Efficient uniform approximation using Random Vector Functional Link networks

Palina Salanevich
Utrecht University
p.salanevich@uu.nl

Olov Schavemaker
Utrecht University
o.p.schavemaker@uu.nl

***Abstract*—A Random Vector Functional Link (RVFL) network is a depth-2 neural network with random inner weights and biases. As only the outer weights of such an architecture need to be learned, the learning process boils down to a linear optimization task, allowing one to sidestep the pitfalls of nonconvex optimization problems. In this paper, we prove that an RVFL with ReLU activation functions can approximate Lipschitz continuous functions provided its hidden layer is exponentially wide in the input dimension. Although it has been established before that such approximation can be achieved in L_2 sense, we prove it for L_∞ approximation error and Gaussian inner weights. To the best of our knowledge, our result is the first of this kind. We give a non-asymptotic lower bound for the number of hidden layer nodes, depending on, among other things, the Lipschitz constant of the target function, the desired accuracy, and the input dimension. Our method of proof is rooted in probability theory and harmonic analysis.**

I. Introduction

In this paper, we examine the approximation capacity of the Random Vector Functional Link (RVFL) network. An RVFL is a depth-2 neural network with random inner weights and biases. More precisely, an RVFL is a random function $N_n : \mathbb{R}^m \to \mathbb{R}$ of the form

$$N_n(x) = \sum_{j=1}^{n} a_j \rho(\langle \mathbb{w}_j, x\rangle + \mathbb{b}_j),$$

where $\mathbb{w}_j$'s and $\mathbb{b}_j$'s are iid random variables, $\rho : \mathbb{R} \to \mathbb{R}$ is the *activation function*, and a_j's are real numbers, chosen or learned so as to have N_n be close to a target function f.

Despite the simplicity of their architecture, RVFL models found their applications in signal classification and regression problems [7], forecasting [10], time-series data prediction [3], and others; for an overview, see [8]. At the same time, theoretical foundation for RVFL networks is still lacking [9, §1]. This paper aims to remedy this discrepancy, bringing us one step closer to understanding more complicated architectures on neural networks, widely used in practical applications.

Since only the outer weights of RVFL architectures need to be optimized, in practice the learning process boils down to a linear optimization task. Indeed, given training data $\{x_p\}_{p=1}^{k}$, we aim to choose a_j's so that

$$\begin{Bmatrix} f(x_1) \\ \vdots \\ f(x_k) \end{Bmatrix} \approx \begin{Bmatrix} \sum_{j=1}^{n} a_j \rho(\langle \mathbb{w}_j, x_1\rangle + \mathbb{b}_j) \\ \vdots \\ \sum_{j=1}^{n} a_j \rho(\langle \mathbb{w}_j, x_k\rangle + \mathbb{b}_j) \end{Bmatrix} =$$
$$\begin{Bmatrix} \rho(\langle \mathbb{w}_1, x_1\rangle + \mathbb{b}_1) & \cdots & \rho(\langle \mathbb{w}_n, x_1\rangle + \mathbb{b}_n) \\ \vdots & \ddots & \vdots \\ \rho(\langle \mathbb{w}_1, x_k\rangle + \mathbb{b}_1) & \cdots & \rho(\langle \mathbb{w}_n, x_k\rangle + \mathbb{b}_n) \end{Bmatrix} \begin{Bmatrix} a_1 \\ \vdots \\ a_n \end{Bmatrix}.$$

Owing to the fact that the learning process boils down to a linear optimization problem, training RVFL networks sidesteps the usual pitfalls of nonconvex optimization problems, such as slow convergence and getting stuck in local optima.

In order to compare our main result to the existing literature, it will be useful to first state an approximate version of our main theorem.

Main Theorem (approximate). *Let K be compact subset of $\mathbb{R}^m$ with at least two elements and circumradius R. Let p be a circumcenter of K and $f : K \to \mathbb{R}$ be ℓ-Lipschitz with $2\zeta = \max f + \min f$. There exist outer weights such that the corresponding RVFL network N_n with n hidden-layer nodes, ReLU activation functions, Gaussian inner weights, and uniformly distributed biases satisfies*

$$\mathbb{P}\Big\{ \max_{x \in K+p} |f(x-p) - \zeta - N_n(x)| > \varepsilon \Big\} \leqslant \eta$$

for any $\eta > 0$ if

$$n \gtrapprox \frac{2e}{\pi} \ln(2/\eta)(2\ell R\sqrt{e}/\varepsilon)^{2m+6} \exp\Big(2d(K)\ln 2 - 2^{1/3} a m^{1/3} + 3\ln m\Big)$$

for large m and small ε such that $\varepsilon^{-1/m} \approx 1$. Here $a \approx -2.3381$ is the first negative zero of the Airy function Ai *and $d(K) \in [1, m]$ is given by Definition 2 below.*

Remark. Although K is in fact allowed to be a low-dimensional compact manifold, the lower bound for n will remain exponential in the ambient dimension, meaning more refined methods will be required to capitalize on the low-dimensional structure insofar possible. Assuming K is a low-dimensional manifold and trying to get a bound that is exponential in the manifold dimension in place of the ambient dimension is a setting commonly adopted to bridge the aforementioned gap between theory and practice [9].

While many results like this exist in the literature, ours offers an improvement on all of those in at least one way. To the best of our knowledge, two of the papers most similar to ours are [9] and [6]; both of them also examined how many hidden-layer nodes are sufficient for an RVFL to be able to approximate (Lipschitz) continuous functions.

In [9], Needell et al. obtained a sufficient number of hidden-layer nodes depending *superexponentially* on the dimension. Unlike us, they use L_2 approximation error and the support of their uniformly distributed weights (and biases) crucially depend(s) on the approximation error, making their distributions not very suitable for practical use.

Our Gaussian weights, which are more common in practice than uniform ones, do not have variances depending on the approximation error. Our result additionally improves on [9, Thm 4.1] in that we measure the error in the L_∞ norm and our bound is only exponential in the dimension. In fact, our paper is, to the best of our knowledge, the first one to study Gaussian weights and L_∞ approximation error.

Hsu et al. not only managed to obtain a sufficient number of hidden-layer nodes that depends exponentially on the dimension, but also a necessary number of hidden-layer nodes, albeit in the L_2 norm [6]. However, unlike our Gaussian weights, which align with common practice, their weight distribution is supported on a discrete subset of the unit sphere, which may be too restrictive for practical use. Their proof method, whilst conceptually similar to ours, differs in many details, such as using Fourier series where we use the Fourier transform.

We would also like to draw attention to [1], which also concerns itself with RVFL networks, although without the RVFL moniker. Whereas the "corrective method" developed therein is very different from our proof method, their use of spectral methods has been a great inspiration for this paper.

Lastly, we would be remiss if we did not to mention that this paper comes with a supplement. Many technical proof details for the lemmas below have been delegated to the supplement for brevity's sake. We highly encourage the interested reader to check it out as well.

Notation

- Square brackets may denote Iverson brackets
- Derivatives may be denoted by a dot atop the function
- j_ν is the first positive zero of the Bessel function J_ν
- $a \approx -2.3381$ symbolizes the first negative zero of the Airy function Ai
- ρ denotes the ReLU function
- sg denotes the sign function
- Integrals without specified integration domains are understood to integrate over all of Euclidean space
- $\mathscr{F}\{\varphi\}(v) \equiv \int \varphi(u) \exp(-i\langle v, u\rangle)\, du$
- δ_X denotes the pdf of a random variable X (cf., δ_0)
- We write $\varphi(\diamondsuit)$ in lieu of the more common $\varphi(\cdot)$
- V_m denotes the volume of the m-dimensional unit ball
- Absolute value bars may denote either the m-dimensional Lebesgue measure or the ℓ_2 norm in any dimension
- $\left\|\diamondsuit\right\|_K \equiv \max\limits_K |\diamondsuit|$

II. Bird's-eye overview

Our approximation procedure essentially comprises four steps. We first extend the ℓ-Lipschitz target function f, which is only defined on some compactum $K \subset \mathbb{R}^m$, to a compactly supported ℓ-Lipschitz function $\tilde{f}$. The next three steps constitute a chain of approximations: $\tilde{f} \approx g \approx h \approx N_n$.

First, we approximate the extension of the target function $\tilde{f}$ with a "smoothed" version g obtained by convolving $\tilde{f}$ with a specifically-constructed approximate delta function. We show that g can be viewed as an "infinite width" depth-2 neural network with Gaussian inner weights. The biases of g, however, are not random, and the activation function is a cosine function.

Secondly, g is approximated by h, which is an "infinite width" RVFL with ReLU activation functions and Gaussian inner weights, and can be viewed as the "infinite width limit" of the desired RVFL.

Lastly, we will use Hoeffding's concentration inequality to approximate h by a finite width counterpart N_n with n hidden-layer nodes.

III. Main result and proof

Before we state our main theorem, the following definitions will prove useful.

Definition 1. Let $K \subset \mathbb{R}^m$ be a compactum with at least two elements. We define its *circumradius* as

$$R = \min_p \max_{u \in K} |u - p|.$$

Note that $R > 0$. Any $p \in \mathbb{R}^m$ achieving the minimum is called a *circumcenter*.

Remark. If $q \notin K + \{x \in \mathbb{R}^m : |x| \leqslant \operatorname{diam}(K)\}$, then

$$\max_{u \in K} |u - q| > \operatorname{diam}(K) \geqslant \min_{p \in K} \max_{u \in K} |u - p|,$$

so every compactum K with at least two elements has at least one circumcenter within $K + \{x \in \mathbb{R}^m : |x| \leqslant \operatorname{diam}(K)\}$.

Definition 2. Let $K \subset \mathbb{R}^m$ be a compactum with at least two elements and circumradius R. We denote

$$d(K) = \lg \frac{|K + \{x \in \mathbb{R}^m : |x| \leqslant R\}|}{|\{x \in \mathbb{R}^m : |x| \leqslant R\}|},$$

where lg is the binary logarithm. Note that $1 \leqslant d(K) \leqslant m$, so that $d(K)$ may be seen as some sort of unfamiliar notion of *effective dimension*.

Our main result is the following.

Main Theorem. *Let $K \subset \mathbb{R}^m$ be compact with at least two elements and circumradius R. Let p be a circumcenter of K and $f : K \to \mathbb{R}$ be ℓ-Lipschitz with* $\max f - \min f = 2M$

and $\zeta = M + \min f$. There exist outer weights such that the corresponding RVFL network N_n with n hidden-layer nodes, ReLU activation functions, inner weight distribution $N(0, \sigma I_m)$ with $\sigma > 0$, and biases uniformly distributed on $\Big[-\sigma R\sqrt{m}, \sigma R\sqrt{m}\Big]$ satisfies

$$\mathbb{P}\Big\{\big\|f(\Diamond - p) - \zeta - N_n\big\|_{K+p} > \varepsilon\Big\} \leqslant \eta$$

for any $\varepsilon, \eta > 0$ if

$$\begin{aligned} n \geqslant \frac{1}{8\pi e} &\ln(2/\eta)(1+\vartheta)^2 \bigg\{1 + \frac{m+1}{m(m+2)}\bigg\}^{2(m+2)} \\ &\Big(2 - 2^{1/3} a m^{-2/3}\Big)^4 \Big(m^2 + 3m + 1\Big)^{2+2/m} \\ &\exp\Big(2d(K)\ln 2 - 2^{1/3} a m^{1/3} - \ln m\Big) \\ &(2\ell R\sqrt{e}/\varepsilon)^{2m+6+2/m}, \end{aligned}$$

where

$$\begin{aligned} \frac{1}{\vartheta} = &\bigg\{\frac{m^2+3m+1}{\varepsilon} \times \frac{2\ell R}{\sqrt{\pi m}} V_m R^m\bigg\}^{1/m} \sqrt{\frac{em}{2\pi}} \\ &\bigg\{1 + \frac{m+1}{m(m+2)}\bigg\} \times \frac{\ell}{\varepsilon}\Big(2 - 2^{1/3} a m^{-2/3}\Big). \end{aligned}$$

Note that $1/\vartheta \approx 2\ell Re/\varepsilon$ for large m and small ε such that $\varepsilon^{-1/m} \approx 1$ in light of the fact that

$$V_m^{1/m} \sim \sqrt{2\pi e/m},$$

which follows readily from applying Stirling's formula to $V_m = \pi^{m/2} / \Gamma(m/2+1)$ [4, (5.19.4)].

WLOG we henceforth suppose that $p = 0$ and $\zeta = 0$, i.e., the image of f is $[-M, M]$. One can interpret ζ as the bias of the outer layer of N_n, which would be convenient in practice. A good way to deal with p in practice would be approximating p as a part of preprocessing.

Proof of the Main Theorem

Our first order of business is extending f. This would make it easier to construct a smooth approximation of f. Let

$$\tilde{f}(x) = \rho\Big(|f(a)| - \ell|x-a|\Big) \operatorname{sg} f(a),$$

where $a \in \operatorname*{argmax}_{u \in K}\Big(|f(u)| - \ell|x-u|\Big)$.

Lemma 1. *$\tilde{f}$ is a compactly supported ℓ-Lipschitz extension of f.*

Proof. To show that $\tilde{f}$ extends f boils down to showing that $x \in K \Rightarrow \tilde{f}(x) = f(x)$. It suffices to show that x maximizes $K \ni u \mapsto |f(u)| - \ell|x-u|$, which is plain from the fact that

$$|f(u)| - |f(x)| \leqslant |f(x) - f(u)| \leqslant \ell|x-u|$$

by the reverse triangle inequality.

Because $\operatorname{supp} \tilde{f} \subseteq K + \{x \in \mathbb{R}^m : |x| \leqslant M/\ell\}$, its compact. Demonstrating the Lipschitz continuity is technical; as such, see the supplement. □

Henceforth, $\tilde{K}$ shall denote the support of $\tilde{f}$.

Now that we have extended f, we can define g which we do as follows

$$g(x) = (2\pi)^{-m} \int F(v) \exp(i\langle v, x\rangle - |v|^2/2\lambda^2) \Psi(v/\lambda)\, dv,$$

where $\lambda = \sigma\Lambda$ for a TBD $\Lambda > 0$ and $F = \mathscr{F}\{\tilde{f}\}$. Moreover, $\Psi(x) = (\omega * \omega)\Big(x/\sqrt{m}\Big)$ with

$$\omega(x) \propto \Big[|x| \leqslant \tfrac{1}{2}\Big] J_\nu\Big(2j_\nu|x|\Big) \Big/ |x|^\nu,$$

where $\nu = m/2 - 1$ and the proportionality constant is such that $\psi = \mathscr{F}^{-1}\{\Psi\}$ is a pdf; see [5, §5] for more details. Note that $t^{-\nu} J_\nu(t) \to 2^{-\nu} / \Gamma(\nu+1)$ as $t \to 0$ [4, (10.7.3)], whereupon $\omega(0)$ is well-defined.

Upon recognizing g as an inverse Fourier transform, we see that g can be interpreted as $\tilde{f}$ convolved with an approximate delta function, that is, g is a smoothed version of $\tilde{f}$, in light of the convolution theorem and the scaling property of the Fourier transform.

Lemma 2. $\big\|\tilde{f} - g\big\|_\infty \leqslant \dfrac{\ell}{\lambda}\Big(2 - 2^{1/3} a m^{-2/3}\Big)\sqrt{m}$*.*

Proof. Let $Z \sim N(0, I_m)$. Essentially, we first show that

$$g(x) = \int \tilde{f}(x - s/\lambda)(\delta_Z * \psi)(s)\, ds.$$

using standard integral manipulation. Since ψ is a pdf,

$$\begin{aligned} |\tilde{f}(x) - g(x)| = \bigg|\tilde{f}(x) - \int \tilde{f}(x - s/\lambda)(\delta_Z * \psi)(s)\, ds\bigg| = \\ \bigg|\int \tilde{f}(x)(\delta_Z * \psi)(s)\, ds - \int \tilde{f}(x - s/\lambda)(\delta_Z * \psi)(s)\, ds\bigg| \leqslant \\ \int |\tilde{f}(x) - \tilde{f}(x - s/\lambda)|(\delta_Z * \psi)(s)\, ds \leqslant \\ \frac{\ell}{\lambda}\int |s|(\delta_Z * \psi)(s)\, ds, \end{aligned}$$

so all that remains is to bound the integral. For the remaining details, see the supplement. □

Before we introduce h, it is useful to rewrite g in a yet different form. Since

$$\begin{aligned} (2\pi)^{-m} \int F(v) \exp(i\langle v, x\rangle - |v|^2/2\lambda^2) \Psi(v/\lambda)\, dv = \\ g(x) = \int \tilde{f}(x - s/\lambda)(\delta_Z * \psi)(s)\, ds \in \mathbb{R}, \end{aligned}$$

the inverse Fourier transform integral equals its own real part, that is, $g(x)$ is equal to

$$
\begin{aligned}
(2\pi)^{-m}\int |F(v)| c(v,x)\exp(-|v|^2/2\lambda^2)\Psi(v/\lambda)\,dv &= \\
(2\pi)^{-m}\lambda^m \int |F(\lambda w)| c(\lambda w, x)\exp(-|w|^2/2)\Psi(w)\,dw &= \\
(2\pi)^{-m/2}\lambda^m \int |F(\lambda w)| c(\lambda w)\delta_Z(w)\Psi(w)\,dw &= \\
(2\pi)^{-m/2}\lambda^m \mathbb{E}\Big(|F(\lambda \mathpzc{n})|\Psi(\mathpzc{n}) c(\lambda \mathpzc{n}, x)\Big)&,
\end{aligned}
$$

where $c(v,x) \coloneqq \cos(\langle v, x\rangle + \arg F(v))$ and $\mathpzc{n} \sim N(0, I_m)$. We now approximate the above expectation by

$$
h(x) = \left\{\frac{\lambda}{\sqrt{2\pi}}\right\}^m \mathbb{E}\Big(|F(\lambda \mathpzc{n})|\Psi(\mathpzc{n})\Big[|\mathpzc{n}| > \vartheta\sqrt{m}\Big] c(\lambda \mathpzc{n}, x)\Big).
$$

Lemma 3. $\|g - h\|_\infty \leqslant \dfrac{2\ell R}{\sqrt{\pi m}} V_m \left\{\dfrac{R\vartheta\lambda}{\sqrt{2\pi/e}}\right\}^m.$

Proof. Since $\|\mathscr{F}\{\Diamond\}\|_\infty \leqslant \|\Diamond\|_1$, and ψ is a pdf,

$$
\begin{aligned}
|g(x) - h(x)| &\leqslant \\
\left\{\frac{\lambda}{\sqrt{2\pi}}\right\}^m \mathbb{E}\Big||F(\lambda \mathpzc{n})|\Psi(\mathpzc{n})\Big[|\mathpzc{n}| \leqslant \vartheta\sqrt{m}\Big] c(\lambda \mathpzc{n}, x)\Big| &\leqslant \\
(2\pi)^{-m/2}\lambda^m \|\tilde{f}\|_1 \mathbb{P}\Big\{|\mathpzc{n}| \leqslant \vartheta\sqrt{m}\Big\}&.
\end{aligned}
$$

The remaining details may be found in the supplement. □

In Section II, we said that h would be an "infinite width" RVFL with ReLU activation functions and Gaussian inner weights. The following lemma corroborates this claim.

Lemma 4. *$h = \mathbb{E}(G(\mathpzc{w}, \mathpzc{b})\rho(\langle \mathpzc{w}, \Diamond\rangle + \mathpzc{b}))$ on K, where*

- $G(w,b) = -2\sigma R\sqrt{m}\Lambda^2(2\pi)^{-m/2}\lambda^m|F(\Lambda w)|\Psi(w/\sigma)\Big[|w| \geqslant \vartheta\sigma\sqrt{m}\Big]\cos(\Lambda b - \arg F(\Lambda w))$*;*
- $\mathpzc{b} \sim \mathrm{Unif}\Big[-\sigma R\sqrt{m}, \sigma R\sqrt{m}\Big]$*;*
- $\mathpzc{w} \sim N(0, \sigma I_m)$.

Proof. The statement follows from straightforward manipulations of unwieldy integrals, that are crucially relying on the fact that the ReLU activation function satisfies "$\ddot{\rho} = \delta_0$". As always, the details can be found in the supplement. □

We now want to approximate our "infinite width" RVFL with a finite width one. Essentially, we are going to utilize Hoeffding's inequality to approximate the expectation h by a sample mean.

There is a problem, however. Using Hoeffding's inequality directly would yield an upper bound for

$$
\mathbb{P}\left\{\left|\frac{1}{n}\sum_{p=1}^n H_p(x) - \mathbb{E}(H(x))\right| > t\right\},
$$

where $H(x) = G(\mathpzc{w}, \mathpzc{b})\rho(\langle \mathpzc{w}, x\rangle + \mathpzc{b})$ and $H_1, \ldots, H_n$ are iid copies of H. Instead, we need an upper bound for

$$
\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{p=1}^n H_p - \mathbb{E}(H)\right\|_K > t\right\}.
$$

Arguably the cleanest solution is to find a measurable selector, that is, a random variable X_n such that

$$
\left|\frac{1}{n}\sum_{p=1}^n H_p(X_n) - \mathbb{E}(H(X_n))\right| = \left\|\frac{1}{n}\sum_{p=1}^n H_p - \mathbb{E}(H)\right\|_K.
$$

Such a measurable selector turns out to exist [2, thm 18.3] and can be used to deduce the following bound.

Lemma 5. *Let $N_n = \frac{1}{n}\sum_{p=1}^n H_p$ and $t > 0$. Then*

$$
\begin{aligned}
\mathbb{P}\Big\{\|N_n - \mathbb{E}(H)\|_K > t\Big\} &\leqslant \\
2\exp\left(-\frac{n}{2}\left(\frac{t}{2R^2\sqrt{m}(2\pi)^{-m/2}\lambda^{m+1}(1+1/\vartheta)\ell|\tilde{K}|}\right)^2\right)&.
\end{aligned}
$$

Proof. Technical; see the supplement. □

We are now ready to put Lemmas 2, 3, and 5 together to yield the desideratum. By setting

$$
\begin{aligned}
\alpha &= \frac{m(m+2)}{m^2+3m+1}; \\
\beta &= \frac{1}{m^2+3m+1}; \\
\Lambda &= \frac{1}{\sigma} \times \frac{\ell}{\alpha\varepsilon}\Big(2 - 2^{1/3}am^{-2/3}\Big)\sqrt{m}; \\
\frac{1}{\vartheta} &= \left\{\frac{1}{\beta\varepsilon} \times \frac{2\ell R}{\sqrt{\pi m}} V_m R^m\right\}^{1/m} \frac{\lambda}{\sqrt{2\pi/e}},
\end{aligned}
$$

we obtain by design that $\|f - g\|_\infty \leqslant \alpha\varepsilon$ and $\|g - h\|_\infty \leqslant \beta\varepsilon$. As such, letting $\gamma = 1 - \alpha - \beta$ yields that

$$
\begin{aligned}
\mathbb{P}\Big\{\|f - N_n\|_K > \varepsilon\Big\} = \mathbb{P}\Big\{\|\tilde{f} - N_n\|_K > \varepsilon\Big\} &\leqslant \\
\mathbb{P}\Big\{\|\tilde{f} - g\|_\infty + \|g - h\|_\infty + \|h - N_n\|_K > \varepsilon\Big\} &\leqslant \\
\mathbb{P}\Big\{\alpha\varepsilon + \beta\varepsilon + \|h - N_n\|_K > \varepsilon\Big\} &= \\
\mathbb{P}\Big\{\|h - N_n\|_K > \gamma\varepsilon\Big\} &\leqslant \\
2\exp\left(-\frac{n}{2}\left(\frac{\gamma\varepsilon}{2R^2\sqrt{m}(2\pi)^{-m/2}\lambda^{m+1}(1+1/\vartheta)\ell|\tilde{K}|}\right)^2\right) &\leqslant \eta,
\end{aligned}
$$

upon plugging in the lower bound for n and doing a lot of simplifications, chief among them being

$$
\begin{aligned}
|\tilde{K}| &\leqslant |K + \{x \in \mathbb{R}^m : |x| \leqslant R\}| \\
&= V_m R^m \frac{|K + \{x \in \mathbb{R}^m : |x| \leqslant R\}|}{|\{x \in \mathbb{R}^m : |x| \leqslant R\}|} \\
&= V_m R^m \exp(d(K)\ln 2).
\end{aligned}
$$

It follows from the fact that $\tilde{K} \subseteq K + \{x \in \mathbb{R}^m : |x| \leqslant M/\ell\}$ and $M \leqslant \ell R$ (this is derived while proving Lemma 3 in the supplement). We also used the following inequalities.

- Since $1 - x \leqslant e^{-x}$, it follows that
$$\left(2 - 2^{-1/3}am^{-2/3}\right)^{2m} = 4^m\left(1 - a(2m)^{-2/3}\right)^{2m} \leqslant 4^m \exp(-2^{1/3}am^{1/3}).$$
- Since $V_m = \pi^{m/2}/\,\Gamma(m/2+1)$ [4, (5.19.4)], inequality [4, (5.6.1)] yields that
$$V_m \leqslant \frac{1}{\sqrt{\pi m}}\left(\frac{2\pi e}{m}\right)^{m/2}.$$
Ergo, $V_m^2 \leqslant \dfrac{1}{\pi m}\left(\dfrac{2\pi e}{m}\right)^m$ and $V_m^{2/m} \leqslant 2\pi e/m$.

Finally, note that N_n is indeed an RVFL with outer weights of the form $G(\mathbb{w}, \mathbb{b})/n$.

Acknowledgements

Palina Salanevich is partially supported by NWO Talent programme Veni ENW grant, file number VI.Veni.212.176.

References

[1] Guy Bresler and Dheeraj Nagaraj. *A Corrective View of Neural Networks: Representation, Memorization and Learning*. 2020. arXiv: 2002.00274v2 `[cs.LG]`.

[2] D Charalambos and Border Aliprantis. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer-Verlag Berlin and Heidelberg GmbH & Company KG, 2013.

[3] CL Philip Chen and John Z Wan. "A rapid learning and dynamic stepwise updating algorithm for flat neural networks and the application to time-series prediction". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 29.1 (1999), pp. 62–72.

[4] *NIST Digital Library of Mathematical Functions*. https://dlmf.nist.gov/, Release 1.1.9 of 2023-03-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

[5] Werner Ehm, Tilmann Gneiting, and Donald Richards. "Convolution Roots of Radial Positive Definite Functions with Compact Support". In: *Transactions of the American Mathematical Society* 356.11 (2004), pp. 4655–4685.

[6] Daniel Hsu et al. *On the Approximation Power of Two-Layer Networks of Random ReLUs*. 2021. arXiv: 2102.02336v2 `[cs.LG]`.

[7] Rakesh Katuwal, Ponnuthurai N. Suganthan, and Le Zhang. "An ensemble of decision trees with random vector functional link networks for multi-class classification". In: *Applied Soft Computing* 70 (2018), pp. 1146–1153.

[8] A. K. Malik et al. *Random vector functional link network: recent developments, applications, and future directions*. 2022. arXiv: 2203.11316v1 `[cs.NE]`.

[9] Deanna Needell et al. *Random Vector Functional Link Networks for Function Approximation on Manifolds*. 2022. arXiv: 2007.15776v2 `[stat.ML]`.

[10] Ling Tang, Yao Wu, and Lean Yu. "A non-iterative decomposition-ensemble learning paradigm using RVFL network for crude oil price forecasting". In: *Applied Soft Computing* 70 (2018), pp. 1097–1108.