Constructive Approximation (2022) 55:369–406
https://doi.org/10.1007/s00365-021-09549-y

CONSTRUCTIVE APPROXIMATION

The Barron Space and the Flow-Induced Function Spaces for Neural Network Models

Weinan E[1,2,3] · Chao Ma[2] · Lei Wu[2]

Received: 14 June 2019 / Accepted: 21 November 2020 / Published online: 24 May 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

One of the key issues in the analysis of machine learning models is to identify the appropriate function space and norm for the model. This is the set of functions endowed with a quantity which can control the approximation and estimation errors by a particular machine learning model. In this paper, we address this issue for two representative neural network models: the two-layer networks and the residual neural networks. We define the Barron space and show that it is the right space for two-layer neural network models in the sense that optimal direct and inverse approximation theorems hold for functions in the Barron space. For residual neural network models, we construct the so-called flow-induced function space and prove direct and inverse approximation theorems for this space. In addition, we show that the Rademacher complexity for bounded sets under these norms has the optimal upper bounds.

Keywords Function space · Neural network · Approximation · Rademacher complexity

Mathematics Subject Classification 65D15 · 68T05 · 46B99

Communicated by Wolfgang Dahmen, Ronald A. DeVore, and Philipp Grohs.

✉ Weinan E
weinan@math.princeton.edu

Chao Ma
chaom@princeton.edu

Lei Wu
leiwu@princeton.edu

1 Department of Mathematics, Princeton University, Princeton, USA

2 Program in Applied and Computational Mathematics, Princeton University, Princeton, USA

3 Beijing Institute of Big Data Research, Beijing, China

1 Introduction

The task of supervised learning is to approximate a function using a given set of data. This type of problem has been the subject of classical numerical analysis and approximation theory for a long time. The theory of splines and the theory of finite element methods are very successful examples of such classical results [8,9], both are concerned with approximating functions using piecewise polynomials. In these theories, one starts from a function in a particular function space, say a Sobolev or Besov space and proceeds to derive optimal error estimates for this function. The optimal error estimates depend on the function norm, and the regularity encoded in the function space as well as the approximation scheme. They are the most important pieces of information for understanding the underlying approximation scheme. When discussing a particular function space, the associated norm is as crucial as the set of functions it contains.

Identifying the right function space that one should use is the most crucial step in this analysis. Sobolev/Besov type spaces are good function spaces for these classical theories since:

1. One can prove direct and inverse approximation theorems for these spaces. Roughly speaking, a function can be approximated by piecewise polynomials with certain convergence rate if and only if the function is in certain Sobolev/Besov space.
2. The functions we are interested in, e.g., solutions of partial differential equations (PDEs), are in these spaces. This is at the heart of the regularity theory for PDEs.

However, these spaces are tied with the piecewise polynomial basis used in the approximation scheme. These approximation schemes suffer from the curse of dimensionality, i.e., the number of parameters needed to achieve certain level of accuracy grows exponentially with dimension. Consequently, Sobolev/Besov type spaces are not the right function spaces for studying machine learning models that can potentially address the curse of dimensionality problem.

Another inspiration for this paper comes from kernel methods. It is well-known that the right function space associated with a kernel method is the corresponding reproducing kernel Hilbert space (RKHS) [1]. RKHS and kernel methods provide one of the first examples for which dimension-independent error estimates can be established.

The main purpose of this paper is to construct and identify the analog of these spaces for two-layer and residual neural network models. For two-layer neural network models, we show that the right function space is the so-called "Barron space." Roughly speaking, a function belongs to the Barron space if and only if it can be approximated by "well-behaved" two-layer neural networks, and the approximation error is controlled by the norm of the Barron space. The analog of the Barron space for deep residual neural networks is the "flow-induced function space" that we construct in the second part of this paper. With the "flow-induced norms," we will establish direct and inverse approximation theorems for these spaces as well as the optimal Rademacher complexity estimates.

One important difference between approximation theory in low and high dimensions is that in high dimensions, the best error rate (or order of convergence) that

one can hope for is the Monte Carlo error rate. Therefore using the error rate as an indicator to distinguish the quality of different approximation schemes or machine learning models is not a good option. The function spaces or the associated norms seem to be a better alternative. We take the viewpoint that a function space is defined by its approximation property using a particular approximation scheme. In this sense, Sobolev/Besov spaces are the result when we consider approximation by piecewise polynomials or wavelets. Barron space is the analog when we consider approximation by two-layer neural networks and the flow-induced function space is the analog when we consider approximation by deep residual networks. The norms that are associated with these new spaces may seem a bit unusual at a first sight, but they arise naturally in the approximation process, as we will see from the direct and inverse approximation theorems presented below.

It should be stressed that the terminologies "space" and "norm" in this paper are used in a loose way. For example, flow-induced norms are a family of quantities that control the approximation error. We do not take effort to investigate whether it is a real norm.

Although this work was motivated by the problem of understanding approximation theory for neural network models in machine learning, we believe that it should have an implication for high dimensional analysis in general. One natural follow-up question is whether one can show that solutions to high dimensional partial differential equations (PDE) belong to the function spaces introduced here. At least for linear parabolic PDEs, the work in [14] suggests that some close analog of the flow-induced spaces should serve the purpose.

In Sect. 2, we introduce the Barron space for two-layer neural networks. Although not all the results in this section are new (some have appeared in various forms in [2,11,15]), they are useful for illustrating our angle of attack and they are also useful for the work in Sect. 3 where we introduce the flow-induced function space for residual networks.

Notations Let $\mathbb{S}^d = \{\boldsymbol{w} \in \mathbb{R}^{d+1} : \|\boldsymbol{w}\|_1 = 1\}$. We define $\hat{\boldsymbol{w}} = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_1}$ if $\boldsymbol{w} \neq 0$ otherwise $\hat{\boldsymbol{w}} = 0$. For simplicity, we fix the domain of interest to be $X = [0, 1]^d$. We denote by $\mathbf{x} \in X$ the input variable, and let $\tilde{\mathbf{x}} = (\mathbf{x}^T, 1)^T$. We sometimes abuse notation and use $f(\mathbf{x})$ (or some other analogs) to denote the function f in order to signify the independent variable under consideration. We use $\|f\|$ to denote the L_2 norm of function f defined by

$$\|f\| = \left(\int_X |f(\mathbf{x})|^2 \mu(d\mathbf{x}) \right)^{\frac{1}{2}},$$

where $\mu(\mathbf{x})$ is a probability distribution on X. We do not specify μ in this paper.

One important point for working in high dimension is the dependence of the constants on the dimension. We will use C to denote constants that are independent of the dimension.

In Sect. 3, the absolute values and powers of matrices and vectors ($|\cdot|$ and $(\cdot)^p$) are understood as being element-wise. The multiplication of two matrices is regular matrix multiplication.

2 The Barron Space

In this section, we define the Barron space and study its properties. The proofs of theorems are postponed to the end of the section.

2.1 Definition of the Barron Space

We will consider functions $f : X \mapsto \mathbb{R}$ that admit the following representation

$$f(\mathbf{x}) = \int_{\Omega} a\sigma(\boldsymbol{b}^T\mathbf{x} + c)\rho(da, d\boldsymbol{b}, dc), \quad \mathbf{x} \in X \tag{1}$$

where $\Omega = \mathbb{R}^1 \times \mathbb{R}^d \times \mathbb{R}^1$, ρ is a probability distribution on (Ω, Σ_Ω), with Σ_Ω being a Borel σ-algebra on Ω, and $\sigma(x) = \max\{x, 0\}$ is the ReLU activation function. This representation can be considered as the continuum analog of two-layer neural networks:

$$f_m(\mathbf{x}; \Theta) := \frac{1}{m}\sum_{j=1}^{m} a_j\sigma(\boldsymbol{b}_j^T\mathbf{x} + c_j),$$

where $\Theta = (a_1, \boldsymbol{b}_1, c_1, \ldots, a_m, \boldsymbol{b}_m, c_m)$ denotes all the parameters. It should be noted that in general, the ρ's for which (1) holds are not unique.

To get some intuition about the representation (1), we write the Fourier representation of a function f as:

$$f(\mathbf{x}) = \int_{\mathbb{R}^d} \hat{f}(\omega)\cos(\omega^T\mathbf{x})d\omega = \int_{\mathbb{R}^1\times\mathbb{R}^d} a\cos(\omega^T\mathbf{x})\rho(da, d\omega),$$
$$\rho(da, d\omega) = \delta(a - \hat{f}(\omega))dad\omega. \tag{2}$$

This can be thought of as the analog of (1) for the case when $\sigma(z) = \cos(z)$ except for the fact that the ρ defined in (2) is not normalizable.

For functions that admit the representation (1), we define its Barron norm:

$$\|f\|_{\mathcal{B}_p} = \inf_{\rho}\left(\mathbb{E}_\rho[|a|^p(\|\boldsymbol{b}\|_1 + |c|)^p]\right)^{1/p}, \quad 1 \le p \le +\infty. \tag{3}$$

Here the infimum is taken over all ρ for which (1) holds for all $\mathbf{x} \in X$, and when $p = \infty$ the norm (3) becomes

$$\inf_{\rho}\max_{(a,\boldsymbol{b},c)\in\mathrm{supp}(\rho)} |a|(\|\boldsymbol{b}\|_1 + |c|).$$

Barron spaces $\mathcal{B}_p$ are defined as the set of continuous functions that can be represented by (1) with finite Barron norm. We name these spaces after Barron to honor his contribution to the mathematical analysis of two-layer neural networks, in particular the work in [4,5,15].

Remark 1 It should be noted that the Barron norm defined here is different from the spectral norm used in Barron's original papers (see for example [4]).

As a consequence of the Hölder's inequality, we trivially have

$$\mathcal{B}_\infty \subset \cdots \mathcal{B}_2 \subset \mathcal{B}_1.$$

However, the opposite is also true for the ReLU activation function we are considering.

Proposition 1 *For any $f \in \mathcal{B}_1$, we have $f \in \mathcal{B}_\infty$ and*

$$\|f\|_{\mathcal{B}_1} = \|f\|_{\mathcal{B}_\infty}.$$

As a consequence, we have that for any $1 \le p \le \infty$, $\mathcal{B}_p = \mathcal{B}_\infty$ and $\|f\|_{\mathcal{B}_p} = \|f\|_{\mathcal{B}_\infty}$. Hence, we can use $\mathcal{B}$ and $\|\cdot\|_{\mathcal{B}}$ to denote the Barron space and Barron norm.

A natural question is: What kind of functions are in the Barron space? The following is a restatement of an important result proved in [15]. It is an extension of the Fourier analysis of two-layer sigmoidal neural networks in Barron's seminal work [4].

Proposition 2 (Theorem 6 in [15]) *Let $f \in C(X)$, the space of continuous functions on X, and assume that f satisfies:*

$$\gamma(f) := \inf_{\hat{f}} \int_{\mathbb{R}^d} \|\omega\|_1^2 |\hat{f}(\omega)| d\omega < \infty,$$

where $\hat{f}$ is the Fourier transform of an extension of f to $\mathbb{R}^d$. Then f admits a representation as in (1)*. Moreover,*

$$\|f\|_{\mathcal{B}} \le 2\gamma(f) + 2\|\nabla f(0)\|_1 + 2|f(0)|. \tag{4}$$

Remark 2 In Section 9 of [4], examples of functions with bounded $\gamma(f)$ are given (e.g., Gaussian, positive definite functions, etc.). [4] used the norm $\int_{\mathbb{R}^d} \|\omega\| |\hat{f}(\omega)| d\omega$, instead of $\gamma(f)$, but the analysis also shows that Gaussian and positive definite functions give rise to finite values of $\gamma(f)$. By Proposition 2, these functions belong to the Barron space.

In addition, the Barron space is also closely related to a family of RKHS. Let $\boldsymbol{w} = (\boldsymbol{b}, c)$. Due to the scaling invariance of $\sigma(\cdot)$, we can assume $\boldsymbol{w} \in \mathbb{S}^d$. Then (1) can be written as

$$\begin{aligned} f(\mathbf{x}) &= \int_{\mathbb{S}^d} a\sigma(\boldsymbol{w}^T \tilde{\mathbf{x}}) \rho(da, d\boldsymbol{w}) = \int_{\mathbb{S}^d} a(\boldsymbol{w}) \sigma(\boldsymbol{w}^T \tilde{\mathbf{x}}) \pi(d\boldsymbol{w}), \\ a(\boldsymbol{w}) &= \frac{\int_{\mathbb{R}} a\rho(a, \boldsymbol{w}) da}{\pi(\boldsymbol{w})}, \quad \pi(\boldsymbol{w}) = \int_{\mathbb{R}} \rho(a, \boldsymbol{w}) da \end{aligned} \tag{5}$$

Moreover,

$$\|f\|_{\mathcal{B}_2}^2 = \inf_{\pi} \mathbb{E}_\pi[|a(\boldsymbol{w})|^2],$$

where the infimum is taken over all π that satisfies (5).

Given a fixed probability distribution π, we can define a kernel:

$$k_\pi(\mathbf{x}, \mathbf{x}') = \mathbb{E}_{w\sim\pi}[\sigma(\boldsymbol{w}^T\tilde{\mathbf{x}})\sigma(\boldsymbol{w}^T\tilde{\mathbf{x}}')]$$

Let $\mathcal{H}_{k_\pi}$ denote the RKHS induced by k_π. Then we have the following proposition.

Proposition 3

$$\mathcal{B} = \bigcup_{\pi\in P(\mathbb{S}^d)} \mathcal{H}_{k_\pi}.$$

2.2 Direct and Inverse Approximation Theorems

With (1), approximating f by two-layer networks becomes a Monte Carlo integration problem.

Theorem 1 *For any $f \in \mathcal{B}$ and $m > 0$, there exists a two-layer neural network $f_m(\cdot;\Theta)$, $f_m(\mathbf{x};\Theta) = \frac{1}{m}\sum_{k=1}^m a_k\sigma(\boldsymbol{b}_k^T\mathbf{x} + c_k)$ (Θ denotes the parameters $\{(a_k, \boldsymbol{b}_k, c_k), k \in [m]\}$ in the neural network), such that*

$$\|f(\cdot) - f_m(\cdot;\Theta)\|^2 \le \frac{3\|f\|_{\mathcal{B}}^2}{m},$$

Furthermore, we have

$$\|\Theta\|_{\mathcal{P}} := \frac{1}{m}\sum_{j=1}^m |a_j|(\|\boldsymbol{b}_j\|_1 + |c_j|) \le 2\|f\|_{\mathcal{B}}.$$

Remark 3 We call $\|\Theta\|_{\mathcal{P}}$ the path norm of two-layer neural network. This is the analog of the Barron norm of functions in $\mathcal{B}$. Hence, when studying approximation properties, it is natural to study two-layer neural networks with bounded path norm.

One can also prove an inverse approximation theorem. To state this result, we define:

$$\mathcal{N}_Q = \left\{ \frac{1}{m}\sum_{k=1}^m a_k\sigma(\boldsymbol{b}_k^T\mathbf{x} + c_k) : \frac{1}{m}\sum_{k=1}^m |a_k|(\|\boldsymbol{b}_k\|_1 + |c_k|) \le Q, m \in \mathbb{N}^+ \right\}.$$

Theorem 2 *Let f^* be a continuous function on X. Assume there exists a constant Q and a sequence of functions $(f_m) \subset \mathcal{N}_Q$ such that*

$$f_m(\mathbf{x}) \to f^*(\mathbf{x})$$

for all $\mathbf{x} \in X$. *Then there exists a probability distribution* ρ^* *on* (Ω, Σ_Ω), *such that*

$$f^*(\mathbf{x}) = \int a\sigma(\boldsymbol{b}^T\mathbf{x} + c)\rho^*(da, d\boldsymbol{b}, dc),$$

for all $\mathbf{x} \in X$. *Furthermore, we have* $f^* \in \mathcal{B}$ *with*

$$\|f^*\|_{\mathcal{B}} \leq Q.$$

2.3 Estimates of the Rademacher Complexity

Next, we show that the Barron spaces we defined have low complexity. We show this by bounding the Rademacher complexity of bounded sets in the Barron spaces.

Definition 1 (*Rademacher complexity*) Given a set of functions $\mathcal{F}$ and n data samples $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, the Rademacher complexity of $\mathcal{F}$ with respect to S is defined as

$$\mathrm{Rad}_n(\mathcal{F}) = \frac{1}{n}\mathbb{E}_{\boldsymbol{\xi}} \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \xi_i f(\mathbf{x}_i),$$

where $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_n)$ is a vector of n *i.i.d.* random variables that satisfy $\mathbb{P}(\xi = 1) = \mathbb{P}(\xi = -1) = \frac{1}{2}$.

The following theorem gives an estimate of the Rademacher complexity of the Barron space. Similar results can be found in [2]. We include the proof in the next section for completeness.

Theorem 3 *Let* $\mathcal{F}_Q = \{f \in \mathcal{B} : \|f\|_{\mathcal{B}} \leq Q\}$. *Then we have*

$$\mathrm{Rad}_n(\mathcal{F}_Q) \leq 2Q\sqrt{\frac{2\ln(2d)}{n}}$$

From Theorem 8 in [6], we see that the above result implies that functions in the Barron spaces can be learned efficiently .

2.4 Barron Space for Non-ReLU Functions and the Space $\mathcal{F}_1$

The definition of the Barron space and Barron norm can be extended to representations (1) with $\sigma(\cdot)$ being a general activation function. Specifically, for any function f with representation

$$f(\mathbf{x}) = \int_\Omega a\tilde{\sigma}(\boldsymbol{b}^T\mathbf{x} + c)\rho(da, d\boldsymbol{b}, dc), \quad \mathbf{x} \in X, \tag{6}$$

where $\tilde{\sigma}$ is an activation function not necessarily ReLU, we define the extended Barron norm (which is denoted by $\|\cdot\|_{\tilde{\mathcal{B}}_p}$) as

$$\|f\|_{\tilde{\mathcal{B}}_p} := \inf_{\rho} \left(\mathbb{E}_{\rho}\left[|a|^p(\|\boldsymbol{b}\|_1 + |c| + 1)^p\right]\right)^{1/p}, \tag{7}$$

where $p \in [1, \infty]$, and the infimum is taken over all ρ for which (6) holds. The extended Barron space $\tilde{\mathcal{B}}_p$ is defined as the set of functions with finite $\tilde{\mathcal{B}}_p$ norm. In this case, since the homogeneity property does not hold for the activation function, $\tilde{\mathcal{B}}_p$ spaces with different p are not equal. The direct approximation theorem and Rademacher complexity control can be proven for $\tilde{\mathcal{B}}_p$ as long as $\tilde{\sigma}$ satisfies

$$\int_{\mathbb{R}} |\tilde{\sigma}''(x)|(|x| + 1)dx < \infty.$$

See [18] for more details.

We deal with general activation functions by approximating them using two-layer ReLU neural networks, and the "+1" term in (7) appears naturally during the approximation process. It is worth mentioning that if $\tilde{\sigma}$ =ReLU the $\tilde{\mathcal{B}}_p$ norms become equivalent with the Barron norm $\|\cdot\|_{\mathcal{B}}$, because of the infimum and the homogeneity property.

In [2], a similar function space $\mathcal{F}_1$ is defined by using the variation norm [16,19]. In [2], signed measures are used to represent the function as follows,

$$f(x) = \int_{\mathcal{V}} \sigma(\boldsymbol{b}^T\mathbf{x} + c)d\mu(\boldsymbol{b}, c), \tag{8}$$

where $\mathcal{V}$ is the support of the signed measure μ. Let S_f denote the set of signed measures such that (8) holds. The $\mathcal{F}_1$ norm of f is given by

$$\|f\|_{\mathcal{F}_1} := \inf_{\mu \in S_f} |\mu|(\mathcal{V}),$$

where $|\mu|(\mathcal{V})$ denotes the total variation of μ. The estimate of Rademacher complexity of $\mathcal{F}_1$ is provided for the ReLU activation function.

For ReLU activation function, $\mathcal{F}_1$ is equivalent with $\mathcal{B}$, and the norms are equal, too [12]. However, for a general activation function (e.g., tanh, sigmoid), the Barron space is different from $\mathcal{F}_1$. $\mathcal{F}_1$ typically requires $(\boldsymbol{b}, c)$ to lie in a compact set, which is generally not true. With $(\boldsymbol{b}, c)$ being in a compact set, the variation norm only considers a and treat features with any $(\boldsymbol{b}, c)$ equivalently. Hence, a very simple feature will have the same variation norm as a complicated feature, which leads to loose bounds for simple functions. On the contrary, the $\tilde{\mathcal{B}}_p$ norms consider $(a, \boldsymbol{b}, c)$ together, and features with different $(\boldsymbol{b}, c)$ make different contributions to the norm.

2.5 Proofs

2.5.1 Proof of Proposition 1

Take $f \in \mathcal{B}_1$. For any $\varepsilon > 0$, there exists a probability measure ρ that satisfies

$$f(\mathbf{x}) = \int_{\Omega} a\sigma(\boldsymbol{b}^T\mathbf{x} + c)\rho(da, d\boldsymbol{b}, dc), \quad \forall\, \mathbf{x} \in X,$$

and

$$\mathbb{E}_\rho\left[|a|(\|\boldsymbol{b}\|_1 + |c|)\right] < \|f\|_{\mathcal{B}_1} + \varepsilon.$$

Let $\Lambda = \{(\boldsymbol{b}, c) : \ \|\boldsymbol{b}\|_1 + |c| = 1\}$, and consider two measures ρ_+ and ρ_- on Λ defined by

$$\rho_+(A) = \int_{\{(a,\boldsymbol{b},c):\ (\hat{\boldsymbol{b}},\hat{c})\in A, a>0\}} |a|(\|\boldsymbol{b}\|_1 + |c|)\rho(da, d\boldsymbol{b}, dc),$$
$$\rho_-(A) = \int_{\{(a,\boldsymbol{b},c):\ (\hat{\boldsymbol{b}},\hat{c})\in A, a<0\}} |a|(\|\boldsymbol{b}\|_1 + |c|)\rho(da, d\boldsymbol{b}, dc),$$

for any Borel set $A \subset \Lambda$, where

$$\hat{\boldsymbol{b}} = \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|_1 + |c|}, \ \ \hat{c} = \frac{c}{\|\boldsymbol{b}\|_1 + |c|}.$$

Obviously $\rho_+(\Lambda) + \rho_-(\Lambda) = \mathbb{E}_\rho\left[|a|(\|\boldsymbol{b}\|_1 + |c|)\right]$, and

$$f(\mathbf{x}) = \int_{\Lambda} \sigma(\boldsymbol{b}^T\mathbf{x} + c)\rho_+(d\boldsymbol{b}, dc) - \int_{\Lambda} \sigma(\boldsymbol{b}^T\mathbf{x} + c)\rho_-(d\boldsymbol{b}, dc).$$

Next, we define extensions of ρ_+ and ρ_- to $\{-1, 1\} \times \Lambda$ by

$$\tilde{\rho}_+(A') = \rho_+(\{(\boldsymbol{b}, c) :\ (1, \boldsymbol{b}, c) \in A'\}),$$
$$\tilde{\rho}_-(A') = \rho_-(\{(\boldsymbol{b}, c) :\ (-1, \boldsymbol{b}, c) \in A'\}),$$

for any Borel sets $A' \subset \{-1, 1\} \times \Lambda$, and let $\tilde{\rho} = \tilde{\rho}_+ + \tilde{\rho}_-$. Then we have $\tilde{\rho}(\{-1, 1\} \times \Lambda) = \mathbb{E}_\rho\left[|a|(\|\boldsymbol{b}\|_1 + |c|)\right]$ and

$$f(\mathbf{x}) = \int_{\{-1,1\}\times\Lambda} a\sigma(\boldsymbol{b}^T\mathbf{x} + c)\tilde{\rho}(da, d\boldsymbol{b}, dc).$$

Therefore, we can normalize $\tilde{\rho}$ to be a probability measure, and

$$\|f\|_{\mathcal{B}_\infty} \le \tilde{\rho}(\{-1, 1\} \times \Lambda) \le \|f\|_{\mathcal{B}_1} + \varepsilon.$$

Taking the limit as $\varepsilon \to 0$, we have $\|f\|_{\mathcal{B}_\infty} \le \|f\|_{\mathcal{B}_1}$. Since $\|f\|_{\mathcal{B}_1} \le \|f\|_{\mathcal{B}_\infty}$ from Hölder's inequality, we conclude that $\|f\|_{\mathcal{B}_1} = \|f\|_{\mathcal{B}_\infty}$. □

2.5.2 Proof of Theorem 3

According to [21], we have the following characterization of $\mathcal{H}_{k_\pi}$:

$$\mathcal{H}_{k_\pi} = \left\{ \int_{\mathbb{S}^d} a(\boldsymbol{w})\sigma(\boldsymbol{w}^T\tilde{\mathbf{x}})d\pi(\boldsymbol{w}) : \mathbb{E}_\pi[|a(\boldsymbol{w})|^2] < \infty \right\}.$$

In addition, for any $h \in \mathcal{H}_{k_\pi}$, $\|h\|^2_{\mathcal{H}_{k_\pi}} = \mathbb{E}_\pi[|a(\boldsymbol{w})|^2]$. It is obvious that for any $\pi \in P(\mathbb{S}^d)$, $\mathcal{H}_{k_\pi} \subset \mathcal{B}_2$, which implies that $\cup_\pi \mathcal{H}_{k_\pi} \subset \mathcal{B}_2$. Conversely, for any $f \in \mathcal{B}_2$, there exists a probability distribution $\tilde{\pi}$ that satisfies

$$f(\mathbf{x}) = \int_{\mathbb{S}^d} a(\boldsymbol{w})\sigma(\boldsymbol{w}^T\tilde{\mathbf{x}})\tilde{\pi}(d\boldsymbol{w}) \quad \forall \mathbf{x} \in X,$$

and $\mathbb{E}_{\tilde{\pi}}[|a(\boldsymbol{w})|^2] \le 2\|f\|^2_{\mathcal{B}_2} < \infty$. Hence we have $f \in \mathcal{H}_{k_{\tilde{\pi}}}$, which implies $\mathcal{B}_2 \subset \cup_\pi \mathcal{H}_{k_\pi}$. Therefore $\mathcal{B}_2 = \cup_\pi \mathcal{H}_{k_\pi}$. Together with Proposition 1, we complete the proof. □

2.5.3 Proof of Theorem 1

Let ε be a positive number such that $\varepsilon < 1/5$. Let ρ be a probability distribution such that $f(\mathbf{x}) = \mathbb{E}_\rho[a\sigma(\boldsymbol{b}^T\mathbf{x} + c)]$ and $\mathbb{E}_\rho[|a|^2(\|\boldsymbol{b}\|_1 + |c|)^2] \le (1+\varepsilon)\|f\|^2_{\mathcal{B}_2}$. Let $\phi(\mathbf{x};\theta) = a\sigma(\boldsymbol{b}^T\mathbf{x} + c)$ with $\theta = (a, \boldsymbol{b}, c) \sim \rho$. Then we have $\mathbb{E}_{\theta\sim\rho}[\phi(\mathbf{x};\theta)] = f(\mathbf{x})$. Let $\Theta = \{\theta_j\}_{j=1}^m$ be i.i.d. random variables drawn from $\rho(\cdot)$, and consider the following empirical average,

$$\hat{f}_m(\mathbf{x};\Theta) = \frac{1}{m}\sum_{j=1}^m \phi(\mathbf{x};\theta_j).$$

Let $\mathcal{E}(\Theta) = \mathbb{E}_{\mathbf{x}}[|\hat{f}_m(\mathbf{x};\Theta) - f(\mathbf{x})|^2]$ be the approximation error. Then we have

$$\begin{aligned}
\mathbb{E}_\Theta[\mathcal{E}(\Theta)] &= \mathbb{E}_\Theta\mathbb{E}_{\mathbf{x}}|\hat{f}_m(\mathbf{x};\Theta) - f(\mathbf{x})|^2 \\
&= \mathbb{E}_{\mathbf{x}}\mathbb{E}_\Theta|\frac{1}{m}\sum_{j=1}^m \phi(\mathbf{x};\theta_j) - f(\mathbf{x})|^2 \\
&= \frac{1}{m^2}\mathbb{E}_{\mathbf{x}}\sum_{j,k=1}^m \mathbb{E}_{\theta_j,\theta_k}[(\phi(\mathbf{x};\theta_j) - f(\mathbf{x}))(\phi(\mathbf{x};\theta_k) - f(\mathbf{x}))] \\
&\le \frac{1}{m^2}\sum_{j=1}^m \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\theta_j}[(\phi(\mathbf{x};\theta_j) - f(\mathbf{x}))^2]
\end{aligned}$$

$$\leq \frac{1}{m}\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\theta\sim\rho}[\phi^2(\mathbf{x};\theta)]$$
$$\leq \frac{(1+\varepsilon)\|f\|_{\mathcal{B}_2}^2}{m}.$$

In addition,

$$\mathbb{E}_\Theta[\|\Theta\|_{\mathcal{P}}] = \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_\Theta[\|a_j\|(\|\boldsymbol{b}_j\|_1 + |c_j|)] \leq (1+\varepsilon)\|f\|_{\mathcal{B}_2}.$$

Define the event $E_1 = \{\mathcal{E}(\Theta) < \frac{3\|f\|_{\mathcal{B}_2}^2}{m}\}$, and $E_2 = \{\|\Theta\|_{\mathcal{P}} < 2\|f\|_{\mathcal{B}_2}\}$. By Markov inequality, we have

$$\mathbb{P}\{E_1\} = 1 - \mathbb{P}\{E_1^c\} \geq 1 - \frac{\mathbb{E}_\Theta[\mathcal{E}(\Theta)]}{3\|f\|_{\mathcal{B}_2}^2/m} \geq \frac{2-\varepsilon}{3}$$
$$\mathbb{P}\{E_2\} = 1 - \mathbb{P}\{E_2^c\} \geq 1 - \frac{\mathbb{E}_\Theta[\|\Theta\|_{\mathcal{P}}]}{2\|f\|_{\mathcal{B}_2}} \geq \frac{1-\varepsilon}{2}.$$

Therefore we have

$$\mathbb{P}\{E_1 \cap E_2\} = \mathbb{P}\{E_1\} + \mathbb{P}\{E_2\} - 1 \geq \frac{2-\varepsilon}{3} + \frac{1-\varepsilon}{2} - 1 = \frac{1-5\varepsilon}{6} > 0.$$

Choose any Θ in $E_1 \cap E_2$. The two-layer neural network model defined by this Θ satisfies both requirements in the theorem. □

2.5.4 Proof of Theorem 2

Without loss of generality, we assume that $\|\boldsymbol{b}\|_1 + |c| = 1$, otherwise due to the scaling invariance of $\sigma(\cdot)$ we can redefine the parameters as follows,

$$a \leftarrow a(\|\boldsymbol{b}\|_1 + |c|), \quad \boldsymbol{b} \leftarrow \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|_1 + |c|}, \quad c \leftarrow \frac{c}{\|\boldsymbol{b}\|_1 + |c|}.$$

Let $\Theta_m = \{(a_k^{(m)}, \boldsymbol{b}_k^{(m)}, c_k^{(m)})\}_{k=1}^m$ be the parameters in the two-layer neural network model f_m and let $A = \sum_{k=1}^m |a_k|$ and $\alpha_k = \frac{|a_k|}{A}$. Then we can define a probability measure:

$$\rho_m = \sum_{k=1}^{m}\alpha_k \delta\left(a - \frac{\operatorname{sign}(a_k^{(m)})A}{m}\right)\delta(\boldsymbol{b} - \boldsymbol{b}_k^{(m)})\delta(c - c_k^{(m)}),$$

which satisfies

$$f_m(\mathbf{x};\Theta_m) = \int a\sigma(\boldsymbol{b}^T\mathbf{x} + c)\rho_m(da, d\boldsymbol{b}, dc).$$

Let

$$K_Q = \{(a, \boldsymbol{b}, c) : |a| \le Q, \|\boldsymbol{b}\|_1 + |c| \le 1\}.$$

It is obvious that $\mathrm{supp}(\rho_m) \subset K_Q$ for all m. Since K_Q is compact, the sequence of probability measure (ρ_m) is tight. By Prokhorov's Theorem, there exists a subsequence (ρ_{m_k}) and a probability measure ρ^* such that ρ_{m_k} converges weakly to ρ^*.

The fact that $\mathrm{supp}(\rho_m) \subset K_Q$ implies $\mathrm{supp}(\rho^*) \subset K_Q$. Therefore, we have

$$\|f^*\|_{\mathcal{B}} = \|f^*\|_{\mathcal{B}_\infty} \le Q.$$

For any $\mathbf{x} \in X$, $a\sigma(\boldsymbol{b}^T\mathbf{x} + c)$ is continuous with respect to $(a, \boldsymbol{b}, c)$ and bounded from above by Q. Since ρ^* is the weak limit of ρ_{m_k}, we have

$$f^*(\mathbf{x}) = \lim_{k\to\infty} \int a\sigma(\boldsymbol{b}^T\mathbf{x} + c) d\rho_{m_k} = \int a\sigma(\boldsymbol{b}^T\mathbf{x} + c) d\rho^*(da, d\boldsymbol{b}, dc).$$

□

2.5.5 Proof of Theorem 3

Let $\boldsymbol{w} = (\boldsymbol{b}^T, c)^T$ and $\tilde{\mathbf{x}} = (\mathbf{x}^T, 1)^T$. For any $\varepsilon > 0$ and $f \in \mathcal{B}$, let $\rho_f^\varepsilon(a, \boldsymbol{w})$ be a distribution such that $f(\mathbf{x}) = \mathbb{E}_{\rho_f^\varepsilon}[a\sigma(\boldsymbol{b}^T\mathbf{x} + c)]$ and $\mathbb{E}_{\rho_f^\varepsilon}[|a|\|\boldsymbol{w}\|_1] < (1 + \varepsilon)\|f\|_{\mathcal{B}}$. Then,

$$\begin{aligned} n\,\mathrm{Rad}_n(\mathcal{F}_Q) &= \mathbb{E}_{\boldsymbol{\xi}}[\sup_{f\in\mathcal{F}_Q} \sum_{i=1}^n \xi_i \mathbb{E}_{\rho_f^\varepsilon}[a\sigma(\boldsymbol{w}^T\mathbf{x}_i)]] \\ &= \mathbb{E}_{\boldsymbol{\xi}}[\sup_{f\in\mathcal{F}_Q} \mathbb{E}_{\rho_f^\varepsilon}[\sum_{i=1}^n \xi_i a\sigma(\boldsymbol{w}^T\mathbf{x}_i)]] \\ &= \mathbb{E}_{\boldsymbol{\xi}}[\sup_{f\in\mathcal{F}_Q} \mathbb{E}_{\rho_f^\varepsilon}[|a|\|\boldsymbol{w}\|_1 |\sum_{i=1}^n \xi_i \sigma(\hat{\boldsymbol{w}}^T\mathbf{x}_i)|]] \\ &\le (1 + \varepsilon) Q \mathbb{E}_{\boldsymbol{\xi}}[\sup_{\|\boldsymbol{w}\|\le 1} |\sum_{i=1}^n \xi_i \sigma(\boldsymbol{w}^T\mathbf{x}_i)|]. \end{aligned} \tag{9}$$

Due to the symmetry, we have

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\xi}}[\sup_{\|\boldsymbol{w}\|\le 1} |\sum_{i=1}^n \xi_i \sigma(\boldsymbol{w}^T\mathbf{x}_i)|] &\le \mathbb{E}_{\boldsymbol{\xi}}[\sup_{\|\boldsymbol{w}\|\le 1} \sum_{i=1}^n \xi_i \sigma(\boldsymbol{w}^T\mathbf{x}_i)] + \mathbb{E}_{\boldsymbol{\xi}}[\sup_{\|\boldsymbol{w}\|\le 1} -\sum_{i=1}^n \xi_i \sigma(\boldsymbol{w}^T\mathbf{x}_i)] \\ &= 2\mathbb{E}_{\boldsymbol{\xi}}[\sup_{\|\boldsymbol{w}\|\le 1} \sum_{i=1}^n \xi_i \sigma(\boldsymbol{w}^T\mathbf{x}_i)] \end{aligned}$$

$$\leq 2\mathbb{E}_{\boldsymbol{\xi}}[\sup_{\|\boldsymbol{w}\|\leq 1}\sum_{i=1}^{n}\xi_i \boldsymbol{w}^T \mathbf{x}_i], \tag{10}$$

where the last inequality follows from the contraction property of Rademacher complexity (see Lemma 26.9 in [22]) and the fact that $\sigma(\cdot)$ is Lipschitz continuous with Lipschitz constant 1. Applying Lemma 26.11 in [22] and plugging (10) into (9), we obtain

$$\mathrm{Rad}_n(\mathcal{F}_Q) \leq 2(1+\varepsilon)Q\sqrt{\frac{2\ln(2d)}{n}}.$$

Taking $\varepsilon \to 0$, we complete the proof. □

3 Flow-Induced Function Spaces

In this section, we carry out a similar program for residual neural networks. Since the limit of these networks give rise to continuous in time flows, the natural function spaces and norms associated with the residual neural networks are also flow-based. For this reason, we call them flow-induced spaces and flow-induced norms, respectively. Similar to what was done in the last section, we establish a natural connection between these function spaces and residual neural networks, by proving direct and inverse approximation theorems. We also prove a complexity bound for the flow-induced space.

We postpone all the proofs to the end of this section.

3.1 The Compositional Law of Large Numbers

We consider residual neural networks defined by

$$\begin{aligned} z_{0,L}(\mathbf{x}) &= \mathbf{V}\mathbf{x}, \\ z_{l+1,L}(\mathbf{x}) &= z_{l,L}(\mathbf{x}) + \frac{1}{L}\mathbf{U}_l\sigma \circ (\boldsymbol{W}_l z_{l,L}(\mathbf{x})), \\ f_L(\mathbf{x};\Theta) &= \boldsymbol{\alpha}^T z_{L,L}(\mathbf{x}), \end{aligned} \tag{11}$$

where $\mathbf{x} \in \mathbb{R}^d$ is the input, $\mathbf{V} \in \mathbb{R}^{D\times d}$, $\boldsymbol{W}_l \in \mathbb{R}^{m\times D}$, $\mathbf{U}_l \in \mathbb{R}^{D\times m}$, $\boldsymbol{\alpha} \in \mathbb{R}^D$ and we use $\Theta := \{\mathbf{V}, \mathbf{U}_1, \ldots, \mathbf{U}_L, \boldsymbol{W}_l, \ldots, \boldsymbol{W}_L, \boldsymbol{\alpha}\}$ to denote all the parameters to be learned from data. Without loss of generality, we will fix $\mathbf{V}$ to be

$$\mathbf{V} = \begin{bmatrix} I_{d\times d} \\ 0_{(D-d)\times d} \end{bmatrix}. \tag{12}$$

We will fix D and m throughout this paper, and when there is no danger for confusion, we will omit Θ in the notation and use $f_L(\mathbf{x})$ to denote the residual network for simplicity.

For two-layer neural networks, if the parameters $\{a_k, \boldsymbol{b}_k, c_k\}$ are i.i.d sampled from a probability distribution ρ, then we have

$$\frac{1}{m}\sum_{k=1}^{m} a_k \sigma(\boldsymbol{b}_k^T \mathbf{x} + c_k) \to \int a\sigma(\boldsymbol{b}^T \mathbf{x} + c)\rho(da, d\boldsymbol{b}, dc),$$

when $m \to \infty$ as a consequence of the law of large numbers. To get some intuition in the current situation, we will first study a similar setting for residual networks in which $\mathbf{U}_l$ and $\boldsymbol{W}_l$ are i.i.d sampled from a probability distribution ρ on $\mathbb{R}^{D\times m} \times \mathbb{R}^{m\times D}$. To this end, we will study the behavior of $z_{L,L}(\cdot)$ as $L \to \infty$. The sequence of mappings we obtain is the repeated composition of many *i.i.d.* random near-identity maps.

The following theorem can be viewed as a compositional version of the law of large numbers. The "compositional mean" is defined with the help of the following ordinary differential equation (ODE) system:

$$\begin{aligned} z(\mathbf{x}, 0) &= \mathbf{V}\mathbf{x}, \\ \frac{d}{dt} z(\mathbf{x}, t) &= \mathbb{E}_{(\mathbf{U},\boldsymbol{W})\sim\rho} \mathbf{U}\sigma(\boldsymbol{W} z(\mathbf{x}, t)). \end{aligned} \tag{13}$$

Theorem 4 *Assume that σ is Lipschitz continuous and*

$$\mathbb{E}_\rho ||\mathbf{U}||\boldsymbol{W}||||_F^2 < \infty. \tag{14}$$

Then, the ODE (13) has a unique solution. For any $\mathbf{x} \in X$, we have

$$z_{L,L}(\mathbf{x}) \to z(\mathbf{x}, 1)$$

in probability as $L \to +\infty$. Moreover, we have

$$\lim_{L\to\infty} \sup_{\mathbf{x}\in X} \mathbb{E}\|z_{L,L}(\mathbf{x}) - z(\mathbf{x}, 1)\|^2 = 0,$$

i.e., the convergence is uniform with respect to $\mathbf{x} \in X$.

This result can be extended to situations when the distribution ρ is time-dependent, which is the right setting in applications.

Theorem 5 *Let $\{\rho_t,\ t \in [0, 1]\}$ be a family of probability distributions on $\mathbb{R}^{D\times m} \times \mathbb{R}^{m\times D}$ with the property that there exist constants c_1 and c_2 such that*

$$\mathbb{E}_{\rho_t} ||\mathbf{U}||\boldsymbol{W}||||_F^2 < c_1$$

and

$$\left|\mathbb{E}_{\rho_t} U\sigma(Wz) - \mathbb{E}_{\rho_s} U\sigma(Wz)\right| \le c_2|t - s||z|$$

for all $s, t \in [0, 1]$. *Let* z *be the solution of the following ODE,*

$$\begin{aligned} z(\mathbf{x}, 0) &= \mathbf{V}\mathbf{x}, \\ \frac{d}{dt} z(\mathbf{x}, t) &= \mathbb{E}_{(\mathbf{U},\boldsymbol{W})\sim\rho_t} \mathbf{U}\sigma(\boldsymbol{W} z(\mathbf{x}, t)). \end{aligned}$$

Then, for any fixed $\mathbf{x} \in \mathbf{X}$, *we have*

$$z_{L,L}(\mathbf{x}) \to z(\mathbf{x}, 1)$$

in probability as $L \to +\infty$. *Moreover, the convergence is uniform in* $\mathbf{x}$.

Similar results have been proved in the context of stochastic approximations, for example in [7,17].

3.2 The Flow-Induced Function Spaces

Motivated by the previous results, we consider the set of functions $f_{\boldsymbol{\alpha},\{\rho_t\}}$ defined by:

$$\begin{aligned} z(\mathbf{x}, 0) &= \mathbf{V}\mathbf{x}, \\ \dot{z}(\mathbf{x}, t) &= \mathbb{E}_{(\mathbf{U},\boldsymbol{W})\sim\rho_t} \mathbf{U}\sigma(\boldsymbol{W} z(\mathbf{x}, t)), \\ f_{\boldsymbol{\alpha},\{\rho_t\}}(\mathbf{x}) &= \boldsymbol{\alpha}^T z(\mathbf{x}, 1), \end{aligned} \tag{15}$$

where $\mathbf{V} \in \mathbb{R}^{D\times d}$ is given in (12), $\mathbf{U} \in \mathbb{R}^{D\times m}$, $\boldsymbol{W} \in \mathbb{R}^{m\times D}$, and $\boldsymbol{\alpha} \in \mathbb{R}^D$. To define a norm for these functions, we consider the following linear ODEs ($p \geq 1$)

$$\begin{aligned} N_p(0) &= \mathbf{e}, \\ \dot{N}_p(t) &= \left(\mathbb{E}_{\rho_t}(|\mathbf{U}||\boldsymbol{W}|)^p\right)^{1/p} N_p(t), \end{aligned} \tag{16}$$

where $\mathbf{e}$ is the all-one vector in $\mathbb{R}^D$. Note that in (16), $|\mathbf{A}|$ and $|\mathbf{A}|^q$ are defined element-wise for matrix $\mathbf{A}$, and the multiplication of $|\mathbf{U}|$ and $|\boldsymbol{W}|$ is the regular matrix multiplication. This linear system of equations has a unique solution as long as the expected value is integrable as a function of t. If f admits a representation as in (15), we can define the $\mathcal{D}_p$ norm of f.

Definition 2 Let f be a function that satisfies $f = f_{\boldsymbol{\alpha},\{\rho_t\}}$ for a pair of $(\boldsymbol{\alpha}, \{\rho_t\})$, then we define

$$\|f\|_{\mathcal{D}_p(\boldsymbol{\alpha},\{\rho_t\})} = |\boldsymbol{\alpha}|^T N_p(1),$$

to be the $\mathcal{D}_p$ norm of f with respect to the pair $(\boldsymbol{\alpha}, \{\rho_t\})$. We define

$$\|f\|_{\mathcal{D}_p} = \inf_{f=f_{\boldsymbol{\alpha},\{\rho_t\}}} |\boldsymbol{\alpha}|^T N_p(1). \tag{17}$$

to be the $\mathcal{D}_p$ norm of f.

As an example, if ρ is constant in t, then the $\mathcal{D}_p$ norm becomes

$$\|f\|_{\mathcal{D}_p} = \inf_{f=f_{\boldsymbol{\alpha},\rho}} |\boldsymbol{\alpha}|^T e^{(\mathbb{E}_\rho(|\mathbf{U}||\boldsymbol{W}|)^p)^{1/p}}\mathbf{e}.$$

Given this definition, the flow-induced function spaces on X are defined as the set of continuous functions that can be represented as $f_{\boldsymbol{\alpha},\{\rho_t\}}$ in (15) with finite $\mathcal{D}_p$ norm. Here we assume that for any $t \in [0,1]$, ρ_t is a probability distribution defined on (Ω, Σ_Ω), $\Omega = \mathbb{R}^{D\times m} \times \mathbb{R}^{m\times D}$, Σ_Ω is the Borel σ-algebra on Ω. We use $\mathcal{D}_p$ to denote these function spaces. It is easy to see $\mathcal{D}_p \subset \mathcal{D}_q$ for $p \geq q$.

Note that in the definitions above, the only condition on $\{\rho_t\}$ is the existence and uniqueness of z defined by (15). Hence, $\{\rho_t\}$ can be discontinuous as a function t. However, the compositional law of large numbers, which is the underlying reason behind the approximation theorem that we will discuss next (Theorem 5), requires $\{\rho_t\}$ to satisfy some continuity condition. To that end, we define the following "Lipschitz coefficient" and "Lipschitz norm" for $\{\rho_t\}$

Definition 3 Given a family of probability distribution $\{\rho_t,\ t \in [0,1]\}$, the "Lipschitz coefficient" of $\{\rho_t\}$, which is denoted by $Lip_{\{\rho_t\}}$, is defined as the infimum of all the number L that satisfies

$$\left|\mathbb{E}_{\rho_t}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}) - \mathbb{E}_{\rho_s}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z})\right| \leq Lip_{\{\rho_t\}}|t-s||\boldsymbol{z}|,$$

and

$$\left|\left\|\mathbb{E}_{\rho_t}|\mathbf{U}||\boldsymbol{W}|\right\|_{1,1} - \left\|\mathbb{E}_{\rho_s}|\mathbf{U}||\boldsymbol{W}|\right\|_{1,1}\right| \leq Lip_{\{\rho_t\}}|t-s|,$$

for any $t, s \in [0,1]$, where $\|\cdot\|_{1,1}$ is the sum of the absolute values of all the entries in a matrix. The "Lipschitz norm" of $\{\rho_t\}$ is defined as

$$\|\{\rho_t\}\|_{Lip} = \left\|\mathbb{E}_{\rho_0}|\mathbf{U}||\boldsymbol{W}|\right\|_{1,1} + Lip_{\{\rho_t\}}.$$

With the Lipschitz norm of $\{\rho_t\}$ defined above, we can introduce another class of function spaces $\tilde{\mathcal{D}}_p$, which independently controls $N_p(1)$ and $\|\{\rho_t\}\|_{Lip}$.

Definition 4 Let f be a function that satisfies $f = f_{\boldsymbol{\alpha},\{\rho_t\}}$ for a pair of $(\boldsymbol{\alpha}, \{\rho_t\})$, then we define

$$\|f\|_{\tilde{\mathcal{D}}_p(\boldsymbol{\alpha},\{\rho_t\})} = |\boldsymbol{\alpha}|^T N_p(1) + \|N_p(1)\|_1 - D + \|\{\rho_t\}\|_{Lip},$$

to be the $\tilde{\mathcal{D}}_p$ norm of f with respect to the pair $(\boldsymbol{\alpha}, \{\rho_t\})$. We define

$$\|f\|_{\tilde{\mathcal{D}}_p} = \inf_{f=f_{\boldsymbol{\alpha},\{\rho_t\}}} \|f\|_{\tilde{\mathcal{D}}_p(\boldsymbol{\alpha},\{\rho_t\})}.$$

to be the $\tilde{\mathcal{D}}_p$ norm of f. The space $\tilde{\mathcal{D}}_p$ is defined as the set of all the continuous functions that admit the representation $f_{\boldsymbol{\alpha},\{\rho_t\}}$ in (15) with finite $\tilde{\mathcal{D}}_p$ norm.

Remark 4 We add a "$-D$" term in the definition of $\tilde{\mathcal{D}}_p$ norm because $\|N_p(1)\|_1 \geq D$ and we want the norm of the zero function to be 0. As was stressed earlier, we use the terminology "norm" loosely, and we do not care whether these are really norms. Strictly speaking, they are just some quantities that can be used to bound approximation/estimation errors.

Next, for residual networks (11), we define a parameter-based norm as a discrete analog of (17). This is similar to the l_1 path norm of the residual network, which is studied in [10,20]

Definition 5 For a residual network defined by (11) with parameters $\Theta = \{\boldsymbol{\alpha}, \mathbf{U}_l, \boldsymbol{W}_l, l = 0, 1, \ldots, L-1\}$, we define the l_1 path norm of Θ to be

$$\|\Theta\|_P = |\boldsymbol{\alpha}|^T \prod_{l=1}^{L} \left(I + \frac{1}{L} |\mathbf{U}_l||\boldsymbol{W}_l| \right) \mathbf{e}.$$

We can also define the analog of the p-norms for $p > 1$ for residual networks. But in this paper, we will only use the l_1 norm defined above.

It is easy to see that $\tilde{\mathcal{D}}_p \subset \mathcal{D}_p$, and for any $f \in \tilde{\mathcal{D}}_p$ we have $\|f\|_{\mathcal{D}_p} \leq \|f\|_{\tilde{\mathcal{D}}_p}$. Moreover, for any $1 \leq q \leq p$, if $f \in \tilde{\mathcal{D}}_p$, then we have $f \in \tilde{\mathcal{D}}_q$ and $\|f\|_{\tilde{\mathcal{D}}_q} \leq \|f\|_{\tilde{\mathcal{D}}_p}$. The next proposition states that Barron space is embedded in $\tilde{\mathcal{D}}_1$.

Proposition 4 *Assume that $D \geq d+2$ and $m \geq 1$. For any function $f \in \mathcal{B}$, we have $f \in \tilde{\mathcal{D}}_1$, and*

$$\|f\|_{\tilde{\mathcal{D}}_1} \leq 2\|f\|_{\mathcal{B}} + 1.$$

Moreover, for any $\varepsilon > 0$, there exists $(\boldsymbol{\alpha}, \{\rho_t\})$ such that ρ_t is fixed for all t, $f = f_{\boldsymbol{\alpha}, \{\rho_t\}}$, and

$$\|f\|_{\tilde{\mathcal{D}}_1(\boldsymbol{\alpha}, \{\rho_t\})} \leq 2\|f\|_{\mathcal{B}} + 1 + \varepsilon.$$

Similar to the results of Proposition 4, we can prove that the composition of two Barron functions belongs to the flow-induced function space, and the norm is bounded by a polynomial of the norms of the two Barron functions.

Proposition 5 *Assume that $D \geq d+3$ and $m \geq 1$. Assume that $g : [0,1]^d \to [0,1] \in \mathcal{B}$, $h : [0,1] \to \mathbb{R}^1 \in \mathcal{B}_1$. Let $f = h \circ g$ be the composition of g and h. Then we have $f \in \mathcal{D}_1$ and*

$$\|f\|_{\mathcal{D}_1} \leq (\|h\|_{\mathcal{B}} + 1)(\|g\|_{\mathcal{B}} + 1).$$

In [13], the authors constructed a sequence of functions $\{f_d : \mathbb{R}^d \to \mathbb{R}\}$ whose spectral norms (4) grow exponentially with respect to d. They also showed that these functions can be expressed as the composition of two functions (one from $\mathbb{R}^d$ to $\mathbb{R}$ and the other from $\mathbb{R}$ to $\mathbb{R}$) whose spectral norms depend only polynomially on the

dimension d. By Proposition 5, the $\mathcal{D}_1$ norms of f_d are bounded by a polynomial of d. This shows that in the high dimensions, the flow-induced norm can be significantly smaller than the spectral norm. Combined with the direct approximation theorem below, this implies that residual networks can better approximate some functions than two-layer networks.

3.3 Direct and Inverse Approximation Theorems

We first prove the direct approximation theorem, which states that functions in $\tilde{\mathcal{D}}_2$ can be approximated by a sequence of residual networks with a $1/L^{1-\delta}$ error rate for any $\delta \in (0, 1)$, and the networks have uniformly bounded path norm.

Theorem 6 *Let $f \in \tilde{\mathcal{D}}_2$, $\delta \in (0, 1)$. Then, there exists an absolute constant C, such that for any*

$$L \geq C \left(m^4 D^6 \|f\|_{\tilde{\mathcal{D}}_2}^5 (\|f\|_{\tilde{\mathcal{D}}_2} + D)^2 \right)^{\frac{3}{\delta}},$$

there is an L-layer residual network $f_L(\cdot; \Theta)$ that satisfies

$$\|f - f_L(\cdot; \Theta)\|^2 \leq \frac{\|f\|_{\tilde{\mathcal{D}}_2}^2}{L^{1-\delta}},$$

and

$$\|\Theta\|_P \leq 9\|f\|_{\tilde{\mathcal{D}}_1}.$$

We can also prove an inverse approximation theorem, which states that any function that can be approximated by a sequence of well-behaved residual networks has to belong to the flow-induced space.

Theorem 7 *Let f be a function defined on X. Assume that there is a sequence of residual networks $\{f_L(\cdot; \Theta_L)\}_{L=1}^{\infty}$ such that $f_L(\mathbf{x}; \Theta) \to f(\mathbf{x})$ for every $\mathbf{x} \in X$ as $L \to \infty$. Assume further that the parameters in $\{f_L(\cdot; \Theta)\}_{L=1}^{\infty}$ are (entry-wise) bounded by c_0. Then, we have $f \in \mathcal{D}_\infty$, and*

$$\|f\|_{\mathcal{D}_\infty} \leq \frac{2e^{m(c_0^2+1)} D^2 c_0}{m}$$

Moreover, if for some constant c_1, $\|f_L\|_{\mathcal{D}_1} \leq c_1$ holds for all $L > 0$, then we have

$$\|f\|_{\mathcal{D}_1} \leq c_1$$

3.4 Bounds for the Rademacher Complexity

Our final result is an upper bound for the Rademacher complexity involving the flow-induced norm. Due to technical difficulties, in this part we consider a family of modified flow-induced function norms $\|\cdot\|_{\hat{\mathcal{D}}_p}$, which is defined as

$$\|f\|_{\hat{\mathcal{D}}_p} = \inf_{f=f_{\boldsymbol{\alpha},\{\rho_t\}}} |\boldsymbol{\alpha}|^T \hat{N}_p(1) + \|\hat{N}_p(1)\|_1 - D + \|\{\rho_t\}\|_{Lip},$$

where $\hat{N}_p(t)$ is given by

$$\begin{aligned} \hat{N}_p(0) &= 2\mathbf{e}, \\ \dot{\hat{N}}_p(t) &= 2\left(\mathbb{E}_{\rho_t}(|\mathbf{U}||\boldsymbol{W}|)^p\right)^{1/p} \hat{N}_p(t). \end{aligned}$$

Denote by $\hat{\mathcal{D}}_p$ the space of functions with finite $\hat{\mathcal{D}}_p$ norm. Then, we have

Theorem 8 *Let $\hat{\mathcal{D}}_p^Q = \{f \in \hat{\mathcal{D}}_p : \|f\|_{\hat{\mathcal{D}}_p} \le Q\}$, then we have*

$$\mathrm{Rad}_n(\hat{\mathcal{D}}_2^Q) \le 18Q\sqrt{\frac{2\log(2d)}{n}}.$$

The difference between the definitions of the spaces $\hat{\mathcal{D}}_p$ and $\mathcal{D}_p$ lies in the factor 2 that appears in $\hat{N}_p$. At this stage, we are not able to remove this factor. It should be noted that this factor of 2 is also present in the "weighted path norm" introduced in [10]. If $\mathbf{U}$, $\boldsymbol{W}$ and $\hat{N}_p(t)$ are scalars, then $\hat{N}_p(t)$ can be upper bounded by $(N_p(t))^2$. However, in the vectorial case this bound does not always hold. Hence, it is unclear how the two spaces $\hat{\mathcal{D}}_p$ and $\mathcal{D}_p$ are related. Clearly we can also develop an approximation theory for the space $\hat{\mathcal{D}}_p$, but we feel it is worthwhile to show that the space $\tilde{\mathcal{D}}_p$ is sufficient for that purpose. We also point out here at this stage, we allow to use different quantities (norms) to control the approximation and estimation errors.

3.5 Proofs

3.5.1 Proof of Theorem 4

To prove convergence, let $t_{l,L} = l/L$, and consider $\mathbf{e}_{l,L}(\mathbf{x}) = \sqrt{L}(z_{l,L}(\mathbf{x}) - z(\mathbf{x}, t_{l,L}))$. We will focus on fixed $\mathbf{x}$ and from now on we omit the dependence on $\mathbf{x}$ in the notations and write instead $\mathbf{e}_{l,L}$, $z_{l,L}$ and $z(t)$, for example. From the definition of $z(t)$, we have

$$\begin{aligned} z(t_{l+1,L}) &= z(t_{l,L}) + \int_{t_{l,L}}^{t_{l+1,L}} \mathbb{E}\mathbf{U}\sigma(\boldsymbol{W}z(t))dt \\ &= z(t_{l,L}) + \frac{1}{L}\mathbf{U}_l\sigma(\boldsymbol{W}_l z(t_{l,L})) + \frac{1}{L}\left(\mathbf{U}_l\sigma(\boldsymbol{W}_l z(t_{l,L})) - \mathbb{E}\mathbf{U}\sigma(\boldsymbol{W}z(t_{l,L}))\right) \end{aligned}$$

$$+\left(\frac{1}{L}\mathbb{E}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(t_{l,L}))-\int_{t_{l,L}}^{t_{l+1,L}}\mathbb{E}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(t))dt\right). \tag{18}$$

Since

$$\boldsymbol{z}_{l+1,L}=\boldsymbol{z}_{l,L}+\frac{1}{L}\mathbf{U}_l\sigma(\boldsymbol{W}_l\boldsymbol{z}_{l,L}), \tag{19}$$

subtract (18) from (19) gives

$$\begin{aligned}\mathbf{e}_{l+1,L}=\mathbf{e}_{l,L}&+\frac{1}{\sqrt{L}}\left(\mathbf{U}_l\sigma(\boldsymbol{W}_l\boldsymbol{z}_{l,L})-\mathbf{U}_l\sigma(\boldsymbol{W}_l\boldsymbol{z}(t_{l,L}))\right)\\&+\frac{1}{\sqrt{L}}\left(\mathbf{U}_l\sigma(\boldsymbol{W}_l\boldsymbol{z}(t_{l,L}))-\mathbb{E}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(t_{l,L}))\right)\\&+\frac{1}{\sqrt{L}}\left(\mathbb{E}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(t_{l,L}))-L\int_{t_{l,L}}^{t_{l+1,L}}\mathbb{E}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(t))dt\right).\end{aligned}$$

Define

$$\begin{aligned}I_{l,L}&=\frac{1}{\sqrt{L}}\left(\mathbf{U}_l\sigma(\boldsymbol{W}_l\boldsymbol{z}_{l,L})-\mathbf{U}_l\sigma(\boldsymbol{W}_l\boldsymbol{z}(t_{l,L}))\right),\\J_{l,L}&=\frac{1}{\sqrt{L}}\left(\mathbf{U}_l\sigma(\boldsymbol{W}_l\boldsymbol{z}(t_{l,L}))-\mathbb{E}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(t_{l,L}))\right),\\K_{l,L}&=\frac{1}{\sqrt{L}}\left(\mathbb{E}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(t_{l,L}))-L\int_{t_{l,L}}^{t_{l+1,L}}\mathbb{E}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(t))dt\right).\end{aligned}$$

Then, we have

$$\mathbf{e}_{l+1,L}=\mathbf{e}_{l,L}+I_{l,L}+J_{l,L}+K_{l,L}. \tag{20}$$

Next, we consider $\|\mathbf{e}_{l,L}\|^2$. From (20), we get

$$\begin{aligned}\|\mathbf{e}_{l+1,L}\|^2&=\|\mathbf{e}_{l,L}\|^2+\|I_{l,L}\|^2+\|J_{l,L}\|^2+\|K_{l,L}\|^2\\&\quad+2\mathbf{e}_{l,L}^TI_{l,L}+2\mathbf{e}_{l,L}^TJ_{l,L}+2\mathbf{e}_{l,L}^TK_{l,L}\\&\quad+2I_{l,L}^TJ_{l,L}+2I_{l,L}^TK_{l,L}+2J_{l,L}^TK_{l,L}\\&\le\|\mathbf{e}_{l,L}\|^2+3\|I_{l,L}\|^2+3\|J_{l,L}\|^2+3\|K_{l,L}\|^2\\&\quad+2\mathbf{e}_{l,L}^TI_{l,L}+2\mathbf{e}_{l,L}^TJ_{l,L}+2\mathbf{e}_{l,L}^TK_{l,L}.\end{aligned} \tag{21}$$

We are going to estimate the expectation of the right hand side of (21) term by term. First, note that $\mathbb{E}\||\mathbf{U}||\boldsymbol{W}|\|_F^2<\infty$, which means $\boldsymbol{z}(t)$ is bounded for $t\in[0,1]$. Hence, we can find a constant $C>0$ that satisfies

$$\mathbb{E}\||\mathbf{U}||\boldsymbol{W}|\|_F\le C,\quad \mathbb{E}\||\mathbf{U}||\boldsymbol{W}|\|_F^2\le C,$$

and $\|\boldsymbol{z}(t)\| \leq C$. In addition, note that for any l, $\mathbf{U}_l$ and $\boldsymbol{W}_l$ are independent with $\mathbf{e}_{l,L}$. Therefore, for $\|I_{l,L}\|^2$, we have

$$\begin{aligned}\mathbb{E}\|I_{l,L}\|^2 &= \frac{1}{L}\mathbb{E}\left\|\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}_{l,L}) - \mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(t_{l,L}))\right\|^2 \\ &\leq \frac{1}{L^2}\mathbb{E}\|\|\mathbf{U}_l\|\|\boldsymbol{W}_l\|\|\mathbf{e}_{l,L}\|\|^2 \\ &\leq \frac{C}{L^2}\mathbb{E}\|\mathbf{e}_{l,L}\|^2.\end{aligned}$$

For the term $\|J_{l,L}\|^2$, we have

$$\mathbb{E}\|J_{l,L}\|^2 \leq \frac{1}{L}\mathbb{E}\|\|\mathbf{U}_l\|\|\boldsymbol{W}_l\|\|\boldsymbol{z}(t_{l,L})\|\|^2 \leq \frac{C^2}{L}.$$

For the term $\|K_{l,L}\|$, since $\mathbb{E}\|\|\mathbf{U}\|\|\boldsymbol{W}\|\|_F \leq C$ and $\|\boldsymbol{z}\| \leq C$, we know that the Lipschitz constant of $\boldsymbol{z}(t)$ is bounded by C^2. Hence, we have

$$\begin{aligned}\|K_{l,L}\| &\leq \sqrt{L}\left\|\int_{t_{l,L}}^{t_{l+1,L}} \mathbb{E}(\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(t_{l,L})) - \mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(t)))dt\right\| \\ &\leq \frac{C^2}{\sqrt{L}}\int_{t_{l,L}}^{t_{l+1,L}} (t - t_{l,L})\mathbb{E}\|\|\mathbf{U}\|\|\boldsymbol{W}\|\|dt \\ &\leq \frac{C^3}{L\sqrt{L}},\end{aligned}$$

which implies that

$$\mathbb{E}\|K_{l,L}\|^2 \leq \frac{C^6}{L^3}.$$

Next, we consider $\mathbf{e}_{l,L}^T I_{l,L}$. We easily have

$$\mathbb{E}\mathbf{e}_{l,L}^T I_{l,L} \leq \frac{1}{L}\mathbb{E}\|\|\mathbf{U}\|\|\boldsymbol{W}\|\|_F\|\mathbf{e}_{l,L}\|^2 \leq \frac{C}{L}\mathbb{E}\|\mathbf{e}_{l,L}\|^2.$$

For $\mathbf{e}_{l,L}^T J_{l,L}$, by the independence of $\mathbf{U}_l$, $\boldsymbol{W}_l$ and $\mathbf{e}_{l,L}$, we have

$$\mathbb{E}\mathbf{e}_{l,L}^T J_{l,L} = 0.$$

Finally, for $\mathbf{e}_{l,L}^T K_{l,L}$, we have

$$\mathbb{E}\mathbf{e}_{l,L}^T K_{l,L} \leq \frac{C^3}{L\sqrt{L}}\sqrt{\mathbb{E}\|\mathbf{e}_{l,L}\|^2} \leq \frac{C^3}{L\sqrt{L}}(\mathbb{E}\|\mathbf{e}_{l,L}\|^2 + 1).$$

Plugging all the estimates above into (21), we obtain

$$\mathbb{E}\|\mathbf{e}_{l+1,L}\|^2 \le \left(1+\frac{2C}{L}+\frac{3C}{L^2}+\frac{2C^3}{L\sqrt{L}}\right)\mathbb{E}\|\mathbf{e}_{l,L}\|^2+\frac{3C^2}{L}+\frac{3C^6}{L^3}+\frac{2C^3}{L\sqrt{L}}.$$

Hence there is an L_0 depending only on C, such that if $L > L_0$, we have

$$\mathbb{E}\|\mathbf{e}_{l+1,L}\|^2 \le \left(1+\frac{3C}{L}\right)\mathbb{E}\|\mathbf{e}_{l,L}\|^2+\frac{4C^2}{L}.$$

Since $\mathbf{e}_{0,L}=0$, by induction we obtain

$$\mathbb{E}\|\mathbf{e}_{L,L}\|^2 \le 4C^2e^{3C},$$

which means

$$\mathbb{E}\|\boldsymbol{z}_{L,L}-\boldsymbol{z}(1)\|^2 \le \frac{4C^2e^{3C}}{L}\to 0,$$

when $L\to\infty$. This implies that $\boldsymbol{z}_{L,L}\to\boldsymbol{z}(1)$ in probability. □

3.5.2 Proof of Theorem 5

The only modification required for the proof of Theorem 5 is in the estimate of $K_{l,L}$. Now $K_{l,L}$ becomes

$$K_{l,L}=\frac{1}{\sqrt{L}}\left(\mathbb{E}_{\rho_{t_{l,L}}}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(t_{l,L}))-L\int_{t_{l,L}}^{t_{l+1,L}}\mathbb{E}_{\rho_t}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(t))dt\right).$$

The conditions of the theorem still guarantee that $\boldsymbol{z}(t)$ is Lipschtiz continuous. Hence, we can find a constant C' such that $\boldsymbol{z}(t)$ is C'-Lipschitz and

$$\mathbb{E}_{\rho_t}\||\mathbf{U}||\boldsymbol{W}|\|\le C',$$

for any $t\in[0,1]$. Hence,

$$\begin{aligned}\|K_{l,L}\| &\le \sqrt{L}\int_{t_{l,L}}^{t_{l+1,L}}\left\|\mathbb{E}_{\rho_{t_{l,L}}}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(t_{l,L}))-\mathbb{E}_{\rho_t}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(t))\right\|dt\\ &\le \sqrt{L}\int_{t_{l,L}}^{t_{l+1,L}}\left\|\mathbb{E}_{\rho_{t_{l,L}}}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(t_{l,L}))-\mathbb{E}_{\rho_t}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(t_{l,L}))\right\|dt\\ &\quad+\sqrt{L}\int_{t_{l,L}}^{t_{l+1,L}}\left\|\mathbb{E}_{\rho_t}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(t_{l,L}))-\mathbb{E}_{\rho_t}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(t))\right\|dt\\ &\le \frac{c_2C'}{L\sqrt{L}}+\frac{C'^2}{L\sqrt{L}}.\end{aligned}\tag{22}$$

From (22), we know that in this case $K_{l,L}$ is of the same order as that in Theorem 4. We can then complete the proof following the same arguments as in the proof of Theorem 4. □

3.5.3 Proof of Proposition 4

Since $f \in \mathcal{B}$, for any $\varepsilon > 0$, there exists a distribution ρ^{ε} that satisfies

$$f(\mathbf{x}) = \int_{\Omega} a\sigma(\boldsymbol{b}^T\mathbf{x} + c)\rho_{\varepsilon}(da, d\boldsymbol{b}, dc)$$
$$\mathbb{E}_{\rho_{\varepsilon}}[|a|(\|\boldsymbol{b}\|_1 + |c|)] \leq \|f\|_{\mathcal{B}} + \varepsilon.$$

Define $\hat{f}$ by

$$\begin{aligned} z(\mathbf{x}, 0) &= \begin{bmatrix} \mathbf{x} \\ 1 \\ 0 \end{bmatrix} \\ \frac{d}{dt} z(\mathbf{x}, t) &= \mathbb{E}_{(a,\boldsymbol{b},c)\sim\rho_{\varepsilon}} \begin{bmatrix} 0 \\ 0 \\ a \end{bmatrix} \sigma([\boldsymbol{b}^T, c, 0]z(\mathbf{x}, t)) \\ \hat{f}(\mathbf{x}) &= \mathbf{e}_{d+2}^T z(\mathbf{x}, 1) \end{aligned} \tag{23}$$

Then, we can easily verify that $\hat{f} = f$. Using ρ_{ε}, we can define probability distribution $\tilde{\rho}_{\varepsilon}$ on $\mathbb{R}^{D\times m} \times \mathbb{R}^{m\times D}$: $\tilde{\rho}_{\varepsilon}$ is concentrated on matrices of the form that appears in (23). Consider $\|f\|_{\tilde{\mathcal{D}}_1(\mathbf{e}_{d+2}, \{\tilde{\rho}_{\varepsilon}\})}$, we have

$$\begin{aligned} \|f\|_{\tilde{\mathcal{D}}_1(\mathbf{e}_{d+2}, \{\tilde{\rho}_{\varepsilon}\})} &= \mathbf{e}_{d+2}^T \exp\left(\mathbb{E}_{\rho}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ |a\boldsymbol{b}^T| & |ac| & 0 \end{bmatrix}\right)\mathbf{e} \\ &\quad + \left\|\exp\left(\mathbb{E}_{\rho}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ |a\boldsymbol{b}^T| & |ac| & 0 \end{bmatrix}\right)\right\|_1 - D \\ &= \mathbf{e}_{d+2}^T \begin{bmatrix} I & 0 & 0 \\ 0 & 1 & 0 \\ \mathbb{E}_{\rho}|a\boldsymbol{b}^T| & \mathbb{E}_{\rho}|ac| & 1 \end{bmatrix}\mathbf{e} + \left\|\begin{bmatrix} I & 0 & 0 \\ 0 & 1 & 0 \\ \mathbb{E}_{\rho}|a\boldsymbol{b}^T| & \mathbb{E}_{\rho}|ac| & 1 \end{bmatrix}\right\|_1 - D \\ &= 2\mathbb{E}_{\rho_{\varepsilon}}|a|(\|\boldsymbol{b}\|_1 + |c|) + 1 \\ &\leq 2\|f\|_{\mathcal{B}} + 2\varepsilon + 1. \end{aligned}$$

Therefore, we have

$$\|f\|_{\tilde{\mathcal{D}}_1} \leq \|f\|_{\tilde{\mathcal{D}}_1(\boldsymbol{\alpha}, \tilde{\rho}_{\varepsilon})} \leq 2\|f\|_{\mathcal{B}} + 2\varepsilon + 1.$$

Taking $\varepsilon \to 0$, we get

$$\|f\|_{\tilde{\mathcal{D}}_1} \leq 2\|f\|_{\mathcal{B}} + 1.$$

Besides, since $\{\tilde{\rho}_\varepsilon\}$ gives the same probability distribution for all $t \in [0, 1]$, we have $Lip_{\{\tilde{\rho}_\varepsilon\}} = 0$. □

3.5.4 Proof of Theorem 6

For any $\varepsilon > 0$, let

$$\sigma^\varepsilon(x) = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\varepsilon^2}} e^{-\frac{(x-y)^2}{2\varepsilon^2}} \sigma(y) dy.$$

Then we have

$$|\sigma^\varepsilon(x) - \sigma(x)| < \varepsilon, \quad |(\sigma^\varepsilon(x))'| \leq 1, \quad |(\sigma^\varepsilon(x))''| \leq \frac{1}{\varepsilon},$$

for all $x \in \mathbb{R}$. For a function $f \in \tilde{\mathcal{D}}_2$, we are going to show that for sufficiently large L there exists an L-layer residual network f_L such that

$$\|f - f_L\|^2 \leq \frac{\|f\|^2_{\tilde{\mathcal{D}}_2}}{L^{1-\delta}},$$

.

To do this, assume that $\boldsymbol{\alpha}$ and $\{\rho_t\}$ satisfy $f = f_{\boldsymbol{\alpha},\{\rho_t\}}$ and $\|f\|_{\tilde{\mathcal{D}}_2(\alpha,\{\rho_t\})} \leq 2\|f\|_{\tilde{\mathcal{D}}_1}$. Let f_L be a residual network in the form (11), and the weights $\boldsymbol{U}_l$, $\boldsymbol{W}_l$ are sampled from $\rho_{l/L}$. Let f^ε and f_L^ε be generated in the same way as f and f_L using instead the activation function σ^ε. Then we have

$$\|f - f_L\|^2 \leq 3\left(\|f - f^\varepsilon\|^2 + \|f^\varepsilon - f_L^\varepsilon\|^2 + \|f_L^\varepsilon - f_L\|^2\right). \tag{24}$$

Before dealing with (24), we first prove the following lemma, which shows that we can pick the family of distributions $\tilde{\rho}_t$ to have compact support.

Lemma 1 *For any $f \in \tilde{\mathcal{D}}_1$ that satisfies the conditions of Theorem 6, and any $\varepsilon > 0$, there exists $\boldsymbol{\alpha}$ and $\{\rho_t\}$, such that $f = f_{\boldsymbol{\alpha},\{\rho_t\}}$ and $\|f\|_{\tilde{\mathcal{D}}_2(\boldsymbol{\alpha},\{\rho_t\})} \leq (1+\varepsilon)\|f\|_{\tilde{\mathcal{D}}_2}$. Moreover, for any $t \in [0, 1]$, we have*

$$\max_{(\boldsymbol{U},\boldsymbol{W})\sim\rho_t} (\||\boldsymbol{U}||\boldsymbol{W}|\|_1) \leq (1+\varepsilon)\|f\|_{\tilde{\mathcal{D}}_2}.$$

Proof of Lemma 1 The proof of Lemma 1 is similar to the proof of Proposition 1. By the definition of $\tilde{\mathcal{D}}_2$, for any $f \in \tilde{\mathcal{D}}_2$ and $\varepsilon > 0$, there exists $\boldsymbol{\alpha}$ and $\{\rho_t\}$ such that

$f = f_{\boldsymbol{\alpha},\{\rho_t\}}$, $\|f\|_{\tilde{\mathcal{D}}_2(\boldsymbol{\alpha},\{\rho_t\})} \le (1+\varepsilon)\|f\|_{\tilde{\mathcal{D}}_2}$, and hence $\|\{\rho_t\}\|_{Lip} \le (1+\varepsilon)\|f\|_{\tilde{\mathcal{D}}_2}$. This means that for any $t \in [0,1]$, we have

$$\left\|\mathbb{E}_{\rho_t}|\mathbf{U}||\boldsymbol{W}|\right\|_1 \le (1+\varepsilon)\|f\|_{\tilde{\mathcal{D}}_2}.$$

Let $\Lambda = \{(\mathbf{U},\boldsymbol{W}):\ \|\boldsymbol{W}\|_1 = 1,\ \||\mathbf{U}||\boldsymbol{W}|\|_1 = 1\}$, and consider a family of measures $\{\rho_t^\Lambda\}$ defined by

$$\rho_t^\Lambda(A) = \int_{(\mathbf{U},\boldsymbol{W}):\ (\tilde{\mathbf{U}},\tilde{\boldsymbol{W}})\in\Lambda} \||\mathbf{U}||\boldsymbol{W}|\|_1 \rho_t(d\mathbf{U}, d\boldsymbol{W}),$$

for any Borel set $A \subset \Lambda$, where

$$\tilde{\mathbf{U}} = \frac{\|\boldsymbol{W}\|_1}{\||\mathbf{U}||\boldsymbol{W}|\|_1}\mathbf{U},\quad \tilde{\boldsymbol{W}} = \frac{\boldsymbol{W}}{\|\boldsymbol{W}\|_1}.$$

It is easy to verify that $\rho_t^\Lambda(\Lambda) = \mathbb{E}_{\rho_t}\||\mathbf{U}||\boldsymbol{W}|\|_1$ and

$$\mathbb{E}_{\rho_t^\Lambda}\tilde{\mathbf{U}}\sigma(\tilde{\boldsymbol{W}}\boldsymbol{z}) = \mathbb{E}_{\rho_t}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z})$$

hold for any $t \in [0,1]$ and $\boldsymbol{z} \in \mathbb{R}^D$. After normalizing $\{\rho_t^\Lambda\}$, we obtain a family of probability distributions $\{\tilde{\rho}_t^\Lambda\}$ on

$$\{(\mathbf{U},\boldsymbol{W}):\ \|\boldsymbol{W}\|_1 = 1,\ \||\mathbf{U}||\boldsymbol{W}|\| = \mathbb{E}_{\rho_t}\||\mathbf{U}||\boldsymbol{W}|\|_1\}.$$

Finally, it is easy to verify that $f = f_{\boldsymbol{\alpha},\{\tilde{\rho}_t^\Lambda\}}$, $\|f\|_{\tilde{\mathcal{D}}_2(\boldsymbol{\alpha},\{\rho_t\})} \le (1+\varepsilon)\|f\|_{\tilde{\mathcal{D}}_2}$, as well as

$$\max_{(\mathbf{U},\boldsymbol{W})\sim\tilde{\rho}_t^\Lambda}(\||\mathbf{U}||\boldsymbol{W}|\|_1) \le (1+\varepsilon)\|f\|_{\tilde{\mathcal{D}}_2}.$$

□

From Lemma 1, without loss of generality we can assume that ρ_t has compact support, and the entries of $(\mathbf{U},\boldsymbol{W})$ sampled from ρ_t for any t are bounded by $2\|f\|_{\tilde{\mathcal{D}}_2}$. Next we proceed to control the three terms on the right-hand side of (24). The following two lemmas give the bounds for the first and third terms.

Lemma 2 $\|f - f^\varepsilon\|^2 \le 4m^2\varepsilon^2\|f\|^4_{\tilde{\mathcal{D}}_2}$.

Proof of Lemma 2 Let $\boldsymbol{z}(t)$ be defined by (15) for fixed $\mathbf{x}$, and $\boldsymbol{z}^\varepsilon(t)$ be the solution of the same ODE after replacing σ by σ^ε. Then, we have $\boldsymbol{z}(0) - \boldsymbol{z}^\varepsilon(0) = 0$, and

$$\begin{aligned}|\boldsymbol{z}(t) - \boldsymbol{z}^\varepsilon(t)| &\le \int_0^t \left|\frac{d}{dt}(\boldsymbol{z}(s) - \boldsymbol{z}^\varepsilon(s))\right| ds\\ &= \int_0^t \left|\mathbb{E}_{\rho_s}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(s)) - \mathbb{E}_{\rho_s}\mathbf{U}\sigma^\varepsilon(\boldsymbol{W}\boldsymbol{z}^\varepsilon(s))\right| ds\end{aligned}$$

$$\begin{aligned}
&\leq \int_0^t \left|\mathbb{E}_{\rho_s}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}(s)) - \mathbb{E}_{\rho_s}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}^{\varepsilon}(s))\right| ds \\
&\quad + \int_0^t \left|\mathbb{E}_{\rho_s}\mathbf{U}\sigma(\boldsymbol{W}\boldsymbol{z}^{\varepsilon}(s)) - \mathbb{E}_{\rho_s}\mathbf{U}\sigma^{\varepsilon}(\boldsymbol{W}\boldsymbol{z}^{\varepsilon}(s))\right| ds \\
&\leq \int_0^t \left(\mathbb{E}_{\rho_s}|\mathbf{U}||\boldsymbol{W}||\boldsymbol{z}(s) - \boldsymbol{z}^{\varepsilon}(s)| + 2\|f\|_{\tilde{\mathcal{D}}_2} m\varepsilon\right) ds.
\end{aligned}$$

Hence, we have

$$|\boldsymbol{z}(1) - \boldsymbol{z}^{\varepsilon}(1)| \leq 2\|f\|_{\tilde{\mathcal{D}}_2} m\varepsilon N_1(1) e,$$

where e is an all-one vector. This gives that

$$\|f - f^{\varepsilon}\|^2 \leq \int_{D_0} \left(|\alpha|^T |\boldsymbol{z}(\mathbf{x}, 1) - \boldsymbol{z}^{\varepsilon}(\mathbf{x}, 1)|\right)^2 d\rho(\mathbf{x}) \leq 4m^2\varepsilon^2\|f\|^4_{\tilde{\mathcal{D}}_2}.$$

□

Lemma 3

$$\mathbb{E}\|f_L - f_L^{\varepsilon}\|^2 \leq 4m^2\varepsilon^2\|f\|^4_{\tilde{\mathcal{D}}_2},$$

where the expectation is taken over the random choice of weights $\{(\mathbf{U}_l, \boldsymbol{W}_l)\}$.

Proof of Lemma 3 Let $\boldsymbol{z}_{l,L}$ be defined by (11) for a fixed $\mathbf{x}$, and $\boldsymbol{z}^{\varepsilon}_{l,L}$ be defined similarly with σ replaced by σ^{ε}. Then, we have $\boldsymbol{z}_{0,L} - \boldsymbol{z}^{\varepsilon}_{0,L} = 0$, and

$$\begin{aligned}
\boldsymbol{z}_{l+1,L} - \boldsymbol{z}^{\varepsilon}_{l+1,L} &= \boldsymbol{z}_{l,L} - \boldsymbol{z}^{\varepsilon}_{l,L} + \frac{1}{L}\left[\mathbf{U}_l\sigma(\boldsymbol{W}_l\boldsymbol{z}_{l,L}) - \mathbf{U}_l\sigma(\boldsymbol{W}_l\boldsymbol{z}^{\varepsilon}_{l,L})\right] \\
&\quad + \frac{1}{L}\left[\mathbf{U}_l\sigma(\boldsymbol{W}_l\boldsymbol{z}^{\varepsilon}_{l,L}) - \mathbf{U}_l\sigma^{\varepsilon}(\boldsymbol{W}_l\boldsymbol{z}^{\varepsilon}_{l,L})\right].
\end{aligned}$$

Taking absolute value gives

$$|\boldsymbol{z}_{l+1,L} - \boldsymbol{z}^{\varepsilon}_{l+1,L}| \leq \left(I + \frac{1}{L}|\mathbf{U}||\boldsymbol{W}|\right)|\boldsymbol{z}_{l,L} - \boldsymbol{z}^{\varepsilon}_{l,L}| + \frac{2\|f\|_{\tilde{\mathcal{D}}_2} m\varepsilon}{L} e,$$

which implies that

$$|\boldsymbol{z}_{L,L} - \boldsymbol{z}^{\varepsilon}_{L,L}| \leq 2\|f\|_{\tilde{\mathcal{D}}_2} m\varepsilon \prod_{l=0}^{L-1}\left(I + \frac{1}{L}|\mathbf{U}||\boldsymbol{W}|\right) e.$$

By Theorem 5, we have

$$\mathbb{E}|f_L(\mathbf{x}) - f_L^{\varepsilon}(\mathbf{x})|^2 \leq 4\|f\|^2_{\tilde{\mathcal{D}}_2} m^2\varepsilon^2\mathbb{E}\left(|\alpha|^T \prod_{l=0}^{L-1}\left(I + \frac{1}{L}|\mathbf{U}||\boldsymbol{W}|\right) e\right)^2$$

$$\leq 4m^2\varepsilon^2\|f\|^4_{\tilde{\mathcal{D}}_2}. \tag{25}$$

Integrating (25) over $\mathbf{x}$ gives the results. □

Proof of Theorem 6 (Continued) With Lemmas 2 and 3, we have

$$\mathbb{E}\|f - f_L\|^2 \leq 24m^2\varepsilon^2\|f\|^4_{\tilde{\mathcal{D}}_2} + 3\mathbb{E}\|f^\varepsilon - f_L^\varepsilon\|^2. \tag{26}$$

To bound $\mathbb{E}\|f^\varepsilon - f_L^\varepsilon\|^2$, let $\mathbf{e}_{l,L} = \sqrt{L}(z^\varepsilon_{l,L} - z^\varepsilon_{t_{l,L}})$, and recall that we can write

$$\mathbf{e}_{l+1,L} = \mathbf{e}_{l,L} + I_{l,L} + J_{l,L} + K_{l,L}, \tag{27}$$

with

$$\begin{aligned}
I_{l,L} &= \frac{1}{\sqrt{L}}\left[\mathbf{U}_l\sigma^\varepsilon(\boldsymbol{W}_l z^\varepsilon_{l,L}) - \mathbf{U}_l\sigma^\varepsilon(\boldsymbol{W}_l z^\varepsilon(t_{l,L}))\right],\\
J_{l,L} &= \frac{1}{\sqrt{L}}\left[\mathbf{U}_l\sigma^\varepsilon(\boldsymbol{W}_l z^\varepsilon(t_{l,L})) - \mathbb{E}_{\rho_{t_{l,L}}}\mathbf{U}\sigma^\varepsilon(\boldsymbol{W} z^\varepsilon(t_{l,L}))\right]\\
K_{l,L} &= \frac{1}{\sqrt{L}}\left[\mathbb{E}_{\rho_{t_{l,L}}}\mathbf{U}\sigma^\varepsilon(\boldsymbol{W} z^\varepsilon(t_{l,L})) - L\int_{t_{l,L}}^{t_{l+1,L}}\mathbb{E}_{\rho_t}\mathbf{U}\sigma^\varepsilon(\boldsymbol{W} z^\varepsilon(t))dt\right].
\end{aligned}$$

For $I_{l,L}$, by the Taylor expansion of $\mathbf{U}_l\sigma^\varepsilon(\boldsymbol{W}_l z^\varepsilon_{l,L})$ at $z^\varepsilon(t_{l,L})$, we get

$$\begin{aligned}
I_{l,L} = &\frac{1}{L}\mathbf{U}_l(\sigma^\varepsilon(\boldsymbol{W}_l z^\varepsilon(t_{l,L})))'\boldsymbol{W}_l\mathbf{e}_{l,L}\\
&+\frac{\mathbf{U}_l(\sigma^\varepsilon(\boldsymbol{W}_l z^\varepsilon(t_{l,L})))''(\boldsymbol{W}_l\mathbf{e}_{l,L})\circ(\boldsymbol{W}_l\mathbf{e}_{l,L})}{L\sqrt{L}},
\end{aligned} \tag{28}$$

where for two vectors α and β, $\alpha\circ\beta$ means element-wise product. For the second term on the right-hand side of (28), we have

$$\left|\mathbf{U}_l(\sigma^\varepsilon(\boldsymbol{W}_l z^\varepsilon(t_{l,L})))''(\boldsymbol{W}_l\mathbf{e}_{l,L})\circ(\boldsymbol{W}_l\mathbf{e}_{l,L})\right| \leq \frac{8\|f\|^3_{\tilde{\mathcal{D}}_2} mD\|\mathbf{e}_{l,L}\|^2}{\varepsilon}e.$$

On the other hand, for $K_{l,L}$ we have

$$|K_{l,L}| \leq \frac{C_2}{L\sqrt{L}}e,$$

for some constant C_2. Hence, we can write (27) as

$$\mathbf{e}_{l+1,L} = \mathbf{e}_{l,L} + \frac{1}{L}\mathbf{U}_l(\sigma^\varepsilon(\boldsymbol{W}_l z^\varepsilon(t_{l,L})))'\boldsymbol{W}_l\mathbf{e}_{l,L} + J_{l,L} + \frac{r_{l,L}}{L\sqrt{L}}, \tag{29}$$

with

$$|r_{l,L}| \le (8\|f\|_{\tilde{\mathcal{D}}_2}^3 mD\varepsilon^{-1}\|\mathbf{e}_{l,L}\|^2 + C_2)e.$$

Next, we consider $\mathbf{e}_{l,L}\mathbf{e}_{l,L}^T$. By (29), we have

$$\begin{aligned}
\mathbf{e}_{l+1,L}\mathbf{e}_{l+1,L}^T = \mathbf{e}_{l,L}\mathbf{e}_{l,L}^T &+ \frac{1}{L}\Big(\mathbf{U}_l(\sigma^\varepsilon(\boldsymbol{W}_l\boldsymbol{z}^\varepsilon(t_{l,L})))'\boldsymbol{W}_l\mathbf{e}_{l,L}\mathbf{e}_{l,L}^T \\
&+\mathbf{e}_{l,L}\mathbf{e}_{l,L}^T\mathbf{U}_l(\sigma^\varepsilon(\boldsymbol{W}_l\boldsymbol{z}^\varepsilon(t_{l,L})))'\boldsymbol{W}_l\Big) \\
&+J_{l,L}J_{l,L}^T + \frac{1}{L^2}\mathbf{U}_l(\sigma^\varepsilon(\boldsymbol{W}_l\boldsymbol{z}^\varepsilon(t_{l,L})))'\boldsymbol{W}_l\mathbf{e}_{l,L}(\mathbf{U}_l(\sigma^\varepsilon(\boldsymbol{W}_l\boldsymbol{z}^\varepsilon(t_{l,L})))'\boldsymbol{W}_l\mathbf{e}_{l,L})^T \\
&+\mathbf{e}_{l,L}J_{l,L}^T + J_{l,L}\mathbf{e}_{l,L}^T + \frac{1}{L\sqrt{L}}\Big(\mathbf{e}_{l,L}r_{l,L}^T + r_{l,L}\mathbf{e}_{l,L}\Big) + \frac{1}{L^3}r_{l,L}r_{l,L}^T \\
&+\frac{1}{L}\mathbf{U}_l(\sigma^\varepsilon(\boldsymbol{W}_l\boldsymbol{z}^\varepsilon(t_{l,L})))'\boldsymbol{W}_l\mathbf{e}_{l,L}J_{l,L}^T + \frac{1}{L}J_{l,L}(\mathbf{U}_l(\sigma^\varepsilon(\boldsymbol{W}_l\boldsymbol{z}^\varepsilon(t_{l,L})))'\boldsymbol{W}_l\mathbf{e}_{l,L})^T \\
&+\frac{1}{L^2\sqrt{L}}\mathbf{U}_l(\sigma^\varepsilon(\boldsymbol{W}_l\boldsymbol{z}^\varepsilon(t_{l,L})))'\boldsymbol{W}_l\mathbf{e}_{l,L}r_{l,L}^T \\
&+\frac{1}{L^2\sqrt{L}}r^{l,L}(\mathbf{U}_l(\sigma^\varepsilon(\boldsymbol{W}_l\boldsymbol{z}^\varepsilon(t_{l,L})))'\boldsymbol{W}_l\mathbf{e}_{l,L})^T \\
&+\frac{1}{L\sqrt{L}}\Big(J_{l,L}r_{l,L}^T + r_{l,L}J_{l,L}^T\Big).
\end{aligned}$$

Taking expectation over the equation above, noting that $J_{l,L}$ is independent with $\mathbf{e}_{l,L}$, and using the bound of $r_{l,L}$ we derived above, we get

$$\begin{aligned}
|\mathbb{E}\mathbf{e}_{l+1,L}\mathbf{e}_{l+1,L}^T| \le |\mathbb{E}\mathbf{e}_{l,L}\mathbf{e}_{l,L}^T| + \frac{1}{L}\Big(A_{l,L}|\mathbb{E}\mathbf{e}_{l,L}\mathbf{e}_{l,L}^T| + |\mathbb{E}\mathbf{e}_{l,L}\mathbf{e}_{l,L}^T|A_{l,L}^T\Big) + \frac{1}{L}\Sigma_{l,L} \\
+\frac{C\|f\|_{\tilde{\mathcal{D}}_2}^3}{L}\left(\frac{mD\mathbb{E}\|\mathbf{e}_{l,L}\|^3}{\sqrt{L}\varepsilon} + \frac{m^2D^2\mathbb{E}\|\mathbf{e}_{l,L}\|^4}{L^2\varepsilon^2}\right)E,
\end{aligned} \tag{30}$$

where

$$A_{l,L} = \mathbb{E}_{\rho_{t_{l,L}}}|\mathbf{U}||\boldsymbol{W}|, \quad \Sigma_{l,L} = \Big|\mathrm{Cov}_{\rho_{t_{l,L}}}\mathbf{U}_l\sigma^\varepsilon(\boldsymbol{W}_l\boldsymbol{z}^\varepsilon(t_{l,L}))\Big|,$$

E is an all-one matrix and C is a constant.

Next, we bound the third and fourth order moments of $\|\mathbf{e}_{l,L}\|$ using its second order moment. This is done by the following lemma.

Lemma 4 *For any L and $1 \le l \le L$, there exists a constant C such that*

$$\mathbb{E}\|\mathbf{e}_{l,L}\|^3 \le CmD^{3/2}\|f\|_{\tilde{\mathcal{D}}_2}\left(\sqrt{\log L} + \frac{D}{\sqrt{L}\varepsilon}\right)\mathbb{E}\|\mathbf{e}_{l,L}\|^2,$$

and

$$\mathbb{E}\|\mathbf{e}_{l,L}\|^4 \le C^2 m^2 D^3 \|f\|^2_{\tilde{\mathcal{D}}_2} \left(\sqrt{\log L} + \frac{D}{\sqrt{L}\varepsilon}\right)^2 \mathbb{E}\|\mathbf{e}_{l,L}\|^2.$$

Proof of Lemma 4 Let $S_{l,L} = \sum_{k=0}^{l-1} J_{k,L}$. Then, $\mathbb{E}S_{l,L} = 0$. Since $J_{l,L}$ are independent for different l, and

$$|J_{l,L}| \le \frac{C' m D}{\sqrt{L}} e$$

holds for all l and some constant C', by Hoeffding's inequality, for any $t > 0$ and $1 \le i \le D$, we have

$$\mathbb{P}(|S_{l,L,i}| \ge t) \le 2\exp(-\frac{t^2}{2C'^2 m^2 D^2}).$$

Here, $S_{l,L,i}$ denotes the i-th entry of the vector $S_{l,L}$. Taking $t = 2C' m D\sqrt{\log L}$, we obtain

$$\mathbb{P}\left(|S_{l,L,i}| \ge 2C' m D\sqrt{\log L}\right) \le \frac{2}{L^2}.$$

This implies

$$\begin{aligned}
\mathbb{P}\left(\|S_{l,L}\| \ge 2C' m D^{3/2}\sqrt{\log L}\right) &= 1 - \mathbb{P}\left(\|S_{l,L}\| < 2C' m D^{3/2}\sqrt{\log L}\right) \\
&\le 1 - \mathbb{P}\left(\bigcup_i \left\{|S_{l,L,i}| < 2C' m D\sqrt{\log L}\right\}\right) \\
&\le 1 - \left(1 - \frac{2}{L^2}\right)^D \\
&\le \frac{2D}{L^2}. \qquad (31)
\end{aligned}$$

Define the event $\mathcal{A}$ by

$$\mathcal{A} = \left\{\|S_{l,L}\| \le 2C' m D^{3/2}\sqrt{\log L},\ i = 1, 2, \ldots, L\right\}.$$

Then by (31) we have

$$\mathbb{P}(\mathcal{A}) \ge 1 - \frac{2D}{L}.$$

Using (29), we have

$$\mathbf{e}_{l,L} = \sum_{k=0}^{l-1} \frac{1}{L} \mathbf{U}_k (\sigma^{\varepsilon}(\boldsymbol{W}_k \boldsymbol{z}^{\varepsilon}(t_{k,L})))' \boldsymbol{W}_k \mathbf{e}_{k,L} + S_{l,L} + \sum_{k=0}^{l-1} \frac{r_{k,L}}{L\sqrt{L}}.$$

Hence, using the bounds of $S_{l,L}$ and $r_{k,L}$, we obtain that there is a constant C such that

$$\|\mathbf{e}_{l,L}\| \le CmD^{3/2} \|f\|_{\tilde{\mathcal{D}}_2} \left(\sqrt{L} + \frac{1}{\sqrt{L}\varepsilon} \right).$$

On the other hand, under event $\mathcal{A}$, using the sharper bound of $S_{l,L}$, we have

$$\|\mathbf{e}_{l,L}\| \le CmD^{3/2} \|f\|_{\tilde{\mathcal{D}}_2} \left(\sqrt{\log L} + \frac{1}{\sqrt{L}\varepsilon} \right).$$

For third-order moment of $\|\mathbf{e}_{l,L}\|$, we have

$$\begin{aligned} \mathbb{E}\|\mathbf{e}_{l,L}\|^3 &\le CmD^{3/2} \|f\|_{\tilde{\mathcal{D}}_2} \left(\left(\sqrt{\log L} + \frac{1}{\sqrt{L}\varepsilon} \right) \mathbb{P}(\mathcal{A}) \right. \\ &\quad \left. + \left(\sqrt{L} + \frac{1}{\sqrt{L}\varepsilon} \right) \mathbb{P}(\mathcal{A}^c) \right) \mathbb{E}\|\mathbf{e}_{l,L}\|^2 \\ &\le CmD^{3/2} \|f\|_{\tilde{\mathcal{D}}_2} \left(\sqrt{\log L} + \frac{D}{\sqrt{L}\varepsilon} \right) \mathbb{E}\|\mathbf{e}_{l,L}\|^2. \end{aligned}$$

Similarly, for fourth-order moment we have

$$\mathbb{E}\|\mathbf{e}_{l,L}\|^4 \le C^2 m^2 D^3 \|f\|^2_{\tilde{\mathcal{D}}_2} \left(\sqrt{\log L} + \frac{D}{\sqrt{L}\varepsilon} \right)^2 \mathbb{E}\|\mathbf{e}_{l,L}\|^2.$$

□

Proof of Theorem 6 (Continued) Applying the results of Lemma 4 to (30) gives

$$\begin{aligned} |\mathbb{E}\mathbf{e}_{l+1,L}\mathbf{e}_{l+1,L}^T| &\le |\mathbb{E}\mathbf{e}_{l,L}\mathbf{e}_{l,L}^T| + \frac{1}{L}\left(A_{l,L}|\mathbb{E}\mathbf{e}_{l,L}\mathbf{e}_{l,L}^T| + |\mathbb{E}\mathbf{e}_{l,L}\mathbf{e}_{l,L}^T|A_{l,L}^T \right) + \frac{1}{L}\Sigma_{l,L} \\ &\quad + \frac{C}{L}\left(m^4 D^5 \|f\|^5_{\tilde{\mathcal{D}}_2} \left(\frac{\sqrt{\log L}}{\sqrt{L}\varepsilon} + \frac{D}{L\varepsilon^2} \right) \mathbb{E}\|\mathbf{e}_{l,L}\|^2 \right) E. \end{aligned} \tag{32}$$

Since $\|f\|_{\tilde{\mathcal{D}}_2} < \infty$, $\Sigma_{l,L}$ is uniformly bounded. Without loss of generality, we can assume $\Sigma_{l,L} \le CE$. Furthermore, assume L is sufficiently large such that

$$\frac{m^4 D^6 \|f\|^5_{\tilde{\mathcal{D}}_2} \mathbb{E}\|\mathbf{e}_{l,L}\|^2}{L^{\delta/3}} \le 1. \tag{33}$$

Then, from (32) we have

$$|\mathbb{E}\mathbf{e}_{l+1,L}\mathbf{e}_{l+1,L}^T| \le |\mathbb{E}\mathbf{e}_{l,L}\mathbf{e}_{l,L}^T| + \frac{1}{L}\left(A_{l,L}|\mathbb{E}\mathbf{e}_{l,L}\mathbf{e}_{l,L}^T| + |\mathbb{E}\mathbf{e}_{l,L}\mathbf{e}_{l,L}^T|A_{l,L}^T\right) + \frac{C}{L}\left(1 + \frac{\sqrt{\log L}}{L^{1/2-\delta/3}\varepsilon} + \frac{D}{L^{1-\delta/3}\varepsilon^2}\right)E,$$

which implies that

$$|\mathbb{E}\mathbf{e}_{l+1,L}\mathbf{e}_{l+1,L}^T| \le C\left(1 + \frac{\sqrt{\log L}}{L^{1/2-\delta/3}\varepsilon} + \frac{D}{L^{1-\delta/3}\varepsilon^2}\right)N_1(1)N_1(1)^T, \tag{34}$$

and thus

$$\mathbb{E}\|\mathbf{e}_{l,L}\|^2 \le e^T|\mathbb{E}\mathbf{e}_{l+1,L}\mathbf{e}_{l+1,L}^T|e \le C\left(1 + \frac{\sqrt{\log L}}{L^{1/2-\delta/3}\varepsilon} + \frac{D}{L^{1-\delta/3}\varepsilon^2}\right)(e^T N_1(1))^2. \tag{35}$$

Note that $e^T N_1(1) = \|N_1(1)\|_1 \le \|f\|_{\tilde{\mathcal{D}}_2} + D$. By (33) and (35), (34) happens if

$$Cm^4D^6\|f\|_{\tilde{\mathcal{D}}_2}^5\left(1 + \frac{\sqrt{\log L}}{L^{1/2-\delta/3}\varepsilon} + \frac{D}{L^{1-\delta/3}\varepsilon^2}\right)(\|f\|_{\tilde{\mathcal{D}}_2} + D)^2 \le L^{\delta/3}.$$

Taking $\varepsilon = L^{-1/2+\delta/3}$, it suffices to have

$$Cm^4D^6\|f\|_{\tilde{\mathcal{D}}_2}^5\left(1 + D + \sqrt{\log L}\right)(\|f\|_{\tilde{\mathcal{D}}_2} + D)^2 \le L^{\delta/3}.$$

In this case, we have

$$\mathbb{E}\|f^\varepsilon - f_L^\varepsilon\|^2 \le \frac{C}{L}\left(1 + D + \sqrt{\log L}\right)\|f\|_{\tilde{\mathcal{D}}_2}^2.$$

Plugging into (26) gives

$$\mathbb{E}\|f - f_L\|^2 \le \frac{24m^2}{L^{1-2\delta/3}}\|f\|_{\tilde{\mathcal{D}}_2}^4 + \frac{3C}{L}\left(1 + D + \sqrt{\log L}\right)\|f\|_{\tilde{\mathcal{D}}_2}^2.$$

When L sufficiently large (larger than polynomial of m, D, $\log L$), we have

$$\mathbb{E}\|f - f_L\|^2 \le \frac{\|f\|_{\tilde{\mathcal{D}}_2}^2}{3L^{1-\delta}}.$$

Note that the bound above holds for any fixed $\mathbf{x} \in \mathbf{X}$. Now, integrating over $\mathbf{x}$, we have

$$\mathbb{E}\|f - f_L\|^2 = \int \mathbb{E}|f(\mathbf{x}) - f_L(\mathbf{x})|^2 d\mu(\mathbf{x}) \le \frac{\|f\|_{\tilde{\mathcal{D}}_2}^2}{3L^{1-\delta}}.$$

By Markov's inequality, with probability no less than $\frac{2}{3}$, the distance between f and f_L can be controlled by

$$\|f - f_L\|^2 \leq \frac{\|f\|^2_{\tilde{\mathcal{D}}_2}}{L^{1-\delta}}. \tag{36}$$

Next, consider the path norm of f_L, which is defined as

$$\|f_L\|_P = \left\| |\alpha| \prod_{l=1}^{L} \left(I + \frac{1}{L} |\mathbf{U}_l||\boldsymbol{W}_l| \right) |\mathbf{V}| \right\|_1.$$

Define a recurrent scheme,

$$\begin{aligned} \mathbf{y}_{0,L} &= \mathbf{V}, \\ \mathbf{y}_{l+1,L} &= \mathbf{y}_{l,L} + \frac{1}{L} |\mathbf{U}_l||\boldsymbol{W}_l| \mathbf{y}_{l,L}. \end{aligned}$$

Using Theorem 5 with σ being the identity function and $\mathbf{U}$ and $\boldsymbol{W}$ replaced by $|\mathbf{U}|$ and $|\boldsymbol{W}|$ respectively, we know that $\||\alpha|^T \mathbf{y}_{L,L}\|_1 \to \|f\|_{\mathcal{D}_1(\rho_t)}$ almost surely. Hence, by taking ρ_t such that $\|f\|_{\mathcal{D}_1(\rho_t)} \leq 2\|f\|_{\mathcal{D}_1}$, we have

$$\mathbb{E}\|f_L\|_P \leq 3\|f\|_{\mathcal{D}_1},$$

when L is sufficiently large. Again using Markov's inequality, with probability no less than $\frac{2}{3}$, we have

$$\mathbb{E}\|f_L\|_P \leq 9\|f\|_{\mathcal{D}_1}. \tag{37}$$

Combining the result above with (36), we know that with probability no less than $\frac{1}{3}$, we have both (36) and (37). Therefore, we can find an f_L that satisfies both (36) and (37). This completes the proof. □

3.5.5 Proof of Theorem 7

For any L, let $f_L(\cdot)$ be the residual network represented by the parameters $\boldsymbol{\alpha}_L$, $\{\mathbf{U}_l^L, \boldsymbol{W}_l^L\}_{l=0}^{L-1}$ and $\mathbf{V}$. Let $z_{l,L}(\mathbf{x})$ be the function represented by the l-th layer of network f_L, then $f_L(\mathbf{x}) = \boldsymbol{\alpha}_L^T z_{L,L}(\mathbf{x})$ for all $\mathbf{x} \in \mathbf{X}$. Since $\boldsymbol{\alpha}_L$ uniformly bounded for all L, there exists a subsequence L_k and $\boldsymbol{\alpha}$ such that

$$\boldsymbol{\alpha}_{L_k} \to \boldsymbol{\alpha},$$

when $k \to \infty$. Without loss of generality, we assume $\boldsymbol{\alpha}_L \to \boldsymbol{\alpha}$.

Let $\mathbf{U}_t^L : [0, 1] \to \mathbb{R}^{D \times m}$ be a piecewise constant function defined by

$$\mathbf{U}_t^L = \mathbf{U}_l^L, \ for\ t \in [\frac{l}{L}, \frac{l+1}{L}),$$

and $\mathbf{U}_1^L = \mathbf{U}_{L-1}^L$. Similarly we can define $\boldsymbol{W}_t^L$. Then, $\{\mathbf{U}_t^L\}$ and $\{\boldsymbol{W}_t^L\}$ are uniformly bounded. Hence, by the fundamental theorem for Young measures [3,23], there exists a subsequence $\{L_k\}$ and a family of probability measure $\{\rho_t, t \in [0,1]\}$, such that for every Caratheodory function F,

$$\lim_{k\to\infty} \int_0^1 F(\mathbf{U}_t^{L_k}, \boldsymbol{W}_t^{L_k}, t)dt = \int_0^1 \mathbb{E}_{\rho_t} F(\mathbf{U}, \boldsymbol{W}, t)dt.$$

Let $\tilde{f} = f_{\boldsymbol{\alpha},\{\rho_t\}}$. We are going to show $\tilde{f} = f$. Let $z_Y(\cdot, t)$ be defined by $z_Y(\mathbf{x}, 0) = \mathbf{V}\mathbf{x}$ and

$$z_Y(\mathbf{x}, t) = z_Y(\mathbf{x}, 0) + \int_0^t \mathbb{E}_{\rho_t} U\sigma(W z_Y(\mathbf{x}, s))ds.$$

Then it suffices to show that

$$\lim_{k\to\infty} z_{L_k,L_k}(\mathbf{x}) \to z_Y(\mathbf{x}, 1), \tag{38}$$

for any fixed $\mathbf{x} \in D_0$.

To prove (38), we first consider the following continuous version of $z_{l,L}$,

$$\begin{aligned} z_L(\mathbf{x}, 0) &= z_{0,L}(\mathbf{x}), \\ \frac{d}{dt} z_L(\mathbf{x}, t) &= \mathbf{U}_t^L \sigma(\boldsymbol{W}_t^L z_L(\mathbf{x}, t)), \end{aligned}$$

and show that $|z_L(\mathbf{x}, 1) - z_{L,L}(\mathbf{x})| \to 0$. To see this, note that

$$z_L(\mathbf{x}, t_{l+1,L}) = z_L(\mathbf{x}, t_{l,L}) + \int_{t_{l,L}}^{t_{l+1,L}} \mathbf{U}_t^L \sigma(\boldsymbol{W}_t^L z_L(\mathbf{x}, s))ds, \tag{39}$$

$$z_{l+1,L}(\mathbf{x}) = z_{l,L}(\mathbf{x}) + \int_{t_{l,L}}^{t_{l+1,L}} \mathbf{U}_t^L \sigma(\boldsymbol{W}_t^L z_{l,L}(\mathbf{x}))ds. \tag{40}$$

Subtracting (39) from (40), and let $\mathbf{e}_{l,L} = z_{l,L}(\mathbf{x}) - z_L(\mathbf{x}, t_{l,L})$, we have

$$\begin{aligned} \mathbf{e}_{l+1,L} &= \mathbf{e}_{l,L} + \int_{t_{l,L}}^{t_{l+1,L}} \left(\mathbf{U}_t^L \sigma(\boldsymbol{W}_t^L z_{l,L}(\mathbf{x})) - \mathbf{U}_t^L \sigma(\boldsymbol{W}_t^L z_L(\mathbf{x}, s))\right) ds \\ &= \mathbf{e}_{l,L} + \int_{t_{l,L}}^{t_{l+1,L}} \left(\mathbf{U}_t^L \sigma(\boldsymbol{W}_t^L z_{l,L}(\mathbf{x})) - \mathbf{U}_t^L \sigma(\boldsymbol{W}_t^L z_L(\mathbf{x}, t_{l,L}))\right) ds \\ &\quad + \int_{t_{l,L}}^{t_{l+1,L}} \left(\mathbf{U}_t^L \sigma(\boldsymbol{W}_t^L z_L(\mathbf{x}, t_{l,L})) - \mathbf{U}_t^L \sigma(\boldsymbol{W}_t^L z_L(\mathbf{x}, s))\right) ds. \end{aligned} \tag{41}$$

Since $\{\mathbf{U}_t^L\}$ and $\{\boldsymbol{W}_t^L\}$ are bounded, we know that $\{z_L(\mathbf{x}, t)\}$ is bounded, and $\{\frac{d}{dt} z_L(\mathbf{x}, t)\}$ is also bounded. Hence, there exists a uniform constant C such that

$$\left\| \mathbf{U}_t^L \sigma(\boldsymbol{W}_t^L z_{l,L}(\mathbf{x})) - \mathbf{U}_t^L \sigma(\boldsymbol{W}_t^L z_L(\mathbf{x}, t_{l,L})) \right\| \leq C \|e_{l,L}\|, \tag{42}$$

$$\left\| \mathbf{U}_t^L \sigma(\boldsymbol{W}_t^L z_L(\mathbf{x}, t_{l,L})) - \mathbf{U}_t^L \sigma(\boldsymbol{W}_t^L z_L(\mathbf{x}, s)) \right\| \leq C|s - t_{l,L}|.. \tag{43}$$

Plugging (42) and (43) into (41), we obtain

$$\|\mathbf{e}_{l+1,L}\| \leq \left(1 + \frac{C}{L}\right) \|\mathbf{e}_{l,L}\| + \frac{C}{L^2}.$$

Therefore, by Gronwall's inequality, $\|\mathbf{e}_{L,L}\| \leq \mathcal{O}(1/L)$, which gives

$$|z_L(\mathbf{x}, 1) - z_{L,L}(\mathbf{x})| \to 0. \tag{44}$$

Now with (44), we only need to show

$$\lim_{k \to \infty} z_{L_k}(\mathbf{x}, 1) \to z_Y(\mathbf{x}, 1),$$

which is equivalent to showing that for any ϵ, there exists $K > 0$ such that for any $k > K$, we have

$$\|z_{L_k}(\mathbf{x}, 1) - z_Y(\mathbf{x}, 1)\| \leq \epsilon.$$

For a large integer N, let $t_{i,N} = i/N$. By the definition of z_Y and z_{L_k}, we have

$$z_{L_k}(\mathbf{x}, t_{i+1,N}) = z_{L_k}(\mathbf{x}, t_{i,N}) + \int_{t_{i,N}}^{t_{i+1,N}} \mathbf{U}_s^{L_k} \sigma(\boldsymbol{W}_s^{L_k} z_{L_k}(\mathbf{x}, s)) ds,$$

and

$$z_Y(\mathbf{x}, t_{i+1,N}) = z_Y(\mathbf{x}, t_{i,N}) + \int_{t_{i,N}}^{t_{i+1,N}} \mathbb{E}_{\rho_t} \mathbf{U} \sigma(\boldsymbol{W} z_Y(\mathbf{x}, s)) ds.$$

Let $r_{i,N}(\mathbf{x}) = z_Y(\mathbf{x}, t_{i,N}) - z_{L_k}(\mathbf{x}, t_{i,N})$, and note that $\{\mathbf{U}_t^{L_k}\}$ and $\{\boldsymbol{W}_t^{L_k}\}$ are bounded, we have

$$\begin{aligned} \|r_{i+1,N}(\mathbf{x})\| \leq & \left(1 + \frac{C}{N}\right) \|r_{i,N}\| + \frac{C}{N^2} \\ & + \left\| \int_{t_{i,N}}^{t_{i+1,N}} \left[\mathbf{U}_s^{L_k} \sigma(\boldsymbol{W}_s^{L_k} z_Y(\mathbf{x}, s)) - \mathbb{E}_{\rho_t} \mathbf{U} \sigma(\boldsymbol{W} z_Y(\mathbf{x}, s)) \right] ds \right\|, \end{aligned}$$

for some constant C. Using the theorem for Young measures [3,23], there exists a sufficiently large K, such that for all $k > K$, we have

$$\left\| \int_{t_{i,N}}^{t_{i+1,N}} \left[\mathbf{U}_s^{L_k} \sigma(\boldsymbol{W}_s^{L_k} z_Y(\mathbf{x}, s)) - \mathbb{E}_{\rho_t} \mathbf{U} \sigma(\boldsymbol{W} z_Y(\mathbf{x}, s)) \right] ds \right\| \leq \frac{1}{N^2},$$

for all $0 \leq i \leq N-1$. By Gronwall's inequality, there exists a constant $\tilde{C}$ such that

$$\|r_{N,N}(\mathbf{x})\| \leq \frac{\tilde{C}}{N}.$$

If we take $N = \epsilon/\tilde{C}$, we have

$$\|z_{L_k}(\mathbf{x}, 1) - z_Y(\mathbf{x}, 1)\| \leq \epsilon,$$

for sufficiently large k. This shows that $f = f_{\boldsymbol{\alpha}, \{\rho_t\}}$.

To bound the $\mathcal{D}_\infty$ norm of f, take F as the indicator function of $\{|\mathbf{U}| \leq c_0, |\boldsymbol{W}| \leq c_0\}^c$ and apply the theorem for Young measures, we obtain that for any $t \in [0, 1]$, the support of ρ_t lies in $\{|\mathbf{U}| \leq c_0, |\boldsymbol{W}| \leq c_0\}$. Hence, $f \in \mathcal{D}_\infty$. To estimate $\|f\|_{\mathcal{D}_\infty}$, consider $N_\infty(t)$ defined by (16), since the elements of $\mathbf{U}$ and $\boldsymbol{W}$ are bounded by c_0, we have

$$\dot{N}_\infty(t) \leq m c_0^2 E N_\infty(t),$$

where E is an all-one $D \times D$ matrix. Therefore, we have

$$N_\infty(1) \leq e^{m c_0^2 E} \mathbf{e} \leq \frac{2 D e^{m(c_0^2+1)}}{m} \mathbf{e}.$$

Since the elements of $\boldsymbol{\alpha}$ are also bounded by c_0, we get

$$\|f\|_{\mathcal{D}_\infty} \leq |\boldsymbol{\alpha}|^T N_\infty(1) \leq \frac{2 D^2 e^{m(c_0^2+1)} c_0}{m}.$$

Finally, if $\|f_L\|_{\mathcal{D}_1} \leq c_1$ holds for all $L > 0$, then using the technique of treating $z_Y(\mathbf{x}, t)$ on $N_1(t)$, we obtain $\|f\|_{\mathcal{D}_1} \leq c_1$. □

3.5.6 Proof of Theorem 8

Similar to the proof of Theorem 6, we can define a discrete analogy of the $\hat{\mathcal{D}}_1$ norm for residual network

$$\|\Theta\|_{\mathrm{WP}} = |\boldsymbol{\alpha}|^T \prod_{l=1}^{L} \left(I + \frac{2}{L} |\mathbf{U}_l| |\boldsymbol{W}_l| \right) \mathbf{e}.$$

Using the same techniques as for the direct approximation theorem, we can show that any functions in $\hat{\mathcal{D}}_2^Q$ can be approximated by a series of residual networks $f_L(\cdot;\Theta_L)$ with depth L tends to infinity and $\|\Theta_L\|_{\mathrm{WP}} \le 9Q$. Here we use WP (weighted path) to denote the discrete norm because this norm is a weighted version of the original path norm and assigns larger weights for those paths going through more non-linearities. Let $\mathcal{F}^Q$ be the set of all residual networks whose weighted path norms are bounded by Q, i.e.,

$$\mathcal{F}^Q = \{f(\cdot;\Theta) : \ f(\cdot;\Theta) \text{ is a residual network and } \|\Theta\|_{\mathrm{WP}} \le Q\},$$

and let $\overline{\mathcal{F}}^Q$ be the closure of $\mathcal{F}^Q$. Then, by the direct approximation results, $\hat{\mathcal{D}}_2^Q \subset \hat{\mathcal{D}}_1^Q \subset \overline{\mathcal{F}}^Q$. Hence, $\mathrm{Rad}_n(\hat{\mathcal{D}}_2^Q) \le \mathrm{Rad}_n(\overline{\mathcal{F}}^{9Q})$. On the other hand, in [10] it is proven that

$$\mathrm{Rad}_n(\overline{\mathcal{F}}^Q) \le 2Q\sqrt{\frac{2\log(2d)}{n}}.$$

Therefore,

$$\mathrm{Rad}_n(\hat{\mathcal{D}}_2^Q) \le 18Q\sqrt{\frac{2\log(2d)}{n}}$$

□

4 Concluding Remarks

As far as the high-dimensional approximation theory is concerned, we are interested in approximation schemes (or machine learning models) that satisfy

$$\|f - f_m\|^2 \le C_0 \frac{\gamma(f)^2}{m}$$

for f is a certain function space $\mathcal{F}$ defined by the particular approximation scheme or machine learning model. Here γ is a functional defined on $\mathcal{F}$, typically a norm for the function space. It plays the role of the variance in the context of Monte Carlo integration. A machine learning model is preferred if its associated function space $\mathcal{F}$ is large and the functional γ is small.

However, practical machine learning models can only work with a finite dataset on which the values of the target function are known. This results in an additional error, the estimation error, in the total error of the machine learning model. The estimation error is controlled by the Rademacher complexity of the hypothesis space, which can be thought of as a truncated version of the space $\mathcal{F}$. It just so happens that for the spaces identified here the Rademacher complexity has the optimal estimates:

$$\mathrm{Rad}_n(\mathcal{F}_Q) \le C_0 \frac{Q}{\sqrt{n}}.$$

This is also true for the RKHS. It is not clear whether this is a coincidence, or there are some more fundamental reasons behind.

Whatever the reason, the combination of these two results imply that the generalization error (also called population risk) should have the optimal scaling $O(1/m) + O(1/\sqrt{n})$ for all three methods: the kernel method, the two-layer neural networks and residual networks. The difference lies in the coefficients hidden in the above expression. These coefficients are the norms of the target function in the corresponding function spaces. In this sense, going from the kernel method to two-layer neural networks and to deep residual neural networks is like a variance reduction process since the value of the norms decreases in this process. In addition, the function space $\mathcal{F}$ expands substantially from some RKHS to the Barron space and to the flow-induced function space.

Acknowledgements The work presented here is supported in part by a gift to Princeton University from iFlytek and the ONR Grant N00014-13-1-0338.

References

1. Aronszajn, N.: Theory of reproducing kernels. Trans. Am. Math. Soc. **68**(3), 337–404 (1950)
2. Bach, F.: Breaking the curse of dimensionality with convex neural networks. J. Mach. Learn. Res. **18**(19), 1–53 (2017)
3. Ball, J.M.: A version of the fundamental theorem for young measures. In: PDEs and continuum models of phase transitions, pp. 207–215. Springer (1989)
4. Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Trans. Inf. Theory **39**(3), 930–945 (1993)
5. Barron, A.R.: Approximation and estimation bounds for artificial neural networks. Mach. Learn. **14**(1), 115–133 (1994)
6. Bartlett, P.L., Mendelson, S.: Rademacher and Gaussian complexities: risk bounds and structural results. J. Mach. Learn. Res. **3**(Nov), 463–482 (2002)
7. Benveniste, A., Métivier, M., Priouret, P.: Adaptive Algorithms and Stochastic Approximations, vol. 22. Springer, Berlin (2012)
8. Ciarlet, P.G.: The finite element method for elliptic problems. Class. Appl. Math. **40**, 1–511 (2002)
9. DeVore, R.A., Lorentz, G.G.: Constructive Approximation, vol. 303. Springer, Berlin (1993)
10. E, W., Ma, C., Wang, Q.: A priori estimates of the population risk for residual networks. arXiv preprint arXiv:1903.02154 (2019)
11. E, W., Ma, C., Lei, W.: A priori estimates of the population risk for two-layer neural networks. Commun. Math. Sci. **17**(5), 1407–1425 (2019). arXiv:1810.06397
12. E, W., Wojtowytsch, S.: Representation formulas and pointwise properties for barron functions. arXiv preprint arXiv:2006.05982 (2020)
13. Eldan, R., Shamir, O.: The power of depth for feedforward neural networks. In: Conference on Learning Theory, pp. 907–940 (2016)
14. Jentzen, A., Salimova, D., Welti, T.: A proof that deep artificial neural networks overcome the curse of dimensionality in the numerical approximation of kolmogorov partial differential equations with constant diffusion and nonlinear drift coefficients. arXiv preprint arXiv:1809.07321 (2018)
15. Klusowski, J.M., Barron, A.R.: Risk bounds for high-dimensional ridge function combinations including neural networks. arXiv preprint arXiv:1607.01434 (2016)
16. Kurková, V., Sanguineti, M.: Bounds on rates of variable-basis and neural-network approximation. IEEE Trans. Inf. Theory **47**(6), 2659–2665 (2001)
17. Kushner, H., Yin, G.G.: Stochastic Approximation and Recursive Algorithms and Applications, vol. 35. Springer, Berlin (2003)
18. Li, Z., Ma, C., Wu, L.: Complexity measures for neural networks with general activation functions using path-based norms. arXiv preprint arXiv:2009.06132 (2020)

19. Mhaskar, H.N.: On the tractability of multivariate integration and approximation by neural networks. J. Complex. **20**(4), 561–590 (2004)
20. Neyshabur, B., Bhojanapalli, S., Mcallester, D., Srebro, N.: Exploring generalization in deep learning. Adv. Neural. Inf. Process. Syst. **30**, 5949–5958 (2017)
21. Rahimi, A., Recht, B.: Uniform approximation of functions with random bases. In: 2008 46th Annual Allerton Conference on Communication, Control, and Computing, pp. 555–561. IEEE (2008)
22. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, Cambridge (2014)
23. Young, L.C.: Lecture on the calculus of variations and optimal control theory, vol. 304. American Mathematical Society (2000)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.