Neurocomputing 171 (2016) 598–604

Contents lists available at ScienceDirect

Neurocomputing

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

Model complexities of shallow networks representing highly varying functions

Věra Kůrková [a], Marcello Sanguineti [b]

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

ARTICLE INFO

Article history:
Received 7 November 2014
Received in revised form
18 May 2015
Accepted 1 July 2015
Communicated by B. Hammer
Available online 14 July 2015

Keywords:
Shallow networks
Model complexity
Highly varying functions
Chernoff Bound
Perceptrons
Gaussian kernel units

ABSTRACT

Model complexities of shallow (i.e., one-hidden-layer) networks representing highly varying multivariable $\{-1,1\}$-valued functions are studied in terms of variational norms tailored to dictionaries of network units. It is shown that bounds on these norms define classes of functions computable by networks with constrained numbers of hidden units and sizes of output weights. Estimates of probabilistic distributions of values of variational norms with respect to typical computational units, such as perceptrons and Gaussian kernel units, are derived via geometric characterization of variational norms combined with the probabilistic Chernoff Bound. It is shown that almost any randomly chosen $\{-1,1\}$-valued function on a sufficiently large d-dimensional domain has variation with respect to perceptrons depending on d exponentially.

1. Introduction

A widely used type of a neural-network architecture is the *one-hidden-layer network*. Typical computational units in the hidden layer are perceptrons, radial, and kernel units. Recently, one-hidden-layer networks have been called *shallow networks*, in contrast to *deep* ones, which contain two or more hidden layers (see, e.g., [1,2]).

A variety of learning algorithms for shallow networks were developed and successfully applied (see, e.g., [3] and the references therein). In addition to applications, theoretical analysis confirmed capabilities of shallow networks. For many types of computational units, shallow networks are known to be *universal approximators*, i.e., they can approximate up to any desired accuracy all continuous or $\mathcal{L}^p$ functions on compact subsets of $\mathbb{R}^d$. In particular, the universal approximation property holds for shallow networks with perceptrons having non-polynomial activation units [4,5], and radial and kernel units satisfying mild conditions [6–9].

However, the universal approximation capability requires potentially unlimited numbers of hidden units. This number, which plays the role of *model complexity* of a network, is a critical factor for practical implementations. Since typical neurocomputing applications deal with many variables, it is particularly important to understand how quickly model complexities of shallow networks grow with increasing input dimensions. Estimates of rates of approximation of various classes of multivariable functions by networks with increasing numbers of hidden units were derived and employed to obtain bounds on model complexities (see, e.g., [10–16] and the references therein).

On the other hand, limitations of computational capabilities of shallow networks are much less understood. Only few lower bounds on rates of approximation by these networks are known. Moreover, the bounds are mostly non-constructive and hold for types of computational units that are not commonly used [17,18]. Also the growth of sizes of weights is not well understood, although it was shown that in some cases, reasonable sizes of weights are more important for successful learning than bounds on the numbers of network units [19].

Recently, new hybrid learning algorithms for deep networks (such as convolutional and graph networks) were developed and applied to various pattern recognition tasks (see, e.g., [1,2,20–24]). However, a theoretical analysis identifying tasks for which shallow networks require considerably larger numbers of units and/or sizes of weights than deep ones is missing. In [25,26], Bengio et al. suggested that a cause of large model complexities of shallow networks might be in the "amount of variations" of functions to be computed and they focused their analysis on the d-dimensional parities on the d-dimensional Boolean cube. Recently, Bianchini and Scarselli [27] investigated

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

http://dx.doi.org/10.1016/j.neucom.2015.07.014
0925-2312/

limitations of shallow networks by employing Betti Numbers from algebraic topology.

In this paper, we investigate model complexity of shallow networks implementing $\{-1,1\}$-valued functions on finite subsets of d-dimensional spaces. Such functions represent binary classification tasks. Following the above-mentioned conjecture by Bengio et al. [25,26] about a connection between "amount of variations" and large model complexities of shallow networks, we investigate variations of functions in terms of *variational norms tailored to network units*. This concept has been successfully used as a tool to characterize classes of functions that can be approximated by networks with reasonable model complexities (see, e.g., [28–33]) and to study infinite-dimensional optimization problems [34–36]. Besides playing a critical role in estimates of rates of approximation of multivariable functions by shallow networks, the size of the variational norm of a function bounds from below the number of hidden units or sizes of output weights in such networks. We compare linear dependence on input dimension of variational norm of the d-dimensional parity with respect to perceptrons with an exponential growth of variational norm of the same function with respect to Gaussian kernel units having centers in the d-dimensional Boolean cube.

Using an argument based on the probabilistic Chernoff Bound, we show that for many common dictionaries, a representation of almost any uniformly randomly chosen $\{-1,1\}$-valued function on a sufficiently large finite domain by a shallow network requires intractably large number of units and/or sizes of output weights. For the dictionary of signum perceptrons, we derive on the network complexity lower bounds that depend on the ratio between the size of the domain of a function to be computed and the input dimension. In particular, we prove that every representation of a randomly chosen function on a discretized d-dimensional cube by a shallow network requires number of units and/or sizes of output weights that depend on d exponentially. A preliminary version of some results appeared in conference proceedings [37,38].

The paper is organized as follows. Section 2 presents basic concepts on shallow networks and dictionaries of computational units. Section 3 introduces variational norms and describes their main properties. In Section 4, lower bounds on variation with respect to Gaussian kernel units are derived. In Section 5, estimates of probabilistic distributions of sizes of G-variations are proven for various dictionaries, including those formed by signum perceptrons and generalized parities. Section 6 is a conclusive discussion.

2. Preliminaries

A widely used network architecture is a *one-hidden-layer network with a single linear output*, also called *shallow network*. Such a network with n hidden units computes input–output functions from the set

$$\operatorname{span}_n G := \left\{ \sum_{i=1}^{n} w_i g_i \,\middle|\, w_i \in \mathbb{R},\ g_i \in G \right\},$$

where G, called *dictionary*, is a set of functions computable by a given type of units. The *linear span* of G is denoted by $\operatorname{span} G$, i.e.,

$$\operatorname{span} G := \left\{ \sum_{i=1}^{n} w_i g_i \,\middle|\, w_i \in \mathbb{R},\ g_i \in G,\ n \in \mathbb{N} \right\}.$$

By X we denote the domain of functions computable by a network. Generally, X is a subset of $\mathbb{R}^d$ and $\operatorname{card} X$ denotes its cardinality. Shallow networks with *perceptrons* compute functions of the form $\sigma(v\cdot\,.+b): X\to\mathbb{R}$, where $\sigma:\mathbb{R}\to\mathbb{R}$ is an *activation function*. We denote by ϑ the *Heaviside activation function*, defined as

$$\vartheta(t) := 0 \text{ for } t<0 \quad \text{and} \quad \vartheta(t) := 1 \text{ for } t\geq 0$$

and by sgn the *signum activation function* $\operatorname{sgn}:\mathbb{R}\to\{-1,1\}$, defined as

$$\operatorname{sgn}(t) := -1 \text{ for } t<0 \quad \text{and} \quad \operatorname{sgn}(t) := 1 \text{ for } t\geq 0.$$

We denote by $H_d(X)$ the dictionary of functions on $X\subset\mathbb{R}^d$ computable by *Heaviside perceptrons*, i.e.,

$$H_d(X) := \{\vartheta(v\cdot\,.+b): X\to\{0,1\} \mid v\in\mathbb{R}^d, b\in\mathbb{R}\}$$

and by $S_d(X)$ the dictionary of functions on X computable by *signum perceptrons*, i.e.,

$$S_d(X) := \{\operatorname{sgn}(v\cdot\,.+b): X\to\{-1,1\} \mid v\in\mathbb{R}^d, b\in\mathbb{R}\}.$$

Note that $H_d(\mathbb{R}^d)$ is the *set of characteristic functions of half-spaces* of $\mathbb{R}^d$. We use the signum activation function as for X finite, all elements from $S_d(X)$ have the same norm $\sqrt{\operatorname{card} X}$, which is often convenient. From the point of view of model complexity, there is only a minor difference between dictionaries of signum and Heaviside perceptrons, as $\operatorname{sgn}(t)=2\vartheta(t)-1$ and $\vartheta(t)=\frac{\operatorname{sgn}(t)+1}{2}$. So any network having n signum perceptrons can be replaced with a network having $n+1$ Heaviside perceptrons and vice versa.

For $X, U\subseteq\mathbb{R}^d$, we denote by

$$K_d^a(X,U) := \{e^{-a\|\,.-u\|^2}: X\to\mathbb{R} \mid u\in U\}$$

the *dictionary of Gaussian kernel units on X with centers in U and width* $1/a$. In *Support Vector Machine* (*SVM*), $U=\{u_i,\ i=1,\dots,l\}$ is the set of points to be classified, among which some play the role of support vectors. When $X=U$, we write shortly $K_d^a(X)$.

By $P_d(\{0,1\}^d)$ we denote the *dictionary of generalized parities* defined as

$$P_d(\{0,1\}^d) := \{p_u^d: \{0,1\}^d\to\{-1,1\} \mid u\in\{0,1\}^d\},$$

where $p_u^d:\{0,1\}^d\to\{-1,1\}$ satisfies for every $u,x\in\{0,1\}^d$

$$p_u^d(x) := (-1)^{u\cdot x}.$$

In the case where $u_i=1$ for all $i=1,\dots,d$, p_u^d is the *d-dimensional parity* and we write shortly $p^d=p_u^d$.

We denote by

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

the *set of all real-valued functions on X* and by

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

the *set of all functions on X* with values in $\{-1,1\}$. It is easy to see that when $\operatorname{card} X=m$ and $X=\{x_1,\dots,x_m\}$ is a linear ordering of X, then the mapping $\iota:\mathcal{F}(X)\to\mathbb{R}^m$ defined as $\iota(f) := (f(x_1),\dots,f(x_m))$ is an isomorphism. So, on $\mathcal{F}(X)$ we have the Euclidean inner product defined as

$$\langle f,g\rangle := \sum_{u\in X} f(u)g(u)$$

and the Euclidean norm $\|f\| := \sqrt{\langle f,f\rangle}$. If $f\in\mathcal{F}(X)$, then

$$f^o := f/\|f\|$$

denotes its normalization. In contrast to the inner product $\langle\cdot,\cdot\rangle$ on $\mathcal{F}(X)$, we denote by $\cdot$ the inner product on $X\subset\mathbb{R}^d$, i.e., for $u,v\in X$,

$$u\cdot v := \sum_{i=1}^{d} u_i v_i.$$

3. Functions highly varying with respect to dictionaries

In this section, we show that the concept of "highly varying function", suggested by Bengio et al. [26] as a possible cause of

intractable complexity of shallow networks computing such functions, must be considered with respect to a type of network units. We formalize this concept in terms of a "variational norm", tailored to a given dictionary. We investigate such a norm with respect to the dictionary formed by Gaussians centered in $\{0,1\}^d$, considered by Bengio et al. [26].

As an example of a class of functions having "high-variations", Bengio et al. [26] considered parities on d-dimensional Boolean cubes $\{0,1\}^d$. They proved that a classification of points in $\{0,1\}^d$ according to their parities by Gaussian kernel units of any fixed width having centers in $\{0,1\}^d$ cannot be accomplished with less than $2^{d/2}$ units. The following theorem is a reformulation of their result [26, Theorem 2.4].

Theorem 3.1. *Let d be a positive integer, $a>0$, and $\{u_i \mid i=1,\dots,2^d\}$ an ordering of the set $\{0,1\}^d$. If for some bias $b\in\mathbb{R}$ and weights $\{w_i \mid i=1,\dots,2^d\}\subset\mathbb{R}$, $\operatorname{sgn}(\sum_{i=1}^{2^d} w_i e^{-a\|x-u_i\|^2}+b)=p^d(x)$ for all $x\in\{0,1\}^d$, then at least 2^{d-1} coefficients w_i are non-zero.*

Theorem 3.1 shows that sometimes the maximal generalization capability (maximal margin) is obtained at the expense of intractably large model complexity. The theorem implies that if, for some $b\in\mathbb{R}$, a function $f\in \operatorname{span}_n F_d^a(\{0,1\}^d)$ satisfies $f(x)-b=p^d(x)$ for all $x\in\{0,1\}^d$, then $n\geq 2^{d-1}$. On the other hand, it is well-known that when units in a shallow network are Heaviside or signum perceptrons, then merely $d+1$ units are sufficient to computate any generalized parity. Indeed, for all $x\in\{0,1\}^d$ one has

$$p_u^d(x)=\sum_{i=0}^{d}(-1)^i\,\vartheta(u\cdot x-i+1/2). \tag{1}$$

Geometrically, p_u^d can be represented as a plane wave orthogonal to the vector u. In particular, the parity p^d can be considered as a plane wave orthogonal to the diagonal of the cube.

The example of parities shows that the effect of high variations of a function on network complexity depends on a type of network units. In theory of approximation of functions by neural networks, the concept of *variation of a function with respect to a dictionary* was introduced by Barron [39] for Heaviside perceptrons as *variation with respect to half-spaces*. In Kůrková [40], this notion was extended to general bounded sets of functions, in particular dictionaries of computational units. For a bounded subset G of a normed linear space $(\mathcal{X},\|\cdot\|_{\mathcal{X}})$, *$G$-variation* (*variation with respect to the set G*), denoted by $\|\cdot\|_G$, is defined as

$$\|f\|_G:=\inf\left\{c\in\mathbb{R}_+ \,\middle|\, \frac{f}{c}\in \mathrm{cl}_{\mathcal{X}}\operatorname{conv}(G\cup -G)\right\},$$

where $-G:=\{-g \mid g\in G\}$, $\mathrm{cl}_{\mathcal{X}}$ denotes the closure with respect to the topology induced by the norm $\|\cdot\|_{\mathcal{X}}$, and conv is the convex hull. For properties of variation and its role in estimates of rates of approximation, see [13,15,29–31,33,36].

The next proposition, which follows easily from the definition of G-variation, shows that $\|f\|_G$ reflects both the number of hidden units and the sizes of output weights in a shallow network with units from G representing f (see [41]).

Proposition 3.2. *Let G be a bounded subset of a normed linear space $(\mathcal{X},\|.\|)$. Then, for every $f\in\mathcal{X}$ one has*

(i) $\|f\|_G\leq\left\{\sum_{i=1}^{k}|w_i| \,\middle|\, f=\sum_{i=1}^{k} w_i g_i, w_i\in\mathbb{R}, g_i\in G, k\in\mathbb{N}\right\}$;

(ii) *for G finite with* $\operatorname{card} G=k$,

$$\|f\|_G=\min\left\{\sum_{i=1}^{k}|w_i| \,\middle|\, f=\sum_{i=1}^{k} w_i g_i, w_i\in\mathbb{R}, g_i\in G\right\}.$$

Hence, any representation of a function with large G-variation by a network with units from a dictionary G must have large number of units and/or the absolute values of some output weights must be large.

To derive lower bounds on variational norms, we shall exploit the following bound from [41] (see also [33]), which shows that functions nearly orthogonal to all elements of a dictionary G have large G-variations. By $G^{\perp}$ is denoted the *orthogonal complement of G*.

Theorem 3.3. *Let $(\mathcal{X},\|.\|_{\mathcal{X}})$ be a Hilbert space and G its bounded subset. Then, for every $f\in\mathcal{X}\setminus G^{\perp}$ one has*

$$\|f\|_G\geq\frac{\|f\|^2}{\sup_{g\in G}|g\cdot f|}.$$

The next proposition summarizes elementary properties of the variation norm (see [13]).

Proposition 3.4. (i) *Let $X\subset\overline{X}\subseteq\mathbb{R}^d$, $\overline{G}\subset\mathcal{F}(\overline{X})$ be a dictionary of functions on $\overline{X}$ and $G\subset\mathcal{F}(X)$ a dictionary on X obtained by restricting the functions from $\overline{G}$ to X. Then, for every $\overline{f}\in\mathcal{F}(\overline{X})$ and $f=\overline{f}|_X\in\mathcal{F}(X)$ the following hold.*

(i) $\|f\|_G\leq\|\overline{f}\|_{\overline{G}}$.

(ii) *Let $X\subset\mathbb{R}^d$, G_1,G_2 be bounded subsets of $\mathcal{F}(X)$. If for some $c>0$ for all $g\in G_1$, $\|g\|_{G_2}\leq c$ then for all f $\|f\|_{G_2(X)}\leq c\|f\|_{G_1(X)}$.*

Proposition 3.4(i) implies that a lower bound on $\|f\|_G$ applies to $\|\overline{f}\|_{\overline{G}}$ and an upper bound on $\|\overline{f}\|_{\overline{G}}$ applies to $\|f\|_G$. In particular, lower bounds for functions on finite subsets of $\mathbb{R}^d$, e.g., discretized cubes such as $\{0,1\}^d$, apply to functions on infinite sets containing them.

The next proposition shows that variation with respect to signum perceptrons is bounded from above by variation with respect to Heaviside perceptrons.

Proposition 3.5. *For every positive integer d and every $X\subseteq\mathbb{R}^d$,*

$$\|\cdot\|_{S_d(X)}\leq\|\cdot\|_{H_d(X)}.$$

Proof. As $\vartheta(e\cdot x+b)=\frac{1}{2}\operatorname{sgn}(e\cdot x+b)+\frac{1}{2}=\frac{1}{2}\operatorname{sgn}(e\cdot x+b)+\frac{1}{2}\operatorname{sgn}(e\cdot(1,\dots,1)+1)$, for every $h\in H_d(X)$ one has $\|h\|_{S_d(X)}\leq 1$. So, by Proposition 3.4(ii) the statement holds. □

4. Variation with respect to Gaussian kernel units

In this section, we investigate variation of the d-dimensional parity with respect to Gaussian kernel units with centers in the Boolean cube $\{0,1\}^d$. We show that the size of this norm increases with d exponentially.

The representation (1) of a generalized parity p_u^d as a shallow network with $d+1$ Heaviside perceptrons and Proposition 3.5 implies that

$$\|p_u^d\|_{S_d(\mathbb{R}^d)}\leq\|p_u^d\|_{H_d(\mathbb{R}^d)}\leq d+1.$$

Hence, variations with respect to signum or Heaviside perceptrons of all d-dimensional generalized parities grow with d only linearly. On the other hand, Theorem 3.1 by Bengio et al. [26] proves that every representation of the parity p^d by a shallow network with Gaussian kernel units with centers in $\{0,1\}^d$ requires at least 2^{d-1} units. The next theorem shows that $K_d^a(\{0,1\}^d)$-variation of p^d grows with d exponentially.

Theorem 4.1. *For every positive integer d and every $a>0$,*

$$\|p^d\|_{K_d^a(\{0,1\}^d)}>2^{d/2}.$$

Proof. By Theorem 3.3,

$$\|p^d\|_{K_d^a(\{0,1\}^d)} \geq \frac{\|p^d\|}{\sup_{g \in F_a^d(\{0,1\}^d)} |\langle p^d, g\rangle|}.$$

For the Gaussian g_0 centered at $(0,\dots,0)$, one has $\langle p^d, g_0\rangle = \sum_{k=0}^{d} (-1)^k \binom{d}{k} e^{-ak}$. By the binomial formula, we have

$$\langle p^d, g_0\rangle = \sum_{k=0}^{d} (-1)^k \binom{d}{k} e^{-ak} = (1-e^{-a})^d.$$

By a suitable transformation of the coordinate system, we obtain the same value of the inner product with p^d for the Gaussian g_x centered at any $x \in \{0,1\}^d$ such that $p^d(x) = 1$. When the Gaussian g_x is centered at x with $p^d(x) = -1$, we get the same absolute value of the inner product by replacing p^d with $-p^d$ and by a transformation of the coordinate system. As $\|p^d\| = 2^{d/2}$, we get $\|p^d\|_{K_d^a(\{0,1\}^d)} \geq \frac{2^{d/2}}{(1-e^{-a})^d} > 2^{d/2}$. □

Theorem 4.1 shows that for every value of $a > 0$, the variation of the d-dimensional parity with respect to the dictionary of Gaussian kernel units of a fixed width $1/a$ with centers in $\{0,1\}^d$ grows at least exponentially with d.

When the proof technique used to derive Theorem 4.1 is applied to a general $\{-1,1\}$-valued function on $\{0,1\}^d$, it provides the following weaker lower bound.

Theorem 4.2. *For every positive integer d, every $a > 0$, and every function $f^d : \{0,1\}^d \to \{-1,1\}$,*

$$\|f^d\|_{K_d^a(\{0,1\}^d)} \geq \left(\frac{2}{(1+e^{-a})^2}\right)^{d/2}.$$

Proof. By Theorem 3.3, $\|f^d\|_{K_d^a(\{0,1\}^d)} \geq \frac{\|f^d\|}{\sup_{g \in F_a^d(\{0,1\}^d)} |\langle f^d, g\rangle|}$. For the Gaussian g_0 centered at $(0,\dots,0)$, we have

$$|\langle f^d, g_0\rangle| = \Big|\sum_{y \in \{0,1\}^d} f(y) e^{-a\|y-x\|^2}\Big| \leq \sum_{k=0}^{d} \binom{d}{k} e^{-ak}$$

and so by the binomial formula we obtain $|\langle f^d, g_0\rangle| \leq (1+e^{-a})^d$. The same argument as in the proof of Theorem 4.1 shows that this upper bound holds for the Gaussian centered g_x at any $x \in \{0,1\}^d$. Hence, $\|f^d\|_{K_d^a(\{0,1\}^d)} \geq \frac{2^{d/2}}{(1+e^{-a})^d} = \left(\frac{2}{(1+e^{-a})^2}\right)^{d/2}$. □

The rate of growth of the lower bound $\left(\frac{2}{(1+e^{-a})^2}\right)^{d/2}$ from Theorem 4.2 depends on the width $1/a$ of the Gaussian kernel. For a "sufficiently narrow" Gaussian, whose width $1/a$ satisfies $e^{-a} < \sqrt{2}-1$, i.e., $a > -\ln(\sqrt{2}-1) \simeq 0.88$, we have $\frac{2}{(1+e^{-a})^2} > 1$ and thus the lower bound from Theorem 4.2 grows exponentially with d. The theorem implies that any shallow network with "sufficiently narrow" Gaussians (i.e., with "large enough" values of a) and centers in $\{0,1\}^d$ representing a signum perceptron or a generalized parity, must have a number of units and/or sizes of some of the output weights that depend on d exponentially.

Bengio et al. [25] proved that when, instead of $\{0,1\}^d$, the set of centers of "sufficiently narrow" Gaussian kernels is in a properly chosen finite sets of points on the diagonal of the cube $[0,1]^d$, then there exists a function on $\{0,1\}^d$ with the same sign as the parity, which can be represented by network with merely $d+1$. Inspection of their proof shows that the variation of this function with respect to such a dictionary is at most $d+1$ (see [25, Remark 4.8]). So, variation with respect to $K_d^a(\{0,1\}^d, U)$ strongly depends on the choice of the set U of centers.

5. Probability distributions of functions with large variations

In this section we consider functions randomly chosen in $\mathcal{B}(X)$ with respect to the uniform distribution; for short, we shall write "uniformly randomly chosen". For such functions, we estimate the probability distributions of their variations with respect to "relatively small" dictionaries G formed by $\{-1,1\}$-valued functions on "sufficiently large" finite subsets X of $\mathbb{R}^d$. For a dictionary $G(X) \subset \mathcal{B}(X)$, we consider the function $\|\cdot\|_{G(X)} : \mathcal{B}(X) \to \mathbb{R}_+$ as a random variable. When X is finite, this random variable has finite range and so it can be interpreted as a discrete random variable.

To obtain a lower bound on probability that a uniformly randomly chosen $\{-1,1\}$-valued function has G-variation greater than or equal to a prescribed value, we shall exploit the *Chernoff Bound* on the probability distribution of sums of independently identically distributed (i.i.d.) random variables with values in $[-1,1]$ [42, p. 393] (see also [43,44]).

Theorem 5.1 (*Chernoff Bound*). *Let m be a positive integer, $Y_1,\dots,Y_m$ i.i.d. random variables with values in $[-1,1]$, and $\lambda > 0$. Then*

$$\Pr\left(\left|\sum_{i=1}^{m} Y_i - E(Y_i)\right| \geq \lambda\right) \leq 2e^{-\lambda^2/2m}.$$

Combining the Chernoff Bound with the geometric lower bound on variational norm provided by Theorem 3.3, we obtain the next theorem.

Theorem 5.2. *Let d be a positive integer, $X \subset \mathbb{R}^d$ with $\operatorname{card} X = m$, $G(X) \subset \mathcal{B}(X)$ with $\operatorname{card} G(X) = k$, f uniformly randomly chosen from $\mathcal{B}(X)$, and $\varepsilon \in (0,1)$. Then*

$$\Pr(\|f\|_{G(X)} \geq 1/\varepsilon) \geq 1 - 2k\, e^{-(m\varepsilon^2)/2}.$$

Proof. Let $W_\varepsilon(G(X)) := \{f \in \mathcal{B}(X) \mid \|f\|_G(X) \geq 1/\varepsilon\}$. By Theorem 3.3, $W_\varepsilon(G(X))$ contains all $f \in \mathcal{B}(X)$ satisfying for all $g \in G$, $\langle f^o, g^o\rangle \leq \varepsilon$, where the superscript "o" denotes normalization (i.e., $f^o := f/\|f\|$ and $g^o := g/\|g\|$). Thus $W_\varepsilon(G(X))$ contains the complement of the set

$$\bigcup_{g \in G(X)} \{f \in \mathcal{B}(X) \mid |\langle f^o, g^o\rangle| \geq \varepsilon\}$$

and so

$$\Pr(f \in W_\varepsilon(G(X))) \geq 1 - \sum_{g \in G(X)} \Pr(|\langle f^o, g^o\rangle| \geq \varepsilon).$$

We show that for every function $h \in \mathcal{B}(X)$ one has $\Pr(|\langle f^o, h^o\rangle| \geq \varepsilon) \leq 2e^{-m\varepsilon^2/2}$.

First, we verify that this holds for the constant function f_1 defined for all $x \in X$ as $f_1(x) := 1$. Let $X := \{x_1,\dots,x_m\}$. For every $h \in \mathcal{B}(X)$, we have $\langle h, f_1\rangle = \sum_{i=1}^{m} h(x_i)$. By the Chernoff Bound applied to i.i.d. variables $Y_i \in \{-1,1\}$, $i = 1,\dots,m$, such that $\Pr(Y_i = 1) = \Pr(Y_i = -1) = \frac{1}{2}$, we get

$$\Pr(|\langle f_1, h\rangle| \geq \lambda) = \Pr\left(\left|\sum_{i=1}^{m} h(x_i)\right| \geq \lambda\right) \leq 2e^{-\lambda^2/2m}.$$

As for all $f \in \mathcal{B}(X)$ one has $\|f\| = \sqrt{m}$, setting $\varepsilon := \lambda/m$, we get

$$\Pr(|\langle h^o, f_1^o\rangle| \geq \varepsilon) \leq 2e^{-\lambda^2/2m} = 2e^{-m\varepsilon^2/2}.$$

Any $f \in \mathcal{B}(X)$ can be obtained from f_1 by a finite sequence of sign-flips $F_x : \mathcal{B}(X) \to \mathcal{B}(X)$ defined as $F_x(f)(x) := -f(x)$ and $F_x(f)(y) := f(y)$ for all $y \neq x$. As the inner product is invariant under sign-flipping, for all $f \in \mathcal{B}(X)$ the probability distribution of inner products $\langle f^o, h^o\rangle$ on $\mathcal{B}(X)$ satisfies $\Pr(|\langle f^o, h^o\rangle| \geq \varepsilon) \leq 2e^{-m\varepsilon^2/2}$. So, for every $g \in G$ we get

$$\Pr(|\langle f^o, g^o\rangle| \geq \varepsilon) \leq 2e^{-m\varepsilon^2/2}.$$

Thus $\Pr(f \in W_\varepsilon(G(X))) \geq 1 - 2ke^{-m\varepsilon^2/2}$. □

Theorem 5.2 can be applied to dictionaries $G(X)$ of functions on $X \subset \mathbb{R}^d$ with $\operatorname{card} X = m$ such that $\operatorname{card} G(X) = k$ is "relatively small"

with respect to the cardinality 2^m of the set of all functions in $\mathcal{B}(X) := \{f \mid f : X \to \{-1,1\}\}$. In particular, for dictionaries $G(X) \subset \mathcal{B}(X)$ of cardinalities at most $e^{\rho(\log_2 m)}$, where ρ is a polynomial, we get the next corollary.

Corollary 5.3. *Let d be a positive integer, $X \subset \mathbb{R}^d$ with* card $X = m$, *$\rho(.)$ a polynomial, $G(X) \subset \mathcal{B}(X)$ with* card $G(X) \le e^{\rho(\log_2 m)}$, *$f$ uniformly randomly chosen in $\mathcal{B}(X)$, and $\varepsilon \in (0,1)$. Then*

$$\Pr\left(\|f\|_{G(X)} \ge 1/\varepsilon\right) \ge 1 - 2e^{-((m\varepsilon^2 - 2\rho(\log_2 m))/2)}.$$

Examples of dictionaries formed by functions on $\{0,1\}^d$ with values in $\{-1,1\}$ such that cardinalities of these dictionaries are "relatively small" with respect to the cardinality 2^{2^d} of the set of all functions in $\mathcal{B}(\{0,1\}^d)$ are the dictionaries $S_d(\{0,1\}^d)$ of signum perceptrons and $P_d(\{0,1\}^d)$ of generalized parities. Obviously, card $P_d(\{0,1\}^d) = 2^d$. The upper bound card $S_d(\{0,1\}^d) \le 2^{d^2}$ is well-known (see, e.g., [45]).

Corollary 5.4. *Let d be a positive integer, f uniformly randomly chosen in $\mathcal{B}(\{0,1\}^d)$, and $\varepsilon > 0$. Then*

(i) $\Pr\left(\|f\|_{P_d(\{0,1\}^d)} \ge 1/\varepsilon\right) \ge 1 - e^{(-2^d\varepsilon^2 + (d+1)\ln 2)/2}$;

(ii) $\Pr\left(\|f\|_{S_d(\{0,1\}^d)} \ge 1/\varepsilon\right) \ge 1 - e^{-(2^d\varepsilon^2 + (d^2+1)\ln 2)/2}$.

Proof.

(i) The estimate follows from Theorem 5.2 applied to a dictionary of cardinality 2^d.

(ii) The estimate follows by Theorem 5.2 combined with a classical result by Schläfli [46] (see also [45, Theorem 13.2, p. 561], [47, p. 33]) showing that for all $d > 1$, card $S_d(\{0,1\}^d) =$ card $H_d(\{0,1\}^d) < 2^{d^2}$, and the expression $2^x = e^{x\ln 2}$. □

For example, setting $\varepsilon = 2^{-d/4}$ we obtain from Corollary 5.4 the lower bound

$$\Pr\left(\|f\|_{S_d(\{0,1\}^d)} \ge 2^{d/4}\right) \ge 1 - e^{-(2^{d/2} + (d^2+1)\ln 2)/2} \tag{2}$$

on the probability that a function f uniformly randomly chosen in $\mathcal{B}(\{0,1\}^d)$ has variation with respect to signum perceptrons larger than $2^{d/4}$. The estimate (2) implies that for growing dimension d, the probabilistic measure of the subset of functions in $\mathcal{B}(\{0,1\}^d)$ having an exponentially increasing lower bound on $S_d(\{0,1\}^d)$-variations approaches 1 at an exponential rate. In other words, $S_d(\{0,1\}^d)$-variations of most functions in $\mathcal{B}(\{0,1\}^d)$ depend on d exponentially.

However, the only concrete example of a function in $\mathcal{B}(\{0,1\}^d)$ with exponentially growing $S_d(\{0,1\}^d)$-variation of which we are aware is the well-known function *inner product mod* 2 from theory of Boolean functions [48]. For every even positive integer d, we denote it by $\beta_d : \{0,1\}^d \to \{0,1\}$. It is defined for all $x \in \{0,1\}^d$ as

$$\beta_d(x) := 1 \text{ if } l(x) \cdot r(x) \text{ is odd} \quad \text{and} \quad \beta_d(x) := 0 \text{ if } l(x) \cdot r(x) \text{ is even},$$

where $l(x), r(x) \in \{0,1\}^{d/2}$ are set for every $i = 1, \ldots, d/2$ as $l(x)_i := x_i$ and $r(x)_i := x_{d/2+i}$. As we are considering functions with values in $\{-1,1\}$, we use $\overline{\beta}_d : \{0,1\}^d \to \{-1,1\}$ defined as

$$\overline{\beta}_d(x) := (-1)^{l(x)\cdot r(x)}. \tag{3}$$

It was shown in [41, Theorem 3.7 and the discussion before Lemma 3.5] that

$$\|\overline{\beta}_d\|_{S_d(\{0,1\}^d)} = \Omega(2^{d/6}),$$

where for two functions $g, h : \mathbb{N} \to \mathbb{R}$ we write $h = \Omega(g(d))$ when there exist a positive constant c and $n_0 \in \mathbb{N}$ such that for all $n \ge n_o$ one has $h(n) \ge c\, g(n)$ [49]. Note that by Proposition 3.4(ii), also $\|\overline{\beta}_d\|_{H_d(\{0,1\}^d)} \ge \Omega(2^{d/6})$.

Theorem 5.2 can be applied to dictionaries of signum perceptrons on general finite domains $X \subset \mathbb{R}^d$. Upper bounds on card $S_d(X)$ follow from estimates of numbers of linearly separable dichotomies (i.e., partitions into two subsets) of finite subsets of $\mathbb{R}^d$. Various such estimates were derived by several authors (see the references in the discussion after [50, Theorem 1]) starting from the results by Schläfli [46]. We use the following estimate, based on a result from [50].

Theorem 5.5. *For every d and every $X \subset \mathbb{R}^d$ such that* card $X = m$,

$$\text{card } S_d(X) \le 2\sum_{i=0}^{d} \binom{m-1}{i}.$$

Proof. The number of linearly separable dichotomies of an arbitrary set of m points in $\mathbb{R}^d$ is smaller than equal to the number of such dichotomies of a set of m points such that no $d+1$ points lie on the same hyperplane. The latter number is bounded from above by $2\sum_{i=0}^{d} \binom{m-1}{i}$ (see, [50, Table 1, row 2]), hence the statement follows. □

By combining Theorems 5.2 and 5.5 with an upper bound on a partial binomial sum, we derive the next corollary.

Corollary 5.6. *Let d be a positive integer, $X \subset \mathbb{R}^d$ with* card $X = m$, *f uniformly randomly chosen in $\mathcal{B}(X)$, and $\varepsilon \in (0,1)$. Then*

$$\Pr\left(\|f\|_{S_d(X)} \ge 1/\varepsilon\right) \ge 1 - 4\frac{m^d}{d!}e^{-m\varepsilon^2/2}.$$

Proof. It is well-known (see [51, p. 43] and [52]) that

$$\sum_{i=0}^{d} \binom{m-1}{i} \le \frac{m^d}{d!}. \tag{4}$$

Eq. (4) together with Theorem 5.5 implies that the cardinality of the set $S_d(X)$ is bounded from above by

$$\text{card } S_d(X) \le 2\frac{m^d}{d!}. \tag{5}$$

By combining the estimate (5) with Theorem 5.2, we obtain the statement. □

Note that a similar estimate as the upper bound stated in Corollary 5.6 for the dictionary $S_d(X)$ of signum perceptrons can be obtained for the dictionary of characteristic functions of d-dimensional balls. This estimate follows by the upper bound $2\sum_{i=0}^{d} \binom{m}{i}$ on its cardinality from [50, Table 1, row 3].

Our last estimate of this section is based on an upper bound on a partial sum of binomials in terms of the *binary entropy function* Υ defined for every $q \in [0,1]$ as

$$\Upsilon(q) := -q\log_2(q) - (1-q)\log_2(1-q).$$

Corollary 5.7. *Let d be a positive integer, $X \subset \mathbb{R}^d$ with* card $X = m$ *such that $d/(m-1) \le 1/2$, and f uniformly randomly chosen in $\mathcal{B}(X)$. Then*

$$\Pr\left(\|f\|_{S_d(X)} \ge 1/\varepsilon\right) \ge 1 - e^{-(m\varepsilon^2/2) + (\Upsilon(d/(m-1))m + 2)\ln 2}.$$

Proof. For every $\alpha \in (0, 1/2]$ and every n, the partial sum of binomials satisfies [53, Lemma 16.19, p. 427].

$$\sum_{i=0}^{\lfloor \alpha n \rfloor} \binom{n}{i} \le 2^{\Upsilon(\alpha)n} \tag{6}$$

So, by setting $\alpha := d/m-1$ we get $\alpha(m-1)=d$ and so by Theorem 5.5 and Eq. (6)

$$\text{card } S_d(X) \leq 2\sum_{i=0}^{d}\binom{m-1}{i} \leq 2^{\Upsilon(d/m-1)(m-1)+1}.$$

Thus

$$\Pr\left(\|f\|_{S_d(X)} \geq 1/\varepsilon\right) \geq 1-2^{\Upsilon(d/m-1)(m-1)+2}e^{-m\varepsilon^2/2}. \quad (7)$$

As $2^x = e^{x\ln 2}$, the inequality (7) proves the statement. □

Note that Corollary 5.7 provides a useful lower bound only when $\Upsilon(d/m-1)$ is sufficiently small, i.e., when the size of the domain m is much larger than the input dimension d. This happens for large d, when the domain X a discretized d-dimensional cube.

6. Discussion

We investigated limitations of capabilities of shallow networks to implement highly varying functions without intractably large growth of model complexity. We followed the suggestion of Bengio et al. [25,26] that a cause of intractable increase of complexity of shallow networks might be "amount of variations of functions" to be implemented. We proposed a formalization of the concept of highly varying function in terms of a variational norm tailored to a particular dictionary of computational units. We showed that this concept, which has been successfully used in nonlinear approximation schemes, plays a useful role also in the investigation of complexity of shallow networks, reflecting both number of network units and sizes of its output weights. Note that the characterization of classes of functions defined by constraints on both number of gates and sizes of output weights also plays an important role in theory of circuit complexity [48].

On the example of d-dimensional parities on $\{0,1\}^d$ studied by Bengio [25,26], we demonstrated that the concept of highly varying functions must be taken with respect to the type of computational units. We proved that variation of the d-dimensional parity with respect to Gaussians with centers in $\{0,1\}^d$ grows at least exponentially, whereas it is well-known and easy to show that its variation with respect to perceptrons grows with d merely linearly.

We investigated probability distributions of sizes of variations with respect to "relatively small" dictionaries. We derived lower bounds on complexity of shallow networks depending on the ratio between the size of the domain of a function to be represented and the input dimension. We proved that for the dictionaries of signum perceptrons and generalized parities on $\{0,1\}^d$, almost any randomly chosen $\{-1,1\}$-valued function has variation depending on d exponentially.

Our results are probabilistic and existential. They merely show that the majority of functions on large domains cannot be tractably computed by shallow networks with commonly used computational units. The concrete construction of functions with large variations is a subject of our future research. Note also that our rather negative result for shallow networks holds for functions randomly chosen in $\mathcal{B}(X)$ with respect to the uniform distribution, whereas it is unlikely that distributions of functions to be computed by neural networks in real-life applications are uniform.

From the practical point of view, our investigation suggests the use of computational models that aggregate various types of units, e.g., perceptrons with radial and kernel units. As our examples illustrate, the choice of computational units has a strong impact on model complexities of shallow networks. We proved that there exist functions whose variations with respect to perceptrons grow with input dimension linearly, whereas variations with respect to kernel units grow exponentially. As an example of such class of functions, we considered generalized parities that, up to a scaling factor, form a Fourier basis of the set of functions on the Boolean cube $\{0,1\}^d$. As an orthogonal set, this basis represents a sparse dictionary and represents a useful tool in the analysis of Boolean functions. Implications of relationships between sparse dictionaries and overcomplete dictionaries such as perceptrons are a subject of our work in progress.

Acknowledgments

V.K. was partially supported by the grant COST LD13002 of the Ministry of Education of the Czech Republic and institutional support of the Institute of Computer Science RVO 67985807. Her visit to the University of Genova was supported by a 2013 GNAMPA-INdAM (Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni – Istituto Nazionale di Alta Matematica) grant for Visiting Professors. M.S. was partially supported by the Progetto di Ricerca di Ateneo 2013 "Processing High-Dimensional data with Applications to Life Sciences", granted by the University of Genova. M.S. is a member of GNAMPA-INdAM.

References

[1] Y. Bengio, Learning deep architectures for AI, Found. Trends Mach. Learn. 2 (2009) 1–127.
[2] G.E. Hinton, S. Osindero, Y.W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (2006) 1527–1554.
[3] T.W.S. Chow, S.Y. Cho, Neural networks and computing: learning algorithms and applications, Imperial College Press, London, 2007.
[4] M. Leshno, V.Y. Lin, A. Pinkus, S. Schocken, Multilayer feedforward networks with a nonpolynomial activation function can approximate any function, Neural Netw. 6 (1993) 861–867.
[5] A. Pinkus, Approximation theory of the MLP model in neural networks, Acta Numer. 8 (1999) 143–195.
[6] J. Park, I. Sandberg, Approximation and radial-basis-function networks, Neural Comput. 5 (1993) 305–316.
[7] H.N. Mhaskar, Versatile Gaussian networks, in: Proceedings of IEEE Workshop of Nonlinear Image Processing, 1995, pp. 70–73.
[8] V. Kůrková, Some comparisons of networks with radial and kernel units, in: A. Villa, W. Duch, P. Érdi, F. Masulli, G. Palm (Eds.), Neural Networks: Brain-inspired Computing and Machine Learning Research, Lecture Notes in Computer Science, vol. 7553, Springer, Berlin, Heidelberg, 2012, pp. II. 17–24.
[9] V. Kůrková, P.C. Kainen, Comparing fixed and variable-width Gaussian networks, Neural Netw. 57 (2014) 23–28.
[10] A.R. Barron, Universal approximation bounds for superpositions of a sigmoidal function, IEEE Trans. Inf. Theory 39 (1993) 930–945.
[11] D. Costarelli, R. Spigler, Multivariate neural network operators with sigmoidal activation functions, Neural Netw. 48 (2013) 72–77.
[12] F. Girosi, G. Anzellotti, Rates of convergence for radial basis functions and neural networks, in: R.J. Mammone (Ed.), Artificial Neural Networks for Speech and Vision, Chapman & Hall, London, 1993, pp. 97–113.
[13] V. Kůrková, M. Sanguineti, Comparison of worst-case errors in linear and neural network approximation, IEEE Trans. Inf. Theory 48 (2002) 264–275.
[14] V. Kůrková, M. Sanguineti, Geometric upper bounds on rates of variable-basis approximation, IEEE Trans. Inf. Theory 54 (2008) 5681–5688.
[15] P.C. Kainen, V. Kůrková, M. Sanguineti, Dependence of computational models on input dimension: tractability of approximation and optimization tasks, IEEE Trans. Inf. Theory 58 (2012) 1203–1214.
[16] H.N. Mhaskar, On the tractability of multivariate integration and approximation by neural networks, J. Complex. 20 (2004) 561–590.
[17] V. Maiorov, On best approximation by ridge functions, J. Approx. Theory 99 (1999) 68–94.
[18] V. Maiorov, A. Pinkus, Lower bounds for approximation by MLP neural networks, Neurocomputing 25 (1999) 81–91.
[19] P.L. Bartlett, The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network, IEEE Trans. Inf. Theory 44 (1998) 525–536.
[20] M. Bianchini, M. Maggini, L. Sarti, F. Scarselli, Recursive neural networks for processing graphs with labelled edges: theory and applications, Neural Netw. 18 (2005) 1040–1050.
[21] A. Pucci, M. Gori, M. Hagenbuchner, F. Scarselli, A.-C. Tsoi, Investigation into the application of graph neural networks to large-scale recommender systems, Syst. Sci. 32 (2006) 17–26.
[22] Y. LeCun, Y. Bengio, Convolutional networks for images, speech, and time series, in: M. Arbib (Ed.), The Handbook of Brain Theory and Neural Networks, MIT Press, Cambridge, MA, 1995, pp. 255–258.

[23] F. Scarselli, M. Gori, A.-C. Tsoi, M. Hagenbuchner, G. Monfardini, Computational capabilities of graph neural networks, IEEE Trans. Neural Netw. 20 (2009) 81–102.
[24] F. Scarselli, M. Gori, A.-C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE Trans. Neural Netw. 20 (2009) 61–80.
[25] Y. Bengio, O. Delalleau, N.L. Roux, The Curse of Dimensionality for Local Kernel Machines, Technical Report 1258, Département d'Informatique et Recherche Opérationnelle, Université de Montréal, 2005. ⟨http://www.iro.umontreal.ca/lisa/pointeurs/tr1258.pdf⟩.
[26] Y. Bengio, O. Delalleau, N.L. Roux, The curse of highly variable functions for local kernel machines. in: Advances in Neural Information Processing Systems, vol. 18, MIT Press, Cambridge, MA, 2006, pp. 107–114.
[27] M. Bianchini, F. Scarselli, On the complexity of neural network classifiers: a comparison between shallow and deep architectures, IEEE Trans. Neural Netw. Learn. Syst. 25 (2014) 1553–1565.
[28] V. Kůrková, High-dimensional approximation and optimization by neural networks, in: J. Suykens, et al. (Eds.), Advances in Learning Theory: Methods, Models, and Applications (NATO Science Series III: Computer and Systems Sciences), vol. 190, IOS Press, Amsterdam, 2003, pp. 69–88.
[29] V. Kůrková, Minimization of error functionals over perceptron networks, Neural Comput. 20 (2008) 250–270.
[30] P.C. Kainen, V. Kůrková, A. Vogt, A Sobolev-type upper bound for rates of approximation by linear combinations of heaviside plane waves, J. Approx. Theory 147 (2007) 1–10.
[31] P.C. Kainen, V. Kůrková, An integral upper bound for neural network approximation, Neural Comput. 21 (10) (2009) 2970–2989.
[32] P.C. Kainen, V. Kůrková, M. Sanguineti, Complexity of Gaussian radial-basis networks approximating smooth functions, J. Complex. 25 (2009) 63–74.
[33] V. Kůrková, Complexity estimates based on integral transforms induced by computational units, Neural Netw. 33 (2012) 160–167.
[34] S. Giulini, M. Sanguineti, Approximation schemes for functional optimization problems, J. Optim. Theory Appl. 140 (2009) 33–54.
[35] G. Gnecco, M. Sanguineti, Estimates of variation with respect to a set and applications to optimization problems, J. Optim. Theory Appl. 145 (2010) 53–75.
[36] G. Gnecco, M. Sanguineti, On a variational norm tailored to variable-basis approximation schemes, IEEE Trans. Inf. Theory 57 (2011) 549–558.
[37] V. Kůrková, M. Sanguineti, Can two hidden layers make a difference? in: M. Tomassini, A. Antonioni, F. Daolio, P. Buesser (Eds.), Proceedings of ICANNGA 2013 Conference on Adaptive and Natural Computing Algorithms, Lecture Notes in Computer Science, vol. 7824, Springer, Berlin, Heidelberg, 2013, pp. 30–39.
[38] V. Kůrková, M. Sanguineti, Complexity of shallow networks representing functions with large variations, in: S. Wermter, C. Weber, W. Duch, T. Honkela, P. Koprinkova-Hristova, S. Magg, G. Palm, A. Villa (Eds.), Proceedings of ICANN 2014 Conference on Artificial Neural Networks and Machine Learning, 8681, Springer, Heidelberg, 2014, pp. 331–338.
[39] A.R. Barron, Neural net approximation, in: K.S. Narendra (Ed.), Proceedings of 7th Yale Workshop on Adaptive and Learning Systems, Yale University Press, New Haven, 1992, pp. 69–72.
[40] V. Kůrková, Dimension-independent rates of approximation by neural networks, in: K. Warwick, M. Kárný (Eds.), Computer-Intensive Methods in Control and Signal Processing, Birkhäuser, Boston, MA, 1997, pp. 261–270.
[41] V. Kůrková, P. Savický, K. Hlaváčková, Representations and rates of approximation of real-valued Boolean functions by neural networks, Neural Netw. 11 (1998) 651–659.
[42] Y. Mansour, Learning Boolean functions via the fourier transform, in: V. Roychowdhury, K. Siu, A. Orlitsky (Eds.), Theoretical Advances in Neural Computation and Learning, Springer, New York, 1994, pp. 391–424.
[43] H. Chernoff, A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, Ann. Math. Stat. 23 (1952) 493–507.
[44] T. Hagerup, C. Rüb, A guided tour of Chernoff bounds, Inf. Process. Lett. 33 (1990) 305–308.
[45] M. Anthony, Neural networks and Boolean functions, in: Y. Crama, P. Hammer (Eds.), Boolean Models and Methods in Mathematics, Computer Science, and Engineering, Vol. II of Encyclopedia of Mathematics and its Applications, vol. 134, Cambridge University Press, New York, 2010, pp. 554–576.
[46] L. Schläfli, Theorie der vielfachen Kontinuität, Zürcher & Furrer, Zürich, 1901.
[47] R. Winder, Single stage threshold logic, in: Proceedings of 2nd Annual Symposium on Switching Circuit Theory and Logical Design, 1961, pp. 321–332.
[48] V. Roychowdhury, K.-Y. Siu, A. Orlitsky, Neural models and spectral methods, in: V. Roychowdhury, K. Siu, A. Orlitsky (Eds.), Theoretical Advances in Neural Computation and Learning, Springer, New York, 1994, pp. 3–36.
[49] D.E. Knuth, Big omicron and big omega and big theta, SIGACT News 8 (1976) 18–24.
[50] T. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, IEEE Trans. Electron. Comput. 14 (1965) 326–334.
[51] R. Rojas, Neural Networks: A Systematic Introduction, Springer, New York, 1996.
[52] R. Winder, Threshold logic, Doctoral Dissertation, Mathematics Department, Princeton University, 1962.
[53] J. Flum, M. Grohe, Parameterized Complexity Theory, Springer, Berlin, Heidelberg, 2006.

Věra Kůrková received PhD in mathematics from Charles University Prague and DrSc (prof.) in theoretical computer science from Academy of Sciences of the Czech Republic. Since 1990 she works as a scientist in the Institute of Computer Science, Prague, in 2002–2009 as the Head of the Department of Theoretical Computer Science. She published many journal papers and book chapters on mathematical theory of neurocomputing and learning and on nonlinear approximation theory. She is a member of the editorial boards of Neural Networks and Neural Processing Letters and in past she served as an associate editor of IEEE Transactions on Neural Networks. She is a member of the Board of European Neural Network Society (ENNS). She was a general chair of the conferences ICANN 2008 and ICANNGA 2001.

Marcello Sanguineti received the "Laurea" (MSc) degree cum laude in Electronic Engineering and the PhD degree in electronic engineering and computer science from the University of Genova, Italy, where he is currently an Associate Professor of Operations Research. He coauthored more than 200 research papers in archival journals, book chapters, and international conference proceedings. He was a member of the Program Committees of several conferences, Chair of the Organizing Committee of the International Conference ICNPAA 2008, and coordinated several national and international research projects on approximate solution of optimization problems. He is a member of the editorial board of Neurocomputing. From 2006 to 2012 he was an Associate Editor of the IEEE Transaction on Neural Networks. His main research interests are infinite-dimensional programming, machine learning, network and team optimization, neural networks for optimization.