NEURAL NETWORKS IN FRÉCHET SPACES

FRED ESPEN BENTH, NILS DETERING, LUCA GALIMBERTI

Abstract. We propose a neural network architecture in infinite dimensional spaces for which we can show the universal approximation property. Indeed, we derive approximation results for continuous functions from a Fréchet space $\mathfrak{X}$ into a Banach space $\mathfrak{Y}$. The approximation results are generalising the well known universal approximation theorem for continuous functions from $\mathbb{R}^n$ to $\mathbb{R}$, where approximation is done with (multilayer) neural networks [16, 28, 20, 35]. Our infinite dimensional networks are constructed using activation functions being nonlinear operators and affine transforms. Several examples are given of such activation functions. We show furthermore that our neural networks on infinite dimensional spaces can be projected down to finite dimensional subspaces with any desirable accuracy, thus obtaining approximating networks that are easy to implement and allow for fast computation and fitting. The resulting neural network architecture is therefore applicable for prediction tasks based on functional data.

1. Introduction

The universal approximation theorem shows that any continuous function from $\mathbb{R}^n$ to $\mathbb{R}$ can be approximated arbitrary well with a one layer neural network. More precisely, for a fixed continuous function $\sigma : \mathbb{R} \to \mathbb{R}$ and $a \in \mathbb{R}^n, \ell, b \in \mathbb{R}$, a *neuron* is a function $\mathcal{N}_{\ell,a,b} \in C(\mathbb{R}^n; \mathbb{R})$ defined by $x \mapsto \ell\sigma(a^\top x + b)$. The universal approximation theorem states conditions on the *activation function* σ such that the linear space of functions generated by the neurons

$$\mathfrak{N}(\sigma) := \text{span}\{\mathcal{N}_{\ell,a,b}; \ell, b \in \mathbb{R}, a \in \mathbb{R}^n\}$$

is dense with respect to the topology of uniform convergence on compacts. This means that for every $f \in C(\mathbb{R}^n; \mathbb{R})$ and compact subset $K \subset \mathbb{R}^n$ and a given $\varepsilon > 0$, there exists $N \in \mathbb{N}$ and $\ell_i, b_i \in \mathbb{R}, a_i \in \mathbb{R}^n$ for $i = 1, \ldots, N$ such that

$$\sup_{x \in K} \left| f(x) - \sum_{i=1}^{N} \mathcal{N}_{\ell_i,a_i,b_i}(x) \right| \leq \varepsilon.$$

Possibly the most widely known property of σ that was shown in Cybenko [16] and Hornik, Stinchcombe, and White [28] to lead to the density of $\mathfrak{N}(\sigma) \subset C(\mathbb{R}^n; \mathbb{R})$ is the *sigmoid* property, which requires σ to be such that $\lim_{t\to\infty} \sigma(t) = 1$ and $\lim_{t\to-\infty} \sigma(t) = 0$. This condition has later been relaxed to a boundedness condition Funahashi [20] and a non-polynomial condition Leshno *et al.* [35]. All these results are on finite-dimensional shallow neural networks that consist of one or two layers with many neurons (bounded depth, arbitrary width). In contrast stands the analysis of networks with arbitrary depth and bounded width which has also recently received attention [38, 23, 30]. We refer the reader to Pinkus [44] for an overview of the earlier literature on neural network approximation theory and to Berner *et al.* [5] for a more recent account. See also Kratsios [32] for a unified approach of approximation results for a wide class of network architectures.

In this paper we are concerned with more general functions $f \in C(\mathfrak{X}; \mathfrak{Y})$, where $\mathfrak{X}$ is an $\mathbb{F}$-Fréchet space, i.e., a Fréchet space over the field $\mathbb{F}$ and $\mathfrak{Y}$ an $\mathbb{F}$-Banach space. We start with $\mathfrak{Y} = \mathbb{F}$. In the definition of a neuron, we replace $a^\top x + b$ by an affine function on $\mathfrak{X}$, the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ by a function in $C(\mathfrak{X}; \mathfrak{X})$, and the scalar ℓ by a linear form. With $\langle \cdot, \cdot \rangle$ the canonical pairing between $\mathfrak{X}'$ and $\mathfrak{X}$ ($\mathfrak{X}'$ denoting the topological dual of $\mathfrak{X}$), for $\ell \in \mathfrak{X}', A \in \mathcal{L}(\mathfrak{X}), b \in \mathfrak{X}$ we then define a neuron $\mathcal{N}_{\ell,A,b}$ by

$$\mathcal{N}_{\ell,A,b}(x) = \langle \ell, \sigma(Ax + b) \rangle$$

Date: February 14, 2023.
Key words and phrases. Neural networks, universal approximation, Fréchet space, activation function.

and ask for conditions on $\sigma : \mathfrak{X} \to \mathfrak{X}$ that ensure that $\mathfrak{N}(\sigma) := \text{span}\{\mathcal{N}_{\ell,A,b}; \ell \in \mathfrak{X}', A \in \mathcal{L}(\mathfrak{X}), b \in \mathfrak{X}\}$ is dense in $C(\mathfrak{X};\mathbb{F})$ under some suitable topology. We thus treat the infinite-dimensional shallow neural network case (bounded depth, arbitrary width).

To indicate the conditions we obtain for σ, recall that any map $\psi \in \mathfrak{X}'$ defines a hyperplane by the set of points $\Psi_0 := \{x \in \mathfrak{X}; \langle\psi, x\rangle = 0\}$. This hyperplane splits the space $\mathfrak{X}$ into the sets $\Psi_- := \{x \in \mathfrak{X}; \langle\psi, x\rangle < 0\}$ and $\Psi_+ := \{x \in \mathfrak{X}; \langle\psi, x\rangle > 0\}$. We show that the main property for the activation function to ensure that $\mathfrak{N}(\sigma)$ is dense in $C(\mathfrak{X};\mathbb{F})$ is, informally, that an $\psi \in \mathfrak{X}'$ exists such that the value $\sigma(x)$ converges, as x moves away from the hyperplane that is defined by ψ. The limiting values on both sides of the hyperplane need to be different. We provide several simple examples of easy to calculate activation functions with the required property. In a second step, we extend our results to $f \in C(\mathfrak{X};\mathfrak{Y})$, where $\mathfrak{Y}$ is an $\mathbb{F}$-Banach space.

While such an approximation result might be of interest in its own, from a practical perspective it is not clear how the functions $\mathcal{N}_{\ell,A,b}$, which involve infinite dimensional quantities, can actually be programmed. We therefore address the question of approximating the maps $\mathcal{N}_{\ell,A,b}$ by finite dimensional, easy to calculate quantities. Under the assumption that the Fréchet space $\mathfrak{X}$ admits a Schauder basis, we show that such an approximation is possible. The resulting neural network has an architecture similar to classical neural networks, with the exception that the activation function is now multidimensional. It does however still permit for an easy to calculate gradient, which is crucial for training the network via a back-propagation algorithm. Finally, we also derive the approximation property for deep neural networks with a given fixed number of layers.

Possible applications of our results are within the area of machine learning, in particular in the many situations where the input of each sample in the training set is actually a function (see e.g. Ramsey and Silverman [45] for an account on functional data analysis and examples). In our accompanying paper [2] we use the results obtained here to derive numerical solutions of partial differential equations for a range of initial conditions or coefficients at once (see Han, Jentzen and E [22], Hutzenthaler *et al.* [29], Cuchiero, Larsson and Teichmann [15], Beck *et al.* [1] for papers on neural networks and partial differential equations). There are other instances where functional data appears naturally. For example grey scale images can be understood as a function $I : [0, 1]^2 \to [0, 1]$. For imagine classification or recognition problems (see Müller, Soto-Ray and Kramer [41] and Tian [49]) one is now interested in approximating the function f that assigns to each image its classification $f(I)$. Additional examples are stock price prediction (see Yu and Yan [52]), option pricing and hedging (see Buehler *et al.* [9] and Benth, Detering and Lavagnini [4]), and many others. We present in this paper an example from commodity markets option pricing.

If the function space of the inputs is a Fréchet space with a Schauder basis, this basis provides structural information about the elements. Traditional neural networks must be of very high dimension (large input dimension, large number of neurons) to approximate a function well. The more variability there is in the function, the larger the number of parameters that is needed. Therefore, instead of using a classical network to approximate a function on a grid, our approach allows one to use information in the basis functions instead to capture the structure and get theoretical convergence results. Our approximation thus focuses on features of the function related to the coefficients in the basis expansion. Moreover, we show that there is a large class of possible activation functions $\sigma : \mathfrak{X} \to \mathfrak{X}$ and a choice that is suitable for the approximation problem at hand can significantly reduce the number of nodes required to approximate a given function f sufficiently well. We refer to our accompanying paper [3] where this idea is used to price flow forward derivatives in energy markets.

Related literature: The approximation with neural networks of functionals and operators that are defined on some general (possibly infinite dimensional) space $\mathfrak{X}$ goes back to Sandberg [46]. In Sandberg [46] in the context of discrete time systems, non-linear functionals on a space of functions from $\mathbb{N} \cup \{0\}$ to $\mathbb{N}$ are approximated with neural networks. In Chen and Chen [11, 12] the authors consider the approximation of non-linear operators defined on infinite dimensional spaces and use these results for approximating the output of dynamical systems. Among other results they approximate functions $f : K \subset \mathfrak{X} \to \mathbb{R}$, where

$\mathfrak{X}$ is Banach, K is compact and f is continuous. In Mhaskar and Hahm [40] the authors derive networks that approximate the functionals on the function spaces $L^p([-1,1]^s)$ for $1 \leq p < \infty$ and $C([-1,1]^s)$ for integer $s \geq 1$. The recent article by Kratisos [32] considers a space $M(\mathfrak{X}, \mathfrak{Y})$ of functions from a metric space $\mathfrak{X}$ to another metric space $\mathfrak{Y}$. Among other results, under the assumption that this functions space is homeomorphic to an infinite-dimensional Fréchet space, the author derives properties of neural network architectures that are dense within this space. We would like to stress however that in the situation we consider in this paper, the domain space $\mathfrak{X}$ is a Fréchet space. The function space $C(\mathfrak{X}, \mathbb{R})$ however is usually not a Fréchet space unless $\mathfrak{X}$ is finite dimensional. Infinitely wide neural networks, with an infinite but countable number of nodes in the hidden layer have been studied in the context of Bayesian learning, Gaussian processes and kernel methods by several authors, see e.g., Neal [43], Williams [51], Cho and Saul [13] and Hazan and Jaakola [24]. Hornik [27] provides approximation results for such infinitely wide networks. Guss and Salakhutdinov [21] prove the universal approximation property for two-layer infinite dimensional neural networks. They show their approximation property for continuous maps between spaces of continuous functions on compacts.

Recently so-called DeepONets for the approximation of operators between Banach spaces of continuous functions on compact subsets of $\mathbb{R}^n$ have been proposed and analyzed by Lu *et al.* [37], and Lauthaler, Mishra and Karniadakis [34]. DeepONets follow a similar structure as the one used in Chen and Chen [12] of a branch net that uses signals to extract information about the functions in the domain, and a trunk net to map to the image. In DeepONets both the branch and trunk nets are deep neural nets. In Kovachki *et al.* [31] and Li *et al.* [36], the authors propose a neural network method to approximate the solution operator that assigns to a coefficient function for a partial differential equation (PDE) its solution function. This leads again to the approximation of an operator between Banach spaces of functions that are defined on a bounded domain in $\mathbb{R}^n$. The neural network that is presented in [31, 36] is tailor-made for the specific problem at hand and its structure is motivated by the Green function which defines the solution to the PDE. We also refer the reader to Kratsios and Bilokopytov [33] for approximations on manifolds in $\mathbb{R}^n$.

Our approach differs in several ways from those previously proposed: We allow for very general spaces $\mathfrak{X}$ going beyond Banach spaces, which extends the scope of applications. We provide an example of a continuous function on Fréchet space in Section 6 where this generality is needed. In contrast to most currently available neural networks for infinite spaces, our architecture focuses on information inherit in the basis decomposition. This decomposition carries important structural information that helps in the learning process. Moreover, the network architectures in the works discussed above have in common an activation function σ with image in $\mathbb{R}$ instead of $\mathfrak{X}$ as we propose it here. Our notion of a neural network in infinite dimensions is additionally motivated by the relationship with controlled ordinary differential equations, which points towards an activation function $\sigma : \mathfrak{X} \to \mathfrak{X}$ rather than the classical one-dimensional maps (possibly on basis coordinates). We refer to E [17] for a connection between ordinary differential equations and deep neural networks. Lastly, our networks are structurally very similar to classical neural networks. This allows to easily adapt widely available and efficient packages as TensorFlow or PyTorch to approximate maps between infinite dimensional spaces. From a theoretical perspective our results allow us to separate two approximations. First, the approximation of arbitrary functions $f \in C(\mathfrak{X}; \mathfrak{Y})$ with a superposition of infinite dimensional neurons, and second, the approximation of the resulting infinite dimensional neural network with computable, finite dimensional quantities.

The outline for the paper is as follows. In Section 2 we derive our first main result Theorem 2.3, which shows that if σ has a property called *discriminatory* property, then $\mathfrak{N}(\sigma)$ is dense in $C(\mathfrak{X}; \mathbb{F})$. The main technical challenge is then to derive conditions that ensure that a given function $\sigma : \mathfrak{X} \to \mathfrak{X}$ is actually discriminatory, which is done in Theorem 2.8. We also provide some first examples of discriminatory functions in this section. We then extend these results in Section 3 to functions $f \in C(\mathfrak{X}; \mathfrak{Y})$, $\mathfrak{Y}$ Banach space. In Section 4 we address the question of finite dimensional approximations to the neural network which can easily be computed and trained. In most generality, only under the assumption that the Fréchet space $\mathfrak{X}$ has a Schauder basis, the approximation is covered in Theorem 4.3.

In Section 5 we cover the approximation with multi-layered neural networks. Finally, in Section 6 we discuss an application from commodity markets, where continuous functions on Fréchet spaces appear.

Acknowledgements

We are grateful to an anonymous referee for the constructive critics leading to a significant improvement of this paper. Fred Espen Benth acknowledges support from SPATUS, a Thematic Research Group funded by UiO:Energy. Luca Galimberti has been supported in part by the grant Waves and Nonlinear Phenomena (WaNP) from the Research Council of Norway.

2. An abstract approximation result

Let $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, and let $\mathfrak{X}$ be an $\mathbb{F}$-Fréchet space. Let $(p_k)_{k\in\mathbb{N}}$ be an increasing sequence of seminorms that generates the topology of $\mathfrak{X}$. We can then consider a metric d on $\mathfrak{X}$ (that generates the same topology) given by

$$d(x,y) := \sum_{k=1}^{\infty} 2^{-k} \frac{p_k(x-y)}{1+p_k(x-y)}, \tag{1}$$

for $x, y \in \mathfrak{X}$.

Let us consider $\sigma : \mathfrak{X} \to \mathfrak{X}$ continuous function. Let $A : \mathfrak{X} \to \mathfrak{X}$ be in $\mathcal{L}(\mathfrak{X})$, i.e. a linear and continuous operator, $b \in \mathfrak{X}$ and $\ell \in \mathfrak{X}'$, where $\mathfrak{X}'$ denotes the topological dual of $\mathfrak{X}$. Let us consider the following function:

$$\mathcal{N}_{\ell,A,b} : \mathfrak{X} \to \mathbb{F}, \quad \mathcal{N}_{\ell,A,b}(x) := \langle \ell, \sigma(Ax+b)\rangle = \ell(\sigma(Ax+b)), \quad x \in \mathfrak{X}, \tag{2}$$

where $\langle\cdot,\cdot\rangle$ is the canonical pairing between $\mathfrak{X}'$ and $\mathfrak{X}$. We will call such function a *neuron*. Every neuron $\mathcal{N}_{\ell,A,b}$ is clearly continuous by composition of continuous maps, i.e. $\mathcal{N}_{\ell,A,b} \in C(\mathfrak{X};\mathbb{F})$, the space of $\mathbb{F}$-valued continuous functions on $\mathfrak{X}$.

We define

$$\mathfrak{N}(\sigma) := \operatorname{span}\{\mathcal{N}_{\ell,A,b};\ \ell \in \mathfrak{X}', A \in \mathcal{L}(\mathfrak{X}), b \in \mathfrak{X}\},$$

namely, we consider all linear combinations of the form

$$\sum_{j=1}^{N} \alpha_j \mathcal{N}_{\ell_j,A_j,b_j}, \quad \alpha_j \in \mathbb{F}, N \in \mathbb{N}.$$

Evidently, $\mathfrak{N}(\sigma) \subset C(\mathfrak{X};\mathbb{F})$. The maps $\mathcal{N}_{\ell_1,A_1,b_1}, \ldots, \mathcal{N}_{\ell_N,A_N,b_N}$ build a *hidden layer* with N neurons.

We endow $C(\mathfrak{X};\mathbb{F})$ with the topology of uniform convergence on compacts. Being $\mathfrak{X}$ metrizable, it is clearly Tychonoff, and in particular completely regular. For a given compact subset $K \subset \mathfrak{X}$, define

$$q_K(f) := \sup_{x\in K} |f(x)|, \quad f \in C(\mathfrak{X};\mathbb{F}).$$

This is a seminorm on $C(\mathfrak{X};\mathbb{F})$. We consider the topology generated by the family of seminorms $\{q_K; K \subset \mathfrak{X},\ \text{compact}\}$, which is the coarsest topology that makes all the seminorms continuous functions on $C(\mathfrak{X};\mathbb{F})$. This is also called the projective topology induced by the maps q_K for K compact or the topology of compact subsets. Thus, we obtain a locally convex topology on $C(\mathfrak{X};\mathbb{F})$, namely $C(\mathfrak{X};\mathbb{F})$ is an $\mathbb{F}$-locally convex space. Conway [14, Proposition 4.1, p. 114] provides us with the following Riesz representation theorem, which we are going to employ in the sequel:

Proposition 2.1. *If $\phi : C(\mathfrak{X};\mathbb{F}) \to \mathbb{F}$ is a continuous and linear functional, then there is a compact set $K \subset \mathfrak{X}$ and a regular Borel measure μ on K such that $\phi(f) = \int_K f\, d\mu$ for every $f \in C(\mathfrak{X};\mathbb{F})$. Conversely, each such measure defines an element of $C(\mathfrak{X};\mathbb{F})'$. (Observe en passant that $|\mu|(K) < \infty$.)*

We recall that for a locally compact space Y equipped with its Borel σ-algebra $\mathcal{B}(Y)$, a positive measure ν on $\mathcal{B}(Y)$ is a regular Borel measure if

(1) $\nu(F) < \infty$ for every $F \subset Y$ compact,

(2) for any $E \in \mathcal{B}(Y)$, $\nu(E) = \sup\{\nu(F); F \subset E, F \text{ compact}\}$,
(3) for any $E \in \mathcal{B}(Y)$, $\nu(E) = \inf\{\nu(U); U \supset E, U \text{ open}\}$.

If ν is complex-valued or signed instead, then it is regular if $|\nu|$ is.

In the following the expression (μ, K) will denote a compact subset $K \subset \mathfrak{X}$ and a regular $\mathbb{F}$-valued Borel measure μ on K. We say that $\sigma : \mathfrak{X} \to \mathfrak{X}$ continuous is discriminatory if for any fixed pair (μ, K)

$$\int_K \langle \ell, \sigma(Ax + b)\rangle \, \mu(dx) = 0$$

for all $\ell \in \mathfrak{X}'$, $A \in \mathcal{L}(\mathfrak{X})$, $b \in \mathfrak{X}$ implies that $\mu = 0$.

Remark 2.2. *It would be tempting, albeit more challenging, to establish our universal approximation result (Thm 2.3) in the space of bounded continuous functions $C_b(\mathfrak{X}; \mathbb{F})$, endowed with the supremum norm (upon imposing suitable boundedness conditions on the non-linearity σ), rather than approximating on a compact subset K as we are doing in this paper. The main obstruction that prevented us from employing this approach is explained by the succeeding observation: If we aim at following Cybenko's blueprint* [16] *(refer to the proof of Thm. 2.3 below) to establish our result, then in that case we would be required to work with the space*

$$rba(\mathfrak{X}) := \{\mu : \mathcal{B}(\mathfrak{X}) \to \mathbb{F};\ \mu(\emptyset) = 0, \textit{finitely additive, finite and regular}\}$$

which is known to be the dual of $C_b(\mathfrak{X}; \mathbb{F})$, i.e. $C_b(\mathfrak{X}; \mathbb{F})' = rba(\mathfrak{X})$. Dealing with finitely additive measures is more involved, because many standard results from classical measure theory cease to hold. In particular, at this stage it is not clear to us to envisage a suitable set of conditions that the non-linearity σ must satisfy in order to be discriminatory (see Def. 2.6).

Nonetheless, we deem this potential extension of our result to be interesting and worthy to be explored (most likely by deviating completely from Cybenko's strategy of proof), and we hope to be able to come back to this question in the future.

The following first main result shows the density of $\mathfrak{N}(\sigma)$ if σ is discriminatory. The result takes inspiration from Cybenko [16] (see also [20], [28] and [35]), where a similar result has been shown for the case $\mathfrak{X} = \mathbb{R}^n$. For general $\mathfrak{X}$ however, showing that a function $\sigma : \mathfrak{X} \to \mathfrak{X}$ is actually discriminatory can be involved. Later, in Theorem 2.8 we therefore state conditions that can easily be verified and give rise to a large family of discriminatory functions.

Theorem 2.3. *Let $\mathfrak{X}$ be an $\mathbb{F}$-Fréchet space, and let $\sigma : \mathfrak{X} \to \mathfrak{X}$ be continuous and discriminatory. Then $\mathfrak{N}(\sigma)$ is dense in $C(\mathfrak{X}; \mathbb{F})$ when equipped with the projective topology with respect to the seminorms q_K. In other words, given $f \in C(\mathfrak{X}; \mathbb{F})$, then, for any compact subset K of $\mathfrak{X}$, and any $\varepsilon > 0$, there exists $\sum_{m=1}^M \alpha_m \mathcal{N}_{\ell_m, A_m, b_m} \in \mathfrak{N}(\sigma)$ with suitable $\alpha_m \in \mathbb{F}, \ell_m \in \mathfrak{X}', A_m \in \mathcal{L}(\mathfrak{X})$ and $b_m \in \mathfrak{X}$ such that*

$$\sum_{m=1}^M \alpha_m \mathcal{N}_{\ell_m, A_m, b_m} \in \{g \in C(\mathfrak{X}; \mathbb{F});\ q_K(g - f) < \varepsilon\}.$$

Proof. We assume that $\mathrm{cl}(\mathfrak{N}(\sigma)) \subsetneq C(\mathfrak{X}; \mathbb{F})$, and observe that $\mathrm{cl}(\mathfrak{N}(\sigma))$ is clearly still a vector subspace.

We choose $u_0 \in C(\mathfrak{X}; \mathbb{F}) \setminus \mathrm{cl}(\mathfrak{N}(\sigma))$. Since the complement of $\mathrm{cl}(\mathfrak{N}(\sigma))$ is open, we may find $n \in \mathbb{N}$, seminorms $q_{K_1}, \dots, q_{K_n}$ on $C(\mathfrak{X}; \mathbb{F})$ and $\varepsilon_1, \dots, \varepsilon_n > 0$ such that

$$\mathcal{U} := \bigcap_{j=1}^n \{u \in C(\mathfrak{X}; \mathbb{F});\ q_{K_j}(u - u_0) < \varepsilon_j\} \subset C(\mathfrak{X}; \mathbb{F}) \setminus \mathrm{cl}(\mathfrak{N}(\sigma)).$$

Clearly $u_0 \in \mathcal{U}, \mathcal{U}$ is convex, open and disjoint from $\mathrm{cl}(\mathfrak{N}(\sigma))$. From one of the Corollaries of the Hahn-Banach Theorem (see e.g. Narici [42, Thm. 8.5.4]) there exists $\phi : C(\mathfrak{X}; \mathbb{F}) \to \mathbb{F}$ linear and continuous such that

$$\phi\big|_{\mathrm{cl}(\mathfrak{N}(\sigma))} = 0, \quad \Re(\phi) > 0 \text{ on } \mathcal{U}.$$

Here, $\Re(\phi)$ means the real part of ϕ. In particular, ϕ is not identically zero. Then by Proposition 2.1, there exists a compact subset $K \subset \mathfrak{X}$ and a regular Borel measure (complex or signed) $\mu \neq 0$ on K such that

$$\phi(f) = \int_K f(x)\,\mu(dx), \quad f \in C(\mathfrak{X};\mathbb{F}).$$

In particular, for any $\ell \in \mathfrak{X}', A \in \mathcal{L}(\mathfrak{X}), b \in \mathfrak{X}$ it holds

$$\int_K \langle \ell, \sigma(Ax+b)\rangle\,\mu(dx) = 0.$$

But σ was assumed to be discriminatory. Thus we infer $\mu = 0$, and this is a contradiction to $\Re(\phi) > 0$ on $\mathcal{U}$. We conclude that $\mathfrak{N}(\sigma)$ is dense in $C(\mathfrak{X};\mathbb{F})$ with respect to the topology of compact subsets of $\mathfrak{X}$. This implies that there exits M and $\alpha_m \in \mathbb{F}, \ell_m \in \mathfrak{X}', A_m \in \mathcal{L}(\mathfrak{X})$ and $b_m \in \mathfrak{X}$ for $m = 1, \ldots, M$ such that (2.3) holds. □

Example 2.4. *Because Theorem 2.3 allows us to approximate continuous functions on compact subsets of $\mathfrak{X}$ with neural networks, let us outline a typical example of an infinite dimensional compact subset. First recall that for $\mathfrak{X}$ Banach space, a subset $S \subset \mathfrak{X}$ is compact if and only if (i) S is closed and bounded, (ii) for all $\varepsilon > 0$, there exists a finite dimensional subspace $\mathfrak{X}_\varepsilon \subset \mathfrak{X}$ such that for all $s \in S$, it holds that $d(s, \mathfrak{X}_\varepsilon) < \varepsilon$. Let now $\mathfrak{X}$ be a separable Hilbert space and let $(e_k)_{k\in\mathbb{N}}$ be an orthonormal basis for $\mathfrak{X}$. Then every $x \in \mathfrak{X}$ can be represented as $x = \sum_{k=1}^{\infty} x_k e_k$ with coefficients $x_k \in \mathbb{F}$. Let us choose $(s_k)_{k\in\mathbb{N}} \in \ell^2$ with $s_k \geq 0$ for all $k \in \mathbb{N}$. Here ℓ^2 denotes the space of square integrable sequences. The set*

$$S := \{x \in \mathfrak{X} : |x_k| \leq s_k, \ \ \forall k \in \mathbb{N}\} \tag{3}$$

is then compact. To see this, first observe that S is clearly bounded. Now, let $y \in cl(S)$. Then we may find a sequence $(x(n))_{n\in\mathbb{N}}$ in S such that $x(n)$ converges to y. This in particular means that $x_k(n)$ converges to y_k for all $k \in \mathbb{N}$. But this implies that $|y_k| \leq s_k$ and hence $y \in S$ and S is closed (i.e., (i) holds). Finally, let $\varepsilon > 0$, then choose $N_\varepsilon \in \mathbb{N}$ such that

$$\sum_{k=N_\varepsilon+1}^{\infty} s_k^2 < \varepsilon^2$$

and set $\mathfrak{X}_\varepsilon := span\{e_1, \ldots, e_{N_\varepsilon}\}$, which is clearly finite dimensional. For any $x \in S$ it holds that

$$\left\| x - \sum_{k=1}^{N_\varepsilon} x_k e_k \right\|^2 = \sum_{k=N_\varepsilon+1}^{\infty} x_k^2 < \varepsilon^2,$$

which clearly implies that $d(x, \mathfrak{X}_\varepsilon) \leq \left\| x - \sum_{k=1}^{N_\varepsilon} x_k e_k \right\| < \varepsilon$ and hence (ii) holds.

For the sequel, we need a boundedness assumption on the activation function σ. First, recall that a set $A \subset \mathfrak{X}$ is von Neumann-bounded if for any $k \in \mathbb{N}$ there exists $c_k > 0$ such that $\sup_{x\in A} p_k(x) \leq c_k$. We assume that the set

$$\sigma(\mathfrak{X}) \subset \mathfrak{X} \tag{4}$$

is von Neumann-bounded.

Remark 2.5. *We have another concept of metric-boundedness available: A subset A of a metric space $(\mathfrak{X}, d)$ is bounded if there exists $R > 0$ such that for all $x_1, x_2 \in A$ it holds $d(x_1, x_2) < R$. This concept is not sufficiently stringent, because $diam(\mathfrak{X}) \leq 1$ under the metric defined in (1), and thus any subset of $\mathfrak{X}$ is bounded. von Neumann-boundedness is more well-suited when one works with metrizable topological vector spaces.*

Assuming von Neumann-boundedness is convenient because it enables us to interchange limits and integrals. Observe that in the case in which $\mathfrak{X}$ is normed, we are back to the classical concept of boundedness.

In view of the von Neumann-boundedness assumption on σ, for any $\ell \in \mathfrak{X}', A \in \mathcal{L}(\mathfrak{X}), b \in \mathfrak{X}$

$$|\mathcal{N}_{\ell,A,b}(x)| \leq C_\ell\, p_{k_\ell}(\sigma(Ax+b)), \quad x \in \mathfrak{X}$$

for some constant $C_\ell \geq 0$ (compare Schaefer [47, Thm. 1.1, p. 74]), and thus, for a constant $C(\ell,\sigma)$ depending on ℓ and σ

$$|\mathcal{N}_{\ell,A,b}(x)| \leq C(\ell,\sigma), \quad x \in \mathfrak{X}.$$

We next investigate under which conditions a non-linear function σ is discriminatory. From now on, we assume that $\mathbb{F} = \mathbb{R}$, because we need that hyperplanes disconnect the space $\mathfrak{X}$. If $\mathbb{F} = \mathbb{C}$, this of course, cannot hold.

We now state a condition that ensures that σ is discriminatory. In order to develop some intuition for this condition, first recall that any $\psi \in \mathfrak{X}' \setminus \{0\}$ defines a hyperplane in $\mathfrak{X}$ by the set $\Psi_0 = \ker(\psi)$. This hyperplane splits $\mathfrak{X}$ between the two half-spaces $\Psi_+ = \{x \in \mathfrak{X}; \langle \psi, x\rangle > 0\}$ and $\Psi_- = \{x \in \mathfrak{X}; \langle \psi, x\rangle < 0\}$, which lie on either side of the hyperplane. It turns out that measures on $\mathcal{B}(\mathfrak{X}) \cap K$ are fully determined by their values on the half-spaces arising from all shifted hyperplanes. If now σ splits the space $\mathfrak{X}$ in the sense that there exists one particular hyperplane Ψ_0 such that on either side of this hyperplane, the function $\sigma(\lambda x)$ converges as $\lambda \to \infty$, then this implies that $\sigma(\lambda x)$ converges pointwise to a function that is constant on both half-spaces separated by Ψ_0. Integrating this pointwise limit over either of those spaces determines the value of the measure on them. The maps $A \in \mathcal{L}(\mathfrak{X}), b \in \mathfrak{X}$ now allow to rotate, shift and project to all possible half-spaces and determine the measure on them (see Lemma 2.10).

The following separating property is the infinite-dimensional counterpart to the well known sigmoidal property for functions from $\mathbb{R}$ to $\mathbb{R}$ (see Cybenko [16]):

Definition 2.6. *Separating property: There exist $\psi \in \mathfrak{X}' \setminus \{0\}$ and $u_+, u_-, u_0 \in \mathfrak{X}$ such that either $u_+ \notin \operatorname{span}\{u_0, u_-\}$ or $u_- \notin \operatorname{span}\{u_0, u_+\}$ and such that*

$$\begin{cases} \lim_{\lambda\to\infty} \sigma(\lambda x) = u_+, & \text{if } x \in \Psi_+ \\ \lim_{\lambda\to\infty} \sigma(\lambda x) = u_-, & \text{if } x \in \Psi_- \\ \lim_{\lambda\to\infty} \sigma(\lambda x) = u_0, & \text{if } x \in \Psi_0 \end{cases} \tag{5}$$

where we have set as above

$$\Psi_+ = \{x \in \mathfrak{X}; \langle \psi, x\rangle > 0\}, \quad \Psi_- = \{x \in \mathfrak{X}; \langle \psi, x\rangle < 0\}$$

and $\Psi_0 = \ker(\psi)$.

We point out that as a particular case of the Separating property we may choose $u_0 = u_- = 0$ and $u_+ \neq 0$ for instance. We now provide a first example of a function σ that fulfills the Separating property. It is in the spirit of the classical Sigmoid activation function. More examples are provided in Section 2.1.

Example 2.7. *We are going to give a construction of a continuous and von Neumann-bounded function $\sigma : \mathfrak{X} \to \mathfrak{X}$ satisfying the Separating property in Definition 2.6, for $u_+, u_-, u_0 \in \mathfrak{X}$ such that either $u_+ \notin \operatorname{span}\{u_0, u_-\}$ or $u_- \notin \operatorname{span}\{u_0, u_+\}$.*

Let us recall this abstract result first: given a metric space (Z, d) and $\emptyset \neq Y \subset Z$, define

$$F_\varepsilon(x) := \max(1 - \varepsilon^{-1} d(x, Y), 0), \quad x \in Z,\ \varepsilon > 0.$$

Then F_ε is Lipschitz continuous, $F_\varepsilon \in [0,1]$ and $F_\varepsilon(x) \to I_Y(x)$ for any $x \in Z$ as $\varepsilon \to 0$.

Consider $\psi \in \mathfrak{X}' \setminus \{0\}$ arbitrary. We approximate with this trick the indicator functions $I_{\{\psi \geq 1\}}$, $I_{\{\psi \leq -1\}}$ and $I_{\{\psi = 0\}}$, obtaining respectively $F_{\varepsilon,1}$, $F_{\varepsilon,-1}$ and $F_{\varepsilon,0}$. The scaling parameter ε is chosen small enough such that the supports of these functions do not meet. This is clearly possible. Indeed: suppose first that $d(\{\psi = 1\}, \{\psi = 0\}) = 0$. Then we might find $(z_n, y_n) \in \{\psi = 1\} \times \{\psi = 0\}$ such that $d(z_n, y_n) \to 0$, namely $p_k(z_n - y_n) \to 0$ for any $k \in \mathbb{N}$. But on the other hand, for some $j \in \mathbb{N}$ and $c_j > 0$

$$1 = |\langle \psi, z_n\rangle - \langle \psi, y_n\rangle| \leq c_j p_j(z_n - y_n) \to 0$$

and thus $d(\{\psi = 1\}, \{\psi = 0\}) > 0$. Since $\operatorname{supp} F_{\varepsilon,1} = \operatorname{cl}(\{\psi \geq 1\}_\varepsilon)$ and $\operatorname{supp} F_{\varepsilon,0} = \operatorname{cl}(\{\psi = 0\}_\varepsilon)$ (for an arbitrary subset Y, Y_ε denotes its ε-neighborhood), for $4\varepsilon < d(\{\psi = 1\}, \{\psi = 0\})$ we obtain that the supports do not meet. The same holds for the other cases.

Define

$$\sigma(x) := F_{\varepsilon,1}(x) u_+ + F_{\varepsilon,-1}(x) u_- + F_{\varepsilon,0}(x) u_0, \quad x \in \mathfrak{X}.$$

Then σ is continuous and von Neumann-bounded, because for any $k \in \mathbb{N}$ and $x \in \mathfrak{X}$ we clearly have

$$p_k(\sigma(x)) \leq p_k(u_+) + p_k(u_-) + p_k(u_0),$$

and the condition (5) *is satisfied.*

The following theorem shows that a function σ that satisfies Definition 2.6 is discriminatory, from which the density of $\mathfrak{N}(\sigma)$ follows by Theorem 2.3.

Theorem 2.8. *Let $\mathfrak{X}$ be a real Fréchet space. Let $\sigma : \mathfrak{X} \to \mathfrak{X}$ be continuous, von Neumann-bounded and satisfying the separating property in Definition 2.6 above. Assume that for a given compact subset $K \subset \mathfrak{X}$ and a given regular Borel measure μ on K it holds*

$$\int_K \langle \ell, \sigma(Ax + b)\rangle \, \mu(dx) = 0$$

for all $\ell \in \mathfrak{X}', A \in \mathcal{L}(\mathfrak{X}), b \in \mathfrak{X}$. Then $\mu = 0$.

Before we can prove Theorem 2.8 we need two preparatory lemmas.

Lemma 2.9. *Given $\phi, \psi \in \mathfrak{X}' \setminus \{0\}$ there exists $z \in \mathfrak{X}$ such that $\phi(z) = 1$ and $\psi(z) \neq 0$.*

Proof. Linearity of ϕ implies that the set $\Phi_+ \cup \Phi_-$, where $\Phi_+ = \{x \in \mathfrak{X}; \langle \phi, x\rangle > 0\}$ and $\Phi_- = \{x \in \mathfrak{X}; \langle \phi, x\rangle < 0\}$, is actually dense. To see this, we need to show that each $x \in \Phi_0 = \ker(\phi)$ can be approximated with a sequence in $\Phi_+ \cup \Phi_-$. Consider $u_n = n^{-1}u \in \mathfrak{X}$ with some $u \in \mathfrak{X}$ such that $\phi(u) = 1$ and define $x_n = x + u_n$. Then clearly $x_n \in \Phi_+$ and $x_n \to x$ and hence we get that $\mathrm{cl}(\Phi_+ \cup \Phi_-) = \mathfrak{X}$. Suppose that ψ vanishes on the set $\Phi_+ \cup \Phi_-$. Again by continuity of ψ we would get $\psi = 0$ identically. Therefore, there must exist $w \in \Phi_+ \cup \Phi_-$ such that $\psi(w) \neq 0$. The element $z = w/\phi(w)$ does the job. □

The next lemma is crucial for the proof of Theorem 2.8 as it allows us to rotate, shift and project to all possible half-spaces and show that the measures on them is zero if certain conditions are satisfied.

Lemma 2.10. *Let $\mathfrak{X}$ be a real Fréchet space. Let $\psi \in \mathfrak{X}'$ be not identically zero. Then, for arbitrary $\gamma \in \mathfrak{X}'$, the equation*

$$\gamma = \psi \circ A$$

is solvable for some $A \in \mathcal{L}(\mathfrak{X})$.

Proof. For arbitrary $\phi \in \mathfrak{X}', t \in \mathbb{R}$ we write $\phi_t := \{x \in \mathfrak{X}; \langle \phi, x\rangle = t\}$. Clearly, we can assume γ not identically zero, otherwise the problem is trivial. Therefore, let $z \in \mathfrak{X}$ be such that $\langle \gamma, z\rangle = 1$ and $\langle \psi, z\rangle \neq 0$. Clearly, such z exists in view of Lemma 2.9 above. Moreover, let $w \in \mathfrak{X}$ such that $\langle \psi, w\rangle = 1$.

Let $\Psi_0 = \ker(\psi)$ and $\Gamma_0 = \ker(\gamma)$. We observe that

$$\mathfrak{X} = \Gamma_0 + \langle z\rangle = \Psi_0 + \langle w\rangle \tag{6}$$

where $\langle z\rangle = \{sz; s \in \mathbb{R}\} \subset \mathfrak{X}$ and $\langle w\rangle = \{sw; s \in \mathbb{R}\} \subset \mathfrak{X}$. Furthermore, $\Gamma_0 \cap \langle z\rangle = \{0\}$ and $\Psi_0 \cap \langle w\rangle = \{0\}$, namely Γ_0 and $\langle z\rangle$, are algebraic complements. The same holds for Ψ_0 and $\langle w\rangle$. Furthermore, Γ_0 and Ψ_0 are closed by continuity, and have codimension one. By Schaefer [47, Prop. 3.5., page 22], it follows that Γ_0 and $\langle z\rangle$ (respectively, Ψ_0 and $\langle w\rangle$) are also topologically complemented.

Therefore, any $x \in \mathfrak{X}$ may be written in a unique way as

$$x = x_{\Gamma_0} + \gamma(x)z = x_{\Psi_0} + \psi(x)w,$$

where $x_{\Gamma_0} \in \Gamma_0, x_{\Psi_0} \in \Psi_0$. We can therefore define the following projections operators:

$$\Pi_{\Gamma_0} : \mathfrak{X} \to \Gamma_0, \quad x \mapsto x_{\Gamma_0},$$

$$\Pi_{\langle z\rangle} : \mathfrak{X} \to \langle z\rangle \quad x \mapsto \gamma(x)z,$$

$$\Pi_{\Psi_0} : \mathfrak{X} \to \Psi_0, \quad x \mapsto x_{\Psi_0},$$

$$\Pi_{\langle w\rangle} : \mathfrak{X} \to \langle w\rangle \quad x \mapsto \psi(x)w.$$

Since ψ, γ and the identity operator are continuous, it follows that $\Pi_{\Psi_0}(x) = x - \psi(x)w$, $\Pi_{\Gamma_0}(x) = x - \gamma(x)z$, $\Pi_{\langle z\rangle}$ and $\Pi_{\langle w\rangle}$ are in $\mathcal{L}(\mathfrak{X})$. Define $A_0 := \Pi_{\Psi_0} \circ \Pi_{\Gamma_0} + \Pi_{\langle w\rangle} \circ \Pi_{\langle z\rangle} \in \mathcal{L}(\mathfrak{X})$. Let $x \in \mathfrak{X}$ arbitrary, and write it as $x = x_{\Gamma_0} + \gamma(x)z$. Write $z = z_{\Psi_0} + \psi(z)w$. Then,

$$A_0 x = \Pi_{\Psi_0} x_{\Gamma_0} + \gamma(x)\Pi_{\langle w\rangle} z = \Pi_{\Psi_0} x_{\Gamma_0} + \gamma(x)\psi(z)w,$$

and

$$\psi(A_0 x) = \gamma(x)\psi(z)\psi(w) = \gamma(x)\psi(z), \quad x \in \mathfrak{X}.$$

But $\psi(z) \neq 0$, and thus $A := \psi(z)^{-1} A_0 \in \mathcal{L}(\mathfrak{X})$ does the job. □

We are now ready to prove Theorem 2.8:

Proof of Theorem 2.8. Consider $\lambda > 0$. Then for any $\ell \in \mathfrak{X}', A \in \mathcal{L}(\mathfrak{X}), b \in \mathfrak{X}$ it holds

$$\int_K \langle \ell, \sigma(\lambda(Ax + b))\rangle \, \mu(dx) = 0.$$

Observe that, as $\lambda \to \infty$, pointwise in $x \in \mathfrak{X}$,

$$\langle \ell, \sigma(\lambda(Ax+b))\rangle \to \begin{cases} \langle \ell, u_+\rangle, & \text{if } Ax + b \in \Psi_+ \\ \langle \ell, u_-\rangle, & \text{if } Ax + b \in \Psi_- \\ \langle \ell, u_0\rangle, & \text{if } Ax + b \in \Psi_0 \end{cases}$$

Since, σ is von Neumann-bounded, then there exists a constant $C(\ell, \sigma)$ such that

$$|\langle \ell, \sigma(\lambda(Ax+b))\rangle| \leq C(\ell, \sigma),$$

uniformly in λ and x. By the Hahn-Jordan decomposition (see Bogachev [6, Thm. 3.1.1., Cor. 3.1.2]), we can write the measure $\mu = \mu_1 - \mu_2$ for two positive measures μ_1, μ_2 on K. This implies that

$$\int_K \langle \ell, \sigma(\lambda(Ax+b))\rangle \mu(dx) = \int_K \langle \ell, \sigma(\lambda(Ax+b))\rangle \mu_1(dx) - \int_K \langle \ell, \sigma(\lambda(Ax+b))\rangle \mu_2(dx)$$

Since we are integrating on the compact set K, and μ is a regular Borel measure, constants are integrable with respect to μ on K. The same holds then for μ_1 and μ_2.

Therefore, by Lebesgue's dominated convergence theorem applied to each integrand above, it follows that

$$\langle \ell, u_+\rangle \mu[K \cap A^{-1}(\Psi_+ - b)] + \langle \ell, u_-\rangle \mu[K \cap A^{-1}(\Psi_- - b)] + \langle \ell, u_0\rangle \mu[K \cap A^{-1}(\Psi_0 - b)] = 0 \tag{7}$$

for any $\ell \in \mathfrak{X}', A \in \mathcal{L}(\mathfrak{X}), b \in \mathfrak{X}$.

Let us first assume that $u_+ \notin \operatorname{span}\{u_0, u_-\}$. Then by the Hahn-Banach theorem (see e.g. Conway [14, Chap IV, Cor. 3.15]) we can choose $\ell \in \mathfrak{X}'$ such that $\langle \ell, u_+\rangle = 1$ and $\langle \ell, u_-\rangle = \langle \ell, u_0\rangle = 0$. This leads us to conclude from (7) that

$$\mu[K \cap A^{-1}(\Psi_+ - b)] = 0$$

for all $A \in \mathcal{L}(\mathfrak{X}), b \in \mathfrak{X}$. Let now $t \in \mathbb{R}$ and $b \in \mathfrak{X}$ such that $t = \psi(-b)$. Then, it is immediate to see that

$$\Psi_+ - b = \psi^{-1}(t, \infty)$$

and thus

$$\mu[K \cap (\psi \circ A)^{-1}(t, \infty)] = 0$$

for each $t \in \mathbb{R}$ and $A \in \mathcal{L}(\mathfrak{X})$. By Lemma 2.10, we therefore deduce that

$$\mu[K \cap \gamma^{-1}(t, \infty)] = 0 \tag{8}$$

for each $t \in \mathbb{R}$ and $\gamma \in \mathfrak{X}'$. In the case that $u_- \notin \operatorname{span}\{u_0, u_+\}$ instead, a similar line of reasoning leads to conclude that

$$\mu[K \cap \gamma^{-1}(-\infty, t)] = 0. \tag{9}$$

Observe in particular that $\mu(K) = 0$. For the sake of convenience, we trivially extend μ to the whole $\mathfrak{X}$, namely

$$\mu_{ext}(E) := \mu(K \cap E), \quad E \in \mathcal{B}(\mathfrak{X})$$

and notice that $|\mu_{ext}|(\mathfrak{X}) = |\mu|(K) < \infty$, where $|\mu_{ext}| = \mu_{ext,1} + \mu_{ext,2}$, and $\mu_{ext} = \mu_{ext,1} - \mu_{ext,2}$ is the Hahn-Jordan decomposition for the extended measure ($\mu_{ext,1}$ and $\mu_{ext,2}$ are positive finite measures on $\mathcal{B}(\mathfrak{X})$). Clearly, then it follows from $\mu(K) = 0$ that $\mu_{ext}(\mathfrak{X}) = 0$. Recall also that $\mathcal{B}(K) = \mathcal{B}(\mathfrak{X}) \cap K$.

Because μ is regular Borel measure, it follows in particular that for every $E \subset K$ and $\varepsilon > 0$, there exists compact $K_\varepsilon \subset K$ such that $|\mu|(E \setminus K_\varepsilon) < \varepsilon$. This property extends to $E \in \mathfrak{X}$ for μ_{ext} as we may use that $|\mu_{ext}|(\cdot) = |\mu|(\cdot \cap K)$ and choose $K_\varepsilon \subset E \cap K$ such that $|\mu|((E \cap K) \setminus K_\varepsilon) < \varepsilon$ and it follows that $|\mu_{ext}|(E \setminus K_\varepsilon) = |\mu_{ext}|((E \cap K) \setminus K_\varepsilon) + |\mu_{ext}|(E \cap K^c) = |\mu|((E \cap K) \setminus K_\varepsilon) < \varepsilon$. This shows that μ_{ext} is a Radon measure in the sense of [7, Def. 7.1.1].

Moreover, (8) or (9) is now telling us that $\mu_{ext} = 0$ on $\sigma(\mathfrak{X}') \subset \mathcal{B}(\mathfrak{X})$, the sigma-algebra generated by all the elements of $\mathfrak{X}'$. We want to show that actually $\mu_{ext} = 0$ on $\mathcal{B}(\mathfrak{X})$ as well. We argue by contradiction and assume there exists $E \in \mathcal{B}(\mathfrak{X})$ such that $\mu_{ext}(E) \neq 0$. In virtue of Bogachev [7, Prop. 7.12.1] we may find $B \in \sigma(\mathfrak{X}')$ such that

$$|\mu_{ext}|(E \Delta B) = 0,$$

namely

$$\mu_{ext,i}(E \Delta B) = 0, \quad i = 1, 2.$$

Since $E \Delta B = (E \cup B) \setminus (E \cap B)$ and $\mu_{ext,i}$ are positive finite measures, we infer

$$\mu_{ext,i}(E \cup B) = \mu_{ext,i}(E \cap B), \quad i = 1, 2,$$

which implies, $i = 1, 2$,

$$\begin{cases} \mu_{ext,i}(E) \leq \mu_{ext,i}(E \cup B) = \mu_{ext,i}(E \cap B) \leq \mu_{ext,i}(E) \\ \mu_{ext,i}(B) \leq \mu_{ext,i}(E \cup B) = \mu_{ext,i}(E \cap B) \leq \mu_{ext,i}(B) \end{cases}$$

and finally $\mu_{ext,i}(E) = \mu_{ext,i}(B)$ for $i = 1, 2$. Therefore,

$$0 \neq \mu_{ext}(E) = \mu_{ext_1}(E) - \mu_{ext,2}(E) = \mu_{ext_1}(B) - \mu_{ext,2}(B) = \mu_{ext}(B)$$

and at the same time $\mu_{ext}(B) = 0$, because $B \in \sigma(\mathfrak{X}')$. Thus, it must hold $\mu_{ext} = 0$ on $\mathcal{B}(\mathfrak{X})$, and hence, $\mu = 0$ on $\mathcal{B}(K)$, which concludes the proof.

□

2.1. **Additional examples of functions with Separating property.** We now provide a few more examples of function that satisfy the Separating property Definition 5. The first example resembles the well known rectified linear activation function (ReLU).

Example 2.11. *We consider the following example: let $(\mathfrak{X}, \|\cdot\|)$ be a real Banach space now. Consider $\psi \in \mathfrak{X}'$ with $\|\psi\| = 1$ (the dual norm). For $R > 0$, let B_R denote the open ball of radius R around the origin. First of all we notice that*

$$d(\mathrm{cl}(B_{R+1}); \{x \in \mathfrak{X}; |\langle \psi, x \rangle| \geq R + 2\}) \geq 1.$$

Indeed, given $y : \|y\| \leq R + 1$ and $x : |\langle \psi, x \rangle| \geq R + 2$, it follows that $\|x\| \geq R + 2$ and thus

$$\|y - x\| \geq |\|y\| - \|x\|| \geq 1.$$

In particular these sets are disjoints.

Set $F_0 := \mathfrak{X} \setminus B_{R+1}$ and $F_1 := \mathrm{cl}(B_R)$: these closed sets are disjoint. Since we are in a normal space, Urysohn's lemma ensures that there exists $\mathcal{U} : \mathfrak{X} \to [0, 1]$ continuous such that

$$\mathcal{U}\big|_{F_1} = 1, \quad \mathcal{U}\big|_{F_0} = 0.$$

In particular, since $\{x \in \mathfrak{X}; |\langle \psi, x \rangle| \geq R + 2\} \subset \mathfrak{X} \setminus cl(B_{R+1}) \subset F_0$, $\mathcal{U} = 0$ on $\{x \in \mathfrak{X} : |\langle \psi, x \rangle| \geq R + 2\}$.

Let $I_{\geq}$ and $I_{\leq}$ be the indicator functions of the sets $\{x \in \mathfrak{X} : \langle \psi, x \rangle \geq R + 2\}$ and $\{x \in \mathfrak{X} : \langle \psi, x \rangle \leq -R - 2\}$ respectively. And let $I^\varepsilon_{\geq}$ and $I^\varepsilon_{\leq}$ be their Lipschitz approximations, as in Example 2.7. Since, with the same notation as above, it holds

$$\mathrm{supp}\, I^\varepsilon_{\geq} = \mathrm{cl}(\{\psi \geq R + 2\}_\varepsilon), \quad \mathrm{supp}\, I^\varepsilon_{\leq} = \mathrm{cl}(\{\psi \leq -R - 2\}_\varepsilon),$$

elementary computations show that

$$\mathrm{supp}\, I^\varepsilon_{\geq} \subset \mathfrak{X} \setminus B_{R+2-\varepsilon}, \quad \mathrm{supp}\, I^\varepsilon_{\leq} \subset \mathfrak{X} \setminus B_{R+2-\varepsilon}$$

and thus for $\varepsilon < 1$

$$\mathrm{supp}\, I^\varepsilon_{\geq} \cap \mathrm{cl}(B_{R+1}) = \emptyset, \quad \mathrm{supp}\, I^\varepsilon_{\leq} \cap \mathrm{cl}(B_{R+1}) = \emptyset.$$

We can also easily get that

$$\operatorname{supp} I^{\varepsilon}_{\geq} \subset \{\psi \geq R + 2 - \varepsilon\}, \quad \operatorname{supp} I^{\varepsilon}_{\leq} \subset \{\psi \leq -R - 2 + \varepsilon\},$$

showing that $\operatorname{supp} I^{\varepsilon}_{\geq} \cap \operatorname{supp} I^{\varepsilon}_{\leq} = \emptyset$.

We choose linearly independent vectors $u_{\geq}$ *and* $u_{\leq}$ *and define*

$$\sigma(x) := \mathcal{U}(x)\, x + I^{\varepsilon}_{\geq}(x)\, u_{\geq} + I^{\varepsilon}_{\leq}(x)\, u_{\leq}, \quad x \in \mathfrak{X}.$$

Then $\sigma \in C(\mathfrak{X};\mathfrak{X})$*, and it is bounded because*

$$\|\sigma(x)\| \leq R + 1 + \|u_{\geq}\| + \|u_{\leq}\|\,, \quad x \in \mathfrak{X}.$$

Clearly, $\sigma(x) = x$ *if* $\|x\| \leq R$.

Moreover, for $x \in \mathfrak{X}$ *such that* $\langle \psi, x\rangle > 0$*, then for all* $\lambda \geq \langle \psi, x\rangle^{-1}(R+2)$ *we have* $\sigma(\lambda x) = u_{\geq}$*. Similarly, for* $x \in \mathfrak{X}$ *such that* $\langle \psi, x\rangle < 0$*, then for all* $\lambda \geq \langle \psi, x\rangle^{-1}(-R-2)$ *we have* $\sigma(\lambda x) = u_{\leq}$*. Finally, if* $\langle \psi, x\rangle = 0$*, then for any* $\lambda > 0$ *we have* $\langle \psi, \lambda x\rangle = 0$*. Thus* $\lambda x \notin \operatorname{supp} I^{\varepsilon}_{\geq} \cup \operatorname{supp} I^{\varepsilon}_{\leq}$ *and so*

$$\sigma(\lambda x) = \mathcal{U}(\lambda x)\lambda x.$$

If $x = 0$*, then* $\sigma(\lambda x) = \sigma(0) = 0$*. If* $x \neq 0$*, then for all* λ *larger than* $\|x\|^{-1}(R+1)$ *it holds* $\sigma(\lambda x) = 0$.

This shows that σ *satisfies* (5).

Example 2.12. *Let us give some further concrete applications of our abstract framework. Let now for the sake of simplicity* $\mathfrak{X}$ *be a real separable Hilbert space with inner product denoted by* $\langle \cdot, \cdot \rangle$ *and corresponding norm by* $\|\cdot\|$*. Further, we denote by* $(e_k)_k$ *an orthonormal basis for* $\mathfrak{X}$*. Any* $x \in \mathfrak{X}$ *may be uniquely written as* $x = \sum_{k\in\mathbb{N}} x_k e_k$*, where* $x_k = \langle e_k, x\rangle$.

Consider $\beta_i \in C(\mathbb{R};\mathbb{R}), i = 1, 2, 3$ *such that*

$$\begin{cases} \lim_{\xi\to\infty} \beta_1(\xi) = 1,\ \lim_{\xi\to-\infty} \beta_1(\xi) = -1,\ \beta_1(0) = 0, \\ \lim_{\xi\to\infty} \beta_2(\xi) = 1,\ \lim_{\xi\to-\infty} \beta_2(\xi) = 1,\ \beta_2(0) = 1, \\ \lim_{\xi\to\infty} \beta_3(\xi) = -1,\ \lim_{\xi\to-\infty} \beta_3(\xi) = 2,\ \beta_3(0) = 0, \end{cases}$$

and define

$$\sigma(x) = \beta_1(x_1)e_1 + \beta_2(x_2)e_2 + \beta_3(x_1)e_3, \quad x \in \mathfrak{X}.$$

Evidently, $\sigma \in C(\mathfrak{X};\mathfrak{X})$*; besides, since* $\|\sigma(x)\|^2 = \beta_1^2(x_1) + \beta_2(x_2)^2 + \beta_3(x_1)^2$*, it holds* $\sup_x \|\sigma(x)\| < \infty$*, because* β_1, β_2 *and* β_3 *are bounded. Thus* σ *is von Neumann-bounded. Consider now the linear bounded functional*

$$\psi(x) := \langle e_1, x\rangle = x_1, \quad x \in \mathfrak{X}.$$

Clearly, $\Psi_+ = \{x \in \mathfrak{X}; x_1 > 0\}, \Psi_- = \{x \in \mathfrak{X}; x_1 < 0\}$ *and* $\Psi_0 = \{x \in \mathfrak{X}; x_1 = 0\}$ *and, as* $\lambda \to \infty$

$$\sigma(\lambda x) \to \begin{cases} e_1 + e_2 - e_3, \ \textit{if } x \in \Psi_+ \\ -e_1 + e_2 + 2e_3, \ \textit{if } x \in \Psi_- \\ e_2, \ \textit{if } x \in \Psi_0 \end{cases}$$

which are linearly independent. We can therefore apply our results to infer that $\mathfrak{N}(\sigma)$ *is dense in* $C(\mathfrak{X};\mathbb{R})$ *with respect to the topology of uniform convergence on the compact subsets of* $\mathfrak{X}$.

We can even go further. By the comment after Definition 2.6 indeed it is enough to consider a function $\beta \in C(\mathbb{R};\mathbb{R})$ *such that*

$$\lim_{\xi\to\infty} \beta(\xi) = 1,\ \lim_{\xi\to-\infty} \beta(\xi) = 0,\ \beta(0) = 0,$$

and arbitrary $z \in \mathfrak{X}$ *in order to define*

$$\sigma(x) = \beta(\psi(x))z = \beta(x_1)z, \quad x \in \mathfrak{X}$$

which still enables us to conclude that $\mathfrak{N}(\sigma)$ *is dense in* $C(\mathfrak{X};\mathbb{R})$*. Example 4.4 below extends this example for more general choices of* ψ*. A natural question now would be to find "optimal"* β *and* z *such that the convergence of the approximation to the function we want to learn is "fast".*

Example 2.13. *The above example can be extended to an activation function that operates on infinitely many different directions $z_j \in \mathfrak{X}$. More precisely, let now $\mathfrak{X}$ be a real Banach space with norm denoted by $\|\cdot\|$. As above, we consider an arbitrary $\psi \in \mathfrak{X}' \setminus \{0\}$. Moreover, suppose we have a sequence $(\beta_j)_{j\in\mathbb{N}} \subset C(\mathbb{R};\mathbb{R})$ such that*

$$\lim_{\xi\to\infty} \beta_j(\xi) = 1, \lim_{\xi\to-\infty} \beta_j(\xi) = 0, \ \beta_j(0) = 0, \ j \in \mathbb{N}$$

and $\sup_j \|\beta_j\|_\infty =: B < \infty$.

Let $(z_j)_{j\in\mathbb{N}} \subset \mathfrak{X}$ be such that $Z := \sum_{j=1}^{\infty} \|z_j\| < \infty$. Set

$$z := \sum_{j=1}^{\infty} z_j \in \mathfrak{X}$$

and assume $z \neq 0$.

We show that the map $\mathfrak{X} \ni x \mapsto \sigma(x) := \sum_{j=1}^{\infty} \beta_j(\psi(x))z_j$ is an activation function.

(1) *Well-defined: since it holds*

$$\sum_{j=1}^{\infty} \|\beta_j(\psi(x))z_j\| = \sum_{j=1}^{\infty} |\beta_j(\psi(x))| \, \|z_j\| \leq \sum_{j=1}^{\infty} \|\beta_j\|_\infty \|z_j\| \leq BZ$$

we have absolute convergence and so $\sigma(x)$ is well-defined.

(2) *Boundedness: $\|\sigma(x)\| \leq \sum_{j=1}^{\infty} \|\beta_j(\psi(x))z_j\| \leq BZ$ for any $x \in \mathfrak{X}$.*

(3) *Continuity: we have*

$$\left\| \sigma(x) - \sum_{j=1}^{N} \beta_j(\psi(x))z_j \right\| = \left\| \sum_{j=N+1}^{\infty} \beta_j(\psi(x))z_j \right\| \leq B \sum_{j=N+1}^{\infty} \|z_j\| ,$$

and thus

$$\sup_{x\in\mathfrak{X}} \left\| \sigma(x) - \sum_{j=1}^{N} \beta_j(\psi(x))z_j \right\| \leq B \sum_{j=N+1}^{\infty} \|z_j\| \to 0$$

as $N \to \infty$, namely the convergence is uniform. Since $x \mapsto \sum_{j=1}^{N} \beta_j(\psi(x))z_j$ is continuous, σ must be continuous as well.

(4) *Separating property: Let $\lambda > 0$. Consider first $x \in \Psi_+$. From the computations just done, we have*

$$\begin{aligned}\|\sigma(\lambda x) - z\| &\leq \left\| \sigma(\lambda x) - \sum_{j=1}^{N} \beta_j(\psi(\lambda x))z_j \right\| + \left\| \sum_{j=1}^{N} \beta_j(\psi(\lambda x))z_j - z \right\| \\ &\leq B \sum_{j=N+1}^{\infty} \|z_j\| + \left\| \sum_{j=1}^{N} \beta_j(\psi(\lambda x))z_j - z \right\| .\end{aligned}$$

Fix $\varepsilon > 0$ and chose $N_\varepsilon \in \mathbb{N}$ such that if $N \geq N_\varepsilon$ it holds $\sum_{j=N+1}^{\infty} \|z_j\| \leq \frac{\varepsilon}{B}$. For such N we have:

$$\|\sigma(\lambda x) - z\| \leq \varepsilon + \left\| \sum_{j=1}^{N} \beta_j(\psi(\lambda x))z_j - z \right\|$$

and thus

$$\limsup_{\lambda\to\infty} \|\sigma(\lambda x) - z\| \leq \varepsilon + \left\| \sum_{j=1}^{N} z_j - z \right\|$$

because evidently as $\lambda \to \infty$

$$\sum_{j=1}^{N} \beta_j(\psi(\lambda x))z_j - z \xrightarrow{\mathfrak{X}} \sum_{j=1}^{N} z_j - z.$$

Hence

$$\limsup_{\lambda\to\infty} \|\sigma(\lambda x) - z\| \le \varepsilon + \left\| \sum_{j=N+1}^{\infty} z_j \right\| \le \varepsilon + \frac{\varepsilon}{B}$$

and by the arbitrariness of ε

$$\lim_{\lambda\to\infty} \|\sigma(\lambda x) - z\| = \limsup_{\lambda\to\infty} \|\sigma(\lambda x) - z\| = 0$$

i.e. $\sigma(\lambda x) \to z$ *as* $\lambda \to \infty$, *if* $x \in \Psi_+$.

The cases $x \in \Psi_-$ *and* $x \in \Psi_0$ *are treated similarly (with* $z = 0$ *now).*

Example 2.14. *In view of the previous example, we further expand on the idea of an activation function operating on each coordinate. Let* $\mathfrak{X}$ *be a separable Hilbert space with an orthonormal basis* $(e_k)_{k\in\mathbb{N}}$ *and inner product naturally denoted* $\langle\cdot,\cdot\rangle$. *For* $x \in \mathfrak{X}$, *we define the activation function as*

$$\sigma(x) = \sum_{k=1}^{\infty} \hat{\sigma}(x_k) e_k$$

where $x_k := \langle x, e_k\rangle$ *and* $\hat{\sigma} : \mathbb{R} \to \mathbb{R}$. *For a linear operator* $A \in L(\mathfrak{X})$, *we can introduce a family of linear functionals* $\rho_k \in \mathfrak{X}'$ *by*

$$\rho_k(x) = \langle Ax, e_k\rangle$$

to obtain

$$Ax + b = \sum_{k=1}^{\infty} (\rho_k(x) + b_k) e_k$$

with $b_k := \langle b, e_k\rangle$. *But then a neuron becomes, with* $\ell = \sum_{k=1}^{\infty} \ell_k e_k \in \mathfrak{X}$,

$$\langle \ell, \sigma(Ax + b)\rangle = \sum_{k=1}^{\infty} \ell_k \hat{\sigma}(\rho_k(x) + b_k) \tag{10}$$

We remark that the representation on the right-hand side above links to infinite wide neural networks. Williams [51] *proposes and studies such networks using weighted integral representations of the infinite layer to encode the sum, and relates such networks to Gaussian processes (see also Cho and Saul* [13]*). As* ℓ *defines a linear functional, we can represent it as an integral operator rather than a sum which shows that our definition of neural networks is a generalisation of this class. Infinitely wide neural networks are based on the approximation results of Hornik* [27].

Observe that we must require $\hat{\sigma}(0) = 0$, *otherwise* $\sigma(0) = \hat{\sigma}(0) \sum_{k=1}^{\infty} e_k \notin \mathfrak{X}$. *Moreover, if* $\hat{\sigma}$ *is Lipschitz continuous, it follows readily that* σ *becomes Lipschitz continuous. We have that,*

$$|\hat{\sigma}(x_k)| = |\hat{\sigma}(x_k) - \hat{\sigma}(0)| \le K|x_k|$$

and therefore $\sigma(x) \in \mathfrak{X}$ *as*

$$\sum_{k=1}^{\infty} \hat{\sigma}^2(x_k) \le K^2 \sum_{k=1}^{\infty} x_k^2 < \infty. \tag{11}$$

To stay within the framework developed in this paper, we also need to have a bounded activation function. However, in the infinite dimensional setting this does not come for free. In light of (11) *one could ask for an activation function* $\hat{\sigma}$ *which is bounded and goes sufficiently fast to zero around the origin. However, let* $\hat{\sigma} = 0$ *on* $[-\varepsilon, \varepsilon]$ *with* $0 < \varepsilon < 1$ *say. Then, for* $x \in \mathfrak{X}$,

$$|\sigma(x)|^2 = \sum_{k:\,|x_k|>\varepsilon} |\hat{\sigma}(x_k)|^2$$

If now $x = \sum_{k=1}^{N} e_k$, *then*

$$|\sigma(x)|^2 = N\ |\hat{\sigma}(1)|^2$$

which blows up when N *grows. It is an interesting question to generalise our activation functions to go beyond boundedness and allow for linear or polynomial growth, say.*

3. Approximation for general codomain

In this section we are going to show that our results can be extended to functions $f \in C(\mathfrak{X};\mathfrak{Y})$ where $(\mathfrak{Y}, \|\cdot\|_{\mathfrak{Y}})$ is an $\mathbb{F}$-Banach space.

As a first step, we need the following simple lemma, which enables us to approximate with our neural network continuous functions from $\mathfrak{X}$ into $\mathbb{F}^d, d \in \mathbb{N}$:

Lemma 3.1. *Let $\mathfrak{X}$ be an $\mathbb{F}$-Fréchet space, and let $\sigma : \mathfrak{X} \to \mathfrak{X}$ be continuous and discriminatory. Then, given $f \in C(\mathfrak{X};\mathbb{F}^d)$, a compact subset K of $\mathfrak{X}$, and $\varepsilon > 0$, there exist $\mathcal{N}^i = \sum_{m=1}^M \alpha_m^i \mathcal{N}_{\ell_m^i, A_m^i, b_m^i} \in \mathfrak{N}(\sigma)$, $i = 1, \dots, d$, with suitable $\alpha_m^i \in \mathbb{F}, \ell_m^i \in \mathfrak{X}', A_m^i \in \mathcal{L}(\mathfrak{X})$ and $b_m^i \in \mathfrak{X}$ such that*

$$\sup_{x\in K} \left\| f(x) - (\mathcal{N}^1(x), \dots, \mathcal{N}^d(x)) \right\|_{\mathbb{F}^d} < \varepsilon$$

where for all $\xi \in \mathbb{F}^d$ we have $\|\xi\|_{\mathbb{F}^d} = \sum_{i=1}^d |\xi^i|$.

Proof. We write $f = (f^1, \dots, f^d)$ with $f^i \in C(\mathfrak{X};\mathbb{F}), i = 1, \dots, d$. Given $K \subset \mathfrak{X}$ and $\varepsilon > 0$, Theorem 2.3 guarantees the existence of $\mathcal{N}^i \in \mathfrak{N}(\sigma)$ such that

$$\sup_{x\in K} \left| f^i(x) - \mathcal{N}^i(x) \right| < \varepsilon/d$$

and we are done. □

We are now ready to prove the following:

Theorem 3.2. *Let $\mathfrak{X}$ be an $\mathbb{F}$-Fréchet space, and let $\sigma : \mathfrak{X} \to \mathfrak{X}$ be continuous and discriminatory. Let $(\mathfrak{Y}, \|\cdot\|_{\mathfrak{Y}})$ be an $\mathbb{F}$-Banach space. Then, given $f \in C(\mathfrak{X};\mathfrak{Y})$, a compact subset K of $\mathfrak{X}$, and $\varepsilon > 0$, there exist $d \in \mathbb{N}$, $v_1, \dots, v_d$ linear independent unit vectors of $\mathfrak{Y}$, $\mathcal{N}^1, \dots \mathcal{N}^d \in \mathfrak{N}(\sigma)$, such that, by defining*

$$\mathcal{N}(x) := \sum_{i=1}^d \mathcal{N}^i(x) v_i, \quad x \in \mathfrak{X},$$

it holds

$$\sup_{x\in K} \|f(x) - \mathcal{N}(x)\|_{\mathfrak{Y}} < \varepsilon.$$

Proof. We recall the following general approximation result (see for example Brezis [8, Ch. 6.1]): given a topological space (Z, τ), an $\mathbb{F}$-Banach space $(\mathfrak{Y}, \|\cdot\|_{\mathfrak{Y}})$ and a continuous map

$$T : Z \to \mathfrak{Y}$$

such that $T(Z)$ is relatively compact in Y, then, given $\varepsilon > 0$ there exists $T_\varepsilon : Z \to \mathfrak{Y}$ continuous, with $T_\varepsilon(Z)$ contained in a finite-dimensional subspace of $\mathfrak{Y}$, and such that

$$\|T_\varepsilon(z) - T(z)\|_{\mathfrak{Y}} < \varepsilon, \quad z \in Z.$$

To apply this result in our present setting, we first restrict f to K

$$f\big|_K : K \to \mathfrak{Y},$$

obtaining a continuous function whose range is compact in $\mathfrak{Y}$. Therefore, we may find $f_\varepsilon : K \to \mathfrak{Y}$ continuous and such that

(1) $f_\varepsilon(K) \subset \operatorname{span}\{v_1, \dots v_d\} \subset \mathfrak{Y}$ for suitable linear independent elements $v_1, \dots, v_d$, whose norm we assume to be equal to 1.
(2) $\sup_{x\in K} \|f(x) - f_\varepsilon(x)\|_{\mathfrak{Y}} = \sup_{x\in K} \left\| f\big|_K(x) - f_\varepsilon(x) \right\|_{\mathfrak{Y}} < \varepsilon/2$.

We set for convenience $V = \operatorname{span}\{v_1, \dots v_d\}$, and we write f_ε as

$$f_\varepsilon(x) = \sum_{i=1}^d f_\varepsilon^i(x) v_i, \quad x \in K$$

with suitable $f_\varepsilon^i \in C(K;\mathbb{F})$, $i = 1, \dots, d$. Being $\mathfrak{X}$ metrizable, it is clearly normal. Therefore, by the Tietze extension theorem (since K is closed), there exist $g_\varepsilon^i \in C(\mathfrak{X};\mathbb{F})$ extensions of f_ε^i, $i = 1, \dots, d$.

We define $g_\varepsilon(x) := \sum_{i=1}^d g_\varepsilon^i(x) v_i$, $x \in \mathfrak{X}$. Then $g_\varepsilon \in C(\mathfrak{X};\mathfrak{Y})$, $g_\varepsilon(\mathfrak{X}) \subset V$ and

$$\sup_{x\in K} \|f(x) - g_\varepsilon(x)\|_{\mathfrak{Y}} < \varepsilon/2.$$

By Lemma 3.1 we may approximate on K

$$\mathfrak{X} \ni x \mapsto (g_\varepsilon^1(x), \dots, g_\varepsilon^d(x)) \in \mathbb{F}^d$$

with $(\mathcal{N}^1, \dots, \mathcal{N}^d)$ such that

$$\sup_{x\in K} \left\|(g_\varepsilon^1(x), \dots, g_\varepsilon^d(x)) - (\mathcal{N}^1(x), \dots, \mathcal{N}^d(x))\right\|_{\mathbb{F}^d} < \varepsilon/2.$$

We define

$$\mathcal{N}(x) := \sum_{i=1}^d \mathcal{N}^i(x) v_i, \quad x \in \mathfrak{X},$$

which has the required property, since we have

$$\begin{aligned}
\sup_{x\in K} \|f(x) - \mathcal{N}(x)\|_{\mathfrak{Y}} &\le \sup_{x\in K} \|f(x) - g_\varepsilon(x)\|_{\mathfrak{Y}} + \sup_{x\in K} \|g_\varepsilon(x) - \mathcal{N}(x)\|_{\mathfrak{Y}} \\
&< \varepsilon/2 + \sup_{x\in K} \sum_{i=1}^d \left|g_\varepsilon^i(x) - \mathcal{N}^i(x)\right| \|v_i\|_{\mathfrak{Y}} \\
&= \varepsilon/2 + \sup_{x\in K} \sum_{i=1}^d \left|g_\varepsilon^i(x) - \mathcal{N}^i(x)\right| \\
&= \varepsilon/2 + \sup_{x\in K} \left\|(g_\varepsilon^1(x), \dots, g_\varepsilon^d(x)) - (\mathcal{N}^1(x), \dots, \mathcal{N}^d(x))\right\|_{\mathbb{F}^d} \\
&< \varepsilon.
\end{aligned}$$

□

4. Approximation with finite dimensional neural networks

In this section we prove a result that ensures that one can approximate a given abstract neural net arbitrary well via a neural network that is constructed from finite dimensional maps and can thus be trained. Of course, this can only work if we can approximate any given $x \in \mathfrak{X}$ sufficiently well with a finite dimensional quantity as otherwise we could not even represent x in a computer. It is therefore plausible that we can derive such results only if some kind of approximation property holds on $\mathfrak{X}$. This approximation property must ensure that one can approximate the identity map on $\mathfrak{X}$ by continuous linear maps of finite rank, uniformly on some subset $K \subset \mathfrak{X}$ of interest. In spaces with a countable Schauder basis $(e_n)_{n\in\mathbb{N}}$, the approximating linear maps are usually the projections $\Pi_N : \mathfrak{X} \to \operatorname{span}\{e_1, \dots, e_N\}$. Unfortunately, not every Fréchet space has a Schauder basis as shown by Enflo [18]. We refer the reader to Schaefer [47, Ch. III, Sec. 9] for a discussion of the approximation property and existence of a Schauder basis for Fréchet space, which was an open problem until answered in [18]. Whenever the space $\mathfrak{X}$ has a Schauder basis, however, we can actually derive an approximation of our abstract neural network with a trainable finite dimensional neural network as we shall see in this section.

To start, we are first going to work in a Banach space setting. Let therefore $\mathfrak{X}$ be a real separable Banach space with norm denoted by $\|\cdot\|$ that admits a normalized Schauder basis $(e_k)_{k\in\mathbb{N}}$, namely each $x \in \mathfrak{X}$ has a unique representation $x = \sum_{k=1}^\infty x_k e_k$ and $\|e_k\| = 1$ for all k. It follows as in Schaefer [47, Thm. 9.6, p. 115] that

$$\Pi_N : \mathfrak{X} \to \operatorname{span}\{e_1, \dots, e_N\}, \quad x \mapsto \sum_{k=1}^N x_k e_k, \quad N \in \mathbb{N}$$

is linear and bounded with $\sup_{N\in\mathbb{N}} \|\Pi_N\|_{op} \le C$ for some suitable constant $C \ge 1$, and that for any $K \subset \mathfrak{X}$ compact we have $\sup_{x\in K} \|x - \Pi_N x\| \to 0$ as $N \to \infty$.

While we know by [18] that there exist Banach spaces without a Schauder basis, it is also true that "all usual separable Banach spaces of Analysis admit a Schauder basis" (see Brezis [8]). For example for the Banach spaces $L^p(\mathbb{R}^n)$, where $1 \le p < \infty$, as well as for the

Sobolev and Besov spaces, a basis is given by wavelets (see Triebel [50]). See Heil [25] for many more examples.

We assume now that the activation function $\sigma : \mathfrak{X} \to \mathfrak{X}$ is Lipschitz, namely

$$\|\sigma(x) - \sigma(y)\| \leq \mathrm{Lip}(\sigma) \|x - y\|, \quad x, y \in \mathfrak{X}. \tag{12}$$

where $0 \leq \mathrm{Lip}(\sigma) < \infty$. Of course since $\mathfrak{X}$ is already a metric space, we do not use the metric d defined in (1), but the one implied by the norm, i.e. $d(x_1, x_2) = \|x_1 - x_2\|$.

Observe also that the activation functions in Example 2.12 become Lipschitz as soon as we impose that the β_i's are Lipschitz. The activation function in Example 2.7 is already Lipschitz, as soon as $\mathfrak{X}$ is assumed to be a normed space, as we are doing here. Therefore, this condition does not seem very restrictive.

We are ready to prove:

Proposition 4.1. *Let $\mathfrak{X}$ be a real separable Banach space that admits a normalized Schauder basis $(e_k)_{k\in\mathbb{N}}$ and let σ be Lipschitz. Let $f \in C(\mathfrak{X}; \mathbb{R})$, $K \subset \mathfrak{X}$ compact and $\varepsilon > 0$. Assume*

$$\mathcal{N}^\epsilon(x) = \sum_{j=1}^{M} \langle \ell_j, \sigma(A_j x + b_j) \rangle, \quad x \in \mathfrak{X}$$

with $\ell_j \in \mathfrak{X}'$, $A_j \in \mathcal{L}(\mathfrak{X})$ and $b_j \in \mathfrak{X}$ such that

$$\sup_{x\in K} |f(x) - \mathcal{N}^\epsilon(x)| < \varepsilon.$$

Fix $\delta > 0$. Then there exists $N_* = N_*(\mathcal{N}^\epsilon, \delta) \in \mathbb{N}$ such that for $N \geq N_*$

$$\sup_{x\in K} \left| f(x) - \sum_{j=1}^{M} \langle \ell_j \circ \Pi_N, \sigma(\Pi_N A_j \Pi_N x + \Pi_N b_j) \rangle \right| < \varepsilon + \delta. \tag{13}$$

Proof. For $j = 1, \dots, M$, $N \in \mathbb{N}$ and $x \in K$ we indeed have

$$\begin{aligned}
&|\langle \ell_j, \sigma(A_j x + b_j)\rangle - \langle \ell_j \circ \Pi_N, \sigma(\Pi_N A_j \Pi_N x + \Pi_N b_j)\rangle| \\
&\quad \leq |\langle \ell_j, \sigma(A_j x + b_j) - \Pi_N \sigma(A_j x + b_j)\rangle| \\
&\qquad + |\langle \ell_j, \Pi_N \sigma(A_j x + b_j) - \Pi_N \sigma(\Pi_N A_j \Pi_N x + \Pi_N b_j)\rangle| \\
&\quad \leq \|\ell_j\| \, \|\sigma(A_j x + b_j) - \Pi_N \sigma(A_j x + b_j)\| \\
&\qquad + \|\ell_j\| \, C \, \|\sigma(A_j x + b_j) - \sigma(\Pi_N A_j \Pi_N x + \Pi_N b_j)\|,
\end{aligned}$$

where in the last line we have used that $\sup_{N\in\mathbb{N}} \|\Pi_N\|_{op} \leq C$. Thus, as far as it concerns the second term, it holds

$$\begin{aligned}
&\|\ell_j\| \, C \, \|\sigma(A_j x + b_j) - \sigma(\Pi_N A_j \Pi_N x + \Pi_N b_j)\| \\
&\quad \leq C \|\ell_j\| \, \mathrm{Lip}(\sigma) \, \|A_j x + b_j - \Pi_N A_j \Pi_N x - \Pi_N b_j\| \\
&\quad \leq C \|\ell_j\| \, \mathrm{Lip}(\sigma) \, \{\|A_j x - \Pi_N A_j x\| + \|\Pi_N A_j x - \Pi_N A_j \Pi_N x\| + \|b_j - \Pi_N b_j\|\} \\
&\quad \leq C \|\ell_j\| \, \mathrm{Lip}(\sigma) \, \{\|A_j x - \Pi_N A_j x\| + C \|A_j x - A_j \Pi_N x\| + \|b_j - \Pi_N b_j\|\} \\
&\quad \leq C \|\ell_j\| \, \mathrm{Lip}(\sigma) \left\{\|A_j x - \Pi_N A_j x\| + C \|A_j\|_{op} \|x - \Pi_N x\| + \|b_j - \Pi_N b_j\|\right\} \\
&\quad \leq C \|\ell_j\| \, \mathrm{Lip}(\sigma) \left\{\sup_{x\in K} \|A_j x - \Pi_N A_j x\| \right. \\
&\qquad\qquad \left. + C \|A_j\|_{op} \sup_{x\in K} \|x - \Pi_N x\| + \|b_j - \Pi_N b_j\| \right\} \\
&\quad = C \|\ell_j\| \, \mathrm{Lip}(\sigma) \left\{\sup_{y\in A_j K} \|y - \Pi_N y\| \right. \\
&\qquad\qquad \left. + C \|A_j\|_{op} \sup_{x\in K} \|x - \Pi_N x\| + \|b_j - \Pi_N b_j\| \right\}.
\end{aligned}$$

Setting for convenience $\sigma_j := \sigma(A_jK + b_j)$, and noticing that it is compact, we eventually arrive at

$$\begin{aligned}
&|\langle \ell_j, \sigma(A_jx + b_j)\rangle - \langle \ell_j \circ \Pi_N, \sigma(\Pi_N A_j \Pi_N x + \Pi_N b_j)\rangle| \\
&\quad \leq \|\ell_j\| \operatorname{Lip}(\sigma) \Bigg\{ \sup_{y \in A_jK} \|y - \Pi_N y\| \\
&\qquad\qquad + C \|A_j\|_{op} \sup_{x \in K} \|x - \Pi_N x\| + \|b_j - \Pi_N b_j\| \Bigg\} \\
&\quad + \|\ell_j\| \sup_{y \in \sigma_j} \|y - \Pi_N y\|
\end{aligned}$$

Observe that $A_jK \subset \mathfrak{X}$ is compact. By the approximation property provided by the Schauder basis $(e_k)_{k\in\mathbb{N}}$, we may find $N(j) \in \mathbb{N}$ such that:

$$\begin{cases}
\sup_{y \in A_jK} \|y - \Pi_N y\| < \frac{\delta}{4M\|\ell_j\| \operatorname{Lip}(\sigma)} \\
\sup_{y \in \sigma_j} \|y - \Pi_N y\| < \frac{\delta}{4M\|\ell_j\|} \\
\sup_{x \in K} \|x - \Pi_N x\| < \frac{\delta}{4M\|\ell_j\| C \|A_j\|_{op} \operatorname{Lip}(\sigma)}, \quad \text{if } \|A_j\|_{op} \neq 0 \\
\|b_j - \Pi_N b_j\| < \frac{\delta}{4M\|\ell_j\| \operatorname{Lip}(\sigma)}
\end{cases}$$

for all $N \geq N(j)$. With this choice, we then have

$$\sup_{x \in K} |\langle \ell_j, \sigma(A_jx + b_j)\rangle - \langle \ell_j \circ \Pi_N, \sigma(\Pi_N A_j \Pi_N x + \Pi_N b_j)\rangle| < \delta/M.$$

Therefore, setting $N_* := \max\{N(1), \dots, N(M)\}$, we conclude that for all $N \geq N_*$

$$\sup_{x \in K} \left| f(x) - \sum_{j=1}^{M} \langle \ell_j \circ \Pi_N, \sigma(\Pi_N A_j \Pi_N x + \Pi_N b_j)\rangle \right| < \varepsilon + \delta.$$

□

We mention that the function $\mathcal{N}^\varepsilon : \mathfrak{X} \to \mathbb{R}$, which is required in the proposition above, exists for instance in view of Theorem 2.3, as soon as we assume that σ is discriminatory.

Remark 4.2. *The terms appearing in the sum in (13) can now easily be programmed in a computer. We see that for large N, it is sufficient to consider the finite dimensional input values $\Pi_N(x)$ instead of x, and then successively the restriction of the operators $\Pi_N A_j, \sigma$ and ℓ_j to $\operatorname{span}\{e_1, \dots, e_N\}$ instead of the maps A_j, σ and ℓ_j for $j = 1, \dots, M$. The maps $\Pi_N A_j, \sigma$ and ℓ_j are finite dimensional when restricted to $\operatorname{span}\{e_1, \dots, e_N\}$ and the sum above thus resembles a classical neural network. However, instead of the typical one dimensional activation function, the function $\Pi_N \circ \sigma$ restricted to $\operatorname{span}\{e_1, \dots, e_N\}$ is multidimensional.*

With an extra effort it is possible to generalize this result to real separable Fréchet spaces that admit Schauder basis. Examples include for instance the Schwartz space of rapidly decreasing functions, for which a basis is given in terms of Hermite functions (see Schwartz [48]) and the Hida test function and distribution space (see Holden *et al.* [26, Def 2.3.2.]).

Let us now see how to do this generalization. Following Meise and Vogt [39, 28.10, p. 331], a Schauder basis for a real separable Fréchet space is a sequence $(e_k)_{k\in\mathbb{N}} \subset \mathfrak{X}$, such that each $x \in \mathfrak{X}$ has a unique representation $x = \sum_{k=1}^{\infty} x_k e_k$. As above, we define

$$\Pi_N : \mathfrak{X} \to \operatorname{span}\{e_1, \dots, e_N\}, \quad x \mapsto \sum_{k=1}^{N} x_k e_k, \quad N \in \mathbb{N}$$

which is linear and bounded. Still from Meise and Vogt [39, 28.10, p. 331], we see that for any $j \in \mathbb{N}$ there exists $m \in \mathbb{N}$ and $C > 0$ such that for any $x \in \mathfrak{X}$

$$\sup_{N \in \mathbb{N}} p_j(\Pi_N x) \leq C p_m(x) \tag{14}$$

Moreover, we can easily see that for any $K \subset \mathfrak{X}$ compact and any $j \in \mathbb{N}$ we have

$$\sup_{x \in K} p_j(x - \Pi_N x) \to 0$$

as $N \to \infty$. Indeed, following Schaefer [47, p. 81] and from (14) we see that

$$\sup_{N\in\mathbb{N}} \sup_{x\in S} p_j\left(\Pi_N x\right) \leq C \sup_{x\in S} p_m(x) < \infty$$

for any $j \in \mathbb{N}$ and $S \subset \mathfrak{X}$ with finite cardinality. Trivially, $\sup_{x\in S} p_j\left(x\right) < \infty$. We therefore deduce that the subset $\{\Pi_N\}_N \cup \{I\} \subset \mathcal{L}(\mathfrak{X})$ is simply bounded, with I being the identity map. By Schaefer [47, Thm 4.2, p. 83], it is equicontinuous, being $\mathfrak{X}$ a Baire space. By Schaefer [47, Thm 4.5, p. 85] we therefore conclude that we have convergence on all precompact subsets of $\mathfrak{X}$.

We are now going to impose the following "graded" Lipschitz condition on the non-linearity σ:

$$\exists k_0 \in \mathbb{N} : \forall k \geq k_0 \ \exists C_k \geq 0 : p_k(\sigma(x) - \sigma(y)) \leq C_k p_k(x-y), \quad x, y \in \mathfrak{X}. \tag{15}$$

Notice that such a map σ is automatically continuous.

We are ready to prove:

Theorem 4.3. *Let $\mathfrak{X}$ be a real separable Fréchet space that admits a Schauder basis $(e_k)_{k\in\mathbb{N}}$ and let σ satisfy condition* (15)*. Let $f \in C(\mathfrak{X};\mathbb{R})$, $K \subset \mathfrak{X}$ compact and $\varepsilon > 0$. Assume*

$$\mathcal{N}^{\varepsilon}(x) = \sum_{j=1}^{M} \langle \ell_j, \sigma(A_j x + b_j)\rangle, \quad x \in \mathfrak{X}$$

with $\ell_j \in \mathfrak{X}'$, $A_j \in \mathcal{L}(\mathfrak{X})$ and $b_j \in \mathfrak{X}$ such that

$$\sup_{x\in K} |f(x) - \mathcal{N}^{\varepsilon}(x)| < \varepsilon.$$

Fix $\delta > 0$. Then there exists $N_* = N_*(\mathcal{N}^{\epsilon}, \delta) \in \mathbb{N}$ such that for $N \geq N_*$

$$\sup_{x\in K} \left| f(x) - \sum_{j=1}^{M} \langle \ell_j \circ \Pi_N, \sigma(\Pi_N A_j \Pi_N x + \Pi_N b_j)\rangle \right| < \varepsilon + \delta.$$

Proof. For $j = 1, \ldots, M$, $N \in \mathbb{N}$ and $x \in K$ we indeed have, for suitable integers $r(\ell_j)$, $t(\ell_j), m(\ell_j, \sigma)$ and $n(\ell_j, \sigma, A_j)$,

$$\begin{aligned}
&|\langle \ell_j, \sigma(A_j x + b_j)\rangle - \langle \ell_j \circ \Pi_N, \sigma(\Pi_N A_j \Pi_N x + \Pi_N b_j)\rangle| \\
&\quad \leq |\langle \ell_j, \sigma(A_j x + b_j) - \Pi_N \sigma(A_j x + b_j)\rangle| \\
&\qquad + |\langle \ell_j, \Pi_N \sigma(A_j x + b_j) - \Pi_N \sigma(\Pi_N A_j \Pi_N x + \Pi_N b_j)\rangle| \\
&\quad \leq C(\ell_j) p_{r(\ell_j)}(\sigma(A_j x + b_j) - \Pi_N \sigma(A_j x + b_j)) \\
&\qquad + C(\ell_j) p_{r(\ell_j)}(\Pi_N \sigma(A_j x + b_j) - \Pi_N \sigma(\Pi_N A_j \Pi_N x + \Pi_N b_j)) \\
&\quad \leq C(\ell_j) p_{r(\ell_j)}(\sigma(A_j x + b_j) - \Pi_N \sigma(A_j x + b_j)) \\
&\qquad + C(\ell_j) \sup_{N\in\mathbb{N}} p_{r(\ell_j)}(\Pi_N \sigma(A_j x + b_j) - \Pi_N \sigma(\Pi_N A_j \Pi_N x + \Pi_N b_j)) \\
&\quad \leq C(\ell_j) p_{r(\ell_j)}(\sigma(A_j x + b_j) - \Pi_N \sigma(A_j x + b_j)) \\
&\qquad + C(\ell_j) C p_{t(\ell_j)}(\sigma(A_j x + b_j) - \sigma(\Pi_N A_j \Pi_N x + \Pi_N b_j)),
\end{aligned}$$

where in the last line we have used the fact that the constant C in (14) is independent of N and x. Therefore, for the second term in the last expression we have

$$\begin{aligned}
&C(\ell_j)\, p_{t(\ell_j)}(\sigma(A_j x + b_j) - \sigma(\Pi_N A_j \Pi_N x + \Pi_N b_j)) \\
&\quad \leq C(\ell_j)\, p_{t(\ell_j)\vee k_0}(\sigma(A_j x + b_j) - \sigma(\Pi_N A_j \Pi_N x + \Pi_N b_j)) \\
&\quad \leq C(\ell_j) C_{t(\ell_j)\vee k_0}\, p_{t(\ell_j)\vee k_0}(A_j x + b_j - \Pi_N A_j \Pi_N x - \Pi_N b_j) \\
&\quad \leq C(\ell_j, \sigma) \left\{ p_{t(\ell_j)\vee k_0}(A_j x - \Pi_N A_j x) + p_{t(\ell_j)\vee k_0}(\Pi_N A_j x - \Pi_N A_j \Pi_N x) \right. \\
&\qquad\qquad \left. + p_{t(\ell_j)\vee k_0}(b_j - \Pi_N b_j) \right\} \\
&\quad \leq C(\ell_j, \sigma) \left\{ p_{t(\ell_j)\vee k_0}(A_j x - \Pi_N A_j x) + C'(\ell_j, \sigma) p_{m(\ell_j,\sigma)}(A_j x - A_j \Pi_N x) \right. \\
&\qquad\qquad \left. + p_{t(\ell_j)\vee k_0}(b_j - \Pi_N b_j) \right\} \\
&\quad \leq C(\ell_j, \sigma) \left\{ p_{t(\ell_j)\vee k_0}(A_j x - \Pi_N A_j x) + C'(\ell_j, \sigma, A_j) p_{n(\ell_j,\sigma,A_j)}(x - \Pi_N x) \right.
\end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad +p_{t(\ell_j)\vee k_0}(b_j - \Pi_N b_j)\Big\} \\
&\le C(\ell_j,\sigma)\Bigg\{\sup_{x\in K} p_{t(\ell_j)\vee k_0}(A_j x - \Pi_N A_j x) \\
&\qquad\qquad +C'(\ell_j,\sigma,A_j)\sup_{x\in K} p_{n(\ell_j,\sigma,A_j)}(x-\Pi_N x) + p_{t(\ell_j)\vee k_0}(b_j - \Pi_N b_j)\Bigg\} \\
&\le C(\ell_j,\sigma)\Bigg\{\sup_{y\in A_j K} p_{t(\ell_j)\vee k_0}(y - \Pi_N y) \\
&\qquad\qquad +C'(\ell_j,\sigma,A_j)\sup_{x\in K} p_{n(\ell_j,\sigma,A_j)}(x-\Pi_N x) + p_{t(\ell_j)\vee k_0}(b_j - \Pi_N b_j)\Bigg\}.
\end{aligned}
$$

Observe that $A_j K \subset \mathfrak{X}$ is compact. Setting for convenience $\sigma_j := \sigma(A_j K + b_j)$, and noticing that it is compact, we eventually arrive at

$$
\begin{aligned}
&|\langle \ell_j, \sigma(A_j x + b_j)\rangle - \langle \ell_j \circ \Pi_N, \sigma(\Pi_N A_j \Pi_N x + \Pi_N b_j)\rangle| \\
&\le C(\ell_j,\sigma)\Bigg\{\sup_{y\in A_j K} p_{t(\ell_j)\vee k_0}(y - \Pi_N y) + \sup_{y\in\sigma_j} p_{r(\ell_j)}(y-\Pi_N y) \\
&\qquad\qquad +C'(\ell_j,\sigma,A_j)\sup_{x\in K} p_{n(\ell_j,\sigma,A_j)}(x-\Pi_N x) + p_{t(\ell_j)\vee k_0}(b_j - \Pi_N b_j)\Bigg\}.
\end{aligned}
$$

By the approximation property provided by the Schauder basis $(e_k)_{k\in\mathbb{N}}$, we may find $N(j)\in\mathbb{N}$ such that:

$$
\begin{cases}
\sup_{y\in A_j K} p_{t(\ell_j)\vee k_0}(y-\Pi_N y) < \frac{\delta}{4MC(\ell_j,\sigma)} \\
\sup_{y\in\sigma_j} p_{r(\ell_j)}(y-\Pi_N y) < \frac{\delta}{4MC(\ell_j,\sigma)} \\
\sup_{x\in K} p_{n(\ell_j,\sigma,A_j)}(x-\Pi_N x) < \frac{\delta}{4MC(\ell_j,\sigma)C'(\ell_j,\sigma,A_j)}, \quad \text{if } A_j \neq 0 \\
p_{t(\ell_j)\vee k_0}(b_j - \Pi_N b_j) < \frac{\delta}{4MC(\ell_j,\sigma)}
\end{cases}
$$

for all $N \ge N(j)$. With this choice, we then have

$$
\sup_{x\in K} |\langle \ell_j, \sigma(A_j x + b_j)\rangle - \langle \ell_j\circ\Pi_N, \sigma(\Pi_N A_j \Pi_N x + \Pi_N b_j)\rangle| < \delta/M.
$$

Therefore, setting $N_* := \max\{N(1),\dots,N(M)\}$, we conclude that for all $N \ge N_*$

$$
\sup_{x\in K}\left| f(x) - \sum_{j=1}^{M} \langle \ell_j\circ\Pi_N, \sigma(\Pi_N A_j \Pi_N x + \Pi_N b_j)\rangle\right| < \varepsilon + \delta.
$$

□

Again, the required function $\mathcal{N}^\varepsilon : \mathfrak{X} \to \mathbb{R}$ exists in view of Theorem 2.3. However, we need to enhance Example 2.12 to show that activation functions σ satisfying condition (15) exist.

Example 4.4. *Let $\mathfrak{X}$ be a real Fréchet space (not necessarily admitting a Schauder basis). Consider a function $\beta \in \mathrm{Lip}(\mathbb{R};\mathbb{R})$ such that*

$$
\lim_{\xi\to\infty}\beta(\xi) = 1, \ \lim_{\xi\to-\infty}\beta(\xi) = 0, \ \beta(0) = 0,
$$

and arbitrary $z\in\mathfrak{X}, z\neq 0$. Let $\psi\in\mathfrak{X}'\setminus\{0\}$. Define

$$
\sigma(x) = \beta(\psi(x))z, \quad x\in\mathfrak{X}.
$$

Evidently, σ is continuous and von Neumann-bounded, because for any $j\in\mathbb{N}$

$$
p_j(\sigma(x)) \le |\beta(\psi(x))|\, p_j(z) \le \|\beta\|_\infty\, p_j(z) < \infty
$$

uniformly in $x \in \mathfrak{X}$. Furthermore, it is clear that σ satisfies (5)*. Let us finally check that condition* (15) *is met. To this aim, let $k \in \mathbb{N}$. We have*

$$\begin{aligned} p_k(\sigma(x) - \sigma(y)) &= |\beta(\psi(x)) - \beta(\psi(y))|\, p_k(z) \\ &\leq \mathrm{Lip}(\beta) p_k(z))\, |\psi(x) - \psi(y)| \\ &\leq \mathrm{Lip}(\beta) p_k(z) C_\psi\, p_{m(\psi)}(x - y) \\ &:= C(\beta, z, \psi; k)\, p_{m(\psi)}(x - y), \quad x, y \in \mathfrak{X} \end{aligned}$$

for some $m(\psi) \in \mathbb{N}$. Therefore, for any $k \geq m(\psi)$, since the seminorms are non-decreasing, we have

$$p_k(\sigma(x) - \sigma(y)) \leq C(\beta, z, \psi; k)\, p_k(x - y), \quad x, y \in \mathfrak{X}.$$

5. Multi-layer Neural Networks

In this section we are going to show that results analogous to Theorems 2.3 and 2.8 hold also for multi-layer (deep) neural networks with a fixed number $n > 1$ of layers. We consider the following n-layer neural network

$$\mathcal{N}_{\ell, A_1, b_1, \ldots, A_n, b_n} : \mathfrak{X} \to \mathbb{F}, \quad \mathcal{N}_{\ell, A_1, b_1, \ldots, A_n, b_n}(x) := \langle \ell, (\sigma \circ T_1 \circ \cdots \circ \sigma \circ T_n)(x) \rangle, \quad x \in \mathfrak{X},$$

with $\ell \in \mathfrak{X}'$, $A_1, \ldots, A_n \in \mathcal{L}(\mathfrak{X})$, $b_1, \ldots, b_n \in \mathfrak{X}$, $\sigma : \mathfrak{X} \to \mathfrak{X}$ continuous, and where we have set

$$T_j(x) := A_j x + b_j, \quad x \in \mathfrak{X},\ j = 1, \ldots, n.$$

Define

$$\mathfrak{N}(\sigma) := \mathrm{span}\{\mathcal{N}_{\ell, A_1, b_1, \ldots, A_n, b_n};\ \ell \in \mathfrak{X}', A_1, \ldots, A_n \in \mathcal{L}(\mathfrak{X}), b_1, \ldots, b_n \in \mathfrak{X}\}.$$

Before embarking on the proof of the density of $\mathfrak{N}(\sigma)$, we need to establish the following result, which will turn out to be very fruitful in the sequel.

Lemma 5.1. *Assume that $\mathfrak{X}$ is a real separable Fréchet space. Let $\sigma : \mathfrak{X} \to \mathfrak{X}$ be continuous and satisfying the following condition: there exist $\psi \in \mathfrak{X}' \setminus \{0\}$ and $0 \neq u_+ \in \mathfrak{X}$ such that*

$$\begin{cases} \lim_{\lambda \to \infty} \sigma(\lambda x) = u_+, & \text{if } x \in \Psi_+ \\ \lim_{\lambda \to \infty} \sigma(\lambda x) = 0, & \text{if } x \in \Psi_- \\ \lim_{\lambda \to \infty} \sigma(\lambda x) = 0, & \text{if } x \in \Psi_0 \end{cases}$$

Let $0 \neq y \in \mathfrak{X}$ be arbitrary. Then there exists $A \in \mathcal{L}(\mathfrak{X})$ such that $\sigma(Ay) \neq 0$.

Proof. We need to distinguish two cases:

(1) $y \in \Psi_0$,
(2) $y \notin \Psi_0$.

In the first case, let $\phi \in \mathfrak{X}' : \phi(y) \neq 0$. By Lemma 2.9, choose z accordingly, i.e. $\phi(z) = 1, \psi(z) \neq 0$. Consider the projection onto $\Phi_0 = \ker(\phi)$

$$\Pi_{\Phi_0} : \mathfrak{X} \to \Phi_0, \quad x \mapsto x_{\Phi_0} = x - \phi(x) z,$$

which we know belongs to $\mathcal{L}(\mathfrak{X})$. Thus, $\psi(\Pi_{\Phi_0} y) = -\phi(y)\psi(z) \neq 0$, namely $\Pi_{\Phi_0} y \notin \Psi_0$. If $\Pi_{\Phi_0} y \in \Psi_+$, set $A = \lambda \Pi_{\Phi_0}$, where $\lambda > 0$. Then $\sigma(\lambda \Pi_{\Phi_0} y) \to u_+ \neq 0$ as $\lambda \to \infty$, and therefore for $\lambda \gg 0$ we obtain $\sigma(Ay) \neq 0$. If on the other hand $\Pi_{\Phi_0} y \in \Psi_-$, set $A = -\lambda \Pi_{\Phi_0}$ this time, to get the same conclusion, i.e. $\sigma(Ay) \neq 0$.

If $y \notin \Psi_0$, then define $A = \pm \lambda I$ with $\lambda \gg 0$, accordingly if $y \in \Psi_+$ or $\in \Psi_-$. □

With this result at hand, we are now ready to prove:

Proposition 5.2. *Let $\mathfrak{X}$ be a real and separable Fréchet space, and let $\sigma : \mathfrak{X} \to \mathfrak{X}$ be von Neumann-bounded and satisfy the conditions of Lemma 5.1. Then $\mathfrak{N}(\sigma)$ is dense in $C(\mathfrak{X}; \mathbb{R})$ with respect to the topology of compact subsets of $\mathfrak{X}$.*

Proof. Evidently, $\mathfrak{N}(\sigma) \subset C(\mathfrak{X}; \mathbb{R})$. Assume once again that $\mathrm{cl}(\mathfrak{N}(\sigma)) \subsetneq C(\mathfrak{X}; \mathbb{R})$. Then, once again we obtain the following

$$\int_K \langle \ell, (\sigma \circ T_1 \circ \cdots \circ \sigma \circ T_n)(x) \rangle \mu(dx) = 0,$$

for all $\ell \in \mathfrak{X}'$, $A_1, \dots, A_n \in \mathcal{L}(\mathfrak{X})$, $b_1, \dots, b_n \in \mathfrak{X}$.

Observe that $\sigma(0) = 0$. Reasoning as in the proof of Proposition 2.8, this time we get that, as $\lambda \to \infty$, pointwise in $x \in \mathfrak{X}$,

$$\langle \ell, (\sigma \circ T_1 \circ \cdots \circ \sigma)(\lambda T_n(x)) \rangle \to \begin{cases} \langle \ell, (\sigma \circ T_1 \circ \cdots \circ \sigma)(T_{n-1}(u_+)) \rangle, \text{ if } T_n(x) \in \Psi_+ \\ \langle \ell, (\sigma \circ T_1 \circ \cdots \circ \sigma)(b_{n-1}) \rangle, \text{ otherwise} \end{cases}$$

and hence, since σ is von Neumann-bounded, by the dominated convergence theorem (for finite signed measures)

$$\begin{aligned} &\langle \ell, (\sigma \circ T_1 \circ \cdots \circ \sigma)(T_{n-1}(u_+)) \rangle \, \mu[K \cap T_n^{-1}(\Psi_+)] \\ &\quad + \langle \ell, (\sigma \circ T_1 \circ \cdots \circ \sigma)(b_{n-1}) \rangle \left\{ \mu[K \cap T_n^{-1}(\Psi_-)] + \mu[K \cap T_n^{-1}(\Psi_0)] \right\} = 0 \end{aligned}$$

for any $\ell \in \mathfrak{X}'$, $A_1, \dots, A_n \in \mathcal{L}(\mathfrak{X})$, $b_1, \dots, b_n \in \mathfrak{X}$.

Choosing $b_1 = b_2 = \cdots = b_{n-1} = 0$ results in

$$\langle \ell, (\sigma \circ A_1 \circ \cdots \circ \sigma \circ A_{n-1})(u_+) \rangle \, \mu[K \cap T_n^{-1}(\Psi_+)] = 0$$

for any $\ell \in \mathfrak{X}'$, $A_1, \dots, A_n \in \mathcal{L}(\mathfrak{X})$, and $b_n \in \mathfrak{X}$. Define iteratively backward

$$\begin{cases} y_{n-1} = \sigma(A_{n-1} u_+) \\ y_j = \sigma(A_j y_{j+1}), \quad j = 1, \dots, n-2, \end{cases}$$

where $A_1, \dots, A_{n-1} \in \mathcal{L}(\mathfrak{X})$ are chosen in such a way that

$$y_{n-1} \neq 0,\, y_{n-2} \neq 0, \dots,\, y_1 \neq 0.$$

This is achievable in virtue of Lemma 5.1. At the last step of the iteration we arrive at

$$\langle \ell, y_1 \rangle \, \mu[K \cap T_n^{-1}(\Psi_+)] = 0$$

for any $\ell \in \mathfrak{X}'$, $A_n \in \mathcal{L}(\mathfrak{X})$ and $b_n \in \mathfrak{X}$, and hence $\mu[K \cap T_n^{-1}(\Psi_+)] = 0$, namely

$$\mu[K \cap A^{-1}(\Psi_+ + b)] = 0$$

for any $A \in \mathcal{L}(\mathfrak{X})$, $b \in \mathfrak{X}$. Following the steps in the proof of Proposition 2.8, we conclude once more that $\mu = 0$ and hence that $\mathfrak{N}(\sigma)$ is dense in $C(\mathfrak{X}; \mathbb{R})$. □

6. An example

In modelling the dynamics of forward rates in fixed-income markets, or forward and futures contract prices in commodity markets, one is concerned with a stochastic process taking values in a suitable space of functions, $(x(t, \cdot))_{t \geq 0}$. Here, for every $t \geq 0$, $x(t, \cdot)$ is a random variable with state space being real-valued functions on $\mathbb{R}_+$, i.e., each sample defines a function $\xi \mapsto x(t, \xi), \xi \geq 0$. The minimal condition on the state space of curves is that they are locally integrable functions, see Carmona and Tehranchi [10] and Filipović [19] for forward rates. Local intergrability allows for defining zero-coupon bond prices, and swap prices in power and gas markets.

Following Benth, Detering and Galimberti [3], the price of a typical financial derivative in the power market can be expressed by the functional

$$F(x) = \mathbb{E}[\chi(x)]$$

where χ is a random field and x is a real-valued function on $\mathbb{R}_+$. In practice, x denotes the current term structure of power forward prices. Following the discussion above, we may choose the space of such functions to be $L^1_{loc} := L^1_{loc}(\mathbb{R}_+)$. Thus, $F : L^1_{loc} \to \mathbb{R}$. The random field χ, may be compactly expressed as (see Benth, Detering and Galimberti [3]),

$$\chi(x) = \mathfrak{P}(Z \mathcal{I}_D(x))),$$

where Z is a real-valued integrable random variable, $\mathfrak{P} : \mathbb{R} \to \mathbb{R}$ is a Lipschitz continuous function (being the option's payoff) and $\mathcal{I}_D$ is a linear functional on L^1_{loc} defined as an integral of x over a compact set $D \subset \mathbb{R}_+$, namely $\mathcal{I}_D(x) = \int_D x(\xi)\, d\xi,\ x \in L^1_{loc}$.

We observe the following lemma, which shows continuity of F with respect to the natural locally convex topology of L^1_{loc}:

Lemma 6.1. *The functional $F(x) = \mathbb{E}[\chi(x)]$ is locally Lipschitz on L^1_{loc}.*

Proof. We first remind that when L^1_{loc} is endowed with the following set of seminorms

$$q_m(x) := \int_0^m |x(\xi)| \, d\xi, \quad x \in L^1_{loc}, \quad m \in \mathbb{N}$$

it is a Fréchet space with metric d given by

$$d(x,y) := \sum_{m=1}^{\infty} 2^{-m} \frac{q_m(x-y)}{1+q_m(x-y)}, \quad x, y \in L^1_{loc}$$

(compare (1)).

Thus, we have for $x, y \in L^1_{loc}$,

$$|F(x) - F(y)| \leq \|\mathfrak{P}\|_{\text{Lip}} \mathbb{E}[|Z|] \int_D |x(\xi) - y(\xi)| d\xi.$$

and hence

$$|F(x) - F(y)| \leq \|\mathfrak{P}\|_{\text{Lip}} \mathbb{E}[|Z|] \, q_{m_0}(x-y)$$

where $D \subset [0, m_0], m_0 \in \mathbb{N}$. Clearly, it also holds

$$|F(x) - F(y)| \leq \|\mathfrak{P}\|_{\text{Lip}} \mathbb{E}[|Z|] 2^{m_0} [1 + q_{m_0}(x-y)] \; d(x,y)$$

But from the general theory of locally convex vector spaces we know that the set $\{y : q_{m_0}(x-y) < 1\}$ is an open neighborhood of x. Thus, for all y in this set we get

$$|F(x) - F(y)| \leq \|\mathfrak{P}\|_{\text{Lip}} \mathbb{E}[|Z|] 2^{m_0+1} \; d(x,y)$$

showing that our function is locally Lipschitz.

□

When considering forwards and options on these, the set D is typically a contractually specified week, month, quarter or year. Due to the continuity result above, we are in the context of our neural networks on a Fréchet space.

Conflict of Interest: The authors declare that they have no conflict of interest.
Data availability: Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

References

[1] C. Beck, M. Hutzenthaler, A. Jentzen, and B. Kuckuck. An overview on deep learning-based approximation methods for partial differential equations, 2021.
[2] F. E. Benth, N. Detering, and L. Galimberti. A functional neural network approach to the Cauchy problem, 2022.
[3] F. E. Benth, N. Detering, and L. Galimberti. Pricing options on flow forwards by neural networks in Hilbert space, 2022.
[4] F. E. Benth, N. Detering, and S. Lavagnini. Accuracy of deep learning in calibrating HJM forward curves. *Digital Finance*, 3(3-4):209–248, 2021.
[5] J. Berner, P. Grohs, G. Kutyniok, and P. Petersen. The modern mathematics of deep learning, 2021.
[6] V. Bogachev. *Measure Theory*. Number v. 1 in Measure Theory. Springer, 2007.
[7] V. Bogachev. *Measure Theory*. Number v. 2 in Measure Theory. Springer, 2007.
[8] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer New York, 2010.
[9] H. Buehler, L. Gonon, J. Teichmann, and B. Wood. Deep hedging. *Quantitative Finance*, 19(8):1271–1291, 2019.
[10] R. A. Carmona and M. R. Tehranchi. *Interest Rate Models: an Infinite Dimensional Stochastic Analysis Perspective*. Springer-Verlag, 2006.
[11] T. Chen and H. Chen. Approximations of continuous functionals by neural networks with application to dynamic systems. *IEEE Transactions on Neural Networks*, 4(6):910–918, 1993.
[12] T. Chen and H. Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks*, 6(4):911–917, 1995.
[13] Y. Cho and L. K. Saul. Kernel methods for deep learning. *Advances in neural information processing systems,*, pages 342–350, 2009.
[14] J. B. Conway. *A Course in Functional Analysis*. Graduate Texts in Mathematics; 96. Springer Science+Business Media, New York, 2nd edition, 2010.

[15] C. Cuchiero, M. Larsson, and J. Teichmann. Deep neural networks, generic universal interpolation, and controlled odes. *SIAM J. Math. Data Sci.*, 2:901–919, 2020.
[16] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.
[17] W. E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017.
[18] P. Enflo. A counterexample to the approximation problem in Banach spaces. *Acta Mathematica*, 130:309 – 317, 1973.
[19] D. Filipović. *Consistency Problems for Heath-Jarrow-Morton Interest Rate Models*. Springer-Verlag, 2001.
[20] K.-I. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2(3):183–192, 1989.
[21] W. H. Guss and R. Salakhutdinov. On universal approximation by neural networks with uniform guarantees on approximation of infinite dimensional maps, 2019.
[22] J. Han, A. Jentzen, and W. E. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34):8505–8510, 2018.
[23] B. Hanin and M. Sellke. Approximating continuous functions by relu nets of minimal width. 10 2017.
[24] T. Hazan and T. Jaakola. Steps toward deep kernel methods from infinite neural networks, 2015.
[25] C. Heil. *A Basis Theory Primer: Expanded Edition*. Applied and Numerical Harmonic Analysis. Birkhäuser Boston, 2011.
[26] H. Holden, B. Oksendal, J. Uboe, and T. Zhang. *Stochastic Partial Differential Equations: A Modeling, White Noise Functional Approach*. Springer, 2010.
[27] K. Hornik. Neural networks for functional approximation and system identification. *Neural Computation*, 9(1):143–159, 1997.
[28] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.
[29] M. Hutzenthaler, A. Jentzen, T. Kruse, and T. A. Nguyen. A proof that rectified deep neural networks overcome the curse of dimensionality in the numerical approximation of semilinear heat equations. *SN Partial Differential Equations and Applications*, 1(2):10, 2020.
[30] P. Kidger and T. Lyons. Universal Approximation with Deep Narrow Networks. In J. Abernethy and S. Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2306–2327. PMLR, 09–12 Jul 2020.
[31] N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar. Neural operator: Learning maps between function spaces, 2021.
[32] A. Kratsios. The universal approximation property. *Annals of Mathematics and Artificial Intelligence*, 89(5):435–469, 2021.
[33] A. Kratsios and I. Bilokopytov. Non-euclidean universal approximation. 2020.
[34] S. Lanthaler, S. Mishra, and G. E. Karniadakis. Error estimates for DeepONets: a deep learning framework in infinite dimensions. *Transactions of Mathematics and Its Applications*, 6(1), 03 2022. tnac001.
[35] 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.
[36] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Neural operator: Graph kernel network for partial differential equations, 2020.
[37] L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.
[38] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang. The expressive power of neural networks: A view from the width. 09 2017.
[39] R. Meise and D. Vogt. *Einführung in die Funktionalanalysis*. Aufbaukurs Mathematik. Vieweg, 1992.
[40] H. N. Mhaskar and N. Hahm. Some new results on neural network approximation. *Neural Networks*, 6(8):1069–1072, 1993.
[41] D. Müller, I. Soto-Rey, and F. Kramer. An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks, 2022.
[42] L. Narici. *Topological vector spaces*. Monographs and textbooks in pure and applied mathematics ; 95. M. Dekker, New York, 1985 - 1985.
[43] R. M. Neal. *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics: 118. Springer Science+Business Media, New York, 1996.
[44] A. Pinkus. Approximation theory of the mlp model in neural networks. *Acta Numerica*, 8:143–195, 1999.
[45] J. O. Ramsey and B. W. Silverman. *Functional Data Analysis*. Springer Science+Business Media, New York, 2nd edition, 2005.
[46] I. Sandberg. Approximation theorems for discrete-time systems. *IEEE Transactions on Circuits and Systems*, 38(5):564–566, 1991.
[47] H. Schaefer. *Topological Vector Spaces*. Elements of mathematics / N. Bourbaki. Springer, 1971.
[48] L. Schwartz. *Théorie des distributions*. Number v. 1-2 in Actualités scientifiques et industrielles. Hermann, 1957.

[49] T. S. Tian. Functional data analysis in brain imaging studies. *Frontiers in psychology*, 1:35–35, 10 2010.
[50] H. Triebel. A note on wavelet bases in function spaces. *Banach Center Publications*, 64(1):193–206, 2004.
[51] C. K. I. Williams. Computing with infinite networks. *Advances in neural information processing systems,*, pages 295–301, 1997.
[52] P. Yu and X. Yan. Stock price prediction based on deep neural networks. *Neural Computing and Applications*, 32(6):1609–1628, 2020.

Fred Espen Benth, University of Oslo, Department of Mathematics, P.O. Box 1053, Blindern, N–0316 Oslo, Norway
Email address: fredb@math.uio.no

Nils Detering, University of California at Santa Barbara, Department of Statistics and Applied Probability, CA 93106 Santa Barbara, USA
Email address: detering@pstat.ucsb.edu

Luca Galimberti, Norwegian University of Science and Technology, Department of Mathematical Sciences, Sentralbygg 2, Gløshaugen, Trondheim, Norway
Email address: luca.galimberti@ntnu.no