S. Krajči (ed.): ITAT 2018 Proceedings, pp. 86–91
CEUR Workshop Proceedings Vol. 2203, ISSN 1613-0073, © 2018 Vera Kůrková and Marcello Sanguineti

Probabilistic Bounds on Complexity of Networks Computing Binary Classification Tasks

Věra Kůrková[1] Marcello Sanguineti[2]

[1] Institute of Computer Science, Czech Academy of Sciences, Prague, Czech Republic
vera@cs.cas.cz,
WWW home page: http://www.cs.cas.cz/∼vera
[2] DIBRIS, University of Genoa, Genoa, Italy
marcello.sanguineti@unige.it
WWW home page: http://www.dist.unige.it/msanguineti/

Abstract: Complexity of feedforward networks computing binary classification tasks is investigated. To deal with unmanageably large number of these tasks on domains of even moderate sizes, a probabilistic model characterizing relevance of the classification tasks is introduced. Approximate measures of sparsity of networks computing randomly chosen functions are studied in terms of variational norms tailored to dictionaries of computational units. Probabilistic lower bounds on these norms are derived using the Chernoff-Hoeffding Bound on sums of independent random variables, which need not be identically distributed. Consequences of the probabilistic results on the choice of dictionaries of computational units are discussed.

1 Introduction

It has long been known that one-hidden-layer (shallow) networks with computational units of many common types can exactly compute any function on a finite domain [8]. In particular, they can perform any binary-classification task. Proofs of theorems on the universal approximation and representation properties of feedforward networks guarantee their power to express wide classes of functions, but do not deal with the efficiency of such representations. Typically, such arguments assume that the number of units is unbounded or is as large as the size of the domain of functions to be computed. For large domains, implementations of such networks might not be feasible.

A proper choice of a network architecture and a type of its units can, in some cases, considerably reduce network complexity. For example, a classification of points in the d-dimensional Boolean cube $\{0,1\}^d$ according to the parity of the numbers of 1's cannot be computed by a Gaussian SVM network with less than 2^{d-1} support vectors [3] (i.e., it cannot be computed by a shallow network with less than 2^{d-1} Gaussian SVM units). On the other hand, it is easy to show that the parity function (as well as any generalized parity, the set of which forms the Fourier basis) can be computed by a shallow network with merely $d+1$ Heaviside perceptrons [12].

The basic measure of sparsity of a network with a single linear output is the number of its nonzero output weights. The number of nonzero entries of a vector in $\mathbb{R}^n$ is called "l_0-pseudonorm". The quotation marks are used as l_0 is not homogenous and its "unit ball" is unbounded and nonconvex. Thus, minimization of the number of nonzero entries of an output-weight vector is a difficult nonconvex optimization task. Minimization of "l_0-pseudonorm" has been studied in signal processing, where it was shown that in some cases it is NP-hard [20].

A good approximation of convexification of "l_0-pseudonorm" is the l_1-norm [17]. In neurocomputing, l_1-norm has been used as a stabilizer in weight-decay regularization techniques [7]. In statistical learning theory, it l_1-norm plays an important role in LASSO regularization [19].

Networks with large l_1-norms of output-weight vectors have either large numbers of units or some of the weights are large. Both are not desirable: implementation of networks with large numbers of units might not be feasible and large output weights might lead to instability of computation. The minimum of the l_1-norms of output-weight vectors of all networks computing a given function is bounded from below by the variational norm tailored to a type of network units, which is a critical factor in estimates of upper bounds on network complexity [10, 11].

To identify and explain design of networks capable of efficient classifications, one has to focus on suitable classes of tasks. Even on a domain of a moderate size, there exists an enormous number of functions representing multi-class or binary classifications. For example, when the size of a domain is equal to 80, then the number of classifications into 10 classes is 10^{80} and when its size is 267, then the number of binary classification tasks is 2^{267}. These numbers are larger than the estimated number 10^{78} of atoms in the observable universe (see, e.g., [15]). Obviously, most classification tasks on such domains are not likely to be relevant for neurocomputing, as they do not model any task of practical interest.

In this paper, we investigate how to choose dictionaries of network units such that binary classification tasks can be efficiently solved. We assume that elements of a finite domain in $\mathbb{R}^d$ represent vectors of features, measurements, or observations for which some prior knowledge is available about probabilities that a presence of each of

these features implies the property described by one of the classes. For example, when vectors in the domain represent ordered sets of medical symptoms, certain values of some of these symptoms might indicate a high probability of some diagnosis, while others might indicate a low probability or be irrelevant.

For sets of classification tasks endowed with product probability distributions, we explore suitability of dictionaries of computational units in terms of values of variational norms tailored to the dictionaries. We analyze consequences of the concentration of measure phenomena which imply that with increasing sizes of function domains, correlations between network units and functions tend to concentrate around their mean or median values. We derive lower bounds on variational norms of functions to be computed and on l_1-norms of output-weight vectors of networks computing these functions. To obtain the lower bounds, we apply the Chernoff-Hoeffding Bound [5, Theorem 1.11] on sums of independent random variables not necessarily identically distributed. We show that when a priori knowledge of classification tasks is limited, then sparsity can only be achieved with large sizes of dictionaries. On the other hand, when such knowledge is biased, then there exist functions with which most functions on a large domain are highly correlated. If some of these functions is close to an element of a dictionary, then most functions can be well approximated by sparse networks with units from the dictionary.

The paper is organized as follows. In Section 2, we introduce basic concepts on feedforward networks, dictionaries of computational units, and approximate measures of network sparsity. In Section 3, we propose a probabilistic model of classification tasks and analyze properties of approximate measures of sparsity using the Chernoff-Hoeffding Bound. In Section 4, we derive estimates of probability distributions of values of variational norms and analyze consequences of these estimates for choice of dictionaries suitable for tasks modeled by the given probabilities. Section 5 is a brief discussion.

2 Approximate measures of network sparsity

We investigate computation of classification tasks represented by binary-valued functions on finite domains $X \subset \mathbb{R}^d$. We denote by

$$\mathscr{B}(X) := \{f \,|\, f : X \to \{-1,1\}\}$$

the *set of all functions on X with values in* $\{-1,1\}$ and by

$$\mathscr{F}(X) := \{f \,|\, f : X \to \mathbb{R}\}$$

the *set of all real-valued functions on X*.

When $\mathrm{card}X = m$ and $X = \{x_1,\ldots,x_m\}$ is a linear ordering of X, then the mapping $\iota : \mathscr{F}(X) \to \mathbb{R}^m$ defined as $\iota(f) := (f(x_1),\ldots,f(x_m))$ is an isomorphism. So, on $\mathscr{F}(X)$ we have the Euclidean inner product defined as

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

and the Euclidean norm $\|f\| := \sqrt{\langle f,f\rangle}$. We consider binary-valued functions with the range $\{-1,1\}$ instead of $\{0,1\}$ as all functions in $\mathscr{B}(X)$ have norms equal to $\sqrt{\mathrm{card}X}$.

A *feedforward network with a single linear output* can compute input-output functions from the set

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

where G, called a *dictionary*, is a parameterized family of functions. In networks with one hidden layer (called *shallow networks*), G is formed by functions computable by a given type of computational units, whereas in networks with several hidden layers (called *deep networks*), it is formed by combinations and compositions of functions representing units from lower layers (see, e.g., [2, 16]).

Formally, the number of hidden units in a shallow network or in the last hidden layer of a deep one can be described as the *number of nonzero entries* of the vector of output weights of the network. In applied mathematics, the number of nonzero entries of a vector $w \in \mathbb{R}^n$, denoted $\|w\|_0$, is called *"l_0-pseudonorm"* as it satisfies the equation

$$\|w\|_0 = \sum_{i=1}^{n} w_i^0.$$

The quotation marks are used because $\|w\|_0$ is neither a norm nor a pseudonorm. Minimization of "l_0-pseudonorm" is a difficult non convex problem as l_0 lacks the homogeneity property of a norm and its "unit ball" is not convex.

Instead of the nonconvex l_0-functional, its approximation by the *l_1-norm*

$$\|w\|_1 = \sum_{i=1}^{n} |w_i|$$

have been used as a stabilizer in weight-decay regularization methods [7]. Some insight into efficiency of computation of a function f by networks with units from a dictionary G can be obtained from investigation of the minima of l_1-norms of all vectors from the set

$$W_f(G) = \{w = (w_1,\ldots,w_n) \,|\, f = \sum_{i=1}^{k} w_i g_i, g_i \in G, n \in \mathbb{N}\}.$$

Minima of l_1-norms of elements of $W_f(G)$ are bounded from below by a norm of f tailored to a dictionary G called *G-variation*. It is defined for a bounded subset G of a normed linear space $(\mathscr{X}, \|.\|)$ as

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

where $-G := \{-g \mid g \in G\}$, $\mathrm{cl}_{\mathscr{X}}$ denotes the closure with respect to the topology induced by the norm $\|\cdot\|_{\mathscr{X}}$, and conv denotes the convex hull. Variation with respect to Heaviside perceptrons (called *variation with respect to half-spaces*) was introduced in [1] and extended to general dictionaries in [9].

It is easy to check (see [13]) that for a finite dictionary G and any f, such that the set $W_f(G)$ non empty, G-variation of f is equal to the minimum of l_1-norms of output-weight vectors of shallow networks with units from G, which compute f, i.e.,

$$\|f\|_G = \min\left\{\|w\|_1 \,\middle|\, w \in W_f(G)\right\}.$$

Thus lower bounds on minima of l_1-norms of output-weight vectors of networks computing a function f can be obtained from lower bounds on variational norms. Such bounds can be derived using the following theorem, which is a special case of a more general result [10] proven using Hahn-Banach theorem. By $G^\perp$ is denoted the *orthogonal complement of* G in the Hilbert space $\mathscr{F}(X)$.

Theorem 1. *Let X be a finite subset of $\mathbb{R}^d$ and G be a bounded subset of $\mathscr{F}(X)$. Then for every $f \in \mathscr{F}(X) \setminus G^\perp$,*

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

So functions which are nearly orthogonal to all elements of a dictionary G have large G-variations. On the other hand, if a function is correlated with some element of G, then it is close to this element and so can be well approximated by an element of G.

3 Probabilistic bounds

When we do not have any prior knowledge about a type of classification tasks to be computed, we have to assume that a network from the class has to be capable to compute any uniformly randomly chosen function on a given domain. Often in practical applications, most of the binary-valued functions o a given domain are not likely to represent tasks of interest. In such cases some knowledge is available that can be expressed in terms of a discrete probability measure on the set of all functions on X.

For a finite domain $X = \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$, a function f in $\mathscr{B}(X)$ can be represented as a vector $(f(x_1), \ldots, f(x_m)) \in \{-1, 1\}^m \subset \mathbb{R}^m$. We assume that for each $x_i \in X$, there exists a known probability $p_i \in [0,1]$ that $f(x_i) = 1$. For $p = (p_1, \ldots, p_m)$, we denote by

$$\rho_p : \mathscr{B}(X) \to [0,1]$$

the *product probability* defined for every $f \in \mathscr{B}(X)$ as

$$\rho_p(f) := \prod_{i=1}^m \rho_{p,i}(f), \tag{1}$$

where $\rho_{p,i}(f) := p_i$ if $f(x_i) = 1$ and $\rho_{p,i}(f) := 1 - p_i$ if $f(x_i) = -1$. It is easy to verify that ρ_p is a probability measure on $\mathscr{B}(X)$.

When $\mathrm{card} X$ is large, the set $\mathscr{F}(X)$ is isometric to a high-dimensional Euclidean space and $\mathscr{B}(X)$ to a high-dimensional Hamming cube. In high-dimensional spaces and cubes various concentration of measure phenomena occur [14]. We apply the Chernoff-Hoeffding Bound on sums of independent random variables, which do not need to be identically distributed [5, Theorem 1.11] to obtain estimates of distributions of inner products of any fixed function $h \in \mathscr{B}(X)$ with functions randomly chosen from $\mathscr{B}(X)$ with probability ρ_p.

Theorem 2 (Chernoff-Hoeffding Bound)**.** *Let m be a positive integer, $Y_1, \ldots, Y_m$ independent random variables with values in real intervals of lengths $c_1, \ldots, c_m$, respectively, $\varepsilon > 0$, and $Y := \sum_{i=1}^m Y_i$. Then*

$$\Pr\left(|Y - E(Y)| \geq \varepsilon\right) \leq e^{-\frac{2\varepsilon^2}{\sum_{i=1}^m c_i^2}}.$$

For a function $h \in \mathscr{B}(X)$ and $p = (p_1, \ldots, p_m)$, where $p_i \in [0,1]$, we denote by

$$\mu(h,p) := E_p\big(\langle h, f\rangle \,|\, f \in \mathscr{B}(X)\big)$$

the *mean value of inner products* of h with f randomly chosen from $\mathscr{B}(X)$ with probability ρ_p, and by $h^o := \frac{h}{\|h\|}$ its normalization. The next theorem estimates the distribution of these inner products.

Theorem 3. *Let $X = \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$, $p = (p_1, \ldots, p_m)$ be such that $p_i \in [0,1]$, $i = 1, \ldots, m$, and $h \in \mathscr{B}(X)$. Then the inner product of h with f randomly chosen from $\mathscr{B}(X)$ with a probability $\rho_p(f)$ satisfies for every $\lambda > 0$*

i) $$\Pr\Big(|\langle f, h\rangle - \mu(h,p)| > m\lambda\Big) \leq e^{-\frac{m\lambda^2}{2}};$$

ii) $$\Pr\Big(|\langle f^\circ, h^\circ\rangle - \frac{\mu(h,p)}{m}| > \lambda\Big) \leq e^{-\frac{m\lambda^2}{2}}.$$

Proof. Let $F_h : \mathscr{B}(X) \to \mathscr{B}(X)$ be an operator composed of sign-flips mapping h to the constant function equal to 1, i.e., $F_h(h)(x_i) = 1$ for all $i = 1, \ldots, m$ and for all $f \in \mathscr{F}(X)$ and all $i = 1, \ldots, m$, $F_h(f)(x_i) = f(x_i)$ if $h(x_i) = 1$ and $F_h(f)(x_i) = -f(x_i)$ if $h(x_i) = -1$. Let $p(h) = (p(h)_1, \ldots, p(h)_m)$ be defined as $p(h)_i = p_i$ if $h(x_i) = 1$ and $p(h)_i = 1 - p_i$ if $h(x_i) = -1$. The inverse operator F_h^{-1} maps the random variable $F_h(f) \in \mathscr{B}(X)$ such that

$$\Pr\left(F_h(f)(x_i) = 1\right) = p(h)_i$$

to the random variable $f \in \mathscr{B}(X)$ such that

$$\Pr(f(x_i)=1)=p_i.$$

Since the inner product is invariant under sign flipping, for every $f\in\mathscr{B}(X)$ we have $\langle f,h\rangle=\langle F_h(f),(1,\dots,1)\rangle=\sum_{i=1}^m F_h(f)(x_i)$. Thus the mean value of the sum of random variables $\sum_{i=1}^m F_h(f)(x_i)$ is $\mu(h,p)$. Applying to this sum the Chernoff-Hoeffding Bound stated in Theorem 2 with $c_1=\cdots=c_m=2$ and $\varepsilon=m\lambda$, we get

$$\Pr\Big(|\sum_{i=1}^m F_h(f)(x_i)-\mu(h,p)|>m\lambda\Big)\le e^{-\frac{m\lambda^2}{2}}.$$

Hence

$$\Pr\Big(|\langle f,h\rangle-\mu(h,p)|>m\lambda\Big)\le e^{-\frac{m\lambda^2}{2}},$$

which proves i).

ii) follows from i) as all functions in $\mathscr{B}(X)$ have norms equal to $\sqrt{m}$. □

Theorem 3 shows that when the domain X is large, most inner products of any given function with functions randomly chosen from $\mathscr{B}(X)$ with a probability ρ_p are concentrated around their mean values. For example, setting $\lambda=m^{-1/4}$, we get $e^{-\frac{m\lambda^2}{2}}=e^{-\frac{m^{-1/2}}{2}}$ which is decreasing exponentially fast with increasing size m of the domain.

4 Dictionaries for efficient classification

Theorem 3 implies that when a dictionary G contains a function h, for which the mean value $\mu(h,p)$ is large, then most functions randomly chosen with respect to the probability distribution ρ_p are correlated with h. Thus most classification tasks characterized by ρ_p can be well approximated by a network with just one element h. A dictionary G is also suitable for a given task when such function h can be well approximated by a small network with units from G.

It is easy to calculate the mean value $\mu(h,p)$ of inner products of a fixed function h from $\mathscr{B}(X)$ with randomly chosen functions from $\mathscr{B}(X)$ with respect to the probability ρ_p.

Proposition 4. *Let $h\in\mathscr{B}(X)$ and $p=(p_1\dots,p_m)$, where $p_i\in[0,1]$ for each $i=1,\dots,m$. Then for a function f randomly chosen in $\mathscr{B}(X)$ according to ρ_p, the mean value of $\langle f,h\rangle$ satisfies*

$$\mu(h,p)=\sum_{i\in I_h}(2p_i-1)+\sum_{i\in J_h}(1-2p_i),$$

where $I_h=\{i\in\{1,\dots,m\}\,|\,h(x_i)=1\}$ and $J_h=\{i\in\{1,\dots,m\}\,|\,h(x_i)=-1\}$.

By Theorem 1, variation with respect to a dictionary of a function is large when the function is nearly orthogonal to all elements of the dictionary. For $G:=\{g_1,\dots,g_k\}$, we define

$$\mu_G(p):=\max_{g_i,\dots,g_k}|\mu(g_i,p)|.$$

The next theorem estimates probability distributions of variational norms in dependence on the size of a dictionary.

Theorem 5. *Let $X=\{x_1,\dots,x_m\}\subset\mathbb{R}^d$, $G=\{g_1,\dots,g_k\}\subset\mathscr{B}(X)$, and $p=(p_1,\dots,p_m)$ such that $p_i\in[0,1]$, $i=1,\dots,m$. Then for every $f\in\mathscr{B}(X)$ randomly chosen according to ρ_p and every $\lambda>0$*

$$\Pr\left(\|f\|_G\ge\frac{m}{\mu_G(p)+m\lambda}\right)>1-ke^{-\frac{m\lambda^2}{2}}.$$

Proof. By Theorem 3 (i), we get

$$\Pr\Big(|\langle f,h\rangle-\mu(h,p)|>m\lambda\ \ \forall h\in G\Big)\le ke^{-\frac{m\lambda^2}{2}}.$$

Hence,

$$\Pr\Big(|\langle f,h\rangle-\mu(h,p)|\le m\lambda\ \ \forall h\in G\Big)>1-ke^{-\frac{m\lambda^2}{2}}.$$

As $|\langle f,h\rangle-\mu(h,p)|\le m\lambda$ implies $|\langle f,h\rangle|\le\mu(h,p)+m\lambda$, we get

$$\Pr\Big(|\langle f,h\rangle|\le\mu(h,p)+m\lambda\ \ \forall h\in G\Big)>1-ke^{-\frac{m\lambda^2}{2}}.$$

So by Theorem 1

$$\Pr\left(\|f\|_G\ge\frac{m}{\mu(h,p)+m\lambda}\ \ \forall h\in G\right)>1-ke^{-\frac{m\lambda^2}{2}}.$$

Since by the definition, for every $h\in G$ one has $\mu_G(p)\ge\mu(h,p)$, we obtain

$$\frac{m}{\mu_G(p)+m\lambda}\le\frac{m}{\mu(h,p)+m\lambda}$$

and so

$$\Pr\left(\|f\|_G\ge\frac{m}{\mu_G(p)+m\lambda}\right)>1-ke^{-\frac{m\lambda^2}{2}}.$$

□

Theorem 5 shows that when for all computational units h in a dictionary G, the mean values $\mu(h,p)$ are small, then for large m almost all functions randomly chosen according to ρ_p are nearly orthogonal to all elements of the dictionary G. For example, setting $\lambda=m^{-1/4}$, we get a probability greater than $1-ke^{-\frac{m^{1/2}}{2}}$ that a randomly chosen function has G-variation greater or equal to $\frac{m}{\mu_G(p)+m^{3/4}}$.

Thus when for large m, $\frac{\mu_G(p)}{m}$ is small, G-variation of most functions is large unless the size k of a dictionary G outweighs the factor $e^{-\frac{m\lambda^2}{2}}$.

Function with large G-variations cannot be computed by networks that have both the number of hidden units and all absolute vales of output weights small.

Corollary 1. *Let $X = \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$, $G = \{g_1, \ldots, g_k\} \subset \mathscr{B}(X)$, and $p = (p_1, \ldots, p_m)$ such that $p_i \in [0,1]$, $i = 1, \ldots, m$. Then for every $f \in \mathscr{B}(X)$ randomly chosen according to ρ_p, and every $\lambda > 0$,*

$$\Pr\left(\min\left\{\|w\|_1 \,|\, w \in W_f(G)\right\} \geq \frac{m}{\mu_G(p) + m\lambda}\right) > 1 - k\, e^{-\frac{m\lambda^2}{2}}.$$

Corollary 1 implies that computation of most classification tasks randomly chosen from $\mathscr{B}(X)$ with the product probability ρ_p either requires to perform an ill-conditioned task by a moderate network or a well-conditioned task by a large network.

In particular, for the uniform distribution $p_i = 1/2$ for all $i = 1, \ldots, m$, for every $h \in \mathscr{B}(X)$ the mean value $\mu(h, p)$ is zero. Thus for any dictionary $G \subset \mathscr{B}(X)$, almost all functions uniformly randomly chosen from $\mathscr{B}(X)$ are nearly orthogonal to all elements of the dictionary. So we get the following two corollaries.

Corollary 2. *Let $X = \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$ and $f \in \mathscr{B}(X)$ be uniformly randomly chosen. Then for every $h \in \mathscr{B}(X)$ and every $\lambda > 0$*

$$\Pr\left(|\langle f, h\rangle| > m\lambda\right) \leq e^{-\frac{m\lambda^2}{2}}.$$

Corollary 3. *Let $X = \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$ and $G = \{g_1, \ldots, g_k\} \subset \mathscr{B}(X)$. Then for every $f \in \mathscr{B}(X)$ uniformly randomly chosen and every $\lambda > 0$*

$$\Pr\left(\|f\|_G \geq \frac{1}{\lambda}\right) \geq 1 - k\, e^{-\frac{m\lambda^2}{2}}.$$

When we do not have any a priori knowledge about the task, we have to assume that the probability on $\mathscr{B}(X)$ is uniform. Corollary 3 shows that unless a dictionary G is sufficiently large to outweigh the factor $e^{-\frac{m\lambda^2}{2}}$, most functions randomly chosen in $\mathscr{B}(X)$ according to ρ_p have G-variations greater or equal to $1/\lambda$. So for small λ and sufficiently large m, most such functions cannot be computed by linear combinations of small numbers of elements of G with small coefficients. Similar situation occurs when probabilities are nearly uniform.

Many common dictionaries used in neurocomputing are relatively small with respect to the factor $e^{-\frac{m^{1/2}}{2}}$. For example, the size of the *dictionary of signum perceptrons* $P_d(X)$ on a set X of m points in $\mathbb{R}^d$ is well-known since the work of Schläfli [18]. He estimated the number of linearly separated dichotomies of m points in $\mathbb{R}^d$. His upper bound states that for every $X \subset \mathbb{R}^d$ such that $\operatorname{card} X = m$,

$$\operatorname{card} P_d(X) \leq 2\sum_{l=1}^{d}\binom{m-1}{l} \leq 2\frac{m^d}{d!}. \qquad (2)$$

(see, e.g., [4]). The set $P_d(X)$ forms only a small fraction of the set of all functions in the set $\mathscr{B}(X)$, whose cardinality is equal to 2^m. Also other dictionaries of $\{-1,1\}$-valued functions generated by dichotomies of m points in $\mathbb{R}^d$ defined by nonlinear separating surfaces (such as hyperspheres or hypercones) are relatively small (see [4, Table I]).

5 Discussion

As the number of binary-valued functions modeling classification tasks grows exponentially with the size of their domains, we proposed to model relevance of such tasks for a give application area by a probabilistic model. For sets of classification tasks endowed with product probability distributions, we investigated complexity of networks computing these tasks. We explored network complexity in terms of approximate measures of sparsity formalized by l_1 and variational norms. For functions on large domains, we analyzed implications of the concentration of measure phenomena for correlations between network units and randomly chosen functions.

We focused on classification tasks characterized by product probabilities. To derive estimates of complexity of networks computing randomly chosen functions we used the Chernoff-Hoeffding Bound on sums of independent random variables. An extension of our analysis to tasks characterized by more general probability distributions is a subject of our future work. To obtain estimates for more general probability distributions, we plan to apply versions of the Chernoff-Hoeffding Bound stated in [6], which hold in situations when random variables are not independent .

Acknowledgments. V. K. was partially supported by the Czech Grant Foundation grant GA18-23827S and institutional support of the Institute of Computer Science RVO 67985807. M. S. was partially supported by a FFABR grant of the Italian Ministry of Education, University and Research (MIUR). He is Research Associate at INM (Institute for Marine Enginering) of CNR (National Research Council of Italy) under the Project PDGP 2018/20 DIT.AD016.001 "Technologies for Smart Communities" and he is a member of GNAMPA-INdAM (Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni - Instituto Nazionale di Alta Matematica).

References

[1] A. R. Barron. Neural net approximation. In K. S. Narendra, editor, *Proc. 7th Yale Workshop on Adaptive and Learning Systems*, pages 69–72. Yale University Press, 1992.

[2] Y. Bengio and A. Courville. Deep learning of representations. In *Handbook of Neural Information Processing*. M. Bianchini and M. Maggini and L. Jain, Berlin, Heideleberg, 2013.

[3] Y. Bengio, O. Delalleau, and N. Le Roux. The curse of highly variable functions for local kernel machines. In *Advances in Neural Information Processing Systems*, volume 18, pages 107–114. MIT Press, 2006.

[4] T. Cover. Geometrical and statistical properties of systems of linear inequalities with applictions in pattern recognition. *IEEE Trans. on Electronic Computers*, 14:326–334, 1965.

[5] B. Doerr. Analyzing randomized search heuristics: Tools from probability theory. In *Theory of Randomized Seach Heuristics - Foundations and Recent Developments*, chapter 1, pages 1–20. World Scientific Publishing, 2011.

[6] D. P. Dubhashi and A. Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, New York, 2009.

[7] T. L. Fine. *Feedforward Neural Network Methodology*. Springer, Berlin Heidelberg, 1999.

[8] Y. Ito. Finite mapping by neural networks and truth functions. *Mathematical Scientist*, 17:69–77, 1992.

[9] V. Kůrková. Dimension-independent rates of approximation by neural networks. In K. Warwick and M. Kárný, editors, *Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality*, pages 261–270. Birkhäuser, Boston, MA, 1997.

[10] V. Kůrková. Complexity estimates based on integral transforms induced by computational units. *Neural Networks*, 33:160–167, 2012.

[11] V. Kůrková and M. Sanguineti. Approximate minimization of the regularized expected error over kernel models. *Mathematics of Operations Research*, 33:747–756, 2008.

[12] V. Kůrková and M. Sanguineti. Model complexities of shallow networks representing highly varying functions. *Neurocomputing*, 171:598–604, 2016.

[13] V. Kůrková, P. Savický, and K. Hlaváčková. Representations and rates of approximation of real-valued Boolean functions by neural networks. *Neural Networks*, 11:651–659, 1998.

[14] M. Ledoux. *The Concentration of Measure Phenomenon*. AMS, Providence, 2001.

[15] H.W. Lin, M. Tegmark, and D. Rolnick. Why does deep and cheap learning work so well? *J. of Statistical Physics*, 168:1223–1247, 2017.

[16] H. N. Mhaskar and T. Poggio. Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, 14:829–848, 2016.

[17] Y. Plan and R. Vershynin. One-bit compressed sensing by linear programming. *Communications in Pure and Applied Mathematics*, 66:1275–1297, 2013.

[18] L. Schläfli. *Theorie der Vielfachen Kontinuität*. Zürcher & Furrer, Zürich, 1901.

[19] T.Hastie, R.Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, London, 2015.

[20] A.M. Tillmann. On the computational intractability of exact and approximate dictionary learning. *IEEE Signal Processing Letters*, 22:45–49, 2015.