Neural Networks 165 (2023) 654–661

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Approximation of classifiers by deep perceptron networks

Věra Kůrková [a,*], Marcello Sanguineti [b,c]

[a] Institute of Computer Science of the Czech Academy of Sciences, Pod Vodárenskou věží 2, 18207 Prague, Czech Republic
[b] DIBRIS University of Genova, Via Opera Pia 13, 16145 Genova, Italy
[c] Institute of Marine Engineering, National Research Council of Italy, 16149 Genova, Italy

ARTICLE INFO

Article history:
Received 31 December 2022
Received in revised form 16 April 2023
Accepted 2 June 2023
Available online 7 June 2023

Keywords:
Approximation by deep networks
Probabilistic bounds on approximation errors
Random classifiers
Concentration of measure
Method of bounded differences
Growth functions

ABSTRACT

We employ properties of high-dimensional geometry to obtain some insights into capabilities of deep perceptron networks to classify large data sets. We derive conditions on network depths, types of activation functions, and numbers of parameters that imply that approximation errors behave almost deterministically. We illustrate general results by concrete cases of popular activation functions: Heaviside, ramp sigmoid, rectified linear, and rectified power. Our probabilistic bounds on approximation errors are derived using concentration of measure type inequalities (method of bounded differences) and concepts from statistical learning theory.

© 2023 Elsevier Ltd. All rights reserved.

1. Introduction

Most research in the theory of neural nets has been concentrated on issues involving continuous and smooth functions on infinite domains. But in practical applications, feedforward networks compute functions on *finite* sets (such as regular grids or scattered vectors in $\mathbb{R}^d$). While many classes of feedforward networks have the *universal representation property* and can exactly compute any function on a finite domain (Ito, 1992), such universality type results have limited applicability as they require the number of network parameters to be as large as the datasets, and these sets are often quite large. Some guidance for choice of network architecture can be obtained by investigating how their approximation capabilities depend on the depth, width, and type of the computational units. For some special types of functions, increase of depth reduces the number of network parameters needed to obtain the same or better accuracy of approximation (Bianchini & Scarselli, 2014; Kůrková, 2018, 2019; Maiorov, 1999; Poggio et al., 2017; Telgarsky, 2016; Yarotsky, 2017). However, the comprehensive theoretical analysis of the impact of network depth is still in its early stages.

Functions on finite domains can be represented as vectors in Euclidean spaces and errors in their approximation as distances between these vectors. Typically, neural networks process large sets of data, so they compute functions on large domains which can be viewed as high-dimensional vectors. Computational difficulties of high-dimensional tasks, called the "curse of dimensionality" (Bellman, 1957), have long been known. On the other hand, the almost deterministic behavior of randomized models depending on large numbers of variables can be attributed to the "blessing of dimensionality" (Donoho & Tanner, 2009; Gonon, Grigoryeva, & Ortega, 2023; Gorban, Makarov, & Tyukin, 2019; Gorban & Tyukin, 2018; Kainen, 1997; Kůrková & Sanguineti, 2016). These phenomena can be explained by rather counterintuitive properties of geometry of high-dimensional spaces. They imply concentration of values of sufficiently smooth functions of many variables around their mean values (Dubhashi & Panconesi, 2009; Gorban, Golubkov, Grechuk, Mirkes, & Tyukin, 2018; Gorban, Tyukin, Prokhorov, & Sofeikov, 2016; Matoušek, 2002; Milman & Schechtman, 1986; Vershynin, 2020).

Numbers of all binary-valued functions even on sets of moderate sizes are too large to be likely that all of them represent some tasks of interest in a given type of application. Thus suitability of classes of networks can be investigated merely for sets of functions that might be relevant for these tasks. Approximation of functions by neural networks have been studied for sets of functions defined by constraints on various norms (see, e.g., Kainen, Kůrková, and Sanguineti (2012) and references therein). In Kůrková and Sanguineti (2019, 2021), we introduced a different approach to restriction of sets of functions to be approximated. We defined relevance of functions for a given application area in terms of a probability distribution on the set of all binary-valued functions.

* Corresponding author.
E-mail addresses: vera@cs.cas.cz (V. Kůrková), marcello.sanguineti@unige.it (M. Sanguineti).

https://doi.org/10.1016/j.neunet.2023.06.004
0893-6080/© 2023 Elsevier Ltd. All rights reserved.

In this paper, we employ this approach to investigate the influence of depths, type of activation functions, and numbers of parameters on the approximation capabilities of deep perceptron with piecewise-polynomial activation functions. We derive probabilistic bounds on approximation errors using concentration-of-measure inequalities applicable to sufficiently smooth functions of the random variables. We show that l_1- and l_2-approximation errors satisfy one of these smoothness conditions (they are coordinate-wise Lipschitz with small coefficients). Applying the McDiarmid Inequality belonging to methods of bounded differences (Dubhashi & Panconesi, 2009; McDiarmid, 1989) we get conditions under which approximation of random classifiers behaves almost deterministically. We prove that deterministic behavior is manifested when networks are "reasonably small" with respect to the size of functions's domain, in particular when the number of input–output functions grows polynomially with its size.

To obtain characteristics of deep perceptron networks that approximate random classifiers in a predictable way, we employ estimates of *growth functions*, which are studied in statistical learning theory in connection with VC-dimension. We derive sufficient conditions for concentration of approximation errors in terms of network depth, effective depth, total number of parameters, and degrees and numbers of pieces of piecewise polynomial activation functions. We illustrate general results in concrete cases of networks with popular activation functions: Heaviside, ramp sigmoid, rectified linear, and rectified power. In each of these special cases, we state conditions on the degree of polynomial growth in terms of network depth, effective depth, and total number of parameters.

The paper is organized as follows. Section 2 contains basic concepts and notations for approximation of functions on finite domains, and it describes the class of deep perceptron networks, whose approximation capabilities we investigate. In Section 3, a probabilistic model of relevance is introduced. Section 4 contains our main results on the approximation of random classifiers by deep perceptron networks. It contains analysis of the influence of network depth, number of parameters, and type of activation functions on the concentration of approximation errors. Proofs are deferred to Section 5. Section 6 is a brief discussion.

2. Preliminaries

2.1. Approximation of functions on finite domains

We investigate approximation of functions on finite subsets $X = \{x_1, \dots, x_m\}$ of $\mathbb{R}^d$ by neural networks. The domain X can model set of vectors of features, which can be scattered in $\mathbb{R}^d$ or form a regular grid.

We denote by

$$\mathcal{F}(X) := \{f \mid f : X \to \mathbb{R}\}$$

the *space of all real-valued functions on* X. $\mathcal{F}(X)$ is isometric to the m-dimensional Euclidean space $\mathbb{R}^m$ and thus a function $f : X \to \mathbb{R}$ can be represented as the m-dimensional vector $(f(x_1), \dots, f(x_m))$. On $\mathcal{F}(X)$, we consider l_2 and l_1-norms defined as

$$\|f\|_2 := \sqrt{\sum_{i=1}^{m} f(x_i)^2}$$

$$\|f\|_1 := \sum_{i=1}^{m} |f(x_i)|,$$

resp. We measure errors in approximation of functions by neural networks as their l_1- and l_2-distances from sets of input–output functions. For $f \in \mathcal{F}(X)$, and $\mathcal{H} \subset \mathcal{S}(X)$, $p = 1, 2$, we denote

$$\|f - \mathcal{H}\|_p = \inf_{h \in \mathcal{H}} \|f - h\|_p.$$

We denote

$$\mathcal{S}(X) := \{f \mid f : X \to \{-1, 1\}\}$$

the set of all functions on X with values in $\{-1, 1\}$ ($\mathcal{S}(X)$ is equivalent to the *Hamming cube*). This set corresponds to all binary classifiers on the domain X. From technical reasons, we consider range $\{-1, 1\}$ instead of $\{0, 1\}$ (so that all binary classifiers have l_2-norms equal $\sqrt{m}$).

For any set of real-valued functions $\mathcal{L}$ we denote by $\mathrm{sgn} \circ \mathcal{L}$ the set of binary-valued functions obtained by composing functions from $\mathcal{L}$ with the *signum function* $\mathrm{sgn}(t) = +1$ for $t \geq 0$ and $\mathrm{sgn}(t) = -1$ for $t < 0$, i.e.,

$$\mathrm{sgn} \circ \mathcal{L} := \{\mathrm{sgn} \circ g \mid g \in \mathcal{L}\}.$$

The following simple proposition shows that lower bounds on errors in approximation of binary classifiers by a set $\mathcal{L}$ of real-valued functions on X (which might be infinite) can be obtained from lower bounds on approximation by the finite set $\mathrm{sgn} \circ \mathcal{L}$, which has cardinality at most $2^{\mathrm{card} X}$.

Proposition 2.1. *Let $X \subset \mathbb{R}^d$ be finite, $\mathcal{L} \subset \mathcal{F}(X)$, and $f \in \mathcal{S}(X) \setminus \mathcal{L}$. Then for $p \in \{1, 2\}$*

$$\|f - \mathcal{L}\|_p \geq \frac{1}{2} \|f - sgn \circ \mathcal{L}\|_p.$$

Proof. It is easy to verify (see Kůrková and Sanguineti (2017, Proposition 3.1)) that for every $h \in \mathcal{F}(X)$, $\|f - h\|_p \geq \frac{1}{2} \|f - \mathrm{sgn}(h)\|_p$. Thus $\inf_{h \in \mathcal{L}} \|f - h\|_p \geq \inf_{h \in \mathrm{sgn} \circ \mathcal{L}} \frac{1}{2} \|f - \mathrm{sgn}(h)\|_p$. □

Proposition 2.1 shows that lower bounds on approximation of binary classifiers by sets $\mathcal{L}$ of real-valued functions (in particular, by networks with linear outputs) can be obtained from lower bounds derived for $\mathrm{sgn} \circ \mathcal{H}$ (in particular, for networks where linear output units are replaced by signum perceptrons that we consider in the next subsection).

2.2. Deep perceptron networks

Multilayer feedforward networks are determined by directed acyclic graphs $\mathcal{G}$, where nodes represent computational units and edges connections between them. The units are arranged in L layers, network inputs are viewed as the layer 0, and for all $l = 1, \dots, L$, units in the lth layer have inputs only from preceding layers $i = 0, \dots, l-1$. The layers $l = 1, \dots, L-1$ are called *hidden layers*. We assume that the last layer L contains a unique output unit.

A biologically inspired computational unit called *perceptron* applies a fixed *activation function* $\psi : \mathbb{R} \to \mathbb{R}$ to affine functions with varying parameters. A perceptron computes functions of the form

$$\psi(v \cdot . + b) : \mathbb{R}^d \to \mathbb{R},$$

where $v \in \mathbb{R}^d$ is called a *weight vector*, $b \in \mathbb{R}$ a *bias*, and $v \cdot x = \sum_{i=1}^{d} v_i x_i$ is the *scalar product* of the weight vector with input vector x (weighted sum of inputs). Originally, perceptrons were endowed with sigmoidal activation functions representing hard or soft threshold, such as

Heaviside $\theta(t) = 0$ for $t \leq 0$ and $\theta(t) = 1$ for $t \geq 1$,

ramp sigmoid $\tau(t) = -0$ for $t \leq 0$, $\tau(t) = t$ for $t \in [0, 1]$, $\tau(t) = 1$ for $t \geq 1$,

logistic sigmoid $\sigma(t) = \frac{1}{1+e^{-x}}$.

Currently *rectified linear* (ReLU) activation function $\rho(t) = \max(0, t)$ is popular, sometimes also *rectified powers* (RePU, ReLU^k) $\rho_k(t) = \max(0, t^k)$ are used.

We investigate approximation of binary classifiers by *multilayer perceptron networks with piecewise polynomial activation functions and a single output with the signum activation*. So we assume that the unique output unit in the last Lth layer is a perceptron with the signum activation function (thus network outputs are in $\{-1, 1\}$), and all units in layers $l = 1, \ldots, L-1$ are perceptrons with piecewise polynomial activation functions $\psi_{l,j}, j = 1, \ldots, k_l, l = 1, \ldots, L-1$. We denote by

$$\mathcal{M} := \mathcal{M}(\mathcal{G}, L, \{k_l, l = 1, \ldots, L-1\}, \{\psi_{l,j}, \\ l = 1, \ldots, L-1, j = 1, \ldots, k_l\}) \quad (1)$$

the parameterized family of $\{-1, 1\}$-valued *input–output functions of the class of networks described above with a fixed graph $\mathcal{G}$, fixed activation functions $\psi_{i,j}, l = 1, \ldots, L-1, j = 1, \ldots, k_l$, and varying network parameters* (weights and biases).

The number L of layers in the above described class of networks is called the *network depth*. A more refined concept, introduced by Bartlett, Harvey, Liaw, and Mehrabian (2019), is its effective depth, which instead of merely counting the number of layers, also takes into account an arrangement of units with respect to the hierarchical structure of the network. The *effective depth*, denoted $\bar{L}$, was defined in Bartlett et al. (2019) for multilayer perceptron networks where all activation functions are piecewise polynomials as

$$\bar{L} := \frac{1}{W} \sum_{l=1}^{L} W_l, \quad (2)$$

where W is the *total number of network parameters*, while the definition of W_l, $l = 1, \ldots, L$ depends on the maximal degree of piecewise polynomial activations $\delta > 0$ as follows:

- for $\delta > 0$, W_l is the number of parameters (weights and biases) at inputs of all layers up to the lth layer (i.e., at the layers $i = 1, \ldots, l$);
- for $\delta = 0$, W_l is the number of parameters (weights and biases) at the inputs of units in the layer l. Note that it follows from the definition of $\bar{L}$ that when all activation functions are piecewise constant (such as Heaviside and signum), then the effective depth is 1.

The concept of the effective depth reflects that units closer to the input layer have a greater influence on the size of sets of network input–output functions than units closer to the last layer. This can be illustrated by the following simple example of three arrangements of n units, all with r parameters, in networks with a single output and depths $L = 3$ and $L = 2$:

(1) the first hidden layer has $n-3$ units, the second hidden layer 2 units,
(2) the first hidden layer has 2 units, the second hidden layer $n-3$ units,
(3) the network has only one hidden layer with $n-1$ units.

It follows from the definition of the effective depth that in the case (1), $\bar{L}_1 = 3 - \frac{4}{n}$, while in (2), $\bar{L}_2 = 2 + \frac{1}{n}$, and in (3), $\bar{L}_3 = 2 - \frac{1}{n}$.

3. Probabilistic bounds on approximation errors

3.1. Probabilistic model of relevance

We assume that there is a probability measure P on the finite set $\mathcal{S}(X)$ of all binary classifiers on X which models their relevance for a given application area. A function $f \in \mathcal{S}(X)$ randomly chosen according to P induces *random variables*

$$Y_1 := f(x_1), \ldots, Y_m := f(x_m)$$

with values in $\{-1, 1\}$.

Many studies of random phenomena assume (often implicitly) that the probability is uniform. Here, we focus on cases when distributions of values of random variables $Y_1, \ldots, Y_m$ can differ, but the variables remain *independent*. Independence of random variables is essential for applications of most theorems on concentration of values of functions of large numbers of random variables. These theorems state that under various smoothness conditions, values of these functions tend to concentrate around their mean values. The assumption of independence of random variables $Y_1, \ldots, Y_m$ implies that P is a *product probability*, i.e., there exist distributions $\mathrm{P}_1, \ldots, \mathrm{P}_m$ such that

$$\mathrm{P}(Y_1, \ldots, Y_m) = \Pi_{i=1}^{m} P_i(Y_i).$$

We investigate l_1- and l_2-distances of random functions chosen according to P from a fixed function $h \in \mathcal{S}(X)$ as functions of random variables defined as

$$\Psi_{h,1}(Y_1, \ldots, Y_m) := \sum_{i=1}^{m} |Y_i - h(x_i)|$$

and

$$\Psi_{h,2}(Y_1, \ldots, Y_m) := \sqrt{\sum_{i=1}^{m} (Y_i - h(x_i))^2}.$$

3.2. Concentration of approximation errors

To derive probabilistic bounds for approximation of random classifiers by deep perceptron networks, we first formulate bounds for general approximating sets in terms of their sizes.

For a set $\mathcal{H} \subset \mathcal{S}(X)$ of binary valued-functions and $\iota = 1, 2$, we denote by

$$\mu_{\mathcal{H},\iota} := \min\{E\|f - h\|_\iota \mid h \in \mathcal{H}\} \quad (3)$$

the *minimum of mean values of l_ι-distances from $\mathcal{H}$*. The following theorem gives probabilistic bounds on deviations from $\mu_{\mathcal{H},\iota}$ of approximation errors of randomly-chosen classifiers. The lower bound depends on both $\mathrm{card}\mathcal{H}$ and $\mathrm{card}X = m$, while the upper bound depends merely on m.

Theorem 3.1. *Let $X \subset \mathbb{R}^d$ be finite with $\mathrm{card}\, X = m$, P be a product probability measure on $\mathcal{S}(X)$, $\mathcal{H} \subset \mathcal{S}(X)$, and $\lambda > 0$. Then for $f \in \mathcal{S}(X)$ randomly chosen according to* P,

(i) $\mathrm{P}\left[\|f - \mathcal{H}\|_\iota \le \mu_{\mathcal{H},\iota} + \lambda\right] > 1 - e^{-\frac{m\lambda^2}{2}}$;

(ii) $\mathrm{P}\left[\mu_{\mathcal{H},\iota} - \lambda \le \|f - \mathcal{H}\|_\iota\right] > 1 - \mathrm{card}\mathcal{H}\, e^{-\frac{m\lambda^2}{2}}$.

The proof of Theorem 3.1, presented in Section 5.1, is based on a concentration of measure type inequality. It employs the McDiarmid Bound (McDiarmid, 1989), which is one of probabilistic bounds holding for functions satisfying smoothes assumptions called *bounded differences conditions* (Dubhashi & Panconesi, 2009).

Note that the probability of the lower bound from Theorem 3.1(ii) is high when the size $\mathrm{card}\mathcal{H}$ of the approximating set $\mathcal{H}$ does not grow with m fast enough to outweigh the decrease of $e^{-\frac{m\lambda^2}{2}}$. Setting $\mathrm{card}\,\mathcal{H} = \kappa(m)$, we can assess the lower bound on the probability by estimating $\kappa(m)\, e^{-\frac{m\lambda^2}{2}}$. When it converges to zero, for large m most approximation errors are in the interval

$$[\mu_{\mathcal{H},\iota} - \lambda, \mu_{\mathcal{H},\iota} + \lambda]. \quad (4)$$

When λ is small enough to guarantee proximity to $\mu_{\mathcal{H},\iota}$ and $\frac{\lambda^2}{2}$ does not outweigh m, then Theorem 3.1 implies concentration of approximation errors. A suitable choice is, for example, $\lambda = m^{-1/4}\,2^{1/2}$. For large m, it is small and $\frac{m\lambda^2}{2} = m^{1/2}$. Thus we obtain the following corollary of Theorem 3.1.

Corollary 3.2. *Let $X \subset \mathbb{R}^d$ be finite with $\operatorname{card} X = m$, P be a product probability measure on $\mathcal{S}(X)$, and $\mathcal{H} \subset \mathcal{S}(X)$ be such that $\operatorname{card}\mathcal{H} = \kappa(m)$. Then for $f \in \mathcal{S}(X)$ randomly chosen according to* P,

(i) $\mathrm{P}\Big[\|f - \mathcal{H}\|_\iota \le \mu_{\mathcal{H},\iota} + m^{-1/4}\Big] > 1 - e^{-m^{1/2}}$;

(ii) $\mathrm{P}\Big[\mu_{\mathcal{H},\iota} - m^{1/4} \le \|f - \mathcal{H}\|_\iota\Big] > 1 - \kappa(m)\,e^{-m^{1/2}}$.

Corollary 3.2 implies concentration of approximation errors for approximating sets $\mathcal{H}$ with $\operatorname{card}\mathcal{H} = \kappa(m)$ being a polynomial. The value of $\mu_{\mathcal{H},\iota}$ is critical for assessment of suitability of the class $\mathcal{H}$ for classification tasks characterized by the probability of relevance P. When $\mu_{\mathcal{H},\iota}$ is large, then Corollary 3.2(ii) shows that the set $\mathcal{H}$ is not suitable. On the other hand if $\mu_{\mathcal{H},\iota}$ is small, then the upper bound from Theorem 3.1(i) implies that almost all randomly chosen classifiers can be well approximated by the set $\mathcal{H}$.

Note that for many choices of an approximating set $\mathcal{H}$, the lower bounds on the probability from Theorem 3.1(ii) and Corollary 3.2(ii) are not likely to be tight, the probability of concentration of approximation errors can be larger. The bounds are proven (see Section 5) assuming that λ-neighborhoods of elements of $\mathcal{H}$ in l_1 or l_2-norm might be disjoint, which often is not the case.

4. Approximation by deep perceptron networks

4.1. Probabilistic bounds

The lower bounds on approximation errors from Theorem 3.1(ii) and Corollary 3.2(ii) are expressed in terms of the size $\operatorname{card}\mathcal{H}$ of an approximating set of functions on a domain X of the size $\operatorname{card} X = m$. We apply these general bounds to neural networks by combining them with estimates of sizes of their sets of input–output functions. We employ *growth functions* studied in statistical learning theory in connection with VC-dimension (Vapnik & Chervonenkis, 1971).

Our main results provide probabilistic bounds on errors in approximation of randomly chosen binary classifiers by deep perceptron networks with piecewise polynomial activations from the class described in Section 2.2.

Theorem 4.1. *Let $X \subset \mathbb{R}^d$ be finite with $\operatorname{card} X = m$, P a product probability measure on $\mathcal{S}(X)$, p, δ be positive integers, $\mathcal{M} := \mathcal{M}(\mathcal{G}, \{k_l, l = 1, \dots, L-1\}, \{\psi_{l,j}, l = 1, \dots, L-1, j = 1, \dots, k_l\})$ the set of all input–output functions of a class of deep perceptron networks with depth L, effective depth $\bar{L}$, total number of parameters W, single output with signum activation and activation functions $\psi_{l,j}$, $j = 1, \dots, k_l$ in layers $l = 1, \dots, L-1$ being piecewise polynomials with $p+1$ pieces and degrees at most δ. Then for $f \in \mathcal{S}(X)$ randomly chosen according to* P, $\iota = 1, 2$, *and* $\lambda > 0$

(i) $\mathrm{P}\Big[\|f - \mathcal{M}\|_\iota \le \mu_{\mathcal{M},\iota} + \lambda\Big] > 1 - e^{-\frac{m\lambda^2}{2}}$,

(ii) when $\bar{L}W \le m$, $\mathrm{P}\Big[\mu_{\mathcal{M},\iota} - \lambda \le \|f - \mathcal{M}\|_\iota\Big] > 1 - (\alpha(p,\delta,L)\,em)^{\bar{L}W}\, e^{-\frac{m\lambda^2}{2}}$, *where* $\alpha(p,\delta,L) := 4p\left(1 + (L-1)\delta^{L-1}\right)$.

The proof of Theorem 4.1, presented in Section 5, combines probabilistic bounds from Theorem 3.1 with an estimate of the growth functions of sets of input–output functions of deep perceptron networks with piecewise linear activation functions $\{\psi_{l,j}\}$.

The bounds from Theorem 4.1 are interesting when they hold with a sufficiently large probability. This happens when the term $e^{-\frac{m\lambda^2}{2}}$ converges to zero faster than $\operatorname{card}\mathcal{M}$ grows with m to infinity. For example, $\lambda = 2^{1/2}\,m^{-1/4}$ is for large m small enough to guarantee proximity to $\mu_{\mathcal{M},\iota}$ and the lower bound $1 - e^{-\frac{m\lambda^2}{2}} = 1 - e^{-m^{1/2}}$ on probability that approximation errors are close to $\mu_{\mathcal{M},\iota}$ converges to 1.

Corollary 4.2. *Let $X \subset \mathbb{R}^d$ be finite with $\operatorname{card} X = m$, P a product probability measure on $\mathcal{S}(X)$, p, δ be positive integers, $\mathcal{M} := \mathcal{M}(\mathcal{G}, \{k_l, l = 1, \dots, L-1\}, \{\psi_{l,j}, l = 1, \dots, L-1, j = 1, \dots, k_l\})$ the set of input–output functions of a class of deep perceptron networks with depth L, effective depth $\bar{L}$, total number of parameters W, single output with signum activation and all activation functions $\psi_{l,j}$, $j = 1, \dots, k_l$ in layers $l = 1, \dots, L-1$ being piecewise polynomials with $p+1$ pieces and degrees at most δ. Then for $f \in \mathcal{S}(X)$ randomly chosen according to* P *and* $\iota = 1, 2$,

(i) $\mathrm{P}\Big[\|f - \mathcal{M}\|_\iota \le \mu_{\mathcal{M},\iota} + 2^{1/2} m^{-1/4}\Big] > 1 - e^{-m^{1/2}}$,

(ii) when $\bar{L}W \le m$, $\mathrm{P}\Big[\mu_{\mathcal{M},\iota} - 2^{1/2} m^{-1/4} \le \|f - \mathcal{M}\|_\iota\Big] > 1 - (\alpha(p,\delta,L)\,em)^{\bar{L}W}\, e^{-m^{1/2}}$ *where* $\alpha(p,\delta,L) := 4p\left(1 + (L-1)\delta^{L-1}\right)$.

Theorem 4.1 and Corollary 4.2 imply conditions on depth, effective depth, and total numbers of parameters of deep perceptron networks with piecewise polynomial activations that guarantee concentration of errors in approximation of randomly chosen classifiers.

Note that the lower bounds from Theorem 4.1(ii) and Corollary 4.2(ii) hold for networks where the product of network effective depth $\bar{L}$ and total number of its parameters W does not exceed the size m of the set of data to be classified. In practical tasks dealing with large data, it is desirable to use networks with reasonably small effective depths and with numbers of parameters much smaller than sizes of sets of data to be classified. Thus the assumption $\bar{L}W \le m$ is not practically restrictive, often even $\bar{L}W \le LW \le m$.

Stronger conditions on network depth L, effective depth $\bar{L}$, and total number of its parameters W should be imposed to obtain bounds holding with high probability. The upper bounds from Theorem 4.1(i) and Corollary 4.2(i) do not depend on the size $\operatorname{card}\mathcal{M}$ of the set of input–output functions, but $\operatorname{card}\mathcal{M}$ plays a crucial role in the lower bounds (ii). When

$$\left(4p(1 + (L-1)\delta^{L-1})\,em\right)^{\bar{L}W} = \left(\alpha(p,\delta,L)\,em\right)^{\bar{L}W}$$

does not outweigh the exponential decay of $e^{-\frac{m\lambda^2}{2}}$ (in particular $e^{-m^{1/2}}$ for $\lambda = 2^{1/2} m^{-1/4}$), then the probability converges to 1 with m increasing.

Assuming that the degree δ and the number p of pieces of activation functions are fixed, Theorem 4.1 and Corollary 4.2 imply conditions on L, $\bar{L}$, and W that guarantee that for a sufficiently large m, approximation of classifiers behaves almost deterministically. More precisely, with a high probability almost all randomly chosen classifiers are close to $\mu_{\mathcal{M},\iota}$ in l_ι-norm for $\iota = 1, 2$. In such cases, suitability of a class of networks computing input–output functions from $\mathcal{M}$ can be assessed according to the value of $\mu_{\mathcal{M},\iota}$. When $\mu_{\mathcal{M},\iota}$ is large, $\mathcal{M}$ is not suitable for a task characterized by the probability P.

4.2. Consequences for some types of activation functions

We analyze conditions that guarantee almost deterministic behavior of approximation of random classifiers by some classes of deep perceptron networks for some choices of popular activation functions.

The next corollary follows directly from Theorem 4.1 by applying it to piecewise polynomial activation functions of degrees $\delta = 0, 1, 2$, numbers of their pieces $p + 1$ where $p = 1, 2$, and from the definition of $\bar{L}$ (in particular, $\bar{L} = 1$ for $\delta = 0$).

Corollary 4.3. *Let $X \subset \mathbb{R}^d$ be finite with* card $X = m$, P *a product probability measure on $\mathcal{S}(X)$, $\mathcal{M} := \mathcal{M}(\mathcal{G}, \{k_l, l = 1, \dots, L-1\}, \{\psi_{l,j}, l = 1, \dots, L-1, j = 1, \dots, k_l\})$, the set of all input–output functions of a class of deep perceptron networks with depth L, effective depth $\bar{L}$, total number of parameters W such that $m \geq \bar{L}W$, and a single output with the signum activation. Then for $f \in \mathcal{S}(X)$ randomly chosen according to* P, $\iota = 1, 2$, *and $\lambda > 0$*

(i) when all $\psi_{l,j}$, $l = 1, \dots, L-1$, $j = 1, \dots, k_l$ are either Heaviside or signum functions,

$$\mathrm{P}\Big[\, \mu_{\mathcal{M},\iota} - \lambda \leq \|f - \mathcal{M}\|_\iota \Big] > 1 - (4em)^W e^{-\frac{m\lambda^2}{2}}; \tag{5}$$

(ii) when all $\psi_{l,j}$, $l = 1, \dots, L-1$, $j = 1, \dots, k_l$, are ReLU *functions $\rho_k(t) = \max(0, t^k)$*

$$\mathrm{P}\Big[\, \mu_{\mathcal{M},\iota} - \lambda \leq \|f - \mathcal{M}\|_\iota \Big] > 1 - (4em)^{\bar{L}W} e^{-\frac{m\lambda^2}{2}}. \tag{6}$$

(iii) when all $\psi_{l,j}$, $l = 1, \dots, L-1$, $j = 1, \dots, k_l$, are ramp sigmoid functions

$$\mathrm{P}\Big[\, \mu_{\iota,\mathcal{M}} - \lambda \leq \|f - \mathcal{M}\|_\iota \Big] > 1 - (8em)^{\bar{L}W} e^{-\frac{m\lambda^2}{2}}; \tag{7}$$

(iv) when all $\psi_{l,j}$, $l = 1, \dots, L-1$, $j = 1, \dots, k_l$, are rectified powers $(\max(0, \cdot))^k$ *for a fixed k*

$$\mathrm{P}\Big[\, \mu_{\iota,\mathcal{M}} - \lambda \leq \|f - \mathcal{M}\|_\iota \Big] > 1 - \Big(4em(1 + (L-1)k^{L-1})\Big)^{\bar{L}W} e^{-\frac{m\lambda^2}{2}}. \tag{8}$$

Corollary 4.3 gives estimates of probabilities of lower bounds on approximation errors in terms of characteristics of classes of deep perceptron networks. Note that the probability of upper bounds on approximation errors depends merely on m, it does not depend on the sizes of sets of input–output functions. For all classes from Corollary 4.3, the probability of upper bounds satisfies

$$\mathrm{P}\Big[\|f - \mathcal{M}\|_\iota \leq \mu_{\iota,\mathcal{M}} + \lambda\Big] > 1 - e^{-\frac{m\lambda^2}{2}}. \tag{9}$$

Analysis of bounds from Corollary 4.3 and the bound (9) provides conditions that imply concentration of errors in approximation of random classifiers by deep perceptron networks with popular activation functions (see Table 1).

(i) In the case of networks with Heavisides or signum activations, $\delta = 0$ and $p = 1$ and thus $\alpha(p, \delta, L) = 4$. As $\bar{L} = 1$, the bound (5) on probability that approximation errors are at least $\mu_{\mathcal{M},\iota} - \lambda$ does not depend on depth L nor on the effective depth $\bar{L}$, it merely depends on the total number W of network parameters. For $\lambda = 2^{1/2} m^{-1/4}$, the lower bound on probability becomes

$$1 - e^{W(\ln(4m)+1) - m^{1/2}}. \tag{10}$$

Thus the lower bound is interesting only when the total number W of parameters satisfies

$$W < \frac{m^{1/2}}{\ln(4m) + 1}. \tag{11}$$

The smaller W than $\frac{m^{1/2}}{\ln(4m)+1}$, the larger probability that errors in approximation of random classifiers are within $2^{1/2} m^{-1/4}$ from $\mu_{\mathcal{M},\iota}$.

(ii) For networks with ReLU activation functions, the condition implying concentration of approximation errors around $\mu_{\mathcal{M},\iota}$ involves the product $\bar{L}W$ of the network effective depth and its total number of parameters. For $\lambda = 2^{1/2} m^{-1/4}$, the lower bound on probability (6) becomes

$$1 - e^{\bar{L}W \ln(4m) + \bar{L}W - m^{1/2}} \tag{12}$$

Table 1
Conditions for bounds on probability from Corollary 4.3.

Activation function	Assumption of Theorem 4.1	Exponent negative for $\lambda = 2^{1/2} m^{-1/4}$
Heaviside and signum	$W \leq m$	$W < \frac{m^{1/2}}{\ln(4m)+1}$
ReLU	$\bar{L}W \leq m$	$\bar{L}W < \frac{m^{1/2}}{\ln(4m)+1}$
ramp sigmoid	$\bar{L}W \leq m$	$\bar{L}W < \frac{m^{1/2}}{\ln(8m)+1}$
RePU = ReLU^k	$\bar{L}W \leq m$	$\bar{L}W < \frac{m^{1/2}}{4m \ln(1+(L-1)k^{L-1})+1}$

Thus the lower bound is greater than zero when

$$\bar{L}W < \frac{m^{1/2}}{\ln(4m) + 1}. \tag{13}$$

The slower $\bar{L}W$ grows than $\frac{m^{1/2}}{\ln(4m)+1}$, the higher probability that approximation errors are concentrated around $\mu_{\mathcal{M},\iota}$. Increasing network effective depth by rearranging its units but keeping the number of total parameters W fixed leads to decreasing probability of almost deterministic behavior of approximation errors.

(iii) The case of networks with ramp sigmoid is similar to the case of networks with ReLU units (it is not surprising, since the ramp sigmoid can be obtained as a linear combination of two ReLU functions). For $\lambda = 2^{1/2} m^{-1/4}$, the right-hand side of (7) becomes

$$1 - e^{\bar{L}W \ln(8m) + \bar{L}W - m^{1/2}}. \tag{14}$$

So the bound is interesting when

$$\bar{L}W < \frac{m^{1/2}}{\ln(8m) + 1}. \tag{15}$$

So the condition on $\bar{L}W$ that influences probability of the concentration of approximation errors is only slightly different from the condition for the case of ReLU units.

The case of networks with rectified power units (iv) is less transparent. For $\lambda = 2^{1/2} m^{-1/4}$ the lower bound on probability becomes

$$1 - e^{\bar{L}W\left(\ln(4m(1+(L-1)k^{L-1}))+1\right) - m^{1/2}}. \tag{16}$$

So the bound is interesting when

$$\bar{L}W < \frac{m^{1/2}}{\ln\left(4m(1 + (L-1)k^{L-1}) + 1\right)}. \tag{17}$$

4.3. Consequences for choice of network architectures

Corollary 4.3 implies conditions on the number of parameters, depth, and effective depth of classes of deep networks with some popular activation functions under which approximation of random classifiers by these perceptron networks behaves almost deterministically. Under these conditions l_1- and l_2-approximation errors are concentrated around minima of mean values of these errors over sets of all network input–output functions. Roughly speaking, the conditions imply that this happens when networks are relatively "small" with respect to the size card $X = m$ of the data to be classified, where "small" concerns total number of parameters, network depth, and effective depth.

For networks with piecewise constant activations (such as Heaviside or signum), the conditions assume that the total number of network parameters W is sufficiently smaller than $\frac{m^{1/2}}{\ln(4m)+1}$. For networks with ReLU activation functions, the same constraint $\frac{m^{1/2}}{\ln(4m)+1}$ applies to the product $\bar{L}W$ instead of W. Thus keeping

the total number of parameters W fixed but rearranging network units in more layers so that its effective depth $\bar{L}$ is increased, the probability of concentration of approximation errors might decrease.

Suitability of a class of deep networks which is "small" enough to satisfy the conditions stated in Corollary 4.3 depends on the values $\mu_{\mathcal{M},\iota}$. If they are large, networks computing functions from $\mathcal{M}$ are not suitable for the task characterized by P because almost all randomly chosen classifiers according to P have large errors. Adding more parameters or adding layers to increase network effective depth so that the conditions on concentration are not satisfied might increase chances of better approximation. With increase of the size of a set of input–output functions, the minimum $\mu_{\mathcal{M},\iota}$ can decrease.

When $\mu_{\mathcal{M},\iota}$ is small, then the class of networks is suitable for tasks characterized by P. Note that the upper bound $\mu_{\mathcal{M},\iota} + \lambda$ holds without any restrictions on W and $\bar{L}$.

For every $h \in \mathcal{S}(X)$ and any product probability P, $E\|f - h\|_1 \leq m$ and $E\|f - h\|_2 \leq \sqrt{2m}$. Among all probabilities, the uniform one gives the largest mean values. Due to symmetry, all mean values are equal, and thus for uniform probability $\mu_{\mathcal{M},1} = m$ and $\mu_{\mathcal{M},2} = \sqrt{2m}$. Without a prior knowledge, we have to assume that the probability P is uniform. Our results show that in such cases, almost all uniformly randomly chosen classifiers of sufficiently large data cannot be well approximated by "small" deep perceptron networks satisfying conditions from Corollary 4.3. Thus universal approximation property requires "large" networks.

5. Methods and proofs

5.1. Method of bounded differences and proof of Theorem 3.1

To prove Theorem 3.1, we employ probabilistic inequalities characteristic for the phenomenon of concentration of measure. Under suitable smoothness conditions, functions of large numbers of random variables exhibit almost deterministic behavior in the sense that their values concentrate more or less tightly around their mean values. One of such smoothness conditions is a version of the Lipschitz property. We call a function

$$\Lambda : A_1 \times \cdots \times A_m \to \mathbb{R}$$

coordinate-wise Lipschitz (CWL) with parameters $c_1, \ldots, c_m$ if for all $i = 1, \ldots, m$ and all vectors $a = (a_1, \ldots, a_m)$, $a' = (a'_1, \ldots, a'_m) \in A_1 \times \cdots \times A_m$, which differ just in the ith coordinate,

$$|\Lambda(a) - \Lambda(a')| \leq c_i. \tag{18}$$

Note that the CWL condition implies Lipschitz continuity on $\{-1, 1\}^m$ with the parameter $\bar{c} = \max_{i=1,\ldots,m} c_i$ with respect to the Hamming distance dist_H measured by the number of entries of two vectors in $\{-1, 1\}^m$ in which they differ. Indeed, if Λ satisfies the CWL condition (18) with parameters $c_1, \ldots, c_m$, then for every $u, v \in \{-1, 1\}^m$ which differ in just k entries, we have a sequence $u = u_0, \ldots, u_k = v$ such that u_j, $j = 0, \ldots, k-1$ differs from u_{j+1} in just one coordinate i_{j+1}. Then $|\Lambda(u) - \Lambda(v)| \leq \sum_{j=1}^{k} c_{i_j} \leq \bar{c}\, k = \bar{c}\, \mathrm{dist}_H(u, v)$.

We use the following probabilistic bound that implies concentration of values of functions of independent random variables satisfying the CWL condition with a sufficiently small l_2-norms $\|c\|_2$ of their vectors of parameters $c = (c_1, \ldots, c_m)$. This bound is known as *McDiarmid Inequality* (McDiarmid, 1989) and it belongs to the class of *methods of bounded differences* (Dubhashi & Panconesi, 2009).

Theorem 5.1. *[Dubhashi and Panconesi (2009, p. 70)] Let $Y_1, \ldots, Y_m$ be independent random variables with values in ranges $A_1, \ldots, A_m$, resp., and $\Phi : A_1 \times \cdots \times A_m \to \mathbb{R}$ be a function satisfying the CWL condition with the vector of parameters $c := (c_1, \ldots, c_m)$. Then for every $t > 0$,*

$$\mathrm{P}\Big[\, |\Phi - E(\Phi)| > t \,\Big] \leq e^{-2t^2/\gamma}, \tag{19}$$

where $\gamma := \sum_{i=1}^{m} c_i^2 = \|c\|^2$.

Proof of Theorem 3.1. To apply Theorem 5.1 to the functions $\Psi_{h,1}$ and $\Psi_{h,2}$, we verify that they satisfy the CWL condition and estimate its parameters. We have to estimate differences between values of $\Psi_{h,\iota}$, $\iota = 1, 2$ with $Y_i = 1$ and with $Y_i = -1$, $i = 1, \ldots, m$.

In the case of l_1, we have for every $i = 1, \ldots, m$,

$$|\Psi_{h,1}(Y_1, \ldots, Y_{i-1}, 1, Y_{i+1}, \ldots, Y_m) - \Psi_{h,1}(Y_1, \ldots, Y_{i-1}, -1, Y_{i+1}, \ldots, Y_m)| = 2.$$

Thus $\Psi_{h,1}$ satisfies the CWL condition with all parameters $c_i = 2$.

In the case of l_2, we have

$$|\Psi_{h,2}(Y_1, \ldots, Y_{i-1}, 1, Y_{i+1}, \ldots, Y_m) - \Psi_{h,2}(Y_1, \ldots, Y_{i-1}, -1, Y_{i+1}, \ldots, Y_m)|$$

$= \sqrt{b+4} - \sqrt{b} < 2$, where $b = \sum_{j \neq i}^{m} (Y_j - h(x_j))^2$.

Hence in both cases $\iota = 1, 2$, we have $\gamma = \sum_{i=1}^{m} c_i^2 = 4m$. Setting $t := m\lambda$ we get $2t^2/c \geq (2m^2\lambda^2)/(4\,m) = (m\lambda^2)/2$. Thus Theorem 5.1 implies for all $h \in \mathcal{H}$ and $\iota = 1, 2$,

$$\mathrm{P}\Big[\, \big|\Phi(h, \iota) - E(\Phi_{h,\iota})\big| > \lambda \,\Big] \leq e^{-\frac{m\lambda^2}{2}}. \tag{20}$$

For every $h \in \mathcal{H}$ and $\iota = 1, 2$, set

$$\mu_{h,\iota} := E\|f - h\|_\iota \quad \text{and} \quad \mu_{\mathcal{H},\iota} := \min_{h \in \mathcal{H}} \mu_{h,\iota}.$$

To prove (i), we choose some $h^*_\iota \in \mathcal{H}$, for which $\mu_{\mathcal{H},\iota} = \mu_{h^*_\iota,\iota}$. By (20)

$$\mathrm{P}\Big[\, \|f - h^*_\iota\|_\iota \leq \mu_{h^*_\iota,\iota} + \lambda \,\Big] > 1 - e^{-\frac{m\lambda^2}{2}}.$$

As $\|f - \mathcal{H}\|_\iota \leq \|f - h^*_\iota\|_\iota \leq \mu_{h^*,\iota} + \lambda = \mu_{\mathcal{H}} + \lambda$, the upper bound follows.

To prove (ii), for every $f \in \mathcal{S}(X)$ denote $h_{f,\iota} \in \mathcal{H}$ such that $\|f - h_{f,\iota}\|_\iota = \|f - \mathcal{H}\|_\iota$ (it exists as $\mathcal{H}$ is finite). Since

$$\mu_{\mathcal{H},\iota} - \lambda \leq \mu_{h_{f,\iota},\iota} - \lambda \leq \|f - \mathcal{H}\|_\iota \leq \mu_{h_{f,\iota},\iota} + \lambda$$

by (20) we get

$$\mathrm{P}\Big[\, \mu_{\mathcal{H},\iota} - \lambda \leq \|f - \mathcal{H}\|_\iota \,\Big] > 1 - \mathrm{card}\mathcal{H}\; e^{-\frac{m\lambda^2}{2}}. \quad \square$$

5.2. Growth functions and proof of Theorem 4.1

Growth of sizes of sets induced on finite domains of increasing sizes by various classes of binary-valued functions has long been studied in statistical learning theory. Vapnik and Chervonenkis (1971) introduced the concept of *growth function* $\Pi_{\mathcal{A}}(m) : \mathbb{N}_+ \to \mathbb{N}_+$ defined for any set of binary-valued functions $\mathcal{A} \subseteq \mathcal{S}(U)$ on any set U as

$$\Pi_{\mathcal{A}}(m) := \max_{X \subset U, \mathrm{card}X = m} \mathrm{card}(\mathcal{A}_{|X}).$$

So $\Pi_{\mathcal{A}}(m)$ measures the maximal number of dichotomies that a given family of functions $\mathcal{A}$ can generate on an m-point subset of U (recall that a *dichotomy* is a partition of a set into two disjoint subsets). In particular, for sets of input–output functions of networks with binary-valued outputs, the growth function gives

an upper bound on the sizes of sets of input–output functions induced on domains the size m.

A classical result by Schläfli (1901) proven already in the 19th century (see also (Cover, 1965)) gives an upper bound on the number of linearly separable dichotomies on m points in $\mathbb{R}^d$, i.e., partitions separated by hyper-planes $H_{e,b} = \{x \in \mathbb{R}^d \,|\, x \cdot e + b = 0\}$, $e \in S^{d-1}$, $b \in \mathbb{R}$. Denoting by Ξ_d the set of characteristic functions of half-spaces $H_{e,b}$ of $\mathbb{R}^d$, Schläfli's upper bound states

$$\Pi_{\Xi_d} \leq 2 \sum_{i=0}^{d} \binom{m-1}{i} \leq 2 \frac{m^d}{d!}. \tag{21}$$

The bound (21) shows that the growth function Π_{Ξ_d} is bounded by a polynomial of degree d, where d is the dimension of the ambient space $\mathbb{R}^d$. It the context of neurocomputing, the set Ξ_d corresponds to the set of all functions computable by perceptrons with the Heaviside activation function.

The classical result (21) was extended to estimates of growth functions of polynomially parameterized families of functions (Goldberg & Jerrum, 1995) and to some sets of input–output functions (Bartlett, Maiorov, & Meir, 1998). Here we employ a recent estimate of growth functions of deep perceptron networks with piecewise polynomial activation functions formulated in terms of their effective depth from Bartlett et al. (2019).

Proof of Theorem 4.1. To prove the statement, we combine Theorem 3.1 with an upper bound on the growth function $\Pi_{\mathcal{M}}(m)$ of the set of input–output functions

$$\mathcal{M} := \mathcal{M}(\mathcal{G}, \{k_l, l = 1, \ldots, L-1\}, \{\psi_{l,j}, l = 1, \ldots, L-1, j = 1, \ldots, k_l\})$$

of deep perceptron networks with an increasing size of the domain m that was proven in Bartlett et al. (2019, Theorem 7 and Remark 9) (see also Bartlett et al. (1998)). It states that under the assumption that $\bar{L}W = \sum_{i=1}^{L} W_i \leq m$,

$$\Pi_{\mathcal{M}}(m) \leq \Pi_{l=1}^{L} 2 \left(\frac{2ek_l p(1 + (l-1)\delta^{l-1})}{W_l} m \right)^{W_l} \leq \left(4ep1 + (L-1)\delta^{L-1}\right)^{\bar{L}W} m^{\bar{L}W}.$$

Thus by Theorem 3.1 the probability is bounded from below by

$$1 - \left(4ep1 + (L-1)\delta^{L-1}\right)^{\bar{L}W} m^{\bar{L}W} e^{-\frac{m\lambda^2}{2}}. \quad \square$$

6. Discussion

Our results show that in approximation of classifiers on large finite domains (such as discretized high-dimensional cubes) by deep perceptron networks, an effect of high-dimensional nature of classifiers can lead to a concentration of approximation errors. We analyzed conditions on network depth, number of its parameters, and types of activation functions under which the approximation errors behave almost deterministically.

Some probabilistic approaches to study of approximation by neural networks assume (often implicitly) a uniform probability. This assumption models situations where no prior knowledge about the tasks of interest is available. It can be used to prove existential results (see, e.g. Maiorov (1999), Telgarsky (2016)). In practical situations, often some vectors of features are rarely classified as positive, while others are more often to be like that. We considered product probability distributions on sets of binary-valued functions to keep variables independent but allow distributions of individual random variables to differ. Independence of random variables is an essential assumption in most theorems on concentration of their values, some of which we exploited. The cases when probability distributions generate random variables with some dependence (which often happens in real applications) are much more difficult to investigate. We initiated their study in Kůrková and Sanguineti (2021) by exploring correlations of random high-dimensional vectors.

Suitability of a class of networks can be assessed from the mean values around which approximation errors are concentrated. These values depend on the type of a probability distribution; they are largest for the uniform one. Uniformity has to be assumed when there is no prior knowledge. In this case, our theorems imply that for networks that are "reasonably small" (sizes of sets of their input–output functions do not grow exponentially with their domains), almost any uniformly randomly chosen function cannot be well approximated.

Our results indicate that the good approximation properties of deep networks mean that in many successful applications, distributions of tasks that such networks perform are highly non-uniform. For non-uniform distributions, approximation errors concentrate around values that can be much smaller than in the uniform case.

As mentioned in Section 4, the lower bounds from Theorem 4.1(ii) and Corollary 4.2(ii) are not tight, probabilities of concentrations of approximation errors might be higher than in our estimates. Some improvements of these bounds might be obtained by taking into account coherence of sets of input–output functions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

V.K. was partially supported by the Czech Science Foundation grant 22-02067S and the institutional support of the Institute of Computer Science RVO 67985807. M.S. is a member of GNAMPA-INdAM (Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni - Instituto Nazionale di Alta Matematica) and Visiting Professor at IMT School for Advanced Studies, Lucca, Italy. He was partially supported by the project of the PDGP DIT.AD021.104 "Optimization and Control Techniques" of the Institute of Marine Engineering, National Research Council of Italy, where he is research associate.

References

Bartlett, P. L., Harvey, N., Liaw, C., & Mehrabian, A. (2019). Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, *20*, 1–17.

Bartlett, P., Maiorov, V., & Meir, R. (1998). Almost linear VC-dimension bounds for piecewise polynomial networks. *Neural Computation*, *10*, 2159–2173.

Bellman, R. (1957). *Dynamic programming*. Princeton University Press.

Bianchini, M., & Scarselli, F. (2014). On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE Transactions on Neural Networks and Learning Systems*, *25*, 1553–1565.

Cover, T. (1965). Geometrical and statistical properties of systems of linear inequalities with applictions in pattern recognition. *IEEE Transactions on Electronic Computers*, *14*, 326–334.

Donoho, D., & Tanner, J. (2009). Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of Royal Society A*, *367*, 4273–4293.

Dubhashi, D. P., & Panconesi, A. (2009). *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press.

Goldberg, P. W., & Jerrum, M. R. (1995). Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, *18*(2), 131–148.
Gonon, L., Grigoryeva, L., & Ortega, J. P. (2023). Approximation bounds for random neural networks and reservoir systems. *Annals of Applied Probability*, *33*(1), 28–69.
Gorban, A. N., Golubkov, A., Grechuk, B., Mirkes, E. M., & Tyukin, I. Y. (2018). Correction of AI systems by linear discriminants: Probabilistic foundations. *Information Sciences*, *466*, 303–322.
Gorban, A. N., Makarov, V. A., & Tyukin, I. Y. (2019). The unreasonable effectiveness of small neural ensembles in high-dimensional brain. *Physics of Life Reviews*, *29*, 55–88.
Gorban, A. N., & Tyukin, I. Y. (2018). Blessing of dimensionality: mathematical foundations of the statistical physics of data. *Philosophical Transactions of Royal Society A*, *376*, 2017–2037.
Gorban, A. N., Tyukin, I. Y., Prokhorov, D. V., & Sofeikov, K. I. (2016). Approximation with random bases: Pro et contra. *Information Sciences*, *364–365*, 129–145.
Ito, Y. (1992). Finite mapping by neural networks and truth functions. *The Mathematical Scientist*, *17*, 69–77.
Kainen, P. C. (1997). Utilizing geometric anomalies of high dimension: when complexity makes computation easier. In K. Warwick, & M. Kárný (Eds.), *Computer-intensive methods in control and signal processing. The curse of dimensionality* (pp. 261–270). Boston, MA: Birkhäuser.
Kainen, P. C., Kůrková, V., & Sanguineti, M. (2012). Dependence of computational models on input dimension: Tractability of approximation and optimization tasks. *IEEE Transaction on Information Theory*, *58*, 1203–1214.
Kůrková, V. (2018). Constructive lower bounds on model complexity of shallow perceptron networks. *Neural Computing and Applications*, *29*, 305–315.
Kůrková, V. (2019). Limitations of shallow networks representing finite mappings. *Neural Computing and Applications*, *31*, 1783–1792.
Kůrková, V., & Sanguineti, M. (2016). Model complexities of shallow networks representing highly varying functions. *Neurocomputing*, *171*, 598–604.
Kůrková, V., & Sanguineti, M. (2017). Probabilistic lower bounds for approximation by shallow perceptron networks. *Neural Networks*, *91*, 34–41.
Kůrková, V., & Sanguineti, M. (2019). Classification by sparse neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(9), 2746–2754.
Kůrková, V., & Sanguineti, M. (2021). Correlations of random classifiers on large data sets. *Softcomputing*, *25*, 12641–12648.
Maiorov, V. E. (1999). On best approximation by ridge functions. *Journal of Approximation Theory*, *99*, 68–94.
Matoušek, J. (2002). *Lectures on discrete geometry*. New York: Springer.
McDiarmid, C. (1989). On the method of bounded differences. In J. Siemons (Ed.), *Surveys in combinatorics* (pp. 148–188). Cambridge: Cambridge University Press.
Milman, V. D., & Schechtman, G. (1986). *Lecture notes in mathematics*: *vol. 1200*, *Asymptotic theory of finite dimensional normed spaces*. Springer-Verlag.
Poggio, T., et al. (2017). Why and when can deep but not shallow networks avoid the curse of dimensionality: A review. *International Journal of Automation and Computing*, *14*(5), 503–519.
Schläfli, L. (1901). *Theorie Der Vielfachen Kontinuität*. Zürich: Zürcher & Furrer.
Telgarsky, M. (2016). Benefits of depth in neural networks. In *Proceedings of machine learning research, Vol. 49* (pp. 1517–1539).
Vapnik, V. N., & Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Doklady Akademii Nauk*, *16*(2), 264–279.
Vershynin, R. (2020). *High-dimensional probability*. Irvine: University of California.
Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks*, *94*, 103–114.