Classification by Sparse Neural Networks

Věra Kůrková and Marcello Sanguineti

***Abstract*— The choice of dictionaries of computational units suitable for efficient computation of binary classification tasks is investigated. To deal with exponentially growing sets of tasks with increasingly large domains, a probabilistic model is introduced. The relevance of tasks for a given application area is modeled by a product probability distribution on the set of all binary-valued functions. Approximate measures of network sparsity are studied in terms of variational norms tailored to dictionaries of computational units. Bounds on these norms are proven using the Chernoff–Hoeffding bound on sums of independent random variables that need not be identically distributed. Consequences of the probabilistic results for the choice of dictionaries of computational units are derived. It is shown that when *a priori* knowledge of a type of classification tasks is limited, then the sparsity may be achieved only at the expense of large sizes of dictionaries.**

***Index Terms*— Binary classification, Chernoff–Hoeffding bound, dictionaries of computational units, feedforward networks, measures of sparsity.**

I. Introduction

IT HAS long been known that one-hidden-layer (shallow) networks with computational units of many common types can approximate up to any desired accuracy every reasonable function on a compact domain (see [1]–[4] and the references therein) and that such networks can exactly compute any function on a finite domain [5]. In particular, they can perform any binary classification task.

Theorems on universal approximation and representation properties of feedforward networks only guarantee the capability of expressing wide classes of functions, but do not deal with the complexity of the approximating networks. Proofs of these theorems assume that the number of network units is unbounded or, in the case of a finite domain, as large as the size of the domain. For large domains, implementations of such networks might not be feasible.

Manuscript received December 27, 2017; revised July 12, 2018; accepted November 27, 2018. The work of V. Kůrková was supported in part by Czech Grant Foundation under Grant GA15-18108S and Grant GA18-23827S and in part by the Institute of Computer Science under Grant RVO 67985807. The work of M. Sanguineti was supported in part by the Italian Ministry of Education, University and Research through the FFABR Grant, in part by the National Research Council of Italy through the PDGP 2018/2020 ex-ISSIA, now INM, DIT.AD016.001 "Technologies for Smart Communities" and in part by the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni–Instituto Nazionale di Alta Matematica. *(Corresponding author: Marcello Sanguineti.)*

V. Kůrková is with the Institute of Computer Science, Czech Academy of Sciences, 182 07 Prague, Czech Republic (e-mail: vera@cs.cas.cz).

M. Sanguineti is with the Department of Computer Science, Bioengineering, Robotics, and Systems Engineering, University of Genoa, 16145 Genoa, Italy (e-mail: marcello.sanguineti@unige.it).

Digital Object Identifier 10.1109/TNNLS.2018.2888517

Proper choice of network architecture and type of units can, in some cases, considerably reduce network complexity, and thus enable the efficient computation of functions on large and high-dimensional domains. For example, 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 support vector machine (SVM) network with less than 2^{d-1} support vectors [6] (i.e., 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 only $d+1$ Heaviside perceptrons (see [7, p. 600], [8]).

In addition, one prefers to choose computational units allowing computation of given tasks by reasonably sparse networks. The basic measure of the sparsity of a shallow network with a single linear output is the number of units in the hidden layer, which can be studied in terms of the number of nonzero output weights. The number of nonzero entries of a vector in $\mathbb{R}^n$ is often called the "l_0-pseudonorm" although actually not a pseudonorm (as it is not homogenous and has a "unit ball" which is unbounded and nonconvex). Thus, the minimization of the number of network units (corresponding to nonzero entries of the network) is a difficult nonconvex optimization task. Minimization of "l_0-pseudonorm." has been studied in signal processing, where such minimization was shown to be NP-hard in some cases [9].

Among l_p-functionals with convex unit balls, l_1 is the closest one to "l_0-pseudonorm." There are other reasons for using l_1-norm. In neurocomputing, it has been used as a stabilizer in weight-decay regularization techniques [10]–[13] and the l_1-norm also plays a role of a stabilizer in LASSO regularization [14] and it has also been studied in compressed sensing [15]–[18]. For other approaches to regularization in learning see [19]–[21] and the references therein.

Networks with large l_1-norms of output-weight vectors have either large numbers of units or some of the weights are large. Either of these properties is undesirable: the implementation of networks with large numbers of units might not be feasible, while large output weights can 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 a variational norm tailored to the type of network units [7], and this variational norm is a critical factor in estimates of upper bounds on network complexity [8], [13], [22].

Even on domains of moderate sizes, there are an enormous number of functions representing multiclass or binary classifications. For example, when the size of a domain is equal to 80, 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 [23]). Obviously, most functions on large domains represent classifications, which are not likely to be relevant for neural network computation, as they do not model any task of practical interest.

In this paper, we propose a probabilistic model of relevance for binary classification tasks. We assume that there are known probabilities that specify, for each point in the domain and each class, the probability that the point belongs to the class. Points in a finite domain in $\mathbb{R}^d$ are typically vectors of features, for which some prior knowledge might be available about probabilities that the presence of such features leads to a property described by one of the classes. For example, when points in a domain represent vectors of some medical symptoms, certain values of these symptoms might indicate a high probability of some diagnosis.

To identify and explain the design of networks suitable for computing tasks characterized by probability distributions, we investigate the network simplicity achievable by using various dictionaries of computational units. We analyze network simplicity in terms of its approximate measures of sparsity and estimate the minima of l_1-norms of output-weight vectors using geometrical properties of variational norms tailored to the computational units. In order to describe the properties of networks suitable for efficient computation of tasks modeled by probability distributions, we study the distributions of variational norms. On large domains, various counterintuitive properties of high-dimensional geometry occur.

We analyze the consequences of the concentration of measure phenomena. Such phenomena 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 the variational norms of functions to be computed and on the l_1-norms of output-weight vectors of networks computing these functions. To obtain such lower bounds, we apply the Chernoff–Hoeffding bound [24, Th. 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 the sparsity can only be achieved with large sizes of dictionaries. On the other hand, when such given knowledge is biased, then there exist functions with which most randomly chosen classification tasks on a large domain are highly correlated. If such functions are close to some elements of a dictionary, then most tasks can be well approximated by sparse networks with units from such a biased dictionary.

A preliminary version of some results appears in a regional Czech-Slovak conference proceedings [25].

This paper is organized as follows. In Section II, we introduce basic concepts and notations. In Section III, we discuss various measures of network sparsity ("l_0-pseudonorm," l_1 norm, and the variational norms tailored to dictionaries) and we analyze their relationships. In Section IV, we introduce a probabilistic model of classification tasks and, by using the Chernoff–Hoeffding Bound, we derive geometrical properties of functions satisfying given probability constraints. In Section V, we derive estimates of probability distributions of values of variational norms and analyze their consequences for the choice of dictionaries suitable for tasks modeled by given probabilities. Section VI contains some conclusions. In Section VII, we discuss our results.

II. Preliminaries

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), G is formed by functions computable by a given type of computational units, whereas, in networks with several hidden layers (called deep), it is formed by combinations and compositions of functions representing units from lower layers (see [26], [27]).

We denote by

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

a set of functions computable by networks with at most n units in the last hidden layer.

Dictionaries are parameterized families of functions of the form

$$G_\phi(X, Y) := \{\phi(\cdot, y) : X \to \mathbb{R} | y \in Y\}$$

where $\phi : X \times Y \to \mathbb{R}$ is a function of two variables: an input vector $x \in X \subseteq \mathbb{R}^d$ and a parameter vector $y \in Y \subseteq \mathbb{R}^s$. When the set of parameters is the whole $\mathbb{R}^s$, we write shortly $G_\phi(X)$.

For a domain $X \subset \mathbb{R}^d$, we denote by

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

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

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

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

In practical applications, domains $X \subset \mathbb{R}^d$ are finite, but their sizes $\mathrm{card}X$ and/or input dimensions d can be quite large. It is easy to see that when $\mathrm{card}X = m$ and $X = \{x_1, \ldots, x_m\}$ is a linear ordering of X, then the mapping $\iota : \mathcal{F}(X) \to \mathbb{R}^m$ defined as $\iota(f) := (f(x_1), \ldots, f(x_m))$ is an isomorphism. Therefore, 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 $\|f\| := \sqrt{\langle f, f \rangle}$.

In theoretical analysis of approximation capabilities of neural networks, it has to be taken into account that the approximation error $\|f - \mathrm{span}_n G\|$ in any norm $\|.\|$ can be made arbitrarily large by multiplying f by a scalar. Indeed, for every $c > 0$, one has

$$\|cf - \mathrm{span}_n G\| = c\|f - \mathrm{span}_n G\|.$$

Thus, approximation errors have to be studied either in sets of normalized functions or in sets of functions of a given fixed norm. Thus, it is convenient to consider binary-valued functions with range $\{-1.1\}$ instead of $\{0, 1\}$. All functions in $\mathcal{B}(X)$ have norms equal to $\sqrt{\text{card}X}$.

III. Sparse Network Approximation

In this section, we investigate approximate measures of network sparsity.

Let

$$f = \sum_{i=1}^{m} w_i g_i \tag{1}$$

be a representation of a function $f : X \to \mathbb{R}$ as an input–output function of a shallow network with a single linear output and units from a dictionary G. On infinite compact domains or on $\mathbb{R}^d$, many dictionaries formed by popular computational units are linearly independent (see [28]). However, on finite domains, often representations of the form (1) are not unique. Linearly dependent dictionaries are called *overcomplete* and networks computing the same input–output functions are called *functionally equivalent*. Nonuniqueness of representations of the form (1) provides flexibility of representations, which can, in some cases allow an improvement of sparsity. Also, in signal processing, where representations of signals as linear combinations of their elements called atoms have been studied, overcomplete dictionaries are often advantageous [29], [30].

The basic measure of the sparsity of a network computing the function (1) is the number of nonzero output weights among $w_1, \ldots, w_m$. In applied mathematics, the number of nonzero entries of a vector $w \in \mathbb{R}^m$ is called *"l_0-pseudonorm"* and denoted $\|w\|_0$. The quotation marks are used because l_0 is neither a norm nor a pseudonorm. Although it satisfies the triangle inequality, it does not satisfy the homogeneity property of a norm $\|\lambda x\| = |\lambda|\|x\|$ for all λ. The quantity $\|w\|_0$ is always an integer, and moreover, the "unit ball" $\{w \in \mathbb{R}^n | \|w\|_0 \leq 1\}$ is nonconvex and unbounded. It is equal to the union of all 1-D subspaces of $\mathbb{R}^m$. For any $r > 0$, the ball of radius r is equal to $\text{span}_k \mathbb{R}^m$, where $k = \lfloor r \rfloor$.

Hence, searching for representations as input–output functions of networks with smallest "l_0-pseudonorms" of vectors of output weights is a nonconvex optimization task. In signal processing, sparse representations of as few atoms as possible have been investigated. It was shown that in some cases solving this nonconvex optimization problem is NP-hard [9]. Finding the sparsest solution to a general underdetermined system of equations is NP-hard [31].

"l_0-pseudonorm" can be approximated by l_p-functionals as

$$\lim_{p \to 0} \|w\|_p = \|w\|_0$$

where

$$\|w\|_p^p = \sum_{i=1}^{m} |w_i|^p.$$

Among l_p-functionals, the one with the smallest $p \in [0, \infty]$ having convex unit ball $\{w \in \mathbb{R}^n | \|w\|_p \leq 1\}$ is l_1 (see Fig. 1).

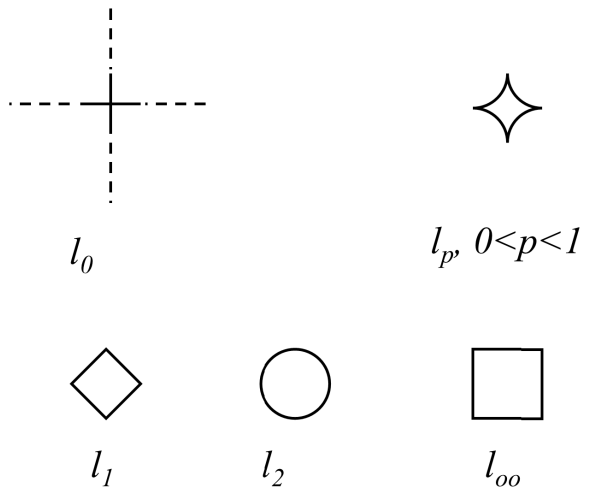

Fig. 1. Shapes of the balls in "l_0-pseudonorm" and in various l_p-norms, $0 < p \leq \infty$.

In neurocomputing, l_1-norm has been used as a stabilizer in weight-decay regularization [12]. Moreover, the optimization problem associated with the search of the representation with minimal l_1-norm is much easier to handle than the one related to "l_0-pseudonorm" [32], [33]. In some cases, a solution with the minimal l_1-norm is also the sparsest solution [34].

The most important property of l_1-norm for investigation of network complexity is its relationship to a norm tailored to a dictionary called *G-variation*. This norm plays a role of a critical factor in estimates of rates of approximation by networks with increasing "l_0-pseudonorms" of output weights and it minimizes l_1-norms of output-weight vectors over all expressions (1) of a given function f as an input–output function of a network with units from a dictionary G. Some insight into efficiency of computation of a function f by networks with units from a dictionary G can be obtained from investigation of minima of l_1-norms of all the vectors from the set

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

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

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

where $-G := \{-g | g \in G\}$, $\text{cl}_{\mathcal{X}}$ denotes the closure with respect to the topology induced by the norm $\|\cdot\|_{\mathcal{X}}$, and conv is the convex hull. Variation with respect to Heaviside perceptrons (called *variation with respect to half-spaces*) was introduced in [35] and extended to general dictionaries in [36]. G-variation is a generalization of the concepts of total variation and l_1-norm. For $d = 1$, variation with respect to half-spaces coincides with total variation up to a constant (see [35] and [37]). The concept of G-variation has been used as a tool for investigation of approximation and optimization by neural networks (see [38]–[43] and the references therein).

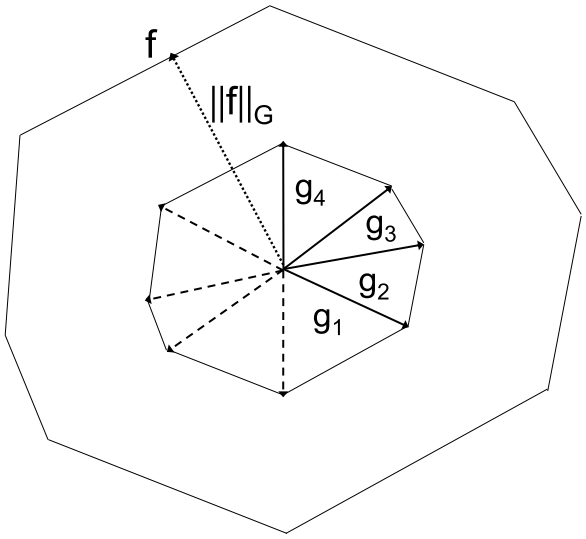

Fig. 2. Unit ball in G-variation norm, for $G = \{g_1, g_2, g_3, g_4\}$, and geometric construction to determine the G-variation norm of an element outside such a ball.

In Fig. 2, the concept of G-variation is illustrated by an example of a finite dictionary G with four elements.

As G-variation is a norm, it can be made arbitrarily large by multiplying a function by a scalar. Also, errors in an approximation of scalar multiples of a given function can be made arbitrarily large or small with proper choices of scalars. Indeed, for every $c > 0$

$$\|cf - \text{span}_n G\| = c\|f - \text{span}_n G\|.$$

Thus, both G-variation and errors in approximation by $\text{span}_n G$ have to be studied either for sets of normalized functions or for sets of functions of a given fixed norm. In this paper, we focus on $\{-1, 1\}$-valued functions on domains of size m, which have all norms equal to $\sqrt{m}$.

For finite dictionaries, the minimum of l_1-norms of output-weight vectors of shallow networks with units from G computing f is equal to $\|f\|_G$. The next proposition follows directly from the definition of G-variation (see [44]).

Proposition 3.1: Let G be a finite subset of $(\mathcal{X}, \|.\|)$ with $\text{card}G = k$. Then, for every $f \in \mathcal{X}$

$$\begin{aligned}\|f\|_G &= \min\left\{\|w\|_1 | w \in W_f(G)\right\} \\ &= \min\left\{\|w\|_1 \middle| f = \sum_{i=1}^{k} w_i g_i, w_i \in \mathbb{R}, g_i \in G\right\}.\end{aligned}$$

Thus, $\|f\|_G$ is equal to the smallest l_1-norm of a representation of a function f as a network with linear output and one hidden layer of units from G. In contrast to "l_0-pseudonorm," l_1-norm can be minimized by various weight-decay regularization algorithms.

Moreover, G-variation and, thus, also l_1-norms of output-weight vectors of all representations of a function f in the form (1) play roles of critical factors in upper bounds on rates of approximation of f by sparse networks with increasing "l_0-pseudonorms." This follows from the Maurey–Jones–Barron theorem [45]. Here, we state a special case of reformulation of this theorem in terms of G-variation from [22], [36]) for the finite dimensional Hilbert space $\mathcal{F}(X) = \{f : X \to \mathbb{R}\}$ which is isometric to $\mathbb{R}^{\text{card}X}$.

Theorem 3.2: Let $X \subset \mathbb{R}^d$ be finite, $\emptyset \neq G \subseteq \mathcal{F}(X)$, $s_G := \max_{g \in G} \|g\|$, and $f \in \mathcal{F}(X)$. Then, for every integer $n \geq 1$, there is a function $f_n \in \text{span}G$ such that $\|f_n\|_0 \leq n$ and

$$\|f - f_n\| \leq \frac{s_G \|f\|_G}{\sqrt{n}}.$$

On the other hand, when G-variation of a function is large, then by Proposition 3.1, any representation of f as an input–output function of a network with a linear output and units from G, must have large number of units or some of output weights must be large. This means that a dictionary G is not well chosen for computation of f. Such computation would require unmanageably large network or be unstable due to large output weights.

To derive lower bounds on G-variation, we employ the geometric properties proven in [22] and [44] via the Hahn–Banach theorem. By $G^\perp$ is denoted the *orthogonal complement of* G in the Hilbert space $\mathcal{F}(X)$.

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

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

Theorem 3.3 shows that functions which are "nearly orthogonal" to all elements of a dictionary G have large G-variations. By Proposition 3.1 computation of such functions requires large numbers of hidden units or large output weights. Therefore, for a given class of tasks to be computed, it is desirable to choose a dictionary in such a way that most functions from the class are correlated with some elements from the dictionary or with a linear combination of a small number of its elements.

IV. Probabilistic Bounds

In this section, we investigate the distribution of variational and l_1-norms of binary-valued functions randomly selected with respect to probability distributions modeling likelihood that functions represent classification tasks from a given application area. When domains are large, various concentration phenomena occur. We analyze their consequences for the choice of dictionaries.

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, many binary-valued functions 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 $X = \{x_1, \ldots, x_m\}$, a function f in $\mathcal{B}(X)$ 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 is a known probability $p_i \in [0, 1]$ of $f(x_i) = 1$. For $p =$

$(p_1, \ldots, p_m)$, we define $\rho_p : \mathcal{B}(X) \to [0, 1]$ such that for every $f \in \mathcal{B}(X)$

$$\rho_p(f) := \prod_{i=1}^{m} \rho_{p,i}(f) \tag{2}$$

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 product probability measure on $\mathcal{B}(X)$.

The set $\mathcal{F}(X)$ is isometric to the Euclidean space $\mathbb{R}^{\mathrm{card}X}$ and $\mathcal{B}(X)$ to the discrete cube $\{-1, 1\}^{\mathrm{card}X}$. When $\mathrm{card}X$ is large, various concentration of measure phenomena occur [46], [47]. To obtain estimates of distributions of inner products of any fixed function $h \in \mathcal{B}(X)$ with functions randomly chosen from $\mathcal{B}(X)$ with probability ρ_p, we use the Chernoff–Hoeffding bound on sums of independent random variables, which do not need to be identically distributed [24, Th. 1.11].

Theorem 4.1 (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 $h \in \mathcal{B}(X)$ and $p = (p_1, \ldots, p_m)$, where $p_i \in [0, 1]$, we denote by

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

the *mean value of the inner products* of h with f randomly chosen from $\mathcal{B}(X)$ with the probability ρ_p. The next theorem estimates the distribution of these inner products. For a function h, we denote $h^\circ := h/\|h\|$.

Theorem 4.2: 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 $h \in \mathcal{B}(X)$. Then, the inner product of h with f randomly chosen from $\mathcal{B}(X)$ with a probability $\rho_p(f)$ satisfies for every $\lambda > 0$

$$1)\ \Pr(|\langle f, h\rangle - \mu(h, p)| > m\lambda) \leq e^{-\frac{m\lambda^2}{2}}$$
$$2)\ \Pr(|\langle f^\circ, h^\circ\rangle - \frac{\mu(h, p)}{m}| > \lambda) \leq e^{-\frac{m\lambda^2}{2}}.$$

Proof: Let $F_h : \mathcal{B}(X) \to \mathcal{B}(X)$ 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 $i = 1, \ldots, m$. 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 $F_h(f) \in \mathcal{B}(X)$ such that

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

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

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

Since the inner product is invariant under sign flipping, for every $f \in \mathcal{B}(X)$, we have $\langle f, h\rangle = \langle F_h(f), (1, \ldots, 1)\rangle = \sum_{i=1}^{m} F_h(f)(x_i)$. Thus, the mean value of the sum of random variables $\sum_{i=1}^{m} F_h(f)(x_i)$ has the mean value $\mu(h, p)$. Applying to this sum the Chernoff–Hoeffding bound stated in Theorem 4.1 with $c_1 = \cdots = c_m = 2$ and $\varepsilon = m\lambda$, we get

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

Hence,

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

which proves 1). 2) follows from 1) as all functions in $\mathcal{B}(X)$ have norms equal to $\sqrt{m}$. □

Theorem 4.2 shows that when the domain X is large, most inner products of any given function with functions randomly chosen from $\mathcal{B}(X)$ with a probability ρ_p are concentrated around their mean value. For example, setting $\lambda = m^{-1/4}$, we get $e^{-(m\lambda^2/2)} = e^{-((m^{-1/2})/2)}$ which is decreasing exponentially fast with increasing size m of the domain.

When G is chosen, in such a way, that it contains a function h, for which the mean value $\mu(h, p)$ is large, then most randomly chosen functions are correlated with h and so can be well approximated by h. Such dictionary is a good choice for performing classification tasks described by the probability ρ_p. 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. In the next section, we analyze the case when the mean values $\mu(g, p)$ are small for all functions in G.

V. Suitability of Dictionaries

In this section, we analyze the properties of dictionaries suitable for efficient computing of classification tasks characterized by prior knowledge in the form of probability distributions.

Assuming that for $p = (p_1, \ldots p_m)$, a probability measure ρ_p on $\mathcal{B}(X)$ is given, we first calculate for any function $h \in \mathcal{B}(X)$, the mean value $\mu(h, p)$ of its inner products with functions randomly chosen with probability ρ_p.

Proposition 5.1: Let $h \in \mathcal{B}(X)$ and $p = (p_1 \ldots, p_m)$, where $p_i \in [0, 1]$ for each $i = 1, \ldots, m$. Then, for a function f randomly chosen in $\mathcal{B}(X)$ 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 $I_h = \{i \in \{1, \ldots, m\} | h(x_i) = 1\}$ and $J_h = \{i \in \{1, \ldots, m\} | h(x_i) = -1\}$.

Proof: 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 $h(x_i) = -1$. Then, $\mu(h, p) = \sum_{i=1}^{m} p(h)_i - \sum_{i=1}^{m} (1 - p(h)_i) = \sum_{i \in I_h} p_i - \sum_{i \in I_h} (1 - p_i) - \sum_{i \in J_h} p_i + \sum_{i \in J_h} (1 - p_i) - \sum_{i \in I_h} (2p_i - 1) + \sum_{i \in J_h} (1 - 2p_i) = \mu(h, p)$. □

For fixed $p = (p_1, \ldots, p_m)$, the quantity $\mu(h, p)$ varies as a function of h. The next proposition bounds its range of variation.

Proposition 5.2: Let $X := \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$ and $p := (p_1, \ldots, p_m)$ be such that $p_i \in [0, 1], i = 1, \ldots, m$. Let $h_p, \bar{h}_p \in \mathcal{B}(X)$ be defined as $h_p(x_i) := 1$ if $p_i \geq 1/2$,

$h_p(x_i) := -1$ if $p_i < 1/2$, and $\bar{h}_p(x_i) := -1$ if $p_i \geq 1/2$, $\bar{h}_p(x_i) := 1$ if $p_i < 1/2$. Then,

$$1)\mu(h_p, p) = \max\{\mu(h, p)|h \in \mathcal{B}(X)\}$$
$$2)\mu(\bar{h}_p, p) = \min\{\mu(h, p)|h \in \mathcal{B}(X)\}.$$

Proof: 1) Let us define $p(h)_i := p_i$ if $h(x_i) = 1$ and $p(h)_i := 1 - p_i$ if $h(x_i) = -1$. As for every $i = 1, \ldots, m$, we have $p(h_p)_i = \max\{p_i, 1 - p_i\} \geq p(h)_i$, the statement follows. 2) is proven analogously. □

By Theorem 3.3, variation with respect to a dictionary of a function is large when the function is nearly orthogonal to all elements of the dictionary. For $G := \{g_1, \ldots, g_k\}$, we denote

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

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

Theorem 5.3: Let $X = \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$, $G = \{g_1, \ldots, g_k\} \subset \mathcal{B}(X)$, and $p = (p_1, \ldots, p_m)$ such that $p_i \in [0, 1], i = 1, \ldots, m$. Then, for every $f \in \mathcal{B}(X)$ randomly chosen according to ρ_p and every $\lambda > 0$

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

Proof: By Theorem 4.2 (i), we get

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

Hence,

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

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

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

Therefore, by Theorem 3.3

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

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

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

and so

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

□

Theorem 5.3 shows that the more biased the sets of functions to be computed, the more chances to find relatively small dictionaries capable to compute or approximate them by reasonably sparse networks.

On the other hand, when for all computational units h in a dictionary G, the mean value $\mu(h, p)$ is 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 probability greater than $1 - ke^{-(m^{1/2}/2)}$ that a randomly chosen function has G-variation greater or equal to $(m/(\mu_G(p) + m^{3/4}))$. Thus, when for large m, $(\mu_G(p)/m)$ is small, G-variation of most functions is large unless the size of the dictionary k outweights the factor $e^{-(m\lambda^2/2)}$.

Functions with large G-variations cannot be computed by networks with a linear output unit which has both the number of elements of G and all absolute values of output weights small.

Corollary 5.4: Let $X = \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$, $G = \{g_1, \ldots, g_k\} \subset \mathcal{B}(X)$, and $p = (p_1, \ldots, p_m)$ such that $p_i \in [0, 1], i = 1, \ldots, m$. Then, for every $f \in \mathcal{B}(X)$ randomly chosen according to ρ_p, and every $\lambda > 0$

$$\Pr\left(\min\left\{\|w\|_1 | f = \sum_{i=1}^{k} w_i g_i\right\} \geq \frac{m}{\mu_G(p) + m\lambda}\right) > 1 - ke^{-\frac{m\lambda^2}{2}}.$$

Corollary 5.4 implies that a computation of most functions either requires to perform an ill-conditioned task by a moderate network or a well-conditioned task by a large network.

As h_p has among all elements of $\mathcal{B}(X)$ the largest mean value of its inner products with randomly chosen functions from $\mathcal{B}(X)$ with respect to ρ_p, for any dictionary G, we have $\mu_G \leq \mu(h_p, p)$. Thus, if $\mu(h_p)$ is small, then unless a dictionary is large enough to outweigh $e^{-(m^{1/2}/2)}$, almost all functions have large variations.

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

Corollary 5.5: Let $X = \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$ and $f \in \mathcal{B}(X)$ be uniformly randomly chosen. Then, for every $h \in \mathcal{B}(X)$ and every $\lambda > 0$

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

Corollary 5.6: Let $X = \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$ and $G = \{g_1, \ldots, g_k\} \subset \mathcal{B}(X)$. Then, for every $f \in \mathcal{B}(X)$ uniformly randomly chosen and every $\lambda > 0$

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

When we do not have any *a priori* knowledge about the task, we have to assume that the probability on $\mathcal{B}(X)$ is uniform. Corollary 5.6 shows that unless a dictionary G is sufficiently large to outweigh the factor $e^{-(m\lambda^2/2)}$, most functions randomly chosen in $\mathcal{B}(X)$ according to ρ_p have G-variations greater or equal to $1/\lambda$. 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 $e^{-(m^{1/2}/2)}$. 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 [48]. Schläfli 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 $\text{card}X = m$

$$\text{card}\, P_d(X) \leq 2\sum_{l=1}^{d} \binom{m-1}{l} \leq 2\frac{m^d}{d!}. \tag{3}$$

(see [49]). The set $P_d(X)$ forms only a small fraction of the set of all functions in the set $\mathcal{B}(X)$, whose cardinality is equal 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 [49, Table I]). Even a large size of a dictionary does not guarantee that all functions have reasonably small variations. Although probabilities that a function has a large inner product with any element of G are small, they may not be independent.

VI. Conclusion

We investigated how to choose dictionaries of network computational units so that binary classification tasks can be efficiently solved. As, with increasing sizes of domains, the number of all classifications tasks becomes unmanageable large, we focused our investigation on tasks characterized by probability distributions, modeling their relevance for a given application area.

To identify efficient network designs, we explored the complexity of networks computing randomly chosen classification tasks. As minimization of the sparsity of a shallow network measured by the number of hidden units (formalized in terms of the "l_0-pseudonorm" of output-weight vectors) is a difficult nonconvex optimization task, we considered l_1-norm as an approximate measure of sparsity. We studied minimization of the l_1-norm of output-weight vectors in terms of variational norms tailored to dictionaries of computational units. We combined geometric properties of variational norms with a concentration of measure phenomena occurring in high-dimensional Euclidean spaces [50]. We explored the effects of increasing sizes of domains of the classification tasks on correlations between these tasks and computational units. Using a version of the Chernoff–Hoeffding Bound on sums of independent but not necessarily uniformly distributed random variables, we proved probabilistic bounds on variational norms of binary-valued functions and l_1-norms of output-weight vectors. We described the distributions of these norms in terms of the size of the domain, the size of the dictionary, and the probability distribution characterizing the type of tasks.

We proved that on large domains there exist functions highly correlated with almost all functions randomly chosen with respect to the probability modeling the prior knowledge. Thus, it is desirable to choose a dictionary containing or well approximating such functions. Otherwise, almost any randomly chosen task requires either a network with a large number of units or computation of the task is unstable as some of the output weights are large.

VII. Discussion

We presented a probabilistic approach to the selection of a suitable dictionary of computational units, assuming that the set of tasks to be computed is endowed with a probability distribution. Another probabilistic approach to the selection of computational units assumes that a function to be computed is fixed, while computational units are chosen randomly. Approximation with random bases has been investigated since 1995 [51]. In some literature, algorithms based on random selection of inner parameters of computational units are called "Extreme Learning Machines" [52]. In [53], a theoretical analysis of random and greedy approximation was given. It was shown therein that "both randomized and deterministic procedures are successful if additional information about the families of function to be approximated is provided. In the absence of such information, one may observe exponential growth of the number of terms needed to approximate the function and/or extreme sensitivity of the outcome of the approximation to parameters."

In the above-mentioned articles on approximation with random bases and extreme learning, there is an implicit assumption that the probability distribution, with respect to which computational units are chosen, is uniform. In this paper, instead, we considered any product distribution, not necessarily uniform. In many real applications, various biases characteristic for tasks of interest can be observed and training data can be imbalanced.

Our work was motivated by questions concerning neurocomputing. However, sparsity also plays an important role in signal processing, where representations of input data as linear combinations of a small number of components (called "atoms") are searched for. Sparse dictionary learning has applications, e.g., in image denoising and classification and video and audio processing (see [54], [55]). Our results can also be applied to the processing of signals on large domains.

We focused on binary classification tasks. An extension of our results to multiclass tasks and computation of general real-valued functions on finite domains is a subject for future work. To reach this goal, it will be necessary to exploit more general versions of the concentration of measure, such as the methods of averaged bounded differences [46]. We believe that such more general methods may allow one to extend the analysis also to tasks characterized by more general probability distributions.

Choice of dictionaries for efficient computation has been investigated here theoretically. For practical approaches to the efficient construction of sparse networks using regularization, we refer the reader to [56]–[58] and references therein.

References

[1] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control, Signals Syst.*, vol. 2, no. 4, pp. 303–314, 1989.

[2] R. Hecht-Nielsen, "Kolmogorov's mapping neural network existence theorem," in *Proc. Int. Joint Conf. Neural Netw.*, San Diego, CA, USA, 1987, pp. 11–14.

[3] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, 1989.

[4] A. Pinkus, "Approximation theory of the MLP model in neural networks," *Acta Numer.*, vol. 8, pp. 143–195, Jan. 1999.

[5] Y. Ito, "Finite mapping by neural networks and truth functions," *Math. Sci.*, vol. 17, pp. 69–77, 1992.

[6] Y. Bengio, O. Delalleau, and N. L. Roux, "The curse of highly variable functions for local kernel machines," in *Advances in Neural Information Processing Systems*, vol. 18. Cambridge, MA, USA: MIT Press, 2006, pp. 107–114.

[7] V. Kůrková and M. Sanguineti, "Model complexities of shallow networks representing highly varying functions," *Neurocomputing*, vol. 171, pp. 598–604, Jan. 2016.

[8] P. C. Kainen, V. Kůrková, and M. Sanguineti, "Dependence of computational models on input dimension: Tractability of approximation and optimization tasks," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 1203–1214, Feb. 2012.

[9] A. M. Tillmann, "On the computational intractability of exact and approximate dictionary learning," *IEEE Signal Process. Lett.*, vol. 22, no. 1, pp. 45–49, Jan. 2015.

[10] T. L. Fine, *Feedforward Neural Network Methodology*. Berlin, Germany: Springer, 1999.

[11] G. Gnecco and M. Sanguineti, "Regularization techniques and suboptimal solutions to optimization problems in learning from data," *Neural Comput.*, vol. 22, no. 3, pp. 793–829, Mar. 2010.

[12] S. Vaiter, G. Peyre, C. Dossal, and J. Fadili, "Robust sparse analysis regularization," *IEEE Trans. Inf. Theory*, vol. 59, no. 4, pp. 2001–2016, Apr. 2013.

[13] V. Kůrková and M. Sanguineti, "Geometric upper bounds on rates of variable-basis approximation," *IEEE Trans. Inf. Theory*, vol. 54, no. 12, pp. 5681–5688, Dec. 2008.

[14] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning With Sparsity: The Lasso and Generalizations*. London, U.K.: Chapman & Hall, 2015.

[15] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[16] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Comp. Rendus Math.*, vol. 346, nos. 9–10, pp. 589–592, May 2008.

[17] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Appl. Comput. Harmon. Anal.*, vol. 31, no. 1, pp. 59–73, Jul. 2011.

[18] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, Mar. 2010.

[19] F. Girosi, "An equivalence between sparse approximation and support vector machines," *Neural Comput.*, vol. 10, no. 6, p. 1455–1480, 1998.

[20] G. Gnecco, M. Gori, and M. Sanguineti, "Learning with boundary conditions," *Neural Comput.*, vol. 25, no. 4, pp. 1029–1106, Apr. 2013.

[21] G. Gnecco, M. Gori, and M. Sanguineti, "Foundations of support constraint machines," *Neural Comput.*, vol. 27, no. 2, pp. 388–480, 2015.

[22] V. Kůrková, "Complexity estimates based on integral transforms induced by computational units," *Neural Netw.*, vol. 33, pp. 160–167, Sep. 2012.

[23] H. W. Lin, M. Tegmark, and D. Rolnick, "Why does deep and cheap learning work so well?" *J. Stat. Phys.*, vol. 168, pp. 1223–1247, Sep. 2017.

[24] B. Doerr, "Analyzing randomized search heuristics: Tools from probability theory," in *Theory of Randomized Search Heuristics—Foundations and Recent Developments*. Singapore: World Scientific, 2011, ch. 1, pp. 1–20.

[25] V. Kůrková and M. Sanguineti, "Probabilistic bounds on complexity of networks computing binary classification tasks," in *Proc. Inf. Technol.-Appl. Theory, CEUR Workshop (ITAT)*, vol. 2203, S. Krajči, Ed. Aachen, Germany: Technical Univ. & CreateSpace Independent Publishing Platform, 2018, pp. 86–91.

[26] Y. Bengio and A. Courville, "Deep learning of representations," in *Handbook on Neural Information Processing*, M. Bianchini, M. Maggini, and L. Jain, Eds. Berlin, Germany: Springer, 2013.

[27] H. N. Mhaskar and T. Poggio, "Deep vs. shallow networks: An approximation theory perspective," *Anal. Appl.*, vol. 14, no. 6, pp. 829–848, 2016.

[28] V. Kůrková and P. C. Kainen, "Comparing fixed and variable-width Gaussian networks," *Neural Netw.*, vol. 57, pp. 23–28, Sep. 2014.

[29] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.

[30] P. Tseng, "Further results on stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 888–899, Feb. 2009.

[31] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.

[32] D. L. Donoho and Y. Tsaig, "Fast solution of ℓ_1-norm minimization problems when the solution may be sparse," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 4789–4812, Nov. 2008.

[33] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3320–3325, Dec. 2003.

[34] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution," *Commun. Pure Appl. Math.*, vol. 59, no. 6, pp. 797–829, 2006.

[35] A. R. Barron, "Neural net approximation," in *Proc. 7th Yale Workshop Adapt. Learn. Syst.*, K. S. Narendra, Ed. New Haven, CT, USA: Yale Univ. Press, 1992, pp. 69–72.

[36] V. Kůrková, "Dimension-independent rates of approximation by neural networks," in *Computer Intensive Methods in Control and Signal Processing: The Curse of Dimensionality*, K. Warwick and M. Kárný, Eds. Boston, MA, USA: Birkhäuser, 1997, pp. 261–270.

[37] V. Kůrková, P. C. Kainen, and V. Kreinovich, "Estimates of the number of hidden units and variation with respect to half-spaces," *Neural Netw.*, vol. 10, no. 6, pp. 1061–1068, Aug. 1997.

[38] G. Gnecco, "A comparison between fixed-basis and variable-basis schemes for function approximation and functional optimization," *J. Appl. Math.*, vol. 2012, 2012, Art. no. 806945.

[39] G. Gnecco, "On the curse of dimensionality in the Ritz method," *J. Optim. Theory Appl.*, vol. 168, no. 2, pp. 488–509, 2016.

[40] G. Gnecco, V. Kůrková, and M. Sanguineti, "Some comparisons of complexity in dictionary-based and linear computational models," *Neural Netw.*, vol. 24, no. 2, pp. 172–182, 2011.

[41] G. Gnecco and M. Sanguineti, "On a variational norm tailored to variable-basis approximation schemes," *IEEE Trans. Inf. Theory*, vol. 57, no. 1, pp. 549–558, Jan. 2011.

[42] V. Kůrková and M. Sanguineti, "Bounds on rates of variable-basis and neural-network approximation," *IEEE Trans. Inf. Theory*, vol. 47, no. 6, pp. 2659–2665, Sep. 2001.

[43] V. Kůrková and M. Sanguineti, "Comparison of worst case errors in linear and neural network approximation," *IEEE Trans. Inf. Theory*, vol. 48, no. 1, pp. 264–275, Jan. 2002.

[44] V. Kůrková, P. Savický, and K. Hlaváčková, "Representations and rates of approximation of real-valued Boolean functions by neural networks," *Neural Netw.*, vol. 11, no. 4, pp. 651–659, 1998.

[45] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 930–945, May 1993.

[46] D. P. Dubhashi and A. Panconesi, *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[47] M. Ledoux, *The Concentration of Measure Phenomenon*. Providence, NJ, USA: AMS, 2001.

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

[49] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Trans. Electron. Comput.*, vol. EC-14, no. 3, pp. 326–334, Jun. 1965.

[50] S. Levy, Ed., *Flavors of Geometry*. Cambridge, U.K.: Cambridge Univ. Press, 1997.

[51] B. Igelnik and Y.-H. Pao, "Stochastic choice of basis functions in adaptive function approximation and the functional-link net," *IEEE Trans. Neural Netw.*, vol. 6, no. 6, pp. 1320–1329, Nov. 1995.

[52] G.-B. Huang, L. Chen, and C.-K. Siew, "Universal approximation using incremental constructive feedforward networks with random hidden nodes," *IEEE Trans. Neural Netw.*, vol. 17, no. 4, pp. 879–892, Jul. 2006.

[53] A. N. Gorban, I. Y. Tyukin, D. V. Prokhorov, and K. I. Sofeikov, "Approximation with random bases: Pro et contra," *Inf. Sci.*, vols. 364–365, pp. 129–145, Oct. 2016.

[54] J.-L. Starck, M. Elad, and D. L. Donoho, "Image decomposition via the combination of sparse representations and a variational approach," *IEEE Trans. Image Process.*, vol. 14, no. 10, pp. 1570–1582, Oct. 2005.

[55] K. Engan, K. Skretting, and H. Husøy, "Family of iterative LS-based dictionary learning algorithms, ILS-DLA, for sparse signal representation," *Digit. Signal Process.*, vol. 17, no. 1, pp. 32–49, Jan. 2007.

[56] X. Qian, H. Huang, X. Chen, and T. Huang, "Efficient construction of sparse radial basis function neural networks using L_1-regularization," *Neural Netw.*, vol. 94, pp. 239–254, Oct. 2017.

[57] S. Scardapane, D. Comminiello, A. Hussain, and A. Uncini, "Group sparse regularization for deep neural networks," *Neurocomputing*, vol. 241, pp. 81–89, Jun. 2017.

[58] D. Vidaurre, C. Bielza, and P. Larrañaga, "Learning an L1-regularized Gaussian Bayesian network in the equivalence class space," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 5, pp. 1231–1242, Oct. 2010.

Věra Kůrková received the Ph.D. degree in mathematics from the Charles University, Prague, Czech Republic, and the DrSc. (Prof.) degree in theoretical computer science from the Czech Academy of Sciences, Prague.

She is currently a Senior Scientist with the Department of Machine Learning, Institute of Computer Science, Czech Academy of Sciences. Her research interests are in mathematical theory of neurocomputing and machine learning. She has published many research papers and several book chapters on capabilities and limitations of shallow and deep networks, dependence of network complexity on increasing dimensionality of computational tasks, nonlinear approximation theory, and connections between theory of inverse problems and generalization in machine learning.

Dr. Kůrková has been a member of the Board of the European Neural Network Society (ENNS) (2008–2016), where she is currently the President (2017–2019). She is also a member of the Editorial Boards of the journals *Neural Networks* and *Neural Processing Letters*. She received the Bolzano Medal for her contribution to mathematical sciences by the Czech Academy of Sciences, in 2010. She was the General Chair or Co-Chair of the European Conferences ICANNGA 2001, ICANN 2008, ICANN 2017, and ICANN 2018, and a Guest Editor of the special issues of the journals *Neural Networks and Neurocomputing*.

Marcello Sanguineti received the Laurea (M.Sc.) degree *(cum laude)* in electronic engineering and the Ph.D. degree in electronic engineering and computer science from the University of Genova, Genova, Italy.

He is currently an Associate Professor of operations research with the University of Genova. He is also a Research Associate at the Institute for Marine Enginering, National Research Council of Italy. He has co-authored over 200 research papers in archival journals, book chapters, and international conference proceedings. His main research interests are infinite-dimensional programming, machine learning, neural networks for optimization, network and team optimization, and affective computing.

Dr. Sanguineti was a member of the Program Committees of several conferences. He was the Chair of the Organizing Committee of the International Conference ICNPAA 2008, and is the Co-Chair of the International Conference on Optimization and Decision Science (Genova), in 2019. He has co-ordinated several national and international research projects on approximate solution of optimization problems. He is a member of the Editorial Boards of the journals *Neurocomputing* and *Neural Processing Letters*. From 2006 to 2012, he was an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS.