International Journal of Applied Mathematics

Volume 36 No. 2 2023, 205-251
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
doi: http://dx.doi.org/10.12732/ijam.v36i2.5

WAVELET NEURAL NETWORKS VERSUS WAVELET-BASED NEURAL NETWORKS

Lubomir T. Dechevsky, Kristoffer M. Tangrand [§]

Faculty of Engineering Science and Technology
UiT – The Arctic University of Norway
P.O. Box 385, N-8505 Narvik, NORWAY

Abstract: This is the first paper in a sequence of studies including also [33] and [34] in which we introduce a new type of *neural networks* (NNs) – *wavelet-based neural networks* (WBNNs) – and study their properties and potential for applications. We begin this study with a comparison to the currently existing type of *wavelet neural networks* (WNNs) and show that WBNNs vastly outperform WNNs. One reason for the vast superiority of WBNNs is their advanced hierarchical tree structure based on *biorthonormal multiresolution analysis* (MRA). Another reason for this is the implementation of our new idea to incorporate the wavelet tree depth into the neural width of the NN. The separation of the roles of wavelet depth and neural depth provides a conceptually and algorithmically simple but very highly efficient methodology for sharp increase in functionality of swarm and deep WBNNs and rapid acceleration of the machine learning process.

AMS Subject Classification: 68T07, 42C40, 46E35, 65T60, 68Q32
Key Words: artificial intelligence, swarm, deep evolutionary, neural networks, wavelet-based, machine learning, fastest learning, maximal compression, neural activation, threshold, decreasing rearrangement, Sobolev embedding

1. Introduction

The purpose of this paper is to propose a new approach to machine learning of geometric manifolds in $\mathbb{R}^n$, where $n = 1, 2, 3, 4, \ldots$ using single-layer or deep

Received: March 12, 2023

[§]Correspondence author

neural networks (NNs) based on Riesz unconditional bases of biorthonormal wavelets.

The first attempt to marry the theory of NNs with wavelet theory dates back to the early 1990s [3]. This initial study gave rise to particularly constructed NNs which were named by the authors of [3] as *wavelet NNs* (WNNs). In the course of the next twenty years, the theory and applications of WNNs were studied by numerous authors. The results of these studies have been summarized in [5] which constitutes a comprehensive account of the current status of the study of WNNs. Before going into the mathematical details of the construction and functioning of WNNs, let us note that this type of networks was introduced relatively early, when wavelet theory was still quite new to the developers of applications in the field of Artificial Intelligence (AI). Due to this, WNNs make use only of a very small subset of the useful properties of wavelet bases. Thus, while the theory of WNNs relies on the basic property of wavelet basis functions that they are dilations and translations of one and the same function, this theory ignores the more advanced properties of wavelet bases related to *Multi-Resolution Analysis* (MRA). As a consequence of this, methods using WNNs are no more than a variant of meshless kernel estimation methods. The typical representative of these meshless methods are the ones using radial basis functions [6]. The only essential difference with the variant of WNNs is that radial basis functions are replaced by tensor-product functions with sufficient number of vanishing moments. Not surprisingly, the mathematical apparatus used with WNNs is identical with the one for radial basis functions: iterative gradient or subgradient optimization methods. Unfortunately, these iterative methods can guarantee providing the global extrema only when the respective criterial functionals (objective functions) are convex. (Of course, in the particular case when the convex criterial functional is quadratic, possibly with linear constraints, in addition to iterative methods there exists also a broad variety of methods of computational linear algebra.) In practical applications, however, the realistic criterial functionals are most often non-convex, with multiple local extrema and saddlepoint singularities. In this general situation, the optimization methods used with WNNs and radial basis functions produce only local extrema which are close to the global extrema only if a very good initial starting point of the iterative algorithm is proposed. The usual defense of this type of results is to claim that all the local extremal values are close in value to the global extremal value. Here is a typical exposition of this type of argument [7]: "..., In particular, it was commonly thought that simple gradient descent would get trapped in poor local minima – weight configurations for which no small change would reduce the average error. In practice, poor local minima are rarely a problem with large networks. Regardless of the initial conditions, the system nearly always reaches solutions of very similar quality. Recent theoretical and empirical results strongly suggest that local minima are not a serious issue in general. Instead, the landscape is packed with a combinatorially large number of saddlepoints where the gradient is zero, and the surface curves up most dimensions and curves down in the remainder. The analysis seems to show that saddlepoints with only a few downward curving directions are present in very large numbers, but almost all of them have very similar values of the objective function. Hence, it does not much matter which of these saddlepoints the algorithm gets stuck at."

Some critical analysis of the above text in [7] is due, as follows. While there is some rationale in the above claims for large and very large sample and network sizes, these claims cannot be accepted even as basic "rules of thumb"; expressions like "rarely", "nearly always", "very similar", "strongly suggest", "not a serious issue in general", "seems to show", "almost all", "does not much matter" are not good replacements for logical quantors. The argument about saddlepoints with only very few negative components in the signature (that is - to use fuzzy terminology in the spirit of [7] - 'saddlepoints which are almost local minima'), is also unconvincing as a qualitative statement without any criteria or means for quantitative measurement. Even if a saddlepoint in a problem with a very large size has only one downward-curving dimension, the respective value of the criterial functional (objective function) can be much larger than the global minimum, if the downward curve is sufficiently steep. The one rigorous conclusion that can be drawn from the above excerpt of [7] is that, once the criterial functional (objective function) ceases to be (globally) convex, iterative gradient/subgradient optimization is no longer a reliable approach to achieving quality learning results. The only way to achieve best (or, at least, sufficiently high) quality results is to start from a very good initial point of the iterative process, but the traditional way of achieving this is by human intervention, i.e., the use of natural, rather than artificial intelligence. In fact, the authors of [7] acknowledge this in another excerpt of their text, as follows: "... The conventional option is to hand design good feature extractors, which requires a considerable amount of engineering skill and domain expertise. But this can all be avoided if good features can be learned automatically using a general-purpose learning procedure. This is the key advantage of deep learning ...". Our comment to this excerpt is that in the general case of non-convex criterial functionals, gradient search only plays the role of an auxiliary tool for improvement of already good results. Achieving these good initial results using AI is thus claimed in [7] to be the main aim of deep learning, and, in general, this aim cannot be attained by only using local optimization methods. The major weakness of WNNs proposed in [3] is that the crucial problem of finding a good initial starting point for iterative local optimization has only been addressed by the vague recommendation that some 'explicit link between the network coefficients and the wavelet transform' should be provided. This weakness persists also in the later developments and upgrades of WNNs discussed in [5]. Our present study shows that for sufficiently large samples this weakness can be overcome, at least partially, via tools from functional analysis. Namely, there is a rigorous mathematical general way to automatically improve the quality of the starting point of the iterative optimization process, valid for all cases when the criterial functional can be interpreted as distance between two mathematical objects. Since, to the best of the authors' knowledge, this systematic approach seems to be new in the context of NNs (and much more certainly so in the specific context of WNNs), we shall outline its main idea already in this early stage of our exposition, as follows. If the metric criterial functional of an optimization problem has an *equivalent metric which can be computed efficiently without the need of iterative optimization*, then this equivalent metric d_1:

$$0 < c_0 d_1(x,y) \leq d(x,y) \leq c_1 d_1(x,y) \tag{1}$$

can be used to generate a consistently good starting point for the optimization of the original metric d, provided that the equivalence constants $c_j,\ j = 0,1$, with $0 < c_0 \leq 1 \leq c_1 < \infty$, do not depend on x, y and, in numerical problems, they are independent of the sample size of the numerical data (in the sequel of this exposition we shall use the notation $d \asymp d_1$). In most practical applications in numerical analysis the metrics d and d_1 are induced by respective equivalent norms or quasinorms [9] or their seminorm variants. This refers not only to deterministic quantities, but also in the indeterministic case, e.g. when considering equivalent risks in statistical estimation. A typical model example, with various applications in deterministic approximation and statistical risk estimation, is the Peetre $\mathbb{K}$-*functional* between Lebesgue space L_p and homogeneous Sobolev space $\dot{W}_p^k,\ 1 \leq p \leq \infty,\ k \in \mathbb{N}$, [9],

$$K(h^k,\ f;\ L_p, \dot{W}_p^k) = \inf_{\varphi \in \dot{W}_p^k} (||f - \varphi||_{L_p} + h^k ||\varphi||_{\dot{W}_p^k}) \tag{2}$$

where $f \in L_p + \dot{W}_p^k$ (the algebraic sum of the two spaces) and h is the step (in applications, related to the sample size). An equivalent seminorm of $K(h^k, f; L_p, \dot{W}_p^k)$ is $||f - f_{k,h}||_{L_p} + h^k ||f_{k,h}||_{\dot{W}_p^k}$ where $f_{k,h}$ is the Steklov mean value of f with parameters k and h [10], the equivalence constants being independent on f and h. The numerical computation of $f_{k,h}$ is based on quadrature formulae and does not involve optimization. Thus, $\varphi_0 = f_{k,h}$ can be used as a starting point of iterative optimization. The quality of φ_0 as initial solution of the optimization problem depends on the size of the equivalence constants c_0 and c_1: if $c_0 = 1 = c_1$ (isometric equivalence) then φ_0 is the exact solution of the optimization problem. In the considered example, c_1 increases rapidly with increase of k and so for a fixed step $h > 0$ (fixed sample) the quality of φ_0 as initial solution deteriorates with the increase of k. Fix $k \in \mathbb{N}$ and let $h \to 0+$ (take sufficiently large sample): since c_j, $j = 0, 1$, do not depend on h (the sample size), for sufficiently large sample sizes φ_0 will be a consistently good starting point of the iterative optimization. Our first new result in Section 2 is to use the above idea for automatic generation of an initial starting point of iterative optimization in the case of WNNs. Although satisfactory from theoretical point of view, the practical usefulness of this generation would be rather limited, because the generated initial solution of the optimization problem would be consistently close to the global optimum only for sufficiently large sample sizes. Nothing is guaranteed for large samples of any a priori fixed size, let alone samples of medium or small size (in our numerical examples in the sequel of this exposition we shall consider sample sizes with $N \geq 2^{12}$ as very large, $2^{10} < N < 2^{12}$ as large, $2^9 \leq N \leq 2^{10}$ as medium-to-large, $2^8 \leq N < 2^9$ as medium, $2^7 \leq N < 2^8$ as medium-to-small, and $1 \leq N < 2^7$ as small). Our conclusion is that WNNs can be used efficiently for finding a consistently good local extremum only for very large sample sizes.

This weakness cannot be overcome within the conceptual construction of WNNs: a more advanced construction of relevant NNs is needed which we shall introduce in the present paper and call *Wavelet-Based Neural Network* (WBNN). The principal difference between WNNs and WBNNs is explained, as follows. Let φ be a scaling function (father wavelet) and ψ be the corresponding wavelet (mother wavelet) obtained by MRA [1],[2], so that for any $j_0 \in \mathbb{Z}$ the functions

$$\varphi_{j_0 k_0}(x_1) = 2^{\frac{j_0}{2}} \varphi(2^{j_0} x_1 - k_0), \quad \psi_{jk}(x_1) = 2^{\frac{j}{2}} \psi(2^j x_1 - k), \tag{3}$$

$x_1 \in \mathbb{R}$, $j = j_0, j_0+1, ...$, $k_0 \in \mathbb{Z}$, $k \in \mathbb{Z}$, form an orthonormal basis of $L_2(\mathbb{R})$ and $\int\limits_{\mathrm{supp}\,\psi} x_1^\lambda \psi(x_1)\, \mathrm{d}x = 0$ holds for all $\lambda = 0, 1, ...$ with $\lambda < r$ for some, henceforward fixed, $r > 0$, where supp ψ is the support of ψ in $\mathbb{R}$.

Let φ be compactly supported in $\mathbb{R}$ (which implies the same for ψ). Let $B^s_{pq}(\mathbb{R}), -\infty < s < +\infty, 0 < p \leq \infty, 0 < q \leq \infty$ be the inhomogeneous Besov space with smoothness index s, metric power index p and metric logarithmic index q (a definition will be given below). Assume that $\varphi \in B^r_{\infty\infty}(\mathbb{R})$, $\psi \in$

$B^r_{\infty\infty}(\mathbb{R})$. Then, since φ and ψ are compactly supported, $\varphi \in B^r_{p\infty}(\mathbb{R})$, $\psi \in B^r_{p\infty}(\mathbb{R})$ holds for every $p : 0 < p \leq \infty$. For $x = (x_1, ..., x_n) \in \mathbb{R}^n$, $k = (k_1, ..., k_n) \in \mathbb{Z}^n$, consider [27]

$$\begin{aligned}
\varphi^{[0]}_{0k}(x) &= \varphi_{0k_1}(x_1)\varphi_{0k_2}(x_2)\varphi_{0k_3}(x_3)...\varphi_{0k_n}(x_n)\\
\psi^{[1]}_{jk}(x) &= \psi_{jk_1}(x_1)\varphi_{jk_2}(x_2)\varphi_{jk_3}(x_3)...\varphi_{jk_n}(x_n)\\
\psi^{[2]}_{jk}(x) &= \varphi_{jk_1}(x_1)\psi_{jk_2}(x_2)\varphi_{jk_3}(x_3)...\varphi_{jk_n}(x_n) \qquad (4)\\
&\quad \cdots\\
\psi^{[2^n-1]}_{jk}(x) &= \psi_{jk_1}(x_1)\psi_{jk_2}(x_2)\psi_{jk_3}(x_3)...\psi_{jk_n}(x_n)\ .
\end{aligned}$$

Denote $\varphi^{[0]} = \varphi^{[0]}_{00}$, $\psi^{[l]} = \psi^{[l]}_{00}$, $l = 1, 2, ..., 2^n - 1$. Then, $\varphi^{[0]} \in B^r_{p\infty}(\mathbb{R}^n)$, $\psi^{[0]} \in B^r_{p\infty}(\mathbb{R}^n$ for any $p : 0 < p \leq \infty$, where $\psi^{[l]}$ is orthogonal to all polynomials of n variables of total degree less than r. Besides, $\{\varphi^{[0]}_{0k}, \psi^{[l]}_{jk}\}_{k\in\mathbb{Z}^n, j=0,1,...,2^n-1}$ is an orthonormal basis of $L_2(\mathbb{R}^n)$.

Moreover, for $f \in B^s_{pq}(\mathbb{R}^n), 0 < p \leq \infty, 0 < q \leq \infty, n(\frac{1}{p} - 1)_+ < s < r$,

$$f(x) = \sum_{k\in\mathbb{Z}^n} \alpha_{0k}\varphi^{[0]}_{0k}(x) + \sum_{j=0}^{\infty}\sum_{k\in\mathbb{Z}^n}\sum_{l=1}^{2^n-1} \beta^{[l]}_{jk}\psi^{[l]}_{jk}(x) \qquad (5)$$

for Lebesgue almost everywhere (Lebesgue – a.e.) $x \in \mathbb{R}^n$ holds, where $\alpha_{0k} =< \varphi^{[0]}_{0k}, f >= \int_{\mathbb{R}^n} \varphi^{[0]}_{0k}(x)f(x)dx$, $\beta^{[l]}_{jk} =< \psi^{[l]}_{jk}, f >$ and $a_+ = \max\{a, 0\}, a \in \mathbb{R}$. Convergence in (5) is in the quasinorm topology of the inhomogeneous Besov space $B^s_{pq}(\mathbb{R}^n)$ and, in view of the lower constraint about s, also in every Lebesgue point of f, i.e., Lebesgue a.e. on $\mathbb{R}^n$. Here, $B^s_{pq}(\mathbb{R}^n)$ admits the following quasinorm in terms of wavelet coefficients:

$$\|f\|_{B^s_{pq}(\mathbb{R}^n)} = \left\{\Big(\sum_{k\in\mathbb{Z}^n}|\alpha_{0k}|^p\Big)^{\frac{q}{p}} + \sum_{j=0}^{\infty}\left[2^{j[s+n(\frac{1}{2}-\frac{1}{p})]}\Big(\sum_{k\in\mathbb{Z}^n}\sum_{l=1}^{2^n-1}|\beta^{[l]}_{jk}|^p\Big)^{\frac{1}{p}}\right]^q\right\}^{\frac{1}{q}} . \qquad (6)$$

The construction introduced in (3-6) above generates an MRA with an orthonormal wavelet basis

$$\{\varphi_{0\mu}, \mu \in \mathbb{Z}\} \cup \{\psi^{(l)}_{j\nu},\ j = 0, 1, ..., \nu \in \mathbb{Z}^n, l = 1, ..., 2^{n-1}.\} \qquad (7)$$

A typical example of such compactly supported wavelets are the Daubechies wavelets [1] which will be the ones used in the remaining part of this paper. It is possible to generalize this construction to generate a broader class of MRAs based on bi-orthonormal wavelets [2]. These are of considerable interest in the case of polynomial spline-wavelets which are the type preferred in image processing for $n = 2$ and surface processing for $n = 3$. In this case it is imperative to use bi-orthonormal and not orthonormal spline-wavelets, because only in the proper bi-orthonormal case can the spline-wavelet be compactly supported. (Moreover, an additional advantage in image processing is that there exist proper bi-orthonormal spline-wavelets which are compactly supported and whose graphs are symmetric.) There are no MRAs with compactly supported orthonormal polynomial spline-wavelet bases. We intend to consider the use of bi-orthonormal compactly supported spline-wavelets in a subsequent publication dedicated to deep image learning.

An important property of (bi)-orthonormal MRAs which follows from (3) is that $j = 0$ in (4) can be replaced by any $j_0 \in \mathbb{Z}$, such that (5) continues to hold true with $j = 0$ replaced by $j = j_0$. In this case, (6) defines an equivalent norm in $B^s_{pq}(\mathbb{R}^n)$ for $p \geq 1$, $q \geq 1$ (quasinorm for $0 < p < 1$ and/or $0 < q < 1$) with equivalence constants dependant on j_0. (The concept of equivalent metrics continues to hold true for quasinorms, because quasinormed abelian groups are metrizable [9, Section 3.10] – see also Section 3.

Consider now the above construction with $j_0 \in \mathbb{Z}$. Let $J \in \mathbb{Z}$ be such that $j_0 < J < \infty$ and consider the truncation $\sum_{j=j_0}^{J}$ of the series $\sum_{j=j_0}^{\infty}$ in (5) and (6). This defines a subspace $V_J \subset B^s_{pq}(\mathbb{R}^n)$ such that

$$V_J = \mathrm{span}\big(\{\varphi_{j_0\mu} : \mu \in \mathbb{Z}\} \cup \{\psi^{[l]}_{j\nu} : l = 1, ..., 2^n - 1, \nu \in \mathbb{Z}^n, j = j_0, ..., J\}\big). \quad (8)$$

Due to the properties of MRA, the following sequence of nested inclusions holds:

$$V_{j_0} \subset V_{j_{0+1}} \subset ... \subset V_J \subset V_{J+1} \subset ... \quad (9)$$

with

$$\overline{\bigcup_{j=j_0}^{\infty} V_j} = L_2(\mathbb{R}^n) \quad (10)$$

holds where $\bar{X}$ is the topological closure in Y of $X \subset Y$, where Y is a complete topological space. (In the case of MRA, the complete topological space Y is L_2 with respect to the topology induced by the inner product in L_2 or, equivalently, the norm in L_2.) Consider also the spaces $W_j = \mathrm{span}\{\psi^{[l]}_{jv} : l = 1, ..., 2^{n-1}, \nu \in \mathbb{Z}^n\}$, where $j = j_0, ..., J$.

By the properties of MRA, $f \in V_J$ admits two equivalent representations, as follows:

$$\sum_{k_J \in \mathbb{Z}^n} \alpha_{Jk_J} \varphi_{Jk_J}(x) = f(x)$$

$$= \sum_{k_{j_0} \in \mathbb{Z}^n} \alpha_{j_0 k_{j_0}} \varphi_{j_0 k_{j_0}}(x) + \sum_{j=j_0}^{J} \sum_{k_j \in \mathbb{Z}^n} \sum_{l_j=1}^{2^n-1} \beta_{jk_j}^{[l_j]} \psi_{jk_j}^{[l_j]}(x), x \in \mathbb{R}^n, \tag{11}$$

where the equalities in (11) are in the sense of Lebesgue – a.e. Invoking the introduced spaces W_j, (11) can be equivalently rewritten as

$$V_J = V_{j_0} \bigoplus W_{j_0} \bigoplus W_{j_{0+1}} \bigoplus \ldots \bigoplus W_J = V_{j_0} \bigoplus \bigoplus_{j=j_0}^{J} W_j. \tag{12}$$

In applications involving processing of numerical data with sample size N, J is chosen dependent on $N : J = J(N)$. Since $\int_{\mathbb{R}^n} \psi(x)dx = 0$, the usual selection of $J(N)$ is such, that the size of the support of $\psi_{Jk_J}^{[l_J]}$ is comparable to the average step h_N between adjacent data points:

$$\text{diam}(\text{supp}\, \psi_{Jk_J}^{[l_J]}) \asymp h_N, \text{where}\, h_N \asymp \frac{1}{N}, \tag{13}$$

with equivalence constants independent of N. In view of the definition of $\psi_{Jk_J}^{[l_J]}$, (13) implies

$$J(N) \asymp \log_2 N \tag{14}$$

with equivalence constants independent of N. With this selection, the father wavelet (scaling function) φ_{Jk_J} acts as a consistent approximation of the Dirac δ-function at the point x, as long as $x \in \text{supp}\, \varphi_{Jk_J}$, and the rate of this approximation improves with the number of consecutive vanishing moments of ψ additional to the condition $\int\limits_{\mathbb{R}} \psi(x)dx = 0$ needed for consistent approximation (these would be $\int\limits_{\mathbb{R}} x\psi(x)dx = 0$, $\int\limits_{\mathbb{R}} x^2\psi(x)dx = 0$ etc.). With this selection, the coefficient α_{Jk_J} is taken to be equal to the value of f at the point where φ_{Jk_J} is concentrated as a δ-function.

So far, we have been considering (3 - 12) for $B_{pg}^s(\mathbb{R}^n)$ of functions defined on the whole space $\mathbb{R}^n$. This means that the subspaces V_j and W_j, $j \in \mathbb{Z}$, are (countably) infinite-dimensional. To make the construction computationally feasible, in numerical applications we limit the consideration to only those $f \in B_{pq}^s(\mathbb{R}^n)$ which are *compactly supported* with diam(supp f) comparable to the diameter of the convex hull of the numerical data set (this numerical data set is finite, therefore its convex hull is a bounded subset of $\mathbb{R}^n$, so its closure is compact $(n < \infty)$). In this case, the subspaces $V_j, W_j, j \in \mathbb{Z}$ are all finite-dimensional, with dimensions depending on j, the distribution of the data set and its sample size N.

Now, let us extend the consideration to include also $f \in B^s_{pq}(\Omega)$, where $\Omega \subset \mathbb{R}^n$ is a nonvoid compact hyper-rectangle, i.e.,

$$\Omega = \mathop{\times}_{i=1}^{n} [a_i, b_i] \tag{15}$$

where " $\times$ " indicates Cartesian product, and $-\infty < a_i < b_i < +\infty$, $i = 1, ..., n$. The construction (3-12) continues to hold also in this case which is very similar to the case of $f \in B^s_{pq}(\mathbb{R}^n)$ such that it is compactly supported, with *supp* f contained in the closure of the convex hull of the numerical data set. In the present case, the finite dimensions of V_j, W_j depend on j, the sizes of $b_i - a_i$, $i = 1, ..., n$, and the sample size N. Moreover, $j_0 \in \mathbb{Z}$ is bounded from below by $diam(\Omega)$:

$$j_0 \asymp \log_2 \mathrm{diam}(\Omega), \tag{16}$$

with constants of equivalence possibly dependent on n, but not on N. There is one notable technical modification: the orthonormal wavelet bases in V_j and W_j, $j = j_0, j_0 + 1, ..., J$, contain *boundary-corrected* wavelet basis functions [11], [12]. For the theory of deep learning of n-dimensional manifolds developed here, there is no principal difference between the case of compactly supported $f \in B^s_{pq}(\mathbb{R}^n)$ and the case of $f \in B^s_{pq}(\Omega)$ with Ω a compact hyper-rectangle in $\mathbb{R}^n$. Therefore, in our application we shall focus on the former one of these two cases, to avoid the construction of boundary-corrected wavelets.

Now, we are in a position to provide a definition of WNNs which is equivalent to the original definition in [3] and [5], but is in a new form which allows an insightful comparison with the new type of WBNNs.

Consider the left-hand side (LHS) of the identity in (11). Define first a 1-layer WNN with its Jk_J-th node being a neuron processing the α_{Jk_J}-th coefficient in the expansion in the LHS of (11). The edges of the WNN's graph are only the ones connecting the input to the Jk_J-th neuron and the ones connecting the Jk_J-th neuron to the output neuron where received results are summed up. In contrast to this construction, repeat the 1-layer NN but with 1–1 correspondence to the α-coefficients in V_{j_0} and the β-coefficients in $W_j, j = j_0, ..., J$ (in the right-hand side (RHS) of (11)). The definition of the edges of the graph of the 1-layer WBNN is the same as with the previous 1-layer WNN construction. The widths of the so-defined 1-layer WNN and WBNN are, of course, the same. A crucial advantage of the WBNN layer is its telescopic ordering which incorporates *the wavelet depth* into the *neural width* of the WBNN layer. Adding *neural depth* to the 1-layer WNN and WBNN is done in one and the same way: the next layers are added as intermediate between the already defined 1st layer and the output neuron, and each intermediate layer has exactly the same structure and ordering as the 1st layer.

As we shall see in the next sections, the learnability conditions and universal approximation theorems for each of WNNs and WBNNs ensure that 1-layer NNs (with the widths specified via the LHS and RHS of (11) and the compactness of $\operatorname{supp} f$) are sufficient for learning every element of the range of the respective approximation theorem. From this point of view, deep WNNs and WBNNs are theoretically redundant, but as we shall see, they provide a highly efficient computing architecture for acceleration of the rate of convergence of the approximation process by using iterative algorithms. A maximally sparse structure of the edges between the l-th and the $(l+1)$-st layer should be used (a neuron on the $(l+1)$-st level is only connected with its corresponding neuron at level l by way of the 1–1 correspondence between levels l and $l+1$). In practice, in the context of WNNs, the intermediate levels of the NN are used for iterative local optimization starting with the initial approximation provided at level 1. Although the performance of the deep WNN is expected to be better than the one of the 1-layer WNN, this can be expected to be noticeable only for very large sample sizes and respective very large number of iterations (very deep WNN). For example, in the case of Fig. 4 of [3] the number of iterations is 10000. In the case of WBNNs, the quality of initial approximation is expected to be very high, due to the efficient use of the wavelet depth within the layer of neural depth 1. As a consequence, in comparison with the very large sample size needed for acceptable performance of WNN, it can be expected for the initial approximation of the 1-layer WBNN to be sufficiently good for large to medium samples sizes, while as the ultimate approximation achieved by a deep WBNN (with the 1-layer WBNN as its initial layer) to be sufficiently good already for medium to small samples.

Notice the distinction we make between *learnability conditions* and *universal approximation theorems* for a given type of single-layer NN computing $f : \mathbb{R}^n \to \mathbb{R}$ for a given $n \in \mathbb{N}$. (To study the general case of parametric manifolds on $\mathbb{R}^n$, i.e., $f : \mathbb{R}^n \to \mathbb{R}^m$, $m \in \mathbb{N}$, $n \in \mathbb{N}$, it is sufficient to study it coordinate by coordinate, i.e., for $m = 1$ only.)

A *universal approximation theorem* (UAT) for a single-layer NN of width N is a *qualitative consistency* result (i.e., refers to existence of convergence in a given topology without specifying *quantitative rates of convergence*) when $N \to +\infty$ under the assumption of an *activation function* of specific type being used in the neural computation. For example, in the case of sigmoid activation, the respective UAT is due to Cybenko [13], [14] [1], and refers to continuous functions, while in the case of *Rectified Linear Unit* (ReLU) activation [4],

[1] In [3] Cybenko's work has been imprecisely and incompletely cited. Here we provide the relevant corrected and complete citation [13], [14]

the respective UAT refers to the more general class of functions in L_1 (see [4, Theorem 1]. The *learnability set* (LS) for a given single-layer NN of width N is the largest set of $f : \mathbb{R}^n \to \mathbb{R}$ which can be approximated by neural computation via this NN when $N \to +\infty$ *without the invocation of a specific activation function*, i.e., when *activation is via the default identity function.*

As mentioned above, both LSs and UATs are qualitative consistency results. In the next sections we shall show that it is possible to obtain quantitative upgrades of LSs and UATs where the consistency results are strengthened to results about concrete rates of approximation.

In Theorem 1 (Section 2) we obtain a new result for the established WNNs: we propose a type of activation which is shown to lead to optimal performance of WNNs and show that even optimal performance of WNNs is vastly outperformed by WBNNs.

In Section 3, in Theorems 2 and 3 we obtain new results about the learnability via WNNs and WBNNs and in Corollary 4 we show that WBNNs can be used to learn efficiently not only any regular distribution in $L_{1,loc}$ but also singular distributions like the Dirac delta and its derivatives. In the same section we provide the general characteristics (i–iii) of the rich diversity of activation operators that can be used in machine learning via WBNNs of univariate and multivariate manifolds in two, three and higher-dimensional spaces. Here we establish the principal differences between non-threshold and threshold activation in learning fractal and piecewise smooth manifolds, respectively.

In Section 4 we briefly address the importance of interconnection and interaction between *swarm AI* and *deep evolutionary AI* and the relevance to computational implementations using CPU and GPU parallelism. In Section 5 we introduce a new activation method based on the concept of *decreasing rearrangement* in functional analysis and function space theory. Theorem 5 is a uniform approximation theorem (UAT) providing qualitative proof of the consistency of the learning process when using the *decreasing rearrangement activation.* Theorem 6 provides an important quantitative upgrade of the UAT in Theorem 5 by showing that decreasing rearrangement activation of WBNNs results in machine learning process which is optimal in two key aspects: *fastest learning* and *maximal compression.* In Section 6 we consider four representative model examples which are then subjected to comprehensive graphical comparison the results of which have been systematized into a collection of conclusions and comments. Section 7 is comprised of the proofs. In the concluding Section 8 we discuss the connection of the present results with the studies in [33], [34], as well as some additional computational and research topics.

2. A new result about WNNs

In this section we shall show how the idea of using equivalent metric (as discussed in Section 1) can be used to generate an initial solution in a single-layer WNN which is a good starting point for local optimization in a deep WNN (having as 1st layer the said single-layer WNN). The universal approximation theorem invoked in [3] is Cybenko's classical result, [13], [14], valid for continuous functions f. As noted in [5], after the publication of [3] more general learnability conditions and universal approximation theorems were derived about WNNs. Namely, results about UATs for ReLU NNs were formulated and proved in [4] in relevance to the larger space of Lebesgue-integrable functions $f \in L_1(\mathbb{R}^n)$. The ideas of the proofs of the results in [4] make it possible to identify the LS for sufficiently wide NNs, including the case of WNNs with (14).

To present this result, here we shall use terminology which will allow us to compare this result to the respective result for WBNNs (derived in section 3 below). Namely, in the case of wavelets on $\mathbb{R}^n$, learnable by sufficiently wide single-layer WNN are all *regular distributions* f in $\mathcal{S}'(\mathbb{R}^n)$ - the dual of the Laurent Schwartz space $\mathcal{S}(\mathbb{R}^n)$ [15]. In the case of boundary-corrected wavelets on Ω-compact hyper-rectangle in $\mathbb{R}^n$ (see Section 1), learnable by sufficiently wide single-layer WNN are all regular distributions $f \in D'(\Omega)$ - the dual of the space $D(\Omega)$ [15] [2]. In both cases, the regular distributions f are exactly the elements of $L_{1,loc}(\Delta)$, $\Delta = \mathbb{R}^n$ or $\Delta = \Omega \subset \mathbb{R}^n$ (for Ω see Section 1). Here, as usual, $L_{1,loc}$ consists of all g defined Lebesgue – a.e. on Δ and such that for every compact subset $C \subset \Delta$ the statement $f\chi_C \in L_1(\Delta)$ holds true, where χ_C is the characteristic function of C :

$\chi_C(x) = 1$ for $x \in C$ and $\chi_C(x) = 0$ for $x \in \Delta \setminus C$.

In the case $\Delta = \Omega$, Ω considered in Section 1 is a compact in $\mathbb{R}^n$, therefore, $L_{1,loc}(\Omega) = L_1(\Omega)$. Notice that in the wavelet context the width of WNN is exponential (solving (14) for N yields the equivalent $N \asymp 2^J$) and sufficiently large for $L_{1,loc}$ to be learnable via single-layer WNN, according to [4]. Therefore, 'deepening' the WNN does not result in increasing the set of learnable functions. However, deep WNN may offer the following advantage: while the universal approximation theorem for sufficiently wide single-layer WNN provides only consistency of the approximation (i.e., convergence exists but may be arbitrarily slow), the use of a deep WNN with the said single-layer WNN forming its 1st layer may result in *increasing the speed (rate) of approximation.*

Now we shall formulate a model problem for which we shall be able to automatically generate a single-layer WNN that is asymptotically optimal with

[2] In [15] a slightly different notation is used: D_Ω instead of $D(\Omega)$ and D'_Ω instead of $D'(\Omega)$.

respect to the paradigm based on (1) in Section 1. According to this paradigm, local optimization with starting point at the automatically generated solution at the 1st layer of a deep WNN is expected to provide the global minimum for asymptotically large sample sizes ($N \to +\infty$).

Model problem. Let $n = 1$. From a random sample with size N, learn the density f of an absolutely continuous cumulative distribution function $F : \mathbb{R} \to [0, 1]$. In this case, $f = F'$, $f \in L_1(\mathbb{R})$, $f \geq 0$ on $\mathbb{R}$, $\int_{\mathbb{R}} |f(x)|dx = \int_{\mathbb{R}} f(x)dx = 1$.

This problem was addressed in [10] and [16], as follows. The density f was approximated by a (father) wavelet expansion, using the basis of a frame more general than a biorthonormal upgrade of LHS in (11). To consider the construction in [10] and [16] strictly in our present context, it is necessary to consider only those particular cases for which the frame is orthonormal, thus corresponding to the LHS in (11). In [16, Remark 2.3.2] were identified all cases when the frame used in [10], [16] is orthonormal: namely, these are exactly the cases in the LHS of (11) where the scaling function φ in (11) is of Haar type, i.e., the normalized characteristic function

$$\varphi(x) = \frac{1}{a} \cdot \chi_{[-\frac{a}{2}, \frac{a}{2}]}(x - x_0), \tag{17}$$

where $a > 0, x_0 \in [-\frac{a}{2}, \frac{a}{2}]$, and a is chosen so, as to match with the selection of j_0 in the RHS of (11).

With this choice of φ in the LHS of (11), the random estimator of f is obtained by plugging in the LHS of (11) the *empirical density*

$$\hat{f}_N(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_i), \tag{18}$$

where $\delta(\cdot)$ is the delta-function and $\{x_i,\ i = 1, ..., N\}$ is the sample data set. Here, as earlier, the selection of the level J in (11) is such that (14) holds. In view of (18), the Jk_J-th neuron in the WNN associated with the basis function φ_{Jk_J} in the LHS of (11) computes the *empirical coefficient*

$$\hat{\alpha}_{Jk_J} = \frac{1}{N} \sum_{i=1}^{N} \varphi_{Jk_J}(X_i),\ k_J \in \mathbb{Z} \tag{19}$$

In [16] the selection of metric in which the risk is measured was relevant to the expectation of "the generalized Cramér-von Mises loss" [16, Sections 2.2 and 2.3]. One very valuable feature of the estimates of this risk obtained in [16] was that they revealed the precise interconnection between the density's smoothness and the weight of its tails as $x \to \pm\infty$. In our present study we made the natural assumption of compactness of the density's support in correspondence with the compactness of the convex hull of the sample data set. For densities with compact support with fixed diameter of the support, the risk estimates in [16] simplify and are only dependent on the smoothness of the density, and the following new result holds true.

Theorem 1. *Under the above assumption about the compact support of the density f, assume $f \in B^s_{p\infty}(\mathbb{R})$, $0 < p \leq \infty$, $0 < s < 2$, and (14) holds. Let the single-layer WNN associated with the Haar-type basis (17) used in (11) be defined as above, with (19) holding true. Let the risk of estimating f via the empirical density $\hat{f}_N$ in (18) be defined as in [16, Section 2.2]. Then,*

(i) The risk $R(f, \hat{f}_N)$ of learning f by the neural computations (19) in the LHS of (11) is:

$$R(f, \hat{f}_N) \asymp N^{-\frac{s}{1+2s}} \tag{20}$$

with constants of equivalence dependent on s and the fixed size of the support of f, but not on the choice of $f \in B^s_{p\infty}(\mathbb{R})$.

(ii) The rate in the RHS of (20) is asymptotically optimal *in the sense of [16, Theorem 2, 3.2].*

As a consequence of Theorem 1, under its assumptions, the single-layer WNN computing (19) generates automatically an element of V_J in (12) which is a good starting point for optimization search in V_J when the sample size N (and $J \asymp \log_2 N$) is asymptotically large ($N \to \infty$). For such N and J, using a deep WNN upgrade of the single-layer WNN (with the latter being the 1st layer of the deep WNN) it is possible to obtain an essential improvement of the learning of f within V_J and, possibly, even obtain globally optimal solution of the iterative optimization search in V_J performed by the deep WNN computing architecture. Thus, we have provided an instance when the equivalent-metric paradigm based on (1) in Section 1 offers an efficient AI alternative of the use of natural intellect in designing good starting point of deep WNN-compatible iterative optimization search, as discussed in [7] – see Section 1. Although results of the type of Theorem 1 provide a rigorous mathematical justification of the use of AI based on deep WNNs, in practice, notable improvement can be generally expected only, or almost only, for very large samples sizes N.

Let us note two additional new features in Theorem 1 and its proof:

- Theorem 1 shows that the optimal estimation rate can be achieved when the activation function is the default identity (i.e., for the empirical density).
- The quantitative result involving rates is achieved in Theorem 1 under less restrictive assumptions for WNNs than the assumptions on generic NNs for the qualitative universal approximation theorem in [4, Theorem 1] in the sense that the latter NN must be *fully connected* (i.e., with *densely distributed edges* between the nodes of the NN) while the former WNN has very sparse edge distribution.

In the remaining part of this exposition, we shall show that the alternative of using WBNNs associated with the RHS of (11) and (12) provides highly efficient AI algorithms with quality practical results achieved already for medium to small sample sizes N.

3. WBNNs: learnability and universal approximation

To study learnability conditions and universal approximation theorems in the case of WBNNs it will be necessary to study some properties of the scale of Besov spaces $B^s_{pq}(\mathbb{R}^n)$ as defined via (5, 6). (The case of $B^s_{pq}(\Omega)$ using boundary-corrected wavelets in (5, 6) can be studied, mutatis mutandis, but our focus will continue to be on the case $\Omega = \mathbb{R}^n$.) To study the necessary aspects of the properties of the Besov-space scale $\{B^s_{pq}(\mathbb{R}^n),\ 0 < p, q \leq \infty, s \in \mathbb{R}\}, n \in \mathbb{N}$, we need some preparation, as follows.

- For the concept of *quasinorm* (or c-norm (c-quasinorm) with $c \geq 1$ being the constant in the quasi-triangle inequality), we refer to [9, Section 3.10]. See also there for the definition of *quasinormed abelian groups.*
- By [9, Lemma 3.10.1], a quasinormed abelian group A with c-quasinorm $||.||_A,:\ c \geq 1$, is *metrizable*, in the sense that the ρ-power A^ρ of A with 1-quasinorm $||.||^\rho_A$, $0 < \rho = \frac{1}{1+\log_2 c} \leq 1$, is a metric space with $d(a,b) = ||a-b||^\rho_A$.
- A necessary and sufficient condition for a normed space A to be *complete* (i.e. for A to be a *Banach space* is given in [9, Lemma 2.2.1].
- The previous item is being generalized in [9, Lemma 3.10.2] to a necessary and sufficient condition for a quasinormed abelian group A to be *complete.*

(If A is not only abelian group, but also a vector space, then, endowed with the quasinorm $||.||_A$, it is called *quasinormed space* and, if it is also complete, *quasi-Banach space*.

- For example, consider the vector space l_2 of all sequences: $= (x_1, ..., x_n, ...), n \in \mathbb{N}, x_j \in \mathbb{R}$ or $x_j \in \mathbb{C}, j \in \mathbb{N}$, with quasinorm $||x||_{l_r} = (\sum_{j=1}^{\infty} |a_j|^r)^{\frac{1}{r}}$, $0 < r < \infty$, or $||x||_{l_\infty} = \max_{j\in\mathbb{N}} |a_j|$, $r = \infty$. By the theory in [9, Section 3.10], l_r is a *Banach space* for $r : 1 \leq r \leq \infty$ and *quasi-Banach space* when $r : 0 < r < 1$. In the latter case, the constant c in the quasi-triangle inequality for $||.||_{l_r}$ is $c = 2^{\frac{1-r}{r}} > 1$; for the power ρ one gets $\rho = r \in (0, 1)$ and the 1-quasinormed abelian group $(l_r)^r$ with 1-quasinorm $||.||_{l_r}^r$ is a metric space with metric $d(a, b) = ||a - b||_{l_r}^r$.

- Using the properties of l_r from the previous item, it is possible to establish that B_{pq}^s, as defined via (5, 6), are Banach spaces for $1 \leq p \leq \infty$, $1 \leq p \leq \infty$, and quasi-Banach spaces when $0 < p < 1$ and/or $0 < q < \infty$ [17].

- Another aspect of the theory of the Besov-space scale which proves to be relevant is *the lifting property* of *the Bessel potential* J^σ, $\sigma \in \mathbb{R}$, in the Besov-space scale. Following the exposition in [17, Section 1.2.1], we define the Fourier transform $\mathcal{F}$ and its inverse $\mathcal{F}^{-1}$ first on $\mathcal{S}(\mathbb{R}^n)$, and then extend it to $\mathcal{S}'(\mathbb{R}^n)$, after which, following [17, Section 2.3.8], we define the Bessel potential $J^\sigma : \mathcal{S}'(\mathbb{R}) \to \mathcal{S}'(\mathbb{R}^n)$, as follows:

$$J^\sigma f = \mathcal{F}^{-1}[(1 + |.|^2)^{-\frac{\sigma}{2}} \mathcal{F} f], f \in \mathcal{S}'(\mathbb{R}^n), \tag{21}$$

  where $\sigma \in \mathbb{R}$. Now the σ*-lifting property* of the Bessel potential in the Besov-space scale can be formulated, as follows [17, Section 2.3.8]. J^σ acts bijectively on $\mathcal{S}'(\mathbb{R}^n)$ and the restriction of J^σ on $\mathcal{S}(\mathbb{R}^n)$ acts bijectively on $\mathcal{S}(\mathbb{R}^n)$. Moreover, if s, p, q are as in (5, 6) and $f \in B_{pq}^{s_1}(\mathbb{R}^n)$, where $s_1 \in \big(-\infty, n(\frac{1}{p} - 1)_+\big] \cup \big[r, \infty\big)$ with $\sigma : s = s_1 + \sigma$, then $J^\sigma f \in B_{pq}^s$ and formulae (5, 6) can be applied on $g = J^\sigma f$ and

$$||J^\sigma f||_{B_{pq}^s(\mathbb{R}^n)} \asymp ||f||_{B_{pq}^{s-\sigma}(\mathbb{R}^n)}. \tag{22}$$

Moreover, using (22) when $s_1 \in (-\infty, n(\frac{1}{p} - 1)_+]$, i.e., for $\sigma > 0$ allows to extend the wavelet-based representation (5) and the quasinorm definition (6) for arbitrary $s \in \mathbb{R}$, i.e., including also singular distributions like the δ-function and its derivatives which are not in $L_{1,loc}$ [1]. Indeed, choose and fix any $s_1 \in \mathbb{R}$ and select and fix σ such that $n(\frac{1}{p}-1)_+ < s-\sigma < r$. Then (5,6) will make sense for f replaced by $J^\sigma f$ and (22) can be used to define an equivalent quasinorm in $B^{s_1}_{pq}(\mathbb{R}^n)$.

Now we are ready to formulate and prove the following results about WBNNs.

Theorem 2. *Using arbitrary samples with size N with $J(N)$ satisfying (14), $f \in \mathcal{S}'(\mathbb{R}^n)$, is learnable for $N \to \infty$ by WNNs if, and only if (iff) f is also learnable by WBNNs, i.e. the learnability sets by WNNs and by WBNNs coincide.*

Theorem 3. *Let N and $J(N)$ be as in Theorem 2, and let $f \in B^s_{pq}(\mathbb{R}^n)$, $0 < p \le \infty$, $0 < q \le \infty$, $s \in \mathbb{R}$. Then, for any $r : n(\frac{1}{p}-1)_+ < r < \infty$ and orthonormal wavelet basis satisfying (3, 4) and for every $\sigma \in \mathbb{R}$ such that $s-\sigma \in (n(\frac{1}{p}-1)_+, r)$ it holds true that $J^\sigma f$ is learnable by the WBNN generated by the said wavelet basis.*

Corollary 4. *The space $B^s_{pq}(\mathbb{R}^n)$ is contained in the learnability set of WBNN for every $s \in \mathbb{R}$, $0 < p \le \infty$, $0 < q \le \infty$.*

Corollary 4 implies that the learnability set of WBNNs contains not only all regular distributions in $\mathcal{S}'(\mathbb{R}^n)$, but also singular distributions, since Besov spaces with $s < 0$ do contain singular distributions.

Theorem 2 now suggests that Corollary 4 extends also to WNNs, but here lies one big difference between the *efficient* use of WBNNs and WNNs. Recovering f from $g = J^\sigma f$ requires approximate numerical computation of $J^{-\sigma} g$ which is very numerically sensitive to errors in the computation of g especially when f can be a singular distribution. Since for a given sample with size N the quality of learning g via WNNs is expected to be much worse than via WBNNs, the deterioration of the recovery of f from g when using WNNs would be much more exacerbated compared to the use of WBNNs so that the only case of σ for which the use of WNNs could be marginally acceptable is the trivial case $\sigma = 0$. (A detailed error analysis of the computations for $\sigma \ne 0$ would require the invocation of aspects of Paley-Wiener theory [19, Sections 4–6], including sampling results of Shannon type [19, Theorem 6.4] which goes beyond the

[1] This fact is relatively easy to derive even in the n-dimensional case, using the theory of Fourier multipliers on $L_p(\mathbb{R}^n)$, $1 \le p \le \infty$, see [9, Section 6.1], [29, Chapter 1] and [41, Introduction].

study of AI aspects considered here.) Theorems 3 and 2 now imply that WNNs can be efficiently used (albeit only marginally for very large sample sizes N only) for learning $f \in B^s_{pq}(\mathbb{R}^n)$ only for the original range s, p, q for which (5) was formulated. Note that for the these values of s, p, q $B^s_{pq}(\mathbb{R}^n) \subset L_{1,loc}(\mathbb{R}^n)$.

The results obtained here for wavelet bases on $\mathbb{R}^n$ can in principle be reformulated for boundary-corrected wavelets on a compact hyperrectangle in $\mathbb{R}^n$, this modification is technically involved. For example, the lifting property of the Bessel potential has to be replaced by a respective property of fractional integro-differential operators of Riemann-Lionville, Grünwald-Letnikov, Caputo and other types under additional assumptions for each of these types [18].

As far as UAT for WBNNs is concerned, it is much more rich and diversified than UAT for WNNs, due to the much more flexible telescopic structure of the RHS in (12). While in the case of WNNs the focus has been only on sigmoid and ReLU activation, in the case of WBNNs there is a great variety of meaningful activation methods, each of which is with its own UAT and own range of practical applications. In this section we study the common features of all these activation methods and provide a classification of these methods into two general subclasses, together with the general range of applications for each of these subclasses.

Any activation method can be defined as a (generally, *nonlinear*) operator Λ acting on the space sum in the RHS of (12) and being of *shrinkage* type, i.e. having the following properties.

(i) The restriction of Λ on V_{j_0} coincides with the identity on V_{j_0}, i.e.,

$$\Lambda(\alpha_{j_0 k_0} \varphi_{j_0 k_0}) = \alpha_{j_0 k_0} \varphi_{j_0 k_0} \tag{23}$$

for every $j_0, k_0, \ldots$.

(ii) Using the Euler representation of $z \in \mathbb{C}$

$$z = |z|(\cos(\arg z) + i \sin(\arg z)), \arg z \in [0, 2\pi), \tag{24}$$

the action of Λ on the space $\bigoplus_{j=j_0}^{J} W_j$ in the RHS of (12) is defined such, that

$$\Lambda(\beta^{[l_j]}_{jk_j} \psi^{[l_j]}_{jk_j}) = \tilde{\beta}^{[l_j]}_{jk_j} \psi^{[l_j]}_{jk_j}, \tag{25}$$

where

$$\arg \tilde{\beta}^{[l_j]}_{jk_j} = \arg \beta^{[l_j]}_{jk_j} \tag{26}$$

$$|\tilde{\beta}^{[l_j]}_{jk_j}| \leq |\beta^{[l_j]}_{jk_j}| \tag{27}$$

for every $(j,\, k_j)$ participating in the formation of $\bigoplus_{j=j_0}^{J} W_j$.

Notice that when $\beta_{jk_j}^{[l_j]} \in \mathbb{R}$ (26) reduces to

$$\operatorname{sgn} \tilde{\beta}_{jk_j}^{[l_j]} = \operatorname{sgn} \beta_{jk_j}^{[l_j]}, \tag{28}$$

where for $x \in \mathbb{R}\backslash\{0\}$

$$\operatorname{sgn} x = \begin{cases} +1, & x > 0; \\ -1, & x < 0; \\ \text{undefined}, & x = 0; \end{cases} \tag{29}$$

and for the case $z = x = 0$ it is convenient to define

$$\arg 0 = \operatorname{sgn} 0 = 0. \tag{30}$$

Clearly, Λ has the special property that it preserves the directrice and respective orientation of every basis function in V_{j_0} and $\bigoplus_{j=j_0}^{J} W_j$. It is also clear that, in general, Λ is nonlinear, since the shrinkage is individual for every basis function. Now, we divide all possible activation methods Λ with properties (i) and (ii) into two disjoint subclasses, as follows.

A. *Non-threshold-type* activation methods have the following additional property:

(iii) for any selection of the coefficient vector $\{\alpha_{j_0k_0}, \beta_{jk_j}^{[l_j]}\}$ in the RHS of (11), such that $\beta_{jk_j}^{[l_j]} \neq 0$ for some triple (j, k_j, l_j), it is fulfilled that $\tilde{\beta}_{jk_j}^{[l_j]} \neq 0$ holds true. (In other words, there is only reducing $|\beta_{jk_j}^{[l_j]}| > 0$ without ever "killing" the coefficient $\tilde{\beta}_{jk_j}^{[l_j]}$, i.e., having $|\tilde{\beta}_{jk_j}^{[l_j]}| = 0$.

B. We shall say that the activation method Λ is of *threshold-type*, if (iii) is not fulfilled for Λ.

In the second part of this study we shall study an important model example of activation of WBNNs resulting in *learning geometric manifolds with compression*. The analysis of concrete examples will show that there is an intrinsic separation of geometric manifolds into ones that are highly compressible and ones that are not. From a geometric point of view, the latter class of manifolds will be identified as *fractal-type* while the former class consists of manifolds of *piecewise-smooth type*. Our forthcoming study [33] of diverse activation methods of both threshold and non-threshold type (type B and A) will demonstrate that activation of threshold type is performing well when learning piecewise-smooth manifolds, while activation of non-threshold type performs well when learning manifolds of fractal type.

4. Particle vs multiagent simulation and swarm vs. deep evolutionary AI

A very new research topic in AI research is establishing connection between *swarm* AI and deep neural networks by the invocation of *evolutionary algorithms* [20], [21]. Scientifically, this is a very new field, but conceptually it appeared in some of the most famous early sci-fi novels [22], [23] (latest English translation [24]) which were written in the first few years after the concept of AI emerged as a term at the workshop "Dartmouth Summer Research Project on Artificial Intelligence" at Darthmouth College, Hannover, NH, USA in 1956, to designate a specialized branch of cybernetics. Both of the novels of Fred Hoyle and Stanisław Lem successfully predicted the development of important modern scientific trends: the latter – nanotechnology and swarm AI; the former – deep learning by AI systems and connections with evolutionary algorithms.

Our present interest to the connection between swarm and deep evolutionary AI is only limited to its computational aspects. From this limited point of view, the above-said connection can be considered as a particular case of particle simulation and multiagent simulation (i.e., simulation of systems involving large numbers of relatively simple agents vs. simulation of systems involving small to moderate number of agents with relatively high level of individual intelligence features). Notice that the most efficient simulation of each of these two types of system is performed on different types of parallel computing architectures.

(a) Swarm of sufficiently broad single-layered NNs (including single-layered WBNNs with (14)) – CPU parallelism – (relatively expensive) large-size multi-CPU supercomputing architectures; e.g., hypercubic [25].

(b) Deep (multi-layered, sufficiently broad) NNs (including deep WBNNs with (14)) – GPU parallelism – (relatively cheap) small-size multi-GPU computing architectures using GPGPU-programming – currently in CUDA, and more recently, Python [28].

Using connections between swarm intelligence and deep NNs [20], [21], it is possible to emulate the performance of the expensive computing architectures in item (a) by the cheap computing architectures in item (b), but at the inevitable price of some loss of efficiency. Ideally, hybrid multi-CPU multi-GPU computing architectures should be recommended.

5. Best activation of WBNNs for fastest learning and maximal compression

In [26] was considered and systematized a rich diversity of threshold and non-threshold wavelet shrinkage methods for *nonparametric statistical estimation* of *densities* and *denoising of nonparametric regression functions.* In [34] we shall show that each of these shrinkage methods gives rise to respective activation of WBNNs, generating highly efficient learning algorithms. Moreover, in some cases these learning algorithms can be shown to be *best possible* with respect to certain aspects which are important for applications. As a model example of the high efficiency of learning with WBNNs, we shall study here the activation induced by only one of these wavelet shrinkage methods, namely, the one discussed in [26, Appendix B10 b)].

For the sake of maximal clarity, we shall consider here only the simplest univariate case $n = 1$. This will be a very clear illustration of the optimal speed of learning and compression in model examples of curve learning in the next sections. The general case $n \in \mathbb{N}$ and some graphical visualization for the cases $n = 2$ and $n = 3$ will be considered in [34].

Assume $f \in B^s_{pq}(\Omega)$ where $\Omega = \mathbb{R}$ or $\Omega = [a, b]$, $-\infty < a < b < \infty$, $0 < p \le \infty, 0 < q \le \infty$ and $\left(\frac{1}{p} - 1\right)_+ < s < r$. Assume also that both the metric power index p_1, the metric logarithmic index q and the smoothness index s are *exact*, that is, $f \notin B^{s_1}_{p_1 q_1}(\Omega)$ for any $p_1 : 0 \le p_1 < p, 0 < q_1 \le q$ and any $s_1 : s_1 > s$.

As usual in our present study, when considering the domain $\mathbb{R}$, we shall be making the default assumption about compactness of $\operatorname{supp} f$ (in the case of boundary-corrected wavelets and $\Omega = [a, b]$ with $-\infty < a < b < +\infty$, we do not need this default assumption, i.e., it is possible that $f(a+) \ne 0$ and/or $f(b-) \ne 0$). For the index triple (p, q, s) consider now the *Sobolev embedding plane* passing through the point (p, q, s), i.e.,

$$\{(\rho, \eta, \sigma) : \sigma - \frac{1}{\rho} = \tau(p, s) = s - \frac{1}{p}, 0 < \rho \le \infty, 0 < \eta \le \infty\}. \tag{31}$$

What is important about this selection is the well-known embedding

$$B^s_{pq}(\Omega) \hookrightarrow B^\sigma_{\rho\eta}(\Omega),\ \sigma - \frac{1}{\rho} = s - \frac{1}{p},\ 0 < p \le \rho \le \infty,\ 0 < q \le \eta \le \infty, \tag{32}$$

where $\Omega = \mathbb{R}$ or $\Omega = [a, b]$.

(For two quasinormed spaces A and B, the notation $B \hookrightarrow A$ ("B is embedded/imbedded in A") means $B \subset A$ and $\exists c \in (0, \infty) : ||b||_B \leq c||b||_A$ for any $b \in B$.)

Due to the Sobolev embedding/imbedding (32), our assumption $f \in B^s_{pq}(\Omega)$ implies

$$f \in B^\sigma_{\rho\eta}(\Omega) \tag{33}$$

for any $(\rho, \eta, \sigma) : \sigma - \frac{1}{\rho} = s - \frac{1}{p}$, $0 < p \leq \rho \leq \infty$, $0 < q \leq \eta \leq \infty$. In [26, Appendix B10 b)] it was explained that for the Besov spaces $B^\sigma_{\rho\eta}$ with (ρ, η, σ) lying on one and the same Sobolev embedding plane, an important part of the function-space theory is related with the so-called *decreasing rearrangement* of f in all Besov spaces $B^\sigma_{\rho\eta}$ with (ρ, η, σ) on the same Sobolev embedding plane. A detailed consideration of the concept of decreasing rearrangement can be found in [9, Section 1.3] and the Peetre-Krée formula [9, Theorem 3.6.1] together with [9, 3.14.5.6 and Theorem 5.2.1 (2) for $q = p$ in local notation] (see also [26, (B8)]. For our purposes in our present context it is sufficient to consider the *normalized decreasing rearrangement* of the β-coefficients in the series (5) and in $\bigoplus_{j=j_0}^J W_j$ in its truncation (11, 12), as follows (compare with [26, Appendix B10 b), items b1 and b2]). Recall that here we consider only the case $n = 1$ in (5-16) - in particular, in (5-8, 11) this implies $l = 1$. Thus, in the sequel of the present definition of *decreasing rearrangement*, we shall be skipping the index l.

b1) Fix $j_0 \in \mathbb{Z}$ (with no loss of generality, it can be assumed that $j_0 = 0$). Consider all (j, k) in $(5, 6)$ such that $\operatorname{supp} \psi_{jk} \cap \operatorname{supp} f \neq \emptyset$. Denote the set of all such (j, k) by $I(f, \psi) = I(f, \psi, j_0)$. It is clear that for every fixed $j = j_0, j_0+1, ...$ in the generalization of (5, 6) involving j_0 the number M_j of elements of $I(f, \psi)$ from the j-th level does not exceed $c(f, \psi)$. 2^j, for some $c(f, \psi) \in (0, \infty)$. Therefore, M_j is finite for any $j = j_0, j_0 + 1, ...$, but $M = \sum_{j=j_0}^{\infty} M_j$ is, generally, *not finite*. On the other hand, for the truncation (11,12) the number $m(j_0, J) = \sum_{j=j_0}^{J} M_j$ is finite, with $M = \lim_{J\to\infty} m(j_0, J)$. Denote by $i(f, \psi, j_0, J)$ the subset of $I(f, \psi, j_0)$ such that (j, k) participates in the formation of the truncation (11) and $\bigoplus_{j=j_0}^J W_j$ in (12).

The number of elements of $i(f, \psi, j_0, J)$ is

$$m(j_0, J) \leq c(f, \psi) \cdot 2^{j_0} \sum_{k=0}^{J-j0} 2^k$$

$$= c(f,\psi)2^{j_0}\cdot\frac{2^{J-j_0+1}-1}{2-1} = c(f,\psi)(2^{J+1}-2^{j_0}) \le 2c(f,\psi)2^J, \tag{34}$$

regardless of the choice of j_0.

b2) Recalling that $\tau = \tau(p,s) = s - \frac{1}{p}$ in (31), for every $(j,k) \in I(f,\psi)$ normalize $|\beta_{jk}|$ by multiplying with the factor $2^{j(\tau+\frac{1}{2})}$ and consider the decreasing rearrangement $\{b_\nu : \nu = 1, ..., M\}$ of the (possibly, infinite) set $\{2^{j(\tau+\frac{1}{2})}|\beta_{jk}| : (j,k) \in I(f,\psi)\}$.

In the case of the truncation (11, 12), we get the subset $\{2^{j(\tau+\frac{1}{2})}|\beta_{jk}| : (j,k) \in i(j_0,J)\}$ which is finite, with number of elements $O_{f,\psi}(2^J)$, according to (34).

For this model case, the activation operator $\Lambda = \Lambda_\delta$ is of threshold type, with threshold $\delta \in (0,\infty)$, defined in the following way. Let the decreasing rearrangement of $I(f,\psi)$ be $\{b_\nu, \nu = 1, ..., M\}$, with $(j,k) \in I(f,\psi)$ being ordered in a respective sequence $\{(j_\nu,k_\nu), \nu = 1, ..., M\}$ where (j_ν,k_ν) corresponds to b_ν for any $\nu = 1, ..., M$.

Then

$$\Lambda(\beta_{j_\nu k_\nu}\psi_{j_\nu k_\nu}) = \begin{cases} 0, & \text{if } 2^{j_\nu(\tau+\frac{1}{2})}|\beta_{j_\nu k_\nu}| \in (0,\delta), \\ \beta_{j_\nu k_\nu}\psi_{j_\nu k_\nu}, & \text{if } 2^{j_\nu(\tau+\frac{1}{2})}|\beta_{j_\nu k_\nu}| \ge \delta. \end{cases} \tag{35}$$

The selection of the threshold δ depends on the concrete context of the learning process. We shall call every threshold activation method designed via the sequence of steps b1) and b2) *a decreasing rearrangement activation method.* For this type of activation method with threshold δ, the UAT corresponds to $\delta \to 0+$ and is given by the following theorem.

Theorem 5. *Let $\delta \to 0+$ in (35), and let f be as assumed above. Then, the summands in the series in the RHS of (5) can be commuted in such a way that (5) becomes*

$$\sum_{k\in\mathbb{Z}} \alpha_{j_0k}\varphi_{j_0k}(x) + \sum_{j=j_0}^{\infty}\sum_{k\in\mathbb{Z}} \beta_{jk}\psi_{jk}(x) = f(x)$$

$$= \sum_{k\in\mathbb{Z}} \alpha_{j_0k}\varphi_{j_0k} + \sum_{\nu=1}^{M} \beta_{j_\nu k_\nu}\psi_{j_\nu k_\nu}(x), \tag{36}$$

where the RHS converges to $f(x)$ Lebesgue – a.e. in $\mathbb{R}$, as well as in the topology of $B^s_{pq}(\mathbb{R})$ and $B^\sigma_{\rho\eta}(\mathbb{R})$ for any $(\rho,\eta,\sigma) : 0 < p \le \rho \le \infty, 0 < q \le \eta \le \infty$, $\sigma - \frac{1}{\rho} = s - \frac{1}{p} = \tau$.

Theorem 5 continues to hold true for boundary–corrected wavelets and $\Omega = [a,b]$, with respective modifications in (5) and (36).

We shall now upgrade the qualitative result of Theorem 5 by formulating a quantitative result which proves to be best possible in a certain sense specified below. Among all Besov spaces with (quasi)norm (6), the ones which are Hilbert spaces are exactly

$$B^s_{pq}(\Omega) \text{ with } p = q = 2 \text{ and } 0 < s < r, \tag{37}$$

where, for $n = 1$ considered here, $\Omega = \mathbb{R}$ or $\Omega = [a,b], -\infty < a < b < +\infty$. Choose arbitrary triple (p,q,s) such that $0 < p \le 2$, $0 < q \le 2$, $(\frac{1}{p} - 1)_+ < s < r$, and consider the respective triple $\rho = \eta = 2$, $\sigma - \frac{1}{\rho} = s - \frac{1}{p}$. For (5, 6) to hold for this choice of (ρ, η, σ) it is necessary that

$$0 \le \sigma < r \tag{38}$$

holds where $\sigma = 0$ corresponds to the case $B^0_{22}(\Omega) = L_2(\Omega)$. Therefore, (5, 6) hold simultaneously for the couples (p,q,s) and $(2,2,\sigma)$, iff the following inequalities and equalities are simultaneously

$$(\frac{1}{p} - 1)_+ < s < r, \quad 0 \le \sigma < r, \quad \sigma = s - \frac{1}{p} + \frac{1}{2}, \quad p \le 2, \quad q \le 2. \tag{39}$$

Solving (39) for p, q and s, we obtain

$$\frac{1}{r + \frac{1}{2}} < p \le 2, \quad 0 < q \le 2, \quad \frac{1}{p} - \frac{1}{2} \le s < r, \tag{40}$$

under which assumptions (38) holds. Consider the orthocomplement

$$O_{j_0\sigma} = O\big(V_{j_0}, B^\sigma_{22}(\Omega)\big) = V_{j_0}^{\perp(B^\sigma_{22}(\Omega))} \tag{41}$$

of V_{j_0} in $B^\sigma_{22}(\Omega)$, with respect to the inner product in $B^\sigma_{22}(\Omega)$, $0 \le \sigma < r$. This orthocomplement is well defined because $\{\varphi_{j_0k_0}, \psi_{jk_j}\}$ is an unconditional Riesz basis in all Besov spaces where (5, 6) hold, and $B^\sigma_{22}(\Omega)$ is a Hilbert space, so that $V_{j_0}^\perp$ is well-defined with respect to the inner product in $B^\sigma_{22}(\Omega)$, $0 \le \sigma < r$. For a given $f \in B^\sigma_{22}(\Omega)$, define $f_{j_0} \in V_{j_0}$ as follows

$$f_{j_0} = \sum_k \alpha_{j_0k}(f)\varphi_{j_0k}. \tag{42}$$

Let $k \in \mathbb{N}$ and consider an arbitrary subspace S_k with $\dim S_k = k$, such that

$$S_k \subset O_{j_0\sigma} \subset B^\sigma_{22}(\Omega), \tag{43}$$

and define the best-approximation functional

$$E_k(f; B^{\sigma}_{22}(\Omega)) = \inf_{s \in S_k} \inf_{S_k \subset O_{j_0\sigma}} ||f - f_{j_0} - s||_{B^{\sigma}_{22}}. \tag{44}$$

Now, we are in the position to formulate the following remarkable result.

Theorem 6. *Assume that f is as in Theorem 5 with the additional assumption that (40) holds. Then,*

$$||f - \sum_{\nu=1}^{k} \beta_{j_\nu k_\nu} \psi_{j_\nu k_\nu}||_{B_{22}^{s-\frac{1}{p}+\frac{1}{2}}(\mathbb{R})} = E_k\big(f; B_{22}^{s-\frac{1}{p}+\frac{1}{2}}(\mathbb{R})\big), k = 1, 2, ... \tag{45}$$

The result (45) holds, *mutatis mutandis*, also for the case of boundary-corrected wavelets and $\Omega = [a, b]$, $-\infty < a < b < +\infty$.

Theorem 6 shows that using a sufficiently broad (i.e., satisfying (14)) single-layered WBNN for learning curves with Besov regularity while using the current activation method results in a learning strategy which is *optimal* in the following two senses:

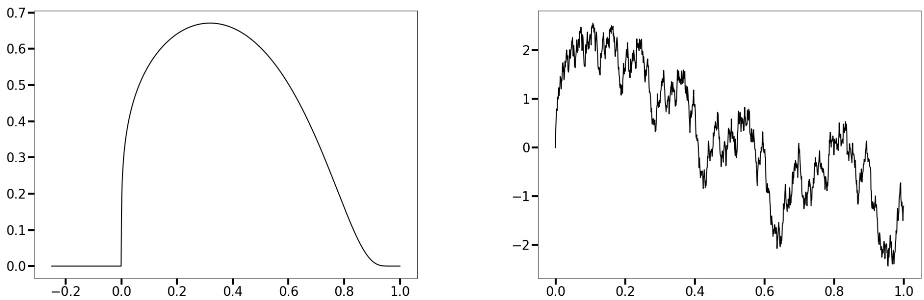
1. *Fastest learning* – using a fixed number of active neurons, the learned function is closest possible to the original, with the distance measuring the closeness being in terms of $|| \cdot ||_{B^{\sigma}_{22}}$, $0 \leq \sigma < r$, that is, taking into account not only position in space ($\sigma = 0$) but also fractional derivatives up to order r.

2. *Maximal compression* – for a benchmark determined by a fixed distance between the target function and its learned version measured in terms of $|| \cdot ||_{B^{\sigma}_{22}}$, the benchmark result is achieved with the fewest possible activated neurons.

In the remaining part of the present study, we shall illustrate graphically aspects 1. and 2. of the optimality of the learning process with WBNNs when the current activation method is being used.

6. Representative model examples

We shall consider in detail graphical visualization related to two model examples which are representative in three important aspects.



(a) "λ-tear" (b) Weierstrass type fractal

Figure 1: Piecewise smooth type vs fractal type manifolds

1. The first one (the "λ-tear") is a typical manifold of *piecewise smooth type*, while the second one (Weierstrass-type curve) is a typical manifold of *fractal type.*
2. For both of them, their *exact* metric power, metric logarithmic and smoothness index of their Besov regularity is known.
3. The exact metric power, metric logarithmic and smoothness index of Besov regularity can be selected to be the *same* for both examples, which allows for insightful graphical comparisons.

Example 1 (see Fig. 1a). *The "λ-tear"*

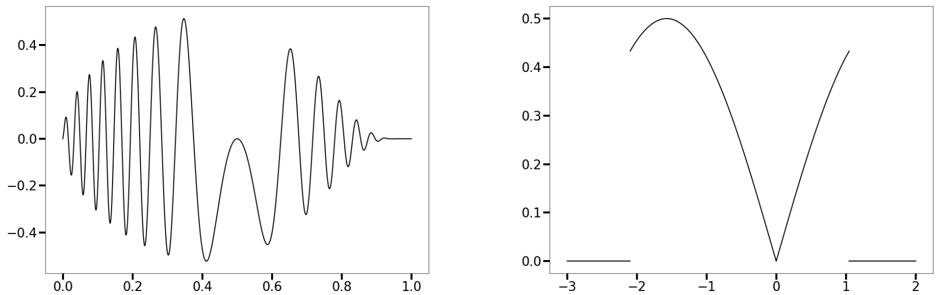
$$f_1(x) = \begin{cases} x_+^\lambda \exp(-\frac{x^2}{1-x^2}) & x \in (0,1), \\ 0 & x \in [-1,0] \cup [1,2], \end{cases} \tag{46}$$

where $\lambda \in (0,1)$. This function is analytic for $x \in [-1,0) \cup (0,1) \cup (1,2]$; it is C^∞, but not analytic at $x = 1$; at $x = 0$ it is only C^0. Its exact Besov regularity for $p : 1 \le p \le \infty$ is $f \in B_{p\infty}^{\lambda+\frac{1}{p}}(\Omega)$, where $\Omega = \mathbb{R}$ or $\Omega = [-1,2]$ [29, Proposition 2.4.2], see also [26, Section 7, Example 1].

Example 2 (see Fig. 1b). *Weierstrass-type curve*

$$f_2(x) = \sum_{k=0}^{\infty} 1.5^{-\tau k} \sin(1.5^k \times 5x), \quad x \in \mathbb{R}, \tag{47}$$

where $\tau \in (0,2)$. For the purpose of comparing with Example 1, we shall consider only the restriction of f_2 onto $\operatorname{supp} f_1$, i.e., for $x \in [0,1]$.



(a) "Double chirp" (b) Sinusoidal density

Figure 2: Curves used to study Besov regularity

The graph of f_2 is a typical *self-similar monofractal* with constant local Hölder index τ and constant local fractal dimension $2-\tau$ which is also its global fractal dimension on $[0,1]$. When considering $\Omega=\mathbb{R}$, for any compactly supported $\chi \in C^\infty(\mathbb{R})$ such that $[0,1] \subset \operatorname{supp}\chi$ and $\chi \equiv 1$ on $[0,1]$, $\chi \cdot f_2 \in B^\tau_{p\infty}(\mathbb{R})$ [29, Proposition 2.4.1 for the imaginary part G_t in local notations, with additional rescaling], see also [26, Section 7, Example 2]. For the restriction $\overline{f}_2 = f_2\Big|_{\operatorname{supp} f_1}$ in the case of boundary-corrected wavelets with $\Omega=[0,1]$, we have directly $\overline{f}_2 \in B^\tau_{p\infty}([0,1])$. For every $p: 1 \le p \le \infty$ this Besov regularity of $\overline{f}_2$ is *exact*. Clearly, when $\tau = \lambda + \frac{1}{p}$, f_1 and f_2 have exactly the same exact Besov regularity.

Besides the detailed comparative study of Examples 1 and 2, we shall study some additional geometric aspects of the learning process on two other model examples: "Double chirp" and "Sinusoidal density".

Example 3 (see Fig. 2a). *"Double chirp"*

$$f_3(x) = \sqrt[4]{x}\exp(-\frac{x^2}{1-x^2})\sin[64\pi x(1-x)], x \in [0,1], \tag{48}$$

Compare also [26, Section 7, Example 3]. The graph of f_3 is very spatially inhomogeneous, containing at the endpoints 0 and 1 two chirps of a very different nature. f_3 in (48) is a product of "λ-tear" for $\lambda=\frac{1}{2}$ and C^∞-smooth function, so its Besov regularity is exactly the same as the Besov regularity of a "λ-tear" (Example 1) for $\lambda = \frac{1}{2}$. Of special interest is to compare how the optimal learning algorithm deals in the spatially different parts of the graph for large, moderate and small samples, or for small, medium or high compression percentage.

Example 4 (see Fig. 2b). *"Sinusoidal density"*

$$f_4(x) = \begin{cases} \frac{1}{2}|\sin x| & x \in [-\frac{2\pi}{3}, \frac{\pi}{3}] \\ 0 & x \in (-\infty, -\frac{2\pi}{3}) \cup (\frac{\pi}{3} + \infty), \end{cases} \tag{49}$$

f_4 is analytic on $(-\infty, -\frac{2\pi}{3}) \cup (-\frac{2\pi}{3}, 0) \cup (0, \frac{\pi}{3}) \cup (\frac{\pi}{3}, +\infty)$; f_4 is C^0 at $x = 0$; f_4 is discontinuous at $x = -\frac{2\pi}{3}$ and $x = \frac{\pi}{3}$.

Because of the presence of the two jumps, $f_4 \in B^{\frac{1}{p}}_{p\infty}(\Omega)$, $1 \le p \le \infty$, $\Omega = \mathbb{R}$ or, for boundary corrected wavelets, $\Omega = [-\pi, \pi]$. This result about the Besov regularity of f_4 follows from the result about Besov regularity of the Heaviside step function which is present in implicit form in the embeddings in [26, Appendix B12b, item (iv)]. The function exhibits considerable spatial inhomogeneity in the neighbourhoods of the three points of singularity ($x = -\frac{2\pi}{3}$, $x = 0$ and $x = \frac{\pi}{3}$). Of particular interest is to compare the performance of the optimal learning algorithm in a neighbourhood of each of the three singularities. The comparison of the performance between the two jump-singularities at $x = -\frac{2\pi}{3}$ and $x = \frac{\pi}{3}$ should include also comparative study of the local Gibbs phenomenon.

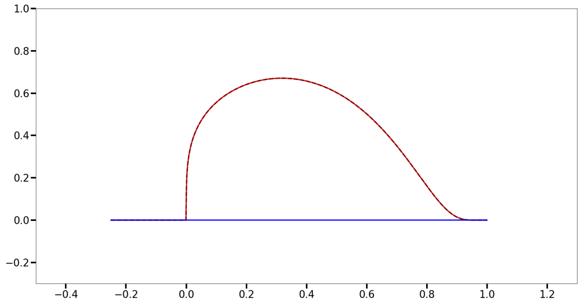
The sample sizes in Figures 1-13 are $N = 2^{10}$ or less.

First we focus on faster learning and maximal compression, according to items 1 and 2 in Section 5. The comparative graphical analysis of Figures 3–6 leads to the following conclusions.

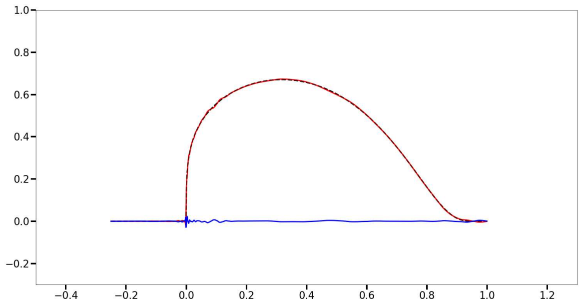
1. Examples 1, 3, 4 are of piecewise smooth type, while Example 2 is of the fractal type.

2. The decreasing rearrangement activation allows very fast learning combined with very high compression rate for the piecewise smooth curves: the quality of learning is superb at 85% compression. In comparison, retaining such high quality of learning for the fractal curve in Figure 2 is possible only at compression rate up to $3 - 4\%$. (See item (a) in each of Figures 3–6.)

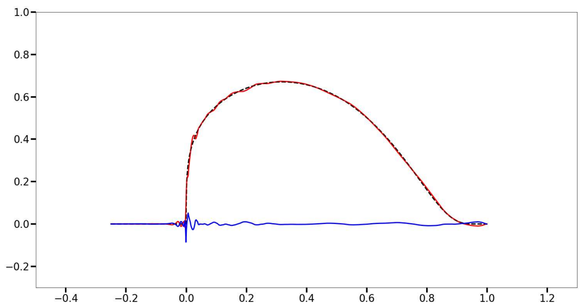
3. Approximation of the target function by the learned one is very good even for superhigh levels of compression ($98 - 99\%$). This also indicates that if the large-to-medium sample size $N = 2^{10}$ be reduced to moderate or even small sizes (cf. Section 1), the rate of learning can be expected to be quite good, while the compression rates will decrease, but remain still quite good.



(a) 85 % compression

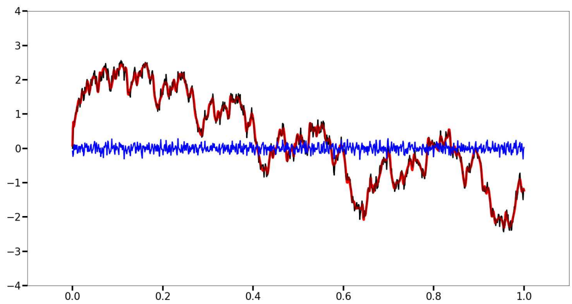


(b) 98 % compression

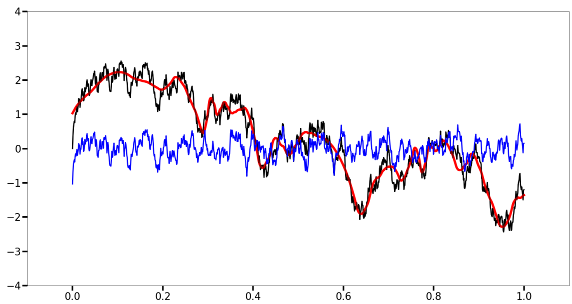


(c) 99 % compression

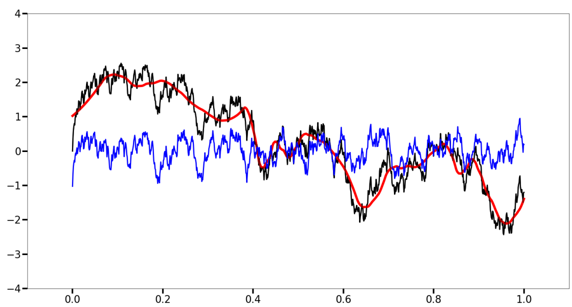
Figure 3: Target function (dashed black), learned function (red) and the error between the two (blue) for the "λ-tear" under high compression.



(a) 85 % compression

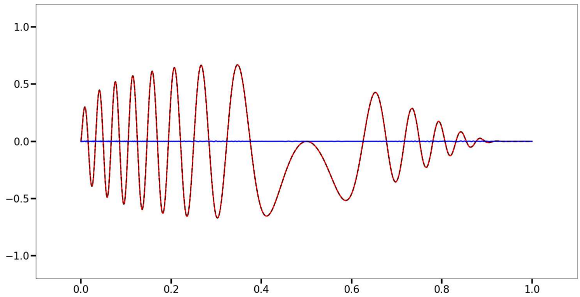


(b) 98 % compression

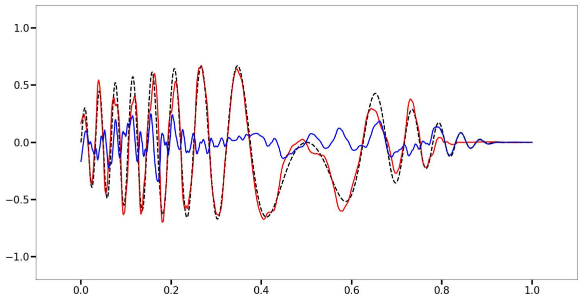


(c) 99 % compression

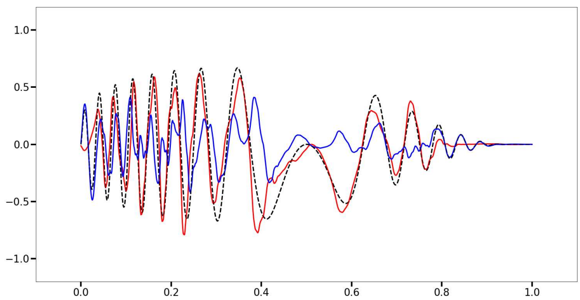
Figure 4: Target function (dashed black), learned function (red) and the error between the two (blue) for the "Weierstrass function" under high compression.



(a) 85 % compression

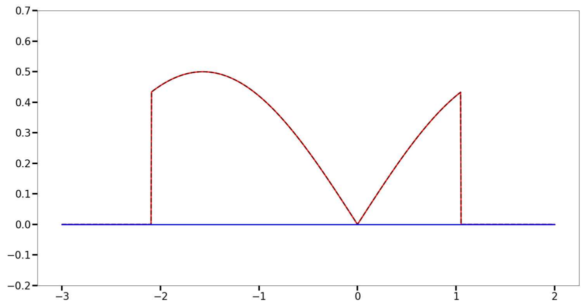


(b) 98 % compression

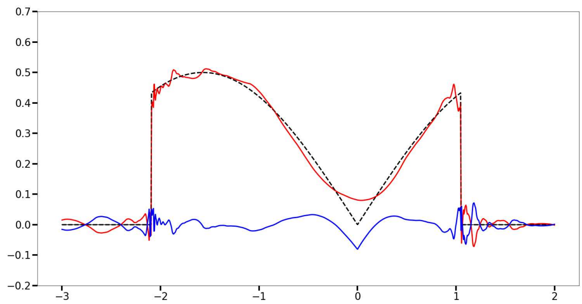


(c) 99 % compression

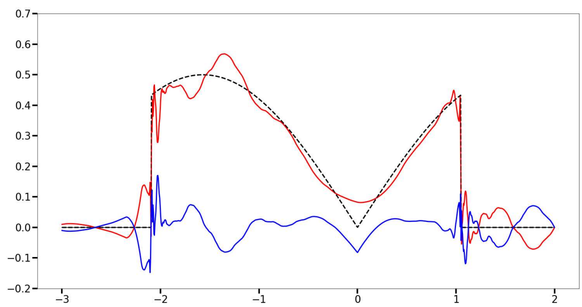
Figure 5: Target function (dashed black), learned function (red) and the error between the two (blue) for “the double chirp” under high compression.



(a) 85 % compression

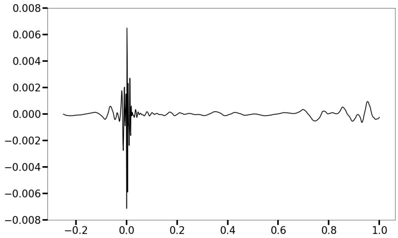


(b) 98 % compression

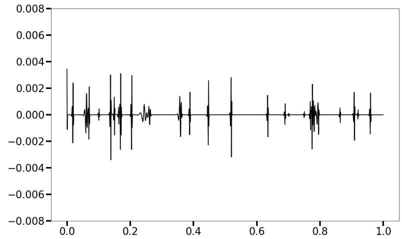


(c) 99 % compression

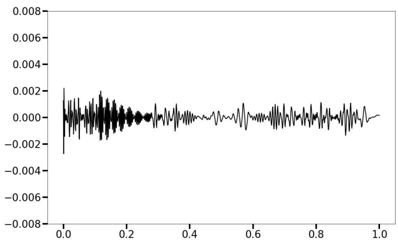
Figure 6: Target function (dashed black), learned function (red) and the error between the two (blue) for “the sinusoidal density” under high compression.



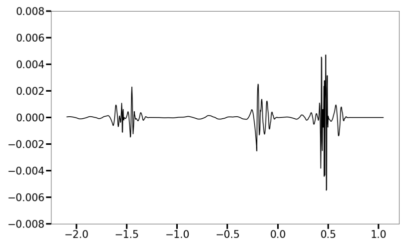
(a) Distribution of benchmark MISE for "λ-tear", at compression rate $\approx$ 99.610%. This is the benchmark MISE for all cases (a) – (d).



(b) Distribution of benchmark MISE for "Weierstrass function", attained at compression rate $\approx$ 3.418%.



(c) Distribution of benchmark MISE for "double chirp", attained at compression rate $\approx$ 83.984%.



(d) Distribution of benchmark MISE for "sinusoidal density", attained at compression rate $\approx$ 83.984%.

Figure 7: Benchmark error distribution for Examples 1-4. Benchmark error was MISE for Example 1 at compression $\approx$ 99.610%.

4. The effect of subjecting the fractal-type curve to high or superhigh compression rates is that the learned curve get smoothed out to a piecewise smooth (few isolated singularities, similar to Examples 1,3 and 4) or even smooth – no singularities at all. In [33] and [34] we shall show that if the fractality of the manifold is due to noise, then, learning the manifold with WBNNs where the decreasing arrangement activation is applied on the noisy (empirical) wavelet β–coefficients results in denoising and high-quality statistical estimation of the manifold. The level of compression resulting from this process can be useful in determining the chances for the manifold to have certain regularity.

The comparison of graphical data for the four examples on Figure 7 leads to the following observations:

5. For piecewise smooth manifolds having isolated singularities of *the first kind* only (left and right onesided limits exist at the singularity) the distribution of the benchmark MISE error is narrowly concentrated in small neighbourhoods of the respective singularities. This also shows that the vector of β–coefficients of such manifolds tends to be sparse, i.e., it contains large in absolute value coefficients on all levels j only in small neighbourhoods of the singularities. Thus, manifolds of this type are highly compressible with threshold activation methods and are the fastest to learn with WBNNs. Typical examples are 1 and 4, see Figure 7 (a) and (d), resp.

6. Piecewise-smooth manifolds with isolated singularities of *the second kind* (at least one of the onesided limits does not exist at the singularity). Typical case of this type of singularity is the presence of a chirp on the side of the missing onesided limit (for example, functions $g(x) = x^a \sin \frac{1}{x^b}$, $x > 0, a > 0, b > 0$). Chirps can be very spatially inhomogeneous and even exhibit some fractal properties (for example, the function g in the above formula, with $a = 0$, has unbounded variation in a neighbourhood of $x = 0$ and the part of its graph in this neighbourhood is infinitely long. Thus, functions with chirps take a somewhat intermediate place between the functions in item 5 and the function in the next item 7. Typical example is 3 – see Figure 7 (c).

7. Fractal-type curves, especially ones with locally constant Hölder index, gather their Besov regularity from a dense vector of β–coefficients where all, or, at least, the vast majority of β–coefficients provide significant contribution which can only be ignored at the price of slowing the rate and decreasing the quality of the learning process. In [33] and [34] we shall show that manifolds of fractal type can be learned well only using non-threshold shrinkage activation. This type of activation produces 0% compression. Threshold activation methods, including the decreasing rearrangement activation, oversmooth the manifold, thereby altering its fractal type. This explains the low compression rate when achieving the benchmark MISE in the typical Example 2 – see Figure 7 (b).

8. (Remark.) One of the main goals of research in [33] and [34] will be to design *hybrid activation strategies for adaptive learning* by composing a sequence of activation strategies of diverse – threshold and non–threshold

– nature which will be achieved by the use of *deep WBNNs*, each single-layer WBNN in which will contribute with its own activation strategy.

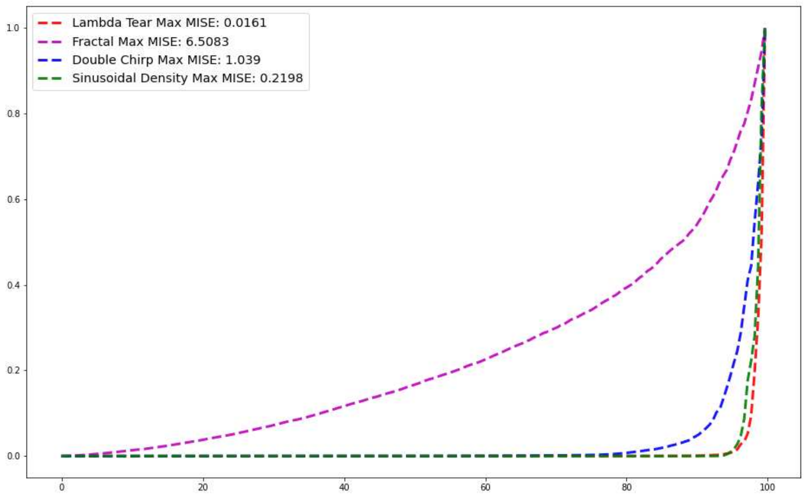
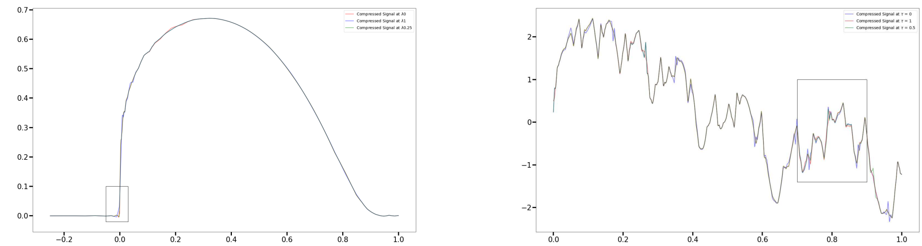


Figure 8: Compression percentage (x-axis) vs standardized relative MISE for Examples 1-4

While Figure 7 provided detailed information about the local distribution of a benchmark MISE, Figure 8 provides an insightful comparison of the ratio between compression percentage and relative MISE for each of the four considered examples.

9. The aspect, in which Figure 8 is most insightful, is the comparative determination of the fractality type of the curves in each of Examples 1–4: "the Weierstrass curve" of Example 2 exhibits markedly fractal behaviour, followed by the "double chirp" of Example 3 exhibiting a somehow 'semi–fractal' behaviour, and with the curves in Example 1 and 4 being of markedly piecewise smooth type.

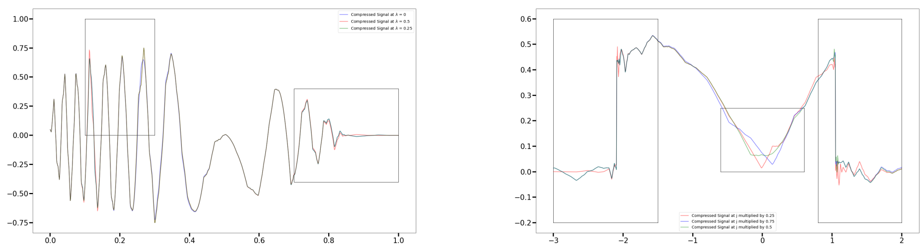
10. (Remark.) Tracing the behaviour of the "double chirp" of Example 3 in Figures 5, 7(c) and 8, some notable differences are observed with "the λ–tear" in Example 1, despite of the fact that they are both of the piecewise



(a) Zoomed region for Example 1 (see Fig. 11a), compression rate 98%

(b) Zoomed region for Example 2 (see Fig. 11b), compression rate 91%

Figure 9: Local zooming for Examples 1 and 2



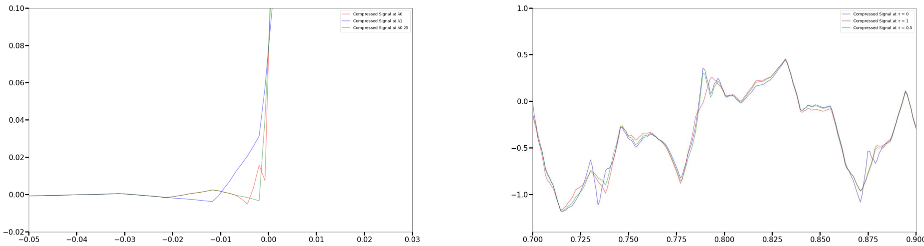
(a) Zoomed regions for Example 3 (from left to right, see Fig. 12 12a) and 12b), resp.), compression rate 96%

(b) Zoomed regions for Example 4 (from left to right, see (13a), 13b) and 13c), resp.), compression rate 98%

Figure 10: Local zooming for Examples 3 and 4

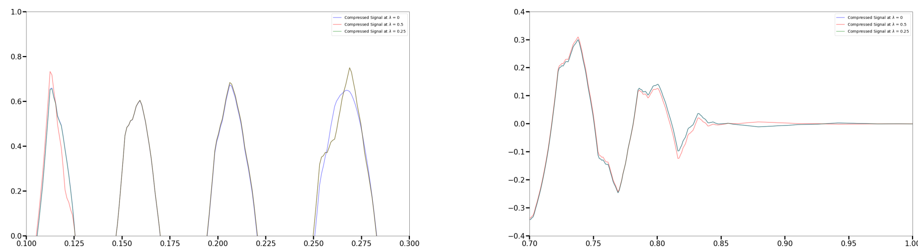
smooth type and, especially, despite of the fact that they have *exactly the same Besov regularity*. The explanation of this phenomenon is that although the Besov norms of f_1 and f_3 with the same exact parameters are both finite, the norm of f_3 is several orders of decimal magnitude larger than the norm of f_1, mainly due to the presence of the factor $64 = 2^6$ in the sine component of the formula (48) for f_3.

In Section 6 the exact Besov regularity of the example was known, and it was used in the construction of the activation operator. What if only approximate information is available about each of the parameters (p, q, s) of the Besov regularity? The algorithmically simplest way to overcome this ambiguity is to use a swarm of sufficiently broad single-layer WBNNs. As discussed in Section



(a) (b)

Figure 11: Region zoom for (a) Example 1, (b) Example 2 (see Fig. 9a and 9b), resp.



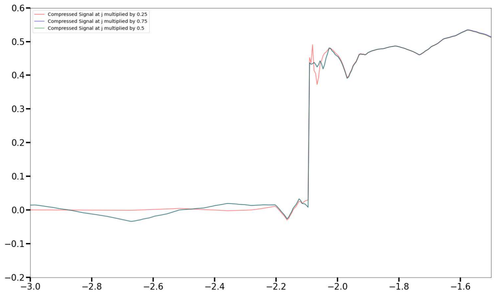
(a) (b)

Figure 12: Region zoom for Example 3: see Fig 10a

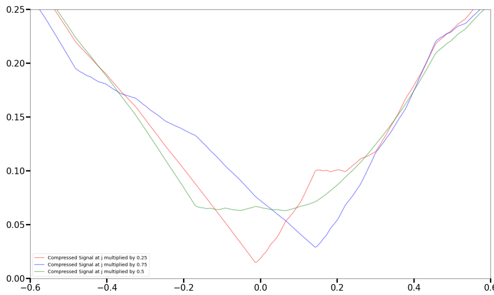
4, using a deep WBNN is possible, but requires adjustments, with some loss of efficiency. This is why here the new research topic about relation between swarm and deep evolutionary AI ([21]) is of great interest.

In Figures 9–13 we produce the graphical results of sequential emulation of the parallel learning process of a 'swarm' of 3 single-layered sufficiently broad (satisfying (14)) WBNNs, one of which is biased towards underestimating the Besov regularity (blue colour), the second one is using the exact Besov regularity information (green colour) and the last one is biased towards overestimating the Besov regularity. The graphical results for Examples 1–4 are presented in Figure 9(a) and 9(b), and Figure 10(a) and 10(b), resp. The rectangular regions on these figures, where the differences are most notable, are marked with window frames.

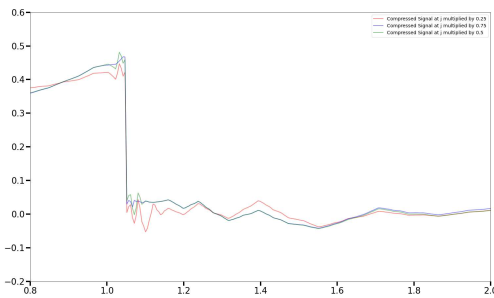
In Figures 11–13 are given zoomings of all windows in Figures 9 and 10, as



(a)



(b)



(c)

Figure 13: Region zooms for Example 4: see Fig 10b

follows.

- The window in Figure 10(a): Figure 11(a)
- The window in Figure 10(b): Figure 11(b)
- The two windows in Figure 10(a): Figure 12(a) and (b)
- The three windows in Figure 10(b): Figure 13(a), (b) and (c)

Some observations from the graphical comparison, as follows:

11. The learning process using WBNNs activated by decreasing rearrangement method proves to be *very robust with respect to errors of estimation of the Besov regularity information* when *manifolds without singularities* are being learned.

12. In the presence of singularities, the robustness in item 11 *decreases*: the more the singularities, the less the robustness, with the maximal deterioration of robustness being in the case of learning fractal-type manifolds. However, due to the relatively uniform distribution over all, or most of the β–coefficients of a fractal manifold (see item 7 above), for the low compression rate in Figure 7(b), the differences between the blue, green and red lines in Figure 9(b) would hardly be noticeable (see also item 13).

13. Due to the robustness properties in items 11 and 12 above, it makes sense to use *large-size swarms of single-layered WBNNs* only in the case of learning fractal-type manifolds, for a reason that will be explained in item 15.

14. Learning of piecewise smooth manifolds with WBNNs activated via the decreasing rearrangement method is robust with respect to small errors of underestimating or overestimating the manifold's Besov regularity, as long as compression rate is none or relatively small. This robustness *rapidly deteriorates* with the increase of the compression rate, but at the same time the quality of learning *deteriorates slowly* with the increase of the compression rate (which is an equivalent way of saying that piecewise smooth manifolds are being learnt fast). This is why in the case of learning a piecewise smooth manifold, small to medium sized swarms of sufficiently broad single-layered WBNNs are expected to be sufficient when the Besov regularity of the manifold is only approximately known, see Figure 11(a) and Figures 12 and 13.

15. In the case of fractal manifolds, the crucial difference is that, unlike the case of piecewise smooth manifolds, the quality of learning of the fractal *deteriorates rapidly together with the rapid deterioration of robustness* with respect to errors in the Besov regularity estimation, when the compression rates increases. This is why, contrary to the comparable compression rates in Figure 7(a) vs. Figure 9(a), Figure 7(c) vs. Figure 10(a) and Figure 7(d) vs. Figure 10(b), resp., there is a sharp difference in the compression rates in Figure 7(b) vs. Figure 9(b): less than 4% vs. more than 90%, resp. Because of this important difference, it can be expected that in the current context large-size swarms of sufficiently broad single-layered WBNNs would be needed to attain high quality of learning fractal manifolds. Notice that in the context of item 4, when the fractality is due to noise, the denoised manifold may still be of fractal type, or it may be of piecewise smooth type. In the context of the present item, in this case determination of the size of the swarm requires special care (this topic will be addressed in [33]).

16. Finally, let us turn our attention on the three singularities in Figure 13. Comparing the three singularities in Figure 13 with the singularity in Figure 11(a), it is observed that the latter singularity is somehow intermediate between the one in Figure 13(b) and the two in Figure 13(a) and (c). This is so, because, on the one hand, the singularities in Figure 13(b) and 11(a) are of the same type – discontinuity of the first derivative f_1', resp. f_4' and, on the other hand, $f_1'(0+) = +\infty$, which forces the graph of f_1 left of 0 to resemble that of the graph of f_4 at the left discontinuity point in Figure 13(a) and, modulo vertical axial symmetry, also the graph of f_4 in Figure 13(c). What is remarkable about the discontinuity singularities of the first kind in Figure 13(a) and (c) (and, to some less expressed extent, also about the singularity in Figure 11(a)) is the presence of *Gibbs phenomenon.* As it can be seen from Figure 3(a) and Figure 6(a), *even at high compression rates* there is *no Gibbs phenomenon at all*, which is due to the selection of *compactly supported* wavelet basis. If in this context a trigonometric basis is used there will be very significant Gibbs phenomenon even at compression rate 0%. Thus, we conclude that in the case of use of compactly supported wavelet basis, notable Gibbs phenomenon may eventually appear at jump points only at superhigh levels of compression, and it is due to uncompressed β-coefficients with low j ($j = j_0$ or j near to j_0). It is also at these superhigh levels of compression, and for the same reason, that errors in the estimation of Besov regularity can result in the decreasing rearrangement activation producing notable

differences in regions with Gibbs phenomenon.

7. Proofs

Proof of Theorem 1. Let first $s : 0 < s < 1$. Using the the definitions and notations of [16, Section 2.2] for ρ, k^*, p and q, the upper bound

$$\exists\, c_1 < +\infty : R(f, \hat{f}_N) \leq c_1 N^{-\frac{s}{1+2s}}, \tag{50}$$

where $c_1 = c_1(s, \mathrm{diam}(\mathrm{supp}(f)))$, follows after taking power $\frac{1}{\rho}$ from the two sides of [16, (2.2.3] where the choice $q = +\infty$ has been made. Under the assumptions of [16, Corollary 2.2.4] on f, taking in consideration the compactness of $\mathrm{supp} f$ implies the equivalence of the assumptions in [16, Corollary 2.2.4] and the current assumption $f \in B^s_{p\infty}(\mathbb{R})$

The lower bound

$$\exists\, c_0 > 0 : R(f, \hat{f}_N) \geq c_0 N^{-\frac{s}{1+2s}}, \tag{51}$$

where $c_0 = c_0(s, \mathrm{diam}(\mathrm{supp} f))$, follows from [16, Theorem 2.3.2] for $q = \infty$. The selection of J in (14) assumes that in all cases considered in [16, Section 2.2], the optimal level k^* defined there, always satisfies

$$j_0 \leq k^* \leq J, \tag{52}$$

which ensures the validity of all upper and lower bounds in [16, Corollaries 2.2.2-11 and Theorem 2.3.2]. Under the assumptions $q = \infty$ and compactness of $\mathrm{supp}\, f$, the norm $||.||_{p,s}$ defined in the formulation of [16, Theorem 2.3.2] can be replaced by the simpler $||.||_{B^s_{p\infty}(\mathbb{R})}$. This proves (i). The optimality of the rate $N^{-\frac{s}{1+2s}}$ in the context of the assumptions of Theorem 1 follows by the standard argument in risk estimation: with the increase of N, the bias term decreases and tends to 0 when $N \to \infty$, while the variance term increases and tends to $+\infty$ when $N \to \infty$. So, the optimal rate in N is achieved when the contributions of the bias and variance terms are equal. Under the assumptions of Theorem 1, it follows from the proof of [16, Theorem 2.2.1, under the assumptions of Corollary 2.2.4] that the rate for which the bias and variance terms are balanced is $N^{-\frac{s}{1+2s}}$. (ii) is proved.

Now let $s : 1 \leq s < 2$. In this case the proof is based on the same line of arguments, but with [16, Corollary 2.2.4] being replaced by by [16, Corollary 2.2.8] and noting that the expression $\frac{s-\frac{1}{r}+\frac{1}{q}}{1+2s-\frac{2}{r}+\frac{2}{q}}$ in [16, (2.2.4)] becomes $\frac{s}{1+2s}$ for $r = q = +\infty$. $\square$

Proof of Theorem 2. Follows straight-forwardly from the chain of equalities in (11). □

Proof of Theorem 3. To prove the theorem for every quadruple $(p,q,s,\sigma) : 0 < p \leq \infty,\ 0 \leq q \leq \infty,\ n(\frac{1}{p}-1)_+ < s+\sigma < r$ we invoke [9, Lemma 3.10.2], as follows. Assume that $g = J^\sigma f$ and $g \in B^{s+\sigma}_{pq}(\mathbb{R}^n)$, that is, the RHS of (6) is finite when s is replaced by $s+\sigma$. Then, by [9, Lemma 3.10.2], the series (5) for g is convergent in the topology of $B^{s+\sigma}_{pq}(\mathbb{R}^n)$, therefore, $g = J^\sigma f$ is learnable by the WBNN generated by the specified wavelet basis.

In the particular case $1 \leq p \leq \infty$, $1 \leq q \leq \infty$, the above proof can be simplified, by using [9, Lemma 2.2.1] instead of [9, Lemma 3.10.2]. □

Proof of Corollary 4. Follows from Theorem 3 by using the lifting property of J^σ. □

Proof of Theorem 5. The RHS of (36) is just a commuted version of its LHS. Since the basis $\{\varphi_{j_0k}, \psi_{jk}\}$ is a Riesz unconditional basis [1] all that has to be shown is that the series in the LHS of (36) is absolutely convergent, because then [9, Lemma 3.10.2] (or, alternatively in the particular case of Theorem 5 when $p \geq 1$ and $q \geq 1$, [9, Lemma 2.2.1]) implies the statement of the theorem. The absolute convergence of the LHS in (36) in B^s_{pq} follows from $f \in B^s_{pq}$, implying the the finiteness of $||\cdot||_{B^s_{pq}}$ in (6). From here, the absolute convergence of the LHS of (36) in $B^\sigma_{\rho\eta}$ for every (ρ,η,σ) specified in the theorem follows from the Sobolev embedding $B^s_{pq} \hookrightarrow B^\sigma_{\rho\eta}$, see (32). □

Proof of Theorem 6. The space B^σ_{22} is a Hilbert space and, by Theorem 5, the RHS of (36) holds true. Because of the Hilbertian geometry of B^σ_{22}, removing the term involving any one $\psi_{j_{\nu_0}k_{\nu_0}}$ from $\sum_{\nu=1}^{M} \beta_{j_\nu k_\nu}\psi_{j_\nu k_\nu}$ in (36) has the geometric meaning of orthogonal projection of

$$\sum_{\nu=1}^{M} \beta_{j_\nu k_\nu}\psi_{j_\nu k_\nu} \in \mathrm{span}\{\psi_{j_\nu k_\nu}\}_{\nu=1}^{M} \text{ onto}$$

$$\left(\sum_{\nu=1}^{\nu_0-1} + \sum_{\nu=\nu_0+1}^{M}\right)\beta_{j_\nu k_\nu}\psi_{j_\nu k_\nu} \in \mathrm{span}\left\{\{\psi_{j_\nu k_\nu}\}_{\nu=1}^{\nu_0-1}\bigcup\{\psi_{j_\nu k_\nu}\}_{\nu=\nu_0+1}^{M}\right\}.$$

Since $|\beta_{j_\nu k_\nu}|$ are ordered as decreasing rearrangement with factor $2^{j_\nu(s-\frac{1}{p}+\frac{1}{2})}$ and $\sigma : \sigma - \frac{1}{2} = s - \frac{1}{p}$, (45) follows, which proves the theorem. Note the following remarkable geometric fact which remained implicit, but is crucial for the proof of the theorem: the basis $\{\varphi_{j_0k_0}, \psi_{jk}\}$ is orthonormal only in $L_2(\sigma = 0)$ but remains orthogonal in B^{σ}_{22} $(0 \le \sigma < 2)$ which has the norm of a weighted l_2–sequence space with weight $2^{j\sigma}$. □

8. Concluding remarks

The present work is the first part of a sequence of studies dedicated to the new WBNNs. The next two parts of this series [33] and [34] are currently in preparation. The main focus in [33] will be on a detailed study of the rich variety of threshold and non-threshold activation methods for learning curves in 2, 3 and higher dimensions. [34] will be dedicated to the diverse problems which arise when learning multivariate multidimensional geometric manifolds (surfaces, volume deformations and manifolds in dimensions higher than 3). One topic will be to reduce the dimensionality of high-dimensional WBNNs to 1- and 2-dimensional WBNNs with full preservation of their functional efficiency. One important application of this approach is to enable the use of GPGPU programming algorithms for learning parametric manifolds with arbitrary number of parameters, immersed in arbitrarily high-dimensional space [38], [39]. Another topic in these two studies is to make progress in understanding the connection between learning and approximation [35] in the context of the new WBNNs. It should be noted that the essence of our new approach in the present paper – (a) separating the roles of wavelet depth and neural depth, (b) incorporating wavelet depth into the WBNN width to achieve consistency of learning, and (c) using the neural depth for accelerating the rate of consistent learning – can in principle be used also in the much more general context of arbitrary tree–based adaptive partitions [40], [36].

We conjecture that Theorem 6 can be generalized for a broader range than $(2, 2, \sigma)$ with $\sigma \in [0, r)$, namely, for the general assumptions on (p, q, s) and ρ, η, σ in (36) of Theorem 5. However, to investigate this conjecture, one needs to resort to a very different and much more technically involved and spacious research approach, beginning with the derivation of *direct inequalities* (Jackson-type, etc.) and *inverse inequalities* (Bernstein/Markov-type, etc.) and then, based on the derived inequalities establish a connection between appropriately selected best-approximation functionals and Peetre K-functionals. We refer to [37, Chapter 3] for an early, but sufficiently complete general exposition of this line of argument.

In conclusion, we note that by focusing on gradient/subgradient iterative optimization method in learning algorithms for NNs in the introduction, we left an important methodological gap which needs to be filled here [30]. Numerical methods for optimal control based on Bellman's principle are very powerful in learning theory, both by swarm and deep evolutionary AI, including optimal control using feedback for supervised problems [30]. Although, theoretically, Bellman's principle allows finding *global extrema* for a very general class of criterial functionals (including non-convex, non-smooth (including non-Lipschitzian) ones), and under complicated sets of constraints (including ones induced by technological standards in real-life engineering problems): computing the/a global extremum is often unfeasible due to the huge computational complexity. So, in many cases, a tradeoff is needed between affordable computational complexity and sufficiently high quality of a local extremum attained [31], [32]. So far, similar to gradient methods, optimal tradeoffs in dynamical programming are also achieved via natural, rather than artificial intelligence. Nevertheless, if a dynamical programming algorithm is being applied in the context of machine learning using WBNNs, we may now have an acceptable automatic alternative.

References

[1] I. Daubechies, *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics, USA (1992).

[2] W. Dahmen, Wavelet and multiscale methods for operator equations, *Acta Numerica*, **6** (1997), 55-228.

[3] Q. Zhang, A. Benveniste, Wavelet networks, *IEEE Trans Neural Networks*, **3**, No 6 (1992), 889-898.

[4] Z. Lu, H. Pu, F. Wang, Z. Hu, L. Wang, The expressive power of neural networks: A view from the width, In: *Advances in Neural Information Processing Systems*, **30** (2017).

[5] A.K. Alexandridis, A.D. Zapranis, Wavelet neural networks: A practical guide, *Neural Networks*, **42** (2013), 1-27.

[6] S. Chen, C.F.N. Cowan, P.M. Grant, Orthogonal least squares learning algorithm for radial basis function networks, *IEEE Transactions on Neural Networks*, **2**, No 2 (1991), 302-309.

[7] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature*, **521**, No 7553 (2015), 436-444.

[8] W. Dahmen, Wavelet and multiscale methods for operator equations, *Acta Numerica*, **6** (1997), 55-228.

[9] J. Bergh, J. Löfström, *Interpolation Spaces. An Introduction*, Grundlehren der Mathematischen Wissenschaften, No. 223, Springer-Verlag, Berlin-New York (1976), x+207.

[10] L. Dechevsky, S. Penev, On shape-preserving probabilistic wavelet approximators, *Stochastic Analysis and Applications*, **15**, No 2 (1997), 187-215.

[11] A. Cohen, I. Daubechies, P. Vial, Multiresolution analysis, wavelets and fast algorithms on an interval, *C.R. Acad. Sci. Paris*, Ser. I Math., **316**, No 5 (1993), 417-421.

[12] A. Cohen, I. Daubechies, P. Vial, Wavelets on the interval and fast wavelet transforms, *Appl. Comp. Harmonic Anal.*, **1**, No 1 (1993), 54-81.

[13] G. Cybenko, Approximation by superpositions of a sigmoidal function, *Math. Control Signal Systems*, **2**, No 4 (1989), 303-314.

[14] G. Cybenko, Correction: "Approximation by superpositions of a sigmoidal function", [Math. Control Signal Systems, 2.4 (1989), pp. 303-314], *Math. Control Signal Systems*, **5**, No 4 (1992), 455.

[15] M. Reed, B. Simon, *Methods of Modern Mathematical Physics, Vol. 1: Functional Analysis*, 2nd Ed., Academic Press [Harcourt Brace Jovanovich, Publishers], New York (1980).

[16] L. Dechevsky, S. Penev, On shape-preserving wavelet estimators of cumulative distribution functions and densities, *Stochastic Analysis and Applications*, **16**, No 3 (1998), 423-462.

[17] H. Triebel, *Theory of Function Spaces*, Monographs in Mathematics, Vol. 78, Birkhäuser Verlag, Basel (1983).

[18] S.G. Samko, A.A. Kilbas, O.I. Marichev, *Fractional Integrals and Derivatives: Theory and Applications*, Gordon and Breach Science Publishers, Yverdon (1993).

[19] M. Frazier, B. Jawerth, G. Weiss, *Littlewood-Paley Theory and the Study of Function Spaces*, CBMS Regional Conf. Ser. in Mathematics, Vol. 79, Amer. Math. Soc., Providence, RI (1991).

[20] H. Iba, *Evolutionary Approach to Machine Learning and Deep Neural Networks. Neuro-evolution and Gene Regulatory Networks*, Springer, Singapore (2018).

[21] H. Iba, *Swarm Intelligence and Deep Evolution. Evolutionary Approach to Artificial Intelligence*, Taylor & Francis CRC Press (2022).

[22] F. Hoyle, *The Black Cloud*, William Heinemann Ltd. (1957).

[23] S. Lem, *Niezwyciezony*, Wydawnictwo MON (1964).

[24] S. Lem, *The Invincible*, The MIT Press (2020).

[25] F.T. Leighton, *Introduction to Parallel Algorithms and Architectures. Arrays, Trees, Hypercubes*, Morgan Kaufman Publishers, Inc., San Mateo, CA (1992).

[26] L.T. Dechevsky, J.O. Ramsay, S.I. Penev, Penalized wavelet estimation with Besov regularity constraints, *Math. Balkanica (N.S.)*, **13**, No 3-4 (1999), 257-376.

[27] L.T. Dechevsky, Atomic decomposition of function spaces and fractional integral and differential operators, *Fract. Calc. Appl. Anal.*, **2**, No 4 (1999), 367-381.

[28] R. Okuta, Y. Unno, D. Nishino, S. Hido, C. Loomis, CuPy: A NumPy-Compatible Library for NVIDIA GPU Calculations, In: *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)* (2017), http://learningsys.org/nips17/assets/papers/paper16.pdf.

[29] P. Brenner, V. Thomée, L.B. Wahlbin, *Besov Spaces and Applications to Difference Methods for Initial Value Problems*, Lect. Notes in Math., No. 434, Springer-Verlag, Berlin-New York (1975).

[30] J. Schmidhuber, Deep learning in neural networks: An overview, *Neural Networks*, **61** (2015), 85-117.

[31] L.T. Dechevsky, L.M. Gulliksen, A multirigid dynamical programming algorithm for discrete dynamical systems and its applications to numerical computation of global geodesics, *Int. J. Pure Appl. Math.*, **33**, No 2 (2006), 257–286.

[32] L.T. Dechevsky, L.M. Gulliksen, Application of a multirigid dynamical programming algorithm to optimal parametrization, and a model solution of an industrial problem, *Int. J. Pure Appl. Math.*, **33**, No 3 (2006), 381-406.

[33] L.T. Dechevsky, L.-E. Person, H. Singh, K.M. Tangrand, Learning non-parametric regression-functions and densities by univariate wavelet-based neural networks (2022).

[34] L.T. Dechevsky, L.-E. Person, H. Singh, K.M. Tangrand, Learning of multidimensional geometric manifolds with wavelet-based neural networks (2022).

[35] K. Jetten, S. Smale, D.-X. Zhou, Learning theory and approximation, In: *Mathematisches Forschungsintitut Oberwolfach Workshop, June 29th - July 5th. Oberfolfach Report 30/2008* (2008), 1655-1705.

[36] P. Binev, A. Cohen, W. Dahmen, R. DeVore, Classification algorithms using adaptive partitioning, *Ann. Statist.*, **42**, No 6 (2014), 2141-2163.

[37] P.P. Petrushev, V.A. Popov, *Rational Approximation of Real Functions*, Encyclopedia of Mathematics and its Applications, No. 28, Cambridge University Press, Cambridge (1987).

[38] L.T. Dechevsky, J. Bratlie, B. Bang, A. Lakså, J. Gundersen, Wavelet-based lossless one- and two-dimensional representation of multidimensional geometric data, In: *AIP Conf. Proc.*, **1410**, Amer. Inst. Phys., Melville, NY (2011), 83-97.

[39] L.T. Dechevsky, J. Bratlie, J. Gundersen, Index mapping between tensor-product wavelet bases of different number of variables, and computing multivariate orthogonal discrete wavelet transforms on graphics processing units, *Lecture Notes in Comput. Sci.*, **7116**, Springer, Heidelberg (2012), 402-410.

[40] P. Binev, A. Cohen, W. Dahmen, R. DeVore, Universal algorithms for learning theory. II. Piecewise polynomial functions, *Constr. Approx.*, **26**, No 2 (2007), 127-152.

[41] M. Hairer, C. Labbé, The reconstruction theorem in Besov spaces, *J. of Functional Analysis*, **273**, No 8 (2017), 2578-2618.