Piecewise-Linear Activations or Analytic Activation Functions: Which Produce More Expressive Neural Networks?

Anonymous authors
Paper under double-blind review

Abstract

Many currently available universal approximation theorems affirm that deep feedforward networks defined using any suitable activation functions can approximating any integrable function locally in L^1-norm. Though different approximation rates are available for deep neural networks defined using other classes of activation functions, there is little explanation for the empirically confirmed advantage that ReLU networks exhibit over their classical (e.g. sigmoidal) counterparts. Our main result demonstrates that deep networks with piecewise linear activation (e.g. ReLU or PReLU) are fundamentally more expressive than deep feedforward networks with analytic (e.g. sigmoid, Swish, GeLU, or Softplus). More specifically, we construct a non-trivial strict refinement of the topology on the space $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ of locally Lebesgue-integrable functions, in which the set of deep ReLU networks with (bilinear) pooling $\text{NN}^{\text{ReLU}+\text{Pool}}$ is dense (i.e. universal) but the set of deep feedforward networks defined using any combination of analytic activation functions with (or without) pooling layers $\text{NN}^{\omega+\text{Pool}}$ is not dense (i.e. not universal). The "separation phenomenon" persists when comparing deep ReLU networks with pooling to classical polynomial functions; which we show are also not dense in this space. Our main result is further explained by *quantitatively* demonstrating that this "separation phenomenon" between the networks in $\text{NN}^{\text{ReLU}+\text{Pool}}$ and those in $\text{NN}^{\omega+\text{Pool}}$ by showing that the networks in $\text{NN}^{\text{ReLU}+\text{Pool}}$ are capable of approximate any compactly supported Lipschitz function while *simultaneously* approximating its essential support; whereas, the networks in $\text{NN}^{\omega+\text{Pool}}$ cannot.

1 Introduction

The classical universal approximation theorems of Hornik et al. (1989) concern neural networks with sigmoidal activation functions, of which the *sigmoid* activation $\sigma(x) \overset{\text{def.}}{=} \frac{e^x}{1+e^x}$ is the prototype. In contrast, in most contemporary universal approximation theorems (Yarotsky, 2018; Gühring et al., 2020a; Lu et al., 2021; Shen et al., 2022; Opschoor et al., 2022) networks with ReLU activation function $\text{ReLU}(x) \overset{\text{def.}}{=} \max\{0,x\}$ are studied. The reason for this is largely the fact that ReLU networks are more popular in practice than σ networks, since ReLU networks tend not to encounter vanishing gradients during training and tend to learn sparser weights and biases after training, than sigmoid networks. Nevertheless, it still remains unclear if ReLU networks are genuinely more expressive than sigmoid networks. This paper takes a first step towards answering this open question.

Our main result can be informally states as follows. Suppose that Alice and Bob both want to approximate a target function f with their deep learning models. Alice has access to a very large state-of-the-art supercomputer with that can train a very deep and wide feedforward network with sigmoid activation function and Bob has a small dated laptop on which they train a deep ReLU network with much fewer neurons than Alice's model. We assume that both have access to an idealized optimizer and an infinite noiseless dataset. Our main result shows that there are non-pathological functions whose "sharp" features cannot be approximately encoded by Alice's model, irrespective of how much more computing power they have access to; however, Bob's modest setup can.

In this way, we demonstrate a *qualitative (or fundamental) gap* between the approximation capabilities of deep feedforward networks with sigmoidal vs. ReLU activation functions. This is in contrast to a *quantitative gap* wherein, given enough of a computational advantage, Alice could in principle approximate f just as well as Bob could. A fortiori, the phenomenon which we study persists even if Alice instead implements deep feedforward networks with any choice of very smooth activation function at each neuron and even if Bob implements any single other piecewise linear activation function (non-affine).

Rigorously, let $\sigma_{\text{PW-Lin}}$ be a piecewise linear activation function with at-least 2 (distinct) pieces and let $\mathrm{NN}^{\mathrm{ReLU+Pool}}$ denote the set of deep feedforward networks mapping $\mathbb{R}^d$ to $\mathbb{R}^D$ with bilinear pooling layer, defined by $\operatorname{Pool}(x_1,\dots,x_{2n}) \overset{\text{def.}}{=} (x_1x_2,\dots,x_{2n-1}x_{2n})$, at their output. Let $\mathrm{NN}^{\omega+\mathrm{Pool}}$ denote the set of deep feedforward networks where *each neuron can have any analytic activation function* (e.g. sigmoid, Swish, GeLU, Softplus, sin, etc...) and any number of bilinear pooling layers (possibly 0). In this paper, we exhibit a topology τ (constructed in Section 3) on the set of *locally integrable functions* from $\mathbb{R}^d$ to $\mathbb{R}^D$, denoted by $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$, in which $\mathrm{NN}^{\mathrm{ReLU+Pool}}$ is dense (i.e. universal) but $\mathrm{NN}^{\omega+\mathrm{Pool}}$ fails to be. Since the topology τ is stronger than the usual topology on $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ (details below) then the gap is not vacuous, in the sense that, density with respect to τ implies density in the classical sense on $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$. However, since τ is strictly stronger than the usual topology on $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ then the converse is generally false. Thus, exhibiting τ reveals the qualitative approximation gap between both neural network models. Let $C^\omega(\mathbb{R})$ denote the set of *analytic* functions from $\mathbb{R}$ to itself; i.e. $\sigma \in C^\omega(\mathbb{R})$ if σ is locally given by a convergent power series. We call a function $\sigma \in C(\mathbb{R})$ piecewise linear with at-least two pieces if $\mathbb{R}$ can be covered by a sequence of intervals on which σ is affine and there is at-least one point at which σ is not differentiable. We prove the following theorem.

Theorem 1 (Separation: Neural Networks with Piecewise-Linear vs. Analytic Activation Functions)**.** *Let $\log_2(d) \in \mathbb{N}_+$. There is a strict refinement τ of the topology on $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ which refines the metric topology on $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ and whose restriction to $L^1(\mathbb{R}^d,\mathbb{R}^D)$ is also a strict refinement of the L^1-norm topology satisfying:*

(i) ***Universality:*** *If $\sigma_{\text{PW-Lin}} \in C(\mathbb{R})$ is piecewise linear with at-least 2 pieces then $\mathrm{NN}^{\sigma_{\text{PW-Lin}}+\mathrm{Pool}}$ is dense in $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ with respect to τ,*

(ii) ***Refinement:*** *$\mathrm{NN}^{\omega+\mathrm{Pool}}$ is not dense in $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ with respect to τ,*

(iii) ***Necessary condition for convergence in τ:*** *For every $n \in \mathbb{N}_+$ and every $f \in L^1(\mathbb{R}^d,\mathbb{R}^D)$ which is essentially supported on $[-n,n]^d$, a sequence $\{f_k\}_{k\in\mathbb{N}^+}$ in $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ converges to f with respect to the separating topology τ only if all but a finite number of f_k are in $[-n,n]^d$,*

(iv) ***Sufficient condition τ-universality:*** *A subset $\mathscr{F}$ of $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ is dense for τ if, for every $f \in \operatorname{Lip}_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ which is essentially compactly supported, there is a sequence $\{f_n\}_{n=1}^\infty$ in $\mathscr{F}$ satisfying*

$$\lim_{n\uparrow\infty} \|f_n - f\|_{L^1(\mathbb{R}^d,\mathbb{R}^D)} = 0 \text{ and } \operatorname{ess-supp}(f) \cup \bigcup_{n=1}^{\infty} \operatorname{ess-supp}(f_n) \subseteq [-n_f-1, n_f+1]^d;$$

where $n_f \overset{\text{def.}}{=} \min\{n \in \mathbb{N}_+ : \operatorname{ess-supp}(f) \subseteq [-n,n]^d\}$.

(v) ***Non-Triviality:*** *For every $f \in \mathrm{NN}^{\sigma_{\text{PW-Lin}}+\mathrm{Pool}}$ the set $\{f\}$ it not open in τ,*

Together Theorem 1 (i) and (ii) show that one can refine the topological conclusion of the classical universal approximation theorem in such a way that neural networks with a non-affine piecewise linear activation function and bilinear pooling layers remain universal, but neural networks with analytic activation function(s) and bilinear pooling layers do not. A subtle point here is given by Theorem 1 (v), which guarantees that this “qualitative separation phenomenon” is not the result of us mandating that any universal approximator in $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ with respect to the separating topology τ must be capable of implementing some network in $\mathrm{NN}^{\sigma_{\text{PW-Lin}}+\mathrm{Pool}}$ (i.e. we have not simply added some non-empty subset of $\mathrm{NN}^{\sigma_{\text{PW-Lin}}+\mathrm{Pool}}$ to the base of the usual topology on $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$).

Instead, Theorem 1 (iii) and (iv) demonstrate and describe genuine qualitative difference in the approximation-theoretic behaviour of neural networks with ReLU activation function and neural networks with analytic activation functions (e.g. tanh, Swish, sigmoid, sin, etc...). Let us now explain further what the qualitative effect described by τ is and how it is captured by the networks in $\mathrm{NN}^{\sigma_{\text{PW-Lin}}+\mathrm{Pool}}$, but not those in $\mathrm{NN}^{\omega+\mathrm{Pool}}$.

The Qualitative Effect Encoded by the Separating Topology τ

The qualitative effect, captured by the separating topology τ in Theorem 1, can be explained by a model's behaviour when approximating compactly supported Lipschitz functions. Specifically, suppose that $f \in L^1(\mathbb{R}^d, \mathbb{R}^D)$ is supported in some compact subset K of $\mathbb{R}^d$ and that $\{f_n\}_{n=1}^{\infty}$ is a sequence in $L^1(\mathbb{R}^d, \mathbb{R}^D)$. The sequence $\{f_n\}_{n=1}^{\infty}$ converges to f in the classical (norm) topology on $L^1_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ if and only if the integrated and normed difference of the f_n with f converges to 0; that is, if and only if the following limit holds

$$\lim_{n\to\infty} \int_{x\in\mathbb{R}^d} \|f_n(x) - f(x)\| dx = 0. \tag{1}$$

In contrast, convergence to a compactly supported Lipschitz function (such as f) in the separating topology τ requires simultaneous approximation of f's value and correct identification of its support, instead of only requiring that f's values are approximated as in the topologies on $L^1(\mathbb{R}^d, \mathbb{R}^D)$ and on $L^1_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$.

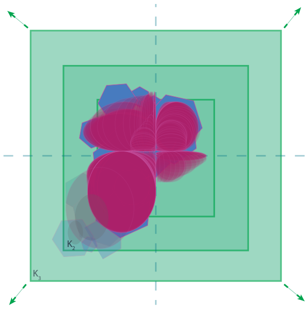

Figure 1: Approximation of a compactly supported Lipschitz function by a ReLU network with bilinear pooling

The 2-dimensional case is illustrated by Figure 1, which shows the target function $f : \mathbb{R}^2 \to \mathbb{R}$ (illustrated in red), an approximation of it by a ReLU network $\hat{f}$ with bilinear pooling (illustrated in blue), and a discretization given by a suitable of compact subsets $\{K_n\}_{n=1}^{\infty}$ of $\mathbb{R}^d$ covering $\mathbb{R}^d$ up to a set of Lebesgue measure 0. The target function's value and the network's output are represented by the vividness (alpha) of the each respective color. We see that ReLU network $\hat{f}$ with bilinear pooling is simultaneously close to the target function f's value and that $\hat{f}$ identifies the correct number of compacts subsets $\{K_1, K_2, K_3\}$ containing target function f is supported (possibly with one extra set; in this case K_3). Moreover, somewhat surprisingly, we will see that this approximation guarantee is true independently of our discretization of $\mathbb{R}^d$ (i.e. our choice of suitable compact subsets $\{K_n\}_{n=1}^{\infty}$ of $\mathbb{R}^d$).

The separating topology τ is constructed in such a way that a model is universal in $L^1_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ for τ only if it can approximate any compactly supported Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}^D$ in sense of Figure 1. In particular, it is necessary but not sufficient for a universal model to contain an expressive subset of compactly supported models. Consequentially, Theorem 1 (ii) is an immediate consequence of this property of the separating topology τ. What is more subtle are Theorem 1 (i) and (iii). Next, we focus on further understanding τ by examining (i), quantitatively.

Quantitative Approximation in the Separating Topology τ

Once τ is build, the crux of our analysis in establishing Theorem 1 reduces to obtaining a refinement of the main result of Shen et al. (2022) capable of expressing the phenomenon in Figure 1. This is our second main finding, which is a *quantitative* approximation result shows that given any compactly supported Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}^D$ we identify a neural network $\hat{f} \in \text{NN}^{\text{ReLU+Pool}}$ which can approximate f's value and its support; *quantitatively*.

A rigorous statement of our result requires some terminology. Denote the d-dimensional Lebesgue measure by μ. The *essential support* of a $f \in L^1_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$, which generalizes the support of a continuous real-valued function to elements in $L^1_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$. The *essential support* of such an $f \in L^1_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ is defined by $\text{ess-supp}(f) \overset{\text{def.}}{=} \mathbb{R}^d - \bigcup \left\{ U \subseteq \mathbb{R}^d : U \text{ open and } \|f\|(x) = 0 \ \mu\text{-a.e. } x \in U \right\}$.We say that an $f \in L^1_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ is *essentially compactly supported* if $\text{ess-supp}(f)$ is contained in a closed and bounded subset of $\mathbb{R}^d$. The regularity of a Lipschitz function $f:\mathbb{R}^d\to\mathbb{R}^D$ (i.e. a function with at-most linear growth) is quantified by its Lipschitz constant $\operatorname{Lip}(f)\overset{\text{def.}}{=}\sup_{x_1,x_2\in\mathbb{R}^d,x_1\neq x_2}\frac{\|f(x_1)-f(x_2)\|}{\|x_1-x2\|}$. The "complexity" of a subset $X\subseteq\mathbb{R}^d$ is quantified both in terms of its size $\operatorname{diam}(X)\overset{\text{def.}}{=}\sup_{x_1,x_2\in X}\|x_1-x_2\|$ and its "fractal dimension" as quantified by its *metric capacity*

$$\operatorname{cap}(X)\overset{\text{def.}}{=}\sup\left\{n\in\mathbb{N}_+:\,(\exists x_1,\dots,x_n\in X),(\exists r>0)\sqcup_{i=1}^N B_2(x_i,r/5)\subset B_2(x_0,r)\right\},$$

where $\sqcup$ denotes the union of *disjoint* subsets of $\mathbb{R}^d$ and where $B_2(x,r)\overset{\text{def.}}{=}\{u\in\mathbb{R}^d:\,\|u-x\|<r\}$. We mention that, for a compact Riemannian manifold, the $\log_2$-metric capacity is always a multiple of the manifold's topological dimension and the $\log_2$-metric capacity of a d-dimensional cube in $\mathbb{R}^d$ is proportional to d; (see (Acciaio et al., 2022, 2.1.3) for further details). We denote the set of polynomial functions from $\mathbb{R}^d$ to $\mathbb{R}^D$ by $\mathbb{R}[x_1,\dots,x_d:D]$.

Theorem 2 (Support Detection and Uniform + τ Approximation of ReLU Networks with Pooling)**.** *Let $f:\mathbb{R}^d\to\mathbb{R}^D$ be Lipschitz and compactly-supported and $\log_2(d)\in\mathbb{N}_+$. For every "width parameter" $N\in\mathbb{N}_+$ and every sequence $\{\varepsilon_n\}_{n=1}^\infty$ in $(0,\infty)$ converging to 0, there is a sequence $\{\hat{f}^{(n)}\}_{n=1}^\infty$ in $\mathrm{NN}^{\mathrm{ReLU+Pool}}$ satisfying:*

(i) ***Quantitative Worst-Case Approximation:*** *for each $n\in\mathbb{N}_+$ $\max_{x\in[n_f,n_f]^d}\left\|f(x)-\hat{f}^{(n)}(x)\right\|\leq\varepsilon_n$,*

(ii) ***Convergence in Separating Topology τ:*** *$\{\mathrm{Pool}\circ\hat{f}^{(n)}\}_{n=1}^\infty$ converges to f in the separating topology τ,*

(iii) ***Support Identification****: $\operatorname{ess-supp}(\hat{f}^{(n)})\subseteq\left[-\sqrt[d]{2^{-d}\varepsilon_n+n_f^d},\sqrt[d]{2^{-d}\varepsilon_n+n_f^d}\right]^d$, where n_f is defined by $n_f\overset{\text{def.}}{=}\min\{n\in\mathbb{N}_+:\,\operatorname{ess-supp}(f)\subseteq[-n,n]^d\}$.*

Moreover, each $\hat{f}^{(n)}$ is specified by:

(iv) ***Width****: $\hat{f}^{(n)}$ has width $C_3+C_4\max\{d\lfloor N^{1/d}\rfloor,N+1\}$,*

(v) ***Depth****: $\hat{f}^{(n)}$ has depth $\frac{\varepsilon_n^{-d/2}}{N\log_3(N+2)^{1/2}}\Big(\log_2(\operatorname{cap}(\operatorname{ess-supp}(f)))\operatorname{diam}(\operatorname{ess-supp}(f))\operatorname{Lip}(f)\Big)^dC_1+C_2$*

(vi) ***Number of bilinear pooling layers:*** *$\hat{f}^{(n)}$ uses $\log_2(d)+1$ bilinear pooling layers.*

where the dimensional constants are $C_1\overset{\text{def.}}{=}c\,2^dD^{3/d}d^d+3d$, $C_2\overset{\text{def.}}{=}+2d+2$, $C_3\overset{\text{def.}}{=}\max\{d(d-1)+2,D\}$, $C_4\overset{\text{def.}}{=}d(D+1)+3^{d+3}$, and where $c>0$ is an absolute constant independent of X,d, D, and f.

Moreover, if f is not identically 0 then there is some sequence $(\varepsilon_n)_{n=1}^\infty$ in $(0,\infty)$ converging to 0 for which no $\hat{f}\in\mathrm{NN}^{\omega+\mathrm{Pool}}\cup\mathbb{R}[x_1,\dots,x_d:D]$ satisfies (i)-(iii) simultaneously.

Omitting constants, the depth of the ReLU networks $\hat{f}^{(n)}$ with pooling in Theorem 2 encodes three factors. The first is the desired approximation quality, with more depth translating to better approximation capacity, and the second is the target function's regularity; both these factors are present in most available quantitative approximation theorems (Yarotsky, 2017b; Gühring et al., 2020a; Jiao et al., 2021; Lu et al., 2021; Shen et al., 2022; Opschoor et al., 2022).

$$\operatorname{Depth}(f^{(n)})\approx\underbrace{\frac{\varepsilon_n^{-d/2}}{N\log_3(N+2)^{1/2}}}_{\text{Approximation Quality}}\quad\underbrace{\Big(\operatorname{Lip}(f)\Big)^d}_{\text{Target's Regularity}}\quad\underbrace{\Big(\log_2(\operatorname{cap}(\operatorname{ess-supp}(f)))\operatorname{diam}(\operatorname{ess-supp}(f))\Big)^d}_{\text{Complexity: Target's Essential Support}}$$

Part of the novelty of Theorem 2 is that it identifies a third impacting the approximation quality of a ReLU network with pooling; namely, the complexity of the target function's support. This third factor can be decomposed into two parts, the diameter of the target function's essential support, which other approximation theorems have also considered Siegel & Xu (2020); Kratsios & Papon (2021), but what is most interesting here is the effect of the *fractal dimension* (via the metric capacity; see Bruè et al. (2021) for details) of the target function's essential support.

In particular, the result shows that functions essentially supported on low-dimensional sets (e.g. low-dimensional latent manifolds) must be simpler to approximate than those with full support. Theorem 1 (i) and variant of (Shen et al., 2022, Theorem 1.1) and of the main result of Yarotsky (2017b) where the approximation has *controlled support* made to match that of the target function. To the best of the authors' knowledge, the result is also the only quantitative universal approximation which encodes the target function f's complexity in terms of its Lipschitz regularity, as well as, the size and dimension of its essential support. We note that, since one can show that τ is not a metric topology, therefore, a *quantitative* counterpart of Theorem 1 does not exist for arbitrary $f \in L^1_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ for the separating topology τ.

Comparison Between Networks in $\mathrm{NN}^{\mathrm{ReLU}+\mathrm{Pool}}$ and Polynomial Regressors
More broadly, we may ask if the phenomenon of Theorem 1 (ii) is true for other analytic machine learning models. In particular, we compare the models in $\mathrm{NN}^{\mathrm{ReLU}+\mathrm{Pool}}$ against classical polynomial regressors, whose universal approximation capabilities are guaranteed by the classical Stone-Weierstrass theorem and its numerous contemporary variants Prolla (1994), Timofte et al. (2018), or of Galindo & Sanchis (2004)). This is because, the set of multivariate polynomial functions from $\mathbb{R}^d$ to $\mathbb{R}^D$, which we denoted by $\mathbb{R}[x_1, \dots, x_d] \overset{\text{def.}}{=} \left\{\sum_{k=0}^K \prod_{i=1}^d \beta_{n,i} x_i^{n_{i,k}} : n_{1,0}, \dots, n_{d,K} \in \mathbb{N}, \beta_{n,i} \in \mathbb{R}, \right\}$, also fails to be dense in $L^1_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ for τ. Therefore, the empirically-observed advantage which deep ReLU networks have over polynomial regression methods and the quantitative edge exhibited over polynomial methods in terms of uniform approximation efficiency Gühring et al. (2020b); Suzuki (2019); Yarotsky & Zhevnerchuk (2020a), carry over to a qualitative gap between the two methods; in the sense of Theorem 1 and the following result.

Theorem 3 (Separation: Deep Networks with Piecewise-Linear Activation Functions vs. Polynomial-Regressors)**.** *The set* $\mathbb{R}[x_1, \dots, x_d]$ *is not dense in* $L^1_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ *for* τ.

Thus, Theorem 1 (ii) and Theorem 3 show that the polynomial regressors model and deep feedforward network with sigmoid activation function cannot approximate a locally integrable functions in the separating topology τ. In contrast, Theorems 1 (i) and 2 guarantees that networks in $\mathrm{NN}^{\mathrm{ReLU}+\mathrm{Pool}}$ can approximate any function in $L^1_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ for the separating topology τ, and Theorem 1 (iii) demonstrates that these results are non-trivial.

1.1 Connection to Other Deep Learning Literature

Our results are perhaps most closely related to Park et al. (2021) which demonstrates, to the best of our knowledge, the only other qualitative gap in the deep learning theory. Namely, therein, the authors identify a minimum width under which all networks become too narrow to approximate any integrable function; equivalently, the set of "very narrow" deep feedforward networks is qualitatively less expressive than the set of "arbitrary deep feedforward networks". Just as our main results are qualitative, the results of Park et al. (2021) can be contrasted against the main result of Shen et al. (2022) which quantifies the exact impacts of depth and width on approximation error of deep feedforward networks.

Our results also add to the recent scrutiny given to deep feedforward networks deploying several activation functions (Jiao et al., 2021; Yarotsky & Zhevnerchuk, 2020b; Beknazaryan, 2021; Yarotsky, 2021; Acciaio et al., 2022). The connection to this branch of deep learning theory happens on two distinct fronts. First $\mathrm{NN}^{\omega+\mathrm{Pool}}$ is clearly a family of deep feedforward networks simultaneously utilizing several activation functions. However, more interesting, is the second connection between networks in $\mathrm{NN}^{\mathrm{ReLU}+\mathrm{Pool}}$ and the approximation theory of deep feedforward networks with generalized ReLU activation function $\mathrm{ReLU}_r(x) \overset{\text{def.}}{=} \max\{x, 0\}^r$, where $r \in \mathbb{R}$ is a *trainable parameter*. This is because, Pool can be implemented by a feedforward network with ReLU_2 activation function, since $x^2 = \mathrm{ReLU}_2(x) + \mathrm{ReLU}_2(-x)$ (where $x \in \mathbb{R}$) and (Kidger & Lyons, 2020, Lemma 4.3) shows that the multiplication map $\mathbb{R}^2 \ni (x_1, x_2) \mapsto x_1 x_2$ can be exactly implemented by a neural network with one hidden layer and with activation function $x \mapsto x^2$. Therefore, any $f \in \mathrm{NN}^{\mathrm{ReLU}+\mathrm{Pool}}$ there are $f_1, \dots, f_I \in \mathrm{NN}^{\mathrm{ReLU}_2} \cup \mathrm{NN}^{\mathrm{ReLU}}$ representing f via

$$f = f_I \circ \cdots \circ f_1.$$

We note that networks with activation function in $\{\mathrm{ReLU}_r\}_{r \in \mathbb{R}}$ have recently rigorous study in Gribonval Rémi et al. (2021) and are related to the the constructive approximation theory of splines where ReLU_r are known as *truncated* *powers* (see (DeVore & Lorentz, 1993, Chapter 5, Equation (1.1))). We also mention that Theorem 2 is related to recent deep learning research considering the approximation of a function or probability measure's support. The former case is considered by Kratsios & Zamanlooy (2022), where the authors consider an exotic neural network architecture specialized in the approximation of piecewise continuous functions in a certain sense. In the latter case, Puthawala et al. (2022) use a GAN-like architecture to approximate probability distributions supported on a low-dimensional manifold by approximating their manifold and the density thereon using a specific neural network architecture. In contrast, our results compare the approximation capabilities of feedforward networks built using different activation functions.

Organization of Paper

This paper is organized as follows. Section 2 reviews the necessary deep learning terminology, measure theoretic, and topological background needed in the formulation of our main result. Section 3 is devoted to the construction of the "separating topology" τ, the examination of its properties so as to ablate the meaning of the qualitative gap in Theorem 1, and then our main result is formally stated. The proofs of all supporting and technical lemmas are relegated to the paper's appendix.

2 Preliminaries

We use $\mathbb{N}_+$ to denote the set of positive integers, fix $d, D \in \mathbb{N}_+$, and let $\|\cdot\|$ denote Euclidean distance on $\mathbb{R}^D$.

To simplify the analysis, we emphasize that d will *always be assumed to be a power of* 2*; i.e.* $d = 2^{d'}$ *where* $d' \in \mathbb{N}_+$.

2.1 Deep Feedforward Networks

Originally introduced by McCulloch & Pitts (1943) as a prototypical model for artificial neural computation, deep feedforward networks have since lead to computational breakthroughs across various areas from biomedical imaging Ronneberger et al. (2015) to quantitative finance Buehler et al. (2019); Jaimungal (2022). Though deep learning tools has become pedestrian in most contemporary scientific computational endeavors, the mathematical foundations of deep learning are still in their early stages.

Therefore, in this paper, we study the approximation-theoretic properties of what is arguably the most basic deep learning model; namely, the *feedforward (neural) network*. These are models which iteratively process inputs in $\mathbb{R}^d$ by repeatedly applying affine transformations (as in linear regression) and simple component-wise non-linearity called *activation functions*, until an output in $\mathbb{R}^D$ is eventually produced.

Our discussion naturally begins with the formal definition of the class of deep feedforward neural networks defined by a (non-empty) *family of (continuous) activation functions* $\Sigma \subseteq C(\mathbb{R})$. In the case where $\Sigma = \{\sigma\}$ is a singleton, one recovers the classical definition of a feedforward network studied in Cybenko (1989); Hornik et al. (1989); Leshno et al. (1993); Yarotsky (2017b); Kidger & Lyons (2020) and when $\Sigma = \{\sigma_r\}_{r\in\mathbb{R}}$ and the map $(r,x) \mapsto \sigma_r(x)$ is Lebesgue a.e. differentiable then one obtains so-called *trainable activation functions* as considered in Cheridito et al. (2021a); Kratsios et al. (2022); Acciaio et al. (2022) of which the $\mathrm{PReLU}_r(x) \overset{\text{def.}}{=} \max\{x, rx\}$ activation function of He et al. (2015) is prototypical. More broadly, neural networks build using families of activation functions Σ exhibiting sub-exponential approximation rates have also recently become increasingly well-studied; e.g. Yarotsky & Zhevnerchuk (2020b); Jiao et al. (2021); Yarotsky (2021); Beknazaryan (2021).

Consider the *bilinear pooling layer*, from computer vision (Lin et al., 2015; Kim et al., 2016; Fang et al., 2019), given for any *even* $n \in \mathbb{N}_+$ and $x \in \mathbb{R}^n$ as

$$\mathrm{Pool}(x) \overset{\text{def.}}{=} \left(x_i x_{n/2+i}\right)_{i=1}^{n/2}.$$

Alternatively, Pool can be thought of as a *masking layer* with non-binary values, similar to the bilinear masking layers or bilinear attention layers used in the computer-vision literature Fang et al. (2019); Lin et al. (2015) or in the low-rank learning literature Kim et al. (2016), or as the *Hadamard product* of the first $n/2$ components of a vector in $\mathbb{R}^n$ with the last n components.

Fix a *depth* $J, d, D \in \mathbb{N}_+$. A function $\hat{f}: \mathbb{R}^d \to \mathbb{R}^D$ is said to be a *deep feedforward network with (bilinear) pooling* if for every $j = 0, \dots, J-1$ there are Boolean *pooling parameters* $\alpha^{(j)} \in \{0,1\}$, $d_{j,2} \times d_{j,1}$-dimensional matrices $A^{(j)}$ with $d_{j+1,1}/2 = d_{j,2}$ if $d_{j,2}$ is even and if $\alpha = 1$ and $d_{j+1,1} = d_{j,2}$ otherwise which are called *weights*, $b^{(j)} \in \mathbb{R}^{d_j}$ and a $c \in \mathbb{R}^{d_J}$ called *biases*, and *activation functions* $\sigma^{(j,i)} \in \Sigma$ such that $\hat{f}$ admits the iterative representation

$$\begin{aligned} \hat{f}(x) &\overset{\text{def.}}{=} x^{(J)} + c && \\ x^{(j+1)} &\overset{\text{def.}}{=} \begin{cases} \operatorname{Pool}(\tilde{x}^{(j+1)}) & : \alpha^{(j)} = 1 \text{ and } d_{j+1} \text{ is even} \\ \tilde{x}^{(j+1)} & : \text{else} \end{cases} && \text{for } j = 0, \dots, J-1 \\ \tilde{x}_i^{(j+1)} &\overset{\text{def.}}{=} \sigma^{(j,i)}((A^{(j)} x^{(j)} + b^{(j)})_i) && \text{for } j = 0, \dots, J-1; i = 1, \dots, d_{j+1} \\ x^{(0)} &\overset{\text{def.}}{=} x. && \end{aligned} \tag{2}$$

We denote by $\mathrm{NN}^{\Sigma+\mathrm{Pool}}$ the set of all deep feedforward networks with pooling and activation functions belonging to Σ. If, in the above notation, $\hat{f}$ is such that $x^{(j+1)} = \tilde{x}^{(j+1)}$ then, we say that $\hat{f}$ is a *deep feedforward network (without pooling)*. The collection of all deep feedforward networks (without pooling) is denoted by NN^{Σ} and activation functions belonging to Σ.

In either case, if Σ consists only of a single activation function σ then, we use $\mathrm{NN}^{\Sigma+\mathrm{Pool}}$ to denote $\mathrm{NN}^{\sigma+\mathrm{Pool}}$. Similarly, if $\Sigma = \{\sigma\}$ then we set $\mathrm{NN}^{\sigma} \overset{\text{def.}}{=} \mathrm{NN}^{\Sigma}$. Let us consider some examples of activation functions.

Example 1 (Piecewise Linear Networks with at-least two distinct pieces)**.** *An activation function $\sigma \in C(\mathbb{R})$ is called piecewise linear with at-least 2 distinct pieces if: there exist $-\infty = t_0 < t_1 < \cdots < t_p < t_{p+1} = \infty$ and some $m_1, \dots, m_p, b_1, \dots, b_p \in \mathbb{R}$ for which*

(i) $\sigma(x) = m_i x + b_i$ for every $t \in (t_i, t_{i+1})$ for each $i = 0, \dots, p$,

(ii) There exist some $i \in \{1, \dots, p\}$ for which $\sigma'(t_i)$ is undefined.

The prototypical example of such an activation function is $\operatorname{ReLU}(x) \overset{\text{def.}}{=} \max\{0, x\}$.

Example 2 (Deep Feedforward Networks with “Adaptive” Analytic Activation Functions (NN^{ω}))**.** *Let $C^{\omega}(\mathbb{R})$ denote the set of a analytic maps from $\mathbb{R}$ to itself. We set $\mathrm{NN}^{\omega} \overset{\text{def.}}{=} \mathrm{NN}^{C^{\omega}(\mathbb{R})}$ and we use $\mathrm{NN}^{\omega+\mathrm{Pool}} \overset{\text{def.}}{=} \mathrm{NN}^{C^{\omega}(\mathbb{R})+\mathrm{Pool}}$*

Remark 1 (Scope of $\mathrm{NN}^{\omega+\mathrm{Pool}}$)**.** *The class NN^{ω} is rather broad and it contains all “classical” feedforward networks with the following common activation functions: the classical* tanh, *logistic, and sigmoid $\sigma_{\mathrm{sigmoid}}(x) \overset{\text{def.}}{=} \frac{e^x}{1+e^x}$ activation functions, the GeLU activation function $\sigma_{\mathrm{GeLU}}(x) \overset{\text{def.}}{=} \frac{1}{2}x(1 + \operatorname{erf}(\frac{x}{2}))$ of Hendrycks & Gimpel (2016), $\sigma_{\mathrm{Softplus}}(x) \overset{\text{def.}}{=} \ln(1 + e^x)$ of Glorot et al. (2011), $\sigma_{\mathrm{Swish}-1}(x) \overset{\text{def.}}{=} \frac{x}{1+e^{-x}}$ of Ramachandran et al. (2018), any polynomial activation function (as used in neural ODEs Cuchiero et al. (2020) literature).*

2.2 Measure Theory

Following (Schwartz, 1966, Chapter 1), we call Borel measurable function $f: \mathbb{R}^d \to \mathbb{R}^D$ is called *locally integrable* if, on each compact subset $K \subset \mathbb{R}^d$ the Lebesgue integral $\int_{x \in K} \|f(x)\| \, dx$ is finite. Let $L^1_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ denote the set of locally integrable functions from $\mathbb{R}^d$ to $\mathbb{R}^D$; with equivalence relation $f \sim g$ if and only if f and g differ only on a set of Lebesgue measure 0. The set $L^1_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ is made into a *complete metric space* by equipping it with the distance function $d_{L^1_{\mathrm{loc}}}$ defined on any two $f, g \in L^1_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ by

$$d_{L^1_{loc}}(f, g) \overset{\text{def.}}{=} \sum_{n=1}^{\infty} \frac{1}{2^n} \frac{\int_{\|x\| \le n} \|(f(x) - g(x))\| \, dx}{1 + \int_{\|x\| \le n} \|(f(x) - g(x))\| \, dx}.$$

The subset of $L^1_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ consisting of all *integrable “functions”*, i.e. all $f \in L^1_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ for which the integral $\int_{x \in \mathbb{R}^d} \|f(x)\| \, dx$ is finite, is denoted by $L^1(\mathbb{R}^d, \mathbb{R}^D)$. The set $L^1(\mathbb{R}^d, \mathbb{R}^D)$ is made into a Banach space, called the *Bochner-Lebesgue* space, by equipping it with the norm $\|f\|_{L^1} \overset{\text{def.}}{=} \int_{x \in \mathbb{R}^d} \|f(x)\| \, dx$.

2.3 Point-Set Topology

In most of analysis one uses the language of *metric spaces*, i.e.: an (abstract) set of points X together with a distance function $d : X^2 \to [0,\infty)$ satisfying certain axioms (see (Heinonen, 2001)), to the similarity of dissimilarity between different mathematical objects. However, not all notions of similarity can be described by a metric structure and this is in particular true for several very finer notions of similarity playing central roles in functional analysis (see Narayanaswami & Saxon (1986)).

In such situations, one instead turns to the notion of a *topology* to qualify closeness of two objects without relying on the quantitative notion of distance defined though by a metric. Briefly, a topology τ_X on a set X is a collection of subsets of X declared as being "open"; we require only that τ_X satisfy certain axioms reminiscent of the familiar open neighborhoods build using balls in metric space theory. Namely, τ_X contains the empty set and the "total" set X, the union of elements in τ_X are again a member of τ_X, and the countable intersection of sets in τ_X are again a set in τ. A *topological space* is a pair (X,τ_X) of a set X and a topology τ_X on X. If clear from the context, we denote (X,τ_X) by X.

Example 3 (Metric Topology on $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$)**.** *The metric topology on $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$, which exists, is the smallest topology on $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ containing all the open balls*

$$B_{L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)}(f,\varepsilon) \overset{\text{def.}}{=} \left\{g \in L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D) : d_{L^1_{\text{loc}}}(f,g) < \varepsilon\right\},$$

where $f \in L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ and $\varepsilon > 0$. We denote this topology by τ_{loc}.

A topology on the subset $L^1(\mathbb{R}^d,\mathbb{R}^D)$ of $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ can always be defined by restricting τ_{loc} as follows.

Example 4 (Subspace Topology on $L^1(\mathbb{R}^d,\mathbb{R}^D)$)**.** *The subspace topology on $L^1(\mathbb{R}^d,\mathbb{R}^D)$, relative to the metric topology on $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$, is the collection $\{U \cap L^1(\mathbb{R}^d,\mathbb{R}^D) : U \in \tau_{\text{loc}}\}$.*

A topology τ'_X on X is said to be strictly *stronger* than another topology τ_X on X if $\tau_X \subset \tau'_X$. The key relation between $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ and $L^1(\mathbb{R}^d,\mathbb{R}^D)$ is that even if former is strictly larger as a set, the topology on the latter induced by the norm $\|\cdot\|_{L^1}$ is strictly stronger than τ_{loc}.

The norm topology on $L^1(\mathbb{R}^d,\mathbb{R}^D)$ is defined as follows.

Example 5 (Norm Topology on $L^1(\mathbb{R}^d,\mathbb{R}^D)$)**.** *The norm topology on $L^1(\mathbb{R}^d,\mathbb{R}^D)$, which exists, is the smallest topology on $L^1(\mathbb{R}^d,\mathbb{R}^D)$ which contains all the open balls*

$$B_{L^1(\mathbb{R}^d,\mathbb{R}^D)}(f,\varepsilon) \overset{\text{def.}}{=} \left\{g \in L^1(\mathbb{R}^d,\mathbb{R}^D) : \|f-g\|_{L^1} < \varepsilon\right\},$$

where $f \in L^1(\mathbb{R}^d,\mathbb{R}^D)$ and $\varepsilon > 0$. We denote this topology by τ_{norm}.

The qualitative statement being put forth by a *universal approximation theorem* (e.g. Leshno et al. (1993); Petrushev (1999); Yarotsky (2017a); Suzuki (2019); Grigoryeva & Ortega (2019); Heinecke et al. (2020); Kidger & Lyons (2020); Zhou (2020); Kratsios & Bilokopytov (2020); Siegel & Xu (2020); Kratsios & Hyndman (2021); Kratsios et al. (2022); Yarotsky (2022)) is a statement about the topological genericness of a machine learning model, such as a neural network model, in specific sets topological "function" spaces. Topological genericness is called *denseness*, and we say that a subset $F \subseteq X$ is dense with respect to a topology τ_X on X if: for every non-empty open subset $U \in \tau_X$ there exists an element $f \in F$ which also belongs to U.

Related is the notion of *convergence* of a sequence in a general topological space X. Let I be a set with a preorder $\preccurlyeq$ (i.e. for every $i,j,k \in I$ $i \preccurlyeq i$ and if $i \preccurlyeq j$ and $j \preccurlyeq k$ then $i \preccurlyeq k$), such that every finite subset of I has an upper-bound with respect to $\preccurlyeq$. A typical example of a directed set is $\mathbb{N}$ equipped with the preorder given by $\leq$. A *net* in a topological space X is a map from a directed set I to X; we denote nets by $(x_i)_{I\in I}$. A typical example of a net is a sequence; in which case the directed set I is the natural numbers with pre-order $\leq$. The next $(x_i)_{I\in I}$ is said to *converge* to an element x of X with respect to the topology τ_X if: for every $U \in \tau_X$ containing x, there exists some $i_U \in I$ such that for every $i \in I$ if $i_U \preccurlyeq I$ then $x_i \in U$.

2.4 Limit-Banach Spaces (LB-Spaces)

Our construction will exploit a special class of *topological vector spaces*, i.e. vector spaces wherein addition and scalar multiplication are continuous operators, formed by inductively *gluing* together ascending sequences of Banach spaces. Specifically, a topological vector space X is a *limit-Banach* space, nearly always referred to as an LB-space in the literature, if first, one can exhibit sequence of strictly nested Banach spaces $\{X_n\}_{n=1}^{\infty}$ (i.e. each X_n is a proper subspace of X_{n+1}) such that

$$X = \cup_{n=1}^{\infty} X_n.$$

Then, the topology on X must be *smallest* topology containing every convex subset $B \subseteq X$ for which $kb \in B$ whenever $k \in [-1,1]$ and $b \in B$, and for every positive integer n, $0 \in B \cap X_n$ and $B \cap X_n$ is an open subset of X_n.

Conversely, given a sequence of strictly nested Banach spaces $\{X_n\}_{n=1}^{\infty}$ one can always form an "optimal" LB-space as follows. Define $X \overset{\text{def.}}{=} \cup_{n=1}^{\infty} X_n$ and equip X with the finest topology making X into an LB-space and such that, for every $n \in \mathbb{N}_+$, the inclusion $X_n \subseteq X$ is continuous. Indeed, as discussed in (Osborne, 2014, Section 3.8), such a topology always exists[1]. We will henceforth refer to X as the *LB-space glued together from* $\{X_n\}_{n=1}^{\infty}$.

A classical example of an LB-space arises when one wants to analyse polynomial functions but does not want to take their closure in some larger space (e.g. a larger space containing power series). We now present this example.

Example 6 (Polynomial Functions)**.** *For every $n \in \mathbb{N}_+$, the set of degree at-most n polynomial functions mapping $\mathbb{R}$ to $\mathbb{R}$ is $\mathbb{R}_n[X] \overset{\text{def.}}{=} \{p(x) = \sum_{i=0}^{n} \beta_i x^i \ \beta \in \mathbb{R}^{n+1}\}$. We make $\mathbb{R}_n[X]$ into a Banach space through its identification the coefficients of polynomials in $\mathbb{R}_n[X]$ with $\mathbb{R}^{n+1}$; i.e. for any polynomial $p(x) = \sum_{i=0}^{n} \beta_i x^i \in \mathbb{R}_n[X]$ we define $\|p\|_n$ by*

$$\|p\|_n \overset{\text{def.}}{=} \left(\beta_0^2 + \cdots + \beta_n^2\right)^{1/2}. \tag{3}$$

Thus, we may consider the $\mathbb{R}[X] \overset{\text{def.}}{=} \cup_{n=1}^{\infty} \mathbb{R}_n[X]$ to be the LB-space glued together from $\{\mathbb{R}_n[X]\}_{n=1}^{\infty}$ and $\mathbb{R}[X]$ consists precisely of all polynomial functions from $\mathbb{R}$ to $\mathbb{R}$ of any degree.

To illustrate the "optimality" of our LB-space, let us compare $\mathbb{R}[X]$ with the smallest Banach space containing every $\{\mathbb{R}_n[X]\}_{n=1}^{\infty}$ as a subspace. Notice that for every positive integer n, $\mathbb{R}_n[X]$ is a subspace of the following Hilbert space of (formal) power-series $\mathbb{R}[[X]] \overset{\text{def.}}{=} \{f(x) = \sum_{i=0}^{\infty} \beta_i x^i : \sum_{i=0}^{\infty} \beta_i^2 < \infty\}$ mapping $\mathbb{R}$ to $[-\infty, \infty]$ and normed by

$$\|f\|_\infty \overset{\text{def.}}{=} \left(\sum_{i=0}^{\infty} \beta_i^2\right)^{1/2}.$$

By construction $\mathbb{R}[X]$ does not contain any function of the form $f(x) = \sum_{i=1}^{\infty} \beta_i x^i$ where an infinite number of β_i are equal to zero while $\mathbb{R}[[X]]$ does contain such functions; e.g. $f(x) \overset{\text{def.}}{=} \sum_{i=0}^{\infty} \frac{x^i}{2^i}$ belongs to $\mathbb{R}[[X]]$ but not to $\mathbb{R}[X]$. In this way, the topological vector space $\mathbb{R}[X]$ is smaller than $\mathbb{R}[[X]]$ precisely because its topology is stronger.

Let us illustrate the topology on the LB-space $\mathbb{R}[X]$. By (Osborne, 2014, Proposition 3.40) we know that a convex subset $U \subseteq \mathbb{R}[X]$ is open if and only if $U \cap \mathbb{R}_n[X]$ is open for the topology on $\mathbb{R}_n[X]$ defined by the norm in equation 3.

Example 6 illustrates the intuition behind *LB-spaces glued together from* $\{X_n\}_{n=1}^{\infty}$; namely, these spaces are "minimal limits" of sequences Banach spaces which contain no new element not already present in the Banach spaces $\{X_n\}_{n=1}^{\infty}$.

3 The Separating Topology τ

We now construct the separating topology τ of Theorem 1 the set $L^1_{\mu,\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$, in three steps. However, before beginning our construction, we fix an arbitrary "good a.e. partition" of $\mathbb{R}^d$. As we will see shortly, the construction of the separating topology τ is independent of the choice of "good a.e. partition" of $\mathbb{R}^d$; and thus, the construction is natural (in the precise algebraic sense describe in Proposition 1, below). However, to establish this surprising algebraic property of the separating topology τ, it is more convenient to describe the construction (for any arbitrary choice of $\{K_n\}_{n=1}^{\infty}$) once and for all.

---

[1] In the language of category theory, X is the colimit of the inductive system $(\{X_n\}_{n=1}^{\infty}, \subseteq)$ in the category of locally-convex topological vector spaces with bounded linear maps as morphisms.

Definition 1 (Good a.e. partition of $\mathbb{R}^d$). *A collection* $\{K_n\}_{n=1}^\infty$ *of compact subsets of* $\mathbb{R}^d$ *is called a good a.e. partition if it satisfies the following conditions:*

(i) The set $\mathbb{R}^d - \cup_{n=1}^\infty K_n$ *has Lebesgue measure* 0*,*

(ii) For every $n \in \mathbb{N}_+$*,* K_n *has positive Lebesgue measure,*

(iii) For each $n, m \in \mathbb{N}_+$*, if* $n \neq m$ *then* $K_n \cap K_m$ *has Lebesgue measure* 0.

For instance, since our construction will be shown to be *independent of our choice* of a good a.e. partition of $\mathbb{R}^d$ made when constructing τ. Once we show this, we may, without loss of generality, henceforth only consider the following partition of $\mathbb{R}^d$. This partition is illustrated in Figure 2.

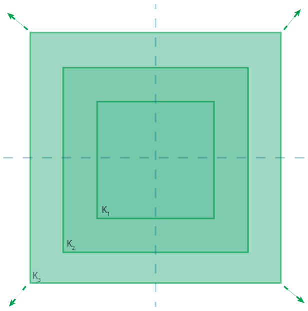

Figure 2: Cubic-Annuli

Example 7 (Good a.e. partition into Cubic Annuli). *For each* $n \in \mathbb{N}_+$ *set* $K_n \overset{\text{def.}}{=} \{x \in \mathbb{R}^d : n < \|x\|_\infty \leq n+1\}$*, where* $\|x\|_\infty \overset{\text{def.}}{=} \max_{i=1,\dots,n} |x_i|$*. Then* $\{K_n\}_{n=1}^\infty$ *is a good a.e. partition of* $\mathbb{R}^d$.

Let us construct the separating topology τ, using a fixed good a.e. partition of $\mathbb{R}^d$ in three steps.
Step 1: Given $\{K_n\}_{n=1}^\infty$ a good a.e. partition of $\mathbb{R}^d$ define the strictly nested sequence of Banach subspaces of $L^1(\mathbb{R}^d, \mathbb{R}^D)$ as follows. For every $n \in \mathbb{N}_+$ let $L^1_n(\mathbb{R}^d, \mathbb{R}^D)$ consist of all $f \in L^1(\mathbb{R}^d, \mathbb{R}^D)$ with ess-supp$(f) \subseteq \cup_{i=1}^n K_i$.
Step 2: The spaces $\{L^1_n(\mathbb{R}^d, \mathbb{R}^D)\}_{n=1}^\infty$ are aggregated into one LB-space, denoted by $L^1_c(\mathbb{R}^d, \mathbb{R}^D)$, whose underlying set is $\bigcup_{n \in \mathbb{N}^+} L^1_n(\mathbb{R}^d, \mathbb{R}^D)$ and equipped with the *finest topology* ensuring that the inclusions $L^1_n(\mathbb{R}^d, \mathbb{R}^D) \subseteq L^1_c(\mathbb{R}^d, \mathbb{R}^D)$ remain continuous.

Remark 2 (Notation and Independence of Choice of Good a.e. Partition of $\mathbb{R}^d$). *The notation* $L^1_c(\mathbb{R}^d, \mathbb{R}^D)$ *does not make any reference to our choice of a good a.e. partition of* $\mathbb{R}^d$ *used to define the space* $L^1_c(\mathbb{R}^d, \mathbb{R}^D)$*. This is because, as we will shortly see in Proposition 1 below, the topology on* $L^1_c(\mathbb{R}^d, \mathbb{R}^D)$ *is independent of our choice of a good a.e. partition of* $\mathbb{R}^d$ *used to define it. However, to formally state that result; we will make use of the notation* $L^1_c(\{K_n\}_{n=1}^\infty, \mathbb{R}^D)$ *emphasizing our choice of* $\{K_n\}_{n=1}^\infty$ *which is a good a.e. partition of* $\mathbb{R}^d$ *used in Steps* 1 *and* 2.

Step 3: Since $L^1_c(\mathbb{R}^d, \mathbb{R}^D)$ does not contain every function in $L^1_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ then, intuitively speaking, we "glue" remaining locally-integrable functions to $L^1_c(\mathbb{R}^d, \mathbb{R}^D)$ by aggregating the topologies on $L^1(\mathbb{R}^d, \mathbb{R}^D)$ and on $L^1_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ to $L^1_c(\mathbb{R}^d, \mathbb{R}^D)$. Rigorously, we define this gluing as follows.

Definition 2 (Separating Topology τ). *The separating topology* τ *on* $L^1_{\mu,loc}(\mathbb{R}^d, \mathbb{R}^D)$ *is smallest*[2] *on* $L^1_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ *containing* $\tau_c \cup \tau_{\text{norm}} \cup \tau_{\text{loc}}$.

Since τ_{norm}, τ_{loc}, and τ_c all exist and since the smallest topology containing a collection of sets[3] must exist (see (Munkres, 2000, page 82)); thus, τ exists. Next, we examine the key properties of τ for our problem. Namely, how it compares to the usual topologies on $L^1_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ and on $L^1(\mathbb{R}^d, \mathbb{R}^D)$, as well as its independence of the choice of good a.e. partition of $\mathbb{R}^d$ used to construct it.

---

[2] I.e. $\tau_c \cup \tau_{\text{norm}} \cup \tau_{\text{loc}}$ is a subbase for the topology τ.

[3] Given a set X and a collection of subsets A of X, the smallest topology τ_A on X containing a A is called the topology generated by A and A is called a subbase of τ_A.

3.1 Properties of the Separating Topology τ

The first indication that the separating topology τ is a natural construction. By which we mean that τ has the somewhat surprising algebraic property that it independent of the good a.e. partition used to build it.

Proposition 1 (The separating topology τ is independent of the choice of good a.e. partition)**.** *Let* $\{K_n\}_{n=1}^{\infty}$ *and* $\{K'_n\}_{n=1}^{\infty}$ *be good a.e. partitions of* $\mathbb{R}^d$*. Then* $L^1_c(\{K_n\}_{n=1}^{\infty}, \mathbb{R}^D) = L^1_c(\{K'_n\}_{n=1}^{\infty}, \mathbb{R}^D)$*. Consequentially,* τ *is independent of the good a.e. partition of* $\mathbb{R}^d$ *used to construct it.*

The significance of Proposition 1 is that it allows us to reduce our entire understanding of the problem, and many of our proofs, to simply considering a single "canonical" good a.e. partition of $\mathbb{R}^d$ which is easy to work with; namely, the Cubic Annuli of Example 7. Briefly, the reason for this is that, given a good a.e. partition of $\mathbb{R}^d$ $\{K_n\}_{n=1}^{\infty}$, the approximation of a compactly supported Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}^D$ in τ requires us to identify the smallest $n \in \mathbb{N}_+$ for which we can identify its support with respect to $\{K_n\}_{n=1}^{\infty}$ which we use to discretize $\mathbb{R}^d$; i.e.

$$\text{ess-supp}(\hat{f}) \subseteq \cup_{i=1}^{n} K_i. \tag{4}$$

Then, we must approximate $\hat{f}$ in the L^1-norm on $\cup_{i=1}^{n+1} K_i$ using our model. The intuitive message of Proposition 1 is that, given any other good a.e. partition of $\mathbb{R}^d$ $\{K'_n\}_{n=1}^{\infty}$, the compactness of $\cup_{i=1}^{n+1} K_i$ implies that there is a smallest $n_1 \in \mathbb{N}_+$ such that $\cup_{i=1}^{n} K_i \subseteq \cup_{j=1}^{n_1} K'_j$ thus, there must be a smallest integer for which equation 4 holds with $\{K'_n\}_{n=1}^{\infty}$ holds in place of $\{K_n\}_{n=1}^{\infty}$. To see the equivalence, arguing similarly, there must exist an (other) $n_2 \in \mathbb{N}_+$ such that $\cup_{j=1}^{n_1} K'_j \subseteq \cup_{i=1}^{n_2} K_i$. Therefore, we may interchangeable identify where the support of a compactly supported Lipschitz function lies using any discretization of $\mathbb{R}^d$ by any choice of good a.e. partition of $\mathbb{R}^d$.

The next result shows that the separating topology τ on $L_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ is strictly finer than the norm metric topology thereon, and its restriction to $L^1(\mathbb{R}^d, \mathbb{R}^D)$ is strictly stronger than the norm topology thereon ((Nagata, 1974, Chapter 2.4)). The approximation-theoretic implication is that fewer members of $L_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ can be approximated by deep learning models in τ than in the other two topologies.

Proposition 2. *The separating topology* τ *is strictly stronger than* τ_{loc}.

The phenomenon of Proposition 2 persists when restricting the separating topology τ to the subset $L^1(\mathbb{R}^d, \mathbb{R}^D)$ of $L^1_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ and comparing it with the norm topology (which is stronger than τ_{loc} restricted to $L^1(\mathbb{R}^d, \mathbb{R}^D)$).

Proposition 3. *The restriction of the separating topology* τ *to* $L^1(\mathbb{R}^d, \mathbb{R}^D)$ *is strictly stronger than the norm topology* τ_{norm} *on* $L^1(\mathbb{R}^d, \mathbb{R}^d)$.

We are now in a position to prove Theorem 1. The next section outlines the main steps in the theorem's derivation, with the details being relegated to our paper's appendix.

4 Outline of the Proof of Theorem 1

To better understand Theorem 1, we overview the principal steps undertaken in its derivation. We begin by establishing the universality of $\text{NN}^{\text{ReLU+Pool}}$ for the separating topology, as guaranteed by Theorem 1 (i). Then, we show the non-universality of $\text{NN}^{\omega+\text{Pool}}$ for the separating topology, given in Theorem 1 (ii). Other curious approximation-theoretic properties of the separating topology are discussed along the way, in order to gain a fuller picture of our main result; such as the failure of the set of polynomial functions to be dense in $L^1_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ for τ.

4.1 Establishing Theorem 1 (i): The universality of $\text{NN}^{\text{ReLU+Pool}}$ in the separating topology

In order to establish Theorem 1 (i), we must first understand how density in $L^1_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$, for the metric topology interacts with density in $L^1_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ for the separating topology. The next lemma accomplishes precisely this, by showing how dense subsets of $L^1_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ for the metric topology can be used to construct dense subsets of $L^1_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ for the separating topology. This construction happens in two phases. First, each "function" in the original dense subset is localized so that it is essentially supported on a part K_n in (any) good a.e. partition $\{K_n\}_{n=1}^{\infty}$

of $\mathbb{R}^d$. Then, each of these localized "functions" are then pieced back together to form a new "function" which is essentially supported on the compact subset $\cup_{i=1}^n K_n$.

Let $\operatorname{Lip}_c(\mathbb{R}^d,\mathbb{R}^D)$ denote the set of "compact support" Lipschitz functions $f:\mathbb{R}^d\to\mathbb{R}^D$; i.e. f is Lipschitz and ess-supp(f) is a compact subset of $\mathbb{R}^d$. The first key observation in the proof of Theorem 1 (i) is that, $\operatorname{Lip}_c(\mathbb{R}^d,\mathbb{R}^D)$ is dense in $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ for the separating topology τ.

Lemma 1 (Density of compactly-supported Lipschitz functions in the separating topology τ)**.** *The set* $\operatorname{Lip}_c(\mathbb{R}^d,\mathbb{R}^D)$ *is dense in* $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ *for the separating topology* τ.

The second key observation, also contained in the next lemma, is a sufficient condition for approximating a "compact support" Lipschitz function with respect to the separating topology τ. Briefly, the approximation of such a function in τ involves the simultaneous approximation of its *outputs* as well as its *essential support*.

Lemma 2 (Approximation of compactly-supported Lipschitz functions in the separating topology τ)**.** *Let* $f\in L^1(\mathbb{R}^d,\mathbb{R}^D)$ *be Lipschitz and* ess-supp(f) *be compact,* $\{K_n\}_{n=1}^\infty$ *be the cubic-annuli of Example 7. If* $\{f_n\}_{n=1}^\infty$ *is a sequence in* $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ *for which there is an* $n_f\in\mathbb{N}_+$ *with*

$$\lim_{n\uparrow\infty}\|f_n-f\|_{L^1(\mathbb{R}^d,\mathbb{R}^D)}=0 \text{ and } \operatorname{ess-supp}(f)\cup\bigcup_{n=1}^\infty \operatorname{ess-supp}(f_n)\subseteq[-n_f-1,n_f+1]^d, \tag{5}$$

then $\{f_n\}_{n=1}^\infty$ *converges to* f *in the separating topology* τ.

Together, Lemmata 2 and 1 provide a sufficient condition for universality with respect to the separating topology. Furthermore the condition is in a sense quantitative. We say in a sense, since the topology τ_c is non-metrizable (see (Narayanaswami & Saxon, 1986, Corollary 3) and consequentially τ is non-metrizable); thus there is no metric describing the approximation of a function in τ. I.e. no genuine quantitative statement is possible[4]

Lemma 3 (Approximation of a compactly essentially-supported functions in the separating topology τ)**.** *Let* $\mathscr{F}\subseteq L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$*. If for every* $f\in\operatorname{Lip}_c(\mathbb{R}^d,\mathbb{R}^D)$ *there exists a sequence* $\{f_n\}_{n=1}^\infty$ *in* $\mathscr{F}$ *satisfying the condition equation 5 then,* $\mathscr{F}$ *is dense in* $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ *for the separating topology* τ.

By Lemma 3, it therefore remains to construct a subset of networks in $\text{NN}^{\text{ReLU+Pool}}$ which can approximate any compactly supported Lipschitz function in the L^1-norm and simultaneously correctly identify its essential support via the cubic annuli partition of $\mathbb{R}^d$. Figure 1 illustrates the main points of the next lemma; namely, if the target function is compactly supported then its output can be closely approximated by a ReLU network which also simultaneously correctly identifiesthe integer n such that the target function is supported in the d-dimensional cube $[-n-1,n+1]^d$.

Accordingly, our next lemma is an extension of the main theorem of Shen et al. (2022), which gives an estimate on the width and depth of the smallest deep ReLU network approximating a Lipschitz map from a compact subset X of $\mathbb{R}^d$ to $\mathbb{R}^D$ (instead of the case where $D=1$ and $X=[0,1]^d$).

Lemma 4 (Uniform approximation of Lipschitz maps on low-dimensional compact subsets of $\mathbb{R}^d$)**.** *Let* $X\subseteq\mathbb{R}^d$ *be non-empty and compact and let* $f:X\to\mathbb{R}^D$ *be Lipschitz. For every "depth parameter"* $L\in\mathbb{N}_+$ *and "width parameter"* $N\in\mathbb{N}_+$ *there exists a* $\hat{f}\in\text{NN}^{\text{ReLU}}$ *satisfying the uniform estimate*

$$\max_{x\in X}\left\|f(x)-\hat{f}(x)\right\|\lesssim\log_2(\operatorname{cap}(X))\operatorname{diam}(\mathrm{X})\ \operatorname{Lip}(f)\ \frac{D^{3/2}d^{1/2}}{N^{2/d}L^{2/d}\log_3(N+2)^{1/d}},$$

where $\lesssim$ *hides an absolute positive constant independent of* X, d, D, *and* f. *Furthermore,* $\hat{f}$ *satisfies*

1. ***Width**:* $\hat{f}$*'s width is at-most* $d(D+1)+3^{d+3}\max\{d\lfloor N^{1/d}\rfloor,N+2\}$
2. ***Depth**:* $\hat{f}$*'s depth is at-most* $D(11L+2d+19)$.

---

[4] Another example of a non-metric universal approximation theorem in the deep learning literature is the universal classification result of (Kratsios & Bilokopytov, 2020, Corollary 3.12)).

In order to apply Lemma 4, we need our approximating model to have support which "matches" the support of the target function $f \in L^1_c(\mathbb{R}^d,\mathbb{R}^D) \overset{\text{def.}}{=} \bigcup_{n\in\mathbb{N}^+} L^1_n(\mathbb{R}^d,\mathbb{R}^D)$ being approximated. The next lemma describes how, given a ReLU network how one can build a new ReLU network with one pooling layer at its output, which coincides with the original network on an arbitrarily cubic-annuli (as in Example 7) and vanishes straightaway outsides the correct number of cubic-annuli $(+1)$.

Lemma 5 (Adjusting a ReLU network to have support on the union of the first $n+1$ cubic annuli)**.** *Let* $\log_2(d) \in \mathbb{N}_+$ *and* $\hat{f} \in \mathrm{NN}^{\mathrm{ReLU}}$ *have depth* $d_{\hat{f}}$ *and width* $w_{\hat{f}}$*. For every* $n \in \mathbb{N}_+$ *and each* $0 < \delta < 1$*, there exists a* $\hat{f}^{\mathrm{pool}} \in \mathrm{NN}^{\mathrm{ReLU+Pool}}$ *with width* $\max\{d(d-1)+2, D\} + w_{\hat{f}}$ *and depth* $2+3d+d_{\hat{f}}$ *satisfying:*

(i) ***Implementation on the cube:*** *For each* $x \in [-n,n]^d$ *it holds that* $\hat{f}(x) = \hat{f}^{\mathrm{pool}}(x)$*,*

(ii) ***Controlled Support:*** $\operatorname{ess-supp}(\hat{f}) \subseteq \left[-\sqrt[d]{2^{-d}\varepsilon+n^d}, \sqrt[d]{2^{-d}\varepsilon+n^d}\right]^d$*,*

(i) ***Control of Error near the Boundary:*** $\left\|\hat{f} - \hat{f}^{\mathrm{pool}}\right\|_{L^1(\mathbb{R}^d,\mathbb{R}^D)} < \varepsilon$*.*

Lemmata 1, 2, and 3 imply that $\mathrm{NN}^{\mathrm{ReLU+Pool}}$ is dense in $L^1_{\mathrm{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ for the separating topology τ only if $\mathrm{NN}^{\mathrm{ReLU+Pool}}$ has a subset which can approximate any essentially compactly-supported Lipschitz function while having almost correct support (as detected by the cubic-annuli partition) as formalized by condition 5. Since Lemma 5 implies that such a subset of networks in $\mathrm{NN}^{\mathrm{ReLU+Pool}}$ exists then, Theorem 1 (i) follows.

Proof of Theorem 1 (i). The result for PW-Lin $=$ ReLU is a direct consequence of Lemmata 4 and 5 applied to Lemma 3. The result for general piecewise linear activation functions with at-least 2 parts follows from the ReLU case by (Yarotsky, 2017b, Proposition 1). This is because (Yarotsky, 2017b, Proposition 1) states that any network in $\mathrm{NN}^{\sigma_{\text{PW-Lin}}}$ can be implemented by a network in $\mathrm{NN}^{\mathrm{ReLU}}$. ☐

We are now equally in a position to prove the first claim in theorem Theorem 2.

Proof of Theorem 2. Since f is compactly essentially-supported, by Lemma 4 there is an $\hat{f}^{\varepsilon_n/2} \in \mathrm{NN}^{\mathrm{ReLU}}$ satisfying

$$\max_{x\in\operatorname{ess-sup}(f)} \left\|f(x) - \hat{f}^{\varepsilon_n/2}(x)\right\| < \frac{\varepsilon_n}{2}, \tag{6}$$

with width $w_{\hat{f}^{\varepsilon_n/2}}$ at-most $d(D+1) + 3^{d+3}\max\{d\lfloor N^{1/d}\rfloor, N+1\}$ and depth $d_{\hat{f}^{\varepsilon_n/2}}$ equal to

$$d_{\hat{f}^{\varepsilon_n/2}} \overset{\text{def.}}{=} \frac{\varepsilon_n^{-d/2}}{N\log_3(N+2)^{1/2}} \Big(2\log_2(\operatorname{cap}(\operatorname{ess-supp}(f)))\operatorname{diam}(\operatorname{ess-supp}(f))\operatorname{Lip}(f)\Big)^d (cD^{3/d}d^d), \tag{7}$$

where $c > 0$ is an absolute constant independent of X, d, D, and f. Set $n_f \overset{\text{def.}}{=} \min\{n \in \mathbb{N}_+ : \operatorname{ess-supp}(f) \subseteq [-n,n]^d\}$ and apply Lemma 5 to $\hat{f}^{\varepsilon_n/2}$ there exists an $\hat{f}^{(n)} \in \mathrm{NN}^{\mathrm{ReLU+Pool}}$ with $\operatorname{ess-supp}(\hat{f}^{(n)}) \subseteq \left[-\sqrt[d]{2^{-d-1}\varepsilon_n+n_f^d}, \sqrt[d]{2^{-d-1}\varepsilon_n+n_f^d}\right]^d$, equal to $\hat{f}^{\varepsilon_n/2}$ on $[-n_f,n_f]^d$ and such that $\left\|\hat{f}^{(n)} - \hat{f}^{\mathrm{pool}}\right\|_{L^1(\mathbb{R}^d,\mathbb{R}^D)} < \frac{\varepsilon_n}{2}$. Therefore, the estimate in equation 6 and implies that

$$\max_{x\in\operatorname{ess-sup}(f)} \left\|f(x) - \hat{f}^{(n)}(x)\right\| \le \max_{x\in\operatorname{ess-sup}(f)} \left\|f(x) - \hat{f}^{(n)}(x)\right\| + \max_{x\in\operatorname{ess-sup}(f)} \left\|\hat{f}^{(n)}(x) - \hat{f}^{\varepsilon_n/2}(x)\right\| \le 2^{-1}\varepsilon_n + 2^{-1}\varepsilon_n = \varepsilon_n.$$

Similarly, equation 6 implies that

$$\|f - \hat{f}^{(n)}\|_{L^1} \le \|f - \hat{f}^{\varepsilon_n/2}\|_{L^1} + \|\hat{f}^{(n)} - \hat{f}^{\varepsilon_n/2}\|_{L^1}$$

and that both $\hat{f}^{(n)}$ and f are essentially-supported in $[-n_f-1,n_f+1]^d$; whence, for each $n\in\mathbb{N}_+$ the condition equation 5 is met. Therefore, Lemma 2 implies that the sequence $\{\hat{f}^{(n)}\}_{n=1}^{\infty}$ in $\mathrm{NN}^{\mathrm{ReLU+Pool}}$ converges to f in the separating topology τ.

It remains to count each of $\hat{f}^{(n)}$'s parameters. By construction, Lemma 5 and the estimate on $w_{\hat{f}\varepsilon/2}$ (below equation 6) imply that $\hat{f}^{(n)}$ has width at-most $\max\{d(d-1)+2,D\}+d(D+1)+3^{d+3}\max\{d\lfloor N^{1/d}\rfloor,N+1\}$. Similarly, Lemma 5 and equation 7 imply that each $\hat{f}^{(n)}$ has depth equal to

$$\frac{\varepsilon_n^{-d/2}}{N\log_3(N+2)^{1/2}}\Big(\log_2(\mathrm{cap}(\text{ess-supp}(f)))\,\mathrm{diam}(\text{ess-supp}(f))\,\mathrm{Lip}(f)\Big)^d(c\,2^dD^{3/d}d^d+3d)+2d+2.$$

Relabeling $C_1\overset{\text{def.}}{=}c\,2^dD^{3/d}d^d+3d$, $C_2\overset{\text{def.}}{=}+2d+2$, $C_3\overset{\text{def.}}{=}\max\{d(d-1)+2,D\}$, $C_4\overset{\text{def.}}{=}d(D+1)+3^{d+3}$, yields the first conclusion. ☐

4.2 Establishing Theorem 1 (ii): The lack of university of $\mathrm{NN}^{\omega+\mathrm{Pool}}$ with respect to the separating topology

The main step in showing that NN^{σ} fails to be dense in $L^1_{\mathrm{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ for the separating topology is the following *necessary condition* for a sequence $\{f_n\}_{n=1}^{\infty}$ in $L^1_{\mathrm{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ to convergence to some essentially compactly supported $f\in L^1_{\mathrm{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ therein with respect to τ.

Proposition 4 (Necessary condition for convergence in the separating topology τ)**.** *Let $n\in\mathbb{N}_+$ and $f\in L^1_n(\mathbb{R}^d,\mathbb{R}^D)$. A sequence $\{f_k\}_{k\in\mathbb{N}^+}$ in $L^1_{\mathrm{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ converges to f with respect to the separating topology τ, only if all but a finite number of f_k are in $L^1_n(\mathbb{R}^d,\mathbb{R}^D)$.*

Together, Proposition 4 and the fact that if any analytic function is 0 on a non-empty open subset of $\mathbb{R}^d$ then it must be identically 0 everywhere on $\mathbb{R}^d$ (see (Griffiths & Harris, 1994, page 1)) imply that no analytic function can converge to an essentially compactly supported "function" in $L^1_{\mathrm{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ with respect to the separating topology.

Lemma 6 (Families of analytic functions cannot be dense with respect to the separating topology τ)**.** *If $\mathscr{F}$ is a set of analytic functions from $\mathbb{R}^d$ to $\mathbb{R}^D$ then*

1. *$\mathscr{F}$ is not dense in $L^1_{loc}(\mathbb{R}^d,\mathbb{R}^D)$ for the separating topology τ.*
2. *If $f:\mathbb{R}^d\to\mathbb{R}^D$ is Lipschitz, is compact essential-supported, and not identically 0 then, is a sequence $\{\varepsilon_n\}_{n=1}^{\infty}$ in $(0,\infty)$ converging to 0 such that no $\hat{f}\in\mathscr{F}$ satisfies both Theorem 2 (i) and (iii).*

The proof of Theorem 1 (ii) is a consequence of Lemma 6 and the observation that any network in $\mathrm{NN}^{\omega+\mathrm{Pool}}$ is an analytic function.

Proof of Theorem 1 (ii). By Lemma 6, the class of analytic functions from $\mathbb{R}^d$ to $\mathbb{R}^D$, denoted by $C^{\omega}(\mathbb{R}^d,\mathbb{R}^D)$, is not dense in $L^1_{\mathrm{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ for the separating topology. Now, the composition and the addition of analytic functions is again analytic. Since every affine function is analytic and since every activation function $\sigma\in C^{\omega}(\mathbb{R})$ is by definition analytic then, every $f\in\mathrm{NN}^{\omega}$ must be analytic. I.e, $\mathrm{NN}^{\omega}\subseteq C^{\omega}(\mathbb{R}^d,\mathbb{R}^D)$. Therefore, NN^{ω} cannot be in $L^1_{\mathrm{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ for the separating topology. ☐

The proof of Theorem 3 now also follows from Lemma 6.

Proof of Theorem 3. Since every polynomial function is analytic then, the result follows from Lemma 6. ☐

Proof of Theorem 2 (Continued). If $f:\mathbb{R}^d\to\mathbb{R}^D$ is Lipschitz, compactly-supported, and not identically 0 then Lemma 6 and the fact that every $\hat{f}\in\mathrm{NN}^{\omega+\mathrm{Pool}}\cup\mathbb{R}[x_1,\dots,x_d]$ is an analytic function implies that Theorem 2 (i)-(iii) cannot all hold simultaneously. This completes the proof of Theorem 2. ☐

We now discuss some technical points surrounding our results, a few of the implications of our findings, and how our analysis could be used to obtain similar constructions for networks designed to approximate solutions to PDEs.

Discussion

In analogy with the results of Yarotsky (2017b), wherein the authors shows that feedforward networks achieve *optimal* approximation rates of a function in L^1, we ask:

Is τ the smallest topology on $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ for which $\mathrm{NN}^{\text{ReLU}+\text{Pool}}$ is dense but $\mathrm{NN}^{\omega+\text{Pool}}$ is not?

A very different qualitative phenomenon manifests in our topological study; namely, there is no optimal topology on $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ exhibiting an "optimal" comparable separating phenomenon exhibited by the *separating topology* τ.

Proposition 5 (Non-Existence Optimal Separating Topology)**.**
There does not exist a topology $\tau^\star$ on $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ such that:

(i) ***Separation:*** $\mathrm{NN}^{\text{ReLU}+\text{Pool}}$ *is dense in* $\tau^\star$ *and* $\mathrm{NN}^{\omega+\text{Pool}}$ *is not dense in* $\tau^\star$*,*

(ii) ***Optimality:*** *If $\tilde{\tau}$ is a topology on $L^1_{\text{loc}}(\mathbb{R}^d,\mathbb{R}^D)$ satisfying (i) then, $\tau^\star \subseteq \tilde{\tau}$.*

There are several explanations for a learning model's success over its alternatives for a *given learning task*. Some of the principle reasons for a model's successful inductive bias are its expressiveness, its ability to generalize well on a given type of problem, and how training dynamics interact with these two properties for a given problem (e.g. the impage of using different initialization schemes as studied by (Martens et al., 2021)).

A key point which we emphasize is that, the type of problem which we have implicitly considered in this paper concerns the approximation of a compactly supported function's output and its support simultaneously. Therefore, we ask the following question from the approximation-theoretic vantage point:

"Are ReLU networks always better than networks with analytic activation functions?"

As one may expect, the answer is a mixed *"yes and no"*. The reason is that is the task is to learn a solution to a PDE (e.g. Han et al. (2018); Beck et al. (2021a;b) physics-informed neural networks Raissi et al. (2019); Shin et al. (2020); Mishra & Molinaro (2021)). Then, the networks should exhibit non-trivial (higher-order) partial derivatives, and the approximation should be in the C^k-norm (for some $k>0$). In such cases, it is known that ReLU networks are less effective than sigmoid, tanh, or SIREN networks; see Markidis (2021) or Hornik et al. (1990); Siegel & Xu (2020); De Ryck et al. (2021). This is the *"no"* part of the answer to the above question.

The *"yes"* part of the answer to the above question is more delicate. Notice that the k-fold anti-derivative of ReLU is the function $\max\{0,t^{k+1}\} \overset{\text{def.}}{=} \mathrm{ReLU}_k$ (called a *truncated power* in (DeVore & Lorentz, 1993, Chapter 5, Equation (1.1))). Therefore, ReLU_k has non-vanishing k first derivatives on the positive half-line. Thus, networks build with it as an activation function have enough flexibility to approximate all the first k partial derivatives of a k-dimension continuously differentiable functions. However, by construction, $\partial^k \mathrm{ReLU}_k = \mathrm{ReLU}$ therefore, if the target function is k-times continuous differentiable and its k^{th}-partial derivatives vanishes outside some compact subset of the input space, then neural networks with ReLU_k activation function should also be able to identify the compact set on which f has non-zero k^{th}-partial derivatives; in the same way as Theorems 1 and 2 showed that neural networks ReLU activation function could identify the support of the "0^{th} order derivative" of a Lipschitz function.

In other words, the authors expect that a construction similar to τ should be possible by replacing the sets $L_n(\mathbb{R}^d,\mathbb{R}^D)$ with some suitable subset of a Sobolev space $\mathscr{W}^{1,1}(\mathbb{R}^d)$ whose elements are functions with $\partial_i^k f = 0$ outside of $[-n,n]^d$ for each $i=1,\dots,d$. The authors plan to explore this extension in a subsequent work.

We conclude our discussion by considering one last question:

"What is the significance of the bi-linear pooling layer?"

Our construction of a network $\hat{f} \in \mathrm{NN}^{\text{ReLU}+\text{Pool}}$ realizing the conclusion of Theorem 2 for a given approximation error $\varepsilon > 0$ relies two distinct ReLU networks which are multiplied together using bi-linear pooling layers. Suppose that $f:\mathbb{R}^d \to \mathbb{R}^D$ is a compactly supported Lipschitz function and let n be the smallest integer for which $\operatorname{ess-supp}(f)$

is contained in the union of the first n Cubic Annuli of Example 7. The role first ReLU network $\hat{f}_{\text{mask}} : \mathbb{R}^d \to \mathbb{R}$ is to implements a piece-wise affine "mask" which takes values 0 outside of $\cup_{i=1}^{n+1} K_i$, value 1 in $\cup_{i=1}^{n} K_i$, and intermediate value in $K_{n+1} - K_n$ just as in the construction of Yarotsky (2017b). The second ReLU network $\hat{f}^{\varepsilon}$ is constructed which approximates the target function $f : \mathbb{R}^d \to \mathbb{R}^D$ uniformly on the compact set ess-supp(f) to ε-precision, and we construct the ReLU network $\hat{f}^{\varepsilon}$ in such a way that its depth and width depend on the dimension and metric capacity of ess-supp(f) as well as on the regularity of the function f.

Lastly, using several bilinear pooling layers we construct the approximating network $\hat{f}$ in Theorem 2 which implements $\hat{f} = \hat{f}_{\text{mask}} \cdot \hat{f}^{\varepsilon}$. Consequentially, $\hat{f} \approx f$ for every $x \in$ ess-supp(f) and it is supported exactly on $\cup_{i=1}^{n+1} K_i$ (i.e.: its support coincides with that of the target function up to our discretization of $\mathbb{R}^d$ as implemented by $\{K_n\}_{n=1}^{\infty}$). The subtle difference in our approach and in the constructions of Yarotsky (2017b); Kidger & Lyons (2020) is that those authors use small ReLU networks to *approximately implement* the multiplication operation $(x_1, x_2) \mapsto x_1 x_2$ instead of the bilinear pooling layers which we use. The issue here is that, their construction does not guarantee that an "approximate product" of $\hat{f}_{\text{mask}}$ and $\hat{f}^{\varepsilon}$ is supported in $\cup_{i=1}^{n+1} K_i$ nor that is has compact support; whence, there is no guarantee with that method that one can construct a deep ReLU network satisfying the conditions of Lemma 2. NB, this is not to say that a construction is impossible; but simply that it remains an *open question*.

Conclusion

This paper showed that deep feedforward networks with piecewise linear activation functions and utilizing bilinear pooling layers are strictly more expressive than deep feedforward networks deploying any number of analytic activation functions and leveraging bilinear pooling layers. We reached this conclusion through two central results, Theorem 1 and Theorem 2.

Theorem 1 provide a topological justification for this claim showing that one can refine topology on $L^1_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^D)$ enough to ensure that $\text{NN}^{\text{ReLU}+\text{Pool}}$ remains universal while $\text{NN}^{\omega+\text{Pool}}$ and $\mathbb{R}[x_1, \dots, x_d : D]$ fail to still be universal. Theorem 2 build on this general result by examining the capability of networks in $\text{NN}^{\text{ReLU}+\text{Pool}}$ to uniformly approximate compactly supported Lipschitz functions while simultaneously identifying the region in $\mathbb{R}^d$ wherein they are supported (as quantified by a good a.e. partition of $\mathbb{R}^d$). In particular, a direct consequence of the analyticity of the maps in $\text{NN}^{\omega+\text{Pool}}$ and $\mathbb{R}[x_1, \dots, x_d : D]$ is that these models do not share this approximation capability.

To the best of our knowledge, this is the first result demonstrating a theoretical approximation-theoretic gap between neural network models utilizing qualitatively different activation functions, beyond just a mildly different approximation rate. We hope that our new proof techniques and our novel constructions can be extended by other community members looking to further understand the approximation-theoretic advantages of different model types.

Bibliography

Beatrice Acciaio, Anastasis Kratsios, and Gudmund Pammer. Metric hypertransformers are universal adapted maps. *arXiv preprint arXiv:2201.13094*, 2022.

Christian Beck, Sebastian Becker, Patrick Cheridito, Arnulf Jentzen, and Ariel Neufeld. Deep splitting method for parabolic PDEs. *SIAM J. Sci. Comput.*, 43(5):A3135–A3154, 2021a. ISSN 1064-8275. doi: 10.1137/19M1297919. URL `https://doi.org/10.1137/19M1297919`.

Christian Beck, Sebastian Becker, Philipp Grohs, Nor Jaafari, and Arnulf Jentzen. Solving the Kolmogorov PDE by means of deep learning. *J. Sci. Comput.*, 88(3):Paper No. 73, 28, 2021b. ISSN 0885-7474. doi: 10.1007/s10915-021-01590-0. URL `https://doi.org/10.1007/s10915-021-01590-0`.

Aleksandr Beknazaryan. Neural networks with superexpressive activations and integer weights. *arXiv preprint arXiv:2105.09917*, 2021.

Elia Bruè, Simone Di Marino, and Federico Stra. Linear lipschitz and c1 extension operators through random projection. *Journal of Functional Analysis*, 280(4):108868, 2021.

H. Buehler, L. Gonon, J. Teichmann, and B. Wood. Deep hedging. *Quant. Finance*, 19(8):1271–1291, 2019. ISSN 1469-7688. doi: 10.1080/14697688.2019.1571683. URL `https://doi.org/10.1080/14697688.2019.1571683`.

Patrick Cheridito, Arnulf Jentzen, and Florian Rossmannek. Efficient approximation of high-dimensional functions with neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2021a.

Patrick Cheridito, Arnulf Jentzen, and Florian Rossmannek. Efficient approximation of high-dimensional functions with neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2021b.

Christa Cuchiero, Martin Larsson, and Josef Teichmann. Deep neural networks, generic universal interpolation, and controlled ODEs. *SIAM J. Math. Data Sci.*, 2(3):901–919, 2020. doi: 10.1137/19M1284117. URL `https://doi.org/10.1137/19M1284117`.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems*, 2(4): 303–314, 1989. ISSN 0932-4194.

Tim De Ryck, Samuel Lanthaler, and Siddhartha Mishra. On the approximation of functions by tanh neural networks. *Neural Networks*, 143:732–750, 2021. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2021.08.015. URL `https://www.sciencedirect.com/science/article/pii/S0893608021003208`.

Ronald A. DeVore and George G. Lorentz. *Constructive approximation*, volume 303 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1993. ISBN 3-540-50627-6.

Jean Dieudonné and Laurent Schwartz. La dualité dans les espaces F et LF. *Ann. Inst. Fourier (Grenoble)*, 1:61–101 (1950), 1949.

Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Bilinear attention networks for person retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8030–8039, 2019.

Jorge Galindo and Manuel Sanchis. Stone-Weierstrass theorems for group-valued functions. *Israel J. Math.*, 141: 341–354, 2004. ISSN 0021-2172. doi: 10.1007/BF02772227. URL `https://doi.org/10.1007/BF02772227`.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL `https://proceedings.mlr.press/v15/glorot11a.html`.

Gribonval Rémi, Kutyniok Gitta, Nielsen Morten, and Voigtlaender Felix. Approximation spaces of deep neural networks. *Constructive Approximation*, forthcoming, 05 2021. ISSN 1432-0940. doi: https://doi.org/10.1007/s00365-021-09543-410.1007/s00365-021-09543-4.

Phillip Griffiths and Joseph Harris. *Principles of algebraic geometry*. Wiley Classics Library. John Wiley & Sons, Inc., New York, 1994. ISBN 0-471-05059-8. doi: 10.1002/9781118032527. URL `https://doi.org/10.1002/9781118032527`. Reprint of the 1978 original.

Lyudmila Grigoryeva and Juan-Pablo Ortega. Differentiable reservoir computing. *J. Mach. Learn. Res.*, 20:Paper No. 179, 62, 2019.

Ingo Gühring, Gitta Kutyniok, and Philipp Petersen. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Anal. Appl. (Singap.)*, 18(5):803–859, 2020a. ISSN 0219-5305. doi: 10.1142/S0219530519410021. URL `https://doi.org/10.1142/S0219530519410021`.

Ingo Gühring, Gitta Kutyniok, and Philipp Petersen. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Anal. Appl. (Singap.)*, 18(5):803–859, 2020b. ISSN 0219-5305. doi: 10.1142/S0219530519410021. URL `https://doi.org/10.1142/S0219530519410021`.

Jiequn Han, Arnulf Jentzen, and Weinan E. Solving high-dimensional partial differential equations using deep learning. *Proc. Natl. Acad. Sci. USA*, 115(34):8505–8510, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1718942115. URL https://doi.org/10.1073/pnas.1718942115.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Andreas Heinecke, Jinn Ho, and Wen-Liang Hwang. Refinement and universal approximation via sparsely connected relu convolution nets. *IEEE Signal Processing Letters*, 27:1175–1179, 2020.

Juha Heinonen. *Lectures on analysis on metric spaces*. Universitext. Springer-Verlag, New York, 2001.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Netw.*, 2(5):359–366, July 1989.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, 3(5):551–560, 1990. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(90)90005-6. URL https://www.sciencedirect.com/science/article/pii/0893608090900056.

Sebastian Jaimungal. Reinforcement learning and stochastic optimisation. *Finance Stoch.*, 26(1):103–129, 2022. ISSN 0949-2984. doi: 10.1007/s00780-021-00467-2. URL https://doi.org/10.1007/s00780-021-00467-2.

Yuling Jiao, Yanming Lai, Xiliang Lu, and Zhijian Yang. Deep neural networks with relu-sine-exponential activations break curse of dimensionality on h\" older class. *arXiv preprint arXiv:2103.00542*, 2021.

Patrick Kidger and Terry Lyons. Universal Approximation with Deep Narrow Networks. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Machine Learning Research*, volume 125, pp. 2306–2327. PMLR, 09–12 Jul 2020.

Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *ICLR*, 2016.

Anastasis Kratsios and Ievgen Bilokopytov. Non-euclidean universal approximation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 10635–10646. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/786ab8c4d7ee758f80d57e65582e609d-Paper.pdf.

Anastasis Kratsios and Cody Hyndman. Neu: A meta-algorithm for universal uap-invariant feature representation. *Journal of Machine Learning Research*, 22(92):1–51, 2021. URL http://jmlr.org/papers/v22/18-803.html.

Anastasis Kratsios and Leonie Papon. Universal approximation theorems for differentiable geometric deep learning, 2021.

Anastasis Kratsios and Behnoosh Zamanlooy. Learning sub-patterns in piecewise continuous functions. *Neurocomputing*, 480:192–211, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2022.01.036. URL https://www.sciencedirect.com/science/article/pii/S092523122200056X.

Anastasis Kratsios, Behnoosh Zamanlooy, Tianlin Liu, and Ivan Dokmanić. Universal approximation under constraints is possible with transformers. In *International Conference on Learning Representations*, pp. 00, 2022. URL https://openreview.net/forum?id=JGO8CvG5S9.

Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861 – 867, 1993.

Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pp. 1449–1457, 2015.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM J. Math. Anal.*, 53(5):5465–5506, 2021. ISSN 0036-1410. doi: 10.1137/20M134695X. URL https://doi.org/10.1137/20M134695X.

Saunders MacLane. *Categories for the working mathematician*. Springer-Verlag, New York-Berlin, 1971. Graduate Texts in Mathematics, Vol. 5.

Stefano Markidis. The old and the new: Can physics-informed deep-learning replace traditional linear solvers? *Frontiers in Big Data*, 4, 2021. ISSN 2624-909X. doi: 10.3389/fdata.2021.669097. URL https://www.frontiersin.org/article/10.3389/fdata.2021.669097.

James Martens, Andy Ballard, Guillaume Desjardins, Grzegorz Swirszcz, Valentin Dalibard, Jascha Sohl-Dickstein, and Samuel S Schoenholz. Rapid training of deep neural networks without skip connections or normalization layers using deep kernel shaping. *arXiv preprint arXiv:2110.01765*, 2021.

Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.*, 5:115–133, 1943. ISSN 0007-4985. doi: 10.1007/bf02478259. URL https://doi.org/10.1007/bf02478259.

Siddhartha Mishra and Roberto Molinaro. Physics informed neural networks for simulating radiative transfer. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 270:107705, 2021.

James R. Munkres. *Topology*. Prentice Hall, Inc., Upper Saddle River, NJ, 2000. ISBN 0-13-181629-2. Second edition of [MR0464128].

Jun-iti Nagata. *Modern general topology*. North-Holland Publishing Co., Amsterdam-London; Wolters-Noordhoff Publishing, Groningen; American Elsevier Publishing Co., New York, revised edition, 1974. Bibliotheca Mathematica, Vol. VII.

P. P. Narayanaswami and Stephen A. Saxon. (LF)-spaces, quasi-Baire spaces and the strongest locally convex topology. *Math. Ann.*, 274(4):627–641, 1986.

J. A. A. Opschoor, Ch. Schwab, and J. Zech. Exponential ReLU DNN expression of holomorphic maps in high dimension. *Constr. Approx.*, 55(1):537–582, 2022. ISSN 0176-4276. doi: 10.1007/s00365-021-09542-5. URL https://doi.org/10.1007/s00365-021-09542-5.

M. Scott Osborne. *Locally convex spaces*, volume 269 of *Graduate Texts in Mathematics*. Springer, Cham, 2014.

Sejun Park, Chulhee Yun, Jaeho Lee, and Jinwoo Shin. Minimum width for universal approximation. *ICLR*, 2021.

Pencho P. Petrushev. Approximation by ridge functions and neural networks. *SIAM J. Math. Anal.*, 30(1): 155–189, 1999. ISSN 0036-1410. doi: 10.1137/S0036141097322959. URL https://doi.org/10.1137/S0036141097322959.

João B. Prolla. On the Weierstrass-Stone theorem. *J. Approx. Theory*, 78(3):299–313, 1994. ISSN 0021-9045. doi: 10.1006/jath.1994.1080. URL https://doi.org/10.1006/jath.1994.1080.

Michael Anthony Puthawala, Matti Lassas, Ivan Dokmanić, and Maarten V. de Hoop. Universal joint approximation of manifolds and densities by simple injective flows, 2022. URL https://openreview.net/forum?id=HUeyM2qVey2.

M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.*, 378: 686–707, 2019. ISSN 0021-9991. doi: 10.1016/j.jcp.2018.10.045. URL https://doi.org/10.1016/j.jcp.2018.10.045.

Prajit Ramachandran, Barret Zoph, and Quoc Le. Searching for activation functions. In *International Conference of Learning Representations*, pp. 00, 2018. URL https://arxiv.org/pdf/1710.05941.pdf.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Raymond A. Ryan. *Introduction to tensor products of Banach spaces*. Springer Monographs in Mathematics. Springer-Verlag London, Ltd., London, 2002.

Laurent Schwartz. *Théorie des distributions*. Publications de l'Institut de Mathématique de l'Université de Strasbourg, No. IX-X. Nouvelle édition, entiérement corrigée, refondue et augmentée. Hermann, Paris, 1966.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Optimal approximation rate of ReLU networks in terms of width and depth. *J. Math. Pures Appl. (9)*, 157:101–135, 2022. ISSN 0021-7824. doi: 10.1016/j.matpur.2021.07.009. URL https://doi.org/10.1016/j.matpur.2021.07.009.

Yeonjong Shin, Jérôme Darbon, and George Em Karniadakis. On the convergence of physics informed neural networks for linear second-order elliptic and parabolic type PDEs. *Commun. Comput. Phys.*, 28(5):2042–2074, 2020. ISSN 1815-2406. doi: 10.4208/cicp.oa-2020-0193. URL https://doi.org/10.4208/cicp.oa-2020-0193.

Jonathan W. Siegel and Jinchao Xu. Approximation rates for neural networks with general activation functions. *Neural Networks*, 128:313 – 321, 2020. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2020.05.019. URL http://www.sciencedirect.com/science/article/pii/S0893608020301891.

Taiji Suzuki. Adaptivity of deep reLU network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, pp. 00, 2019. URL https://openreview.net/forum?id=H1ebTsActm.

Vlad Timofte, Aida Timofte, and Liaqat Ali Khan. Stone-Weierstrass and extension theorems in the nonlocally convex case. *J. Math. Anal. Appl.*, 462(2):1536–1554, 2018. ISSN 0022-247X. doi: 10.1016/j.jmaa.2018.02.056. URL https://doi.org/10.1016/j.jmaa.2018.02.056.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103 – 114, 2017a. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2017.07.002.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017b. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2017.07.002. URL https://www.sciencedirect.com/science/article/pii/S0893608017301545.

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 639–649. PMLR, 06–09 Jul 2018.

Dmitry Yarotsky. Elementary superexpressive activations. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11932–11940. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/yarotsky21a.html.

Dmitry Yarotsky. Universal approximations of invariant maps by neural networks. *Constr. Approx.*, 55(1): 407–474, 2022. ISSN 0176-4276. doi: 10.1007/s00365-021-09546-1. URL https://doi.org/10.1007/s00365-021-09546-1.

Dmitry Yarotsky and Anton Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13005–13015. Curran Associates, Inc., 2020a.

Dmitry Yarotsky and Anton Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13005–13015. Curran Associates, Inc., 2020b.

Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Appl. Comput. Harmon. Anal.*, 48(2): 787–794, 2020. ISSN 1063-5203. doi: 10.1016/j.acha.2019.06.004. URL https://doi.org/10.1016/j.acha.2019.06.004.