) inclusion map of $\partial \mathrm{Im}(\rho) \times [0,1)$ into $\mathscr{Y}'$, such that for any other topological space $\mathscr{Y}''$ satisfying (15) there exists a continuous map $I : \mathscr{Y}' \to \mathscr{Y}''$. In particular, this means that since ψ is a homeomorphism then Ψ must also be a homeomorphism.

Therefore, $C(\mathscr{X}, \Psi(\mathrm{Im}(\rho)))$ is dense in $C(\mathscr{X}, \mathscr{Y}')$ if and only if $C(\mathscr{X}, \mathrm{Im}(\rho))$ is dense in $C(\mathscr{X}, \mathscr{Y})$. Thus, without loss of generality we establish the result in the former setting where we have identified U with $\partial \mathrm{Im}(\rho) \times [0,1)$. When required by the context, we use the notation $x = (u^x, t^x) \in \partial \mathrm{Im}(\rho) \times [0,1) = \Psi[U]$.

Accordingly, for any fixed $f \in C(\mathscr{X}, \mathscr{Y}')$ we define a sequence $\{f_n\}_{n \in \mathbb{N}}$ in $C(\mathscr{X}, \Psi(\mathrm{Im}(\rho)))$ converging to f. For every $s \in (0,1)$ define the continuous functions $\tilde{\psi}_s : \partial \mathrm{Im}(\rho) \times [0,1) \to \partial \mathrm{Im}(\rho) \times (0,1)$ by

$$\tilde{\psi}_s(u,t) \triangleq \begin{cases} (u,t) & : t > s \\ (u,s) & : t \le s. \end{cases}$$

Using the family $\{\tilde{\psi}_s\}_{s \in (0,1)}$ we define the sequence of functions $\{\psi_n\}_{n \in \mathbb{N}}$ from $\mathscr{Y}$ to $\mathrm{Im}(\rho)$ by

$$\psi_n(x) \triangleq \begin{cases} x & : x \notin U \\ (u^x, \frac{1}{n}) & : x \in U \text{ and } t^x \le \frac{1}{n}. \end{cases}$$

Finally, for every $n \in \mathbb{N}$, we set $f_n \triangleq \psi_n \circ f$. Note that, for each $n \in \mathscr{N}$, since $\psi(\partial \mathrm{Im}(\rho)) = \mathrm{Im}(\rho) \times \{0\}$ then $f_n \in C(\mathscr{X}, \Psi(\mathrm{Im}(\rho)))$. Note also that, by construction $f_n(x) = f(x)$ for every $x \notin \tilde{X}$, where $\tilde{X} \triangleq f^{-1}[\partial \mathrm{Im}(\rho) \times [0,1)]$, and both f_n and f map $\tilde{X}$ to $\partial \mathrm{Im}(\rho) \times [0,1)$. Thus, it is enough to show that f_n converges to f in $C(\tilde{X}, \partial \mathrm{Im}(\rho) \times [0,1))$.

Since ϕ is a continuous injection into the locally-compact space $\mathbb{R}^m$ then $\mathscr{X}$, and therefore $\tilde{X}$, are locally-compact. Therefore, $C(\tilde{X}, \partial \mathrm{Im}(\rho) \times [0,1))$ is homeomorphic to $C(\tilde{X}, \partial \mathrm{Im}(\rho)) \times C(\tilde{X}, [0,1))$. By [43, Section 19.2, Exercise 6], f_n converges to f on the product space $C(\tilde{X}, \partial \mathrm{Im}(\rho)) \times C(\tilde{X}, [0,1))$ if and only if $p_i(f_n)$ converges to $p_i(f)$ for $i = 1,2$ where p_1 is the canonical projection of $C(\tilde{X}, \partial \mathrm{Im}(\rho)) \times C(\tilde{X}, [0,1))$ onto $C(\tilde{X}, \partial \mathrm{Im}(\rho))$ and p_2 is the canonical projection of $C(\tilde{X}, \partial \mathrm{Im}(\rho)) \times C(\tilde{X}, [0,1))$ onto $C(\tilde{X}, [0,1))$. First observed that,

$$p_1(f_n) = f = p_1(f),$$

for each $n \in \mathbb{N}$. Next,

$$p_2(f_n) = \frac{1}{n} I_{[0,\frac{1}{n}]} + t I_{[\frac{1}{n}, \infty)} \text{ and } p_2(f) = t.$$

Since $[0,1)$ is topologized with the (relative) Euclidean topology which is metric then [43, Theorem 46.8] implies that compact-open topology on $C(\tilde{X}, [0,1))$ agrees with the topology of uniform convergence on compacts in $C(\tilde{X}, [0,1))$. Fix $\varepsilon > 0$ and $n \ge \frac{1}{\varepsilon}$. Thus, for every compact $\tilde{K} \subseteq \tilde{X}$

$$\sup_{x \in \tilde{K}} \| p_2(f_n)(x) - p_2(f)(x) \| = \max_{t \in [0,\frac{1}{n}]} \left| \frac{1}{n} - t \right| \le \frac{1}{n} < \varepsilon.$$

Therefore, $p_2(f_n)$ converges to $p_2(f)$ in the compact-open topology on $C(\tilde{X}, [0,1))$. Hence, f_n converges to f in the compact-open topology on $C(\tilde{X}, \partial \mathrm{Im}(\rho)) \times C(\tilde{X}, [0,1))$ and therefore on $C(\tilde{X}, \partial \mathrm{Im}(\rho) \times [0,1))$. Since

$$\left\{ U_{\tilde{K}, \tilde{O}} : \emptyset \neq \tilde{K} \subseteq \tilde{\mathscr{X}} \text{ compact and } \emptyset \neq \tilde{O} \subseteq \partial \mathrm{Im}(\rho) \times [0,1) \text{ open} \right\}$$
$$U_{\tilde{K}, \tilde{O}} \triangleq \left\{ f \in C(\tilde{\mathscr{X}}, \partial \mathrm{Im}(\rho) \times [0,1)) : f(\tilde{K}) \subseteq \tilde{O} \right\} \tag{16}$$

is a sub-base for the compact-open topology on $C(\tilde{\mathscr{X}}, \partial \mathrm{Im}(\rho) \times [0,1))$ then by definition of convergence, for every $\tilde{K}_1, \dots, \tilde{K}_n \subseteq \tilde{\mathscr{X}}$ compact and $\tilde{O}_1, \dots, \tilde{O}_n \subseteq \partial \mathrm{Im}(\rho) \times [0,1)$ open if $f \in \bigcap_{i=1}^n U_{\tilde{K}_i, \tilde{O}_i}$ then there exists some $N \in \mathbb{N}$ for which $f_N \in \bigcap_{i=1}^n U_{\tilde{K}_i, \tilde{O}_i}$.

Since every compact subset $K \subseteq \mathscr{X}$ is relatively compact in $\tilde{X}$ and every open subset $O \subseteq \mathscr{Y}$ is relatively compact in $\partial \mathrm{Im}(\rho)) \times C(\tilde{X}, [0,1))$ then if $f \in \bigcap_{i=1}^n V_{K_i, O_i}$ where V_{K_i, O_i} are as in (3) then

$$f(K_i) \subseteq O_i \qquad i = 1, \ldots, n.$$

Therefore, there exists some $N \in \mathbb{N}$ satisfying

$$f_N(K_i \cap \tilde{\mathscr{X}}) \subseteq O_i \cap \partial \mathrm{Im}(\rho) \times [0,1) \qquad i = 1, \ldots, n.$$

However, by construction, $f_N = f$ on $\mathscr{X} - \tilde{\mathscr{X}}$ and therefore on each $K_i - K_i \cap \mathscr{X}$. Therefore,

$$f_N(K_i) \subseteq O_i \qquad i = 1, \ldots, n.$$

Thus, $f_N \in \bigcap_{i=1}^n V_{K_i, O_i}$ and f_n converge to f in $C(\mathscr{X}, \mathscr{Y})$ for the compact-open topology. $\qquad \square$

We may now prove the following result and its consequences.

Theorem (General Version). *Suppose that $\mathscr{F}$ is dense in $C(\mathbb{R}^m, \mathbb{R}^n)$. If Assumptions 3.1 and 3.2 hold then $\mathscr{F}_{\rho, \phi}$ is dense in $C(\mathscr{X}, \mathscr{Y})$.*

Proof. Since Assumptions 3.1 and 3.2 hold and since $\mathscr{F}$ is dense in $C(\mathbb{R}^m, \mathbb{R}^n)$ then $\mathscr{F}_{\rho, \phi}$ is dense in $C(\mathscr{X}, \mathrm{Im}(\rho))$ according to Lemma (B.3). By Lemma B.4, $C(\mathscr{X}, \mathrm{Im}(\rho))$ is dense in $C(\mathscr{X}, \mathscr{Y})$. Since density is transitive, then $\mathscr{F}_{\rho, \phi}$ is dense in $C(\mathscr{X}, \mathscr{Y})$. $\qquad \square$

Theorem (Sharpness of Assumption 3.1). *Let $\mathscr{X}$ be a metrizable manifold with boundary, let ϕ be continuous, and $\mathscr{F} \subseteq C(\mathbb{R}^m, \mathbb{R}^n)$. Then $\mathscr{F}_{1_{\mathbb{R}^n}, \phi}$ is dense in $C(\mathscr{X}, \mathbb{R}^n)$ if and only if ϕ is injective.*

Proof. Let $\phi : \mathscr{X} \to \mathbb{R}^m$ be a continuous function.

Assume that $\mathscr{F}$ is dense in $C(\mathbb{R}^m, \mathbb{R}^n)$. Since $1_{\mathbb{R}^n}$ is a homeomorphism of $\mathbb{R}^n$ onto itself it satisfies Assumption 3.2 (i) and (ii). Since $\mathrm{Im}(1_{\mathbb{R}^n}) = \emptyset$ then Assumption 3.2 (iii) holds. Therefore, Theorem 3.3 implies that $\mathscr{F}_{1_{\mathbb{R}^n}, \phi}$ is dense in $C(\mathscr{X}, \mathbb{R}^n)$.

Conversely, assume that ϕ is not injective. Suppose that $\mathscr{F}_{1_{\mathbb{R}^n}, \phi}$ is dense in $C(\mathscr{X}, \mathbb{R}^n)$. Since ϕ is not injective, then there exists distinct $x_1^\star, x_2^\star \in \mathscr{X}$ such that

$$\phi(x_1^\star) \neq \phi(x_2^\star). \tag{17}$$

Since $\mathscr{X}$ is metrizable then [43, Theorem 32.2] implies that $\mathscr{X}$ is normal; i.e. : if $K_1, K_2 \subseteq \mathscr{X}$ are disjoint closed subsets then there exist disjoint open subsets $U_1, U_2 \subseteq \mathscr{X}$ such that $K_i \subseteq U_i$ for $i = 1, 2$. Since $\mathscr{X}$ is normal then [43, Urysohn's Lemma; Theorem 33.1] implies that $C(\mathscr{X}, \mathbb{R})$ separates points in $\mathscr{X}$; i.e.: for every distinct x_1, x_2 there exists some $\tilde{f}_{x_1, x_2} \in C(\mathscr{X}, \mathbb{R})$ such that

$$\tilde{f}_{x_1, x_2}(x_1) \neq \tilde{f}_{x_1, x_2}(x_2). \tag{18}$$

Since $\mathbb{R}$ is a metric subspace of $\mathbb{R}^m$ then the inclusion map $\iota : \mathbb{R} \to \mathbb{R}^m$ taking x to $(x, 0, \ldots, 0)$ is continuous and by definition it is injective. Therefore, by (18) the function $f \triangleq \iota \circ \tilde{f}_{x_1^\star, x_2^\star} : \mathscr{X} \to \mathbb{R}^n$ satisfies

$$f(x_1) \neq f(x_2). \tag{19}$$

Since $\mathscr{F}_{1_{\mathbb{R}^n}, \phi}$ was assumed to be dense in $C(\mathscr{X}, \mathbb{R}^n)$ then there exists a sequence $\{f_k\}_{k \in \mathbb{N}}$ in $\mathscr{F}_{1_{\mathbb{R}^n}, \phi}$ converging to f in $C(\mathscr{X}, \mathbb{R}^n)$. Since $\mathscr{X}$ is locally-compact then [43, Theorem 46.10] implies that the evaluation function $e : \mathscr{X} \times C(\mathscr{X}, \mathbb{R}^n)$ mapping $(x, g) \mapsto g(x)$ is continuous. Since $\mathbb{R}^m$ is a metric space then $C(\mathscr{X}, \mathbb{R}^m)$ is a metric and since continuity in metric spaces is equivalent to sequential continuity then in particular

$$\lim_{k \uparrow \infty} e(x_i^\star, f_k) = e\left(x_i^\star, \lim_{k \uparrow \infty} f_k\right) = e(x_i^\star, f) = f(x_i^\star), \tag{20}$$

for $i = 1, 2$. Since, for each $k \in \mathbb{N}$, $f_k \in \mathscr{F}_{1_{\mathbb{R}^n}, \phi}$ then there exists some $g_k \in C(\mathbb{R}^m, \mathbb{R}^n)$ satisfying $f_k = g_k \circ \phi$. Hence, (17) implies that, for each $k \in \mathbb{N}$, $f_k(x_1^\star) = g_k \circ \phi(x_1^\star) = g_k \circ \phi(x_2^\star) = f_k(x_2^\star)$; thus,(20) implies that

$$f(x_2^\star) = \lim_{k \uparrow \infty} e(x_2^\star, f_k) = e(x_2^\star, f_k) = f_k(x_2^\star) = f_k(x_1^\star) = e(x_1^\star, f_k) = \lim_{k \uparrow \infty} e(x_1^\star, f_k) = e(x_1^\star, f) = f(x_1^\star).$$

However, $f(x_1^\star) \neq f(x_2^\star)$ according to (19); a contradiction. Therefore, $\mathscr{F}_{1_{\mathbb{R}^n}, \phi}$ cannot be dense in $C(\mathscr{X}, \mathbb{R}^n)$. $\qquad \square$

Corollary. *If ϕ is a continuous injective map, ρ is a surjective covering projection, and $\mathcal{F}$ is dense in $C(\mathbb{R}^d, \mathbb{R}^D)$ then $\mathcal{F}_{\rho,\phi}$ is dense in $C(\mathcal{X}, \mathcal{Y})$. In particular, ϕ and ρ may be homeomorphisms.*

Proof. Since ρ is a continuous surjection then $\mathrm{Int}(\rho)$ is trivially dense in $\mathcal{Y}$, since

$$\mathcal{Y} - \mathrm{Int}(\rho) = \mathcal{Y} - \mathcal{Y} = \emptyset.$$

Therefore, Assumptions 3.2 (ii) holds. Similarly, since $\mathcal{Y} - \mathrm{Int}(\rho) = \emptyset$ then $\partial\mathrm{Im}(\rho) = \emptyset$. Since the Cartesian product between any set and an empty-set is the empty-set, then let

$$U \triangleq \emptyset = \partial\mathrm{Im}(\rho) = \partial\mathrm{Im}(\rho) \times [0,1).$$

Therefore, $\psi(x) = \emptyset$ satisfies Assumption 3.2 (ii). By assumption, ρ is continuous and has a section. Therefore Assumption 3.2 (i.a) holds. Thus, Assumptions 3.2 holds.

Likewise, ϕ was taken to be continuous and injective. Therefore, Assumption 3.1 holds. Thus, the result follows from Theorem 3.3 since $\mathcal{F}$ is dense in $C(\mathbb{R}^m, \mathbb{R}^n)$.

Note, that if ϕ and ρ are homeomorphisms then both are continuous bijections. Moreover, ρ has a continuous two-sided inverse, and in particular, a continuous section. $\square$

Proposition B.5 (Homeomorphic case is Sharp). *Let ϕ and ρ be homeomorphisms. Then $\mathcal{F}$ is dense in $C(\mathbb{R}^m, \mathbb{R}^n)$ if and only if $\mathcal{F}_{\rho,\phi}$ is dense in $C(\mathcal{X}, \mathcal{Y})$.*

Proof. Since every homeomorphism ρ has a continuous section, namely ϕ^{-1}, then by Corollary 3.6 if $\mathcal{F}$ is dense in $C(\mathbb{R}^m, \mathbb{R}^n)$ then so is $\mathcal{F}_{\rho,\phi}$. Since ϕ is a homeomorphism and $\mathbb{R}^m$ is a locally-compact Hausdorff space then $\mathcal{X}$ is locally-compact and since ρ is also a homeomorphism then similarly $\mathcal{Y}$ is a locally-compact Hausdorff space, since $\mathbb{R}^n$ is. Therefore, by [43, Exercise 46.10]the maps $\Phi_{\rho,\phi} : C(\mathbb{R}^m, \mathbb{R}^n) \to C(\mathcal{X}, \mathcal{Y})$ mapping $f \mapsto \rho \circ f \circ \phi$ and $\Psi_{\rho,\phi} : C(\mathcal{X}, \mathcal{Y})$ mapping $g \mapsto \rho^{-1} \circ g \circ \phi^{-1}$ are continuous with respect to the compact-open topology, because ρ, ρ^{-1}, ϕ, and ϕ^{-1} are all continuous. Observe that $\Phi_{\rho,\phi} \circ \Psi_{\rho,\phi} = 1_{C(\mathcal{X}, \mathcal{Y})}$, the identity map on $C(\mathcal{X}, \mathcal{Y})$, and $\Psi_{\rho,\phi} \circ \Phi_{\rho,\phi} = 1_{C(\mathbb{R}^m, \mathbb{R}^n)}$, the identity map on $C(\mathbb{R}^m, \mathbb{R}^n)$. Hence, $\Phi_{\rho,\phi}$ is a homeomorphism with inverse $\Psi_{\rho,\phi}$. Now, since $\mathcal{F}_{\rho,\phi} = \Phi(\mathcal{F})$, $\mathcal{F} = \Psi_{\rho,\phi}(\mathcal{F}_{\rho,\phi})$, and since homeomorphisms take dense sets to dense sets then $\mathcal{F}$ is dense in $C(\mathbb{R}^m, \mathbb{R}^n)$ if and only if $\mathcal{F}_{\rho,\phi}$ is dense in $C(\mathcal{X}, \mathcal{Y})$. $\square$

Corollary. *If ϕ is a continuous injective map, ρ is a continuous surjection with a section, $\mathcal{X}$ is connected and simply connected, and $\mathcal{F}$ is dense in $C(\mathbb{R}^d, \mathbb{R}^D)$ then $\mathcal{F}_{\rho,\phi}$ is dense in $C(\mathcal{X}, \mathcal{Y})$.*

Proof. The proof is identical to the proof of Corollary 3.8 with the only difference that Assumption 3.2 (i.b) holds by assumption instead of Assumption 3.2 (i.a). $\square$

We now establish Theorem 3.10.

Theorem (Geometric Version). *Let $\mathcal{Y}$ be a metrizable manifold with boundary, for which $\mathrm{Int}(\mathcal{Y})$ is a smooth manifold, $\mathcal{X}$ is locally-compact, and $\mathcal{F}$ is dense in $C(\mathbb{R}^m, \mathbb{R}^n)$. If ϕ satisfies Assumption 3.1 and ρ satisfies Assumption 3.9 then $\mathcal{F}_{\rho,\phi}$ is dense in $C(\mathcal{X}, \mathcal{Y})$.*

Proof. For the first portion of the proof we show that $C(\mathbb{R}^n, \mathrm{Im}(\rho))$ is dense in $C(\mathbb{R}^n, \mathcal{Y})$. This is achieved in the following steps. First, $C(\mathbb{R}^n, \mathrm{Im}(\rho))$ is shown to be dense in $C(\mathbb{R}^n, \mathrm{Int}(\mathcal{Y}))$, then that $C(\mathbb{R}^n, \mathrm{Int}(\mathcal{Y}))$ is dense in $C(\mathbb{R}^n, \mathcal{Y})$, and then by the transitivity of density it follows that that $C(\mathbb{R}^n, \mathrm{Im}(\rho))$ is dense in $C(\mathbb{R}^n, \mathcal{Y})$.

Since $\mathbb{R}^n$ and $\mathrm{Int}(\mathcal{Y})$ are smooth manifolds without boundary then, [25, Theorem 2.2] implies that $C^\infty(\mathbb{R}^n, \mathcal{Y})$ is dense in $C(\mathbb{R}^n, \mathcal{Y})$ for a strictly stronger topology than the topology of uniform convergence on compacts. Thus, in particular, $C^\infty(\mathbb{R}^n, \mathcal{Y})$ is dense in $C(\mathbb{R}^n, \mathcal{Y})$ for the topology of uniform convergence on compacts. By the transitivity of density it is therefore sufficient to show, under Assumption 3.9, that $C^\infty(\mathbb{R}^n, \mathrm{Im}(\rho))$ is dense in $C^\infty(\mathbb{R}^n, \mathrm{Int}(\mathcal{Y}))$ to conclude that $C(\mathbb{R}^n, \mathrm{Im}(\rho))$ is dense in $C(\mathbb{R}^n, \mathrm{Int}(\mathcal{Y}))$.

If $\mathrm{Int}(\mathcal{Y}) = \mathrm{Im}(\rho)$ then the claim holds vacuously. Therefore, assume that $\mathrm{Int}(\mathcal{Y}) \neq \mathrm{Im}(\rho)$. By definition, a smooth map $f : \mathbb{R}^n \to \mathrm{Int}(\mathcal{Y})$ is transverse to some $\mathrm{Int}(\mathcal{Y}) - \mathrm{Im}(\rho)$ (i.e.: the inclusion map $\iota_{\mathrm{Int}(\mathcal{Y})-\mathrm{Im}(\rho)} : \mathrm{Int}(\mathcal{Y}) - \mathrm{Im}(\rho) \to \mathrm{Int}(\mathcal{Y}))$ if for every $y \in \mathrm{Int}(\mathcal{Y}) - \mathrm{Im}(\rho)$

$$\mathrm{Im}(df_x) + T_{f(x)}(\mathrm{Int}(\mathcal{Y}) - \mathrm{Im}(\rho)) = T_{f(x)}(\mathrm{Int}(\mathcal{Y})). \tag{21}$$

However, Assumption 3.9 (ii) requires that $\dim(\text{Int}(\mathscr{Y})) - \dim(\text{Int}(\mathscr{Y}) - \text{Im}(\rho)) > n$, which implies that (21) can only hold when, for each $y \in \text{Int}(\mathscr{Y}) - \text{Im}(\rho)$

$$\{0\} = \text{Im}(df_x) \cap T_{f(x)}(\text{Int}(\mathscr{Y}) - \text{Im}(\rho)). \tag{22}$$

By [22, Chapter 1, Excersize 4] , (22) implies that $f(\mathbb{R}^n) \cap \text{Int}(\mathscr{Y}) - \text{Im}(\rho) = \emptyset$. Therefore, $f \in C^\infty(\mathbb{R}^n, \text{Int}(\mathscr{Y}))$ is transversal to $\text{Int}(\mathscr{Y}) - \text{Im}(\rho)$ only if $f \in C(\mathbb{R}^n, \text{Im}(\rho))$. Since the set of all $f \in C^\infty(\mathbb{R}^n, \text{Int}(\mathscr{Y}))$ is dense in $C^\infty(\mathbb{R}^n, \text{Int}(\mathscr{Y}))$, by [25, Theorem 2.1], then $C^\infty(\mathbb{R}^n, \text{Im}(\rho))$ is dense in $C^\infty(\mathbb{R}^n, \text{Int}(\mathscr{Y}))$. Consequentially, $C(\mathbb{R}^n, \text{Im}(\rho))$ is dense in $C(\mathbb{R}^n, \text{Int}(\mathscr{Y}))$.

Consider the inclusion map $\iota_{\text{Int}(\mathscr{Y})} \to \mathscr{Y}$. Then $\iota_{\text{Int}(\mathscr{Y})}$ is an embedding of $\text{Int}(\mathscr{Y})$ into with section the identity map and it is dense in $\mathscr{Y}$. Therefore, Assumption 3.2 (i)-(ii) hold. Assumption 3.2 (iii) is precisely the definition of a $\partial \text{Im}(\iota_{\text{Int}(\mathscr{Y})}) = \partial \text{Int}(\mathscr{Y})$ being collared (see [8, Section II; page 332]). By definition, $\partial \text{Int}(\mathscr{Y})$ is the *boundary*, in the sense of manifolds with boundary, of the manifold with boundary $\mathscr{Y}$. Since $\iota_{\text{Int}(\mathscr{Y})}$ is a surjection onto $\text{Int}(\mathscr{Y})$ then Assumption 3.2 (iii) states that the boundary of the manifold with boundary $\mathscr{Y}$ must be collared. However, since $\mathscr{Y}$ is metrizable then this is guaranteed by [8, Theorem 2]. Thus, Assumption 3.2 (iii) holds. Therefore, Assumption 3.2 holds and Lemma B.4 guarantees that $C(\mathbb{R}^n, \text{Int}(\mathscr{Y}))$ is dense in $C(\mathbb{R}^n, \mathscr{Y})$. Hence, $C(\mathbb{R}^n, \text{Im}(\rho))$ is dense in $C(\mathbb{R}^n, \mathscr{Y})$.

Since $\mathscr{F}$ is dense in $C(\mathbb{R}^m, \mathbb{R}^n)$, since the identity map satisfies Assumption 3.1, then Assumption 3.2 implies that Lemma B.3 applies. Whence, $\mathscr{F}_{\rho, 1_{\mathbb{R}^m}}$ is dense in $C(\mathbb{R}^n, \text{Im}(\rho))$. Therefore, $\mathscr{F}_{\rho, 1_{\mathbb{R}^m}}$ is dense in $C(\mathbb{R}^n, \mathscr{Y})$. Applying Lemma B.2, we conclude that $\mathscr{F}_{\rho, \phi}$ is dense in $C(\mathscr{X}, \text{Im}(\rho))$ and therefore in $C(\mathscr{X}, \mathscr{Y})$ for the compact-open topology by Lemma B.4.

Finally, notice that, since Assumptions 3.2 (i)-(ii) were assumed to hold then Theorem 3.3 implies that $\mathscr{F}_{\rho, \phi}$ is dense in $C(\mathscr{X}, \mathscr{Y})$ for the co topology. Since every singleton is closed in $\mathbb{R}^m$ and ϕ is continuous and injective then every $x \in \mathscr{X}$, $\{x\}$ is the continuous pre-image of the singleton $\{\phi(x)\} \in \mathbb{R}^m$ by ϕ. Since ϕ is continuous, then $\{x\}$ is closed therefore [43, Theorem 17.8] implies that $\mathscr{X}$ is Hausdorff. Since $\mathscr{X}$ was assumed to be locally-compact and $\mathscr{Y}$ is metrizable then [43, Theorem 46.8] implies that the co topology and the ucc topology on $C(\mathscr{X}, \mathscr{Y})$ coincide for any metric topologizing $\mathscr{Y}$. $\qquad \square$

Theorem B.6 (Universal Classification: General Case). *Let $\{0,1\}^n$ be equipped with the n-fold product of the Sierpiński topology on $\{0,1\}$, ϕ satisfy Assumption 3.1, $\rho : \mathbb{R}^n \to (0,1)^n$ be a homeomorphism, $\alpha \in (0,1)$, and $\mathscr{F} \subseteq C(\mathbb{R}^m, \mathbb{R}^n)$ be dense. Let $\{\mathscr{X}_i\}_{i=1}^n$ be a set of open subsets of a metric space $\mathscr{X}$ and let $\hat{h}$ be its associated ideal classifier defined by (6). Then the following hold:*
(i) (Hard-Soft Decomposition) There exist continuous functions $\hat{s}_i \in C(\mathscr{X}, [0,1])$ such that

$$\hat{h} = I_{(\alpha,1]} \bullet (s_1, \ldots, s_n) \qquad \hat{s}_i^{-1}[(\alpha,1]] = \mathscr{X}_i, (\forall i = 1, \ldots, n)$$

(ii) (Universal Classification) There exists a sequence $\{f_k\}_{k \in \mathbb{N}}$ in $\mathscr{F}$ such that:
(a) (Soft Classification) For each non-empty compact subset $\kappa \subseteq \mathscr{X}$ and every $\varepsilon > 0$, there is some $K \in \mathbb{N}^+$ such that

$$\sup_{x \in \kappa} \max_{i=1, \ldots, n} |\rho \circ f_k \circ \phi(x)_i - \hat{s}_i(x_i)| < \varepsilon, \qquad (\forall k \geq K)$$

(b) (Hard Classification) $I_{(\alpha,1]} \bullet \rho \circ f_k \circ \phi$ converges to $\hat{h}$ in $C(\mathscr{X}, \{0,1\}^n)$ for the co-topology.

Furthermore, $\mathscr{F}_{\rho, \phi}$ is dense in $C(\mathscr{X}, [0,1]^n)$.

Proof. Since $\mathscr{X}$ is a metric space and since each $\mathscr{X}_i$ is open then by [2, Corollary 3.19] $\mathscr{X}_i$ is an open F_σ set, i.e.: an open set which is the countable intersection of closed sets. By [15, Corollary 1.5.13] there exists continuous function $\tilde{s}_i : \mathscr{X} \to [0,1]$ such that $\tilde{s}_i^{-1}[(0,1]] = \mathscr{X}_i$. Since $\alpha \in (0,1)$, then for each $i = 1, \ldots, n$, let $s_i(x) = \alpha + (1-\alpha)\tilde{s}_i(x)$. Then, s_i is continuous from $\mathscr{X}$ to $[0,1]$ and satisfies $s_i^{-1}[(\alpha,1]] = \mathscr{X}_i$. Define $\hat{s} \triangleq (s_1, \ldots, s_n)$. Note $\hat{s}$ is continuous from $\mathscr{X}$ to $[0,1]^n$ since its components continuously map $\mathscr{X}$ to $[0,1]$. Next, define the map

$$\Phi_\alpha : [0,1]^n \to \{0,1\}^n$$
$$x \to (I_{(\alpha,1]}(x_i))_{i=1}^n,$$

and note that $\Phi \circ \hat{s} = I_{(\alpha,1]} \bullet \hat{s} = \hat{h}$, by construction. Thus, (i) holds.

Since ϕ is an embedding of $\mathscr{X}$ into $\mathbb{R}^m$ for which $\phi(\mathscr{X})$ is a retract, ρ is a homeomorphism of $\mathbb{R}^n$ onto $(0,1)^n$, and since $[0,1]^n$ is a metrizable manifold with boundary whose interior is $\text{Int}\left([0,1]^n\right) = (0,1)^n$ then Theorem 3.10 implies that $\mathscr{F}_{\rho,\phi}$ is dense in $C(\mathscr{X},[0,1]^n)$.

In particular, this implies that, since $\hat{s} \in C(\mathscr{X},[0,1]^n)$, and $\mathscr{F}_{\rho,\phi}$ is dense therein, then there exists a sequence $\{f_k\}_{k\in\mathbb{N}}$ in $\mathscr{F}$ such that, for every non-empty compact-subset $\kappa \subseteq \mathscr{X}$, every $\tilde{\varepsilon} > 0$

$$\sup_{x\in\kappa} \|\rho \circ f_k \circ \phi(x)_i - s_i(x_i)\| < \tilde{\varepsilon} \qquad (\forall k \geq K). \tag{23}$$

Applying [12, Theorem 3.1] to (23) yields a constant $C > 0$, independent of $x, \{f_k\}_{k\in\mathbb{N}}$, and of $\hat{s}$, satisfying

$$\sup_{x\in\kappa} \max_{i=1,\dots,n} |\rho \circ f_k \circ \phi(x)_i - s_i(x_i)| \leq C \sup_{x\in\kappa} \|\rho \circ f_k \circ \phi(x)_i - s_i(x_i)\| < C\varepsilon \qquad (\forall k \geq K).$$

Setting $\varepsilon \triangleq C\tilde{\varepsilon}$ yields (ii.a). Thus, we only need to verify (ii.b).

Equip $\{0,1\}$ with the Sierpiński topology $\{\emptyset, \{1\}, \{0,1\}\}$ and denote this space by S. Then, for any T_1 space, and any open set U, the indicator function I_U of U is continuous to $\{0,1\}$, see [54, Chapter 7]. In particular, since $[0,1]$ with the Euclidean topology is a metric space, then in particular, it is T_1. Moreover, for any $\alpha \in (0,1)$, the set $(\alpha,1]$ is open in $[0,1]$. Therefore, the map $I_{(\alpha,1]} : [0,1] \to S$ is continuous and thus the map Φ_α is continuous. By [43], since post-composition by continuous functions defines a continuous map between $C(\mathscr{X},[0,1]^n)$ and $C(\mathscr{X},\{0,1\}^n)$ when both are equipped with the compact-open topology then the map

$$\Phi : C(\mathscr{X},[0,1]^n) \to C(\mathscr{X},\{0,1\}^n)$$
$$f \to \Phi_\alpha \circ f \qquad ,$$

is continuous. Since continuous functions preserve convergent sequences and since $\{\rho \circ f_k \circ \phi\}_{k\in\mathbb{N}}$ converges to $\hat{s}$ in the compact-open topology on $C(\mathscr{X},[0,1]^n)$ then

$$\{\Phi(\rho \circ f_k \circ \phi)\}_{k\in\mathbb{N}} = \{\Phi_\alpha \circ \rho \circ f_k \circ \phi\}_{k\in\mathbb{N}} = \left\{I_{(\alpha,1]} \bullet \rho \circ f_k \circ \phi\right\}_{k\in\mathbb{N}}$$

converges to $\Phi(\hat{s}) = \Phi_\alpha \circ \hat{s} = I_{(\alpha,1]} \bullet \hat{s}$ in the compact-open topology on $C(\mathscr{X},0,1^n)$. This verifies (ii.b). Thus, (ii) holds. $\qquad\square$

Corollary (Universal Classification: Deep Feed-Forward Networks). *Let $\{\mathscr{X}_i\}_{i=1}^n$ be open subsets of $\mathscr{X}$, and $\hat{h}$ be their associated ideal classifier. Let $\phi : \mathscr{X} \to \mathbb{R}^n$ be a continuous injective feature map. Let σ be a continuous, locally-bounded, and non-constant activation function. Let ρ either be the component-wise logistic function. Then there exists a sequence $\{f_k\}_{k\in\mathbb{N}^+}$ of DNNs satisfying the conclusions of Theorem 3.11.*

Proof. By [36] the set $\mathscr{N}\mathscr{N}^\sigma$ of all feed-forward DNNs is dense in $C(\mathbb{R}^m,\mathbb{R}^n)$. Moreover, the identity map $1_{\mathbb{R}^m}$ on $\mathbb{R}^m$ satisfies Assumption 3.1. Furthermore, both the soft-max

$$x \to \left(\frac{e^{x_j}}{\sum_{i=1}^n e^{x_i}}\right)_{j=1}^n$$

and the component-wise logistic function

$$x \to \left(\frac{e^{x_j}}{1+e^{x_j}}\right)_{j=1}^n,$$

are continuous bijections with continuous inverses, from $\mathbb{R}^n$ onto $(0,1)^n$. In particular, they satisfy Assumption 3.2. Therefore Theorem 3.11 applies and the conclusion holds. $\qquad\square$

Corollary (Universal Classification: Deep CNNs). *Let $2 \leq s \leq n$, $\{\mathscr{X}_i\}_{i=1}^n$ be open subsets of $\mathscr{X}$, and $\hat{h}$ be their associated ideal classifier. Let $\phi : \mathscr{X} \to \mathbb{R}^n$ be a continuous injective feature map and let $\rho : \mathbb{R} \to (0,1)$ be the logistic function. Then there is a sequence of deep CNNS $\{f_k\}_{k\in\mathbb{N}^+}$ in $Conv_{\rho,\phi}^s$ satisfying the conclusion of Theorem 3.11.*

Proof. Since $Conv^s$ with these specifications is dense in $C(\mathbb{R}^d,\mathbb{R})$ by [57], ϕ is a continuous injection, and ρ is a homeomorphism of $\mathbb{R}$ onto $(0,1)$ then the result follos from Theorem 3.11. $\qquad\square$

Corollary (Cartan-Hadamard Version). *Let $\mathscr{F}$ be dense in $C(\mathbb{R}^m,\mathbb{R}^n)$, let $\mathscr{M}$ and $\mathscr{N}$ be Cartan-Hadamard manifolds of dimension m and n. Then, $\mathscr{F}_{\mathrm{Log}_p^{\mathscr{M}},\mathrm{Exp}_q^{\mathscr{N}}}$ is dense in $C(\mathscr{M},\mathscr{N})$.*

Proof. Since $\mathscr{N}$ and $\mathscr{M}$ are Hadamard manifolds then the Cartan-Hadamard Theorem, see [31, Corollary 6.9.1], implies that $\mathrm{Exp}_p^{\mathscr{M}}$ and $\mathrm{Log}_q^{\mathscr{N}}$ are diffeomorphisms and in particular are homeomorphisms. Since $\mathscr{N}$ and $\mathscr{M}$ have no boundary then the result follows from Corollary 3.8. $\qquad\square$

Corollary (Universal Approximation for Symmetric Positive-Definite Matrices). *Let $d,D \in \mathbb{N}^+$ and $\mathscr{F} \subseteq C(\mathbb{R}^{d(d+1)/2},\mathbb{R}^{D(D+1)/2})$ be ucc-dense. Then, for any $A \in P_d^+$ and $B \in P_D^+$, $\mathscr{F}_{\mathrm{Log}_A,\mathrm{Exp}_B}$ is ucc-dense in $C(P_d^+,P_D^+)$. In particular, if σ is a continuous, locally-bounded, and non-polynomial activation function then $\mathscr{N}\mathscr{N}^{\sigma}_{\mathrm{Log}_A,\mathrm{Exp}_B}$ is ucc-dense in $C(P_d^+,P_D^+)$.*

Proof. As discussed in [7, Section 3.3], the space P_d^+ under the metric d_+ is a complete, connected, and simply connected Riemannian manifold with non-positive curvature. Therefore, Corollary 3.14 applies. $\qquad\square$

Corollary (Hyperbolic Neural Networks are Universal). *Let σ be a continuous, non-polynomial, locally-bounded activation function and $c > 0$. Then for every $g \in C(\mathbb{D}_c^m,\mathbb{D}_c^n)$, every $\varepsilon > 0$, and every compact subset $K \subseteq \mathbb{D}_c^m$ there exists a hyperbolic neural network $g_{\varepsilon,K,c}$ in (10) satisfying*

$$\sup_{x \in K} d_c(g(x),g_{\varepsilon,K,c}) < \varepsilon.$$

Proof. Since $\mathbb{D}_c^n$ is a complete, connected, simply connected, of non-positive sectional curvature Riemannian manifold then it is of Cartan-Hadamard type. Therefore, the result follows directly form Corollary 3.14. $\qquad\square$

Theorem. *If Assumption 3.17 holds, then there exists a measurable subset $\Omega' \subseteq \left\{ \omega \in \Omega : \overline{\mathscr{N}\mathscr{N}^{\sigma}_{2,k}(\omega)} = C(\mathbb{R}^d,\mathbb{R}^D) \right\}$ satisfying $\mathbb{P}(\Omega') = 1$.*

Proof. Let $\mathrm{Mat}_{k,l}$ denote the collection of $k \times l$ matrices with real coefficients, identified with the Euclidean space $\mathbb{R}^{kl}$. By Assumption 3.17 (ii) σ is a injective continuous function. Therefore so is the function

$$\sigma_i : \mathbb{R}^{d_{i+1}} \to \mathbb{R}^{d_{i+1}}$$
$$(x_j)_{j=1}^{d_{i+1}} \mapsto (\sigma(x_j))_{j=1}^{d_{i+1}}.$$

Fix $\omega \in \Omega$. Since $d_i = d_{i+1}$, then the map $W_i(\cdot,\omega)$ is affine injection if and only if $A_i(\omega)$ is of rank d_i. Since the composition of continuous injections is again continuous then the map

$$\begin{aligned} \phi(x,\omega) : \mathbb{R}^d &\to \mathbb{R}^{d_k} \\ x &\mapsto \Sigma_k \circ A_k(x,\omega) \circ \cdots \circ \Sigma_1 \circ A_1(x,\omega), \end{aligned} \tag{24}$$

is continuous and injective if each $A_i(\omega)$ is of full rank. Equivalently, it is enough to show that the smallest singular value of A_i, which we denote by $\lambda^{\star}(A_i)$ to avoid confusion with the notation for activation functions, should be bounded away from 0 with probability 1.

Since each A_i is a random $d_{i+1} \times d_i$ matrix and $d_{i+1} \geq d_i$, then it contains a random $d_i \times d_i$ (square) sub-matrix. It is enough to show that this random sub-matrix is of full-rank. Therefore, without loss of generality, we assume that $d_{i+1} = d_i$ for all $i = 1,\dots,k$.

Since each A_i is a random $d_i \times d_i$ square matrix and Assumptions 3.17 hold then [53, Theorem 1.3 (Universality for the Least Singular Value)] applies whence. Therefore,

$$\begin{aligned} \mathbb{P}\left(\lambda^{\star}(A_i) > 0 \,\forall (i=1,\dots,k)\right) &= 1 - \prod_{i=1}^{k} \mathbb{P}\left(\lambda^{\star}(A_i) > 0\right) \\ &= 1 - \lim_{t \downarrow 0^+} \prod_{i=1}^{k} \mathbb{P}\left(\lambda^{\star}(A_i) \leq \frac{t}{\sqrt{d_i}}\right) \\ &= \left(\int_0^t \frac{1+\sqrt{x}}{2\sqrt{x}} e^{-\frac{x}{2}-\sqrt{x}} dx + O\left(\frac{1}{x^c}\right) \right)^k \\ &= 0, \end{aligned}$$

where $c > 0$ is an absolute constant independent of ξ. Therefore, the set

$$\Omega' \triangleq \{\omega \in \Omega : \lambda_i(A_i) > 0 \, \forall i = 1,\ldots,k\} \subseteq \left\{\omega \in \Omega : \overline{\mathcal{NN}^\sigma(\omega)} = C(\mathbb{R}^d, \mathbb{R}^D)\right\}$$

is Σ-measurable and $\mathbb{P}(\Omega') = 1$. This concludes the proof. $\qquad\square$

Corollary (Sub-Gaussian Case with Sigmoid Activation)**.** *Let $X_i = Z_i$ for each $i = 1,\ldots,k$ be independent sub-Gaussian random-variables, $\sigma(x) = \frac{1}{1+e^{-x}}$, and $d_i = d$ for each $i = 1,\ldots,k$. Then the conclusion of Theorem 3.18 holds.*

Proof. Since the X_i and Z_i are sub-Gaussian random variables, then by [9] all their moments are finite. This verifies Assumption 3.17 (iv). Since they are assumed to be standardized then Assumption 3.17 (iii) holds. Since the sigmoid activation function is continuous and monotonically increasing then Assumption 3.17 (ii) holds. Lastly, Assumption 3.17 (i) holds by construction. Therefore, Theorem 3.18 applies and the conclusion follows. $\qquad\square$

Corollary (Bernoulli Case with PReLU Activation)**.** *Suppose that for every $i, j = 1,\ldots,k$, X_i and Z_j i.i.d. copies of a random variable taking values $\{-1,1\}$ with probabilities $\{\frac{1}{2}, \frac{1}{2}\}$. Let $d_i = d$ for each $i = 1,\ldots,k$ and σ be the PReLU activation function of [23]. Then Assumptions 3.17 are met; thus the conclusion of Theorem 3.18 holds.*

Proof. Since the PReLU activation function is continuous and strictly increasing, then Assumption 3.17 (ii) holds. Since, Bernoulli random variables have all finite C^{th}-moments, for $C > 0$, then Assumption 3.17 (iv) holds. Since X_i and Z_i are taken to be standardized, then Assumption 3.17 (iii) holds. Assumption 3.17 (i) holds by hypothesis. Therefore, the result follows from Theorem 3.18. $\quad\square$

Proposition. *Let σ be a continuous and strictly increasing activation function, $J \in \mathbb{N}_+$, $A_1,\ldots,A_J$ be $m \times m$ matrices, and $b_1,\ldots,b_J \in \mathbb{R}^d$. Let $\phi(x) \triangleq \phi_J(x)$ where*

$$\phi_j(x) \triangleq \sigma \bullet \left(\exp(A_j)\phi_{j-1}(x) + b_j\right) \qquad \phi_0(x) \triangleq x, \quad j = 1,\ldots,J, \tag{25}$$

where $\exp$ is the matrix exponential. Then ϕ satisfies Assumption 3.1.

Proof. Since σ is strictly increasing then it is injective. Therefore, the map $\phi_1(x) \triangleq (\sigma(x_k))_{k=1}^m$ is continuous and injective from $\mathbb{R}^m$ to itself. Next, for each $1 \le j \le J$ with $j \in \mathbb{N}_+$, each A_j is a $d \times d$ matrix then its exponential $\exp(A_j)$ is in the general linear group (see [33]) and therefore it is an invertible $m \times m$ matrix; hence, the map $\phi_{2,j}(x) \triangleq \exp(A_j)x$ is a continuous bijection from $\mathbb{R}^m$ to itself. Finally, for each $b_j \in \mathbb{R}^m$, the map $\phi_3(x) \triangleq x + b_j$ is a continuous bijection, since it is affine with inverse $y \mapsto y - b_j$; hence, $\phi_{3,j}$ is a continuous injection from $\mathbb{R}^m$ to itself. Since the composition of continuous functions is again continuous and the composition of injective functions is again injective then the map $\phi = \phi_1 \circ \phi_{3,J} \circ \phi_{2,J} \circ \cdots \circ \phi_1 \circ \phi_{3,1} \circ \phi_{2,1}$ of (25) is a continuous injection; hence, it satisfies Assumption 3.1. $\qquad\square$

Proposition. *In the setting of Proposition 3.21, if σ is also surjective then ϕ is a homeomorphism, and in particular it satisfies Assumption 3.2.*

Proof. By Proposition 3.21 the map ϕ is continuous and injective. Since σ is a continuous strictly increasing function then it is injective. Moreover by [26] it is a homeomorphism onto its image. Since σ was assumed to be a surjection then this means that σ is a homeomorphism from $\mathbb{R}$ to itself. Therefore, the map $\prod_{i=1}^m \sigma : \mathbb{R}^m \to \mathbb{R}^m$ sending x to $(\sigma(x_i))_{i=1}^m$ is a homeomorphism by [43, Theorem 19.6]. In the proof of Proposition 3.21, above, it was shown that for each $1 \le j \le J$, $j \in \mathbb{N}_+$, the maps $x \mapsto \exp(A_j)x$ and $x \mapsto x + b_j$ are homeomorphisms. Therefore, since the composition of homeomorphisms is again a homeomorphism then ϕ is a homeomorphism. In particular, it satisfies Assumption 3.2 since ϕ^{-1} exists and is continuous (thus Assumption 3.2 (i) holds), $\text{Im}\,\phi = \mathbb{R}^m$ (thus Assumption 3.2 (ii) holds), and since $\partial\,\text{Im}\,\phi = \emptyset$ (thus Assumption 3.2 (iii) holds). $\qquad\square$

Proposition (Graphs of Continuous Functions are Good Feature Maps)**.** *Let $d \in \mathbb{N}_+$ and $g \in C(\mathbb{R}^m, \mathbb{R}^d)$. Then $\phi_g(x) \triangleq (x, g(x))$ satisfies Assumption 3.1.*

Proof. Let identity map $1_{\mathbb{R}^m}(x) \triangleq x$ is continuous from $\mathbb{R}^m$ to itself and by hypothesis the map $g : \mathbb{R}^m \to \mathbb{R}^d$ is continuous, therefore, by [43, Theorem 19.6] the map $1_{\mathbb{R}^m} \times g : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m \times \mathbb{R}^d = \mathbb{R}^{m+d}$ is continuous. A consequence of the same result shows that, for topological spaces $X, Y,$ and Z and a function $F : Z \to X \times Y$ is continuous, when $X \times Y$ is equipped with the product topology, if and only if both $\pi_1 \circ F : Z \to X$ and $\pi_2 \circ F : Z \to Y$ are continuous, where $\pi_1(x, y) = x$ and $\pi_2(x, y) = y$. Therefore, when $X = Y = Z = \mathbb{R}^m$ and $F(x) \triangleq (x, x)$ then

$$\pi_1 \circ F = 1_{\mathbb{R}^m} = \pi_2 \circ F, \tag{26}$$

and since $1_{\mathbb{R}^m}$ is continuous then (26) implies that F is continuous. Since the composition of continuous functions is again continuous, then $(1_{\mathbb{R}^m} \times g) \circ F(x) = (x, g(x))$ is continuous from $\mathbb{R}^m$ to $\mathbb{R}^{m+d}$. Lastly, notice that if $x_1 \neq x_2$ in $\mathbb{R}^m$ then $\|x_1 - x_2\| > 0$ and therefore

$$\|(x_1, g(x_1)) - (x_2, g(x_2))\| \geq \|x_1 - x_2\| > 0.$$

Hence, $(x_1, g(x_1)) \neq (x_2, g(x_2))$ in $\mathbb{R}^{m+d}$ and therefore, $(1_{\mathbb{R}^m} \times g) \circ F$ is a continuous and injective functions. Therefore, Assumption 3.1 holds. $\qquad\square$

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, Princeton, NJ, 2008. ISBN 978-0-691-13298-3. doi: 10.1515/9781400830244. URL https://doi.org/10.1515/9781400830244. With a foreword by Paul Van Dooren.

[2] C. D. Aliprantis and K. C. Border. *Infinite dimensional analysis*. Springer, Berlin, third edition, 2006. ISBN 978-3-540-32696-0; 3-540-32696-0. A hitchhiker's guide.

[3] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/arjovsky17a.html.

[4] N. Ay, J. Jost, H. V. Lê, and L. Schwachhöfer. *Information geometry*, volume 64 of *Series of Modern Surveys in Mathematics*. Springer, Cham, 2017. ISBN 978-3-319-56477-7; 978-3-319-56478-4. doi: 10.1007/978-3-319-56478-4. URL https://doi.org/10.1007/978-3-319-56478-4.

[5] M. Baes, C. Herrera, A. Neufeld, and P. Ruyssen. Low-rank plus sparse decomposition of covariance matrices using neural network parametrization. *arXiv preprint arXiv:1908.00461*, 2019.

[6] S. Bonnabel and R. Sepulchre. Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM J. Matrix Anal. Appl.*, 31(3):1055–1070, 2009. ISSN 0895-4798. doi: 10.1137/080731347. URL https://doi.org/10.1137/080731347.

[7] S. Bonnabel, A. Collard, and R. Sepulchre. Rank-preserving geometric means of positive semi-definite matrices. *Linear Algebra Appl.*, 438(8):3202–3216, 2013. ISSN 0024-3795. doi: 10.1016/j.laa.2012.12.009. URL https://doi.org/10.1016/j.laa.2012.12.009.

[8] M. Brown. Locally flat imbeddings of topological manifolds. *Ann. of Math. (2)*, 75:331–341, 1962. ISSN 0003-486X. doi: 10.2307/1970177. URL https://doi.org/10.2307/1970177.

[9] V. V. Buldygin and J. V. Kozačenko. Sub-Gaussian random variables. *Ukrain. Mat. Zh.*, 32(6):723–730, 1980. ISSN 0041-6053.

[10] J. Clark and D. Crisan. On a robust version of the integral representation formula of nonlinear filtering. *Probability theory and related fields*, 133(1):43–56, 2005.

[11] R. Connelly. A new proof of Brown's collaring theorem. *Proc. Amer. Math. Soc.*, 27:180–182, 1971. ISSN 0002-9939. doi: 10.2307/2037284. URL https://doi.org/10.2307/2037284.

[12] J. B. Conway. *A course in functional analysis*, volume 96 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1990. ISBN 0-387-97245-5.

[13] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.*, 2(4):303–314, 1989.

[14] A. J. De Jong et al. The stacks project. *Algebraic stacks*, 2013. URL http://stacks.math.columbia.edu.

[15] R. Engelking. General topology. 6:viii+529, 1989. Translated from the Polish by the author.

[16] A. Faragó and G. Lugosi. Strong universal consistency of neural network classifiers. *IEEE Transactions on Information Theory*, 39(4):1146–1151, 1993.

[17] P. T. Fletcher. Geodesic regression and the theory of least squares on Riemannian manifolds. *Int. J. Comput. Vis.*, 105(2):171–185, 2013.

[18] O. Ganea, G. Bécigneul, and T. Hofmann. Hyperbolic neural networks. In *Advances in neural information processing systems*, pages 5345–5355, 2018.

[19] A. Geron. https://github.com/ageron/handson-ml/tree/master/datasets/housing. Accessed: 2020-05-15.

[20] L. Gonon, L. Grigoryeva, and J. Ortega. Approximation bounds for random neural networks and reservoir systems. *ArXiv*, abs/2002.05933, 2020.

[21] C. R. Guilbault and C. J. Tirel. On the dimension of $\mathscr{Z}$-sets. *Topology Appl.*, 160(13):1849–1852, 2013. ISSN 0166-8641. doi: 10.1016/j.topol.2013.07.017. URL https://doi.org/10.1016/j.topol.2013.07.017.

[22] V. Guillemin and A. Pollack. *Differential topology*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1974.

[23] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[24] P. Henry-Labordere. *Analysis, geometry, and modeling in finance: Advanced methods in option pricing*. CRC Press, 2008.

[25] M. W. Hirsch. *Differential topology*, volume 33 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1994. ISBN 0-387-90148-5. Corrected reprint of the 1976 original.

[26] H. Hoffmann. On the continuity of the inverses of strictly monotonic functions. *Irish Math. Soc. Bull.*, (75):45–57, 2015.

[27] H. Hopf and W. Rinow. Ueber den Begriff der vollständigen differentialgeometrischen Fläche. *Comment. Math. Helv.*, 3(1):209–225, 1931. ISSN 0010-2571. doi: 10.1007/BF01601813. URL https://doi.org/10.1007/BF01601813.

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

[29] J.-i. Itoh and M. Tanaka. The dimension of a cut locus on a smooth riemannian manifold. *Tohoku Mathematical Journal, Second Series*, 50(4):571–575, 1998.

[30] J. Johnson. Deep, skinny neural networks are not universal approximators. 2018.

[31] J. Jost. *Riemannian geometry and geometric analysis*. Universitext. Springer, Cham, seventh edition, 2017. ISBN 978-3-319-61859-3; 978-3-319-61860-9. doi: 10.1007/978-3-319-61860-9. URL https://doi.org/10.1007/978-3-319-61860-9.

[32] Kaggle. California housing prices. https://www.kaggle.com/camnugent/california-housing-prices. Accessed: 2020-05-15.

[33] A. W. Knapp. *Lie groups beyond an introduction*, volume 140. Springer, 2002.

[34] A. Kratsios. Neurips2020 non euclidean universal approximation example dnn layer comparisons. https://github.com/AnastasisKratsios/NeurIPS2020_Non_Euclidean_Universal_Approximation_Example_DNN_Layer_Comparisons. Accessed: 2020-10-19.

[35] D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Boguná. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106, 2010.

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

[37] J. Lott. Some geometric calculations on Wasserstein space. *Comm. Math. Phys.*, 277(2):423–437, 2008. ISSN 0010-3616. doi: 10.1007/s00220-007-0367-3. URL https://doi.org/10.1007/s00220-007-0367-3.

[38] C. Louart, Z. Liao, R. Couillet, et al. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

[39] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang. The expressive power of neural networks: A view from the width. In *Advances in neural information processing systems*, pages 6231–6239, 2017.

[40] A. Mallasto and A. Feragen. Learning from uncertain curves: The 2-wasserstein metric for gaussian processes. In *Advances in Neural Information Processing Systems*, pages 5660–5670, 2017.

[41] G. Meyer, S. Bonnabel, and R. Sepulchre. Regression on fixed-rank positive semidefinite matrices: a riemannian approach. *Journal of Machine Learning Research*, 12(Feb):593–625, 2011.

[42] C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(Dec):2651–2667, 2006.

[43] J. Munkres. *Topology*. Pearson Education, 2014.

[44] T. Munzner. H3: Laying out large directed graphs in 3d hyperbolic space. In *Proceedings of VIZ'97: Visualization Conference, Information Visualization Symposium and Parallel Rendering Symposium*, pages 2–10. IEEE, 1997.

[45] M. Nickel and D. Kiela. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6338–6347. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7213-poincare-embeddings-for-learning-hierarchical-representations.pdf.

[46] S. Park, C. Yun, J. Lee, and J. Shin. Minimum width for universal approximation. *arXiv preprint arXiv:2006.08859*, 2020.

[47] X. Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *J. Math. Imaging Vis.*, 25(1):127–154, 2006.

[48] X. Pennec. Statistical computing on manifolds: from riemannian geometry to computational anatomy. In *LIX Fall Colloquium on Emerging Trends in Visual Computing*, pages 347–386. Springer, 2008.

[49] X. Pennec, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. *International Journal of computer vision*, 66(1):41–66, 2006.

[50] W. Rudin. *Real and complex analysis*. Tata McGraw-Hill Education, 1987.

[51] F. Sala, C. De Sa, A. Gu, and C. Re. Representation tradeoffs for hyperbolic embeddings. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4460–4469, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[52] E. H. Spanier. *Algebraic topology*. Springer-Verlag, New York, 1995. ISBN 0-387-94426-5. Corrected reprint of the 1966 original.

[53] T. Tao and V. Vu. Random matrices: the distribution of the smallest singular values. *Geom. Funct. Anal.*, 20(1):260–297, 2010. ISSN 1016-443X.

[54] P. Taylor. Foundations for computable topology. In *Foundational theories of classical and constructive mathematics*, pages 265–310. Springer, 2011.

[55] H. Toruńczyk. A correction of two papers concerning Hilbert manifolds: Concerning locally homotopy negligible sets and characterization of ℓ^2-manifolds. *Fund. Math.*, 125(1):89–93, 1985. ISSN 0016-2736. doi: 10.4064/fm-125-1-89-93. URL https://doi.org/10.4064/fm-125-1-89-93.

[56] A. A. Ungar. *Analytic hyperbolic geometry*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2005. ISBN 981-256-457-8. doi: 10.1142/9789812703279. URL https://doi.org/10.1142/9789812703279. Mathematical foundations and applications.

[57] D.-X. Zhou. Universality of deep convolutional neural networks. *Applied and computational harmonic analysis*, 48(2):787–794, 2020.