CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.20-211

Vol. **1**, No. 3, pp. 387-440
September 2020

On the Banach Spaces Associated with Multi-Layer ReLU Networks: Function Representation, Approximation Theory and Gradient Descent Dynamics

Weinan E[1,2,3,*] and Stephan Wojtowytsch[2,*]

[1] *Department of Mathematics, Princeton University, Princeton, NJ 08544, USA.*
[2] *Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544, USA.*
[3] *Beijing Institute of Big Data Research, Beijing, P.R. China.*

Received 1 September 2020; Accepted 21 September 2020

Abstract. We develop Banach spaces for ReLU neural networks of finite depth L and infinite width. The spaces contain all finite fully connected L-layer networks and their L^2-limiting objects under bounds on the natural path-norm. Under this norm, the unit ball in the space for L-layer networks has low Rademacher complexity and thus favorable generalization properties. Functions in these spaces can be approximated by multi-layer neural networks with dimension-independent convergence rates.

The key to this work is a new way of representing functions in some form of expectations, motivated by multi-layer neural networks. This representation allows us to define a new class of continuous models for machine learning. We show that the gradient flow defined this way is the natural continuous analog of the gradient descent dynamics for the associated multi-layer neural networks. We show that the path-norm increases at most polynomially under this continuous gradient flow dynamics.

AMS subject classifications: 68T07, 46E15, 26B35, 35Q68, 34A12, 26B40

Key words: Barron space, multi-layer space, deep neural network, representations of functions, machine learning, infinitely wide network, ReLU activation, Banach space, path-norm, continuous gradient descent dynamics, index representation.

1 Introduction

It is well-known that neural networks can approximate any continuous function on a compact set arbitrarily well in the uniform topology as the number of trainable parameters increase [9, 26, 32]. However, the number and magnitude of the parameters required

*Corresponding author. *Email addresses:* weinan@math.princeton.edu (W. E), stephanw@princeton.edu (S. Wojtowytsch)

may make this result unfeasible for practical applications. Indeed it has been shown to be the case when two-layer neural networks are used to approximate general Lipschitz continuous functions [22]. It is therefore necessary to ask which functions can be approximated *well* by neural networks, by which we mean that as the number of parameters goes to infinity, the convergence rate should not suffer from the curse of dimensionality.

In classical approximation theory, the role of neural networks was taken by (piecewise) polynomials or Fourier series and the natural function spaces were Hölder spaces, (fractional) Sobolev spaces, or generalized versions thereof [33]. In the high-dimensional theories characteristic for machine learning, these spaces appear inappropriate (for example, approximation results of the kind discussed above do not hold for these spaces) and other concepts have emerged, such as reproducing kernel Hilbert spaces for random feature models [37], Barron spaces for two-layer neural networks [4,17–19,22,23,29], and the flow-induced space for residual neural network models [18].

In this article, we extend these ideas to networks with several hidden (infinitely wide) layers. The key is to find how functions in these spaces should be represented and what the right norm should be. Our most important results are:

1. There exists a class of Banach spaces associated with multi-layer neural networks which has low Rademacher complexity (i.e. multi-layer functions in these spaces are easily learnable).

2. The neural tree spaces introduced here are the appropriate function spaces for the corresponding multi-layer neural networks in terms of direct and inverse approximation theorems.

3. The gradient flow dynamics is well defined in a much simpler subspace of the corresponding neural tree space. Functions in this space admit an intuitive representation in terms of compositions of expectations. The path norm increases at most polynomially in time under the natural gradient flow dynamics. Since the path-norm controls the generalization gap, this slow increase suggests that gradient flow training does not lead to overfitting.

These results justify our choice of function representation and the norm.

Neural networks are parametrized by weight matrices which share indices only between adjacent layers. To understand the approximation power of neural networks, we rearrange the index structure of weights in a tree-like fashion and show that the approximation problem under path-norm bounds remains unchanged. This approach makes the problem more linear and easier to handle from the approximation perspective, but is unsuitable when describing training dynamics. To address this discrepancy, we introduce a subspace of the natural function spaces for very wide multi-layer neural networks (or neural trees) which automatically incorporates the structure of neural networks. For this subspace, we investigate the natural training dynamics and demonstrate that the path-norm increases at most polynomially during training.

Although the function representation and function spaces are motivated by developing an approximation theory for multi-layer neural network models, once we have them, we can use them as our starting point for developing alternative machine learning models and algorithms. In particular, we can extend the program proposed in [19] on continuous formulations of machine learning to function representations developed here. As an example, we show that gradient descent training for multi-layer neural networks can be recovered as the discretization of a natural continuous gradient flow.

The article is organized as follows. In the remainder of the introduction, we discuss the philosophy behind this study and the continuous approach to machine learning. In Section 2, we motivate the 'neural tree' approach, introduce an abstract class of function spaces and study their first properties. A special instance of this class tailored to multi-layer networks is studied in greater detail in Section 3. A class of function families with an explicit network structure is introduced in Section 4. While Sections 2 and 3 are written from the approximation perspective, Section 5 is devoted to the study of gradient flow optimization of multi-layer networks and its relation to the function spaces we introduce. We conclude the article with a brief discussion of our results and some open questions in Section 6. Technical results from measure theory which are needed in the article are gathered in the appendix.

1.1 Conventions and notation

Let $K \subseteq \mathbb{R}^d$ be a compact set. Then we denote by $C^0(K)$ the space of continuous functions on K and by $C^{0,\alpha}(K)$ the space of α-Hölder continuous functions for $\alpha \in (0,1]$. In particular $C^{0,1}$ is the space of Lipschitz-continuous functions. The norms are denoted as

$$\|f\|_{C^0(K)} = \sup_{x\in K} |f(x)|, \quad [f]_{C^{0,\alpha}(K)} = \sup_{x,y\in K, x\neq y} \frac{|f(x)-f(y)|}{|x-y|^\alpha}, \quad \|f\|_{C^{0,\alpha}} = \|f\|_{C^0} + [f]_{C^{0,\alpha}}.$$

Since all norms on $\mathbb{R}^d$ are equivalent, the space of Hölder- or Lipschitz-continuous functions does not depend on the choice of norm on $\mathbb{R}^d$. The Hölder constant $[\cdot]_{C^{0,\alpha}}$ however does depend on it, and using different ℓ^p-norms leads to a dimension-dependent factor. In this article, we always consider $\mathbb{R}^d$ equipped with the ℓ^∞-norm, which seems more natural on common domains such as $[0,1]^d$.

Let X be a Banach space. Then we denote by B^X the closed unit ball in X. If X, Y are Banach spaces, we write $X \hookrightarrow Y$ to express that X embeds continuously into Y. Furthermore, a review of notations, terminologies and results relating to measure theory can be found in the appendix.

Frequently and without comment, we identify $x \in \mathbb{R}^d$ with $(x,1) \in \mathbb{R}^{d+1}$. This allows us to simplify notation and treat affine maps as linear. In particular, for $x \in \mathbb{R}^d$ and $w \in \mathbb{R}^{d+1}$ we simply write $w^T x = \sum_{i=1}^d w_i x_i + w_{d+1}$.

2 Generalized Barron spaces

We begin by reviewing multi-layer neural networks.

2.1 Neural networks and neural trees

A *fully connected L-layer neural network* is a function of the type

$$f(x)=\sum_{i_L=1}^{m_L} a_{i_L}^{L}\sigma\left(\sum_{i_{L-1}=1}^{m_{L-1}} a_{i_L i_{L-1}}^{L-1}\sigma\left(\sum\cdots\sigma\left(\sum_{i_1=1}^{m_1} a_{i_2 i_1}^{1}\sigma\left(\sum_{i_0=1}^{d+1} a_{i_1 i_0}^{0}x_{i_0}\right)\right)\right)\right), \tag{2.1}$$

where the parameters a_{ij}^{ℓ} are referred to as the *weights* of the neural network, m_ℓ is the *width* of the ℓ-th layer, and $\sigma:\mathbb{R}\to\mathbb{R}$ is a non-polynomial activation function. For the purposes of this article, we take σ to be the *rectifiable linear unit* $\sigma(z)=\mathrm{ReLU}(z)=\max\{z,0\}$.

Deep neural networks are complicated functions of both their input x and their weights, where the weights of one layer only share an index with neighbouring layers, leading to parameter reuse. For simplicity, consider a network with two hidden layers

$$f(x)=\sum_{i_2=1}^{m_2} a_{i_2}^{2}\sigma\left(\sum_{i_1=1}^{m_1} a_{i_2 i_1}^{1}\sigma\left(\sum_{i_0=1}^{d+1} a_{i_1 i_0}^{0}x_{i_0}\right)\right),$$

and note that f can also be expressed as

$$f(x)=\sum_{i_2=1}^{m_2} a_{i_2}^{2}\sigma\left(\sum_{i_1=1}^{m_1} a_{i_2 i_1}^{1}\sigma\left(\sum_{i_0=1}^{d+1} b_{i_2 i_1 i_0}^{0}x_{i_0}\right)\right), \tag{2.2}$$

with $b_{i_2 i_1 i_0}^{0}\equiv a_{i_1 i_0}^{0}$. In this way, an index in the outermost layer gets its own set of parameters for deeper layers, eliminating parameter sharing. The function parameters are arranged in a tree-like structure rather than a network with many cross-connections. On the other hand, a function of the form (2.2) can equivalently be expressed as

$$f(x)=\sum_{i_2=1}^{m_2} a_{i_2}^{2}\sigma\left(\sum_{j_1=1}^{m_1 m_2} c_{i_2 j_1}^{1}\sigma\left(\sum_{i_0=1}^{d+1} c_{j_1 i_0}^{0}x_{i_0}\right)\right),$$

with

$$c_{i_2 j_1}^{1}=\begin{cases} a_{i_2,j_1-(i_2-1)m_1}^{1} & \text{if } (i_2-1)m_1<j_1\le i_2 m_1,\\ 0 & \text{else,}\end{cases}$$

$$c_{j_1 i_0}^{0}=b_{\lfloor j_1/m_1\rfloor+1,\,j_1-\lfloor j_1/m_1\rfloor,\,i_0}.$$

The cost of rearranging a three-dimensional index set into a two-dimensional one is listing a number of zero-elements explicitly in the preceding layer instead of implicitly. Conversely, if we rearrange a two-dimensional index set into a three-dimensional one, we

need to repeat the same weight multiple times. For deeper trees, the index sets become even higher-dimensional, and the re-arrangement introduces even more trivial branches or redundancies. Nevertheless, we note that the space of finite neural networks of depth L

$$\mathcal{F}_\infty := \left\{ \sum_{i_L=1}^{\infty} a_{i_L}^L \sigma\left(\sum_{i_{L-1}=1}^{\infty} a_{i_L i_{L-1}}^{L-1} \sigma\left(\sum \cdots \sigma\left(\sum_{i_1=1}^{\infty} a_{i_2 i_1}^1 \sigma\left(\sum_{i_0=1}^{d+1} a_{i_1 i_0}^0 x_{i_0} \right) \right) \right) \right) \right| a_{ij}^l = 0 \quad \text{for all but finitely many } i,j,l \Bigg\}$$

and the space of finite neural trees of depth L

$$\widetilde{\mathcal{F}}_\infty := \left\{ \sum_{i_L=1}^{\infty} a_{i_L}^L \sigma\left(\sum_{i_{L-1}=1}^{\infty} a_{i_L i_{L-1}}^{L-1} \sigma\left(\sum \cdots \sigma\left(\sum_{i_1=1}^{\infty} a_{i_L \cdots i_2 i_1}^1 \sigma\left(\sum_{i_0=1}^{d+1} a_{i_L \cdots i_1 i_0}^0 x_{i_0} \right) \right) \right) \right) \right| a_{i_L \cdots i_k}^l = 0 \quad \text{for all but finitely many } l,\, i_1, \cdots, i_L \Bigg\}$$

are identical.

Remark 2.1. We note that this perspective is only admissible concerning approximation theory. For gradient flow-based training algorithms, it makes a huge difference

- whether parameters are reused or not,
- which set of weights that induces a certain function is chosen, and
- how the magnitude of the weights is distributed across the layers (using the invariance $\sigma(z) = \lambda^{-1}\sigma(\lambda z)$ for $\lambda > 0$).

A perspective more adapted to the training of neural networks is presented in Section 5.

For given weights a_{ij}^l or $a_{i_L \cdots i_l}^l$, we consider the *path-norm proxy*, which is defined as

$$\|f\|_{pnp} = \sum_{i_L} \cdots \sum_{i_0} |a_{i_L}^L \cdots a_{i_1 i_0}^0| \qquad \text{or} \qquad \|f\|_{pnp} = \sum_{i_L} \cdots \sum_{i_0} |a_{i_L}^L \cdots a_{i_L \cdots i_0}^0|,$$

respectively. Knowing the weights, the sum is easy to compute and it naturally controls the Lipschitz norm of the function f.

When we train a function f to approximate values $y_i = f^*(x_i)$ at data points x_i, the path-norm proxy controls the *generalization error*, as we will show below. If the path-norm proxy of f is very large, the function values $f(x_i)$ heavily depend on cancellations between the partial sums with positive and negative weights in the outermost layer. In the extreme case, these partial sums may be several orders of magnitude larger than $f(x_i)$.

In that situation, the function values $f^*(x)$ and $f(x)$ may be entirely different for unseen data points x, even if they are close on the training sample $\{x_i\}_{i=1}^N$. On the other hand, we will show below that functions with low path-norm proxy generalize well. Thus controlling the path-norm proxy effectively means controlling the generalization error, either directly or indirectly. We will make this more precise below.

While the path-norm proxy is easy to compute from the weights of a network, it is a quantity related to the parameterization of a function, not the function itself. The map from the weights a_{ij}^l to the *realization* f of the network as in (2.1) is highly non-injective. The *path-norm* of a function f is the infimum of the path-norm proxies over all sets of weights of an L-layer neural network which have the realization f.

2.2 Definition of generalized Barron spaces

Let σ be the rectified linear unit, i.e. $\sigma(z) = \max\{z,0\}$. ReLU is a popular activation function for neural networks and has two useful properties for us: It is positively one-homogeneous and Lipschitz continuous with Lipschitz constant 1.

Let $K \subseteq \mathbb{R}^d$ be a compact set and X be a Banach space such that

1. X embeds continuously into the space $C^{0,1}(K)$ of Lipschitz-functions on K and
2. the closed unit ball B^X in X is closed in the topology of $C^0(K)$.

Recall the following corollary to the Arzelà-Ascoli theorem.

Lemma 2.1. [11, Satz 2.42] *Let $u_n : K \to \mathbb{R}$ be a sequence of functions such that $\|u_n\|_{C^{0,1}(K)} \leq 1$. Then there exists $u \in C^{0,1}(K)$ and a subsequence u_{n_k} such that $u_{n_k} \to u$ strongly in $C^{0,\alpha}(K)$ for all $\alpha < 1$ and*

$$\|u\|_{C^{0,1}(K)} \leq \liminf_{k\to\infty} \|u_{n_k}\|_{C^{0,1}(K)} \leq 1.$$

Thus B^X is pre-compact in the separable Banach space $C^0(K)$. Since B^X is C^0-closed, it is compact, so in particular a Polish space. A brief review of measure theory in Polish spaces and related topics used throughout the article is given in Appendix A.

Let μ be a finite signed measure on the Borel σ-algebra of B^X (with respect to the C^0-norm). Then μ is a signed Radon measure. The vector-valued function

$$B^X \to C^0(K), \qquad g \mapsto \sigma(g)$$

is continuous and thus μ-integrable in the sense of Bochner integrals. We define

$$\begin{aligned} f_\mu &= \int_{B^X} \sigma\big(g(\cdot)\big)\, \mu(\mathrm{d}g), \\ \|f\|_{X,K} &= \inf\Big\{ \|\mu\|_{\mathcal{M}(B^X)} : \mu \in \mathcal{M}(B^X) \text{ s.t. } f = f_\mu \text{ on } K \Big\}, \\ \mathcal{B}_{X,K} &= \big\{ f \in C^0(K) : \|f\|_{X,K} < \infty \big\}. \end{aligned} \tag{2.3}$$

Here $\mathcal{M}(B^X)$ denotes the space of (signed) Radon measures on B^X. The first integral can equivalently be considered as a Lebesgue integral pointwise for every $x \in K$ or as a Bochner integral. We will show below that $\mathcal{B}_{X,K}$ is a normed vector space of (Lipschitz-)continuous functions on K. We call $\mathcal{B}_{X,K}$ the generalized Barron space modelled on X.

Remark 2.2. The construction of the function space $\mathcal{B}_{X,K}$ above resembles the approach to *Barron spaces* for two-layer networks [4, 17, 18, 23]. Note that Barron spaces are distinct from the class of functions considered by Barron in [5], which is sometimes referred to as *Barron class*. While Barron spaces are specifically designed for applications concerning neural networks, the Barron class is defined in terms of spectral properties and a subset of Barron space for almost every activation function of practical importance.

Example 2.1. If X is the space of affine functions from $\mathbb{R}^d$ to $\mathbb{R}$ (which is isomorphic to $\mathbb{R}^{d+1}$), the $\mathcal{B}_{X,K}$ is the usual Barron space for two-layer neural networks as described in [17, 18, 23].

Due to Lemma 2.1, we may choose $X = C^{0,1}(K)$.

Example 2.2. If $X = C^{0,1}(K)$, then $\mathcal{B}_{X,K} = C^{0,1}(K)$ and the norms are equivalent to within a factor of two. For $f \in C^{0,1}(K)$, we represent

$$\begin{aligned} f &= \|f\|_{C^{0,1}(K)} \sigma\left(\frac{f}{\|f\|_{C^{0,1}(K)}}\right) - \|f\|_{C^{0,1}(K)} \sigma\left(-\frac{f}{\|f\|_{C^{0,1}(K)}}\right) \\ &= \int_{B^X} \sigma(g) \left(\|f\|_{C^{0,1}} \cdot \delta_{\frac{f}{\|f\|_{C^{0,1}}}} - \|f\|_{C^{0,1}} \cdot \delta_{-\frac{f}{\|f\|_{C^{0,1}}}} \right) (\mathrm{d}g). \end{aligned}$$

These examples are on opposite sides of the spectrum with X being either the least complex non-trivial space or the largest admissible space. Spaces of deep neural networks lie somewhere between those extremes.

Remark 2.3. For the classical Barron space, we usually consider measures supported on the unit sphere in the finite-dimensional space X. If X is infinite-dimensional, typically only the unit ball in X is closed (and thus compact) in C^0, but not the unit sphere. For mathematical convenience, we choose the compact setting.

2.3 Properties

Let us establish some first properties of generalized Barron spaces.

Theorem 2.1. *The following are true.*

1. *$\mathcal{B}_{X,K}$ is a Banach-space.*
2. *$X \hookrightarrow \mathcal{B}_{X,K}$ and $\|f\|_{\mathcal{B}_{X,K}} \leq 2\|f\|_X$.*

3. *$\mathcal{B}_{X,K} \hookrightarrow C^{0,1}(K)$ and the closed unit ball of $\mathcal{B}_{X,K}$ is a closed subset of $C^0(K)$.*

Proof. Since $X \hookrightarrow C^{0,1}(K)$, we know that there exist $C_1, C_2 > 0$ such that

$$\|g\|_{C^0(K)} \leq C_1 \|g\|_X, \qquad [g]_{C^{0,1}(K)} \leq C_2 \|g\|_X \qquad \forall\, g \in X.$$

Banach space. By construction, $\mathcal{B}_{X,K}$ is isometric to the quotient space $\mathcal{M}(B^X)/N_K$ where

$$N_K = \left\{ \mu \in \mathcal{M}(B^X) \,\middle|\, \int_{B^X} \sigma\big(g(x)\big)\, \mu(\mathrm{d}g) = 0 \;\forall\, x \in K \right\}.$$

In particular, $\mathcal{B}_{X,K}$ is a normed vector space with the norm $\|\cdot\|_{X,K}$. The map

$$\mathcal{M}(B^X) \to C^0(K), \qquad \mu \mapsto f_\mu = \int_{B^X} \sigma(g)\, \mu(\mathrm{d}g)$$

is continuous as

$$\begin{aligned} \left\| \int_{B^X} \sigma(g)\, \mu(\mathrm{d}g) \right\|_{C^0(K)} &\leq \int_{B^X} \|g\|_{C^0(K)} |\mu|(\mathrm{d}g) \\ &\leq C_1 \|\mu\|_{\mathcal{M}(B^X)} \end{aligned}$$

by the properties of Bochner spaces. Thus N_K is the kernel of a continuous linear map, i.e. a closed subspace. We conclude that $\mathcal{B}_{X,K}$ is a Banach space [7, Proposition 11.8].

X embeds into $\mathcal{B}_{X,K}$. For $g \in X$ with $\|g\|_X = 1$ consider $\mu = \delta_g - \delta_{-g}$ and observe that

$$f_\mu = \sigma(g) - \sigma(-g) = g, \qquad \|\mu\|_{\mathcal{M}(B^X)} = 2.$$

The general case follows by homogeneity.

$\mathcal{B}_{X,K}$ embeds into $C^{0,1}$. We have already shown that $\|f_\mu\|_{C^0(K)} \leq C_1 \|\mu\|_{\mathcal{M}(B^X)}$. By taking the infimum over μ, we find that $\|f\|_{C^0(K)} \leq R \|f\|_{\mathcal{B}_{X,K}}$. Furthermore, for any $x \neq x' \in K$ we have

$$\begin{aligned} |f_\mu(x) - f_\mu(x')| &\leq \int_{B^X} \big|\sigma\big(g(x)\big) - \sigma\big(g(x')\big)\big|\, |\mu|(\mathrm{d}g) \\ &\leq \int_{B^X} \big|g(x) - g(x')\big|\, |\mu|(\mathrm{d}g) \\ &\leq \int_{B^X} [g]_{C^{0,1}} |x - x'|\, |\mu|(\mathrm{d}g) \\ &\leq C_2 \|\mu\|_{\mathcal{M}(B^X)} |x - x'|. \end{aligned}$$

We can now take the infimum over μ.

Now assume that $(f_n)_{n\in\mathbb{N}}$ is a sequence such that $\|f_n\|_{X,K}\leq 1$ for all $n\in\mathbb{N}$. Choose a sequence of measures μ_n such that $f_n = f_{\mu_n}$ and $\|\mu_n\|\leq 1+\frac{1}{n}$ for all $n\in\mathbb{N}$. By the compactness theorem for Radon measures (see Theorem A.5 in the appendix), there exists a subsequence μ_{n_k} and a Radon measure μ on B^X such that $\mu_{n_k}\rightharpoonup\mu$ as Radon measures and $\|\mu\|\leq 1$.

By definition, the weak convergence of Radon measures implies that

$$\int_{B^X} F(g)\mu_{n_k}(\mathrm{d}g)\to\int_{B^X} F(g)\mu(\mathrm{d}g)\qquad\forall\, F\in C(B^X).$$

Using $F(g)=\sigma(g(x))$, we find that $f_{\mu_{n_k}}\to f_\mu$ pointwise. In particular, if f_{μ_n} converges to a limit $\tilde{f}$ uniformly, then $\tilde{f}=f\in B^{\mathcal{B}_{X,K}}$, i.e. the unit ball of $\mathcal{B}_{X,K}$ is closed in the C^0-topology. □

The last property establishes that $\mathcal{B}_{X,K}$ satisfies the same properties which we imposed on X, i.e. we can repeat the construction and consider $\mathcal{B}_{\mathcal{B}_{X,K},K}$.

Remark 2.4. We have shown in [23] that if K is an infinite set, Barron space is generally neither separable nor reflexive. In particular, $\mathcal{B}_{X,K}$ is not expected to have either of these properties in the more general case.

2.4 Rademacher complexities

We show that generalized Barron spaces have a favorable property from the perspective of statistical learning theory.

A convenient (and sometimes realistic) assumption is that all data samples accessible to a statistical learner are drawn from a distribution $\mathbb{P}$ independently. The pointwise Monte-Carlo error estimate follows from the law of large numbers which shows that for a fixed function f and data distribution $\mathbb{P}$, we have

$$\left|\mathbb{E}_{(X_1,\cdots,X_N)\sim\pi^N}\left[\sum_{i=1}^N f(X_i)-\int f(x)\mathbb{P}(\mathrm{d}x)\right]\right|\leq\frac{C_f}{\sqrt{N}}.$$

Typically, the uniform error over a function class is much larger than the pointwise error. For example for the class of one-Lipschitz functions

$$\begin{aligned}&\left|\mathbb{E}_{(X_1,\cdots,X_N)\sim\pi^N}\sup_{[f]_{C^{0,1}}\leq 1}\left[\sum_{i=1}^N f(X_i)-\int f(x)\mathbb{P}(\mathrm{d}x)\right]\right|\\ =&\mathbb{E}_{(X_1,\cdots,X_N)\sim\pi^N}\left[W_1\left(\mathbb{P},\frac{1}{N}\sum_{i=1}^N\delta_{X_i}\right)\right]\end{aligned}$$

is the expected 1-Wasserstein distance between $\mathbb{P}$ and the empirical measure of N independent sample points drawn from it. If $\mathbb{P}$ is the uniform distribution on $[0,1]^d$, this decays like $N^{-1/d}$ and thus much slower than $N^{-1/2}$ [22,24].

For Barron-type spaces, the Monte-Carlo error rate may be attained *uniformly* on the unit ball of $\mathcal{B}_{X,K}$. This is established using the Rademacher complexity of a function class. Rademacher complexities essentially decouple the sign and magnitude of oscillations around the mean by introducing additional randomness in a problem. For general information on Rademacher complexities, see [40, Chapter 26].

Definition 2.1. *Let $S = \{x_1, \cdots, x_N\}$ be a set of points in K. The* Rademacher complexity *of $\mathcal{H} \subseteq C^{0,1}(K)$ on S is defined as*

$$\mathrm{Rad}(\mathcal{H}; S) = \mathbb{E}_\xi \left[\sup_{h \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^N \xi_i h(x_i) \right], \tag{2.4}$$

where the ξ_i are iid random variables which take the values 1 and -1 with probability $1/2$ each.

The ξ_i are either referred to as symmetric Bernoulli or Rademacher variables, depending on the author.

Theorem 2.2. *Consider the unit ball $B^{\mathcal{B}_{X,K}}$ of $\mathcal{B}_{X,K}$. Let S be any sample set in $\mathbb{R}^d$. Then*

$$\mathrm{Rad}(B^{\mathcal{B}_{X,K}}; S) \leq 2\,\mathrm{Rad}(B^X, S).$$

Proof. Define the function classes $\mathcal{H}_1 = \{\sigma(g) : g \in B^X\}$, $\mathcal{H}_2 = \{-\sigma(g) : g \in B^X\}$ and $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$. All three are compact in C^0.

We decompose $\mu = \mu^+ - \mu^-$ in its mutually singular positive and negative parts and write $f = f_\mu$ in $\mathcal{B}_{X,K}$ as

$$\begin{aligned} f_\mu(x) &= \int_{B^X} \sigma(g(x))\, \mu^+(\mathrm{d}g) + \int_{B^X} -\sigma(g(x))\, \mu^-(\mathrm{d}g) \\ &= \int_{\mathcal{H}_1} h(x)\, (\rho_\sharp^+ \mu^+)(\mathrm{d}h) + \int_{\mathcal{H}_2} h(x)\, (\rho_\sharp^- \mu^-)(\mathrm{d}h) \\ &= \int_{\mathcal{H}} h(x)\, \hat{\mu}(\mathrm{d}h), \end{aligned}$$

where $\rho^\pm : B^X \to \mathcal{H}$ is given by $g \mapsto \pm\sigma(g)$ and $\hat{\mu} = \rho_\sharp^+ \mu^+ + \rho_\sharp^- \mu^-$. In particular, we note that $\hat{\mu}$ is a non-negative measure and $\|\hat{\mu}\| = \|\mu\|$. We conclude that the closed unit ball in $\mathcal{B}_{X,K}$ is the closed convex hull of $\mathcal{H}$.

Since σ is 1-Lipschitz, the contraction lemma [40, Lemma 26.9] implies that $\mathrm{Rad}(\mathcal{H}_1; S) \leq \mathrm{Rad}(B^X; S)$. Due to [40, Lemma 26.7], we find that

$$\begin{aligned} \mathrm{Rad}(B^{\mathcal{B}_{X,K}}; S) &= \mathrm{Rad}(\mathcal{H}; S) \\ &= \mathrm{Rad}(\mathcal{H}_1 \cup (-\mathcal{H}_1); S) \\ &\leq \mathrm{Rad}(\mathcal{H}_1; S) + \mathrm{Rad}(-\mathcal{H}_1; S) \\ &= 2\,\mathrm{Rad}(\mathcal{H}_1; S) \\ &= 2\,\mathrm{Rad}(B^X; S), \end{aligned}$$

since for any ξ, the supremum is non-negative. ∎

For a priori estimates, it suffices to bound the expected Rademacher complexity. However, the use of randomness in the problem is complicated, and most known bounds work on any suitably bounded sample set.

Example 2.3. If $\mathcal{H}_{lin}$ is the class of linear functions on $\mathbb{R}^d$ with ℓ^1-norm smaller or equal to 1 and S is any sample set of N elements in $[-1,1]^d$, then

$$\mathrm{Rad}(\mathcal{H}_{lin};S) \leq \sqrt{\frac{2\log(2d)}{N}},$$

see [40, Lemma 26.11]. If $\mathcal{H}_{aff}$ is the unit ball in the class of affine functions $x \mapsto w^T x + b$ with the norm $|w|_{\ell^1} + |b|$, we can simply extend x to $(x,1)$ and see that

$$\mathrm{Rad}(\mathcal{H}_{aff};S) \leq \sqrt{\frac{2\log(2d+2)}{N}}.$$

We show that Monte-Carlo rate decay is the best possible result for Rademacher complexities under very weak conditions.

Example 2.4. Let $\mathcal{F}$ be a function class which contains the constant functions $f \equiv \alpha$ and $f \equiv \beta$ for $\alpha,\beta \in \mathbb{R}$. Then there exists $c > 0$ such that

$$\mathrm{Rad}(\mathcal{F};S) \geq c\frac{|\alpha-\beta|}{\sqrt{N}}$$

for any sample set S with N elements. Up to scaling and a constant shift (which does not affect the complexity), we may assume that $\beta = 1, \alpha = -1$. Then

$$\begin{aligned}\mathrm{Rad}(\mathcal{F};S) &\geq \mathbb{E}_\xi \frac{1}{m} \sup_{f=\pm 1} \sum_{i=1}^m \xi_i f(x_i) \\ &= \mathbb{E}_\xi \frac{1}{m} \left| \sum_{i=1}^m \xi_i \right| \\ &\sim \frac{1}{\sqrt{2\pi m}}\end{aligned}$$

by the central limit theorem.

3 Banach spaces for multi-layer neural networks

3.1 Neural tree spaces

In this section, we discuss feed-forward neural networks of infinite width and finite depth L. Let $K \subseteq \mathbb{R}^d$ be a fixed compact set. Consider the following sequence of spaces.

1. $\mathcal{W}^0(K) = (\mathbb{R}^d)^* \oplus \mathbb{R} \cong \mathbb{R}^{d+1}$ is the space of affine functions from $\mathbb{R}^d$ to $\mathbb{R}$ (restricted to K).

2. For $L \geq 1$, we set $\mathcal{W}^L(K) = \mathcal{B}_{\mathcal{W}^{L-1}(K),K}$.

Since we consider $\mathbb{R}^d$ to be equipped with the ℓ^∞-norm, we take $\mathcal{W}^0$ to be equipped with its dual, the ℓ^1-norm. Up to a dimension-dependent normalization constant, this does not affect the analysis.

Thus $\mathcal{W}^L$ is the function space for $L+1$-layer networks (i.e. networks with L hidden layers/nonlinearities). Here we use inductively that $\mathcal{W}^L$ embeds into $C^{0,1}(K)$ continuously and that the unit ball of $\mathcal{W}^L$ is C^0-closed because the same properties held true for $\mathcal{W}^{L-1}$. Due to the tree-like recursive construction, we refer to $\mathcal{W}^L$ as neural tree space (with L layers).

Here and in the following, we often assume that K is a fixed set and will suppress it in the notation $\mathcal{W}^L = \mathcal{W}^L(K)$.

Remark 3.1. For a network with one hidden layer, by construction the coefficients in the inner layer are ℓ^∞-bounded, while the outer layer is bounded in ℓ^1 (namely as a measure). Due to the homogeneity of the ReLU activation function, the bounds can be easily achieved and the function space is not reduced compared to just requiring the path-norm proxy to be finite.

For other activation functions, an ℓ^∞-bound on the coefficients in the inner layer may restrict the space of functions which can be approximated. In particular, if σ is C^k-smooth, then $x \mapsto a\sigma(w^T x)$ is C^k-smooth uniformly in $w \in B_R(0) \subseteq \mathbb{R}^{d+1}$. As a consequence, the space of σ-activated two-layer networks whose inner layer coefficients are ℓ^∞-bounded embeds continuously into C^k. At least if $k > d/2$, it follows from [5] that this space is smaller than the space of functions which can be approximated by σ-activated two-layer networks with uniformly bounded path-norm (see also [23, Theorem 3.1]).

It is likely that neural tree spaces with more general activation require parametrization by Radon measures on entire Banach spaces of functions. For networks with a single hidden layer, some results in this direction were presented in the appendix of [22]. While Radon measures on $\mathbb{R}^{d+2}$ are less convenient than those on S^{d+1}, many results can be carried over since $\mathbb{R}^{d+2}$ is locally compact.

The situation is very different for networks with two hidden layers. The space $X = \mathcal{W}^1$ on which $\mathcal{W}^2 = \mathcal{B}_X$ is modelled is infinite-dimensional, dense in C^0, and not locally compact in the C^0-topology. The restriction to the compact set B^X simplifies the analysis considerably.

3.2 Embedding of finite networks

The space $\mathcal{W}^L$ contains all finite networks with $L \geq 1$ hidden layers.

Theorem 3.1. *Let*

$$f(x)=\sum_{i_L=1}^{m_L} a_{i_L}^L \sigma\left(\sum_{i_{L-1}=1}^{m_{L-1}} a_{i_L i_{L-1}}^{L-1} \sigma\left(\sum \cdots \sigma\left(\sum_{i_1=1}^{m_1} a_{i_2 i_1}^1 \sigma\left(\sum_{i_0=1}^{d+1} a_{i_1 i_0}^0 x_{i_0}\right)\right)\right)\right). \tag{3.1}$$

Then $f \in \mathcal{W}^L$ *and*

$$\|f\|_{\mathcal{W}^L} \leq \sum_{i_L=1}^{m_L} \cdots \sum_{i_1=1}^{m_1} \sum_{i_0=1}^{d+1} \left|a_{i_L}^L a_{i_L i_{L-1}}^{L-1} \cdots a_{i_1 i_0}^0\right|. \tag{3.2}$$

Proof. The statement is obvious for $L=1$ as

$$f(x)=\sum_{i=1}^{m_1} a_{i_1} \sigma\left(\sum_{i_0=1}^{d+1} a_{i_1,i_0} x_{i_0}\right)=\int_{S^d} \sigma(w^T x)\left(\sum_{i=1}^{m} a_i |w_i| \cdot \delta_{w_i/|w_i|}\right)(\mathrm{d}w)$$

is a classical Barron function, where we simplified notation by setting $w_i = (a_{i1},\cdots,a_{i(d+1)}) \in \mathbb{R}^{d+1}$. We proceed by induction.

Let f be like in (3.1). By the induction hypothesis, for any fixed $1 \leq i_L \leq m_L$, the function

$$g_{I_L}(x) := \sum_{i_{L-1}=1}^{m_{L-1}} a_{i_L i_{L-1}}^{L-1} \sigma\left(\sum \cdots \sigma\left(\sum_{i_1=1}^{m_1} a_{i_2 i_1}^1 \sigma\left(\sum_{i_0=1}^{d+1} a_{i_1 i_0}^0 x_{i_0}\right)\right)\right)$$

lies in $\mathcal{W}^{L-1}$ with the appropriate norm bound. We note that

$$f(x)=\sum_{i_L=1}^{m_L} a_{i_L} \sigma\big(g_{i_L}(x)\big)=\sum_{i_L=1}^{m_L} \bar{a}_{i_L} \sigma\big(\bar{g}_{i_L}(x)\big),$$

where

$$\bar{g}_{i_L}=\frac{g_{i_L}}{\sum_{i_{L-1}=1}^{m_L} \cdots \sum_{i_1=1}^{m_1} \sum_{i_0=1}^{d+1} \left|a_{i_{L-1}}^L a_{i_{L-1} i_{L-2}}^{L-1} \cdots a_{i_1 i_0}^0\right|},$$

$$\bar{a}_{i_L}=a_{i_L} \sum_{i_{L-1}=1}^{m_L} \cdots \sum_{i_1=1}^{m_1} \sum_{i_0=1}^{d+1} \left|a_{i_{L-1}}^L a_{i_{L-1} i_{L-2}}^{L-1} \cdots a_{i_1 i_0}^0\right|.$$

It follows that $f \in \mathcal{W}^L$ with appropriate norm bounds. □

3.3 Inverse approximation

We show that $\mathcal{W}^L$ does not only contain all finite ReLU networks with L hidden layers, but also their limiting objects.

Theorem 3.2 (Compactness Theorem). *Let f_n be a sequence of functions in $\mathcal{W}^L$ such that $C^L := \liminf_{n\to\infty} \|f_n\|_{\mathcal{W}^L} < \infty$. Then there exists $f \in \mathcal{W}^L$ and a subsequence f_{n_k} such that $\|f\|_{\mathcal{W}^L} \leq C^L$ and $f_{n_k} \to f$ strongly in $C^{0,\alpha}(K)$ for all $\alpha < 1$.*

Proof. The result is trivial for $L=0$ since $\mathcal{W}^0$ is a finite-dimensional linear space. Using the third property from Theorem 2.1 inductively, we find that $\mathcal{W}^L$ embeds continuously into $C^{0,1}$, thus compactly into $C^{0,\alpha}$ for all $\alpha<1$. This establishes the existence of a convergent subsequence. Since $B^{\mathcal{W}^L}$ is C^0-closed, it follows that the limit lies in $\mathcal{W}^L$. □

Corollary 3.1 (Inverse Approximation Theorem)**.** *Let*

$$f_n(x)=\sum_{i_L=1}^{m_{n,L}} a_{i_L}^{n,L}\sigma\left(\sum_{i_{L-1}=1}^{m_{n,L-1}} a_{i_L i_{L-1}}^{n,L-1}\sigma\left(\sum\cdots\sigma\left(\sum_{i_1=1}^{m_{n,1}} a_{i_2 i_1}^{n,1}\sigma\left(\sum_{i_0=1}^{d+1} a_{i_1 i_0}^{n,0}x_{i_0}\right)\right)\right)\right)$$

be finite L-layer network functions such that

$$\sup_{n\in\mathbb{N}}\sum_{i_L=1}^{m_{n,L}}\cdots\sum_{i_1=1}^{m_{n,1}}\sum_{i_0=1}^{d+1}\left|a_{i_L}^{n,L}a_{i_L i_{L-1}}^{n,L-1}\cdots a_{i_1 i_0}^{n,0}\right|<\infty.$$

If $\mathbb{P}$ is a compactly supported probability measure and $f\in L^1(\mathbb{P})$ such that $f_n\to f$ in $L^1(\mathbb{P})$, then $f\in\mathcal{W}^L(\mathrm{spt}\mathbb{P})$ and

$$\|f\|_{\mathcal{W}^L(\mathrm{spt}\mathbb{P})}\leq\liminf_{n\to\infty}\sum_{i_L=1}^{m_{n,L}}\cdots\sum_{i_1=1}^{m_{n,1}}\sum_{i_0=1}^{d+1}\left|a_{i_L}^{n,L}a_{i_L i_{L-1}}^{n,L-1}\cdots a_{i_1 i_0}^{n,0}\right|. \tag{3.3}$$

Proof. Follows from Theorems 3.2 and 3.1. □

In particular, we make no assumption whether the width of any layer goes to infinity, or at what rate. The path-norm does not control the number of (non-zero) weights of a network.

3.4 Direct approximation

In Sections 3.2 and 3.3, we showed that $\mathcal{W}^L$ is large enough to contain all finite ReLU networks with L hidden layers and their limiting objects, even in weak topologies. In this section, we prove conversely that $\mathcal{W}^L$ is small enough such that every function can be approximated by finite networks with L hidden layers (with rate independent of the dimensionality), i.e. $\mathcal{W}^L$ is the smallest suitable space for these objects.

In fact, we prove a stronger result with an approximation rate in a reasonably weak topology. The rate however depends on the number of layers. Recall the following result on convex sets in Hilbert spaces.

Lemma 3.1. [5, Lemma 1] *Let $\mathcal{G}$ be a set in a Hilbert space H such that $\|g\|_H\leq R$ for all $g\in\mathcal{G}$. If f is in the closed convex hull of $\mathcal{G}$, then for every $m\in\mathbb{N}$ and $\varepsilon>0$, there exist m elements $g_1,\cdots,g_m\in\mathcal{G}$ such that*

$$\left\|f-\frac{1}{m}\sum_{i=1}^m g_i\right\|_H\leq\frac{R+\varepsilon}{\sqrt{m}}. \tag{3.4}$$

The result is attributed to Maurey in [5] and proved using the law of large numbers.

Theorem 3.3. *Let $\mathbb{P}$ be a probability measure with compact support $\mathrm{spt}(\mathbb{P}) \subseteq B_R(0)$. Then for any $L \geq 1$, $f \in \mathcal{W}^L$ and $m \in \mathbb{N}$, there exists a finite L-layer ReLU network*

$$f_m(x) = \sum_{i_L=1}^{m} a_{i_L}^L \sigma\left(\sum_{i_{L-1}=1}^{m^2} a_{i_L i_{L-1}}^{L-1} \sigma\left(\sum_{i_{L-2}=1}^{m^3} \cdots \sigma\left(\sum_{i_1=1}^{m^L} a_{i_2 i_1}^1 \sigma\left(\sum_{i_0=1}^{d+1} a_{i_1 i_0}^0 x_{i_0} \right) \right) \right) \right) \tag{3.5}$$

such that

1. $$\|f_m - f\|_{L^2(\mathbb{P})} \leq \frac{L(2+R)\|f\|_{\mathcal{W}^L}}{\sqrt{m}}; \tag{3.6}$$

2. *the norm bound*

$$\sum_{i_L=1}^{m} \cdots \sum_{i_1=1}^{m^L} \sum_{i_0=1}^{d+1} \left| a_{i_L}^L a_{i_L i_{L-1}}^{L-1} \cdots a_{i_1 i_0}^0 \right| \leq \|f\|_{\mathcal{W}^L} \tag{3.7}$$

holds.

Remark 3.2. Note that the width of deep layers increases rapidly. This is due to the fact that we construct an approximating network inductively. The procedure leads to a tree-like structure where parameters are not shared, but every neuron in the ℓ-th layer has its own set of parameters in the $\ell+1$-th layer and $a_{i_\ell i_{\ell-1}} = 0$ for all other parameter pairings. This is equivalent to standard architectures from the perspective of approximation theory under path-norm bounds, since the path norm does not control the number of neurons.

The total number of parameters in the network of the direct approximation theorem is

$$\begin{aligned} M &= m + m \cdot m^2 + \cdots + m^{L-1} \cdot m^L + m^L(d+1) \\ &= \sum_{\ell=0}^{L-1} m^{2\ell+1} + m^L(d+1) \\ &= m \frac{1-m^{2L}}{1-m^2} + m^L(d+1) \\ &\sim m^{2L-1} \end{aligned}$$

by the geometric sum. Thus the decay rate in the direct approximation theorem is of the order $M^{-\frac{1}{2(2L-1)}}$. This recovers the Monte-Carlo rate $M^{-1/2}$ in the case $L=1$ [18, Theorem 4], but quickly degenerates as L increases. Part of the problem is that the rapidly branching structure combined with neural network indexing induces explicitly listed zeros in the set of weights as explained in Section 2.1. A neural tree expressing the same function would require only $\sim (d+L)m^L$ weights.

Note, however, that the approximation rate is independent of dimension d. In this sense, we are not facing a curse of dimensionality, but a curse of depth.

It is unclear whether this rate can be improved in the general setting. Functions in Barron space are described as the expectation of a suitable quantity, while multi-layer functions are described as iterated conditional expectations and non-linearities. In this setting, it is not obvious whether the Monte-Carlo rate should be expected.

Proof of Theorem 3.3. Without loss of generality $\|f\|_{\mathcal{W}^L}=1$. Since $\mathcal{W}^L \hookrightarrow C^{0,1}$ with constant 1, we find that $\|f\|_{L^2(\mathbb{P})} \leq (1+R)\|f\|_{\mathcal{W}^L}$ for all $f \in \mathcal{W}^L$.

Recall from the proof of Theorem 2.2 that the unit ball of $\mathcal{W}^L$ is the closed convex hull of the class $\mathcal{H} = \{\pm\sigma(g) : \|g\|_{\mathcal{W}^{L-1}} \leq 1\}$. Thus by Lemma 3.1 there exist $g_1, \cdots g_M \in \mathcal{W}^{L-1}$ and $\varepsilon_i \in \{-1,1\}$ such that

$$\left\| f - \frac{1}{m}\sum_{i=1}^m \varepsilon_i \sigma(g_i(x)) \right\|_{L^2(\mathbb{P})} < \frac{2+R}{\sqrt{m}}.$$

If $L=1$, g_i is an affine linear map and $f_m(x) = \sum_{i=1}^m \frac{\varepsilon_i}{m}\sigma(g_i(x))$ is a finite neural network. Thus the theorem is established for $L=1$.

We proceed by induction. Assume that the theorem has been proved for $L-1 \geq 1$. Then we note that $\|g_i\|_{\mathcal{W}^{L-1}} \leq 1$, so for $1 \leq i \leq m$ we can find a finite $L-1$-layer network $\tilde{g}_i$ such that

$$\begin{aligned}
&\left\| f - \frac{1}{m}\sum_{i=1}^m \varepsilon_i \sigma(\tilde{g}_i(x)) \right\|_{L^2(\mathbb{P})} \\
\leq &\left\| f - \frac{1}{m}\sum_{i=1}^m \varepsilon_i \sigma(g_i(x)) \right\|_{L^2(\mathbb{P})} + \frac{1}{m}\sum_{i=1}^m \|g_i - \tilde{g}_i\|_{L^2(\mathbb{P})} \\
\leq &\frac{2+R}{\sqrt{m}} + \frac{m}{m}\frac{(L-1)(2+R)}{\sqrt{m}}.
\end{aligned}$$

We merge the m trees associated with $\tilde{g}_i$ into a single tree, increasing the width of each layer by a factor of m, and add an outer layer of width m with coefficients $a_{i_L} = \frac{\varepsilon_{i_L}}{m}$. □

Remark 3.3. Let $p \in [2,\infty)$. Then by interpolation

$$\begin{aligned}
\|f - f_m\|_{L^p(\mathbb{P})} &\leq \|f - f_m\|_{L^2(\mathbb{P})}^{\frac{2}{p}} \|f - f_m\|_{L^\infty(\mathbb{P})}^{1-\frac{2}{p}} \\
&\leq C\|f\|_{\mathcal{W}^L} m^{-1/p}.
\end{aligned}$$

Corollary 3.2. *For every compact set K and $f \in \mathcal{W}^L(K)$, there exists a sequence of finite neural networks with L hidden layers*

$$f_n(x) = \sum_{i_L=1}^{m_{n,L}} a_{i_L}^{n,L} \sigma\left(\sum_{i_{L-1}=1}^{m_{n,L-1}} a_{i_L i_{L-1}}^{n,L-1} \sigma\left(\sum_{i_{L-2}} \cdots \sigma\left(\sum_{i_1=1}^{m_{n,1}} a_{i_2 i_1}^{n,1} \sigma\left(\sum_{i_0=1}^{d+1} a_{i_1 i_0}^{n,0} x_{i_0} \right) \right) \right) \right)$$

such that $\|f_n\|_{\mathcal{W}^L} \leq \|f\|_{\mathcal{W}^L}$ *and* $f_n \to f$ *in* $C^{0,\alpha}(K)$ *for every* $\alpha < 1$.

Proof. We take $R > 0$ such that $K \subseteq B_R(0)$ and take $\mathbb{P}$ to be the uniform distribution on $B_R(0)$. By Theorem 3.3, we can approximate f in $L^2(\mathbb{P})$ with the norm bound $\|f_n\|_{\mathcal{W}^L} \leq \|f\|_{\mathcal{W}^L}$.

By compactness, we find that f_n converges to a limit in $C^{0,\alpha}(\overline{B_R(0)})$ for all $\alpha < 1$, which coincides with the $L^2(\mathbb{P})$-limit f. In particular, f_n converges in $C^{0,\alpha}(K)$. We can eliminate the ε in the norm bound by a diagonal sequence argument. □

Remark 3.4. The direct and indirect approximation theorems show that neural tree spaces are the correct function spaces for neural networks under path-norm bounds. The construction of vector spaces and proofs made ample use of the equivalence between neural networks and neural trees. It is tempting to try to force more classical neural network structures by prescribing that the width of all layers tends to infinity at the same rate. However, this does not change the approximation spaces since in the direct approximation theorem, we can repeat a function from the approximating sequence multiple times until the width of the most restrictive layer is sufficiently large to pass to the next element in the sequence. A more successful approach is discussed in Section 4.

3.5 Composition of multi-layer functions

Let $f \in (\mathcal{W}^L(K))^k$ be an L-layer function with values in $\mathbb{R}^k$. Since K is compact and f is continuous, $f(K)$ is also compact. Let $g \in \mathcal{W}^\ell(f(K))$ be an ℓ-layer function on $f(K)$.

Lemma 3.2. $g \circ f \in \mathcal{W}^{L+\ell}(K)$ *and*

$$\|g \circ f\|_{\mathcal{W}^{L+\ell}(K)} \leq \|g\|_{\mathcal{B}^\ell(f(K))} \sup_{1 \leq i \leq k} \|f_i\|_{\mathcal{B}^L(K)}. \tag{3.8}$$

Proof. We proceed by induction. First consider the case $\ell = 0$. Then $g(x) = w^T f(x)$, so $w^T f(x) = \sum_{i=1}^k w_i f_i(x)$ is a (weighted) sum of L-layer functions, i.e. an L-layer function. By the triangle inequality we have

$$\begin{aligned}
\|g \circ f\|_{\mathcal{W}^L} &\leq \sum_{i=1}^k |w_i| \|f_i\|_{\mathcal{W}^L} \\
&\leq \|w\|_{\ell^1} \sup_{1 \leq i \leq k} \|f_i\|_{\mathcal{W}^L} \\
&= \|g\|_{\mathcal{W}^0} \sup_{1 \leq i \leq k} \|f_i\|_{\mathcal{W}^L}.
\end{aligned}$$

Now assume that the theorem has been proved for $\ell - 1$ with $\ell \geq 1$. To avoid double superscripts, denote by B^ℓ the closed unit ball in $\mathcal{W}^\ell(f(K))$. Let $g(z) = \int_{B^{\ell-1}} \sigma(h(z))\, \mu(\mathrm{d}h)$.

Then

$$\begin{aligned}(g\circ f) &= \int_{B^{\ell-1}} \sigma\big((h\circ f)\big)\,\mu(\mathrm{d}h)\\ &= \left(\sup_{1\le i\le k}\|f_i\|_{\mathcal{B}^L}\right)\int_{B^{\ell-1}}\sigma\left(\frac{h\circ f}{\sup_{1\le i\le k}\|f_i\|_{\mathcal{B}^L}}\right)\mu(\mathrm{d}h)\\ &= \left(\sup_{1\le i\le k}\|f_i\|_{\mathcal{B}^L}\right)\int_{B^{L+\ell-1}}\sigma(j(\cdot))\,(F_\sharp\mu)(\mathrm{d}j),\end{aligned}$$

where

$$F: B^{\ell-1}\to B^{L+\ell-1},\qquad F(h)=\frac{h\circ f}{\sup_{1\le i\le k}\|f_i\|_{\mathcal{W}^L}}$$

is well-defined by the induction hypothesis. By definition, $g\circ f\in\mathcal{W}^{L+\ell}$ with the appropriate norm bound. □

For generalized Barron spaces, we showed that $\|f\|_{X,K}\le 2\|f\|_X$ for all $f\in X$, thus by induction $\|f\|_{\mathcal{W}^{\ell+L}}\le 2^\ell\|f\|_{\mathcal{W}^L}$ for $L\ge 1$. We show that this naive bound can be improved to be independent of the number of additional layers.

Lemma 3.3. *Let $\ell, L\ge 1$ and $f\in\mathcal{W}^L(K)$. Then $f\in\mathcal{W}^{\ell+L}(K)$ and $\|f\|_{\mathcal{W}^{\ell+L}(K)}\le 2\|f\|_{\mathcal{W}^L(K)}$.*

Proof. Without loss of generality, $\|f\|_{\mathcal{W}^L(K)}\le 1$. We note that $g_1=\sigma(f)$ and $g_2=\sigma(-f)$ are both in the unit ball of $\mathcal{W}^{L+1}$ and non-negative, i.e. $g_1=\sigma(g_1)$ and $g_2=\sigma(g_2)$. Thus g_1, g_2 are also in the unit ball of $\mathcal{W}^{L+2}$. By induction, we observe that $\|g_i\|_{\mathcal{W}^{L+\ell}(K)}\le 1$ for all $\ell\ge 1$, $i=1,2$ and thus

$$\|f\|_{\mathcal{W}^{L+\ell}(K)}=\|g_1+g_2\|_{\mathcal{W}^{L+\ell}(K)}\le\|g_1\|_{\mathcal{W}^{L+\ell}(K)}+\|g_2\|_{\mathcal{W}^{L+\ell}(K)}\le 2. \tag{3.9}$$

This completes the proof. □

3.6 Rademacher complexity

Considering statistical learning theory, neural tree spaces inherit the convenient properties of the space of affine functions. These convenient properties are one of the reasons why we study the path-norm in the first place. Recall the definition and discussion of Rademacher complexities from Section 2.4.

Lemma 3.4. *For every L, and every set of N points $S\subseteq[-1,1]^d$, the hypothesis class $\mathcal{H}^L$ given by the closed unit ball in $\mathcal{W}^L$ satisfies the Rademacher complexity bound*

$$\mathrm{Rad}\left(\mathcal{H}^L; S\right)\le 2^L\sqrt{\frac{2\log(2d+2)}{N}}. \tag{3.10}$$

Proof. This follows directly from Example 2.3 and Theorem 2.2 by induction. □

The complexity bound has an immediate application in statistical learning theory.

Corollary 3.3 (Generalization gap)**.** *Let $\mathbb{P}$ be any probability distribution supported on $[-1,1]^d \times \mathbb{R}$ and $(X_1,Y_1)\cdots,(X_N,Y_N)$ be iid random variables with law $\mathbb{P}$. Consider the hypothesis space $\mathcal{H} = \{h \in \mathcal{W}^L(K) : \|h\|_{\mathcal{W}^L(K)} \leq 1\}$. Assume that $\ell : \mathbb{R}\times\mathbb{R} \to [0,\bar{c}]$ is a bounded loss function. Then, with probability at least $1-\delta$ over the choice of data points $X_1,\cdots,X_N$, the estimate*

$$\begin{aligned}&\sup_{h\in\mathcal{H}}\left|\frac{1}{N}\sum_{i=1}\ell\big(h(X_i),Y_i\big) - \int_{[-1,1]^d\times\mathbb{R}} \ell\big(h(x),y\big)\,\mathbb{P}(\mathrm{d}x\otimes\mathrm{d}y)\right| \\ \leq& 2^{L+1}\sqrt{\frac{2\log(2d+2)}{N}} + \bar{c}\sqrt{\frac{2\log(2/\delta)}{N}} \end{aligned} \tag{3.11}$$

holds.

Proof. This follows directly from Lemma 3.4 and [40, Theorem 26.5]. □

Thus it is easy to "learn" a multi-layer function with low path norm in the sense that a relatively small size of sample data points is sufficient to understand whether the function has low population risk or not. More sophisticated methods can provide dimension-dependent decay rates $1/2+1/(2d+2)$ of the generalization error at the expense of constants scaling like $\sqrt{d}$ instead of $\log(d)$ [6, Remark 1].

3.7 Generalization error estimates for regularized model

As an application, we prove that empirical risk minimization with explicit regularization is a successful strategy in learning multi-layer functions. For technical reasons, we work with a bounded modification of L^2-risk instead of the mean squared error functional.

Let $\mathbb{P}$ be a probability measure on $[-1,1]^d$ and $S = \{x_1,\cdots,x_N\}$ be a set of samples drawn iid from $\mathbb{P}$. Denote

$$\mathcal{R},\mathcal{R}_n : \mathcal{W}^L \to \mathbb{R}, \qquad \mathcal{R}(f) = \int_{\mathbb{R}^d} \ell(x,f(x))\,\mathbb{P}(\mathrm{d}x), \qquad \mathcal{R}_N(f) = \frac{1}{N}\sum_{i=1}^N \ell(x_i,f(x_i)),$$

where the loss function ℓ satisfies

$$\ell(x,y) \leq \min\big\{\bar{c}, |y-f^*(x)|^2\big\}.$$

For finite neural networks with weights $(a^L,\cdots,a^0) \in \mathbb{R}^{m_L}\times\cdots\times\mathbb{R}^{m_1\times d}$ we denote

$$\begin{aligned}&\widehat{\mathcal{R}}_N(a^L,\cdots,a^0) = \mathcal{R}_N(f_{a^L,\cdots,a^0}), \\ &f_{a^L,\cdots,a^0}(x) = \sum_{i_L=1}^{m_L} a^L_{i_L}\sigma\left(\sum_{i_{L-1}=1}^{m_{L-1}} a^{L-1}_{i_L i_{L-1}}\sigma\left(\sum_{\cdots}\cdots\sigma\left(\sum_{i_1=1}^{m_1} a^1_{i_2 i_1}\sigma\left(\sum_{i_0=1}^{d+1} a^0_{i_1 i_0}\,x_{i_0}\right)\right)\right)\right).\end{aligned}$$

Theorem 3.4 (Generalization error)**.** *Assume that the target function satisfies $f^* \in \mathcal{W}^L$. Let $\mathcal{F}_m$ be the class of neural networks with architecture like in Theorem* 3.3*. The minimizer $f_m \in \mathcal{F}_m$ of the regularized risk functional*

$$\widehat{\mathcal{R}}_n(a^L,\cdots,a^0)+\frac{9L^2}{m}\left[\sum_{i_L=1}^{m_L}\cdots\sum_{i_0=1}^{d+1}\left|a^L_{i_L}a^{L-1}_{i_Li_{L-1}}\cdots a^0_{i_1i_0}\right|\right]^2$$

satisfies the risk bound

$$\mathcal{R}(f_m)\leq\frac{18L^2\|f^*\|^2_{\mathcal{W}^L}}{m}+2^{L+3/2}\|f^*\|_{\mathcal{W}^L}\sqrt{\frac{2\log(2d+2)}{N}}+\bar{c}\sqrt{\frac{2\log(2/\delta)}{N}}. \tag{3.12}$$

The first term comes from the direct approximation theorem. The explicit scaling in L looks unproblematic, but recall that the network requires $\sim m^{2L-1}$ parameters. The second term stems from the Rademacher bound and is subject to the 'curse of depth'. An improvement in either term would lead to better a priori estimates. The third term is purely probabilistic and unproblematic.

Proof of Theorem 3.4. Denote $\lambda=\lambda_m=9L^2m^{-1}$ and let $\hat{f}_m=f_{\hat{a}_L,\cdots,\hat{a}_0}$ be like in Theorem 3.3, i.e.

$$\|\hat{f}_m-f^*\|_{L^2(\mathbb{P}_n)}\leq\frac{3L\|f^*\|_{\mathcal{W}^L}}{\sqrt{m}},\qquad\sum_{i_L,\cdots,i_0}\left|a^L_{i_L}\cdots a^0_{i_1i_0}\right|\leq\|f^*\|_{\mathcal{W}^L}.$$

Then by definition

$$\begin{aligned}&\widehat{\mathcal{R}}_n(a^L,\cdots,a^0)+\lambda\left[\sum_{i_L=1}^{m_L}\cdots\sum_{i_0=1}^{d+1}\left|a^L_{i_L}a^{L-1}_{i_Li_{L-1}}\cdots a^0_{i_1i_0}\right|\right]^2\\ \leq&\widehat{\mathcal{R}}_n(\hat{a}^L,\cdots,\hat{a}^0)+\lambda\left[\sum_{i_L=1}^{m_L}\cdots\sum_{i_0=1}^{d+1}\left|\hat{a}^L_{i_L}\hat{a}^{L-1}_{i_Li_{L-1}}\cdots\hat{a}^0_{i_1i_0}\right|\right]^2\\ \leq&\frac{9L^2\|f^*\|^2_{\mathcal{W}^L}}{m}+\lambda\|f^*\|^2_{\mathcal{W}^L}.\end{aligned}$$

In particular

$$\|f_{a^L,\cdots,a^0}\|^2_{\mathcal{W}^L}\leq\left[\sum_{i_0=1}^{d+1}\left|a^L_{i_L}a^{L-1}_{i_Li_{L-1}}\cdots a^0_{i_1i_0}\right|\right]^2\leq\frac{2\lambda\|f^*\|^2_{\mathcal{W}^L}}{\lambda}=2\|f^*\|^2_{\mathcal{W}^L}.$$

The Rademacher complexity is the supremum of linear random variables, so $\mathrm{Rad}(B^L_R;S)=R\cdot\mathrm{Rad}(B^L_1;S)$ where B^L_R denotes the ball of radius R centered at the origin in $\mathcal{W}^L$. We

conclude that, with probability at least $1-\delta$ over the draw of the training sample, we have

$$\begin{aligned}\mathcal{R}(f_{a^L,\cdots,a^0}) &\leq \mathcal{R}_n(f_{a^L,\cdots,a^0}) + 2^{L+3/2}\|f^*\|_{\mathcal{W}^L}\sqrt{\frac{2\log(2d+2)}{N}} + \bar{c}\sqrt{\frac{2\log(2/\delta)}{N}} \\ &= \frac{18L^2\|f^*\|_{\mathcal{W}^L}^2}{m} + 2^{L+3/2}\|f^*\|_{\mathcal{W}^L}\sqrt{\frac{2\log(2d+2)}{N}} + \bar{c}\sqrt{\frac{2\log(2/\delta)}{N}}.\end{aligned}$$

This completes the proof. □

Remark 3.5. Since $\|f\|_{L^\infty} \leq \big(1+\sup_{x\in K}|x|\big)\|f\|_{\mathcal{W}^L}$ for all $f \in \mathcal{W}^L(K)$, we can repeat the argument for the loss function $\ell(x,y) = |y-f^*(x)|^2$, which is a priori unbounded, but can be modified outside of the interval which f_m, f^* take values in due to the a priori norm bound. The constant $\bar{c}$ in (3.12) in this case is

$$\bar{c} = 4\|f^*\|^2_{\mathcal{W}^L([-1,1]^d)}.$$

Remark 3.6. For large L, these bounds degenerate rapidly. In [6], the authors show that under the stronger condition that a balanced version of the path-norm (which measures the average weights of incoming and outcoming paths at all nodes in all layers), a better bound on the Rademacher complexity is available. The balanced path norm achieves control over cancellations and the balancing of weights at different layers.

Heuristically, the proof proceeds as follows: Let $S = \{x_1,\cdots,x_N\}$ be a sample set in $[-1,1]^d$ and the hypothesis space $\mathcal{H}$ be given by the unit ball in $\mathcal{W}^L$. By the direct approximation theorem, there exists a network with $O(m^{2L})$ weights which approximates f with $\|f\|_{\mathcal{W}^L} \leq 1$ to accuracy $\sim \frac{L}{\sqrt{m}}$ in $L^2(\mathbb{P}_N)$ where $\mathbb{P}_N$ is the uniform measure on S. Thus the covering number $\overline{N}_{\varepsilon,L^2(\mathbb{P}_N)}(\mathcal{H})$ of $\mathcal{H}$ in the $L^2(\mathbb{P}_N)$-distance *should* scale like $L\varepsilon^{-4L}$.

Since $f(x_1),\cdots f(x_N) \subset B_{\sqrt{m}}(0) \subseteq \mathbb{R}^N$ (with respect to the Euclidean distance), the Rademacher complexity can be bounded by

$$\begin{aligned}\mathrm{Rad}(\mathcal{H};S) &\leq \frac{2^{-K}\sqrt{m}}{\sqrt{m}} + \frac{6\sqrt{m}}{m}\sum_{i=1}^{K} 2^{-i}\sqrt{\log\big(\overline{N}_{2^{-i}\sqrt{m},L^2(\mathbb{P}_N)}\big)} \\ &\leq 2^{-K} + \frac{C}{\sqrt{m}}\sum_{i=1}^{K} 2^{-i}\sqrt{\log\big(Lm^{-2L}2^{4iL}\big)} \\ &\approx \frac{2^{-K}}{\sqrt{m}} + \frac{C\sqrt{L}}{\sqrt{m}}\sum_{i=1}^{K} 2^{-i}\sqrt{i}\end{aligned}$$

for any $K \in \mathbb{N}$ using [40, Lemma 27.1]. Taking $K \to \infty$, only $\sqrt{L}$ enters in the estimate. The point in the proof that needs to be made rigorous is the connection between covering the parameter space with an ε-fine net and covering the function class with an ε-fine net. For a neural network

$$f(x) = \varepsilon^2\sigma\left(\frac{1}{\varepsilon}x\right)$$

the path-norm is bounded, but an ε-small change in the outer layer would lead to a large change in the function space. Thus a balanced version of the path-norm is needed. In some cases, this may be possible through rescaling layers, but see Remark 4.5 for a possible obstruction. Similar ideas are explored below in Section 4.2, although we do not estimate the Rademacher complexity explicitly.

The ability to obtain Rademacher estimates from covering also suggests that improvements in the direct approximation theorem may not be possible, since the complexity of the function classes should increase with increasing depth.

Unfortunately, the convenience in learning functions comes at a price when considering the approximation power of neural tree spaces as described in [22, Corollary 3.4] for general function classes of low complexity.

Corollary 3.4. *For any $d \geq 3$, there exists a 1-Lipschitz function ϕ on $[0,1]^d$ such that*

$$\limsup_{t\to\infty} \left(t^\gamma \inf_{\|f\|_X \leq t} \|\phi - f\|_{L^2(Q)} \right) = \infty \tag{3.13}$$

for all $\gamma > \frac{2}{d-2}$.

Thus to approximate even relatively regular functions in a fairly weak topology up to accuracy ε, the path-norm of a network with L hidden layers may have to grow (almost) as quickly as $\varepsilon^{-\frac{d-2}{2}}$ independently of L. In particular, increasing the depth of an infinitely wide network does not increase the approximation power sufficiently to approximate general Lipschitz functions (while the path norm remains bounded by the same constant).

3.8 Countably wide neural networks

Let us briefly comment on another natural concept of infinitely wide neural networks. The space of countably wide networks

$$f(x) = \sum_{i_L=1}^{\infty} a_{i_L}^L \sigma \left(\sum_{i_{L-1}=1}^{\infty} a_{i_L i_{L-1}}^{L-1} \sigma \left(\sum_{i_{L-2}} \cdots \sigma \left(\sum_{i_1=1}^{\infty} a_{i_2 i_1}^1 \sigma \left(\sum_{i_0=1}^{d+1} a_{i_1 i_0}^0 x_{i_0} \right) \right) \right) \right)$$

equipped with the path-norm

$$\|f\| = \inf_a \sum_{i_L} \cdots \sum_{i_0} \left| a_{i_L}^L \cdots a_{i_1 i_0}^0 \right|$$

is a subspace of $\mathcal{W}^L$ by the same reasoning as Theorem 3.1 and the fact that the cross-product of a finite number of countable sets is countable. Like in the introduction, we can show that the spaces of countably wide neural networks and neural trees coincide.

Unlike finite neural networks, countable networks form a vector space. The space of countably wide networks is a proper subspace of $\mathcal{W}^L$ which contains all finite neural networks. The direct approximation theorem implies that the unit ball in the space of countably wide neural networks is not closed in weaker topologies like $C^{0,\alpha}$ or L^p. Thus the space of countably wide neural networks is not suitable from the perspective of variational analysis.

Intuitively, any convergent infinite sum contains a finite number of macroscopic terms and an infinite tail of rapidly decaying terms. Thus at initialization and throughout training, a scale difference would exist in a countable neural network between leading order neurons and tail neurons. This is not a useful way to think of neural networks where parameters in a fixed layer are typically chosen randomly from the same distribution and then optimized by gradient flow-type algorithms. It should be noted however that common schemes like Xavier initialization [25] choose the weights in a fashion which makes the path-norm grows beyond all bounds as the number of neurons goes to infinity.

4 Indexed representation of arbitrarily wide neural networks

4.1 Neural networks with general index sets

The spaces considered above are a bit abstract. In this section, we discuss a more concrete representation for a subspace of $\mathcal{W}^L$. As we show below, this subspace is invariant under the gradient flow dynamics. For all practical purposes, it might just be the right set of functions that we need to consider.

In [23, Section 2.8], we showed that $f:\mathbb{R}^d\to\mathbb{R}$ is a Barron function if and only if there exist measurable maps $a,b:(0,1)\to\mathbb{R}$ and $w:(0,1)\to\mathbb{R}^d$ such that

$$f(x)=f_{a,w,b}(x)=\int_0^1 a_\theta\sigma\big(w_\theta^T x+b_\theta\big)\,\mathrm{d}\theta.$$

Furthermore

$$\|f\|_{\mathcal{B}(K)}=\inf\left\{\int_0^1|a|\left[|w|+|b|\right](\theta)\,\mathrm{d}\theta\,\middle|\,f=f_{a,w,b}\text{ on }K\right\}.$$

Thus we can think of Barron space as replacing the finite sum over neurons by an integral and replacing the index set $\{1,\cdots,m\}$ by the (continuous) unit interval. We extend the approach to multi-layer networks in some generality.

Definition 4.1. *For $0\le i\le L$, let $(\Omega_i,\mathcal{A}_i,\pi^i)$ be probability spaces where $\Omega_0=\{0,\cdots,d\}$ and π^0 is the normalized counting measure. Consider measurable functions $a^L:\Omega_L\to\mathbb{R}$ and $a^i:\Omega_{i+1}\times\Omega_i\to\mathbb{R}$ for $0\le i\le L-1$. Then define*

$$\begin{aligned}&f_{a^L,\cdots,a^0}(x)\\&=\int_{\Omega_L}a^{(L)}_{\theta_L}\sigma\left(\int_{\Omega_{L-1}}\cdots\sigma\left(\int_{\Omega_1}a^1_{\theta_2,\theta_1}\sigma\left(\int_{\Omega_0}a^0_{\theta_1,\theta_0}x_{\theta_0}\pi^0(\mathrm{d}\theta_0)\right)\pi^1(\mathrm{d}\theta_1)\right)\cdots\pi^{(L-1)}(\mathrm{d}\theta_{L-1})\right)\pi^L(\mathrm{d}\theta_L).\end{aligned}\tag{4.1}$$

Consider the norm

$$\begin{aligned}&\|f\|_{\Omega_L,\cdots,\Omega_0;K}\\ =&\inf\left\{\int_{\prod_{i=0}^L\Omega_i}\left|a_{\theta_L}^{(L)}\cdots a_{\theta_1\theta_0}^{(0)}\right|\left(\pi^L\otimes\cdots\otimes\pi^0\right)\left(\mathrm{d}\theta_L\otimes\cdots\otimes\mathrm{d}\theta_0\right)\middle| f=f_{a^L,\cdots,a^0}\text{ on }K\right\}.\end{aligned}\tag{4.2}$$

As usual, we set

$$X_{\Omega_L,\cdots,\Omega_0;K}=\{f\in C^{0,1}(K):\|f\|_{\Omega_L,\cdots,\Omega_0;K}<\infty\}.\tag{4.3}$$

We call $X_{\Omega_L,\cdots,\Omega_0;K}$ the class of neural networks over K modeled on the index spaces $\Omega_i=(\Omega_i,\mathcal{A}_i,\pi^i)$.

The representation in (4.1) can also be written as:

$$f(\mathbf{x})=\mathbb{E}_{\theta_L\sim\pi_L}a_{\theta_L}^{(L)}\sigma(\mathbb{E}_{\theta_{L-1}\sim\pi_{L-1}}\cdots\sigma(\mathbb{E}_{\theta_1\sim\pi_1}a_{\theta_2,\theta_1}^1\sigma(a_{\theta_1}^0\cdot\mathbf{x}))\cdots).\tag{4.4}$$

Representing functions as some form of expectations is the starting point for the continuous formulation of machine learning.

As we mentioned above, $\mathcal{W}^1(K)=X_{(0,1),\{0,\cdots,d\}}$ where the unit interval is equipped with Lebesgue measure. The collection of *finite* neural networks is realized when all sigma-algebras $\mathcal{A}_i$ contain only finitely many sets (in particular, if all probability spaces are finite). In this situation, $X_{\Omega_L,\cdots,\Omega_0}$ is not even a vector space.

Lemma 4.1. *1. For $L\geq 2$ and any selection of probability spaces $\Omega_L,\cdots,\Omega_1$, the space of neuronal embeddings $X_{\Omega_L,\cdots,\Omega_0;K}$ is a subset of the neural tree space $\mathcal{W}^L(K)$.*

2. *If $\Omega_i=(0,1)$ and π^i is Lebesgue measure for all $i\geq 1$, then $X_{\Omega_L,\cdots,\Omega_0;K}$ is a vector-space and $\|\cdot\|_{\Omega_L,\cdots,\Omega_0;K}$ is a norm on it.*

3. *If $\Omega_i=(0,1)$ and π^i is Lebesgue measure for all $i\geq 1$, then $X_{\Omega_L,\cdots,\Omega_0;K}$ contains all finite neural networks with L hidden layers. In particular, $X_{\Omega_L,\cdots,\Omega_0;K}$ is a subspace of $\mathcal{W}^L(K)$ which is dense in $\mathcal{W}^L(K)$ with respect to the $C^{0,\alpha}$-topology for all $\alpha<1$ and consequently in $L^p(\mathbb{P})$ for any probability measure on K, $p\in[1,\infty]$.*

4. *$X_{(0,1),(0,1),\{0,\cdots,d\};K}$ contains Barron space $\mathcal{W}^1(K)$ and*

$$\|f\|_{(0,1),(0,1),\{0,\cdots,d\};K}\leq 2\|f\|_{\mathcal{W}^1(K)}\qquad\forall\, f\in\mathcal{W}^1(K).$$

5. *Let $f\in\left(\mathcal{W}^1(K)\right)^k$ be a vector-valued Barron function and $g\in\mathcal{W}^1(\mathbb{R}^k)$ a scalar-valued Barron function. Then the composition $g\circ f$ lies in $X_{(0,1),(0,1),\{0,\cdots,d\};K}$ and*

$$\|g\circ f\|_{(0,1),(0,1),\{0,\cdots,d\};K}\leq\|g\|_{\mathcal{W}^1(\mathbb{R}^k)}\sum_{i=1}^k\|f_i\|_{\mathcal{W}^1(K)}.\tag{4.5}$$

In particular, this includes

- *the absolute value/positive part/negative of a Barron function,*
- *the pointwise maximum/minimum of two Barron functions,*
- *the product of two Barron functions.*

Proof. **First claim.** This can be proved exactly like Theorem 3.1.

Second claim. For any choice of parameter spaces Ω_i, the set of functions $X_{\Omega_L,\cdots,\Omega_0;K}$ is a balanced cone, i.e. if $f \in X_{\Omega_L,\cdots,\Omega_0;K}$ then $\lambda f \in X_{\Omega_L,\cdots,\Omega_0;K}$ for all $\lambda \in \mathbb{R}$. It remains to show that $X_{\Omega_L,\cdots,\Omega_0;K}$ is closed under function addition. Let

$$f(x) = \int_0^1 a^L_{\theta_L} \sigma\left(\int_0^1 \cdots \sigma\left(\int_0^1 a^1_{\theta_2\theta_1}\sigma\left(\frac{1}{d+1}\sum_{\theta_0=1}^{d+1} a^0_{\theta_1\theta_0} x_{\theta_0}\right)\right)\right),$$
$$g(x) = \int_0^1 b^L_{\theta_L} \sigma\left(\int_0^1 \cdots \sigma\left(\int_0^1 b^1_{\theta_2\theta_1}\sigma\left(\frac{1}{d+1}\sum_{\theta_0=1}^{d+1} b^0_{\theta_1\theta_0} x_{\theta_0}\right)\right)\right).$$

Then

$$(f+g)(x) = \int_0^1 c^L_{\theta_L} \sigma\left(\int_0^1 \cdots \sigma\left(\int_0^1 c^1_{\theta_2\theta_1}\sigma\left(\frac{1}{d+1}\sum_{\theta_0=1}^{d+1} c^0_{\theta_1\theta_0} x_{\theta_0}\right)\right)\right),$$

where

$$c^L_\theta = \begin{cases} 2a^L_{2\theta}, & \theta \in (0,1/2), \\ 2b^L_{2\theta-1}, & \theta \in (1/2,1), \end{cases} \qquad c^\ell_{\theta\xi} = \begin{cases} 4a^\ell_{2\theta,2\xi}, & \theta,\xi \in (0,1/2), \\ 4b^\ell_{2\theta-1,2\xi-1}, & \theta,\xi \in (1/2,1), \\ 0, & \text{else.} \end{cases}$$

Essentially, we construct two parallel networks that are added in the final layer and otherwise do not interact. The pre-factors stem from the fact that we re-arrange a mean-field index set and could be eliminated if we chose more general measure spaces (e.g. $\mathbb{Z}$ or $\mathbb{R}$) as index sets.

Third claim. Any finite neural network can be written as a mean field neural network

$$f(x) = \frac{1}{m_L}\sum_{i_L=1}^{m_L} a^L_{i_L}\sigma\left(\frac{1}{m_{L-1}}\sum_{i_{L-1}=1}^{m_{L-1}} a^{L-1}_{i_L i_{L-1}}\sigma\left(\cdots\sigma\left(\frac{1}{m_1}\sum_{i_1=1}^{m_1} a^1_{i_2 i_1}\sigma\left(\frac{1}{d+1}\sum_{i_0=1}^{d+1} a^0_{i_1 i_0} x_{i_0}\right)\right)\right)\right)$$

Define the functions

$$a^L : (0,1) \to \mathbb{R}, \quad a^L(s) = a^L_i \quad \text{for } \frac{i-1}{m_L} \le s < \frac{i}{m_L},$$
$$a^\ell : (0,1)^2 \to \mathbb{R}, \quad a^\ell(r,s) = a^\ell_{ij} \quad \text{for } \frac{i-1}{m_{\ell+1}} \le r < \frac{i}{m_L}, \frac{j-1}{m_\ell} \le s < \frac{i}{m_\ell},$$

for $0 \le \ell < L$. Then $f = f_{a^L,\cdots,a^0}$.

Fourth claim. Let f be a Barron function. Then, according to [23, Section 2.8], f can be written as

$$f(x)=\int_0^1 \bar{a}^1_\theta \sigma\left(\frac{1}{d+1}\sum_{i=1}^{d+1} a^0_{\theta,i}x_i\right)\mathrm{d}\theta.$$

For $\bar{a}^1, a^0 \in L^2(0,1)$. In particular,

$$f(x)=\int_0^1 a^2_{\theta_2}\sigma\left(\int_0^1 a^1_{\theta_2\theta_1}\sigma\left(\frac{1}{d+1}\sum_{i=1}^{d+1} a^0_{\theta_1,i}x_i\right)\mathrm{d}\theta_1\right)\mathrm{d}\theta_2,$$

where

$$a^2_{\theta_2}=\begin{cases}2, & \theta_2<1/2,\\ -2, & \theta_2>1/2,\end{cases}\qquad a^1_{\theta_2\theta_1}=\begin{cases}\bar{a}_{\theta_1}, & \theta_2<1/2,\\ -\bar{a}_{\theta_1}, & \theta_2>1/2.\end{cases}$$

Fifth claim. Let $f_{k+1}\equiv 1$. For $1\le i\le k$, let

$$f_i(x)=\int_0^1 a^i_s\sigma\left(\frac{1}{d+1}\sum_{j=1}^{d+1} b^i_{s,j}x_j\right)\mathrm{d}s,$$

$$g(y)=\int_0^1 c_t\sigma\left(\frac{1}{k+1}\sum_{l=1}^{k+1} d_{t,l}y_l\right)\mathrm{d}t.$$

Then

$$\begin{aligned}(g\circ f)(x)&=\int_0^1 c_t\sigma\left(\frac{1}{k+1}\sum_{i=1}^{k+1} d_{t,i}\int_0^1 a^i_s\sigma\left(\frac{1}{d+1}\sum_{j=1}^{d+1} b^i_{s,j}x_j\right)\mathrm{d}s\right)\mathrm{d}t\\ &=\int_0^1 c_t\sigma\left(\int_0^1 \bar{a}_{ts}\sigma\left(\frac{1}{d+1}\sum_{j=1}^{d+1}\bar{b}_{s,j}x_j\right)\mathrm{d}s\right)\mathrm{d}t,\end{aligned}$$

where

$$\bar{a}_{ts}=d_{t,i}a^i_{(k+1)(s-\frac{i-1}{k+1})}\quad\text{for }\frac{i-1}{k+1}\le s\le\frac{i}{k+1},\qquad \bar{b}_{s,j}=b^i_{(k+1)(s-\frac{i-1}{k+1}),j}\quad\text{for }\frac{i-1}{k+1}.$$

For the special cases observe that

$$g(z)=\sigma(z),\qquad g(z_1,z_2)=\max\{z_1,z_2\}=z_1+\sigma(z_2-z_1)$$

are Barron functions, thus the first two claims are immediate. Furthermore

$$\tilde{g}(z)=\max\{0,z\}^2=\int_{\mathbb{R}} 1_{0,\infty}2\sigma(z-\xi)\mathrm{d}\xi$$

is a Barron function on bounded intervals, and so is $z\mapsto z^2$. The Barron functions f_1, f_2 are continuous on a compact set K and hence bounded. It follows that

$$f_1f_2=\frac{1}{4}\left[(f_1+f_2)^2-(f_1-f_2)^2\right]\in X_{(0,1),(0,1),\{0,\cdots,d\};K}.$$

This completes the proof. □

In particular, $X_{(0,1),(0,1),\{0,\cdots,d\};K}$ contains many functions which are not in Barron space (compare [23, Remark 5.12]).

Remark 4.1. The unit interval with Lebesgue measure is a probability space with two convenient properties for our purposes:

1. For any finite collection of numbers $0 \leq \alpha_1, \cdots, \alpha_N \leq 1$ such that $\sum_{i=1}^N \alpha_i = 1$, there exist disjoint measurable subsets $I_i \subseteq (0,1)$ such that $\mathcal{L}(I_i) = \alpha_i$ for all $1 \leq i \leq N$. This allows us to embed finite networks of arbitrary width (and can be extended to countable sums).

2. There exist measurable bijections between the unit interval and many index sets which appear larger at first sight. By rearranging decimal representations, we may for example construct a measurable bijection between $(0,1)$ and $(0,1)^d$ for any $d \geq 1$. Using a hyperbolic tangent or similar for rescaling, we can further show that a measurable bijection between $(0,1)$ and $\mathbb{R}^d$ exists. Furthermore, using the characteristic function of a probability measure π on $(0,1)$, we can find a measurable map $\phi : (0,1) \to (0,1)$ such that $\phi_\sharp \pi$ is Lebesgue measure. For details, see e.g. [23, Section 2.8].

The entire analysis remains valid for any index set with these two properties. We describe a more natural (but also more complicated) approach in Section 4.4.

Remark 4.2. Let $(\Omega_\ell, \mathcal{A}_\ell, \pi^\ell)$, $(\widetilde{\Omega}_\ell, \widetilde{\mathcal{A}}_\ell, \widetilde{\pi}^\ell)$ be families of probability spaces for $0 \leq \ell \leq \Omega_L$ and $\phi^\ell : \Omega_\ell \to \widetilde{\Omega}_\ell$ measurable bijections such that $\widetilde{\pi}^\ell = \phi^\ell_\sharp \pi^\ell$. Then the spaces

$$X_{\Omega_L, \cdots, \Omega_0; K} = X_{\widetilde{\Omega}_L, \cdots, \widetilde{\Omega}_0; K}$$

coincide and the norms induced by network representations with the different index spaces agree.

Remark 4.3. We never used that the measures π^i are probability measures (or even finite). More general measures could be used on the index set. In particular, the analysis of this section also applies to the space of countably wide neural networks (which corresponds to the integers with the counting measure).

There currently seems little gain in pursuing that generality, and we will remain in the natural mean field setting of networks indexed by probability spaces.

4.2 Networks with Hilbert weights

We can bound the path-norm by a more convenient expression. At first glance, it looks as though the weight functions a^i are required to satisfy a restrictive integrability condition like $a^i \in L^{L+1}(\pi^L \otimes \cdots \otimes \pi^0)$. This can be weakened significantly by using the neural-network structure in which indices for layers separated by one intermediate layer are independent.

Lemma 4.2. *For any f, the path-norm is bounded by*

$$\|f\|_{\Omega_L,\cdots,\Omega_0;K} \leq \inf\left\{ \|a^L\|_{L^2(\pi^L)} \prod_{i=0}^{L-1} \|a^i\|_{L^2(\pi^{i+1}\otimes\pi^i)} \,\middle|\, a^i \text{ s.t. } f = f_{a^L,\cdots,a^0} \text{ on } K \right\}. \tag{4.6}$$

Proof. To simplify notation, we denote $\pi^i(\mathrm{d}\theta_i) = \mathrm{d}\theta_i$ as in the case of the unit interval. The proof goes through in the general case. We quickly observe that for a network with two hidden layers, we can easily bound

$$\begin{aligned}
&\int_{\Omega_2\times\Omega_1\times\Omega_0} \left|a^2_{\theta_2} a^1_{\theta_2\theta_1} a^0_{\theta_1\theta_0}\right| \mathrm{d}\theta_2\mathrm{d}\theta_1\mathrm{d}\theta_0 \\
&= \int_{\Omega_2\times\Omega_1\times\Omega_0} \left|a^2_{\theta_2} a^0_{\theta_1\theta_0}\right| \left|a^1_{\theta_2\theta_1}\right| \mathrm{d}\theta_2\mathrm{d}\theta_1\mathrm{d}\theta_0 \\
&\leq \left(\int_{\Omega_2\times\Omega_1\times\Omega_0} \left|a^2_{\theta_2} a^0_{\theta_1\theta_0}\right|^2 \mathrm{d}\theta_2\mathrm{d}\theta_1\mathrm{d}\theta_0\right)^{\frac{1}{2}} \left(\int_{\Omega_2\times\Omega_1\times\Omega_0} \left|a^1_{\theta_2\theta_1}\right|^2 \mathrm{d}\theta_2\mathrm{d}\theta_1\mathrm{d}\theta_0\right)^{\frac{1}{2}} \\
&= \left(\int_{\Omega_2} \left|a^2_{\theta_2}\right|^2 \mathrm{d}\theta_2\right)^{\frac{1}{2}} \left(\int_{\Omega_2\times\Omega_1} \left|a^1_{\theta_2\theta_1}\right|^2 \mathrm{d}\theta_2\mathrm{d}\theta_1\right)^{\frac{1}{2}} \left(\int_{\Omega_1\times\Omega_0} \left|a^0_{\theta_1\theta_0}\right|^2 \mathrm{d}\theta_1\mathrm{d}\theta_0\right)^{\frac{1}{2}}.
\end{aligned}$$

In the general case, we set $\Omega_{L+1} = \{0\}$ to simplify notation. the argument follows as above by

$$\begin{aligned}
&\|f_{a^L,\cdots,a^0}\|_{\Omega_L,\cdots,\Omega_0;K} \\
&\leq \int_{\prod_{i=0}^{L+1}\Omega_i} \left|a^{(L)}_{\theta_{L+1}\theta_L} \cdots a^{(0)}_{\theta_1\theta_0}\right| \mathrm{d}\theta_L\cdots\mathrm{d}\theta_0 \\
&= \int_{\prod_{i=0}^{L+1}\Omega_i} \left|\prod_{i=0}^{\lfloor L/2\rfloor} a^{2i}_{\theta_{2i+1}\theta_{2i}}\right| \left|\prod_{i=0}^{\lfloor (L-1)/2\rfloor} a^{2i+1}_{\theta_{2i+2}\theta_{2i+1}}\right| \mathrm{d}\theta_L\cdots\mathrm{d}\theta_0 \\
&\leq \left(\int_{\prod_{i=0}^{L+1}\Omega_i} \left|\prod_{i=0}^{\lfloor L/2\rfloor} a^{2i}_{\theta_{2i+1}\theta_{2i}}\right|^2 \mathrm{d}\theta_L\cdots\mathrm{d}\theta_0\right)^{\frac{1}{2}} \left(\int_{\prod_{i=0}^{L+1}\Omega_i} \left|\prod_{i=0}^{\lfloor L-1/2\rfloor} a^{2i+1}_{\theta_{2i+2}\theta_{2i+1}}\right|^2 |\mathrm{d}\theta_L\cdots\mathrm{d}\theta_0\right)^{\frac{1}{2}} \\
&= \|a^L\|_{L^2(\pi^L)} \prod_{i=0}^{L-1} \|a^i\|_{L^2(\pi^{i+1}\times\pi^i)}.
\end{aligned}$$

We may now take the infimum over all coefficient functions. □

The lemma allows us to analyze networks in a convenient fashion using only L^2-norms. In numerical simulations, explicit regularization by penalizing L^2-norms provides a smoother alternative to penalizing the path-norm directly. Note that the proof is built on the network index structure and does not extend to neural trees.

Lemma 4.3. *The realization map*

$$F: L^2(\pi^L) \times L^2(\pi^L \otimes \pi^{L-1}) \cdots \times L^2(\pi^1 \otimes \pi^0) \to C^0(K), \qquad F(a_L, \cdots, a_0) = f_{a^L, \cdots, a^0} \tag{4.7}$$

is locally Lipschitz-continuous.

Proof. For $L=1$ and $x \in K$, note that

$$\begin{aligned}
&\left| f_{a^1,a^0}(x) - f_{\bar{a}^1,\bar{a}^0}(x) \right| \\
&= \left| \int_{\Omega_1} a^1_{\theta_1} \sigma\left(\frac{1}{d+1} \sum_{\theta_0=1}^{d+1} a^0_{\theta_1\theta_0} x_{\theta_0} \right) - \bar{a}^1_{\theta_1} \sigma\left(\frac{1}{d+1} \sum_{\theta_0=1}^{d+1} \bar{a}^0_{\theta_1\theta_0} x_{\theta_0} \right) \pi^1(\mathrm{d}\theta_1) \right| \\
&\leq \int_{\Omega_1} \left| a^1_{\theta_1} - \bar{a}^1_{\theta_1} \right| \left| \sigma\left(\frac{1}{d+1} \sum_{\theta_0=1}^{d+1} a^0_{\theta_1\theta_0} x_{\theta_0} \right) \right| \\
&\qquad + \left| \bar{a}^1_{\theta_1} \right| \left| \sigma\left(\frac{1}{d+1} \sum_{\theta_0=1}^{d+1} a^0_{\theta_1\theta_0} x_{\theta_0} \right) - \sigma\left(\frac{1}{d+1} \sum_{\theta_0=1}^{d+1} \bar{a}^0_{\theta_1\theta_0} x_{\theta_0} \right) \right| \pi^1(\mathrm{d}\theta_1) \\
&\leq \|a^1 - \bar{a}^1\|_{L^2(\Omega_1)} \|a^0\|_{L^2(\Omega_1 \times \Omega_0)} \sup_{x \in K} |x| + \|\bar{a}^1\|_{L^2(\Omega_1)} \|a^0 - \bar{a}^0\|_{L^2(\Omega_1 \times \Omega_0)} \sup_{x \in K} |x|.
\end{aligned}$$

The general case follows analogously by induction. □

We define a third class of spaces for the L^2-approach.

Definition 4.2. *For $0 \leq i \leq L$, let $(\Omega_i, \mathcal{A}_i, \pi^i)$ be a probability space where $\Omega_0 = \{0, \cdots, d\}$ and π^0 is the normalized counting measure. Let $a^L \in L^2(\pi^L)$ and $a^i \in L^2(\pi^{i+1} \otimes \pi^i)$ for $0 \leq i \leq L-1$. Then define like in* (4.1)

$$\begin{aligned}
&f_{a^L, \cdots, a^0}(x) \\
&= \int_{\Omega_L} a^{(L)}_{\theta_L} \sigma\left(\int_{\Omega_{L-1}} \cdots \sigma\left(\int_{\Omega_1} a^1_{\theta_2,\theta_1} \sigma\left(\int_{\Omega_0} a^0_{\theta_1,\theta_0} x_{\theta_0} \pi^0(\mathrm{d}\theta_0) \right) \pi^1(\mathrm{d}\theta_1) \right) \cdots \pi^{(L-1)}(\mathrm{d}\theta_{L-1}) \right) \pi^L(\mathrm{d}\theta_L).
\end{aligned}$$

We define the class of neural networks over K with Hilbert weights over the index spaces $\Omega_i = (\Omega_i, \mathcal{A}_i, \pi^i)$ *as the image of $L^2(\pi^L) \times \cdots \times L^2(\pi^1 \otimes \pi^0)$ under the realization map* (4.7) *and denote it by*

$$\mathcal{W}_{\pi^L, \cdots, \pi^0}(K) = \left\{ f: K \to \mathbb{R} \,\middle|\, \exists\, a^L \in L^2(\pi^L), a^\ell \in L^2(\pi^\ell \otimes \pi^\ell) \text{ s.t. } f \equiv f_{a^L, \cdots, a^0} \text{ on } K \right\}.$$

The function class is equipped with the measure of complexity

$$Q_{\pi^L, \cdots, \pi^0; K}(f) = \inf \left\{ \|a^L\|_{L^2(\pi^L)} \prod_{i=0}^{L-1} \|a^i\|_{L^2(\pi^{i+1} \otimes \pi^i)} \,\middle|\, a^i \text{ s.t. } f = f_{a^L, \cdots, a^0} \text{ on } K \right\}. \tag{4.8}$$

We declare a notion of convergence on $\mathcal{W}_{\pi^L,\cdots,\pi^0}(K)$ by the convergence of the weight functions in the L^2-strong topology. To avoid pathological cases, we normalize the weights across layers. Using the homogeneity of σ, note that $f_{a^L,\cdots,a^0} = f_{\lambda_\ell a^\ell,\cdots,\lambda_0 a^0} \in \mathcal{W}_{\pi^L,\cdots,\pi^0}(K)$ for $\lambda_i > 0$ such that $\prod_{i=0}^L \lambda_i = 1$. In particular, we may assume without loss of generality that

$$\|a^\ell\|_{L^2} = \left(\prod_{i=0}^L \|a^i\|_{L^2}\right)^{\frac{1}{L+1}}$$

for all $\ell \geq 1$.

Definition 4.3. *We say that a sequence of functions $f_n \in \mathcal{W}_{\pi^L,\cdots,\pi^0}(K)$ converges weakly to a limit $f \in \mathcal{W}_{\pi^L,\cdots,\pi^0}(K)$ if there exist coefficient functions $a^{L,n},\cdots,a^{0,n}$ for $n \in \mathbb{N}$ and $a^L,\cdots,a^0$ such that*

1. *$f_n = f_{a^{L,n},\cdots,a^{0,n}}$ for all $n \in \mathbb{N}$ and $f = f_{a^L,\cdots,a^0}$.*
2. *$\|a^{\ell,n}\| = \left(\prod_{i=0}^L \|a^{i,n}\|_{L^2}\right)^{\frac{1}{L+1}}$ for all $n \in \mathbb{N}$ and $0 \leq \ell \leq L$.*
3. *$\limsup_{n\to\infty} \left[\prod_{i=0}^L \|a^{i,n}\|_{L^2} - Q(f_n)\right] = 0$.*
4. *$a^{\ell,n} \to a^\ell$ in the L^2-strong topology for all $0 \leq \ell \leq n$.*

To evaluate the notion of convergence, consider the case $L = 1$ and write (a,w) for (a^1,a^0). We interpret a^0 as an $\mathbb{R}^{d+1}$-valued function on $(0,1)$ rather than a scalar function on $(0,1)\times\{0,\cdots,d\}$. Then it is easy to see that

$$(a^n,w^n) \to (a,w) \quad \text{strongly in } L^2(0,1) \quad \Rightarrow \quad (a^n,w^n)_\sharp \mathcal{L} \to (a,w)_\sharp \mathcal{L} \quad \text{in Wasserstein.}$$

The inverse statement holds up to a rearrangement of the index set. The Wasserstein distance is associated with the weak convergence of measures, while the topology of Barron space is associated with the norm topology for the total variation norm (strong convergence). This justifies the terminology of 'weak convergence' of arbitrarily wide neural networks.

Weak convergence is locally metrizable, but not induced by a norm. A relaxed version of convergence described above is metrizable by the distance function

$$d_{HW}(f,g) = \inf\left\{ \sum_{\ell=0}^L \|a^{\ell,f} - a^{\ell,g}\|_{L^2(\pi^\ell)} \,\middle|\, a^{L,f},\cdots,a^{0,g} \text{ s.t. } f = f_{a^{L,f},\cdots,a^{0,f}}, g = f_{a^{L,g},\cdots,a^{0,g}} \text{ and } \right.$$
$$\left. \|a^{\ell,h}\| \equiv \left(\prod_{i=0}^L \|a^{i,h}\|_{L^2}\right)^{\frac{1}{L+1}} \leq 2Q(h)^{\frac{1}{L+1}} \text{ for } h \in \{f,g\} \right\}. \quad (4.9)$$

The third condition has been weakened from $\prod_{i=0}^{L}\|a^{i,n}\|_{L^2} - Q(f_n) \to 0$ to $Q(f_n) \leq \prod_{i=0}^{L}\|a^{i,n}\|_{L^2} \leq 2Q(f_n)$. The normalization is required to ensure that functions in which one layer can be chosen identical do not have zero distance by shifting all weight to the one layer. Which mode of convergence is superior to another remains to be seen. Equipped with the Hilbert weight metric d_{HW}, the spaces $\mathcal{W}_{\pi^L,\cdots,\pi^0}(K)$ are complete.

To avoid the unwieldy terminology of arbitrarily wide neural networks with Hilbert weights, we introduce the following simpler terminology.

Definition 4.4. *The metric spaces* $(\mathcal{W}_{\pi^L,\cdots,\pi^0}(K), d_{HW})$ *equipped with the metric* d_{HW} *from* (4.9) *are called* multi-layer spaces *for short.*

Remark 4.4. As seen in Lemma 4.2, the inclusions

$$\mathcal{W}_{\pi^L,\cdots,\pi^0}(K) \;\subseteq\; X_{\Omega^L,\cdots,\Omega^0;K} \;\subseteq\; \mathcal{W}^L(K) \tag{4.10}$$

hold. The last three points of Lemma 4.1 hold with $\mathcal{W}_{\pi^L,\cdots,\pi^0}(K)$ in place of $X_{\Omega^L,\cdots,\Omega^0;K}$. We note however that the functions

$$c_\theta^L = \begin{cases} 2a_{2\theta}^L, & \theta \in (0,1/2), \\ 2b_{2\theta-1}^L, & \theta \in (1/2,1), \end{cases} \qquad c_{\theta\xi}^\ell = \begin{cases} 4a_{2\theta,2\xi}^\ell, & \theta,\xi \in (0,1/2), \\ 4b_{2\theta-1,2\xi-1}^\ell, & \theta,\xi \in (1/2,1), \\ 0, & \text{else}, \end{cases}$$

satisfy

$$\|c^L\|_{L^2(0,1)}^2 = 2\left[\|a^L\|_{L^2(0,1)}^2 + \|b^L\|_{L^2(0,1)}^2\right],$$
$$\|c^\ell\|_{L^2\left((0,1)^2\right)} = 4\left[\|a^\ell\|_{L^2\left((0,1)^2\right)} + \|b^\ell\|_{L^2\left((0,1)^2\right)}\right].$$

In particular, if $\Omega^\ell = (0,1)$ and π^ℓ is Lebesgue measure for all $1 \leq \ell \leq L$, then $\mathcal{W}_{\pi^L,\cdots,\pi^0}(K)$ is a linear space, but both $Q_{\pi^L,\cdots,\pi^0;K}$ and d_{HW} generally fail to be a norm.

Remark 4.5. It is not clear whether the inclusions in (4.10) are necessarily strict. In the case of Barron space, it is easily possible to normalize by replacing

$$a_{\theta_1}^1 \mapsto \frac{a_{\theta_1}^1}{\rho_{\theta_1}}, \qquad a_{\theta_1\theta_0}^0 \mapsto \rho_{\theta_1} a_{\theta_1\theta_0}^0$$

such that both layers have the same magnitude in $L^2(0,1)$, even if they are only assumed to be measurable with finite path-norm a priori. For multiple layers, this may not be possible. Let

$$a_s \equiv 1, \qquad b_{st} = f(s-t), \qquad c_t \equiv 1, \tag{4.11}$$

where f is a one-periodic function on $\mathbb{R}$ which is in $L^1(0,1)$, but not $L^2(0,1)$. Then any normalization

$$a_s \mapsto \frac{a_s}{\rho_s}, \qquad b_{st} \mapsto \rho_s \tilde{\rho}_t b_{st}, \qquad c_t \mapsto \frac{c_t}{\tilde{\rho}_t}$$

fails to make b L^2-integrable. Whether or not this can be compensated by choosing other weights with the same realization remains an open question.

4.3 Networks with two hidden layers

We investigate the space $X_{(0,1),(0,1),\{0,\cdots,d\};K}$ and $\mathcal{W}_{\mathcal{L}^1,\mathcal{L}^1,\pi^0}(K)$ more closely where π^0 denotes counting measure and $\mathcal{L}^1$ is the Lebesgue measure on $(0,1)$. In general, any network modelled on probability spaces $\Omega_2,\Omega_1,\Omega_0$ can be written as

$$\begin{aligned} f(x) &= \int_{\Omega_2} a^2_{\theta_2}\sigma\left(\int_{\Omega_1} a^1_{\theta_2,\theta_1}\sigma\left(\sum_{\theta_0=1}^{d+1} a^0_{\theta_1,\theta_0}x_{\theta_0}\right)\pi^1(\mathrm{d}\theta_1)\right)\pi^2(\mathrm{d}\theta_2)\\ &= \int_{\Omega_2} a^2_{\theta_2}\rho_{\theta_2}\sigma\left(\int_{\Omega_1}\frac{a^1_{\theta_2,\theta_1}|w_{\theta_1}|}{\rho_{\theta_2}}\sigma\left(\frac{w^T_{\theta_1}}{|w_{\theta_1}|}(x,1)\right)\pi^1(\mathrm{d}\theta_1)\right)\pi^2(\mathrm{d}\theta_2)\\ &= \int_{\Omega_2} a^2_{\theta_2}\rho_{\theta_2}\sigma\left(\int_{\mathbb{R}\times S^d}\tilde{a}\sigma(\tilde{w}^T x)\,(\Psi(\theta_2,\cdot)_\sharp\pi^1)(\mathrm{d}\tilde{a}\otimes\mathrm{d}\tilde{w})\right)\pi^2(\mathrm{d}\theta_2), \end{aligned}$$

where $w_\theta = (a^0_{\theta_1,1},\cdots,a^0_{\theta_1,d+1})$ and

$$\Psi:\Omega_2\times\Omega_1\to\mathbb{R}\times S^d,\qquad \Psi(\theta_2,\theta_1)=\left(\frac{a^1_{\theta_2\theta_1}|w_{\theta_1}|}{\rho_{\theta_2}},\frac{w_{\theta_1}}{|w_{\theta_1}|}\right).$$

Since the second component of Ψ does not depend on θ_2, the marginal $\overline{\pi}$ of $\Psi(\theta_2,\cdot)_\sharp\pi^1$ on the sphere is independent of θ_2. We can therefore write

$$\int_{\mathbb{R}\times S^d}\tilde{a}\sigma(\tilde{w}^T x)\,(\Psi(\theta_2,\cdot)_\sharp\pi^1)(\mathrm{d}\tilde{a}\otimes\mathrm{d}\tilde{w}) = \int_{S^d}\bar{a}^{\theta_2}(w)\sigma(w^T x)\,\bar{\pi}(\mathrm{d}w)$$

by integrating in the a-direction and making $\bar{a}$ a function of w (see [23, Section 2.3] for the technical details). Thus

$$f(x)=\int_{\Omega_2} a^2_{\theta_2}\rho_{\theta_2}\sigma\left(\int_{S^d}\bar{a}^{\theta_2}(w)\sigma(w^T x)\,\bar{\pi}(\mathrm{d}w)\right)\pi^2(\mathrm{d}\theta_2).$$

We can in particular choose $\rho\geq 0$ such that

$$\int_{S^d}\left|\bar{a}^{\theta_2}(w)\right|\bar{\pi}(\mathrm{d}w)\leq\int_{\Omega^1}\frac{|a^1_{\theta_2\theta_1}||w_{\theta_1}|}{\rho_{\theta_2}}\pi^1(\mathrm{d}\theta_1)=\frac{1}{\rho_{\theta_2}}\int_{\Omega^1}|a^1_{\theta_2\theta_1}||w_{\theta_1}|\,\pi^1(\mathrm{d}\theta_1)\leq 1$$

for all $\theta_2\in\Omega_2$. Then the map

$$F:\Omega_2\to B^X,\qquad \theta_2\mapsto f^{\theta_2}=\int_{S^d}\bar{a}^{\theta_2}(w)\sigma(w^T x)\,\bar{\pi}(\mathrm{d}w)$$

is well-defined and Bochner integrable. In particular

$$\begin{aligned} f(x) &= \int_{\Omega_2} a^{(2)}_{\theta_2}\rho_{\theta_2}\sigma\big(f^{\theta_2}(x)\big)\,\pi^2(\mathrm{d}\theta_2)\\ &= \int_{B^X}\sigma(g(x))\,\mu(\mathrm{d}g), \end{aligned}$$

where $\mu = F_\sharp\left((a^{(2)}\rho)\cdot\pi^2\right)$. By construction, μ is concentrated on the subspace $Y_{\bar\pi}$ of Barron functions which can be represented with an L^1-density with respect to the measure $\bar\pi$, by which we mean that $|\mu|(B^X \setminus Y_{\bar\pi}) = 0$. This equation can be sensibly interpreted since any measure can be extended to a potentially larger σ-algebra containing all null sets.

Thus general functions in $\mathcal{W}^2(K)$ and $X_{(0,1),(0,1),\{0,\cdots,d\};K}$ both take the form

$$f(x) = \int_{B^X} \sigma(g(x))\,\mu(\mathrm{d}g),$$

where B^X is the unit ball in Barron space, but in the second case, μ is concentrated on a subspace $Y_{\bar\pi}$. This space is a quotient of $L^1(\bar\pi)$ by a closed subspace and thus closed in Barron space, but may be dense in $C^0(K)$. If $\bar\pi$ is the uniform distribution on S^d, then $Y_{\bar\pi}$ is dense in C^0 since $L^1(\bar\pi)$ is dense in the space of Radon measures on S^d with respect to the weak topology.

Claim: There is no distribution $\bar\pi$ on S^d such that every Barron function can be expressed with an L^1-density with respect to $\bar\pi$ if K is the closure of an open set.

Proof of claim: Barron space is not separable since

$$\|\sigma(w_1^T\cdot) - \sigma(w_2^T\cdot)\|_{\mathcal{B}^1(K)} \geq [\sigma(w_1^T\cdot) - \sigma(w_2^T\cdot)]_{C^{0,1}(K)} \geq 1$$

if one of the hyperplanes $\{x : w_{1/2}^T x = 0\}$ intersects the interior of K. This is the case for uncountably many $w \in S^d$. On the other hand, $L^1(\bar\pi)$ (and also its quotient by the kernel of the realization map) is separable for any Radon measure. Thus the two spaces cannot coincide. □

The claim can be phrased and proved in greater generality if K is a manifold or similar. We note that for fixed $\bar\pi$, the space $Y_{\bar\pi}$ embeds continuously into $C^{0,1}(K)$, but its unit ball is not closed in $C^0(K)$. Nevertheless, we may consider the space

$$\mathcal{B}_{Y_{\bar\pi},K} = \left\{ f_\mu(x) = \int_{B^X \cap Y_{\bar\pi,K}} \sigma(g(x))\,\mu(\mathrm{d}g) \,\middle|\, \mu \text{ admissible} \right\},$$

where admissible measures are finite (signed) Radon measures for which $Y_{\bar\pi}$ is measurable. Every distribution $\bar\pi$ on S^d can be obtained as the push-forward of Lebesgue measure on the unit interval along a measurable map $\phi : (0,1) \to S^d$, see e.g. [23, Section 2.8]. Thus the associated space of neural networks with two hidden layers is

$$X_{(0,1),(0,1),\{0,\cdots,d\};K} = \bigcup_{\bar\pi} \mathcal{B}_{Y_{\bar\pi},K} =: \widetilde{\mathcal{W}}^2(K),$$

where the union is over all probability distributions $\bar\pi$ on S^d. Thus the first layer of $f \in \mathcal{W}^2$ is wide enough to contain the entire unit ball of $\mathcal{W}^1$, while the first layer of $f \in \widetilde{\mathcal{W}}^2$ can only express a separable subset of the unit ball in $\mathcal{W}^1$. The question whether this reduces expressivity or whether in fact $\mathcal{W}^2 = \widetilde{\mathcal{W}}^2$ remains open.

Finally, consider the space $\mathcal{W}_{\mathcal{L}^1,\mathcal{L}^1,\pi^0}(K)$ where the weights of a function satisfy

$$a^2 \in L^2(0,1), \quad a^1 \in L^2\big((0,1)^2\big), \quad a^0 \in L^2\big((0,1)\times\{0,\cdots,d\}\big) = L^2\big((0,1);\mathbb{R}^d\big).$$

We proceed as before, but normalize with respect to L^2 rather than L^1/L^∞. Again, we can consider the maps

$$\Psi : (0,1)\times(0,1) \to \mathbb{R}^{d+2}, \qquad (\theta_2,\theta_1) \mapsto \big(a^1_{\theta_2\theta_1}, a^0_{\theta_1}\big)$$

and note as before that

$$\int_0^1 a^1_{\theta_2,\theta_1} \sigma\left(\sum_{\theta_0=1}^{d+1} a^0_{\theta_1,\theta_0} x_{\theta_0}\right) \mathrm{d}\theta_1 = \int_{\mathbb{R}^{d+1}} \bar{a}^{\theta_2}(w)\sigma(w^T x)\,\bar{\pi}(\mathrm{d}w),$$

where this time $\bar{a} \in L^2(\bar{\pi})$ for almost all $\theta_2 \in (0,1)$. Thus the first layer of $f \in \mathcal{W}_{\mathcal{L}^1,\mathcal{L}^1,\pi^0}$ takes values in a single reproducing kernel Hilbert space $\mathcal{H}_{\bar{\pi}}$ associated to the kernel

$$k_{\bar{\pi}}(x,x') = \int_{\mathbb{R}^{d+1}} \sigma(w^T x)\sigma(w^T x')\,\bar{\pi}(\mathrm{d}w)$$

while the first layer of $f \in \mathcal{W}^2$ may be wide enough to contain every function in the unit ball of Barron space. Again, the relationship between the function spaces remains open.

4.4 Natural index sets

In this section, we focus on the natural index set for $\mathcal{W}_{\pi^L,\cdots,\pi^0}(K)$. Above, we allowed the index spaces Ω_i to be generic or focused on the case $\Omega_i = (0,1)$. While $(0,1)$ is simple and mathematically convenient, it is not a natural choice. First consider the simpler case of neural networks with a single hidden layer. The classical representation in this case is

$$f(x) = \int_{\mathbb{R}\times\mathbb{R}^{d+1}} a\sigma(w^T x)\,\pi(\mathrm{d}a\otimes\mathrm{d}w)$$

for some distribution π on $\mathbb{R}^{d+2}$, see [23] and the sources cited therein. Using the scaling invariance $\sigma(\cdot) = \lambda^{-1}\sigma(\lambda\cdot)$ if necessary, we may assume that

$$\int_{\mathbb{R}^{d+2}} |a|^2 + |w|^2\,\pi(\mathrm{d}a\otimes\mathrm{d}w) < \infty.$$

Then we set $\Omega_1 = \mathbb{R}^{d+2}, \Omega_0 = \{0,\cdots,d\}$ and

$$a^1_{\theta_1} = (\theta_1)_1, \qquad a^0_{\theta_1,\theta_0} = (\theta_1)_{1+\theta_0},$$

i.e. we index $\mathbb{R}^{d+2}$ by itself. In this equation, $(\theta_1)_i$ denotes the i-the component of the vector $\theta_1 \in \mathbb{R}^{d+2}$.

For networks with more than one hidden layer, the output of the first layer is vector-valued. The preceding analysis determined that the first hidden layer takes values in the reproducing kernel Hilbert space $\mathcal{H}_{\bar{\pi}}$. It thus seems reasonable at first glance to choose $\mathcal{H}_{\bar{\pi}}$ as an index space for the second hidden layer. This intuition is flawed since the output of the first hidden layer is an RKHS function of x, a variable which is fixed when calculating the output of the network and inaccessible to the second hidden layer. The previous observation has no bearing on the inner workings of neural networks, but only on the approximation power of functions described by a given neural network architecture.

Pursuing a different route, we note that π is a Radon measure on $\mathbb{R}\times\mathbb{R}^{d+1}$ where $\mathbb{R}$ is the output and $\mathbb{R}^{d+1}$ the input layer (interpreting x as $(x,1)$). For networks with two hidden layers, we note that

$$\begin{aligned}
&\left\| \int_{\Omega_1} a^1_{\theta_2\theta_1} \sigma\left(\int_{\Omega_0} a^0_{\theta_1\theta_0} x_{\theta_0} \pi^0(\mathrm{d}\theta_0) \right) \pi^1(\mathrm{d}\theta_1) \right\|^2_{L^2(\pi_2)} \\
&\leq \int_{\Omega_2} \left(\int_{\Omega_1} \left| a^1_{\theta_2\theta_1} \right|^2 \pi^1(\mathrm{d}\theta_1) \right)^{\frac{2}{2}} \left(\int_{\Omega_1}\int_{\Omega_0} \left| a^0_{\theta_1\theta_0} \right|^2 \pi^0(\mathrm{d}\theta_0)\,\pi^1(\mathrm{d}\theta_1) \right)^{\frac{2}{2}} \pi^2(\mathrm{d}\theta_2) \sup_{x\in K} |x|^2 \\
&= \|a^1\|_{L^2(\pi^2\otimes\pi^1)} \|a^0\|_{L^2(\pi^1\otimes\pi^0)} \sup_{x\in K} |x|^2
\end{aligned}$$

for all $x\in K$. We can thus view a neural network with two hidden layers and parameter functions a^2,a^1,a^0 as a composition of linear and non-linear maps in the following way:

1. Let π^1 be the distribution of vectors $w := (a^0_{\theta_1\theta_0})_{\theta_0=1}^{d+1}$ on $\mathbb{R}^{d+1}$ and $A^1:\mathbb{R}^d \to L^2(\pi^1)$ is the affine map described by

$$(A^1 x)_{\theta_1} = \int_{\Omega_0} a^0_{\theta_1\theta_0} x_{\theta_0} = \frac{1}{d+1} \sum_{\theta_0=1}^{d+1} a^0_{\theta_1\theta_0} x_{\theta_0}.$$

   We may use $\mathbb{R}^{d+1}$ as its own index set, i.e. $a^0_{\theta_1:} = \theta_1$. To emphasize the fact that index set and distribution are natural, we denote $w = \frac{1}{d+1}\theta_1$, $\bar{\pi} = \pi^1$.

2. The non-linearity σ acts on $L^2(\pi^1)$ by pointwise application.

3. Let $(\Omega_2,\mathcal{A}_2,\pi^2)$ be a general probability space used as an index set. The linear map $A^2: L^2(\pi^1) \to L^2(\pi^2)$ is given by

$$(A^2 f)_{\theta_2} = \int_{\Omega_1} a^1_{\theta_2\theta_1} f_{\theta_1}\, \pi^1(\mathrm{d}\theta_1) = \langle a^1_{\theta_2:}, f\rangle_{L^2(\pi^1)},$$

   where $a^1_{\theta_2:}(\theta_1) = a^1_{\theta_2\theta_1}$.

4. The non-linearity σ acts on $L^2(\pi^2)$ by pointwise application.

5. The map $A^3 : L^2(\pi^2) \to \mathbb{R}$ is given by

$$A^3 f = \int_{\Omega_2} a^2_{\theta_2} f_{\theta_2} \, \pi^2(\mathrm{d}\theta_2).$$

Then

$$\begin{aligned} f(x) &= \left(A^3 \circ \sigma \circ A^2 \circ \sigma \circ A^1\right)(x) \\ &= \int_{\Omega_2} a^2_{\theta_2} \sigma \left(\left\langle a^1_{\theta_2:}, \sigma\left(\frac{1}{d+1} \langle a^0_{\theta_1:}, x \rangle_{\mathbb{R}^{d+1}} \right) \right\rangle_{L^2(\pi^1)} \right) \pi(\mathrm{d}\theta_2) \\ &= \int_{\mathbb{R} \times L^2(\bar{\pi})} \tilde{a} \sigma \left(\langle \tilde{h}, \sigma(w^T x) \rangle_{L^2(\bar{\pi})} \right) (H_\sharp \pi^2)(\mathrm{d}\tilde{a} \otimes \mathrm{d}\tilde{h}), \end{aligned}$$

where

$$H : \Omega_2 \to \mathbb{R} \times L^2(\bar{\pi}), \qquad \theta_2 \mapsto (a^2_{\theta_2}, a^1_{\theta_2:}).$$

Thus we may in a natural way interpret

- $\Omega_0 = \{0, \cdots, d\}$ with the normalized counting measure.
- $\Omega_1 = \mathbb{R}^{d+1} = L^2(\Omega_0)$. $\bar{\pi} = \pi^1$ can be any probability distribution on Ω_1 with finite second moments.
- $\Omega_2 = \mathbb{R} \times L^2(\bar{\pi})$ and π^2 is a probability distribution with finite second moments.

More generally, we set

- $\Omega_0 = \{0, \cdots, d\}$ with the normalized counting measure $\bar{\pi}^0$.
- $\Omega_\ell = L^2(\bar{\pi}^{\ell-1})$ and a measure $\bar{\pi}^\ell$ with finite second moments on Ω_ℓ for $1 \leq \ell \leq L-1$.
- $\Omega_L = \mathbb{R} \times L^2(\bar{\pi}^{L-1})$ and a measure $\bar{\pi}^L$ with finite second moments on Ω_L.

The outermost index space Ω_L has the additional factor $\mathbb{R}$ compared to Ω_ℓ because both the first and the last operations in a neural network are linear. Note that Ω_ℓ is a Polish space for every ℓ by induction.

All considerations above were for fixed x. As x varies, a neural network with L hidden layers takes the form $f(x) = (z^L \circ \cdots \circ z^1)(x)$ where

1. $z^1 \in C^{0,1}(\mathrm{spt}\mathbb{P}, \Omega_1)$, $z^0(w,x) = w^T x = \mathbb{E}_{w_i \sim \pi_0} w_i x_i$ where we interpret $w \in \Omega_1 = L^2(\pi_0)$.
2. $x^\ell \in C^{0,1}(\mathrm{spt}\mathbb{P}, \Omega_{\ell+1})$ is defined by $z^\ell(y,f) = \langle f, \sigma(y) \rangle_{\pi^{\ell-1}}$ where $y \in \pi^{\ell-1}$ is the output of the previous layer and $f \in \pi^{\ell-1}$ is the natural index of z^ℓ. Thus $z^\ell(\cdot, y) \in L^2(\pi^{\ell-1}) = \Omega_\ell$.
3. $z^L(y) = \int_{\Omega_L} \tilde{a} \sigma(\langle f, y \rangle_{\pi^{L-1}}) \, \pi^L(\mathrm{d}\tilde{a} \otimes \mathrm{d}f)$.

All natural index spaces above are separable Hilbert spaces and therefore isomorphic to each other (for all ℓ for which Ω_ℓ if infinite-dimensional) and to both $L^2(0,1)$ and ℓ^2. However, the application of the non-linearity σ in L^2 and ℓ^2 is not invariant under Hilbert-space isomorphisms. It makes a big difference whether we take the positive part of a function $f \in L^2(0,1)$ set all negative Fourier-coefficients of a function to zero. Luckily, natural isomorphisms preserve the structure of continuous neural network models as in Remark 4.2.

5 Optimization of the continuous network model

We now study gradient flows for the risk functionals in the continuous setting. We will restrict ourselves to the indexed representation with L^2-weights. The most natural optimization algorithm for weight-functions $a^\ell \in L^2((0,1)^2)$ is the L^2-gradient flow. We show that the usual gradient descent dynamics of neural network training can be recovered as discretizations of the continuous optimization algorithm. In this sense, we follow the philosophy of designing optimization algorithms for continuous models and discretizing them later which was put forth in [19]. We present our findings in the simplest possible setting.

5.1 Discretizations of the continuous gradient flow

We now show that a natural discretization of the continuous gradient flow recovers the gradient descent dynamics for the usual multi-layer neural networks with the "mean-field" scaling. This is a general feature of Vlasov type dynamics.

The following computations are purely formal, assuming that solutions to all ODEs proposed below exist – the issue of existence and uniqueness of solutions is briefly discussed in Appendix B. The arguments however are based on an identity and energy dissipation property which are expected to be stable when considering generalized solutions. For smooth activation functions σ, all computations can be made rigorous and solutions exist.

Lemma 5.1. *Consider a discretized version of the continuous indexed representation:*

$$f(x) = \frac{1}{m_L} \sum_{i_L=1}^{m_L} a_{i_L}^L \sigma\left(\frac{1}{m_{L-1}} \sum_{i_{L-1}=1}^{m_{L-1}} a_{i_L i_{L-1}}^{L-1} \sigma\left(\cdots \sigma\left(\frac{1}{m_1} \sum_{i_1=1}^{m_1} a_{i_2 i_1}^1 \sigma\left(\frac{1}{d+1} \sum_{i_0=1}^{d+1} a_{i_1 i_0}^0 x_{i_0} \right) \right) \right) \right).$$

Define functions

$$a^L : (0,1) \to \mathbb{R}, \quad a^L(s) = a_i^L \quad \text{for } \frac{i-1}{m_L} \leq s < \frac{i}{m_L},$$

$$a^\ell : (0,1)^2 \to \mathbb{R}, \quad a^\ell(r,s) = a_{ij}^\ell \quad \text{for } \frac{i-1}{m_{\ell+1}} \leq r < \frac{i}{m_L}, \frac{j-1}{m_\ell} \leq s < \frac{i}{m_\ell},$$

for $0 \leq \ell < L$. Then $f = f_{a^L,\cdots,a^0}$ and the coefficient functions $a^L,\cdots,a^0$ evolve by the L^2-gradient flow of

$$\mathcal{R}(a^L,\cdots,a^0) = \int_{\mathbb{R}^d} \ell\big(f_{a^L,\cdots,a^0}(x),y\big)\,\mathbb{P}(\mathrm{d}x\otimes\mathrm{d}y)$$

if and only if the parameters a_i^L, a_{ij}^ℓ evolve by the time-rescaled gradient flows

$$\begin{aligned}
\dot{a}_i^L &= -m_L \partial_{a_i^L} \mathcal{R}\big(a_{i_L}^L,\cdots,a_{i_1 i_0}^0\big),\\
\dot{a}_{ij}^\ell &= -m_{\ell+1} m_\ell \partial_{a_{ij}^\ell} \mathcal{R}\big(a_{i_L}^L,\cdots,a_{i_1 i_0}^0\big), \qquad 0 \leq i \leq L-1,
\end{aligned} \tag{5.1}$$

where the risk of finitely many weights is defined accordingly.

Passing to a single index set $(0,1)$ for all layers, we lose the information about the scaling of the width and compensate by prescribing layer-wise learning rates which lead to balanced training velocities.

Proof. The proof for networks with one hidden layer can be found in [22, Lemma 2.8]. To simplify the presentation, we focus on the case of two hidden layers. The general case follows the same way. Consider the network

$$f(x) = \frac{1}{M}\sum_{i=1}^{M} a_i \sigma\left(\frac{1}{m}\sum_{j=1}^{m} b_{ij}\sigma\left(\frac{1}{d+1}\sum_{k=1}^{d+1} c_{jk}x_k\right)\right)$$

and compute the gradient

$$\begin{aligned}
&\nabla_{a_i,b_{ij},c_{jk}} \mathcal{R}(a,b,c)\\
&= \nabla_{a_i,b_{ij},c_{jk}} \int_{\mathbb{R}^d} \ell\big(f_{a,b,c}(x),y\big)\,\mathbb{P}(\mathrm{d}x\otimes\mathrm{d}y)\\
&= \int_{\mathbb{R}^d} (\partial_1\ell)\big(f_{a,b,c}(x),y\big)\,\nabla_{a_i,b_{ij},c_{jk}} f_{a,b,c,}(x)\,\mathbb{P}(\mathrm{d}x\otimes\mathrm{d}y)\\
&= \int_{\mathbb{R}^d} (\partial_1\ell)\big(f_{a,b,c}(x),y\big) \begin{pmatrix} \frac{1}{M}\sigma\left(\frac{1}{m}\sum_{j=1}^m b_{ij}\sigma\left(\frac{1}{d+1}\sum_{k=1}^{d+1} c_{jk}x_k\right)\right)\\ \frac{1}{M}a_i\sigma'\left(\frac{1}{m}\sum_l b_{il}\sigma(\cdots)\right)\frac{1}{m}\sigma\left(\frac{1}{d+1}\sum_{k=1}^{d+1} c_{ik}x_k\right)\\ \frac{1}{M}\sum_{i=1}^M a_i\sigma'(\cdots)\frac{1}{m}\sigma'\left(\frac{1}{d+1}\sum_{l=1}^{d+1} c_{jl}x_l\right)\frac{1}{d+1}x_k \end{pmatrix}\mathbb{P}(\mathrm{d}x\otimes\mathrm{d}y)\\
&= \int_{\mathbb{R}^d} (\partial_1\ell)\big(f_{a,b,c}(x),y\big) \begin{pmatrix} \frac{1}{M}\sigma\big(f_{b_{i:},c}(x)\big)\\ \frac{1}{Mm}a_i\sigma'\big(f_{b_{i:},c}(x)\big)\,\sigma\left(f_{c_{j:}}(x)\right)\\ \frac{1}{m(d+1)}\frac{1}{M}\sum_{i=1}^M a_i\sigma'\big(f_{b_{i:},c}(x)\big)\,\sigma'\left(f_{c_{j:}}(x)\right) \end{pmatrix}\mathbb{P}(\mathrm{d}x\otimes\mathrm{d}y),
\end{aligned}$$

where

$$f_{b_{i:},c}(x) = \frac{1}{m}\sum_{j=1}^{m} b_{ij}\sigma\left(\frac{1}{d+1}\sum_{k=1}^{d+1} c_{jk}x_k\right) \quad \text{and} \quad f_{c_{j:}}(x) = \frac{1}{d+1}\sum_{l=1}^{d+1} c_{jl}x_l.$$

Equally, we can compute the L^2-gradient by taking variations

$$\begin{aligned}
\delta_{a;\phi}\mathcal{R}(a,b,c) &= \lim_{h\to 0}\frac{\mathcal{R}(a+h\phi,b,c)-\mathcal{R}(a,b,c)}{h}\\
&= \int_{\mathbb{R}^d}\lim_{h\to 0}\frac{\ell\big(f_{a+h\phi,b,c}(x),y\big)-\ell\big(f_{a,b,c}(x),y\big)}{h}\mathbb{P}(\mathrm{d}x\otimes\mathrm{d}y)\\
&= \int_{\mathbb{R}^d}(\partial_1\ell)\big(f_{a,b,c}(x),y\big)\lim_{h\to 0}\frac{f_{a+h\phi,b,c}(x)-f_{a,b,c}(x)}{h}\mathbb{P}(\mathrm{d}x\otimes\mathrm{d}y)\\
&= \int_{\mathbb{R}^d}(\partial_1\ell)\big(f_{a,b,c}(x),y\big)\int_0^1\phi(s)\sigma\big(f_{b_{s:},c}(x)\big)\,\mathrm{d}s\mathbb{P}(\mathrm{d}x\otimes\mathrm{d}y)\\
&= \int_0^1\left(\int_{\mathbb{R}^d}(\partial_1\ell)\big(f_{a,b,c}(x),y\big)\,\sigma\big(f_{b_{s:},c}(x)\big)\,\mathbb{P}(\mathrm{d}x\otimes\mathrm{d}y)\right)\phi(s)\,\mathrm{d}s \qquad (5.2)
\end{aligned}$$

since $f_{a,b,c}$ is linear in a. Thus the L^2-gradient of $\mathcal{R}$ with respect to a is represented by the L^2-function

$$\delta_a\mathcal{R}(a,b,c;s) = \int_{\mathbb{R}^d}(\partial_1\ell)\big(f_{a,b,c}(x),y\big)\,\sigma\big(f_{b_{s:},c}(x)\big)\,\mathbb{P}(\mathrm{d}x\otimes\mathrm{d}y),$$

where again

$$f_{b_{s:},c}(x) = \int_0^1 b_{st}\left(\frac{1}{d+1}\sum_{i=1}^{d+1}c_{ti}x_i\right)\mathrm{d}t.$$

Using the chain rule instead of linearity, we compute

$$\begin{aligned}
&\delta_{b;\phi}\mathcal{R}(a,b,c)\\
&= \int_{\mathbb{R}^d}(\partial_1\ell)\big(f_{a,b,c}(x),y\big)\lim_{h\to 0}\frac{f_{a+h\phi,b,c}(x)-f_{a,b,c}(x)}{h}\mathbb{P}(\mathrm{d}x\otimes\mathrm{d}y)\\
&= \int_{\mathbb{R}^d}(\partial_1\ell)(\cdots)\int_0^1 a_s\lim_{h\to 0}\frac{\sigma\Big(\int_0^1\big(b_{s,t}+h\phi_{s,t}\big)\sigma(f_{c_{t:}}(x))\,\mathrm{d}t\Big)-\sigma\Big(\int_0^1 b_{s,t}\sigma(f_{c_{t:}}(x))\,\mathrm{d}t\Big)}{h}\mathrm{d}s\mathbb{P}(\mathrm{d}x\otimes\mathrm{d}y)\\
&= \int_{\mathbb{R}^d}(\partial_1\ell)(\cdots)\int_0^1 a_s\sigma'(f_{b_{s:}c}(x))\int_0^1\phi_{s,t}\sigma(f_{c_{t:}}(x))\,\mathrm{d}s\mathrm{d}t\mathbb{P}(\mathrm{d}x\otimes\mathrm{d}y)\\
&= \int_{(0,1)^2}\phi_{s,t}\left(\int_{\mathbb{R}^d}(\partial_1\ell)(\cdots)\int_0^1 a_s\sigma'(f_{b_{s:}c}(x))\sigma(f_{c_{t:}}(x))\right)\mathrm{d}s\mathrm{d}t
\end{aligned}$$

and obtain

$$\begin{aligned}
\delta_b\mathcal{R}(a,b,c;s,t) &= \int_{\mathbb{R}^d}(\partial_1\ell)\big(f_{a,b,c}(x),y\big)\,a_s\sigma'\big(f_{b_{s:}c}(x)\big)\,\sigma(f_{c_{t:}}(x))\,\mathbb{P}(\mathrm{d}x\otimes\mathrm{d}y),\\
\delta_c\mathcal{R}(a,b,c;t) &= \int_{\mathbb{R}^d}(\partial_1\ell)\big(f_{a,b,c}(x),y\big)\int_0^1 a(s)\sigma'(f_{b_{s:}c}(x))b(s,t)\sigma'(f_{c_{t:}}(x))\,\frac{x_i}{d+1}\mathrm{d}s\mathbb{P}(\mathrm{d}x\otimes\mathrm{d}y).
\end{aligned}$$

We can now see by comparing the terms that the gradient flow of a finite number of weights, interpreted as a step function, is a solution to the L^2-gradient flow under the appropriate time-scaling.

The general case for deep neural networks follows the same way, in which case

$$\begin{aligned}&\delta_{a^\ell}\mathcal{R}(a^L,\cdots,a^0;\theta_{\ell+1},\theta_\ell)\\ =&\int_{\mathbb{R}^d}(\partial_1\ell)\big(f_{a^L,\cdots,a^0}(x),y\big)\int_{(0,1)^{L-\ell-1}}a^L_{\theta_L}\sigma'(f_{a^{L-1}_{\theta_{L:}}\cdots a^0}(x))\cdots a^{\ell+1}_{\theta_{\ell+1}\theta_\ell}\\ &\qquad\sigma'(f_{a^\ell_{\theta_{\ell:}}\cdots a^0}(x))\sigma(f_{a^{\ell-1}_{\theta_{\ell:}}\cdots a^0}(x))\,\mathrm{d}\theta_L\cdots\mathrm{d}\theta_{\ell+2}\mathbb{P}(\mathrm{d}x\otimes\mathrm{d}y).\end{aligned}$$

This completes the proof. □

If the learning rates are not adapted to the layer width, the weights of different layers may move at different rates. In the natural time scaling, some layers would evolve at positive speed while others would remain frozen at their initial position in the limit. In particular, if the width of the two outermost layers goes to infinity, the index set of the second layer has size $m_L m_{L-1} \gg m_L$, meaning that the outermost layer would move much faster. In [3], the authors consider the opposite extreme where the coefficients of the first and last layers are frozen and only intermediate layers evolve (with $m_\ell \equiv m$ for all ℓ).

Remark 5.1. Alternative proposals for multi-layer network training in mean field scaling [3, 35, 36, 39]. In this article, we opted for a particularly simple description of wide multi-layer networks and the natural extension of gradient descent dynamics. All results proved here hold for networks with finite layers of any width and therefore should remain valid more generally for another description of the parameter distribution associated to infinitely wide multi-layer networks.

5.2 Growth of the path norm

Assuming existence of the gradient-flow evolution for the moment, we prove that the path-norm of an arbitrarily wide neural network increases at most polynomially in time under natural training dynamics. First, we consider the second moments.

Lemma 5.2. *Consider the risk functional*

$$\mathcal{R}(a^L,\cdots a^0)=\int_{\mathbb{R}^d}\ell\big(f_{a^L,\cdots,a^0}(x),y\big)\,\mathbb{P}(\mathrm{d}x\otimes\mathrm{d}y),$$

where $\ell:\mathbb{R}\times\mathbb{R}\to[0,\infty)$ *is a sufficiently smooth loss function and* $\mathbb{P}$ *is a compactly supported data distribution. Then*

$$\left\|a^i(t)\right\|_{L^2(\pi^{i+1}\otimes\pi^i)}\leq\left\|a^i(0)\right\|_{L^2(\pi^{i+1}\otimes\pi^i)}+\sqrt{\mathcal{R}\big(a^L(0),\cdots a^0(0)\big)}\,t^{1/2}. \tag{5.3}$$

Proof. We calculate

$$\begin{aligned}
&\frac{d}{dt}\int_{\Omega_{i+1}\times\Omega_i}\left(a^i_{\theta_{i+1}\theta_i}(t)\right)^2(\pi^{i+1}\otimes\pi^i)(\mathrm{d}\theta_{i+1}\otimes\mathrm{d}\theta_i)\\
=&2\int_{\Omega_{i+1}\times\Omega_i}a^i_{\theta_{i+1}\theta_i}(t)\frac{da^i_{\theta_{i+1}\theta_i}(t)}{dt}\mathrm{d}\theta_{i+1}\mathrm{d}\theta_i\\
\leq&2\left(\int_{\Omega_{i+1}\times\Omega_i}\left(a^i_{\theta_{i+1}\theta_i}(t)\right)^2\mathrm{d}\theta_{i+1}\mathrm{d}\theta_i\right)^{\frac{1}{2}}\left(\int_{\Omega_{i+1}\times\Omega_i}\left(\frac{d}{dt}a^i_{\theta_{i+1}\theta_i}(t)\right)^2\mathrm{d}\theta_{i+1}\mathrm{d}\theta_i\right)^{\frac{1}{2}}
\end{aligned}$$

so

$$\begin{aligned}
\frac{d}{dt}\|a_i\|_{L^2(\pi^{i+1}\otimes\pi^i)}&=\frac{\frac{d}{dt}\|a_i\|^2_{L^2(\pi^{i+1}\otimes\pi^i)}}{2\|a_i\|_{L^2(\pi^{i+1}\otimes\pi^i)}}\\
&\leq\left\|\frac{d}{dt}a^i\right\|_{L^2(\pi^{i+1}\otimes\pi^i)}\\
&\leq\left|\frac{d}{dt}\mathcal{R}(a^L,\cdots a^0)\right|^{\frac{1}{2}}
\end{aligned}$$

since the L^2-gradient flow naturally satisfies the energy dissipation identity

$$\frac{d}{dt}\mathcal{R}(a^L,\cdots a^0)=-\sum_{i=0}^{L}\left\|\frac{d}{dt}a^i\right\|^2_{L^2(\pi^{i+1}\otimes\pi^i)}.$$

Thus

$$\begin{aligned}
&\left\|a^i(t)\right\|_{L^2(\pi^{i+1}\otimes\pi^i)}\\
\leq&\left\|a^i(0)\right\|_{L^2(\pi^{i+1}\otimes\pi^i)}+\int_0^t\frac{d}{ds}\|a_i(s)\|_{L^2(\pi^{i+1}\otimes\pi^i)}\mathrm{d}s\\
\leq&\left\|a^i(0)\right\|_{L^2(\pi^{i+1}\otimes\pi^i)}+\left(\int_0^t 1\mathrm{d}s\right)^{\frac{1}{2}}\left(\int_0^t\left|\frac{d}{ds}\mathcal{R}\left(a^L(s),\cdots a^0(s)\right)\right|\mathrm{d}s\right)^{\frac{1}{2}}\\
\leq&\left\|a^i(0)\right\|_{L^2(\pi^{i+1}\otimes\pi^i)}+\sqrt{\mathcal{R}\left(a^L(0),\cdots a^0(0)\right)}\,t^{1/2}
\end{aligned}$$

since the risk is monotone decreasing and bounded from below by zero. □

Remark 5.2. Like in [42, Lemma 3.3], a more careful analysis shows that the increase in the L^2-norm actually satisfies the stronger estimate

$$\lim_{t\to\infty}\frac{\|a^i(t)\|_{L^2}}{t^{1/2}}=0.$$

The proof of this result is based on the energy dissipation identity which characterizes weak solutions to gradient flows.

Corollary 5.1. *Assume that* $\|a^i(0)\|_{L^2(\pi^{i+1}\otimes\pi^i)} \leq C_0$ *for all* $i=0,\cdots,L$ *and some constant* $C_0>0$. *Then*

$$\|f_{a^L(t),\cdots,a^0(t)}\|_{\Omega_L,\cdots,\Omega_0;K} \leq \left(C_0+\sqrt{\mathcal{R}\big(a^L(0),\cdots a^0(0)\big)}\,t^{1/2}\right)^{L+1} \tag{5.4}$$

for all $t>0$.

Proof. Follows from Lemmas 4.2 and 5.2. □

As such, neural tree spaces are also the relevant class of function spaces for suitably initialized neural networks which are trained by a gradient descent algorithm. Like in [41, Theorem 2], the slow increase of the norm together with the poor approximation property from Corollary 3.4 implies that the training of multi-layer networks may be subject to the curse of dimensionality when trying to approximate general Lipschitz functions in $L^2(\mathbb{P})$ for a truly high-dimensional data-distribution $\mathbb{P}$.

Corollary 5.2. *Consider population and empirical risk functionals*

$$\mathcal{R}(a^L,\cdots,a^0)=\frac{1}{2}\int_{[0,1]^d}(f_{a^L,\cdots,a^0}-f^*)^2(x)\,\mathrm{d}x, \qquad \mathcal{R}_n(a^L,\cdots,a^0)=\frac{1}{2n}\sum_{i=1}^n(f_\pi-f^*)^2(x_i),$$

where f^* *is a Lipschitz-continuous target function and the points* x_i *are iid samples from the uniform distribution on* $[0,1]^d$. *There exists* f^* *satisfying*

$$\sup_{x\in[0,1]^d}\big|f^*(x)\big|+\sup_{x\neq y}\frac{|f^*(x)-f^*(y)|}{|x-y|}\leq 1$$

such that the weight functions of $a^L,\cdots,a^0$ *evolving by* L^2*-gradient flow of either* $\mathcal{R}_n$ *or* $\mathcal{R}$ *satisfy*

$$\limsup_{t\to\infty}\big[t^\gamma\mathcal{R}(a^L(t),\cdots,a^0(t))\big]=\infty$$

for all $\gamma>\frac{2L}{d-2}$.

6 Conclusion

The classical function spaces which have been proved very successful in low-dimensional analysis (Sobolev, BV, BD, $\cdots$) seem ill-equipped to tackle problems in machine learning. The situation has been partially remedied in some cases by introducing the function spaces associated to different models, like reproducing kernel Hilbert spaces for random feature models, Barron space for two-layer neural networks or the flow-induced function space for infinitely deep ResNets [18].

In this article, we introduced several function classes for fully connected multi-layer feed-forward networks:

1. The neural tree spaces $\mathcal{W}^L(K)$ for questions related to approximation theory and variational analysis,

2. the classes of arbitrarily wide neural networks modelled on general index spaces $\Omega_L, \cdots, \Omega_0$, which we denoted by $X_{\Omega_L,\cdots,\Omega_0;K}$, and

3. the classes of arbitrarily wide neural networks modelled on general index spaces with Hilbert weights (or multi-layer spaces), which we denoted by $\mathcal{W}_{\pi^L,\cdots,\pi^0}(K)$.

The key to the definition of these spaces is the representation of functions.

Neural tree spaces are built using a tree-like index structure, and network weights have no natural meaning. This point of view thus cannot encompass training algorithms which operate on network weights. By analogy with classical approximation theory, we can think of finite neural networks as polynomials (finitely parametrized functions) and of neural tree spaces as Sobolev or Besov classes obtained as the closure under a weak norm, but too general for classical Taylor series. We denoted these by $\mathcal{W}$ for 'wide' structures.

The classes of arbitrarily wide neural networks are introduced as very general function classes which exhibit the natural neural network structure via generalized index spaces. In the general class of arbitrarily wide networks, weight functions are assumed to be merely measurable with integrable products, which is a too large space to study training dynamics. The restriction of the multi-layer norm to this space is a natural norm, and the closure of the unit ball in the space of arbitrarily wide neural networks and neural tree space coincides.

To study training dynamics, we consider the space of arbitrarily wide neural networks with Hilbert weights, where the L^2-inner product induces a gradient flow in the natural way. The restriction of the path norm does not control the L^2-magnitude of the weight functions, so we studied a different measure of complexity on this function space (which is not usually a norm). The complexity measure was seen to bound the path-norm from above and to grow at most like $t^{\frac{L+1}{2}}$ in time under gradient flow training.

It is immediate that $\mathcal{W}_{\pi^L,\cdots,\pi^0}(K) \subseteq X_{\Omega_L,\cdots,\Omega_0;K} \subseteq \mathcal{W}^L(K)$ with inclusions that are strict if the index spaces are finite. Whether the inclusions are strict in the general case, is not clear. In the case of three-layer networks, they can be interpreted as the spaces in which the first hidden layer is wide enough to output Barron space, a separable subspace of Barron space and a reproducing kernel Hilbert space respectively. All three spaces contain all Barron functions and their compositions.

One naturally asks which one of these spaces is most suited for describing multi-layer neural networks. An ideal space should (1) be complete, (2) have a nice approximation theory, (3) have a low Rademacher complexity, and (4) most importantly, be concrete enough so that one can make use of the function representation for practical purposes. At this point, we cannot prove any of the spaces introduced here satisfies all these requirements. Our feeling is that the space $\mathcal{W}_{\pi^L,\cdots,\pi^0}(K)$ for sufficiently large index spaces

$(\Omega_\ell, \mathcal{A}_\ell, \pi^\ell)$ might be the most promising one, even though at this point it is only a metric vector space, not a normed space (see Definitions 4.3 and 4.4 and the surrounding paragraphs.). However, it seems to be the most relevant space for practical purposes.

A number of questions remain open.

1. Beyond first observations, the relationship between the neural tree spaces $\mathcal{W}^L(K)$ and its subspace $\mathcal{W}_{\pi^L,\cdots,\pi^0}(K)$ for sufficiently expressive index sets remains unexplored. The first space is suited for variational and approximation problems, while the second is a natural object for mean-field training. It is an important question how much of the hypothesis space we can explore using natural training dynamics.

   Even for networks with two hidden layers, only heuristic observations about $\mathcal{W}^2(K)$ and its subspaces $\mathcal{W}_{\mathcal{L},\mathcal{L},\pi^0;K}$ and $\widetilde{\mathcal{W}}^2 = X_{(0,1),(0,1),\{0,\cdots,d\};K}$ of network-like functions are available. Whether the two can be treated in a unified perspective remains to be seen.

2. The direct approximation theorem holds for neural tree spaces, but not with the Monte-Carlo rate(in terms of free parameters). Whether a better rate can be achieved for functions in neural tree space for $L \gg 1$ (or at least a space of arbitrarily wide neural networks) remains an important open problem.

3. The properties of the complete metric vector spaces $\mathcal{W}_{\pi^L,\cdots\pi^0}(K)$ have not been studied yet.

4. We defined a monotonically increasing sequence of spaces $\mathcal{W}^L$ for $L \in \mathbb{N}$. Examples 2.1 and 2.2 show that $\mathcal{B}_{X,K}$ may much larger than X or exactly the same, depending on X. Concerning neural networks, it is clear that $\mathcal{W}^1$ is much larger than $\mathcal{W}^0$. In [23], we give an easy to check criterion which implies that a function is not in $\mathcal{W}^1$ and provide examples of functions which are in $\mathcal{W}_{\mathcal{L},\mathcal{L},\pi^0}(K) \subseteq \mathcal{W}^2$, but not $\mathcal{W}^1$. Beyond this, the relationship between the spaces $\mathcal{W}^\ell$ and $\mathcal{W}^L$ for $\ell < L$ is largely unexplored.

5. In this paper, we considered the minimization of an integral risk functional. A more classical problem in numerical analysis concerns the discretization of variational problems and partial differential equations. In both applications, a key component is the approximation of a solution f^* of the problem by functions f_m in a finitely parameterized hypothesis class (Galerkin spaces or neural networks). Often, the approximation rate $\|f_m - f^*\| \leq m^{-\alpha}$ of solutions f_m of the discretized problem to the true solution depends on the properties of f^* (as well as the choice of norm).

   For many variational problems and partial differential equations, a priori estimates on the solutions in Sobolev or Hölder spaces are available. The regularity of f^* is therefore understood, as well as the expected rate of convergence $f_m \to f$.

In machine learning, a regularity theory of this type is generally missing. It is often unclear in which function space the minimizer of a well-posed risk functional should lie, and thus equally unclear what type of machine learning model to use (random feature model, shallow neural network, deep neural network, ResNet, $\cdots$). A regularity theory which bounds the necessary number of layers in a neural network from above or below even for specific learning applications is not yet available.

As shown in Corollary 5.2, gradient descent may converge very slowly if the target function does not lie in the correct target space and $\frac{L}{d} \ll 1$.

6. Even assuming that the solution to a variational problem is known explicitly, it remains difficult to decide whether it lies in $\mathcal{W}^L$ for a given L. Only for $L=1$ a positive criterion is given in [5] and a negative criterion following [23, Theorem 5.4]. In general, it remains hard to check whether a function can be expressed as a neural network of depth L.

7. In this article, we focused on fully connected networks with infinitely wide layers. The theory for other types of neural networks (convolutional, recurrent, residual) will be the subject of future work.

   Starting with the articles [13, 15, 27, 31], deep ResNets have been modeled as discretizations of an ODE flow (sometimes referred to as 'neural ODEs'). A function space for infinitely deep residual networks with skip-connections after *every* layer has been proposed in [18]. In this model, the width of incremented layers is constant, but the width of the residual block may go to infinity. The case of ResNets which are both very wide and very deep and have skip-connections every $\ell \geq 1$ layers is currently unexplored.

   As demonstrated in Example 2.4, Rademacher complexity cannot give a significantly better generalization bound for the space of convolutional networks than for the space of fully connected networks. Despite many heuristic explanations, the factors contributing to the success of convolutional networks in image processing have not been understood rigorously (for non-linear activation functions).

8. Even for finite neural networks with ReLU activation and more than one hidden layer, we are not aware of rigorous results for the existence of solutions to the gradient flow equations in any strong or weak formulation.

9. In many applications, neural networks are initialized with parameters that scale in such a way that the path-norm grows beyond all bounds as the number of neurons increases. Learning rates may not be adapted to the width of the layers in applications, and the scaling invariance $\sigma(z) \equiv \lambda\sigma(\lambda^{-1}z)$ for $\lambda > 0$ may lead to coefficients which are of very different magnitude on different layers. In this situation, our analysis does not apply, and it can be shown rigorously in some cases that very wide networks of fixed depth may behave like linear models [1, 10, 12, 20, 21, 28].

These analyses typically make use of over-parametrization by assuming that the network has many more neurons than the data set has training samples. In this scaling regime, the correct function spaces and training dynamics for wide networks under population risk are generally unexplored.

Acknowledgment

This work was supported in part by a gift of iFlytek to Princeton University.

Appendices

A A brief review of measure theory

We briefly review some notions of measure theory used throughout the article. We assume familiarity with the basic notions of topology, measure theory, and functional analysis (metrics, topologies, σ-algebras, measures, Banach spaces, dual spaces, weak topologies, $\cdots$). Further background material can be found e.g. in [7, 16, 30, 34, 43].

A.1 General measure theory

Let $(X,\mathcal{A})$ be a measurable space. A signed measure is a map $\mu:\mathcal{A}\to\mathbb{R}\cup\{-\infty,\infty\}$ such that for any collection $\{A_i\}_{i\in\mathbb{Z}}$ of measurable disjoint sets we have

$$\mu\left(\bigcup_{i=1}^{\infty}A_i\right)=\sum_{i=1}^{\infty}\mu(A_i)$$

(σ-additivity), assuming that the right hand side is defined. A signed measure μ admits a Hahn decomposition $\mu=\mu^+-\mu^-$ where μ^+,μ^- are mutually singular (non-negative) measures. All proofs for this section can be found in [30, Chapter 7.5] for proofs in this section. Being mutually singular means that there exist measurable sets A_+,A_- such that

$$\mu^+(A^+)=\mu^+(X),\quad \mu^-(A^+)=0,\qquad \mu^+(A^-)=0,\quad \mu^-(A^-)=\mu^-(X),$$

i.e. μ_+,μ_- "live" on different subset of X. The (non-negative) measure $|\mu|=\mu^++\mu^-$ is called the total variation measure of μ. The total variation norm of μ is defined as

$$\|\mu\|=|\mu|(X)=\mu^+(A^+)+\mu^-(A^-)=\sup_{A,A'\in\mathcal{A}}\mu(A)-\mu(A').$$

Let X,Y be measurable spaces, $\phi:X\to Y$ a measurable map and μ a (signed) measure on X. Then we define the push-forward $\phi_\sharp\mu$ of μ along ϕ by $(\phi_\sharp\mu)(A)=\mu(\phi^{-1}(A))$ for all measurable $A\subseteq Y$. Note that by definition

$$\int_X f(\phi(x))\,\mu(\mathrm{d}x)=\int_Y f(y)\,(\phi_\sharp\mu)(\mathrm{d}y)\qquad\forall\, f:Y\to\mathbb{R}.$$

Furthermore, $\|\phi_\sharp\mu\| \leq \|\mu\|$ (since the images $\phi(A^+)$ and $\phi(A^-)$ may intersect non-trivially) and $\|\phi_\sharp\mu\| = \|\mu\|$ if μ is a (non-negative) measure (since no cancellations can occur).

A.2 Measure theory and topology

All measurable spaces considered in this article have compatible topological and measure theoretic structures. The following kind of spaces have proved to be well suited for many applications.

Definition A.1. *A* Polish space *is a second countable topological space X such that there exists a metric d on X which induces the topology of X and such that (X,d) is a complete metric space.*

In particular, compact metric spaces are Polish. Since Polish spaces are metrizable, being second countable and separable is equivalent here.

Lemma A.1. [16, Appendix A.22] *Let X,Y be Polish spaces. The following are Polish spaces.*

1. *An open subset $U \subseteq X$ with the subspace topology.*
2. *A closed subset $U \subseteq X$ with the subspace topology.*
3. *$X \times Y$ with the product topology.*

All but the first point are trivial. If U is a non-empty open set, note that the metric

$$d_U(x,x') = d(x,x') + \big|f_U(x) - f_U(x')\big|, \qquad f_U(x) = \frac{1}{\operatorname{dist}(x,\partial U)}$$

induces the same topology as d on U and is complete if d is complete on X. There are various compatibility notions between the topological structure and measure theoretic structure of a space X.

Definition A.2. *Let X be a Hausdorff space (so that compact sets are closed $\Rightarrow$ Borel).*

1. *The Borel σ-algebra is the σ-algebra generated by the collection of open subsets of X. We will always assume that measures are defined on a the Borel σ-algebra.*
2. *A measure μ is called* locally finite *if every set $x \in X$ has a neighbourhood U such that $\mu(U) < \infty$. Locally finite measures are also referred to as* Borel measures.
3. *A measure μ is called* inner regular *if*

$$\mu(A) = \sup\{\mu(K) \,|\, K \subseteq A, K \text{ is compact}\}$$

for all measurable sets A. An inner regular Borel measure is called a Radon measure.

4. *A measure μ is called* outer regular *if*

$$\mu(U) = \inf\{\mu(U) \,|\, A \subseteq U, U \text{ is open}\}$$

for all measurable sets A. A measure is called regular *if it is both inner and outer regular.*

5. *A measure μ is called* moderate *if $X = \bigcup_{k=1}^{\infty} U_k$ where the U_k are open sets of finite measure.*

On Polish spaces, most measures of importance are Radon measures. The following result is due to Ulam.

Theorem A.1. [16, Kapitel VIII, Satz 1.16] *Let X be a Polish space. Then every Borel measure μ on X is moderate and regular (in particular, a Radon measure).*

For Radon measures, we can define the analogue of the support of a function to capture the set the measure 'sees'.

Definition A.3. *Let μ be a Radon measure. We set*

$$\operatorname{spt}(\mu) = \bigcap_{K \text{ closed}, \, \mu(X \setminus K) = 0} K.$$

The support of a measure is closed. Note that the measure $\mu = \sum_{i=1}^{\infty} a_i \delta_{q_i}$ has support $\mathbb{R}$ if a_i is a summable sequence of positive numbers and q_i is an enumeration of $\mathcal{Q}$. We say that μ *concentrates* on $\mathcal{Q}$ since $\mu(\mathbb{R} \setminus \mathcal{Q}) = 0$. The support of a measure μ can be significantly larger than a set on which μ concentrates.

A.3 Continuous functions on metric spaces

In many analysis classes, the space of continuous functions on $[0,1]$ is shown to be separable as a corollary to the Stone-Weierstrass theorem with the dense set of polynomials with rational coefficients. This can be shown in a simpler way and greater generality.

Theorem A.2. *Let X be a compact metric space and $C(X)$ the space of continuous real-valued functions on X with the supremum norm. Then $C(X)$ is separable.*

Proof. Since X is compact, it has a countable dense subset $\{x_n\}_{n \in \mathbb{N}}$. Consider a family of continuous functions $\eta_{n,m} : X \to [0,1]$ such that

$$\eta_{n,m}(x) = \begin{cases} 1, & d(x, x_n) \leq \frac{1}{m}, \\ 0, & d(x, x_n) \geq \frac{2}{m}. \end{cases}$$

Denote

$$\mathcal{F}_{n,m} = \left\{ \sum_{i=1}^{n} \sum_{j=1}^{m} a_{i,j} \eta_{i,j}(x) \,\middle|\, a_{i,j} \in \mathbb{Q} \,\forall\, i,j \in \mathbb{N} \right\}, \qquad \mathcal{F} = \bigcup_{n,m=1}^{\infty} \mathcal{F}_{n,m}.$$

Then $\mathcal{F}$ is a countable subset of $C(X)$. If $f : X \to \mathbb{R}$ is continuous, it is uniformly continuous, and it is easy to see by contradiction that f can be approximated uniformly by functions in $\mathcal{F}$. ∎

Remark A.1. The same holds for the space of continuous functions from a compact metric space X into a separable metric space Y with the metric

$$d(f,g) = \sup_{x\in X} d_Y(f(x),g(x))$$

and more generally on locally compact Hausdorff spaces and the compact-open topology on the space of continuous maps.

A.4 Measure theory and functional analysis

Radon measures allow a convenient functional analytic interpretation due to the following Riesz representation theorem. We only invoke the theorem in the special case of compact spaces and note that compact metric spaces are both locally compact and separable. The same result holds in greater generality, which we shall avoid to focus on the setting where the space of continuous functions is a Banach space.

Theorem A.3. [2, Theorem 1.54] *Let X be a compact metric space and $C(X;\mathbb{R}^m)$ the space of all continuous $\mathbb{R}^m$-valued functions on X. Let L be a continuous linear functional on $C(X;\mathbb{R}^m)$. Then there exist a (non-negative) Radon measure μ and a μ-measurable function $\nu : X \to S^{m-1}$ such that*

$$L(f) = \int_X \langle f(x),\nu(x)\rangle \,\mu(\mathrm{d}x) \qquad \forall\, f\in C(X;\mathbb{R}^m).$$

Furthermore, $\|L\|_{C(X;\mathbb{R}^m)^*} = \|\mu\|$.

Denote by $\mathcal{A}$ the Borel σ-algebra of X. The function

$$\nu\cdot\mu : \mathcal{A}\to\mathbb{R}^m, \qquad (\nu\cdot\mu)(A) = \int_A \nu(x)\,\mu(\mathrm{d}x)$$

is called a *vector valued Radon measure* if $m\geq 2$ (and a *signed Radon measure* if $m=1$). Vector-valued Radon measures are σ-additive on the Borel σ-algebra. The measure μ is called the total variation measure of $\nu\cdot\mu$. In the following, we will denote vector-valued Radon measures simply by μ and the total variation measure by $|\mu|$, like we did before for signed measures. The theorem admits the following interpretation and extension.

Theorem A.4. *The dual space of $C(X;\mathbb{R}^m)$ is the space of $\mathbb{R}^m$-valued Radon measures $\mathcal{M}(X;\mathbb{R}^m)$ with the norm*

$$\|\mu\|_{\mathcal{M}(X;\mathbb{R}^m)} = |\mu|(X).$$

We denote the space of $\mathbb{R}^m$-valued Radon measures by $\mathcal{M}(X;\mathbb{R}^m)$ and $\mathcal{M}(X;\mathbb{R}) =: \mathcal{M}(X)$.

Definition A.4. *We say that a sequence of (signed, vector-valued) Radon measures μ_n converges weakly to μ and write $\mu_n \rightharpoonup \mu$ if*

$$\int_X f(x)\,\mu_n(\mathrm{d}x) \to \int_X f(x)\,\mu(\mathrm{d}x) \qquad \forall\, f\in C(X) = C(X;\mathbb{R}).$$

In this terminology, the weak convergence of Radon measures coincides with weak* convergence in the dual space of $C(X)$. By the Banach-Alaoglu theorem [7, Theorem 3.16], the unit ball of $\mathcal{M}(X)$ is compact in the weak* topology. Since $C(X)$ is separable, the weak* topology of $\mathcal{M}(X)$ is metrizable [7, Theorem 3.28]. Thus if μ_n is a bounded sequence in $\mathcal{M}(X)$, there exists a weakly convergent subsequence. This establishes the *compactness theorem for Radon measures*.

Theorem A.5. *Let μ_n be a sequence of (signed, vector-valued) Radon measures such that $\|\mu_n\| \leq 1$. Then there exists a (signed, vector-valued) Radon measure μ such that $\mu_n \rightharpoonup \mu$.*

A good exposition in the context of Euclidean spaces can be found in [14, Chapter 1] with arguments which can be applied more generally.

A.5 Bochner integrals

Bochner integrals are a generalization of Lebesgue integrals to functions with values in Banach spaces. A quick introduction can be found e.g. in [43, Chapter V, part 5] or [38, Kapitel 2.1].

Definition A.5. *Let $(X, \mathcal{A}, \mu)$ be a measure space and Y a Banach space. A function $f : X \to Y$ is called* Bochner-measurable *if there exists a sequence of step functions $f_n = \sum_{i=1}^{n} y_i \chi_{A_i}$ with $y_i \in Y, A_i \in \mathcal{A}$ such that $f_n \to f$ pointwise μ-almost everywhere.*

For real-valued functions, Bochner-measurability coincides with the usual notion of measurability.

Lemma A.2. *Let X be a compact metric space, $\mathcal{A}$ its Borel sigma algebra, μ a measure on $\mathcal{A}$ and Y a Banach space. Then every continuous function $f : X \to Y$ is uniformly continuous and thus Bochner-measurable.*

A function f is *Bochner-integrable* if the integrals $\sum_{i=1}^{n} \mu(A_i) y_i$ of the approximating sequence f_n converge and do not depend on the choice of f_n.

Lemma A.3. *Let X be a compact metric space, $\mathcal{A}$ its Borel sigma algebra, μ a finite measure on $\mathcal{A}$ and Y a Banach space. Then every continuous function $f : X \to Y$ is additionally bounded and thus Bochner-integrable.*

Bochner-integrals are linked to Lebesgue-integrals in the following way.

Lemma A.4. *Let f be a Bochner-measurable function. Then f is Bochner-integrable if and only if $\|f\| : X \to \mathbb{R}$ is Lebesgue-integrable. Furthermore,*

$$\left\| \int_X f(x) \mu(\mathrm{d}x) \right\|_Y \leq \int_X \|f(x)\|_Y \mu(\mathrm{d}x).$$

If μ is a finite signed measure, these notions generalize in the obvious way.

Definition A.6. *Let $(\Omega,\mathcal{A},\mu)$ be a measure space, $p\in[1,\infty]$ and X a Banach space. Then the Bochner space $L^p(\Omega;X)$ is the space of all Bochner-measurable functions $f:\Omega\to X$ such that $\|f\|\in L^p(\Omega)$.*

The following is proved in the unnumbered example following [38, Lemma 1.23]. The claim is formulated in the special case where Ω_1 is an interval and $\Omega_2\subseteq\mathbb{R}^d$, but the proof holds more generally.

Lemma A.5. *Let $(\Omega_i,\mathcal{A}_i,\mu_i)$ be measure spaces for $i=1,2$. Then $f\in L^p(\mu_1\otimes\mu_2)$ if and only if the function*

$$F:\Omega_1\to L^p(\Omega_2),\qquad \big[F(\omega_1)\big](\omega_2)=f(\omega_1,\omega_2)$$

is well-defined and in $L^p(\Omega_1,L^p(\Omega_2))$.

Furthermore, we recall the following immediate result, which we will apply in conjunction with the previous lemma in the special case that $H=L^2(0,1)$.

Lemma A.6. *If H is a Hilbert space, so is $L^2(\Omega;H)$ with the inner production*

$$\langle f,g\rangle_{L^2(H)}=\int_\Omega\langle f(\omega),g(\omega)\rangle\,\mu(\mathrm{d}\omega).$$

B On the existence and uniqueness of the gradient flow

For networks with smooth activation functions, the preceding analysis can be justified rigorously. We briefly discuss some obstacles in the case of ReLU activation.

Example B.1. Generically, solutions of gradient flow training for ReLU-activation are non-unique, even for functions with one hidden layer. We consider a network with one hidden layer, one neuron, and a risk functional with one data point:

$$f_{a,b}(x)=a\sigma(b^1x-b^2),\qquad \mathcal{R}(a,b)=\big|f_{a,b}(1)-1\big|^2=\big|a(b^1-b^2)_+-1\big|^2.$$

If a,b is initialized as $a_0=1$, $b_0=(1,1)$, then one solution of the gradient flow inclusion is constant in time. This solution is obtained as the limit of gradient flow training for regularized activation functions σ^ε satisfying $(\sigma^\varepsilon)'(0)=0$. Another solution is the solution (a,b) of ReLU training is

$$\begin{pmatrix}\dot a_t\\ \dot b_t^1\\ \dot b_t^2\end{pmatrix}=-\nabla_{a,b}\big|a(b^1-b^2)-1\big|^2=-2\big(a(b^1-b^2)-1\big)\begin{pmatrix}b^1-b^2\\ a\\ -a\end{pmatrix},$$

for which the risk decays to zero. This is obtained as the limit of approximating gradient flows associated to σ^ε with $(\sigma^\varepsilon)'(0)=1$.

As the training dynamics are non-unique, the Picard-Lindelöff theorem cannot apply. In [42, Lemma 3.1], we showed that the situation can be remedied by considering gradient flows of population risk for suitably regular data distributions $\mathbb{P}$. A key ingredient of the proof is that for fixed w, $\sigma'(w^T x)$ is well-defined except on a hyper-plane in $\mathbb{R}^d$, which we assume to be $\mathbb{P}$-null sets. An existence proof based on the Peano existence theorem is also presented in a specific context in [8].

This argument cannot be extended to networks with multiple hidden layers since terms of the form $\sigma'(f(x))$ occur where f can be a general Barron function (or even more general for deep networks). The level sets of Barron functions may be highly irregular and even for C^1-smooth Barron functions, Sard's theorem need not apply [23, Remark 3.2]. In particular, for any data distribution $\mathbb{P}$, we can find a non-constant Barron function f such that $\mathbb{P}(\{f=0\})>0$. It thus appears inevitable to consider a class of weak solutions based on energy dissipation properties or differential inclusions. We note that the proofs in this article are based on purely formal identities and the energy dissipation property. We thus expect the results to remain valid for suitable generalized solutions.

References

[1] S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pages 8139–8148, 2019.

[2] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of bounded variation and free discontinuity problems*, volume 254. Clarendon Press Oxford, 2000.

[3] D. Araùjo, R. I. Oliveira, and D. Yukimura. A mean-field limit for certain deep neural networks. *arXiv:1906.00193 [math.ST]*, 2019.

[4] F. Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[5] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

[6] A. R. Barron and J. M. Klusowski. Approximation and estimation for high-dimensional deep learning networks. *arXiv preprint arXiv:1809.03090*, 2018.

[7] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.

[8] L. Chizat and F. Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *arxiv:2002.04486 [math.OC]*, 2020.

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

[10] S. S. Du, J. D. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. *arXiv:1811.03804 [cs.LG]*, 2018.

[11] M. Dobrowolski. *Angewandte Funktionalanalysis: Funktionalanalysis, Sobolev-Räume und elliptische Differentialgleichungen*. Springer-Verlag, 2010.

[12] S. S. Du, X. Zhai, B. Poczos, and A. Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv:1810.02054 [cs.LG]*, 2018.

[13] W. E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017.

[14] L. C. Evans and R. F. Gariepy. *Measure theory and fine properties of functions*. CRC press, 2015.
[15] W. E, J. Han, and Q. Li. A mean-field optimal control formulation of deep learning. *Research in the Mathematical Sciences*, 6:, arXiv:1807.01083 [math.OC], 07 2018.
[16] J. Elstrodt. *Maß-und Integrationstheorie*, volume 7. Springer, 1996.
[17] W. E, C. Ma, and L. Wu. A priori estimates of the population risk for two-layer neural networks. *Comm. Math. Sci.*, 17(5):1407 – 1425 (2019), arxiv:1810.06397 [cs.LG] (2018).
[18] W. E, C. Ma, and L. Wu. Barron spaces and the compositional function spaces for neural network models. *arXiv:1906.08039 [cs.LG]*, 2019.
[19] W. E, C. Ma, and L. Wu. Machine learning from a continuous viewpoint. *arxiv:1912.12777 [math.NA]*, 2019.
[20] W. E, C. Ma, and L. Wu. A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics. *Sci. China Math.*, https://doi.org/10.1007/s11425-019-1628-5, arXiv:1904.04326 [cs.LG] (2019).
[21] W. E, C. Ma, Q. Wang, and L. Wu. Analysis of the gradient descent algorithm for a deep neural network model with skip-connections. *arXiv:1904.05263 [cs.LG]*, 2019.
[22] W. E and S. Wojtowytsch. Kolmogorov width decay and poor approximators in machine learning: Shallow neural networks, random feature models and neural tangent kernels. *arXiv:2005.10807 [math.FA]*, 2020.
[23] W. E and S. Wojtowytsch. Representation formulas and pointwise properties for Barron functions. *In preparation*, 2020.
[24] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.
[25] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.
[26] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.
[27] E. Haber and L. Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):014004, 2017.
[28] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.
[29] J. M. Klusowski and A. R. Barron. Risk bounds for high-dimensional ridge function combinations including neural networks. *arXiv preprint arXiv:1607.01434*, 2016.
[30] A. Klenke. *Wahrscheinlichkeitstheorie*, volume 1. Springer, 2006.
[31] Q. Li, L. Chen, C. Tai, and W. E. Maximum principle based algorithms for deep learning. *The Journal of Machine Learning Research*, 18(1):5998–6026, 2017.
[32] 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.
[33] G. Lorentz. *Approximation of Functions*. Holt, Rinehart and Winston, New York, 1966.
[34] J. R. Munkres. *Topology: a First Course*. Prentice-Hall, 1974.
[35] P.-M. Nguyen. Mean field limit of the learning dynamics of multilayer neural networks. *arXiv:1902.02880 [cs.LG]*, 2019.
[36] P.-M. Nguyen and H. T. Pham. A rigorous framework for the mean field limit of multilayer neural networks. *arXiv:2001.11443 [cs.LG]*, 2020.

[37] A. Rahimi and B. Recht. Uniform approximation of functions with random bases. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pages 555–561. IEEE, 2008.
[38] M. Růžička. *Nichtlineare Funktionalanalysis: Eine Einführung*. Springer-Verlag, 2006.
[39] J. Sirignano and K. Spiliopoulos. Mean field analysis of deep neural networks. *arXiv:1903.04440 [math.PR]*, 2019.
[40] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
[41] S. Wojtowytsch and W. E. Can shallow neural networks beat the curse of dimensionality? A mean field training perspective. *arXiv:2005.10815 [cs.LG]*, 2020.
[42] S. Wojtowytsch. On the global convergence of gradient descent training for two-layer Relu networks in the mean field regime. *arXiv:2005.13530 [math.AP]*, 2020.
[43] K. Yosida. *Functional analysis*. Springer Science & Business Media, 2012.