frontiers | Frontiers in Applied Mathematics and Statistics

TYPE Original Research
PUBLISHED 17 April 2024
DOI 10.3389/fams.2024.1284706

OPEN ACCESS

EDITED BY
HanQin Cai,
University of Central Florida, United States

REVIEWED BY
Jingyang Li,
The University of Hong Kong, Hong Kong SAR, China
Dong Xia,
Hong Kong University of Science and Technology, Hong Kong SAR, China

*CORRESPONDENCE
Deanna Needell
deanna@math.ucla.edu

RECEIVED 28 August 2023
ACCEPTED 21 March 2024
PUBLISHED 17 April 2024

CITATION
Needell D, Nelson AA, Saab R, Salanevich P and Schavemaker O (2024) Random vector functional link networks for function approximation on manifolds.
Front. Appl. Math. Stat. 10:1284706.
doi: 10.3389/fams.2024.1284706

Random vector functional link networks for function approximation on manifolds

Deanna Needell[1]*, Aaron A. Nelson[2], Rayan Saab[3], Palina Salanevich[4] and Olov Schavemaker[4]

[1]Department of Mathematics, University of California, Los Angeles, Los Angeles, CA, United States, [2]Department of Mathematical Sciences, United States Air Force Academy, Colorado Springs, CO, United States, [3]Department of Mathematics and Halıcıoğlu Data Science Institute, University of California, San Diego, San Diego, CA, United States, [4]Mathematical Institute, Utrecht University, Utrecht, Netherlands

The learning speed of feed-forward neural networks is notoriously slow and has presented a bottleneck in deep learning applications for several decades. For instance, gradient-based learning algorithms, which are used extensively to train neural networks, tend to work slowly when all of the network parameters must be iteratively tuned. To counter this, both researchers and practitioners have tried introducing randomness to reduce the learning requirement. Based on the original construction of Igelnik and Pao, single layer neural-networks with random input-to-hidden layer weights and biases have seen success in practice, but the necessary theoretical justification is lacking. In this study, we begin to fill this theoretical gap. We then extend this result to the non-asymptotic setting using a concentration inequality for Monte-Carlo integral approximations. We provide a (corrected) rigorous proof that the Igelnik and Pao construction is a universal approximator for continuous functions on compact domains, with approximation error squared decaying asymptotically like $O(1/n)$ for the number n of network nodes. We then extend this result to the non-asymptotic setting, proving that one can achieve any desired approximation error with high probability provided n is sufficiently large. We further adapt this randomized neural network architecture to approximate functions on smooth, compact submanifolds of Euclidean space, providing theoretical guarantees in both the asymptotic and non-asymptotic forms. Finally, we illustrate our results on manifolds with numerical experiments.

KEYWORDS

machine learning, feed-forward neural networks, function approximation, smooth manifold, random vector functional link

1 Introduction

In recent years, neural networks have once again triggered an increased interest among researchers in the machine learning community. So-called deep neural networks model functions using a composition of multiple hidden layers, each transforming (possibly non-linearly) the previous layer before building a final output representation [1–5]. In machine learning parlance, these layers are determined by sets of *weights* and *biases* that can be tuned so that the network mimics the action of a complex function. In particular, a single layer feed-forward neural network (SLFN) with n nodes may be regarded as a parametric function $f_n : \mathbb{R}^N \to \mathbb{R}$ of the form

$$f_n(x) = \sum_{k=1}^{n} v_k \rho(\langle w_k, x \rangle + b_k), \quad x \in \mathbb{R}^N.$$

Here, the function $\rho\colon \mathbb{R} \to \mathbb{R}$ is called an activation function and is potentially non-linear. Some typical examples include the sigmoid function $\rho(z) = \frac{1}{1+\exp(-z)}$, ReLU $\rho(z) = \max\{0, z\}$, and sign functions, among many others. The parameters of the SLFN are the number of nodes $n \in \mathbb{N}$ in the the hidden layer, the input-to-hidden layer weights and biases $\{w_k\}_{k=1}^n \subset \mathbb{R}^N$ and $\{b_k\}_{k=1}^n \subset \mathbb{R}$ (resp.), and the hidden-to-output layer weights $\{v_k\}_{k=1}^n \subset \mathbb{R}$. In this way, neural networks are fundamentally parametric families of functions whose parameters may be chosen to approximate a given function.

It has been shown that any compactly supported continuous function can be approximated with any given precision by a single layer neural network with a suitably chosen number of nodes [6], and harmonic analysis techniques have been used to study stability of such approximations [7]. Other recent results that take a different approach directly analyze the capacity of neural networks from a combinatorial point of view [8, 9].

While these results ensure existence of a neural network approximating a function, practical applications require construction of such an approximation. The parameters of the neural network can be chosen using optimization techniques to minimize the difference between the network and the function $f\colon \mathbb{R}^N \to \mathbb{R}$ it is intended to model. In practice, the function f is usually not known, and we only have access to a set $\{(x_k, f(x_k))\}_{k=1}^m$ of values of the function at finitely many points sampled from its domain, called a *training set*. The approximation error can be measured by comparing the training data to the corresponding network outputs when evaluated on the same set of points, and the parameters of the neural network f_n can be *learned* by minimizing a given loss function $\mathcal{L}(x_1, \ldots, x_m)$; a typical loss function is the sum-of-squares error

$$\mathcal{L}(x_1, \ldots, x_m) = \frac{1}{m} \sum_{k=1}^{m} |f(x_k) - f_n(x_k)|^2.$$

The SLFN which approximates f is then determined using an optimization algorithm, such as back-propagation, to find the network parameters which minimize $\mathcal{L}(x_1, \ldots, x_m)$. It is known that there exist weights and biases which make the loss function vanish when the number of nodes n is at least m, provided the activation function is bounded, non-linear, and has at least one finite limit at either $\pm\infty$ [10].

Unfortunately, optimizing the parameters in SLFNs can be difficult. For instance, any non-linearity in the activation function can cause back-propagation to be very time-consuming or get caught in local minima of the loss function [11]. Moreover, deep neural networks can require massive amounts of training data, and so are typically unreliable for applications with very limited data availability, such as agriculture, healthcare, and ecology [12].

To address some of the difficulties associated with training deep neural networks, both researchers and practitioners have attempted to incorporate randomness in some way. Indeed, randomization-based neural networks that yield closed form solutions typically require less time to train and avoid some of the pitfalls of traditional neural networks trained using back-propagation [11, 13, 14]. One of the popular randomization-based neural network architectures is the Random Vector Functional Link (RVFL) network [15, 16], which is a single layer feed-forward neural network in which the input-to-hidden layer weights and biases are selected randomly and independently from a suitable domain and the remaining hidden-to-output layer weights are learned using training data.

By eliminating the need to optimize the input-to-hidden layer weights and biases, RVFL networks turn supervised learning into a purely linear problem. To see this, define $\rho(X) \in \mathbb{R}^{n\times m}$ to be the matrix whose jth column is $\{\rho(\langle w_k, x_j\rangle + b_k)\}_{k=1}^n$ and $f(X) \in \mathbb{R}^m$ the vector whose jth entry is $f(x_j)$. Then, the vector $v \in \mathbb{R}^n$ of hidden-to-output layer weights is the solution to the matrix-vector equation $f(X) = \rho(X)^T v$, which can be solved by computing the Moore-Penrose pseudoinverse of $\rho(X)^T$. In fact, there exist weights and biases that make the loss function vanish when the number of nodes n is at least m, provided the activation function is smooth [17].

Although originally considered in the early- to mid-1990s [15, 16, 18, 19], RVFL networks have had much more recent success in several modern applications, including time-series data prediction [20], handwritten word recognition [21], visual tracking [22], signal classification [23, 24], regression [25], and forecasting [26, 27]. Deep neural network architectures based on RVFL networks have also made their way into more recent literature [28, 29], although traditional, single layer RVFL networks tend to perform just as well as, and with lower training costs than, their multi-layer counterparts [29].

Even though RVFL networks are proving their usefulness in practice, the supporting theoretical framework is currently lacking [see 30]. Most theoretical research into the approximation capabilities of deep neural networks centers around two main concepts: universal approximation of functions on compact domains and point-wise approximation on finite training sets [17]. For instance, in the early 1990s, it was shown that multi-layer feed-forward neural networks having activation functions that are continuous, bounded, and non-constant are universal approximators (in the L^p sense for $1 \leq p < \infty$) of continuous functions on compact domains [31, 32]. The most notable result in the existing literature regarding the universal approximation capability of RVFL networks is due to Igelnik and Pao [16] in the mid-1990s, who showed that such neural networks can universally approximate continuous functions on compact sets; the noticeable lack of results since has left a sizable gap between theory and practice. In this study, we begin to bridge this gap by further improving the Igelnik and Pao result, and bringing the mathematical theory behind RFVL networks into the modern spotlight. Below, we introduce the notation that will be used throughout this study, and describe our main contributions.

1.1 Notation

For a function $f\colon \mathbb{R}^N \to \mathbb{R}$, the set $\text{supp}(f) \subset \mathbb{R}^N$ denotes the support of f. We denote by $C_c(\mathbb{R}^N)$ and $C_0(\mathbb{R}^N)$ the classes of continuous functions mapping $\mathbb{R}^N$ to $\mathbb{R}$ whose support sets are compact and vanish at infinity, respectively. Given a set $S \subset \mathbb{R}^N$, we define its radius to be $\text{rad}(S) := \sup_{x\in S} \|x\|_2$; moreover, if $d\mu$ denotes the uniform volume measure on S, then we write $\text{vol}(S) := \int_S d\mu$ to represent the volume of S. For any probability distribution $P\colon \mathbb{R}^N \to [0, 1]$, a random variable X distributed

according to P is denoted by $X \sim P$, and we write its expectation as $\mathbb{E}X := \int_{\mathbb{R}^N} X \mathrm{d}P$. The open ℓ_p ball of radius $r > 0$ centered at $x \in \mathbb{R}^N$ is denoted by $B_p^N(x, r)$ for all $1 \leq p \leq \infty$; the ℓ_p unit-ball centered at the origin is abbreviated B_p^N. Given a fixed $\delta > 0$ and a set $S \subset \mathbb{R}^N$, a minimal δ-net for S, which we denote $\mathcal{C}(\delta, S)$, is the smallest subset of S satisfying $S \subset \cup_{x \in \mathcal{C}(\delta,S)} B_2^N(x, \delta)$; the δ-covering number of S is the cardinality of a minimal δ-net for S and is denoted $\mathcal{N}(\delta, S) := |\mathcal{C}(\delta, S)|$.

1.2 Main results

In this study, we analyze the uniform approximation capabilities of RVFL networks. More specifically, we consider the problem of using RVFL networks to estimate a continuous, compactly supported function on N-dimensional Euclidean space.

The first theoretical result on approximating properties of RVFL networks, due to Igelnik and Pao [16], guarantees that continuous functions may be universally approximated on compact sets using RVFL networks, provided the number of nodes $n \in \mathbb{N}$ in the network goes to infinity. Moreover, it shows that the mean square error of the approximation vanishes at a rate proportional to $1/n$. At the time, this result was state-of-the-art and justified how RVFL networks were used in practice. However, the original theorem is not technically correct. In fact, several aspects of the proof technique are flawed. Some of the minor flaws are mentioned in Li et al. [33], but the subsequent revisions do not address the more significant issues which would make the statement of the result technically correct. We address these issues in this study, see Remark 1. Thus, our first contribution to the theory of RVFL networks is a corrected version of the original Igelnik and Pao theorem:

Theorem 1 ([16]). Let $f \in C_c(\mathbb{R}^N)$ with $K := \mathrm{supp}(f)$ and fix any activation function ρ, such that either $\rho \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ with $\int_{\mathbb{R}} \rho(z)\mathrm{d}z = 1$ or ρ is differentiable with $\rho' \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ and $\int_{\mathbb{R}} \rho'(z)\mathrm{d}z = 1$. For any $\varepsilon > 0$, there exist distributions from which input weights $\{w_k\}_{k=1}^n$ and biases $\{b_k\}_{k=1}^n$ are drawn, and there exist hidden-to-output layer weights $\{v_k\}_{k=1}^n \subset \mathbb{R}$ that depend on the realization of weights and biases, such that the sequence of RVFL networks $\{f_n\}_{n=1}^\infty$ is defined by

$$f_n(x) := \sum_{k=1}^{n} v_k \rho(\langle w_k, x\rangle + b_k) \quad \text{for } x \in K$$

satisfies

$$\mathbb{E} \int_K |f(x) - f_n(x)|^2 \mathrm{d}x < \varepsilon + O(1/n),$$

as $n \to \infty$.

For a more precise formulation of Theorem 1 and its proof, we refer the reader to Theorem 5 and Section 3.1.

Remark 1.

1. Even though in Theorem 1 we only claim existence of the distribution for input weights $\{w_k\}_{k=1}^n$ and biases $\{b_k\}_{k=1}^n$, such a distribution is actually constructed in the proof. Namely, for any $\varepsilon > 0$, there exist constants $\alpha, \Omega > 0$ such that the random variables

$$\begin{aligned} w_k &\sim \mathrm{Unif}\left([-\alpha\Omega, \alpha\Omega]^N\right); \\ y_k &\sim \mathrm{Unif}\,(K); \\ u_k &\sim \mathrm{Unif}\left([-\tfrac{\pi}{2}(2L+1), \tfrac{\pi}{2}(2L+1)]\right), \\ &\text{where } L := \lceil \tfrac{2N}{\pi} \mathrm{rad}(K)\Omega - \tfrac{1}{2} \rceil, \end{aligned}$$

   are independently drawn from their associated distributions, and $b_k := -\langle w_k, y_k\rangle - \alpha u_k$.
2. We note that, unlike the original theorem statement in Igelnik and Pao [16], Theorem 1 does not show exact convergence of the sequence of constructed RVFL networks f_n to the original function f. Indeed, it only ensures that the limit f_n is ε-close to f. This should still be sufficient for practical applications since, given a desired accuracy level $\varepsilon > 0$, one can find values of α, Ω, and n such that this accuracy level is achieved on average. Exact convergence can be proved if one replaces α and Ω in the distribution described above by sequences $\{\alpha_n\}_{n=1}^\infty$ and $\{\Omega_n\}_{n=1}^\infty$ of positive numbers, both tending to infinity with n. In this setting, however, there is no guaranteed rate of convergence; moreover, as n increases, the ranges of the random variables $\{w_k\}_{k=1}^n$ and $\{u_k\}_{k=1}^n$ become increasingly larger, which may cause problems in practical applications.
3. The approach we take to construct the RVFL network approximating a function f allows one to compute the output weights $\{v_k\}_{k=1}^n$ exactly (once the realization of random parameters is fixed), in the case where the function f is known. For the details, we refer the reader to Equations 6, 8 in the proof of Theorem 1. If we only have access to a training set that is sufficiently large and uniformly distributed over the support of f, these formulas can be used to compute the output weights approximately, instead of solving the least squares problem.
4. Note that the normalization $\int_{\mathbb{R}} \rho(z)\mathrm{d}z = 1$ of the activation function can be replaced by the condition $\int_{\mathbb{R}} \rho(z)\mathrm{d}z \neq 0$. Indeed, in the case when $\rho \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ and $\int_{\mathbb{R}} \rho(z)\mathrm{d}z \notin \{0, 1\}$, one can simply use Theorem 1 to approximate $\frac{1}{\int_{\mathbb{R}} \rho(z)\mathrm{d}z} f$ by a sequence of RVFL network with the activation function $\frac{1}{\int_{\mathbb{R}} \rho(z)\mathrm{d}z} \rho$. Mutatis mutandis in the case when $\int_{\mathbb{R}} \rho'(z)\mathrm{d}z' \notin \{0, 1\}$. More generally, this trick allows any of our theorems to be applied in the case $\int_{\mathbb{R}} \rho(z)\mathrm{d}z \neq 0$.

One of the drawbacks of Theorem 1 is that the mean square error guarantee is asymptotic in the number of nodes used in the neural network. This is clearly impractical for applications, and so it is desirable to have a more explicit error bound for each fixed number n of nodes used. To this end, we provide a new, non-asymptotic version of Theorem 1, which provides an error guarantee with high probability whenever the number of network nodes is large enough, albeit at the price of an additional Lipschitz requirement on the activation function:

Theorem 2. Let $f \in C_c(\mathbb{R}^N)$ with $K := \mathrm{supp}(f)$ and fix any activation function $\rho \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ with $\int_{\mathbb{R}} \rho(z)\mathrm{d}z = 1$. Suppose further that ρ is κ-Lipschitz on $\mathbb{R}$ for some $\kappa > 0$. For any $\varepsilon > 0$ and $\eta \in (0, 1)$, suppose that $n \geq C(N, f, \rho)\varepsilon^{-1} \log(\eta^{-1}/\varepsilon)$,

where $C(N,f,\rho)$ is independent of ε and η and depends on f, ρ, and superexponentially on N. Then, there exist distributions from which input weights $\{w_k\}_{k=1}^n$ and biases $\{b_k\}_{k=1}^n$ are drawn, and there exist hidden-to-output layer weights $\{v_k\}_{k=1}^n \subset \mathbb{R}$ that depend on the realization of weights and biases, such that the RVFL network defined by

$$f_n(x) := \sum_{k=1}^{n} v_k \rho(\langle w_k, x\rangle + b_k) \quad \text{for } x \in K$$

satisfies

$$\int_K |f(x) - f_n(x)|^2 \mathrm{d}x < \varepsilon$$

with probability at least $1-\eta$.

For simplicity, the bound on the number n of the nodes on the hidden layer here is rough. For a more precise formulation of this result that contains a bound with explicit constant, we refer the reader to Theorem 6 in Section 3.2. We also note that the distribution of the input weight and bias here can be selected as described in Remark 1.

The constructions of RVFL networks presented in Theorems 1, 2 depend heavily on the dimension of the ambient space $\mathbb{R}^N$. If N is small, this dependence does not present much of a problem. However, many modern applications require the ambient dimension to be large. Fortunately, a common assumption in practice is the support of the signals of interest lies on a lower-dimensional manifold embedded in $\mathbb{R}^N$. For instance, the landscape of cancer cell states can be modeled using non-linear, locally continuous "cellular manifolds;" indeed, while the ambient dimension of this state space is typically high (e.g., single-cell RNA sequencing must account for approximately 20,000 gene dimensions), cellular data actually occupies an intrinsically lower dimensional space [34]. Similarly, while the pattern space of neural population activity in the brain is described by an exponential number of parameters, the spatiotemporal dynamics of brain activity lie on a lower-dimensional subspace or "neural manifold" [35]. In this study, we propose a new RVFL network architecture for approximating continuous functions defined on smooth compact manifolds that allows to replace the dependence on the ambient dimension N with dependence on the manifold intrinsic dimension. We show that RVFL approximation results can be extended to this setting. More precisely, we prove the following analog of Theorem 2.

Theorem 3. Let $\mathcal{M} \subset \mathbb{R}^N$ be a smooth, compact d-dimensional manifold with finite atlas $\{(U_j,\phi_j)\}_{j\in J}$ and $f \in C(\mathcal{M})$. Fix any activation function $\rho \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ with $\int_{\mathbb{R}} \rho(z)\mathrm{d}z = 1$ such that ρ is κ-Lipschitz on $\mathbb{R}$ for some $\kappa > 0$. For any $\varepsilon > 0$ and $\eta \in (0,1)$, suppose $n \geq C(d,f,\rho)\varepsilon^{-1}\log(\eta^{-1}/\varepsilon)$, where $C(d,f,\rho)$ is independent of ε and η and depends on f, ρ, and superexponentially on d. Then, there exists an RVFL-like approximation f_n of the function f with a parameter selection similar to the Theorem 1 construction that satisfies

$$\int_{\mathcal{M}} |f(x) - f_n(x)|^2 \mathrm{d}x < \varepsilon$$

with probability at least $1-\eta$.

For a the construction of the RVFL-like approximation f_n and a more precise formulation of this result and an analog of Theorem 1 applied to manifolds, we refer the reader to Section 3.3.1 and Theorems 7, 8. We note that the approximation f_n here is not obtained as a single RVFL network construction, but rather as a combination of several RVFL networks in local manifold coordinates.

1.3 Organization

The remaining part of the article is organized as follows. In Section 2, we discuss some theoretical preliminaries on concentration bounds for Monte-Carlo integration and on smooth compact manifolds. Monte-Carlo integration is an essential ingredient in our construction of RVFL networks approximating a given function, and we use the results listed in this section to establish approximation error bounds. Theorem 1 is proven in Section 3.1, where we break down the proof into four main steps, constructing a limit-integral representation of the function to be approximated in Lemmas 3, 4, then using Monte-Carlo approximation of the obtained integral to construct an RVFL network in Lemma 5, and, finally, establishing approximation guarantees for the constructed RVFL network. The proofs of Lemmas 3, 4, and 5 can be found in Sections 3.5.1, 3.5.2, and 3.5.3, respectively. We further study properties of the constructed RVFL networks and prove the non-asymptotic approximation result of Theorem 2 in Section 3.2. In Section 3.3, we generalize our results and propose a new RVFL network architecture for approximating continuous functions defined on smooth compact manifolds. We show that RVFL approximation results can be extended to this setting by proving an analog of Theorem 1 in Section 3.3.2 and Theorem 3 in Section 3.5.5. Finally, in Section 3.4, we provide numerical evidence to illustrate the result of Theorem 3.

2 Materials and methods

In this section, we briefly introduce supporting material and theoretical results which we will need in later sections. This material is far from exhaustive, and is meant to be a survey of definitions, concepts, and key results.

2.1 A concentration bound for classic Monte-Carlo integration

A crucial piece of the proof technique employed in Igelnik and Pao [16], which we will use repeatedly, is the use of the Monte-Carlo method to approximate high-dimensional integrals. As such, we start with the background on Monte-Carlo integration. The following introduction is adapted from the material in Dick et al. [36].

Let $f\colon \mathbb{R}^N \to \mathbb{R}$ and $S \subset \mathbb{R}^N$ a compact set. Suppose we want to estimate the integral $I(f,S) := \int_S f\mathrm{d}\mu$, where μ is the uniform measure on S. The classic Monte Carlo method does this by an

equal-weight cubature rule,

$$I_n(f,S) := \frac{\mathrm{vol}(S)}{n}\sum_{j=1}^{n} f(x_j),$$

where $\{x_j\}_{j=1}^n$ are independent identically distributed uniform random samples from S and $\mathrm{vol}(S) := \int_S d\mu$ is the volume of S. In particular, note that $\mathbb{E}I_n(f,S) = I(f,S)$ and

$$\mathbb{E}I_n(f,S)^2 = \frac{1}{n}\big(\mathrm{vol}(S)I(f^2,S) + (n-1)I(f,S)^2\big).$$

Let us define the quantity

$$\sigma(f,S)^2 := \frac{I(f^2,S)}{\mathrm{vol}(S)} - \frac{I(f,S)^2}{\mathrm{vol}^2(S)}. \tag{1}$$

It follows that the random variable $I_n(f)$ has mean $I(f,S)$ and variance $\mathrm{vol}^2(S)\sigma(f,S)^2/n$. Hence, by the Central Limit Theorem, provided that $0 < \mathrm{vol}^2(S)\sigma(f,S)^2 < \infty$, we have

$$\lim_{n\to\infty}\mathbb{P}\Big(|I_n(f,S) - I(f,S)| \le \frac{C\varepsilon(f,S)}{\sqrt{n}}\Big) = (2\pi)^{-1/2}\int_{-C}^{C} e^{-x^2/2}\mathrm{d}x$$

for any constant $C > 0$, where $\varepsilon(f,S) := \mathrm{vol}(S)\sigma(f,S)$. This yields the following well-known result:

Theorem 4. For any $f \in L^2(S,\mu)$, the mean-square error of the Monte Carlo approximation $I_n(f,S)$ satisfies

$$\mathbb{E}\big|I_n(f,S) - I(f,S)\big|^2 = \frac{\mathrm{vol}^2(S)\sigma(f,S)^2}{n},$$

where the expectation is taken with respect to the random variables $\{x_j\}_{j=1}^n$ and $\sigma(f,S)$ is defined in Equation 1.

In particular, Theorem 4 implies $\mathbb{E}\big|I_n(f,S) - I(f,S)\big|^2 = O(1/n)$ as $n\to\infty$.

In the non-asymptotic setting, we are interested in obtaining a useful bound on the probability $\mathbb{P}(|I_n(f,S) - I(f,S)| \ge t)$ for all $t > 0$. The following lemma follows from a generalization of Bennett's inequality (Theorem 7.6 in [37]; see also [38, 39]).

Lemma 1. For any $f \in L^2(S)$ and $n \in \mathbb{N}$, we have

$$\mathbb{P}\Big(|I_n(f,S) - I(f,S)| \ge t\Big) \le 3\exp\left(-\frac{nt}{CK}\log\left(1 + \frac{Kt}{\mathrm{vol}(S)I(f^2,S)}\right)\right)$$

for all $t > 0$ and a universal constant $C > 0$, provided $|\mathrm{vol}(S)f(x)| \le K$ for almost every $x \in S$.

2.2 Smooth, compact manifolds in Euclidean space

In this section, we review several concepts of smooth manifolds that will be useful to us later. Many of the definitions and results that follow can be found, for instance, in Shaham et al. [40]. Let $\mathcal{M} \subset \mathbb{R}^N$ be a smooth, compact d-dimensional manifold. A *chart* for $\mathcal{M}$ is a pair (U,ϕ) such that $U \subset \mathcal{M}$ is an open set and $\phi : U \to \mathbb{R}^d$ is a homeomorphism. One way to interpret a chart is as a tangent space at some point $x \in U$; in this way, a chart defines a Euclidean coordinate system on U via the map ϕ. A collection $\{(U_j,\phi_j)\}_{j\in J}$ of charts defines an *atlas* for $\mathcal{M}$ if $\cup_{j\in J}U_j = \mathcal{M}$. We now define a special collection of functions on $\mathcal{M}$ called a *partition of unity*.

Definition 1. Let $\mathcal{M} \subset \mathbb{R}^N$ be a smooth manifold. A *partition of unity* of $\mathcal{M}$ with respect to an open cover $\{U_j\}_{j\in J}$ of $\mathcal{M}$ is a family of non-negative smooth functions $\{\eta_j\}_{j\in J}$ such that for every $x \in \mathcal{M}$, we have $1 = \sum_{j\in J}\eta_j(x)$ and, for every $j \in J$, $\mathrm{supp}(\eta_j) \subset U_j$.

It is known that if $\mathcal{M}$ is compact, there exists a partition of unity of $\mathcal{M}$ such that $\mathrm{supp}(\eta_j)$ is compact for all $j \in J$ [see 41]. In particular, such a partition of unity exists for any open cover of $\mathcal{M}$ corresponding to an atlas.

Fix an atlas $\{(U_j,\phi_j)\}_{j\in J}$ for $\mathcal{M}$, as well as the corresponding, compactly supported partition of unity $\{\eta_j\}_{j\in J}$. Then, we have the following useful result [see 40, Lemma 4.8].

Lemma 2. Let $\mathcal{M} \subset \mathbb{R}^N$ be a smooth, compact manifold with atlas $\{(U_j,\phi_j)\}_{j\in J}$ and compactly supported partition of unity $\{\eta_j\}_{j\in J}$. For any $f \in C(\mathcal{M})$, we have

$$f(x) = \sum_{\{j\in J\,:\,x\in U_j\}} (\hat{f}_j \circ \phi_j)(x)$$

for all $x \in \mathcal{M}$, where

$$\hat{f}_j(z) := \begin{cases} f(\phi_j^{-1}(z))\,\eta_j(\phi_j^{-1}(z)) & z \in \phi_j(U_j) \\ 0 & \text{otherwise.}\end{cases}$$

In later sections, we use the representation of Lemma 2 to integrate functions $f \in C(\mathcal{M})$ over $\mathcal{M}$. To this end, for each $j \in J$, let $D\phi_j(y)$ denote the differential of ϕ_j at $y \in U_j$, which is a map from the tangent space $T_y\mathcal{M}$ into $\mathbb{R}^d$. One may interpret $D\phi_j(y)$ as the matrix representation of a basis for the cotangent space at $y \in U_j$. As a result, $D\phi_j(y)$ is necessarily invertible for each $y \in U_j$, and so we know that $|\det(D\phi_j(y))| > 0$ for each $y \in U_j$. Hence, it follows by the change of variables theorem that

$$\begin{aligned}\int_{\mathcal{M}} f(x)\mathrm{d}x &= \int_{\mathcal{M}} \sum_{\{j\in J\,:\,x\in U_j\}} (\hat{f}_j \circ \phi_j)(x)\mathrm{d}x \\ &= \sum_{j\in J}\int_{\phi_j(U_j)} \frac{\hat{f}_j(z)}{|\det(D\phi_j(\phi_j^{-1}(z)))|}\mathrm{d}z. \end{aligned} \tag{2}$$

3 Results

In this section, we prove our main results formulated in Section 1.2 and also use numerical simulations to illustrate the RVFL approximation performance in a low-dimensional submanifold setup. To improve readability of this section, we postpone the proofs of technical lemmas till Section 3.5.

3.1 Proof of Theorem 1

We split the proof of the theorem into two parts, first handling the case $\rho \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ and second, addressing the case $\rho' \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$.

3.1.1 Proof of Theorem 1 when $\rho \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$

We begin by restating the theorem in a form that explicitly includes the distributions that we draw our random variables from.

Theorem 5 ([16]). Let $f \in C_c(\mathbb{R}^N)$ with $K := \text{supp}(f)$ and fix any activation function $\rho \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ with $\int_{\mathbb{R}} \rho(z)\mathrm{d}z = 1$. For any $\varepsilon > 0$, there exist constants $\alpha, \Omega > 0$ such that the following holds: If, for $k \in \mathbb{N}$, the random variables

$$\begin{aligned} w_k &\sim \text{Unif}\left([-\alpha\Omega, \alpha\Omega]^N\right); \\ y_k &\sim \text{Unif}\,(K); \\ u_k &\sim \text{Unif}\left([-\tfrac{\pi}{2}(2L+1), \tfrac{\pi}{2}(2L+1)]\right), \\ &\quad \text{where } L := \lceil \tfrac{2N}{\pi}\text{rad}(K)\Omega - \tfrac{1}{2} \rceil, \end{aligned}$$

are independently drawn from their associated distributions, and

$$b_k := -\langle w_k, y_k \rangle - \alpha u_k,$$

then there exist hidden-to-output layer weights $\{v_k\}_{k=1}^n \subset \mathbb{R}$ (that depend on the realization of the weights $\{w_k\}_{k=1}^n$ and biases $\{b_k\}_{k=1}^n$) such that the sequence of RVFL networks $\{f_n\}_{n=1}^\infty$ defined by

$$f_n(x) := \sum_{k=1}^n v_k \rho(\langle w_k, x \rangle + b_k) \quad \text{for } x \in K$$

satisfies

$$\mathbb{E} \int_K |f(x) - f_n(x)|^2 \mathrm{d}x \leq \varepsilon + O(1/n).$$

as $n \to \infty$.

Proof. Our proof technique is based on that introduced by Igelnik and Pao and can be divided into four steps. The first three steps essentially consist of Lemma 3, Lemma 4, and Lemma 5, and the final step combines them to obtain the desired result. First, the function f is approximated by a convolution, given in Lemma 3. The proof of this result can be found in Section 3.5.1.

Lemma 3. Let $f \in C_0(\mathbb{R}^N)$ and $h \in L^1(\mathbb{R}^N)$ with $\int_{\mathbb{R}^N} h(z)\mathrm{d}z = 1$. For $\Omega > 0$, define

$$h_\Omega(y) := \Omega^N h(\Omega y). \tag{3}$$

Then, we have

$$f(x) = \lim_{\Omega \to \infty} (f * h_\Omega)(x) \tag{4}$$

uniformly for all $x \in \mathbb{R}^N$.

Next, we represent f as the limiting value of a multidimensional integral over the parameter space. In particular, we replace $(f * h_\Omega)(x)$ in the convolution identity (Equation 4) with a function of the form $\int_K F(y)\rho(\langle w, x \rangle + b(y))\mathrm{d}y$, as this will introduce the RVFL structure we require. To achieve this, we first use a truncated cosine function in place of the activation function ρ and then switch back to a general activation function.

To that end, for each fixed $\Omega > 0$, let $L = L(\Omega) := \lceil \frac{2N}{\pi}\text{rad}(K)\Omega - \frac{1}{2} \rceil$ and define $\cos_\Omega : \mathbb{R} \to [-1, 1]$ by

$$\cos_\Omega(x) := \begin{cases} \cos(x) & x \in [-\frac{1}{2}(2L+1)\pi, \frac{1}{2}(2L+1)\pi], \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Moreover, introduce the functions

$$\begin{aligned} F_{\alpha,\Omega}(y, w, u) &:= \frac{\alpha}{(2\pi)^N} f(y) \cos_\Omega(u) \prod_{j=1}^N \phi(w(j)/\Omega), \\ b_\alpha(y, w, u) &:= -\alpha(\langle w, y \rangle + u) \end{aligned} \tag{6}$$

where $y, w \in \mathbb{R}^N$, $u \in \mathbb{R}$, and $\phi = A * A$ for *any* even function $A \in C^\infty(\mathbb{R})$ supported on $[-\frac{1}{2}, \frac{1}{2}]$ s.t. $\|A\|_2 = 1$. Then, we have the following lemma, a detailed proof of which can be found in Section 3.5.2.

Lemma 4. Let $f \in C_c(\mathbb{R}^N)$ and $\rho \in L^1(\mathbb{R})$ with $K := \text{supp}(f)$ and $\int_{\mathbb{R}} \rho(z)\mathrm{d}z = 1$. Define $F_{\alpha,\Omega}$ and b_α as in Equation 6 for all $\alpha > 0$. Then, for $L := \lceil \frac{2N}{\pi}\text{rad}(K)\Omega - \frac{1}{2} \rceil$, we have

$$f(x) = \lim_{\Omega \to \infty} \lim_{\alpha \to \infty} \int_{K(\Omega)} F_{\alpha,\Omega}(y, w, u)\rho\big(\alpha\langle w, x \rangle + b_\alpha(y, w, u)\big)\mathrm{d}y\mathrm{d}w\mathrm{d}u \tag{7}$$

uniformly for every $x \in K$, where $K(\Omega) := K \times [-\Omega, \Omega]^N \times [-\frac{\pi}{2}(2L+1), \frac{\pi}{2}(2L+1)]$.

The next step in the proof of Theorem 5 is to approximate the integral in Equation 7 using the Monte-Carlo method. Define $v_k := \frac{\text{vol}(K(\Omega))}{n} F_{\alpha,\Omega}\left(y_k, \frac{w_k}{\alpha}, u_k\right)$ for $k = 1, \ldots, n$, and the random variables $\{f_n\}_{n=1}^\infty$ by

$$f_n(x) := \sum_{k=1}^n v_k \rho\big(\langle w_k, x \rangle + b_k\big). \tag{8}$$

Then, we have the following lemma that is proven in Section 3.5.3.

Lemma 5. Let $f \in C_c(\mathbb{R}^N)$ and $\rho \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ with $K := \text{supp}(f)$ and $\int_{\mathbb{R}} \rho(z)\mathrm{d}z = 1$. Then, as $n \to \infty$, we have

$$\mathbb{E} \int_K \left| \int_{K(\Omega)} F_{\alpha,\Omega}(y, w, u)\rho\big(\alpha\langle w, x \rangle + b_\alpha(y, w, u)\big)\mathrm{d}y\mathrm{d}w\mathrm{d}u - f_n(x) \right|^2 \mathrm{d}x = O(1/n), \tag{9}$$

where $K(\Omega) := K \times [-\Omega, \Omega]^N \times [-\frac{\pi}{2}(2L+1), \frac{\pi}{2}(2L+1)]$ and $L := \lceil \frac{2N}{\pi}\text{rad}(K)\Omega - \frac{1}{2} \rceil$.

To complete the proof of Theorem 5, we combine the limit representation (Equation 7) with the Monte-Carlo error guarantee (Equation 9) and show that, given any $\varepsilon > 0$, there exist $\alpha, \Omega > 0$ such that

$$\mathbb{E}\int_K |f(x) - f_n(x)|^2 \mathrm{d}x \leq \varepsilon + O(1/n)$$

as $n \to \infty$. To this end, let $\varepsilon' > 0$ be arbitrary and consider the integral $I(x; p)$ given by

$$I(x;p) := \int_{K(\Omega)} \Big(F_{\alpha,\Omega}(y,w,u)\rho\big(\alpha\langle w,x\rangle + b_\alpha(y,w,u)\big)\Big)^p \mathrm{d}y\mathrm{d}w\mathrm{d}u \tag{10}$$

for $x \in K$ and $p \in \mathbb{N}$. By Equation 7, there exist $\alpha, \Omega > 0$ such that $|f(x) - I(x; 1)| < \varepsilon'$ holds for every $x \in K$, and so it follows that

$$\big|f(x) - f_n(x)\big| < \varepsilon' + \big|I(x;1) - f_n(x)\big|$$

for every $x \in K$. Jensen's inequality now yields that

$$\mathbb{E}\int_K |f(x) - f_n(x)|^2 \mathrm{d}x \leq 2\mathrm{vol}(K)(\varepsilon')^2 + 2\mathbb{E}\int_K \big|I(x;1) - f_n(x)\big|^2 \mathrm{d}x. \tag{11}$$

By Equation 9, we know that the second term on the right-hand side of Equation 11 is $O(1/n)$. Therefore, we have

$$\mathbb{E}\int_K |f(x) - f_n(x)|^2 \mathrm{d}x \leq 2\mathrm{vol}(K)(\varepsilon')^2 + O(1/n),$$

and so the proof is completed by taking $\varepsilon' = \sqrt{\varepsilon/2\mathrm{vol}(K)}$ and choosing $\alpha, \Omega > 0$ accordingly.

3.1.2 Proof of Theorem 1 when $\rho' \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$

The full statement of the theorem is identical to that of Theorem 5 albeit now with $\rho' \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$, so we omit it for brevity. Its proof is also similar to the proof of the case where $\rho \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ with some key modifications. Namely, one uses an integration by parts argument to modify the part of the proof corresponding to Lemma 4. The details of this argument are presented in Section 3.5.4.

3.2 Proof of Theorem 2

In this section, we prove the non-asymptotic result for RVFL networks in $\mathbb{R}^N$, and we begin with a more precise statement of the theorem that makes all the dimensional dependencies explicit.

Theorem 6. Consider the hypotheses of Theorem 5 and suppose further that ρ is κ-Lipschitz on $\mathbb{R}$ for some $\kappa > 0$. For any

$$0 < \delta < \frac{\sqrt{\varepsilon}}{8\sqrt{2N}\kappa\alpha^2 M\Omega(\Omega/\pi)^N \mathrm{vol}^{3/2}(K)(\pi + 2N\mathrm{rad}(K)\Omega)},$$

suppose

$$n \geq \frac{c\Sigma\alpha(\Omega/\pi)^N(\pi + 2N\mathrm{rad}(K)\Omega)\log(3\eta^{-1}\mathcal{N}(\delta,K))}{\sqrt{\varepsilon}\log\big(1 + \frac{\sqrt{\varepsilon}}{\Sigma\alpha(\Omega/\pi)^N(\pi+2N\mathrm{rad}(K)\Omega)}\big)},$$

where $M := \sup_{x\in K}|f(x)|$, $c > 0$ is a numerical constant, and Σ is a constant depending on f and ρ, and let parameters $\{w_k\}_{k=1}^n$, $\{b_k\}_{k=1}^n$, and $\{v_k\}_{k=1}^n$ be as in Theorem 5. Then, the RVFL network defined by

$$f_n(x) := \sum_{k=1}^n v_k \rho(\langle w_k, x\rangle + b_k) \quad \text{for } x \in K$$

satisfies

$$\int_K |f(x) - f_n(x)|^2 \mathrm{d}x < \varepsilon$$

with probability at least $1 - \eta$.

Proof. Let $f \in C_c(\mathbb{R}^N)$ with $K := \mathrm{supp}(f)$ and suppose $\varepsilon > 0$, $\eta \in (0,1)$ are fixed. Take an arbitrarily κ-Lipschitz activation function $\rho \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$. We wish to show that there exists an RVFL network $\{f_n\}_{n=1}^\infty$ defined on K that satisfies the

$$\int_K |f(x) - f_n(x)|^2 \mathrm{d}x < \varepsilon$$

with probability at least $1 - \eta$ when n is chosen sufficiently large. The proof is obtained by modifying the proof of Theorem 5 for the asymptotic case.

We begin by repeating the first two steps in the proof of Theorem 5 from Sections 3.5.1, 3.5.2. In particular, by Lemma 4 we have the representation given by Equation 4, namely,

$$f(x) = \lim_{\Omega\to\infty}\lim_{\alpha\to\infty}\int_{K(\Omega)} F_{\alpha,\Omega}(y,w,u)\rho\big(\alpha\langle w,x\rangle + b_\alpha(y,w,u)\big)\mathrm{d}y\mathrm{d}w\mathrm{d}u$$

holds uniformly for all $x \in K$. Hence, if we define the random variables f_n and I_n from Section 3.5.3 as in Equations 8, 29, respectively, we seek a uniform bound on the quantity

$$|f(x) - f_n(x)| \leq |f(x) - I(x;1)| + |I_n(x) - I(x;1)|$$

over the compact set K, where $I(x; 1)$ is given by Equation 10 for all $x \in K$. Since Equation 7 allows us to fix $\alpha, \Omega > 0$ such that

$$|f(x) - I(x;1)| = \Big|f(x) - \int_{K(\Omega)} F_{\alpha,\Omega}(y,w,u)\rho\big(\alpha\langle w,x\rangle + b_\alpha(y,w,u)\big)\mathrm{d}y\mathrm{d}w\mathrm{d}u\Big| < \sqrt{\frac{\varepsilon}{2\mathrm{vol}(K)}}$$

holds for every $x \in K$ simultaneously, the result would follow if we show that, with high probability, $|I_n(x) - I(x;1)| < \sqrt{\varepsilon/2\mathrm{vol}(K)}$ uniformly for all $x \in K$. Indeed, this would yield

$$\int_K |f(x) - f_n(x)|^2 \mathrm{d}x \leq 2\int_K |f(x) - I(x;1)|^2 \mathrm{d}x + 2\int_K |I_n(x) - I(x;1)|^2 \mathrm{d}x < \varepsilon$$

with high probability. To this end, for $\delta > 0$, let $\mathcal{C}(\delta, K) \subset K$ denote a minimal δ-net for K, with cardinality $\mathcal{N}(\delta, K)$. Now, fix $x \in K$ and consider the inequality

$$|I_n(x) - I(x;1)| \le \underbrace{|I_n(x) - I_n(z)|}_{(*)} + \underbrace{|I_n(z) - I(z;1)|}_{(**)} + \underbrace{|I(x;1) - I(z;1)|}_{(***)}, \tag{12}$$

where $z \in \mathcal{C}(\delta, K)$ is such that $\|x - z\|_2 < \delta$. We will obtain the desired bound on Equation 12 by bounding each of the terms $(*)$, $(**)$, and $(***)$ separately.

First, we consider the term $(*)$. Recalling the definition of I_n, observe that we have

$$\begin{aligned}(*) &= \frac{\mathrm{vol}(K(\Omega))}{n}\Big|\sum_{k=1}^{n} F_{\alpha,\Omega}(y_k,w_k,u_k)\Big(\rho\big(\alpha\langle w_k,x\rangle + b_\alpha(y_k,w_k,u_k)\big) \\ &\quad - \rho\big(\alpha\langle w_k,z\rangle + b_\alpha(y_k,w_k,u_k)\big)\Big)\Big| \\ &\le \frac{\alpha M \mathrm{vol}(K(\Omega))}{(2\pi)^N n}\sum_{k=1}^{n}\Big|\rho\big(\alpha\langle w_k,x\rangle + b_\alpha(y_k,w_k,u_k)\big) \\ &\quad - \rho\big(\alpha\langle w_k,z\rangle + b_\alpha(y_k,w_k,u_k)\big)\Big| \\ &\le \alpha M(2\pi)^{-N}\mathrm{vol}(K(\Omega))R_{\alpha,\Omega}(x,z),\end{aligned}$$

where $M := \sup_{x\in K}|f(x)|$ and we define

$$R_{\alpha,\Omega}(x,z) := \sup_{\substack{y\in K \\ w\in[-\Omega,\Omega]^N \\ u\in[-(L+\frac{1}{2})\pi,(L+\frac{1}{2})\pi]}} \Big|\rho\big(\alpha\langle w,x\rangle + b_\alpha(y,w,u)\big) - \rho\big(\alpha\langle w,z\rangle + b_\alpha(y,w,u)\big)\Big|.$$

Now, since ρ is assumed to be κ-Lipschitz, we have

$$\begin{aligned}&\Big|\rho\big(\alpha\langle w,x\rangle + b_\alpha(y,w,u)\big) - \rho\big(\alpha\langle w,z\rangle + b_\alpha(y,w,u)\big)\Big| \\ &\quad = \Big|\rho\Big(\alpha\big(\langle w,x-y\rangle - u\big)\Big) - \rho\Big(\alpha\big(\langle w,z-y\rangle - u\big)\Big)\Big| \le \kappa\alpha\big|\langle w,x-z\rangle\big|\end{aligned}$$

for any $y \in K$, $w \in [-\Omega,\Omega]^N$, and $u \in [-(L+\frac{1}{2})\pi,(L+\frac{1}{2})\pi]$. Hence, an application of the Cauchy–Schwarz inequality yields $R_{\alpha,\Omega}(x,z) \le \kappa\alpha\Omega\delta\sqrt{N}$ for all $x \in K$, from which it follows that

$$(*) \le M\sqrt{N}\kappa\delta\alpha^2\Omega(2\pi)^{-N}\mathrm{vol}(K(\Omega)) \tag{13}$$

holds for all $x \in K$.

Next, we bound $(***)$ using a similar approach. Indeed, by the definition of $I(\,\cdot\,;1)$, we have

$$\begin{aligned}(***) &= \Big|\int_{K(\Omega)} F_{\alpha,\Omega}(y,w,u)\Big(\rho\big(\alpha\langle w,x\rangle + b_\alpha(y,w,u)\big) \\ &\quad - \rho\big(\alpha\langle w,z\rangle + b_\alpha(y,w,u)\big)\Big)\mathrm{d}y\mathrm{d}w\mathrm{d}u\Big| \\ &\le \frac{\alpha M\|\phi\|_\infty^N}{(2\pi)^N}\int_{K(\Omega)}\Big|\rho\big(\alpha\langle w,x\rangle + b_\alpha(y,w,u)\big) \\ &\quad - \rho\big(\alpha\langle w,z\rangle + b_\alpha(y,w,u)\big)\Big|\mathrm{d}y\mathrm{d}w\mathrm{d}u \\ &\le \alpha M(2\pi)^{-N}\mathrm{vol}(K(\Omega))R_{\alpha,\Omega}(x,z).\end{aligned}$$

Using the fact that $R_{\alpha,\Omega}(x,z) \le \kappa\alpha\Omega\delta\sqrt{N}$ for al $x \in K$, it follows that

$$(***) \le M\sqrt{N}\kappa\delta\alpha^2\Omega(2\pi)^{-N}\mathrm{vol}(K(\Omega)) \tag{14}$$

holds for all $x \in K$, just like Equation 13.

Notice that the Equations 13, 14 are deterministic. In fact, both can be controlled by choosing an appropriate value for δ in the net $\mathcal{C}(\delta, K)$. To see this, fix $\varepsilon' > 0$ arbitrarily and recall that $\mathrm{vol}(K(\Omega)) = (2\Omega)^N\pi(2L+1)\mathrm{vol}(K)$. A simple computation then shows that $(*) + (***) < \varepsilon'$ whenever

$$\begin{aligned}\delta &< \frac{\varepsilon'}{4\sqrt{N}\kappa\alpha^2 M\Omega(\Omega/\pi)^N\mathrm{vol}(K)(\pi + 2N\mathrm{rad}(K)\Omega)} \\ &< \frac{\varepsilon'}{2\sqrt{N}\kappa\alpha^2 M\Omega(\Omega/\pi)^N\pi(2L+1)\mathrm{vol}(K)}.\end{aligned} \tag{15}$$

We now bound $(**)$ uniformly for $x \in K$. Unlike $(*)$ and $(***)$, we cannot bound this term deterministically. In this case, however, we may apply Lemma 1 to

$$g_z(y,w,u) := F_{\alpha,\Omega}(y,w,u)\rho\big(\alpha\langle w,z\rangle + b_\alpha(y,w,u)\big),$$

for any $z \in \mathcal{C}(\delta, K)$. Indeed, $g_z \in L^2(K(\Omega))$ because $F_{\alpha,\Omega} \in L^2(K(\Omega))$ and $\rho \in L^\infty(\mathbb{R})$. Then, Lemma 1 yields the tail bound

$$\begin{aligned}\mathbb{P}\big((**) \ge t\big) &= \mathbb{P}\Big(|I_n(g_z,K(\Omega)) - I(g_z,K(\Omega))| \ge t\Big) \\ &\le 3\exp\Big(-\frac{nt}{Bc}\log\big(1 + \frac{Bt}{\mathrm{vol}(K(\Omega))I(g_z^2,K(\Omega))}\big)\Big) \\ &= 3\exp\Big(-\frac{nt}{Bc}\log\big(1 + \frac{Bt}{\mathrm{vol}(K(\Omega))I(z;2)}\big)\Big)\end{aligned}$$

for all $t > 0$, where $c > 0$ is a numerical constant and

$$\begin{aligned}B &:= 2\alpha M(\Omega/\pi)^N(\pi + 2N\mathrm{rad}(K)\Omega)\|\rho\|_\infty\mathrm{vol}(K) \\ &\ge \alpha M(\Omega/\pi)^N\pi(2L+1)\|\rho\|_\infty\mathrm{vol}(K) \\ &= \alpha M(2\pi)^{-N}\|\rho\|_\infty\mathrm{vol}(K(\Omega)) \\ &\ge \max_{z\in\mathcal{C}(\delta,K)}\|g_z\|_\infty\mathrm{vol}(K(\Omega)).\end{aligned}$$

By taking

$$C := 2M\|\rho\|_\infty\mathrm{vol}(K) \quad \text{and} \quad \Sigma := 2C\sqrt{2\mathrm{vol}(K)},$$

we obtain $B = C\alpha(\Omega/\pi)^N(\pi + 2N\text{rad}(K)\Omega)$ and

$$\max_{z\in\mathcal{C}(\delta,K)} \text{vol}(K(\Omega))I(z;2) \leq \left(\alpha M(2\pi)^{-N}\|\rho\|_\infty \text{vol}(K(\Omega))\right)^2 \leq B^2.$$

If we choose the number of nodes such that

$$n \geq \frac{Bc\log(3\eta^{-1}\mathcal{N}(\delta,K))}{t\log(1+t/B)}, \tag{16}$$

then a union bound yields $(**) < t$ simultaneously for all $z \in \mathcal{C}(\delta,K)$ with probability at least $1-\eta$. Combined with the bounds from Equations 13, 14, it follows from Equation 12 that

$$|I_n(x) - I(x;1)| < \varepsilon' + t$$

simultaneously for all $x \in K$ with probability at least $1-\eta$, provided δ and n satisfy Equations 15, 16, respectively. Since we require $|I_n(x) - I(x;1)| < \sqrt{\varepsilon/2\text{vol}(K)}$, the proof is then completed by setting $\varepsilon' + t = \sqrt{\varepsilon/2\text{vol}(K)}$ and choosing δ and n accordingly. In particular, it suffices to choose $\varepsilon' = t = \frac{1}{2}\sqrt{\varepsilon/2\text{vol}(K)} = C\sqrt{\varepsilon}/\Sigma$, so that Equations 15, 16 become

$$\delta < \frac{\sqrt{\varepsilon}}{8\sqrt{2N}\kappa\alpha^2 M\Omega(\Omega/\pi)^N \text{vol}^{3/2}(K)(\pi + 2N\text{rad}(K)\Omega)},$$
$$n \geq \frac{c\Sigma\alpha(\Omega/\pi)^N(\pi + 2N\text{rad}(K)\Omega)\log(3\eta^{-1}\mathcal{N}(\delta,K))}{\sqrt{\varepsilon}\log\left(1 + \frac{\sqrt{\varepsilon}}{\Sigma\alpha(\Omega/\pi)^N(\pi+2N\text{rad}(K)\Omega)}\right)},$$

as desired.

Remark 2. The implication of Theorem 6 is that, given a desired accuracy level $\varepsilon > 0$, one can construct a RVFL network f_n that is ε-close to f with high probability, provided the number of nodes n in the neural network is sufficiently large. In fact, if we assume that the ambient dimension N is fixed here, then δ and n depend on the accuracy ε and probability η as

$$\delta \lesssim \sqrt{\varepsilon} \quad \text{and} \quad n \gtrsim \frac{\log(\eta^{-1}\mathcal{N}(\delta,K))}{\sqrt{\varepsilon}\log\left(1+\sqrt{\varepsilon}\right)}.$$

Using that $\log(1+x) = x + O(x^2)$ for small values of x, the requirement on the number of nodes behaves like

$$n \gtrsim \frac{\log\left(\eta^{-1}\mathcal{N}\left(\sqrt{\varepsilon},K\right)\right)}{\varepsilon}$$

whenever ε is sufficiently small. Using a simple bound on the covering number, this yields a coarse estimate of $n \gtrsim \varepsilon^{-1}\log(\eta^{-1}/\varepsilon)$.

Remark 3. If we instead assume that N is variable, then, under the assumption that f is Hölder continuous with exponent β, one should expect that $n = \omega(N^{2\beta N})$ as $N \to \infty$ (in light of Remark 10 and in conjunction with Theorem 6 with $\log(1 + 1/x) \approx 1/x$ for large x). In other words, the number of nodes required in the hidden layer is superexponential in the dimension. This dependence of n on N may be improved by means of more refined proof techniques. As for α, if follows from Remark 12 that $\alpha = \Theta(1)$ as $N \to \infty$ provided $\int_{\mathbb{R}}|\nu\rho(\nu)|\text{d}\nu < \infty$.

Remark 4. The κ-Lipschitz assumption on the activation function ρ may likely be removed. Indeed, since $(***)$ in Equation 12 can be bounded instead by leveraging continuity of the L^1 norm with respect to translation, the only term whose bound depends on the Lipschitz property of ρ is $(*)$. However, the randomness in I_n (that we did not use to obtain the bound in Equation 13) may be enough to control $(*)$ in most cases. Indeed, to bound $(*)$, we require control over quantities of the form $\left|\rho\left(\alpha\left(\langle w_k, x - y_k\rangle - u_k\right)\right) - \rho\left(\alpha\left(\langle w_k, z - y_k\rangle - u_k\right)\right)\right|$. For most practical realizations of ρ, this difference will be small with high probability (on the draws of y_k, w_k, u_k), whenever $\|x - z\|_2$ is sufficiently small.

3.3 Results on sub-manifolds of Euclidean space

The constructions of RVFL networks presented in Theorems 5, 6 depend heavily on the dimension of the ambient space $\mathbb{R}^N$. Indeed, the random variables used to construct the input-to-hidden layer weights and biases for these neural networks are N-dimensional objects; moreover, it follows from Equations 15, 16 that the lower bound on the number n of nodes in the hidden layer depends superexponentially on the ambient dimension N. If the ambient dimension is small, these dependencies do not present much of a problem. However, many modern applications require the ambient dimension to be large. Fortunately, a common assumption in practice is that signals of interest have (e.g., manifold) structure that effectively reduces their complexity. Good theoretical results and algorithms in a number of settings typically depend on this induced smaller dimension rather than the ambient dimension. For this reason, it is desirable to obtain approximation results for RVFL networks that leverage the underlying structure of the signal class of interest, namely, the domain of $f \in C_c(\mathbb{R}^N)$.

One way to introduce lower-dimensional structure in the context of RVFL networks is to assume that $\text{supp}(f)$ lies on a subspace of $\mathbb{R}^N$. More generally, and motivated by applications, we may consider the case where $\text{supp}(f)$ is actually a submanifold of $\mathbb{R}^N$. To this end, for the remainder of this section, we assume $\mathcal{M} \subset \mathbb{R}^N$ to be a smooth, compact d-dimensional manifold and consider the problem of approximating functions $f \in C(\mathcal{M})$ using RVFL networks. As we are going to see, RVFL networks in this setting yield theoretical guarantees that replace the dependencies of Theorems 5, 6 on the ambient dimension N with dependencies on the manifold dimension d. Indeed, one should expect that the random variables $\{w_k\}_{k=1}^n$, $\{b_k\}_{k=1}^n$ are essentially d-dimensional objects (rather than N-dimensional) and that the lower bound on the number of network nodes in Theorem 6 scales as a (superexponential) function of d rather than N.

3.3.1 Adapting RVFL networks to d-manifolds

As in Section 2.2, let $\{(U_j,\phi_j)\}_{j\in J}$ be an atlas for the smooth, compact d-dimensional manifold $\mathcal{M} \subset \mathbb{R}^N$ with the corresponding compactly supported partition of unity $\{\eta_j\}_{j\in J}$. Since $\mathcal{M}$ is compact, we assume without loss of generality that $|J| < \infty$. Indeed, if we additionally assume that $\mathcal{M}$ satisfies the property that there exists an $r > 0$ such that, for each $x \in \mathcal{M}$, $\mathcal{M} \cap B_2^N(x,r)$

is diffeomorphic to an ℓ_2 ball in $\mathbb{R}^d$ with diffeomorphism close to the identity, then one can choose an atlas $\{(U_j, \phi_j)\}_{j\in J}$ with $|J| \lesssim 2^d T_d \mathrm{vol}(\mathcal{M}) r^{-d}$ by intersecting $\mathcal{M}$ with ℓ_2 balls in $\mathbb{R}^N$ of radii $r/2$ [40]. Here, T_d is the so-called thickness of the covering and there exist coverings such that $T_d \lesssim d\log(d)$.

Now, for $f \in C(\mathcal{M})$, Lemma 2 implies that

$$f(x) = \sum_{\{j\in J\,:\,x\in U_j\}} (\hat{f}_j \circ \phi_j)(x) \tag{17}$$

for all $x \in \mathcal{M}$, where

$$\hat{f}_j(z) := \begin{cases} f(\phi_j^{-1}(z))\,\eta_j(\phi_j^{-1}(z)) & z \in \phi_j(U_j) \\ 0 & \text{otherwise.} \end{cases}$$

As we will see, the fact that $\mathcal{M}$ is smooth and compact implies $\hat{f}_j \in C_c(\mathbb{R}^d)$ for each $j \in J$, and so we may approximate each $\hat{f}_j$ using RVFL networks on $\mathbb{R}^d$ as in Theorems 5, 6. In this way, it is reasonable to expect that f can be approximated on $\mathcal{M}$ using a linear combination of these low-dimensional RVFL networks. More precisely, we propose approximating f on $\mathcal{M}$ via the following process:

1. For each $j \in J$, approximate $\hat{f}_j$ uniformly on $\phi_j(U_j) \subset \mathbb{R}^d$ using a RVFL network $\tilde{f}_{n_j}$ as in Theorems 5, 6;
2. Approximate f uniformly on $\mathcal{M}$ by summing these RVFL networks over J, i.e.,

$$f(x) \approx \sum_{\{j\in J\,:\,x\in U_j\}} (\tilde{f}_{n_j} \circ \phi_j)(x)$$

for all $x \in \mathcal{M}$.

3.3.2 Main results on d-manifolds

We now prove approximation results for the manifold RVFL network architecture described in Section 3.3.1. For notational clarity, from here onward, we use $\lim_{\{n_j\}_{j\in J}\to\infty}$ to denote the limit as each n_j tends to infinity simultaneously. The first theorem that we prove is an asymptotic approximation result for continuous functions on manifolds using the RVFL network construction presented in Section 3.3.1. This theorem is the manifold-equivalent of Theorem 5.

Theorem 7. Let $\mathcal{M} \subset \mathbb{R}^N$ be a smooth, compact d-dimensional manifold with finite atlas $\{(U_j, \phi_j)\}_{j\in J}$ and $f \in C(\mathcal{M})$. Fix any activation function $\rho \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ with $\int_{\mathbb{R}} \rho(z)\mathrm{d}z = 1$. For any $\varepsilon > 0$, there exist constants $\alpha_j, \Omega_j > 0$ for each $j \in J$ such that the following holds. If, for each $j \in J$ and for $k \in \mathbb{N}$, the random variables

$$\begin{aligned} w_k^{(j)} &\sim \mathrm{Unif}([-\alpha_j\Omega_j, \alpha_j\Omega_j]^d); \\ y_k^{(j)} &\sim \mathrm{Unif}(\phi_j(U_j)); \\ u_k^{(j)} &\sim \mathrm{Unif}([-\tfrac{\pi}{2}(2L_j+1), \tfrac{\pi}{2}(2L_j+1)]), \\ &\quad \text{where } L_j := \lceil \tfrac{2d}{\pi}\mathrm{rad}(\phi_j(U_j))\Omega_j - \tfrac{1}{2}\rceil, \end{aligned}$$

are independently drawn from their associated distributions, and

$$b_k^{(j)} := -\langle w_k^{(j)}, y_k^{(j)}\rangle - \alpha_j u_k^{(j)},$$

then there exist hidden-to-output layer weights $\{v_k^{(j)}\}_{k=1}^{n_j} \subset \mathbb{R}$ such that the sequences of RVFL networks $\{\tilde{f}_{n_j}\}_{n_j=1}^\infty$ defined by

$$\tilde{f}_{n_j}(z) := \sum_{k=1}^{n_j} v_k^{(j)} \rho\big(\langle w_k^{(j)}, z\rangle + b_k^{(j)}\big), \quad \text{for } z \in \phi_j(U_j)$$

satisfy

$$\mathbb{E}\int_{\mathcal{M}} \left| f(x) - \sum_{\{j\in J\,:\,x\in U_j\}} (\tilde{f}_{n_j} \circ \phi_j)(x)\right|^2 \mathrm{d}x \le \varepsilon + O(1/\min_{j\in J} n_j)$$

as $\{n_j\}_{j\in J} \to \infty$.

Proof. We wish to show that there exist sequences of RVFL networks $\{\tilde{f}_{n_j}\}_{n_j=1}^\infty$ defined on $\phi_j(U_j)$ for each $j \in J$, which together satisfy the asymptotic error bound

$$\mathbb{E}\int_{\mathcal{M}} \left| f(x) - \sum_{\{j\in J\,:\,x\in U_j\}} (\tilde{f}_{n_j} \circ \phi_j)(x)\right|^2 \mathrm{d}x \le \varepsilon + O(1/\min_{j\in J} n_j)$$

as $\{n_j\}_{j\in J} \to \infty$. We will do so by leveraging the result of Theorem 5 on each $\phi_j(U_j) \subset \mathbb{R}^d$.

To begin, recall that we may apply the representation given by Equation 17 for f on each chart (U_j, ϕ_j); the RVFL networks $\tilde{f}_{n_j}$ we seek are approximations of the functions $\hat{f}_j$ in this expansion. Now, as $\mathrm{supp}(\eta_j) \subset U_j$ is compact for each $j \in J$, it follows that each set $\phi_j(\mathrm{supp}(\eta_j))$ is a compact subset of $\mathbb{R}^d$. Moreover, because $\hat{f}_j(z) \neq 0$ if and only if $z \in \phi_j(U_j)$ and $\phi_j^{-1}(z) \in \mathrm{supp}(\eta_j) \subset U_j$, we have that $\hat{f}_j = \hat{f}_j|_{\phi_j(\mathrm{supp}(\eta_j))}$ is supported on a compact set. Hence, $\hat{f}_j \in C_c(\mathbb{R}^d)$ for each $j \in J$, and so we may apply Lemma 4 to obtain the uniform limit representation given by Equation 7 on $\phi_j(U_j)$, that is,

$$\hat{f}_j(z) = \lim_{\Omega_j\to\infty}\lim_{\alpha_j\to\infty} \int_{K(\Omega_j)} F_{\alpha_j,\Omega_j}(y,w,u)\rho\big(\alpha_j\langle w, z\rangle + b_{\alpha_j}(y,w,u)\big)\mathrm{d}y\mathrm{d}w\mathrm{d}u,$$

where we define

$$K(\Omega_j) := \phi_j(U_j) \times [-\Omega_j, \Omega_j]^d \times [-\tfrac{\pi}{2}(2L_j+1), \tfrac{\pi}{2}(2L_j+1)].$$

In this way, the asymptotic error bound that is the final result of Theorem 5, namely

$$\mathbb{E}\int_{\phi_j(U_j)} |\hat{f}_j(z) - \tilde{f}_{n_j}(z)|^2 \mathrm{d}z \le \varepsilon_j + O(1/n_j) \tag{18}$$

holds. With these results in hand, we may now continue with the main body of the proof.

Since the representation given by Equation 17 for f on each chart (U_j, ϕ_j) yields

$$\left| f(x) - \sum_{\{j\in J\,:\,x\in U_j\}} (\tilde{f}_{n_j}\circ\phi_j)(x)\right| \le \sum_{\{j\in J\,:\,x\in U_j\}} \left|(\hat{f}_j\circ\phi_j)(x) - (\tilde{f}_{n_j}\circ\phi_j)(x)\right|$$

for all $x \in \mathcal{M}$, Jensen's inequality allows us to bound the mean square error of our RVFL approximation by

$$\begin{aligned}\mathbb{E}\int_{\mathcal{M}}\left|f(x) - \sum_{\{j\in J:\, x\in U_j\}}(\tilde{f}_{n_j}\circ\phi_j)(x)\right|^2 \mathrm{d}x \\ \leq |J|\cdot\underbrace{\mathbb{E}\int_{\mathcal{M}}\sum_{\{j\in J:\, x\in U_j\}}\left|(\hat{f}_j\circ\phi_j)(x) - (\tilde{f}_{n_j}\circ\phi_j)(x)\right|^2\mathrm{d}x}_{(*)}\end{aligned} \tag{19}$$

To bound $(*)$, note that the change of variables given by Equation 2 implies

$$\begin{aligned}\int_{\mathcal{M}}\sum_{\{j\in J:\, x\in U_j\}}\left|(\hat{f}_j\circ\phi_j)(x) - (\tilde{f}_{n_j}\circ\phi_j)(x)\right|^2\mathrm{d}x \\ = \sum_{j\in J}\int_{\phi_j(U_j)}\frac{\left|\hat{f}_j(z) - \tilde{f}_{n_j}(z)\right|^2}{|\det(D\phi_j(\phi_j^{-1}(z)))|}\mathrm{d}z\end{aligned}$$

for each $j \in J$. Defining $\beta_j := \inf_{y\in U_j}|\det(D\phi_j(y))|$, which is necessarily bounded away from zero for each $j \in J$ by compactness of $\mathcal{M}$, we therefore have

$$(*) \leq \sum_{j\in J}\beta_j^{-1}\mathbb{E}\int_{\phi_j(U_j)}\left|\hat{f}_j(z) - \tilde{f}_{n_j}(z)\right|^2\mathrm{d}z.$$

Hence, applying Equation 18 for each $j \in J$ yields

$$(*) \leq \sum_{j\in J}\beta_j^{-1}\left(\varepsilon_j + O(1/n_j)\right) = \sum_{j\in J}\frac{\varepsilon_j}{\beta_j} + O(1/\min_{j\in J} n_j) \tag{20}$$

because $\sum_{j\in J} 1/n_j \leq |J|/\min_{j\in J} n_j$. With the bound given by Equation 20 in hand, Equation 19 becomes

$$\mathbb{E}\int_{\mathcal{M}}\left|f(x) - \sum_{\substack{\{j\in J:\\ x\in U_j\}}}(\tilde{f}_{n_j}\circ\phi_j)(x)\right|^2\mathrm{d}x \leq |J|\sum_{j\in J}\frac{\varepsilon_j}{\beta_j} + O(1/\min_{j\in J} n_j)$$

as $\{n_j\}_{j\in J}\to\infty$, and so the proof is completed by taking each $\varepsilon_j > 0$ in such a way that

$$\varepsilon = |J|\sum_{j\in J}\frac{\varepsilon_j}{\beta_j},$$

and choosing $\alpha_j, \Omega_j > 0$ accordingly for each $j \in J$.

Remark 5. Note that the neural-network architecture obtained in Theorem 7 has the following form in the case of a generic atlas. To obtain the estimate of $f(x)$, the input x is first "pre-processed" by computing $\phi_j(x)$ for each $j \in J$ such that $x \in U_j$, and then put through the corresponding RVFL network. However, using the Geometric Multi-Resolution Analysis approach from Allard et al. [42] (as we do in Section 3.4), one can construct an approximation (in an appropriate sense) of the atlas, with maps ϕ_j being linear. In this way, the pre-processing step can be replaced by the layer computing $\phi_j(x)$, followed by the RVFL layer f_j. We refer the reader to Section 3.4 for the details.

The biggest takeaway from Theorem 7 is that the same asymptotic mean-square error behavior we saw in the RVFL network architecture of Theorem 5 holds for our RVFL-like construction on manifolds, with the added benefit that the input-to-hidden layer weights and biases are now d-dimensional random variables rather than N-dimensional. Provided the size of the atlas $|J|$ is not too large, this significantly reduces the number of random variables that must be generated to produce a uniform approximation of $f \in C(\mathcal{M})$.

One might expect to see a similar reduction in dimension dependence for the non-asymptotic case if the RVFL network construction of Section 3.3.1 is used. Indeed, our next theorem, which is the manifold-equivalent of Theorem 6, makes this explicit:

Theorem 8. Let $\mathcal{M} \subset \mathbb{R}^N$ be a smooth, compact d-dimensional manifold with finite atlas $\{(U_j, \phi_j)\}_{j\in J}$ and $f \in C(\mathcal{M})$. Fix any activation function $\rho \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ such that ρ is κ-Lipschitz on $\mathbb{R}$ for some $\kappa > 0$ and $\int_{\mathbb{R}}\rho(z)\mathrm{d}z = 1$. For any $\varepsilon > 0$, there exist constants $\alpha_j, \Omega_j > 0$ for each $j \in J$ such that the following holds. Suppose, for each $j \in J$ and for $k = 1, ..., n_j$, the random variables

$$\begin{aligned}w_k^{(j)} &\sim \mathrm{Unif}([-\alpha_j\Omega_j, \alpha_j\Omega_j]^d); \\ y_k^{(j)} &\sim \mathrm{Unif}(\phi_j(U_j)); \\ u_k^{(j)} &\sim \mathrm{Unif}([-\tfrac{\pi}{2}(2L_j+1), \tfrac{\pi}{2}(2L_j+1)]), \\ &\quad \text{where } L_j := \lceil \tfrac{2d}{\pi}\mathrm{rad}(\phi_j(U_j))\Omega_j - \tfrac{1}{2}\rceil,\end{aligned}$$

are independently drawn from their associated distributions, and

$$b_k^{(j)} := -\langle w_k^{(j)}, y_k^{(j)}\rangle - \alpha_j u_k^{(j)}.$$

Then, there exist hidden-to-output layer weights $\{v_k^{(j)}\}_{k=1}^{n_j} \subset \mathbb{R}$ such that, for any

$$0 < \delta_j < \frac{\sqrt{\varepsilon}}{8|J|\sqrt{d\mathrm{vol}(\mathcal{M})}\kappa\alpha_j^2 M_j\Omega_j(\Omega_j/\pi)^d\mathrm{vol}(\phi_j(U_j))(\pi + 2d\mathrm{rad}(\phi_j(U_j))\Omega)},$$

and

$$n_j \geq \frac{2c|J|\sqrt{\mathrm{vol}(\mathcal{M})}C^{(j)}\alpha_j(\Omega_j/\pi)^d(\pi+2d\mathrm{rad}(\phi_j(U_j))\Omega_j)\log(3|J|\eta^{-1}\mathcal{N}(\delta_j,\phi_j(U_j)))}{\sqrt{\varepsilon}\log\left(1+\frac{\sqrt{\varepsilon}}{2|J|\sqrt{\mathrm{vol}(\mathcal{M})}C^{(j)}\alpha_j(\Omega_j/\pi)^d(\pi+2d\mathrm{rad}(\phi_j(U_j))\Omega_j)}\right)},$$

where $M_j := \sup_{z\in\phi_j(U_j)}|\hat{f}_j(z)|$, $c > 0$ is a numerical constant, and $C^{(j)} := 2M_j\|\rho\|_\infty\mathrm{vol}(\phi_j(U_j))$, the sequences of RVFL networks $\{\tilde{f}_{n_j}\}_{n_j=1}^\infty$ defined by

$$\tilde{f}_{n_j}(z) := \sum_{k=1}^{n_j} v_k^{(j)}\rho\left(\langle w_k^{(j)}, z\rangle + b_k^{(j)}\right), \quad \text{for } z \in \phi_j(U_j)$$

satisfy

$$\int_{\mathcal{M}}\left|f(x) - \sum_{\{j\in J:\, x\in U_j\}}(\tilde{f}_{n_j}\circ\phi_j)(x)\right|^2\mathrm{d}x < \varepsilon$$

with probability at least $1 - \eta$.

Proof. See Section 3.5.5.

As alluded to earlier, an important implication of Theorems 7, 8 is that the random variables $\{w_k^{(j)}\}_{k=1}^{n_j}$ and $\{b_k^{(j)}\}_{k=1}^{n_j}$ are d-dimensional objects for each $j \in J$. Moreover, bounds for δ_j and n_j now have superexponential dependence on the manifold dimension d instead of the ambient dimension N. Thus, introducing the manifold structure removes the dependencies on the ambient dimension, replacing them instead with the intrinsic dimension of $\mathcal{M}$ and the complexity of the atlas $\{(U_j, \phi_j)\}_{j\in J}$.

Remark 6. The bounds on the covering radii δ_j and hidden layer nodes n_j needed for each chart in Theorem 8 are not optimal. Indeed, these bounds may be further improved if one uses the local structure of the manifold, through quantities such as its curvature and reach. In particular, the appearance of $|J|$ in both bounds may be significantly improved upon if the manifold is locally well-behaved.

3.4 Numerical simulations

In this section, we provide numerical evidence to support the result of Theorem 8. Let $\mathcal{M} \subset \mathbb{R}^N$ be a smooth, compact d-dimensional manifold. Since having access to an atlas for $\mathcal{M}$ is not necessarily practical, we assume instead that we have a suitable approximation to $\mathcal{M}$. For our purposes, we will use a Geometric Multi-Resolution Analysis (GMRA) approximation of $\mathcal{M}$ (see [42]; and also, e.g., [43] for a complete definition).

A GMRA approximation of $\mathcal{M}$ provides a collection $\{(\mathcal{C}_j, \mathcal{P}_j)\}_{j\in\{1,\dots J\}}$ of centers $\mathcal{C}_j = \{c_{j,k}\}_{k=1}^{K_j} \subset \mathbb{R}^N$ and affine projections $\mathcal{P}_j = \{P_{j,k}\}_{k=1}^{K_j}$ on $\mathbb{R}^N$ such that, for each $j \in \{1, \dots, J\}$, the pairs $\{(c_{j,k}, P_{j,k})\}_{k=1}^{K_j}$ define d-dimensional affine spaces that approximate $\mathcal{M}$ with increasing accuracy in the following sense. For every $x \in \mathcal{M}$, there exists $\widetilde{C}_x > 0$ and $k' \in \{1, \dots, K_j\}$ such that

$$\|x - P_{j,k'}x\|_2 \leq \widetilde{C}_x 2^{-j} \tag{21}$$

holds whenever $\|x - c_{j,k'}\|_2$ is sufficiently small.

In this way, a GMRA approximation of $\mathcal{M}$ essentially provides a collection of approximate tangent spaces to $\mathcal{M}$. Hence, a GMRA approximation having fine enough resolution (i.e., large enough j) is a good substitution for an atlas. In practice, one must often first construct a GMRA from empirical data, assumed to be sampled from appropriate distributions on the manifold. Indeed, this is possible, and yields the so-called empirical GMRA, studied in Maggioni et al. [44], where finite-sample error bounds are provided. The main point is that given enough samples on the manifold, one can construct a good GMRA approximation of the manifold.

Let $\{(c_{j,k}, P_{j,k})\}_{k=1}^{K_j}$ be a GMRA approximation of $\mathcal{M}$ for refinement level j. Since the affine spaces defined by $(c_{j,k}, P_{j,k})$ for each $k \in \{1, \dots, K_j\}$ are d-dimensional, we will approximate f on $\mathcal{M}$ by projecting it (in an appropriate sense) onto these affine spaces and approximating each projection using an RVFL network on $\mathbb{R}^d$. To make this more precise, observe that, since each affine projection acts on $x \in \mathcal{M}$ as $P_{j,k}x = c_{j,k} + \Phi_{j,k}(x - c_{j,k})$ for some othogonal projection $\Phi_{j,k} \colon \mathbb{R}^N \to \mathbb{R}^N$, for each $k \in \{1, \dots K_j\}$, we have

$$f(P_{j,k}x) = f\big(c_{j,k} + \Phi_{j,k}(x - c_{j,k})\big) = f\big((I_N - \Phi_{j,k})c_{j,k} + U_{j,k}D_{j,k}V_{j,k}^T x\big),$$

where $\Phi_{j,k} = U_{j,k}D_{j,k}V_{j,k}^T$ is the compact singular value decomposition (SVD) of $\Phi_{j,k}$ (i.e., only the left and right singular vectors corresponding to non-zero singular values are computed). In particular, the matrix of right-singular vectors $V_{j,k} \colon \mathbb{R}^d \to \mathbb{R}^N$ enables us to define a function $\hat{f}_{j,k} \colon \mathbb{R}^d \to \mathbb{R}$, given by

$$\hat{f}_{j,k}(z) := f\big((I_N - \Phi_{j,k})c_{j,k} + U_{j,k}D_{j,k}z\big), \qquad z \in \mathbb{R}^d,$$

which satisfies $\hat{f}_{j,k}(V_{j,k}^T x) = f(P_{j,k}x)$ for all $x \in \mathcal{M}$. By continuity of f and Equation 21, this means that for any $\varepsilon > 0$, there exists $j \in \mathbb{N}$ such that $|f(x) - \hat{f}_{j,k}(V_{j,k}^T x)| < \varepsilon$ for some $k \in \{1, \dots, K_j\}$. For such $k \in \{1, \dots, K_j\}$, we may therefore approximate f on the affine space associated with $(c_{j,k}, P_{j,k})$ by approximating $\hat{f}_{j,k}$ using a RFVL network $\tilde{f}_{n_{j,k}} \colon \mathbb{R}^d \to \mathbb{R}$ of the form

$$\tilde{f}_{n_{j,k}}(z) := \sum_{\ell=1}^{n_{j,k}} v_\ell^{(j,k)} \rho\big(\langle w_\ell^{(j,k)}, z\rangle + b_\ell^{(j,k)}\big), \tag{22}$$

where $\{w_\ell^{(j,k)}\}_{\ell=1}^{n_{j,k}} \subset \mathbb{R}^d$ and $\{b_\ell^{(j,k)}\}_{\ell=1}^{n_{j,k}} \subset \mathbb{R}$ are random input-to-hidden layer weights and biases (resp.) and the hidden-to-output layer weights $\{v_\ell^{(j,k)}\}_{\ell=1}^{n_{j,k}} \subset \mathbb{R}$ are learned. Choosing the activation function ρ and random input-to-hidden layer weights and biases as in Theorem 8 then guarantees that $|f(P_{j,k}x) - \tilde{f}_{n_{j,k}}(V_{j,k}^T x)|$ is small with high probability whenever $n_{j,k}$ is sufficiently large.

In light of the above discussion, we propose the following RVFL network construction for approximating functions $f \in C(\mathcal{M})$: Given a GMRA approximation of $\mathcal{M}$ with sufficiently high resolution j, construct and train RVFL networks of the form given by Equation 22 for each $k \in \{1, \dots, K_j\}$. Then, given $x \in \mathcal{M}$ and $\varepsilon > 0$, choose $k' \in \{1, \dots, K_j\}$ such that

$$c_{j,k'} \in \arg\min_{c_{j,k} \in \mathcal{C}_j} \|x - c_{j,k}\|_2$$

and evaluate $\tilde{f}_{n_{j,k'}}(x)$ to approximate $f(x)$. We summarize this algorithm in Algorithm 1. Since the structure of the GMRA approximation implies $\|x - P_{j,k'}x\|_2 \leq C_x 2^{-2j}$ holds for our choice of $k' \in \{1, \dots, K_j\}$ [see 43], continuity of f and Lemma 5 imply that, for any $\varepsilon > 0$ and j large enough,

$$\begin{aligned}|f(x) - \tilde{f}_{n_{j,k'}}(V_{j,k'}^T x)| \leq |f(x) - \hat{f}_{j,k'}(V_{j,k'}^T x)| + |\hat{f}_{j,k'}(V_{j,k'}^T x) \\ - \tilde{f}_{n_{j,k'}}(V_{j,k'}^T x)| < \varepsilon\end{aligned}$$

holds with high probability, provided $n_{j,k'}$ satisfies the requirements of Theorem 8.

Remark 7. In the RVFL network construction proposed above, we require that the function f be defined in a sufficiently large region around the manifold. Essentially, we need to ensure that f

Given: $f \in C(\mathcal{M})$; GMRA approximation $\{(c_{j,k}, P_{j,k})\}_{k=1}^{K_j}$ of $\mathcal{M}$ at scale j
Output: $y^\sharp \approx f(x)$ for any $x \in \mathcal{M}$

Step 1: For each $k \in \{1, \ldots, K_j\}$, construct and train[1] an RVFL network $\tilde{f}_{n_{j,k}}$ of the form given by Equation 22
Step 2: For any $x \in \mathcal{M}$, find $c_{j,k'} \in \arg\min_{c_{j,k} \in \mathcal{C}_j} \|x - c_{j,k}\|_2$
Step 3: Set $y^\sharp = \tilde{f}_{n_{j,k'}}(x)$

Algorithm 1. Approximation algorithm.

is continuously defined on the set $S := \mathcal{M} \cup \widehat{\mathcal{M}}_j$, where $\widehat{\mathcal{M}}_j$ is the scale-j GMRA approximation

$$\widehat{\mathcal{M}}_j := \{P_{j,k_j(z)} z : \|z\|_2 \leq \mathrm{rad}(\mathcal{M})\} \cap B_2^N(0, \mathrm{rad}(\mathcal{M})).$$

This ensures that f can be evaluated on the affine subspaces given by the GMRA.

To simulate Algorithm 1, we take $\mathcal{M} = \mathbb{S}^2$ embedded in $\mathbb{R}^{20}$ and construct a GMRA up to level $j_{\max} = 15$ using 20,000 data points sampled uniformly from $\mathcal{M}$. Given $j \leq j_{\max}$, we generate RVFL networks $\hat{f}_{n_{j,k}} : \mathbb{R}^2 \to \mathbb{R}$ as in Equation 22 and train them on $V_{j,k}^T(B_2^N(c_{j,k}, r) \cap T_{j,k})$ using the training pairs $\{(V_{k,j}^T x_\ell, f(P_{j,k} x_\ell))\}_{\ell=1}^p$, where $T_{k,j}$ is the affine space generated by $(c_{j,k}, P_{j,k})$. For simplicity, we fix $n_{j,k} = n$ to be constant for all $k \in \{1, \ldots, K_j\}$ and use a single, fixed pair of parameters $\alpha, \Omega > 0$ when constructing all RVFL networks. We then randomly select a test set of 200 points $x \in \mathcal{M}$ for use throughout all experiments. In each experiment (i.e., point in Figure 1), we use Algorithm 1 to produce an approximation $y^\sharp = \tilde{f}_{n_{j,k'}}(x)$ of $f(x)$. Figure 1 displays the mean relative error in these approximations for varying numbers of nodes n; to construct this plot, f is taken to be the exponential $f(x) = \exp(\sum_{k=1}^N x(k))$ and ρ the hyperbolic secant function. Notice that for small numbers of nodes, the RVFL networks are not very good at approximating f, regardless of the choice of $\alpha, \Omega > 0$. However, the error decays as the number of nodes increases until reaching a floor due to error inherent in the GMRA approximation. Hence, as suggested by Theorem 3, to achieve a desired error bound of $\varepsilon > 0$, one needs to only choose a GMRA scale j such that the inherent error in the GMRA (which scales like 2^{-j}) is less than ε, then adjust the parameters α_j, Ω_j, and $n_{j,k}$ accordingly.

Remark 8. As we just mentioned, the error can only decay so far due to the resolution of the GMRA approximation. However, that is not the only floor in our simulation; indeed, the ε in Theorem 3 is determined by the α_j's and Ω_j's, which we kept fixed (see the caption of Figure 1). Consequently, the stagnating accuracy as n increases, as seen in Figure 1, is also predicted by Theorem 3. Since the solid and dashed lines seem to reach the same floor, the floor due to error inherent in the GMRA approximation seems to be the limiting error term for RVFL networks with large numbers of nodes.

1 The construction and training of RVFL networks is left as a "black box" procedure. How to best choose a specific activation function $\rho(z)$ and train each RVFL network given by Equation 22 is outside of the scope of this study. The reader may, for instance, select from the range of methods available for training neural networks.

Remark 9. Utilizing random inner weights and biases resulted in us needing to approximate the atlas to the manifold. To this end, knowing the computational complexity of the GMRA approximation would be useful in practice. As it turns out in Liao and Maggioni [45], calculating the GMRA approximation has computational complexity $O(C^d N m \log(m))$, where m is the number of training data points and $C > 0$ is a numerical constant.

3.5 Proofs of technical lemmas

3.5.1 Proof of Lemma 3

Observe that h_Ω defined in Equation 3 may be viewed as a multidimensional bump function; indeed, the parameter $\Omega > 0$ controls the width of the bump. In particular, if Ω is allowed to grow very large, then h_Ω becomes very localized near the origin. Objects that behave in this way are known in the functional analysis literature as approximate δ-functions:

Definition 2. A sequence of functions $\{\varphi_t\}_{t>0} \subset L^1(\mathbb{R}^N)$ are called *approximate* (or *nascent*) δ-*functions* if

$$\lim_{t\to\infty} \int_{\mathbb{R}^N} \varphi_t(x) f(x) \mathrm{d}x = f(0)$$

for all $f \in C_c(\mathbb{R}^N)$. For such functions, we write $\delta_0(x) = \lim_{t\to\infty} \varphi_t(x)$ for all $x \in \mathbb{R}^N$, where δ_0 denotes the N-dimensional Dirac δ-function centered at the origin.

Given $\varphi \in L^1(\mathbb{R}^N)$ with $\int_{\mathbb{R}^N} \varphi(x)\mathrm{d}x = 1$, one may construct approximate δ-functions for $t > 0$ by defining $\varphi_t(x) := t^N \varphi(tx)$ for all $x \in \mathbb{R}^N$ [46]. Such sequences of approximate δ-functions are also called *approximate identity sequences* [47] since they satisfy a particularly nice identity with respect to convolution, namely, $\lim_{t\to\infty} \|f * \varphi_t - f\|_1 = 0$ for all $f \in C_c(\mathbb{R}^N)$ [see 47, Theorem 6.32]. In fact, such an identity holds much more generally.

Lemma 6. [46, Theorem 1.18] Let $\varphi \in L^1(\mathbb{R}^N)$ with $\int_{\mathbb{R}^N} \varphi(x)\mathrm{d}x = 1$ and for $t > 0$ define $\varphi_t(x) := t^N \varphi(tx)$ for all $x \in \mathbb{R}^N$. If $f \in L^p(\mathbb{R}^N)$ for $1 \leq p < \infty$ (or $f \in C_0(\mathbb{R}^N) \subset L^\infty(\mathbb{R}^N)$ for $p = \infty$), then $\lim_{t\to\infty} \|f * \varphi_t - f\|_p = 0$.

To prove Equation 4, it would suffice to have $\lim_{\Omega\to\infty} \|f * h_\Omega - f\|_\infty = 0$, which is really just Lemma 6 in case $p = \infty$. Nonetheless, we present a proof by mimicking [46] for completeness. Moreover, we will use a part of proof in Remark 10 below.

Lemma 7. Let $h \in L^1(\mathbb{R}^n)$ with $\int_{\mathbb{R}^N} h(x)\mathrm{d}x = 1$ and define $h_\Omega \in L^1(\mathbb{R}^N)$ as in Equation 3 for all $\Omega > 0$. Then, for all $f \in C_0(\mathbb{R}^N)$, we have

$$\lim_{\Omega\to\infty} \sup_{x\in\mathbb{R}^N} \left|(f * h_\Omega)(x) - f(x)\right| = 0.$$

Proof. By symmetry of the convolution operator in its arguments, we have

$$\begin{aligned}\sup_{x\in\mathbb{R}^N} \left|(f * h_\Omega)(x) - f(x)\right| &= \sup_{x\in\mathbb{R}^N} \left| \int_{\mathbb{R}^N} f(y) h_\Omega(x - y)\mathrm{d}y - f(x)\right| \\ &= \sup_{x\in\mathbb{R}^N} \left| \int_{\mathbb{R}^N} f(x - y) h_\Omega(y)\mathrm{d}y - f(x)\right|.\end{aligned}$$

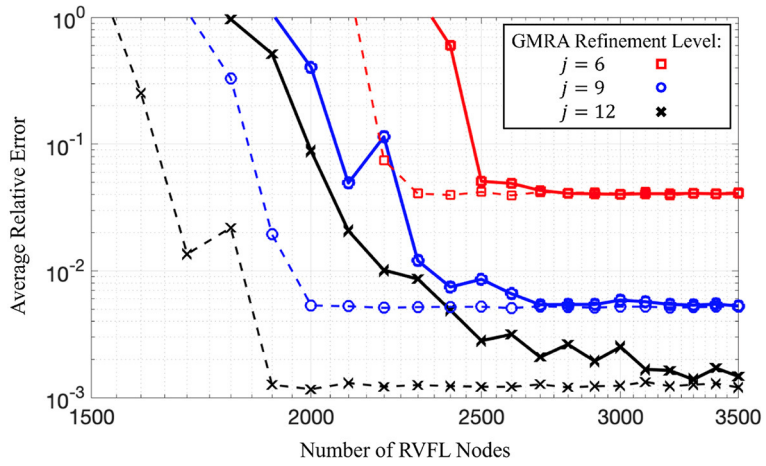

FIGURE 1
Log-scale plot of average relative error for Algorithm 1 as a function of the number of nodes n in each RVFL network. Black (cross), blue (circle), and red (square) lines correspond to GMRA refinement levels $j = 12$, $j = 9$, and $j = 6$ (resp.). For each j, we fix $\alpha_j = 2$ and vary $\Omega_j = 10, 15$ (solid and dashed lines, resp.). Reconstruction error decays as a function of n until reaching a floor due to error in the GMRA approximation of $\mathcal{M}$. The code used to obtain these numerical results is available upon direct request sent to the corresponding author.

Since a simple substitution yields $1 = \int_{\mathbb{R}^N} h(x)\mathrm{d}x = \int_{\mathbb{R}^N} h_\Omega(x)\mathrm{d}x$, it follows that

$$\begin{aligned}\sup_{x\in\mathbb{R}^N} \left|(f * h_\Omega)(x) - f(x)\right| &= \sup_{x\in\mathbb{R}^N} \left| \int_{\mathbb{R}^N} \big(f(x-y) - f(x)\big) h_\Omega(y)\mathrm{d}y \right| \\ &\le \int_{\mathbb{R}^N} |h_\Omega(y)| \sup_{x\in\mathbb{R}^N} \left|f(x) - f(x-y)\right| \mathrm{d}y.\end{aligned}$$

Finally, expanding the function h_Ω, we obtain

$$\begin{aligned}&\sup_{x\in\mathbb{R}^N} \left|(f * h_\Omega)(x) - f(x)\right| \\ &\le \int_{\mathbb{R}^N} \left(\Omega^N |h(\Omega y)|\right) \sup_{x\in\mathbb{R}^N} \left|f(x) - f(x-y)\right| \mathrm{d}y \\ &= \int_{\mathbb{R}^N} |h(z)| \sup_{x\in\mathbb{R}^N} \left|f(x) - f(x - z/\Omega)\right| \mathrm{d}z,\end{aligned}$$

where we have used the substitution $z = \Omega y$. Taking limits on both sides of this expression and observing that

$$\int_{\mathbb{R}^N} |h(z)| \sup_{x\in\mathbb{R}^N} \left|f(x) - f(x - z/\Omega)\right| \mathrm{d}z \le 2\|h\|_1 \sup_{x\in\mathbb{R}^N} |f(x)| < \infty,$$

using the Dominated Convergence Theorem, we obtain

$$\begin{aligned}&\lim_{\Omega\to\infty} \sup_{x\in\mathbb{R}^N} \left|(f * h_\Omega)(x) - f(x)\right| \\ &\quad\le \textstyle\int_{\mathbb{R}^N} |h(z)| \lim_{\Omega\to\infty} \sup_{x\in\mathbb{R}^N} \left|f(x) - f(x - z/\Omega)\right| \mathrm{d}z.\end{aligned}$$

So, it suffices to show that, for all $z \in \mathbb{R}^N$,

$$\lim_{\Omega\to\infty} \sup_{x\in\mathbb{R}^N} \left|f(x) - f(x - z/\Omega)\right| = 0.$$

To this end, let $\varepsilon > 0$ and $z \in \mathbb{R}^N$ be arbitrary. Since $f \in C_0(\mathbb{R}^N)$, there exists $r > 0$ sufficiently large such that $|f(x)| < \varepsilon/2$ for all $x \in \mathbb{R}^N \setminus \overline{B(0,r)}$, where $\overline{B(0,r)} \subset \mathbb{R}^N$ is the closed ball of radius r centered at the origin. Let $\mathcal{B} := \overline{B(0, r + \|z/\Omega\|_2)}$, so that for each $x \in \mathbb{R}^N \setminus \mathcal{B}$ we have both x and $x - z/\Omega$ in $\mathbb{R}^N \setminus \overline{B(0,r)}$. Thus, both $|f(x)| < \varepsilon/2$ and $|f(x - z/\Omega)| < \varepsilon/2$, implying that

$$\sup_{x\in\mathbb{R}^N\setminus\mathcal{B}} \left|f(x) - f(x - z/\Omega)\right| < \varepsilon.$$

Hence, we obtain

$$\begin{aligned}&\lim_{\Omega\to\infty} \sup_{x\in\mathbb{R}^N} \left|f(x) - f(x - z/\Omega)\right| \\ &\quad\le \lim_{\Omega\to\infty} \max\left\{ \sup_{x\in\mathcal{B}} \left|f(x) - f(x - z/\Omega)\right|, \right. \\ &\qquad \left. \sup_{x\in\mathbb{R}^N\setminus\mathcal{B}} \left|f(x) - f(x - z/\Omega)\right| \right\} \\ &\quad\le \max\left\{\varepsilon, \lim_{\Omega\to\infty} \sup_{x\in\mathcal{B}} \left|f(x) - f(x - z/\Omega)\right| \right\}.\end{aligned}$$

Now, as $\mathcal{B}$ is a compact subset of $\mathbb{R}^N$, the continuous function f is uniformly continuous on $\mathcal{B}$, and so the remaining limit and supremum may be freely interchanged, whereby continuity of f yields

$$\lim_{\Omega\to\infty} \sup_{x\in\mathcal{B}} \left|f(x) - f(x - z/\Omega)\right| = \sup_{x\in\mathcal{B}} \lim_{\Omega\to\infty} \left|f(x) - f(x - z/\Omega)\right| = 0.$$

Since $\varepsilon > 0$ may be taken arbitrarily small, we have proved the result.

Remark 10. While Lemma 7 does the approximation we aim for, it gives no indication of how fast

$$\sup_{x\in\mathbb{R}^N} \left|(f * h_\Omega)(x) - f(x)\right|$$

decays in terms of Ω or the dimension N. Assuming $h(z) = g(z(1)) \cdots g(z(N))$ for some non-negative g (which is how we will

choose h in Section 3.5.2) and f to be β-Hölder continuous for some fixed $\beta \in (0,1)$ yields that

$$\begin{aligned}
&\sup_{x\in\mathbb{R}^N} \left|(f * h_\Omega)(x) - f(x)\right| \\
&\leq \int_{\mathbb{R}^N} |h(z)| \sup_{x\in\mathbb{R}^N} \left|f(x) - f(x - z/\Omega)\right| \mathrm{d}z \\
&\lesssim \Omega^{-\beta} \int_{\mathbb{R}^N} \|z\|_2^\beta h(z)\mathrm{d}z \\
&\leq \Omega^{-\beta} \left(\int_{\mathbb{R}^N} \left(z(1)^2 + \cdots + z(N)^2\right) h(z)\mathrm{d}z \right)^{\beta/2} \\
&\leq \Omega^{-\beta} \left(N \max_{j\in\{1,\ldots,N\}} \int_{\mathbb{R}^N} z(j)^2 h(z)\mathrm{d}z \right)^{\beta/2} \\
&= (\sqrt{N}/\Omega)^\beta \left(\max_{j\in\{1,\ldots,N\}} \int_{\mathbb{R}} z(j)^2 g(z(j))\mathrm{d}z(j) \right)^{\beta/2} \\
&= (\sqrt{N}/\Omega)^\beta \left(\int_{\mathbb{R}} z(1)^2 g(z(1))\mathrm{d}z(1) \right)^{\beta/2} \\
&\lesssim (\sqrt{N}/\Omega)^\beta
\end{aligned}$$

where the third inequality follows from Jensen's inequality.

3.5.2 Proof of Lemma 4: the limit-integral representation

Let $A \in C^\infty(\mathbb{R})$ be *any* even function supported on $[-\frac{1}{2}, \frac{1}{2}]$ s.t. $\|A\|_2 = 1$. Then, $\phi = A * A \in C^\infty(\mathbb{R})$ is an even function supported on $[-1,1]$ s.t. $\phi(0) = 1$. Lemma 3 implies that

$$f(x) = \lim_{\Omega\to\infty} (f * h_\Omega)(x) \tag{23}$$

uniformly in $x \in K$ for *any* $h \in L^1(\mathbb{R}^N)$ satisfying $\int_{\mathbb{R}^N} h(z)\mathrm{d}z = 1$. We choose

$$h(z) = \frac{1}{(2\pi)^N} \int_{\mathbb{R}^N} \exp(i\langle w, z\rangle) \prod_{j=1}^N \phi(w(j))\mathrm{d}w$$

which the reader may recognize as the (inverse) Fourier transform of $\prod_{j=1}^N \phi(w(j))$. As we announced in Remark 10, $h(z) = g(z(1)) \cdots g(z(N))$, where (using the convolution theorem)

$$\begin{aligned}
g(z(j)) &= \frac{1}{2\pi} \int_{\mathbb{R}} \exp(iw(j)z(j))\phi(w(j))\mathrm{d}w(j) \\
&= \frac{1}{2\pi} \int_{\mathbb{R}} \exp(iw(j)z(j))(A * A)(w(j))\mathrm{d}w(j) \\
&= 2\pi \left(\frac{1}{2\pi} \int_{\mathbb{R}} \exp(iw(j)z(j))A(w(j))\mathrm{d}w(j) \right)^2 \geq 0
\end{aligned}$$

Moreover, since g is the Fourier transform of an even function, h is real-valued and also even. In addition, since ϕ is smooth, h decays faster than the reciprocal of any polynomial (as follows from repeated integration by parts and the Riemann–Lebesgue lemma), so $h \in L^1(\mathbb{R}^N)$. Thus, Fourier inversion yields

$$\int_{\mathbb{R}^N} h(z)\mathrm{d}z = \int_{\mathbb{R}^N} \exp(-i\langle w, z\rangle)h(z)\mathrm{d}z\Big|_{w=0} = \prod_{j=1}^N \phi(0) = 1,$$

which justifies our application of Lemma 3. Expanding the right-hand side of Equation 23 (using the scaling property of the Fourier transform) yields that

$$\begin{aligned}
(f * h_\Omega)(x) &= \int_{\mathbb{R}^N} f(y)h_\Omega(x-y)\mathrm{d}y \\
&= \frac{1}{(2\pi)^N} \int_K f(y) \int_{\mathbb{R}^N} \exp(i\langle w, x-y\rangle) \\
&\quad \prod_{j=1}^N \phi(w(j)/\Omega)\mathrm{d}w\mathrm{d}y \\
&= \frac{1}{(2\pi)^N} \int_K \int_{[-\Omega,\Omega]^N} f(y)\cos(\langle w, x-y\rangle) \\
&\quad \prod_{j=1}^N \phi(w(j)/\Omega)\mathrm{d}w\mathrm{d}y
\end{aligned} \tag{24}$$

because ϕ is even and supported on $[-1,1]$. Since Equation 24 is an iterated integral of a continuous function over a compact set, Fubini's theorem readily applies, yielding

$$\begin{aligned}
f(x) &= \lim_{\Omega\to\infty} (f * h_\Omega)(x) \\
&= \lim_{\Omega\to\infty} \frac{1}{(2\pi)^N} \int_{K\times[-\Omega,\Omega]^N} f(y)\cos(\langle w, x-y\rangle) \prod_{j=1}^N \phi(w(j)/\Omega)\mathrm{d}y\mathrm{d}w.
\end{aligned}$$

Since $|\langle w, x-y\rangle| \leq \|x-y\|_1 \|w\|_\infty \leq 2N\mathrm{rad}(K)\Omega \leq (L+\frac{1}{2})\pi$, it follows that

$$\begin{aligned}
f(x) = \lim_{\Omega\to\infty} \frac{1}{(2\pi)^N} \int_{K\times[-\Omega,\Omega]^N} &f(y)\cos_\Omega(\langle w, x-y\rangle) \\
&\prod_{j=1}^N \phi(w(j)/\Omega)\mathrm{d}y\mathrm{d}w
\end{aligned} \tag{25}$$

where $\cos_\Omega$ is defined in Equation 5.

With the representation given by Equation 25 in hand, we now seek to reintroduce the general activation function ρ. To this end, since $\cos_\Omega \in C_c(\mathbb{R}) \subset C_0(\mathbb{R})$ we may apply the convolution identity given by Equation 4 with f replaced by $\cos_\Omega$ to obtain $\cos_\Omega(z) = \lim_{\alpha\to\infty}(\cos_\Omega * h_\alpha)(z)$ uniformly for all $z \in \mathbb{R}$, where $h_\alpha(z) = \alpha\rho(\alpha z)$. Using this representation of $\cos_\Omega$ in Equation 25, it follows that

$$\begin{aligned}
f(x) = \lim_{\Omega\to\infty} \frac{1}{(2\pi)^N} \int_{K\times[-\Omega,\Omega]^N} &f(y)\Big(\lim_{\alpha\to\infty} \left(\cos_\Omega * h_\alpha\right)\left(\langle w, x-y\rangle\right)\Big) \\
&\prod_{j=1}^N \phi(w(j)/\Omega)\mathrm{d}y\mathrm{d}w
\end{aligned}$$

holds uniformly for all $x \in K$. Since f is continuous and the convolution $\cos_\Omega * h_\alpha$ is uniformly continuous and uniformly bounded in α by $\|\rho\|_1$ (see below), the fact that the domain $K \times [-\Omega,\Omega]^N$ is compact then allows us to bring the limit as α tends to infinity outside the integral in this expression via the Dominated Convergence Theorem, which gives us

$$f(x) = \lim_{\Omega\to\infty}\lim_{\alpha\to\infty}\frac{1}{(2\pi)^N}\int_{K\times[-\Omega,\Omega]^N} f(y)\big(\cos_\Omega * h_\alpha\big)\big(\langle w, x-y\rangle\big)\prod_{j=1}^{N}\phi(w(j)/\Omega)\mathrm{d}y\mathrm{d}w \tag{26}$$

uniformly for every $x \in K$. The uniform boundedness of the convolution follows from the fact that

$$\begin{aligned}(\cos_\Omega * h_\alpha)(z) &= \int_{\mathbb{R}} \cos_\Omega(z-u)h_\alpha(u)\mathrm{d}u \\ &= \int_{\mathbb{R}} \cos_\Omega(z-v/\alpha)\rho(v)\mathrm{d}v,\end{aligned} \tag{27}$$

where $v = \alpha u$.

Remark 11. It should be noted that we are unable to swap the order of the limits in Equation 26, since $\cos_\Omega$ is not in $C_0(\mathbb{R})$ when Ω is allowed to be infinite.

Remark 12. Complementing Remark 10, we will now elucidate how fast

$$|\cos_\Omega(z) - (\cos_\Omega * h_\alpha)(z)|$$

decays in terms of α. Using the fact that $\int_{\mathbb{R}}\rho(z)\mathrm{d}z = 1$, Equation 27 and the triangle inequality allows us to bound the absolute difference above by

$$\int_{\mathbb{R}} |\cos_\Omega(z) - \cos_\Omega(z - v/\alpha)|\cdot|\rho(v)|\mathrm{d}v.$$

Since $\cos_\Omega$ is 1-Lipschitz, it follows that the above integral is bounded by $\int_{\mathbb{R}}|v\rho(v)|\,dv/\alpha$.

To complete this step of the proof, observe that the definition of $\cos_\Omega$ allows us to write

$$\begin{aligned}(\cos_\Omega * h_\alpha)(z) &= \alpha\int_{\mathbb{R}}\cos_\Omega(u)\rho\big(\alpha(z-u)\big)\mathrm{d}u \\ &= \alpha\int_{-\frac{\pi}{2}(2L+1)}^{\frac{\pi}{2}(2L+1)}\cos_\Omega(u)\rho\big(\alpha(z-u)\big)\mathrm{d}u\end{aligned} \tag{28}$$

By substituting Equation 28 into Equation 26, we then obtain

$$f(x) = \lim_{\Omega\to\infty}\lim_{\alpha\to\infty}\frac{\alpha}{(2\pi)^N}\int_{K(\Omega)} f(y)\cos_\Omega(u)\rho\Big(\alpha\big(\langle w, x-y\rangle - u\big)\Big)\prod_{j=1}^{N}\phi(w(j)/\Omega)\mathrm{d}y\mathrm{d}w\mathrm{d}u$$

uniformly for all $x \in K$, where $K(\Omega) := K\times[-\Omega,\Omega]^N\times[-\frac{\pi}{2}(2L+1),\frac{\pi}{2}(2L+1)]$. In this way, recalling that $F_{\alpha,\Omega}(y,w,u) := \frac{\alpha}{(2\pi)^N}f(y)\cos_\Omega(u)\prod_{j=1}^N\phi(w(j)/\Omega)$, and $b_\alpha(y,w,u) := -\alpha(\langle w,y\rangle + u)$ for $y, w \in \mathbb{R}^N$ and $u \in \mathbb{R}$, we conclude the proof.

3.5.3 Proof of Lemma 5: Monte-Carlo integral approximation

The next step in the proof of Theorem 5 is to approximate the integral in Equation 7 using the Monte-Carlo method. To this end, let $\{y_k\}_{k=1}^n$, $\{w_k\}_{k=1}^n$, and $\{u_k\}_{k=1}^n$ be independent samples drawn uniformly from K, $[-\Omega,\Omega]^N$, and $[-\frac{\pi}{2}(2L+1),\frac{\pi}{2}(2L+1)]$, respectively, and consider the sequence of random variables $\{I_n(x)\}_{n=1}^\infty$ defined by

$$I_n(x) := \frac{\mathrm{vol}(K(\Omega))}{n}\sum_{k=1}^{n}F_{\alpha,\Omega}(y_k,w_k,u_k)\rho\big(\alpha\langle w_k,x\rangle + b_\alpha(y_k,w_k,u_k)\big) \tag{29}$$

for each $x \in K$, where we note that $\mathrm{vol}(K(\Omega)) = (2\Omega)^N\pi(2L+1)\mathrm{vol}(K)$. If we also define

$$I(x;p) := \int_{K(\Omega)}\Big(F_{\alpha,\Omega}(y,w,u)\rho\big(\alpha\langle w,x\rangle + b_\alpha(y,w,u)\big)\Big)^p\mathrm{d}y\mathrm{d}w\mathrm{d}u$$

for $x \in K$ and $p \in \mathbb{N}$, then we want to show that

$$\mathbb{E}\int_K |I(x;1) - I_n(x)|^2\mathrm{d}x = O(1/n) \tag{30}$$

as $n \to \infty$, where the expectation is taken with respect to the joint distribution of the random samples $\{y_k\}_{k=1}^n$, $\{w_k\}_{k=1}^n$, and $\{u_k\}_{k=1}^n$. For this, it suffices to find a constant $C(f,\rho,\alpha,\Omega,N) < \infty$ independent of n satisfying

$$\int_K \mathbb{E}|I(x;1) - I_n(x)|^2\mathrm{d}x \le \frac{C(f,\rho,\alpha,\Omega,N)}{n}.$$

Indeed, an application of Fubini's theorem would then yield

$$\mathbb{E}\int_K |I(x;1) - I_n(x)|^2\mathrm{d}x \le \frac{C(f,\rho,\alpha,\Omega,N)}{n},$$

which implies Equation 30. To determine such a constant, we first observe by Theorem 4 that

$$\mathbb{E}|I(x;1) - I_n(x)|^2 = \frac{\mathrm{vol}^2(K(\Omega))\sigma(x)^2}{n},$$

where we define the variance term

$$\sigma(x)^2 := \frac{I(x;2)}{\mathrm{vol}(K(\Omega))} - \frac{I(x;1)^2}{\mathrm{vol}^2(K(\Omega))}$$

for $x \in K$. Since $\|\phi\|_\infty = 1$ (see Lemma 8 below), it follows that

$$|F_{\alpha,\Omega}(y,w,u)| = \frac{\alpha}{(2\pi)^N}|f(y)|\cdot|\cos_\Omega(u)|\prod_{j=1}^{N}|\phi(w(j)/\Omega)| \le \frac{\alpha M}{(2\pi)^N}$$

for all $y, w \in \mathbb{R}^N$ and $u \in \mathbb{R}$, where $M := \sup_{x\in K}|f(x)| < \infty$, we obtain the following simple bound on the variance term

$$\sigma(x)^2 \le \frac{I(x;2)}{\mathrm{vol}(K(\Omega))} \le \frac{\alpha^2M^2}{(2\pi)^{2N}\mathrm{vol}(K(\Omega))}\int_{K(\Omega)}\Big|\rho\big(\alpha\langle w,x\rangle + b_\alpha(y,w,u)\big)\Big|^2\mathrm{d}y\mathrm{d}w\mathrm{d}u.$$

Since we assume $\rho \in L^\infty(\mathbb{R})$, we then have

$$\int_K \mathbb{E}|I(x;1) - I_n(x)|^2 \mathrm{d}x = \frac{\mathrm{vol}^2(K(\Omega))}{n} \int_K \sigma(x)^2 \mathrm{d}x$$
$$\leq \frac{\alpha^2 M^2 \mathrm{vol}(K(\Omega))}{(2\pi)^{2N} n} \int_{K\times K(\Omega)} \left|\rho\left(\alpha\langle w, x\rangle + b_\alpha(y,w,u)\right)\right|^2 \mathrm{d}x\mathrm{d}y\mathrm{d}w\mathrm{d}u$$
$$= \frac{\alpha^2 M^2 \mathrm{vol}^2(K(\Omega))\mathrm{vol}(K)\|\rho\|_\infty^2}{(2\pi)^{2N} n}.$$

Substituting the value of $\mathrm{vol}(K(\Omega))$, we obtain

$$C(f,\rho,\alpha,\Omega,N) := \alpha^2 M^2 (\Omega/\pi)^{2N} \pi^2 (2L+1)^2 \mathrm{vol}^3(K)\|\rho\|_\infty^2$$

is a suitable choice for the desired constant.

Now that we have established Equation 30, we may rewrite the random variables $I_n(x)$ in a more convenient form. To this end, we change the domain of the random samples $\{w_k\}_{k=1}^n$ to $[-\alpha\Omega, \alpha\Omega]^N$ and define the new random variables $\{b_k\}_{k=1}^n \subset \mathbb{R}$ by $b_k := -(\langle w_k, y_k\rangle + \alpha u_k)$ for each $k = 1, \ldots, n$. In this way, if we denote

$$v_k := \frac{\mathrm{vol}(K(\Omega))}{n} F_{\alpha,\Omega}\left(y_k, \frac{w_k}{\alpha}, u_k\right)$$

for each $k = 1, \ldots, n$, the random variables $\{f_n\}_{n=1}^\infty$ defined by

$$f_n(x) := \sum_{k=1}^n v_k \rho\left(\langle w_k, x\rangle + b_k\right)$$

satisfy $f_n(x) = I_n(x)$ for every $x \in K$. Combining this with Equation 30, we have proved Lemma 5.

Lemma 8. $\|\phi\|_\infty = 1$.

Proof. It suffices to prove that $|\phi(z)| \leq 1$ for all $z \in \mathbb{R}$ because $\phi(0) = 1$. By Cauchy–Schwarz,

$$|\phi(z)| = \left|\int_{\mathbb{R}} A(u)A(z-u)\mathrm{d}u\right|$$
$$\leq \sqrt{\int_{\mathbb{R}} A(u)A(u)\mathrm{d}u \int_{\mathbb{R}} A(z-u)A(z-u)\mathrm{d}u}$$
$$= \sqrt{\int_{\mathbb{R}} A(u)A(0-u)\mathrm{d}u \int_{\mathbb{R}} A(v)A(-v)\mathrm{d}v} = \sqrt{\phi(0)\phi(0)} = 1$$

because A is even.

3.5.4 Proof of Theorem 1 when $\rho' \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$

Let $f \in C_c(\mathbb{R}^N)$ with $K := \mathrm{supp}(f)$ and suppose $\varepsilon > 0$ is fixed. Take the activation function $\rho : \mathbb{R} \to \mathbb{R}$ to be differentiable with $\rho' \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$. We wish to show that there exists a sequence of RVFL networks $\{f_n\}_{n=1}^\infty$ defined on K which satisfy the asymptotic error bound

$$\mathbb{E}\int_K |f(x) - f_n(x)|^2 \mathrm{d}x \leq \varepsilon + O(1/n)$$

as $n \to \infty$. The proof of this result is a minor modification of second step in the proof of Theorem 5.

If we redefine $h_\alpha(z)$ as $\alpha\rho'(\alpha z)$, then Equation 26 plainly still holds and Equation 28 reads

$$(\cos_\Omega * h_\alpha)(z) = \alpha \int_{\mathbb{R}} \cos_\Omega(u)\rho'\left(\alpha(z-u)\right)\mathrm{d}u.$$

Recalling the definition of $\cos_\Omega$ in Equation 5 and integrating by parts, we obtain

$$(\cos_\Omega * h_\alpha)(z) = \alpha \int_{\mathbb{R}} \cos_\Omega(u)\rho'\left(\alpha(z-u)\right)\mathrm{d}u$$
$$= -\int_{-\frac{\pi}{2}(2L+1)}^{\frac{\pi}{2}(2L+1)} \cos_\Omega(u)d\rho(\alpha(z-u))$$
$$= -\cos_\Omega(u)\rho(\alpha(z-u))\Big|_{-\frac{\pi}{2}(2L+1)}^{\frac{\pi}{2}(2L+1)}$$
$$+ \int_{-\frac{\pi}{2}(2L+1)}^{\frac{\pi}{2}(2L+1)} \rho(\alpha(z-u))d\cos_\Omega(u)$$
$$= -\int_{\mathbb{R}} \sin_\Omega(u)\rho\left(\alpha(z-u)\right)\mathrm{d}u$$

for all $z \in \mathbb{R}$, where $L := \lceil \frac{2N}{\pi}\mathrm{rad}(K)\Omega - \frac{1}{2}\rceil$ and $\sin_\Omega : \mathbb{R} \to [-1,1]$ is defined analogously to Equation 5. Substituting this representation of $(\cos_\Omega * h_\alpha)(z)$ into Equation 26 then yields

$$f(x) = \lim_{\Omega\to\infty}\lim_{\alpha\to\infty} \frac{-\alpha}{(2\pi)^N} \int_{K(\Omega)} f(y)\sin_\Omega(u)\rho\left(\alpha(\langle w, x-y\rangle - u)\right)$$
$$\prod_{j=1}^N \phi(w(j)/\Omega)\mathrm{d}y\mathrm{d}w\mathrm{d}u$$

uniformly for every $x \in K$. Thus, if we replace the definition of $F_{\alpha,\Omega}$ in Equation 6 by

$$F_{\alpha,\Omega}(y,w,u) := \frac{-\alpha}{(2\pi)^N} f(y)\sin_\Omega(u)\prod_{j=1}^N \phi(w(j)/\Omega)$$

for $y, w \in \mathbb{R}^N$ and $u \in \mathbb{R}$, we again obtain the uniform representation given by Equation 7 for all $x \in K$. The remainder of the proof proceeds from this point exactly as in the proof of Theorem 5.

3.5.5 Proof of Theorem 8

We wish to show that there exist sequences of RVFL networks $\{\tilde{f}_{n_j}\}_{n_j=1}^\infty$ defined on $\phi_j(U_j)$ for each $j \in J$, which together satisfy the error bound

$$\int_{\mathcal{M}} \left|f(x) - \sum_{\{j\in J\,:\,x\in U_j\}} (\tilde{f}_{n_j} \circ \phi_j)(x)\right|^2 \mathrm{d}x < \varepsilon$$

with probability at least $1-\eta$ for $\{n_j\}_{j\in J}$ sufficiently large. The proof is obtained by showing that

$$\left|f(x) - \sum_{\{j\in J\,:\,x\in U_j\}} (\tilde{f}_{n_j} \circ \phi_j)(x)\right| < \sqrt{\frac{\varepsilon}{\mathrm{vol}(\mathcal{M})}} \tag{31}$$

holds uniformly for $x \in \mathcal{M}$ with high probability.

We begin as in the proof of Theorem 7 by applying the representation given by Equation 17 for (f on each chart (U_j, ϕ_j), which gives us

$$\begin{aligned}\left|f(x) - \sum_{\{j\in J:\, x\in U_j\}} (\tilde{f}_{n_j} \circ \phi_j)(x)\right| \\ \leq \sum_{\{j\in J:\, x\in U_j\}} \left|(\hat{f}_j \circ \phi_j)(x) - (\tilde{f}_{n_j} \circ \phi_j)(x)\right| \end{aligned} \tag{32}$$

for all $x \in \mathcal{M}$. Now, since we have already seen that $\hat{f}_j \in C_c(\mathbb{R}^d)$ for each $j \in J$, Theorem 6 implies that for any $\varepsilon_j > 0$, there exist constants $\alpha_j, \Omega_j > 0$ and hidden-to-output layer weights $\{v_k^{(j)}\}_{k=1}^{n_j} \subset \mathbb{R}$ for each $j \in J$ such that for any

$$\delta_j < \frac{\sqrt{\varepsilon_j}}{8\sqrt{2d}\kappa\alpha_j^2 M_j \Omega_j (\Omega_j/\pi)^d \mathrm{vol}^{3/2}(\phi_j(U_j))(\pi + 2d\mathrm{rad}(\phi_j(U_j))\Omega)} \tag{33}$$

we have

$$\left|\hat{f}_j(z) - \tilde{f}_{n_j}(z)\right| < \sqrt{\frac{\varepsilon_j}{2\mathrm{vol}(\phi_j(U_j))}}$$

uniformly for all $z \in \phi_j(U_j)$ with probability at least $1-\eta_j$, provided the number of nodes n_j satisfies

$$n_j \geq \frac{c\Sigma^{(j)}\alpha_j(\Omega_j/\pi)^d(\pi + 2d\mathrm{rad}(\phi_j(U_j))\Omega_j)\log(3\eta_j^{-1}\mathcal{N}(\delta_j, \phi_j(U_j)))}{\sqrt{\varepsilon_j}\log\left(1 + \frac{\sqrt{\varepsilon_j}}{\Sigma^{(j)}\alpha_j(\Omega_j/\pi)^d(\pi+2d\mathrm{rad}(\phi_j(U_j))\Omega_j)}\right)}, \tag{34}$$

where $c > 0$ is a numerical constant and $\Sigma^{(j)} := 2C^{(j)}\sqrt{2\mathrm{vol}(\phi_j(U_j))}$. Indeed, it suffices to choose

$$v_k^{(j)} := \frac{\mathrm{vol}(K(\Omega_j))}{n_j} F_{\alpha_j,\Omega_j}\left(y_k^{(j)}, \frac{w_k^{(j)}}{\alpha_j}, u_k^{(j)}\right)$$

for each $k = 1, \ldots, n_j$, where

$$K(\Omega_j) := \phi_j(U_j) \times [-\alpha_j\Omega_j, \alpha_j\Omega_j]^d \times [-\tfrac{\pi}{2}(2L_j+1), \tfrac{\pi}{2}(2L_j+1)]$$

for each $j \in J$. Combined with Equation 32, choosing δ_j and n_j satisfying Equations 33, 34, respectively, then yields

$$\begin{aligned}\left|f(x) - \sum_{\{j\in J:\, x\in U_j\}} (\tilde{f}_{n_j} \circ \phi_j)(x)\right| < \sum_{\{j\in J:\, x\in U_j\}} \sqrt{\frac{\varepsilon_j}{2\mathrm{vol}(\phi_j(U_j))}} \\ \leq \sum_{j\in J} \sqrt{\frac{\varepsilon_j}{2\mathrm{vol}(\phi_j(U_j))}}\end{aligned}$$

for all $x \in \mathcal{M}$ with probability at least $1 - \sum_{\{j\in J:\, x\in U_j\}} \eta_j \geq 1 - \sum_{j\in J}\eta_j$. Since we require that Equation 31 holds for all $x \in \mathcal{M}$ with probability at least $1-\eta$, the proof is then completed by choosing $\{\varepsilon_j\}_{j\in J}$ and $\{\eta_j\}_{j\in J}$, such that

$$\varepsilon = \frac{\mathrm{vol}(\mathcal{M})}{2}\left(\sum_{j\in J}\sqrt{\frac{\varepsilon_j}{\mathrm{vol}(\phi_j(U_j))}}\right)^2 \quad \text{and} \quad \eta = \sum_{j\in J}\eta_j.$$

In particular, it suffices to choose

$$\varepsilon_j = \frac{2\mathrm{vol}(\phi_j(U_j))\,\varepsilon}{|J|^2\mathrm{vol}(\mathcal{M})}$$

and $\eta_j = \eta/|J|$ for each $j \in J$, so that Equations 33, 34 become

$$\begin{aligned}\delta_j &< \frac{\sqrt{\varepsilon}}{8|J|\sqrt{d\mathrm{vol}(\mathcal{M})}\kappa\alpha_j^2 M_j\Omega_j(\Omega_j/\pi)^d\mathrm{vol}(\phi_j(U_j))(\pi + 2d\mathrm{rad}(\phi_j(U_j))\Omega)}, \\ n_j &\geq \frac{2c|J|\sqrt{\mathrm{vol}(\mathcal{M})}C^{(j)}\alpha_j(\Omega_j/\pi)^d(\pi + 2d\mathrm{rad}(\phi_j(U_j))\Omega_j)\log(3|J|\eta^{-1}\mathcal{N}(\delta_j, \phi_j(U_j)))}{\sqrt{\varepsilon}\log\left(1 + \frac{\sqrt{\varepsilon}}{2|J|\sqrt{\mathrm{vol}(\mathcal{M})}C^{(j)}\alpha_j(\Omega_j/\pi)^d(\pi+2d\mathrm{rad}(\phi_j(U_j))\Omega_j)}\right)},\end{aligned}$$

as desired.

4 Discussion

The central topic of this study is the study of the approximation properties of a randomized variation of shallow neural networks known as RVFL. In contrast with the classical single-layer neural networks, training of an RVFL involves only learning the output weights, while the input weights and biases of all the nodes are selected at random from an appropriate distribution and stay fixed throughout the training. The main motivation for studying the properties of such networks is as follows:

1. Random weights are often utilized as an initialization for a NN training procedure. Thus, establishing the properties of the RVFL networks is an important first step toward understanding how random weights are transformed during training.
2. Due to their much more computationally efficient training process, the RVFL networks proved to be a valuable alternative to the classical SLFNs. They were successfully used in several modern applications, especially those that require frequent re-training of a neural network [20, 26, 27].

Despite their practical and theoretical importance, results providing rigorous mathematical analysis of the properties of RVFLs are rare. The work of Igelnik and Pao [16] showed that RVFL networks are universal approximators for the class of continuous, compactly supported functions and established the asymptotic convergence rate of the expected approximation error as a function of the number of nodes in the hidden layer. While this result served as a theoretical justification for using RVFL networks in practice, a close examination led us to the conclusion that the proofs in Igelnik and Pao [16] contained several technical errors.

In this study, we offer a revision and a modification of the proof methods from Igelnik and Pao [16] that allow us to prove a corrected, slightly weaker version of the result announced by Igelnik and Pao. We further build upon their work and show a non-asymptotic probabilistic (instead of on average) approximation result, which gives an explicit bound on the number of hidden layer nodes that are required to achieve the desired approximation accuracy with the desired level of certainty (that is, with high enough probability). In addition to that, we extend the obtained result to the case when the function is supported on a compact, low-dimensional sub-manifold of the ambient space.

While our study closes some of the gaps in the study of the approximation properties of RVFL, we believe that it just starts

the discussion and opens many directions for further research. We briefly outline some of them here.

In our results, the dependence of the required number n of the nodes in the hidden layer on the dimension N of the domain is superexponential, which is likely an artifact of the proof methods we use. We believe this dependence can be improved to be exponential by using a different, more refined approach to the construction of the limit-integral representation of a function. A related interesting direction for future research is to study how the bound on n changes for more restricted classes of (e.g., smooth) functions.

Another important direction that we did not discuss in this study is learning the output weights and studying the robustness of the RVFL approximation to the noise in the training data. Obtaining provable robustness guarantees for an RVFL training procedure would be a step toward the robustness analysis of neural networks.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

DN: Writing – original draft, Writing – review & editing. AN: Writing – original draft, Writing – review & editing. RS: Writing – original draft, Writing – review & editing. PS: Writing – original draft, Writing – review & editing. OS: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. DN was partially supported by NSF DMS 2108479 and NSF DMS 2011140. RS was partially supported by a UCSD senate research award and a Simons fellowship. PS was partially supported by NSF DMS 1909457.

Acknowledgments

The authors thank F. Krahmer, S. Krause-Solberg, and J. Maly for sharing their GMRA code, which they adapted from that provided by M. Maggioni.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

The views expressed in this article are those of the authors and do not reflect the official policy or position of the U.S. Air Force, Department of Defense, or U.S. Government.

References

1. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*. (2012). p. 1097–105.

2. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2015). p. 1–9. doi: 10.1109/CVPR.2015.7298594

3. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition*. (2016). p. 770–778. doi: 10.1109/CVPR.2016.90

4. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2017). p. 4700–4708. doi: 10.1109/CVPR.2017.243

5. Yang Y, Zhong Z, Shen T, Lin Z. Convolutional neural networks with alternately updated clique. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2018). p. 2413–2422. doi: 10.1109/CVPR.2018.00256

6. Barron AR. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans Inf Theory*. (1993) 39:930–45. doi: 10.1109/18.256500

7. Candés EJ. Harmonic analysis of neural networks. *Appl Comput Harm Analysis*. (1999) 6:197–218. doi: 10.1006/acha.1998.0248

8. Vershynin R. Memory capacity of neural networks with threshold and ReLU activations. *arXiv preprint arXiv:200106938*. (2020).

9. Baldi P, Vershynin R. The capacity of feedforward neural networks. *Neural Netw*. (2019) 116:288–311. doi: 10.1016/j.neunet.2019.04.009

10. Huang GB, Babri HA. Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions. *IEEE Trans Neural Netw*. (1998) 9:224–9. doi: 10.1109/72.655045

11. Suganthan PN. Letter: On non-iterative learning algorithms with closed-form solution. *Appl Soft Comput*. (2018) 70:1078–82. doi: 10.1016/j.asoc.2018.07.013

12. Olson M, Wyner AJ, Berk R. Modern Neural Networks Generalize on Small Data Sets. In: *Proceedings of the 32Nd International Conference on Neural Information Processing Systems. NIPS'18*. Curran Associates Inc. (2018). p. 3623–3632.

13. Schmidt WF, Kraaijveld MA, Duin RPW. Feedforward neural networks with random weights. In: *Proceedings 11th IAPR International Conference on Pattern Recognition. Vol. II. Conference B: Pattern Recognition Methodology and Systems*. (1992). p. 1–4.

14. Te Braake HAB, Van Straten G. Random activation weight neural net (RAWN) for fast non-iterative training. *Eng*

Appl Artif Intell. (1995) 8:71–80. doi: 10.1016/0952-1976(94)00056-S

15. Pao YH, Takefuji Y. Functional-link net computing: theory, system architecture, and functionalities. *Computer*. (1992) 25:76–9. doi: 10.1109/2.144401

16. Igelnik B, Pao YH. Stochastic choice of basis functions in adaptive function approximation and the functional-link net. *IEEE Trans Neur Netw*. (1995) 6:1320–9. doi: 10.1109/72.471375

17. Huang GB, Zhu QY, Siew CK. Extreme learning machine: theory and applications. *Neurocomputing*. (2006) 70:489–501. doi: 10.1016/j.neucom.2005.12.126

18. Pao YH, Park GH, Sobajic DJ. Learning and generalization characteristics of the random vector functional-link net. *Neurocomputing*. (1994) 6:163–80. doi: 10.1016/0925-2312(94)90053-1

19. Pao YH, Phillips SM. The functional link net and learning optimal control. *Neurocomputing*. (1995) 9:149–64. doi: 10.1016/0925-2312(95)00066-F

20. Chen CP, Wan JZ, A. rapid learning and dynamic stepwise updating algorithm for flat neural networks and the application to time-series prediction. *IEEE Trans Syst Man Cybern B*. (1999) 29:62–72. doi: 10.1109/3477.740166

21. Park GH, Pao YH. Unconstrained word-based approach for off-line script recognition using density-based random-vector functional-link net. *Neurocomputing*. (2000) 31:45–65. doi: 10.1016/S0925-2312(99)00149-6

22. Zhang L, Suganthan PN. Visual tracking with convolutional random vector functional link network. *IEEE Trans Cybern*. (2017) 47:3243–53. doi: 10.1109/TCYB.2016.2588526

23. Zhang L, Suganthan PN. Benchmarking ensemble classifiers with novel co-trained kernel ridge regression and random vector functional link ensembles [research frontier]. *IEEE Comput Intell Mag*. (2017) 12:61–72. doi: 10.1109/MCI.2017.2742867

24. Katuwal R, Suganthan PN, Zhang L. An ensemble of decision trees with random vector functional link networks for multi-class classification. *Appl Soft Comput*. (2018) 70:1146–53. doi: 10.1016/j.asoc.2017.09.020

25. Vuković N, Petrović M, Miljković Z. A comprehensive experimental evaluation of orthogonal polynomial expanded random vector functional link neural networks for regression. *Appl Soft Comput*. (2018) 70:1083–1096. doi: 10.1016/j.asoc.2017.10.010

26. Tang L, Wu Y, Yu L, A. non-iterative decomposition-ensemble learning paradigm using RVFL network for crude oil price forecasting. *Appl Soft Comput*. (2018) 70:1097–108. doi: 10.1016/j.asoc.2017.02.013

27. Dash Y, Mishra SK, Sahany S, Panigrahi BK. Indian summer monsoon rainfall prediction: A comparison of iterative and non-iterative approaches. *Appl Soft Comput*. (2018) 70:1122–34. doi: 10.1016/j.asoc.2017.08.055

28. Henríquez PA, Ruz G. Twitter sentiment classification based on deep random vector functional link. In: 2018 *International Joint Conference on Neural Networks (IJCNN)*. IEEE (2018). p. 1–6. doi: 10.1109/IJCNN.2018.8489703

29. Katuwal R, Suganthan PN, Tanveer M. Random vector functional link neural network based ensemble deep learning. *arXiv preprint arXiv:190700350*. (2019).

30. Zhang Y, Wu J, Cai Z, Du B, Yu PS. An unsupervised parameter learning model for RVFL neural network. *Neural Netw*. (2019) 112:85–97. doi: 10.1016/j.neunet.2019.01.007

31. Hornik K. Approximation capabilities of multilayer feedforward networks. *Neural Netw*. (1991) 4:251–7. doi: 10.1016/0893-6080(91)90009-T

32. Leshno M, Lin VY, Pinkus A, Schocken S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Netw*. (1993) 6:861–7. doi: 10.1016/S0893-6080(05)80131-5

33. Li JY, Chow WS, Igelnik B, Pao YH. Comments on "Stochastic choice of basis functions in adaptive function approximation and the functional-link net" [with reply]. *IEEE Trans Neur Netw*. (1997) 8:452–4. doi: 10.1109/72.557702

34. Burkhardt DB, San Juan BP, Lock JG, Krishnaswamy S, Chaffer CL. Mapping phenotypic plasticity upon the cancer cell state landscape using manifold learning. *Cancer Discov*. (2022) 12:1847–59. doi: 10.1158/2159-8290.CD-21-0282

35. Mitchell-Heggs R, Prado S, Gava GP, Go MA, Schultz SR. Neural manifold analysis of brain circuit dynamics in health and disease. *J Comput Neurosci*. (2023) 51:1–21. doi: 10.1007/s10827-022-00839-3

36. Dick J, Kuo FY, Sloan IH. High-dimensional integration: the quasi-Monte Carlo way. *Acta Numerica*. (2013) 22:133–288. doi: 10.1017/S0962492913000044

37. Ledoux M. The Concentration of Measure Phenomenon. Providence, RI: American Mathematical Soc. (2001).

38. Massart P. *About the constants in Talagrand's deviation inequalities for empirical processes*. Technical Report, Laboratoire de statistiques, Universite Paris Sud. (1998).

39. Talagrand M. New concentration inequalities in product spaces. *Invent Mathem*. (1996) 126:505–63. doi: 10.1007/s002220050108

40. Shaham U, Cloninger A, Coifman RR. Provable approximation properties for deep neural networks. *Appl Comput Harmon Anal*. (2018) 44:537–57. doi: 10.1016/j.acha.2016.04.003

41. Tu LW. *An Introduction to Manifolds*. New York: Springer. (2010). doi: 10.1007/978-1-4419-7400-6_3

42. Allard WK, Chen G, Maggioni M. Multi-scale geometric methods for data sets II: geometric multi-resolution analysis. *Appl Comput Harmon Anal*. (2012) 32:435–62. doi: 10.1016/j.acha.2011.08.001

43. Iwen MA, Krahmer F, Krause-Solberg S, Maly J. On recovery guarantees for one-bit compressed sensing on manifolds. Discr Comput Geom. (2018) 65:953–998. doi: 10.1109/SAMPTA.2017.8024465

44. Maggioni M, Minsker S, Strawn N. Multiscale dictionary learning: non-asymptotic bounds and robustness. *J Mach Learn Res*. (2016) 17:43–93.

45. Liao W, Maggioni M. Adaptive Geometric Multiscale Approximations for Intrinsically Low-dimensional Data. *J Mach Learn Res*. (2019) 20:1–63.

46. Stein EM, Weiss G. *Introduction to Fourier Analysis on Euclidean Spaces*. London: Princeton University Press (1971). doi: 10.1515/9781400883899

47. Rudin W. *Functional Analysis*. New York: McGraw-Hill (1991).