Comparison of Worst Case Errors in Linear and Neural Network Approximation

Věra Kůrková and Marcello Sanguineti

***Abstract*—Sets of multivariable functions are described for which worst case errors in linear approximation are larger than those in approximation by neural networks. A theoretical framework for such a description is developed in the context of nonlinear approximation by fixed versus variable basis functions. Comparisons of approximation rates are formulated in terms of certain norms tailored to sets of basis functions. The results are applied to perceptron networks.**

***Index Terms*—Complexity of neural networks, curse of dimensionality, high-dimensional optimization, linear and nonlinear approximation, rates of approximation.**

Summary of Notation

$(X, \|.\|)$	Normed linear space.
$B_r(\|.\|)$	Ball of radius r in $(X, \|.\|)$.
X_n	n-dimensional subspace of X.
G	Subset of $(X, \|.\|)$.
G^0	Set of normalized elements of G.
$cl\, G$	Closure of G with respect to $\|.\|$.
$int\, G$	Interior of G with respect to $\|.\|$.
∂G	Boundary of G with respect to $\|.\|$.
$c\,G$	$\{cg\colon g \in G\}$, $c \in \mathbb{R}$.
$G(c)$	$\{wg\colon g \in G, w \in \mathbb{R}, \|w\| \leq c\}$, $c > 0$.
s_G	$\sup_{g\in G} \|g\|$.
ν_G	Minkowski functional of G.
$span\, G$	Linear span of G.
$span_k G$	Set of all linear combinations of at most k elements of G.
$conv\, G$	Convex hull of G.
$conv_k G$	Set of all convex combinations of at most k elements of G.
e_M	Error functional of the set M in $(X, \|.\|)$.
$\delta(Y, M)$	Deviation of the set Y from the set M in $(X, \|.\|)$.
$\delta_{G,k}$	Deviation of the set Y from $span_k G$ in $(X, \|.\|)$.
$d_n(Y)$	Kolmogorov n-width of the set Y in $(X, \|.\|)$.
$\beta_n(Y)$	Bernstein n-width of the set Y in $(X, \|.\|)$.
G_ϕ	$G_\phi = \{\phi(a, .)\colon a \in A \subseteq \mathbb{R}^p\}$.
$span_k G_\phi$	Set of functions computable by a ϕ-network with k hidden units.
ϑ	Heaviside function.
κ	Ramp function.
J	Closed interval in $\mathbb{R}$.
$P_d(\psi, J)$, $P_d(\psi)$	Set of functions on J^d computable by ψ-perceptrons, $\psi\colon \mathbb{R} \to \mathbb{R}$.
$H_d(J)$, H_d	$H_d(J) = P_d(\vartheta, J)$, $H_d = P_d(\vartheta)$.
$\|.\|_{1,A}$	l_1 norm with respect to A.
$\|.\|_G$	G-variation with respect to $\|.\|$.
$\|.\|_{H_d}$	Variation with respect to half-spaces (Heaviside perceptrons).
$\|.\|_{S_d}$	Variation with respect to signum perceptrons.
$\|.\|_{R_d}$	Variation with respect to ramp perceptrons.
$V(\psi, J)$	Total variation of $\psi\colon J \to \mathbb{R}$.
$s_{A,G}$	$\sup_{\alpha\in A} \|\alpha\|_G$.

Manuscript received December 12, 1999; revised July 24, 2001. This work was supported in part by NATO under Grant PST.CLG.976870 (project "Approximation and Functional Optimization by Neural Networks"). The work of V. Kůrková was supported in part by GA ČR Grant 201/99/0092. The work of M. Sanguineti was supported in part by the Italian Ministry of University and Research (project "New Techniques for the Identification and Adaptive Control of Industrial Systems"). The material in this paper was presented in part at the XIV International Symposium on Mathematical Theory of Networks and Systems, Perpignan, France, June 19–23, 2000 and at the International Joint Conference on Neural Networks, Como, Italy, July 24–27, 2000.

V. Kůrková is with the Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague 8, Czech Republic (e-mail: vera@cs.cas.cz).

M. Sanguineti is with the Department of Communications, Computer, and System Sciences (DIST), University of Genoa, 16145 Genoa, Italy (e-mail: marcello@dist.unige.it).

Communicated by G. Lugosi, Associate Editor for Classification, Nonparametric Estimation, and Neural Networks.

Publisher Item Identifier S 0018-9448(02)00004-4.

I. Introduction

In many applications, where the approximate solution of multivariable optimization problems is required, proper use of neural networks as approximators helps to cope with the "curse of dimensionality" [1] and so to prevent the optimization task from becoming unmanageably complex with a growing number of variables. This is the case, for example, with dynamic programming [2], parametric approximation of decision strategies in optimization problems with high-dimensional state or output spaces [3], pattern recognition [4], approximate minimization of functionals [5], etc.

While theoretical investigations of approximation by neural networks have mostly focused on the existence of an arbitrarily close approximation and on how accuracy depends on a network's complexity, the difference between linear and neural approximators has remained less understood.

The first result attempting to explain the advantages of neurocomputing methods is Barron's [6] comparison of the worst case errors in linear and neural network approximation. He described sets of multivariable functions for which the $\mathcal{L}_2$ approximation error by one-hidden-layer sigmoidal perceptron networks is bounded from above by $O(1/\sqrt{k})$, where k is the number of

network hidden units, while the $\mathcal{L}_2$ error of the best linear approximator is bounded from below by $O(1/(d\sqrt[d]{n}))$, where n is the dimension of the linear approximating subspace and d is the number of variables of the functions to be approximated. As the number of free parameters of functions computable by one-hidden-layer perceptron networks with k hidden units is $k(d+2)$, Barron compares an upper bound of the order of $O(1/\sqrt{k})$ with a lower bound of the order of $O(1/(d\sqrt[d]{k(d+2)}))$.

Kainen, Kůrková, and Vogt [7], [8] have initiated the study of the comparison of properties of projections (best approximation operators) in linear and neural network approximation. They have shown that many useful properties of best approximation operators, like uniqueness, homogeneity, and continuity, are not satisfied by neural networks, and have suggested that this loss might allow improved approximation rates (as the arguments proving the slow rates of linear approximators are based on such properties).

In this paper, we improve and extend the results by Barron [6] on the comparison of rates of approximation. Generally, *rates of approximation* describe the tradeoff between the accuracy of approximation and the "complexity" of approximating functions. When such functions belong to a parameterized family, their complexity can be measured by the lengths of parameter vectors (depending on the number of variables and, e.g., on the degree of a polynomial or a rational function, on the number of knots in a spline, on the number of hidden units in a neural network, etc.).

To describe sets of multivariable functions, for which worst case errors in linear approximation are larger than those in approximation by neural networks, we investigate such errors in a general framework of fixed- versus variable-basis approximation. We call *fixed-basis approximation* an approximation scheme where the approximating functions are elements of finite-dimensional subspaces generated by the first n elements of a fixed basis, while in *variable-basis approximation*, the approximating functions are linear combinations of all k-tuples of elements of a given set. For example, algebraic and trigonometric polynomials belong to fixed-basis approximation, whereas free-node splines [9, Ch. 13], trigonometric polynomials with free frequencies [11], and feedforward neural networks belong to the variable-basis family.

Within the general framework of fixed- and variable-basis approximation, we derive estimates of worst case errors, formalized for the fixed-basis functions by the concept of *Kolmogorov n-width* (infimum of deviations from n-dimensional linear subspaces) and, for the variable ones, by the *deviation from the union of finite-dimensional subspaces* generated by all k-tuples of functions from a given basis. Considering relatively "small" (of the order of $O(1/\sqrt{k})$) upper bounds on such deviations of balls in certain norms, we investigate "large" lower bounds on their Kolmogorov widths. The norms we use to define these balls are tailored to various sets of variable-basis functions. The class of such norms includes l_1 norm [12], "spectral" norms [13], [6], [14], a generalization of total variation [15], [16], etc. We investigate several methods for deriving lower bounds on the Kolmogorov widths of balls in norms from this class. The lower bounds are formulated in terms of either the Bernstein width or the "capacity" of the basis (in the sense that its convex hull has an orthogonal subset containing, for any positive integer r, at least r^d functions with norms greater than or equal to $1/r$).

Applying these estimates to balls in norms tailored to perceptrons with periodic or sigmoidal activations, we obtain classes of multivariable functions for which neural networks outperform linear methods. Functions from such classes can be approximated by perceptron networks having k hidden units within an accuracy of the order of $O(1/\sqrt{k})$, while, for some periodic activations, no increase in the dimension of a linear approximating subspace can decrease the worst case error below a constant (Proposition 12). For sigmoidal activations, the worst case error in linear approximation is bounded from below by a quantity of the form $1/(4d\sqrt[d]{2k(d+2)})$ (Theorem 2 and Corollaries 4 and 5).

The paper is organized as follows. Section II contains basic concepts concerning approximation in normed linear spaces and feedforward neural networks. Section III describes approximation rates of the order of $O(1/\sqrt{k})$ by one-hidden-layer neural networks with k computational units, in terms of balls in certain norms tailored to such units. To compare these rates with those achievable using linear approximation schemes, in Section IV, we investigate methods of estimation of the Kolmogorov widths of balls in the above-mentioned norms. In Section V, the tools developed in the previous sections are applied to perceptron networks.

II. Preliminaries

A. Approximation in Normed Linear Spaces

For basic concepts concerning functional analysis and topology see, e.g., [17].

In this paper, a *normed linear space* is assumed to be real and is denoted by $(X, \|.\|)$ or merely X when there is no ambiguity on the norm. The dimension of X is denoted by $\dim X$. By $\|.\|_p$, $p \in [1, \infty]$, we denote the $\mathcal{L}_p$ norm with respect to the Lebesgue measure. $\mathbb{R}$ denotes the set of real numbers, $\mathbb{R}_+$ the set of positive real numbers, and $\mathbb{N}_+$ the set of positive integers.

If $(X, \|.\|)$ is a normed linear space, $B_r(\|.\|)$ denotes the ball of radius r, i.e., $B_r(\|.\|) = \{f \in X\colon \|f\| \leq r\}$.

If G is a subset of $(X, \|.\|)$, G^0 denotes the set of its *normalized* elements, i.e.,

$$G^0 = \left\{ g^0 = \frac{g}{\|g\|}\colon 0 \neq g \in G \right\}.$$

The *closure* of G is denoted by clG; G is *dense* in X if $clG = X$. The *interior* and the *boundary* of G are denoted by $int\, G$ and ∂G, respectively. For $c \in \mathbb{R}$, we define

$$cG = \{cg\colon g \in G\}$$

and for c positive

$$G(c) = \{wg\colon g \in G, w \in \mathbb{R}, |w| \leq c\}.$$

G is called *homogeneous* if $cG = G$ for all $c \in \mathbb{R}$, $c \neq 0$. If $G = -G$, then G is called *centrally symmetric*. If $G = G(1)$, then G is called *balanced*. $G(1)$ is called the *balanced hull of* G.

The *Minkowski functional* ν_G of a subset G of a linear space X is defined for all $f \in X$ as

$$\nu_G(f) = \inf \{c \in \mathbb{R}_+ \colon f \in cG\}$$

(see, e.g., [18, p. 131]). When G is balanced and convex, ν_G is a norm on $\{f \in X \colon \nu_G(f) < \infty\}$.

The *linear span* of G is denoted by $span\, G$, i.e.,

$$span\, G = \left\{ \sum_{i=1}^{k} w_i g_i \colon w_i \in \mathbb{R},\, g_i \in G,\, k \in \mathbb{N}_+ \right\}.$$

$span_k G$ denotes the set of all linear combinations of at most k elements of G, i.e.,

$$span_k G = \left\{ \sum_{i=1}^{k} w_i g_i \colon w_i \in \mathbb{R},\, g_i \in G \right\}.$$

The *convex hull* of G is denoted by $conv\, G$, i.e.,

$$conv G = \left\{ \sum_{i=1}^{k} a_i g_i \colon a_i \in [0, 1], \sum_{i=1}^{k} a_i = 1,\, g_i \in G,\, k \in \mathbb{N}_+ \right\}.$$

$conv_k G$ denotes the set of all convex combinations of at most k elements of G, i.e.,

$$conv_k G = \left\{ \sum_{i=1}^{k} a_i g_i \colon a_i \in [0, 1], \sum_{i=1}^{k} a_i = 1,\, g_i \in G \right\}.$$

A set G is called *convex* if $G = conv\, G$.

The *error functional* $e_M \colon X \to [0, +\infty)$ of a subset M of $(X, \|.\|)$ is defined as

$$e_M(f) = \|f - M\| = \inf_{g \in M} \|f - g\|.$$

For any normed linear space $(X, \|.\|)$ and $M \subseteq X$, e_M is uniformly continuous but does not need to be linear (see, e.g., [19, pp. 139–140 and p. 391]).

The *worst case approximation error* is formalized by the concept of *deviation* of a set Y of functions to be approximated from a set M of approximating functions, defined as

$$\delta(Y, M) = \sup_{f \in Y} e_M(f) = \sup_{f \in Y} \inf_{g \in M} \|f - g\|.$$

We do not specify the dependence of deviation on $(X, \|.\|)$, as it will be clear from the context. Note that deviation describes the "size" of the smallest neighborhood of M containing Y: $\delta(Y, M)$ is the infimum of all $\varepsilon > 0$, for which

$$Y \subseteq U_\varepsilon(M) = \{f \in X \colon \|f - M\| \le \varepsilon\}.$$

Proposition 1: Let $(X, \|.\|)$ be a normed linear space and let Z, Y, and M be its subsets. Then

i) if $Y \subseteq Z$, then $\delta(Y, M) \le \delta(Z, M)$;

ii) $\delta(Y, M) = \delta(cl\, Y, M)$;

iii) if M is homogeneous, then, for any $0 \neq c \in \mathbb{R}$, $\delta(cY, M) = |c| \delta(Y, M)$;

iv) if M is convex, then $\delta(Y, M) = \delta(conv\, Y, M)$.

Proof: i) and ii) follow directly from the definition of deviation and from the continuity of the error functional.

iii) holds as

$$\sup_{g \in cY} \|g - M\| = \sup_{f \in Y} \|cf - M\| = |c| \sup_{f \in Y} \|f - M\|.$$

iv)

$$\delta(Y, M) = \inf\{\varepsilon > 0 : Y \subseteq U_\varepsilon(M)\}$$

where

$$U_\varepsilon(M) = \{f \in X \colon \|f - M\| \le \varepsilon\}.$$

If $Y \subseteq U_\varepsilon(M)$, then $conv\, Y \subseteq U_\varepsilon(conv\, M) = U_\varepsilon(M)$. Hence, $\delta(Y, M) = \delta(conv\, Y, M)$. □

An approximation is called *linear* when the approximating functions belong to a *linear subspace*, often generated by the first n elements of a given linearly ordered set (for example, the set of all polynomials of order at most $n - 1$, generated by the first n elements of the set $\{x^{i-1} \colon i \in \mathbb{N}_+\}$). We call such an approximation scheme *fixed-basis approximation*, in contrast to *variable-basis approximation*, where the approximating functions are linear combinations of all k-tuples of elements of a given set G. They form the set $span_k G$ of all linear combinations of at most k elements of G, which is the *union of finite-dimensional subspaces* generated by k-tuples of elements of G (e.g., approximation by trigonometric polynomials with free frequencies, G being the set of sines and cosines with arbitrary frequencies). The number of parameters of $span_k\, G$ depends on k and on the number of parameters of the elements of G.

B. One-Hidden-Layer Feedforward Neural Networks

Feedforward neural networks compute parameterized sets of functions dependent both on the type of computational units and on the type of their interconnections. *Computational units* compute functions $\phi \colon \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ of two vector variables: an *input vector* and a *parameter vector*. ϕ corresponds to the type of unit and p and d correspond to the dimensions of the *parameter space* and of the *input space*, respectively.

We call *ϕ-networks* one-hidden-layer feedforward networks with hidden units computing a function ϕ and a single linear output unit. Thus, ϕ-networks compute functions of the form

$$\sum_{i=1}^{k} w_i \phi(a_i, .)$$

where $a_i \in A \subseteq \mathbb{R}^p$. Let us denote by

$$G_\phi = \{\phi(a, \cdot) \colon a \in A \subseteq \mathbb{R}^p\}$$

the parameterized set of functions corresponding to the computational unit ϕ. A ϕ-network with k hidden units can generate as its input–output functions all the elements of $span_k\, G_\phi$, which is the union of all at most k-dimensional subspaces spanned by k-tuples of elements of G_ϕ. Thus, $span_k\, G_\phi$ belongs to variable-basis approximation. Note that the number of free parameters in a ϕ-network with k computational units is $k(p + 1)$.

Standard types of hidden units are perceptrons. A *perceptron* with an *activation function* $\psi \colon \mathbb{R} \to \mathbb{R}$ computes functions of the form

$$\phi((v, b), x) = \psi(v \cdot x + b) \colon \mathbb{R}^{d+1} \times \mathbb{R}^d \to \mathbb{R}$$

where $v \in \mathbb{R}^d$ is an *input weight vector* and $b \in \mathbb{R}$ is a *bias*. Let J be a closed interval in $\mathbb{R}$. We denote by

$$P_d(\psi, J) = \left\{ f\colon J^d \to \mathbb{R}\colon f(x) = \psi(v \cdot x + b),\ v \in \mathbb{R}^d,\ b \in \mathbb{R} \right\}$$

the set of functions on J^d computable by ψ-perceptrons. When it is clear from the context what J is considered, we shall simply write $P_d(\psi)$ instead of $P_d(\psi, J)$. $span_k P_d(\psi, J)$ represents the set of functions on J^d computable by ψ-perceptron networks with k hidden units, and $span\, P_d(\psi, J)$ denotes the set of functions on J^d computable by such networks with any number of hidden units. The number of free parameters in a perceptron network with k perceptrons is $k(d+2)$.

C. Rates of Approximation

Rates of approximation describe the tradeoff between the accuracy of approximation and the "complexity" of approximating functions. When a class of such functions is represented as the union of a nested sequence of sets of parameterized functions, the complexity corresponds to the increasing length of a parameter vector. Let $\{M_j\colon j \in \mathbb{N}_+\}$ be a sequence of nested subsets of a normed linear space $(X, \|.\|)$. The *rate of approximation* of a subset Y of X by $\{M_j\colon j \in \mathbb{N}_+\}$ can be investigated in terms of worst case errors, corresponding to the deviations $\delta(Y, M_j)$.

If $\bigcup_{j\in\mathbb{N}_+} M_j$ is dense in X, then, for any $f \in X$, the sequence $\{e_{M_j}(f)\colon j \in \mathbb{N}_+\}$ converges to 0. In practical applications, this convergence has to be sufficiently fast to guarantee the desired accuracy of approximation for j small enough so that the functions from M_j have a moderate number of parameters. In the case of functions of d variables, sometimes it happens that deviations are of the order of $O(1/\sqrt[d]{j})$. In such a case, to achieve accuracy within ε, approximating functions with complexity of the order of $O(1/\varepsilon^d)$ are needed. Such exponential dependence of complexity on the number of variables is called the *curse of dimensionality* [1].

In fixed-basis approximation, the nested sets M_n are n-dimensional subspaces. The number of free parameters is then equal to n (the free parameters are only the coefficients of the linear combinations of the first n fixed-basis functions). To describe a theoretical lower bound on linear approximation, Kolmogorov [20] investigated the infimum of deviations over all n-dimensional subspaces of X. He introduced the concept of *n-width* (which was later called *Kolmogorov n-width*) of a set Y, defined as

$$d_n(Y) = \inf_{X_n} \delta(Y, X_n) = \inf_{X_n} \sup_{f\in Y} \inf_{g\in X_n} \|f - g\|,$$

where the leftmost infimum is taken over all n-dimensional subspaces X_n of X. For example, in $(\mathcal{L}_p([0,1]^d), \|.\|_p)$ the Kolmogorov widths of certain balls in Sobolev norms defined in terms of a fixed degree of smoothness exhibit the curse of dimensionality [21, pp. 232–233]. However, if the requirements on smoothness are appropriately increased with the number d of variables, then the curse of dimensionality can be avoided.

In the case of variable-basis approximation by ϕ-networks with k computational units, M_k corresponds to $span_k G_\phi$. The number of free parameters is equal to $k(p+1)$, where p is the number of free parameters of each hidden unit. Given a set $Y \subseteq (X, \|.\|)$ of functions to be approximated, to evaluate the rates of variable-basis approximation by ϕ-networks with k units in the hidden layer we shall investigate the deviation $\delta(Y, span_k G_\phi)$. As we are interested in the comparison of approximation rates by ϕ-networks having k computational units with the rates achievable by the optimal linear approximators with the same number of free parameters, we shall compare $\delta(Y, span_k G)$ with the Kolmogorov width $d_{k(p+1)}(Y)$.

III. Deviation From $span_k G$ of Balls in G-Variation

A. Properties of Deviation From Unions of Finite-Dimensional Subspaces

To derive tools for estimating rates of approximation by variable-basis functions, we investigate the properties of sets of functions of the form $span_k G$, where G is a subset of a normed linear space $(X, \|.\|)$. This approximation scheme includes nonlinear trigonometric approximation (i.e., approximation by trigonometric polynomials with free frequencies; see, e.g., [11]) as well as free-node splines (see, e.g., [9, Ch. 13]) and one-hidden-layer feedforward neural networks. Multilayer feedforward networks with a single linear output unit and k units in the last hidden layer belong to this approximation scheme as well, but they correspond to more complex sets G, which depend on the number of units in the previous hidden layers.

To simplify the notation, we shall denote the deviation of Y from $span_k G$ by $\delta_{G,k}$, i.e.,

$$\delta_{G,k}(Y) = \delta(Y, span_k G).$$

The following proposition states the basic properties of $\delta_{G,k}$ that follow directly from its definition and from Proposition 1 (note that $span_k G$ is homogeneous).

Proposition 2: Let $(X, \|.\|)$ be a normed linear space and Y, Z and G be its subsets. Then, for any positive integer k,

i) if $Y \subseteq Z$, then $\delta_{G,k}(Y) \le \delta_{G,k}(Z)$;
ii) $\delta_{G,k}(Y) \ge \delta_{G,k+1}(Y)$;
iii) $\delta_{G,k}(Y) = \delta_{G,k}(cl\, Y)$;
iv) for any $c \in \mathbb{R}$, $\delta_{G,k}(cY) = |c|\delta_{G,k}(Y)$.

B. Variation With Respect to a Set of Functions

Sets of multivariable functions with upper bounds of the order of $O(1/\sqrt{k})$ on the deviation from $span_k G$ can be described in terms of a norm tailored to a given set G.

For a subset G of a normed linear space $(X, \|.\|)$, *G-variation* (*variation with respect to the set G*), denoted by $\|.\|_G$, is defined as the Minkowski functional of the set $cl\ conv\ (G \cup -G)$, i.e.,

$$\|f\|_G = \inf \left\{ c \in \mathbb{R}_+\colon \frac{f}{c} \in cl\ conv\ (G \cup -G) \right\}.$$

Note that

$$\|f\|_G = \inf\{c \in \mathbb{R}_+\colon f \in cl\ conv\ G(c)\}$$

as $conv\ (G \cup -G) = conv\ G(1)$. G-variation is a norm on the subspace $\{f \in X\colon \|f\|_G < \infty\} \subseteq X$. It was defined by Kůrková [22] as an extension of Barron's [15] concept of variation with respect to half-spaces. Note that $\|f\|_G$ represents

the minimum "dilation" of the set G guaranteeing that f is contained in the closure of the convex symmetric hull of the "dilated" set.

When X is finite-dimensional, all norms on X are equivalent. Hence, in such a case, G-variation does not depend on the norm on X. In the infinite-dimensional case, G-variation in general depends on the choice of a norm on X. To simplify the notation, we shall not write such dependence explicitly, as the norm will be clear from the context.

The following proposition states the basic properties of G-variation.

Proposition 3: Let $(X, \|.\|)$ be a normed linear space, G and F be its subsets, and $s_G = \sup_{g\in G} \|g\|$. Then

i) for all $f \in X$

$$\|f\| \le s_G \|f\|_G;$$

ii) if G is finite with $card\, G = m$ and $f \in span\, G$, then

$$\|f\|_G = \min\left\{\sum_{i=1}^{m} |w_i| \colon f = \sum_{i=1}^{m} w_i g_i,\ g_i \in G,\ w_i \in \mathbb{R}\right\};$$

iii) $\|.\|_G \le c\|.\|_F$ if and only if, for all $h \in F$, $\|h\|_G \le c$.

Proof:

i) Follows from the definitions of $\|.\|_G$ and s_G.

For ii), see [23, Proposition 2.3].

iii) Follows from elementary properties of norms and from the fact that $\|h\|_G \le c$ for all $h \in B_1(\|.\|_F)$ if and only if it holds for all $h \in F$. □

G-variation is a generalization of l_1 norm. Let A be an orthonormal basis of a separable Hilbert space $(X, \|.\|)$. The l_1 *norm with respect to* A of $f \in X$ is defined as

$$\|f\|_{1,A} = \sum_{\alpha\in A} |f \cdot \alpha|.$$

The following relationship between A-variation and l_1 norm with respect to A has been shown in [12].

Proposition 4 [12]: Let $(X, \|.\|)$ be a separable Hilbert space and A its orthonormal basis. Then $\|.\|_A = \|.\|_{1,A}$.

Thus, when A is an orthonormal basis, the unit ball in A-variation coincides with the unit ball in l_1 norm with respect to A.

C. Upper Bounds on the Deviations of Balls in G-Variation

Some insights into the properties of sets of multivariable functions that can be approximated by neural networks with rates of the order of $O(1/\sqrt{k})$ have been obtained by Jones [13] and Barron [6]. The same estimate of approximation rates had earlier been proved by Maurey (see [24]). Using the concept of G-variation, Kůrková [22], [25] has reformulated Barron's [6] improvement of Jones' result [13] in the following way.

Theorem 1: Let $(X, \|.\|)$ be a Hilbert space, G its subset, and $s_G = \sup_{g\in G} \|g\|$. Then, for any $f \in X$ and for any positive integer k

$$\|f - span_k G\| \le \sqrt{\frac{(s_G\|f\|_G)^2 - \|f\|^2}{k}}.$$

Since $span_k G = span_k G^0$ (G^0 denotes the set of the normalized elements of G), Theorem 1 implies

$$\|f - span_k G\| \le \sqrt{\frac{\|f\|^2_{G^0} - \|f\|^2}{k}}.$$

As an immediate corollary, we get a description of sets of multivariable functions that can be approximated by $span_k G$ with rates of the order of $O(1/\sqrt{k})$.

Corollary 1: Let $(X, \|.\|)$ be a Hilbert space, G its subset, and $s_G = \sup_{g\in G} \|g\|$. Then, for any positive integer k

$$\delta_{G,k}(B_1(\|.\|_G)) \le \frac{s_G}{\sqrt{k}} \quad \text{and} \quad \delta_{G,k}(B_1(\|.\|_{G^0})) \le \frac{1}{\sqrt{k}}.$$

If the elements of $(X, \|.\|)$ are functions of d variables, then Corollary 1 implies that functions in the unit ball in G^0-variation can be approximated by elements of $span_k G$ with a rate that does not depend on d. This estimate of rate of approximation is sometimes called "dimension-independent" (see, e.g., [26]). However, this term is misleading as, with an increasing number d of variables, the condition of being in the unit ball in G^0-variation becomes more and more constraining (see [23] for examples of functions with variations dependent on d even exponentially).

Note that the upper bounds on $\delta_{G,k}$ in terms of G-variation are not restricted to Hilbert spaces. In [27], the result by Maurey, Jones, and Barron has been extended to $\mathcal{L}_p$ spaces, for $p \in (1, \infty)$, with a slightly worse rate of approximation (of the order of $O(k^{-1/q})$, where $q = \max\{p, p/(p-1)\}$). There also exist extensions to $\mathcal{L}_\infty$ (see, e.g., [15], [28]–[30], and [23]). An interesting improvement has been derived in [31], combining a concept from metric entropy theory with a probabilistic argument. The tightness of the bound $O(1/\sqrt{k})$ has been investigated in [15], [31], [23], [32] and [12].

IV. Kolmogorov Widths of Balls in G-Variation

A. Basic Properties of the Kolmogorov Widths of Balls in G-Variation

The following proposition summarizes the basic properties of the Kolmogorov n-width (see [33, pp. 132-133] and [21, p. 10]).

Proposition 5: Let $(X, \|.\|)$ be a normed linear space and Y and Z be its subsets. Then, for all positive integers n

i) if $Y \subseteq Z$, then $d_n(Y) \le d_n(Z)$;

ii) $d_n(Y) \ge d_{n+1}(Y)$;

iii) $d_n(cl\, Y) = d_n(Y)$;

iv) for any $c \in \mathbb{R}$, $d_n(cY) = |c| d_n(Y)$;

v) $d_n(conv\, Y) = d_n(Y)$;

vi) $d_n(Y^0) \inf_{f\in Y} \|f\| \le d_n(Y) \le d_n(Y^0) \sup_{f\in Y} \|f\|$;

vii) if $Y \subseteq Z$, then $d_n(Z) - \delta(Z, Y) \le d_n(Y) \le d_n(Z)$.

Thus, the Kolmogorov width of a set is equal to the Kolmogorov width of its closed, convex, balanced hull. As a convex balanced set determines a norm on the space X in which it forms the unit ball (via the Minkowski functional), the Kolmogorov width is essentially a property of balls in various norms on X. It represents the best possible accuracy that can be achieved when such balls are approximated linearly.

To describe sets of functions for which variable-basis approximation by ϕ-networks outperforms linear methods, we shall consider the unit balls in G_ϕ-variations as the sets of functions to be approximated and we shall find conditions on ϕ that guarantee that $\delta_{G,k}(B_1(\|.\|_{G_\phi}))$ is smaller than $d_{k(p+1)}(B_1(\|.\|_{G_\phi}))$ (recall that $k(p+1)$ is the number of free parameters in a ϕ-network with k computational units). We shall start by investigating lower bounds on the Kolmogorov widths of balls in variation with respect to a set G. The following proposition summarizes the basic properties of $d_n(B_1(\|.\|_G))$.

Proposition 6: Let $(X, \|.\|)$ be a normed linear space and G and F be its subsets. Then for any positive integer n

i) $d_n(B_1(\|.\|_G)) = d_n(G)$;

ii) if $F \subseteq G$, then $d_n(B_1(\|.\|_F)) \leq d_n(B_1(\|.\|_G))$;

iii) if $0 \neq s_{F,G} = \sup_{h\in F} \|h\|_G < \infty$, then

$$d_n\left(B_1(\|.\|_G)\right) = d_n(G) \geq \frac{1}{s_{F,G}}\, d_n(F) = \frac{1}{s_{F,G}}\, d_n\left(B_1(\|.\|_F)\right).$$

Proof:

i) Follows from Proposition 5 iii), v), vii), and from $B_1(\|.\|_G) = cl\ conv\,(G \cup -G)$.

ii) As $F \subseteq G$ implies $\|.\|_G \leq \|.\|_F$, we have $B_1(\|.\|_F) \subseteq B_1(\|.\|_G)$.

iii) Follows from Proposition 3 iii) and Proposition 5 iv), noting that $\|.\|_G \leq c\|.\|_F$ implies $c\,B_1(\|.\|_G) \supseteq B_1(\|.\|_F)$. □

The first of these elementary properties has an important consequence. It implies that any estimate of the worst case error in linear approximation of the unit ball in G-variation also applies to G itself. Thus, the speed of decrease of $d_n(G)$ can be evaluated using $d_n\left(B_1(\|.\|_G)\right)$. To derive a lower bound on the Kolmogorov n-width of $B_1(\|.\|_G)$ might be easier than for G.

B. Lower Bounds in Terms of the Bernstein Width

As pointed out by Proposition 3 i), for any subset G of a normed linear space $(X, \|.\|)$, $\|.\| \leq s_G\|.\|_G$, where $s_G = \sup_{g\in G} \|g\|$. Thus, when $s_G \neq 0$, the unit ball in $\|.\|$ contains the ball of radius $1/s_G$ in G-variation. When also the unit ball in G-variation contains a ball of some nonzero radius in $\|.\|$ (i.e., it has a nonempty interior in the topology induced on X by $\|.\|$), then the norms $\|.\|_G$ and $\|.\|$ are equivalent. In such a case, the Bernstein width can be used to estimate the Kolmogorov width of $B_1(\|.\|_G)$.

The *Bernstein n-width* of a subset Y of a normed linear space $(X, \|.\|)$ is defined as

$$\beta_n(Y) = \sup_{X_{n+1}} \sup\left\{r \in \mathbb{R}_+ \colon B_r\left(\|.\|^{X_{n+1}}\right) \subseteq Y\right\}$$

where the leftmost supremum is taken over all $(n+1)$-dimensional subspaces of X, and $\|.\|^{X_{n+1}}$ denotes the restriction of $\|.\|$ to X_{n+1} (see, e.g., [21, p. 13]). We do not specify the dependence of the Bernstein width on $(X, \|.\|)$, as it will be clear from the context. If Y is closed, convex, and centrally symmetric, then, for all integers n

$$\beta_n(Y) = \sup_{X_{n+1}} \inf_{f\in\partial(Y\cap X_{n+1})} \|f\|.$$

An argument based on the Borsuk Antipodality Theorem (see, e.g., [21, p. 11]) shows that the Kolmogorov n-width is bounded from below by the Bernstein n-width. More precisely, for any closed, convex, centrally symmetric subset Y of a Banach space $(X, \|.\|)$ and for any positive integer n, $d_n(Y) \geq \beta_n(Y)$ [21, p. 13]. To obtain from this estimate a lower bound on

$$d_{k(p+1)}(G_\phi) = d_{k(p+1)}(B_1(\|.\|_{G_\phi}))$$

larger than the upper bound on $\delta_{G_\phi,k}(B_1(\|.\|_{G_\phi}))$ guaranteed by Corollary 1, $\beta_{k(p+1)}(G_\phi)$ has to be larger than $s_{G_\phi}/\sqrt{k}$.

Let A be a countable orthonormal basis of a separable Hilbert space and, for any $(\alpha_1, \ldots, \alpha_{n+1}) \in A^{n+1}$, where $n < card\, A$, let

$$X_{n+1} = span\{\alpha_1, \ldots, \alpha_{n+1}\}$$

and

$$A_{n+1} = \{\alpha_1, \ldots, \alpha_{n+1}\}.$$

Then

$$B_1(\|.\|_A) \cap X_{n+1} = B_1(\|.\|_{A_{n+1}}).$$

By Proposition 4, we have $B_1(\|.\|_{1,A}) = B_1(\|.\|_A)$, hence we get, for all $n < card\, A$

$$\beta_n(B_1(\|.\|_A)) = \frac{1}{\sqrt{n+1}}.$$

Let

$$\beta(Y) = \inf_{f\in\partial Y} \|f\|.$$

The following bound is obtained by an elementary argument, which is a slight extension of a result from [33, p. 133].

Proposition 7: Let $(X, \|.\|)$ be a Banach space and G be its subset such that $\|f\|_G < \infty$ for any $f \in X$. Then, for any positive integer n such that $n < \dim X$

$$d_n(G) = d_n\left(B_1(\|.\|_G)\right) \geq \beta\left(B_1(\|.\|_G)\right) = \inf\left\{\|f\| \colon \|f\|_G = 1\right\}.$$

Proof: Let X_n be an n-dimensional subspace of X and $h \in X - X_n$. Then there exists $g \in X_n$ such that $\|h - X_n\| = \|h - g\|$ (see, e.g., [19, p. 98]). Let

$$f = \frac{h-g}{\|h-g\|_G}.$$

We shall show that $\|f - 0\| = \|f - X_n\|$. As for all $g' \in X_n$ also $g - \|h-g\|_G\, g' \in X_n$, we have

$$\|f\| = \frac{\|h-g\|}{\|h-g\|_G} \leq \frac{\|h-g-\|h-g\|_G\, g'\|}{\|h-g\|_G} = \|f - g'\|$$

hence $e_{X_n}(f) = \|f - X_n\| = \|f\|$. As $\|f\|_G = 1$, using Proposition 6 we get

$$d_n(G) = d_n\left(B_1(\|.\|_G)\right) \geq \|f\| \geq \inf\{\|f\| \colon \|f\|_G = 1\} = \beta\left(B_1(\|.\|_G)\right). \qquad \square$$

To illustrate Proposition 7, consider $\mathbb{R}^m$ with the l_2 norm and an orthonormal basis A. As by Proposition 4 $B_1(\|.\|_A) = B_1(\|.\|_{1,A})$, we get

$$d_n(A) = d_n\left(B_1(\|.\|_A)\right) \geq \frac{1}{\sqrt{m}}$$

for any $n < m$.

Note that if the unit ball in G-variation has an empty interior with respect to $\|.\|$, then the method of estimation of $d_n(B_1(\|.\|_G))$ based on Proposition 7 gives the trivial lower bound equal to zero.

C. Lower Bounds on the Kolmogorov Widths of Orthogonal Sets

Even when the unit ball in G-variation contains no balls in the norm $\|.\|$, it might contain a ball of nonzero radius in A-variation for some set A, the Kolmogorov width of which can be estimated from below. In particular, for A orthonormal, we can use the following estimate, which is a slight improvement of a bound obtained by Barron [6, p. 942, Lemma 6].

Proposition 8: Let $(X, \|.\|)$ be a Hilbert space and A be its orthonormal subset. If A is infinite, then, for all positive integers n

$$d_n(A) \geq 1.$$

If A is finite of cardinality m, then, for all positive integers $n \leq m$

$$d_n(A) \geq \sqrt{1 - \frac{n}{m}}.$$

Proof: Let

$$X_n = span\{h_1, \cdots, h_n\}$$

where $\{h_1, \cdots, h_n\}$ is an orthonormal subset of X, and let $p_{X_n}: X \to X_n$ be the best approximation mapping (projection) from X to X_n (see, e.g., [19, p. 139]). Then, for any $\alpha \in A$, we have

$$e^2_{X_n}(\alpha) = \|\alpha - p_{X_n}(\alpha)\|^2 = 1 - \|p_{X_n}(\alpha)\|^2$$

and

$$\|p_{X_n}(\alpha)\|^2 = \sum_{j=1}^{n} (\alpha \cdot h_j)^2.$$

For any $m \in \mathbb{N}_+$, let A_m be a subset of A of cardinality $m \geq n$. Then

$$\sum_{\alpha \in A_m} \|p_{X_n}(\alpha)\|^2 = \sum_{j=1}^{n} \sum_{\alpha \in A_m} (\alpha \cdot h_j)^2 \leq \sum_{j=1}^{n} \|h_j\|^2 = n.$$

Hence, there exists $\alpha_m \in A_m$ such that $\|p_{X_n}(\alpha_m)\|^2 \leq n/m$. Thus,

$$e_{X_n}(\alpha_m) \geq \sqrt{1 - \frac{n}{m}}.$$

So, if A is finite of cardinality m, we have $d_n(A) \geq \sqrt{1 - n/m}$. If A is infinite, then, for any $m \in \mathbb{N}_+$, we have $d_n(A) \geq \sqrt{1 - n/m}$, hence $d_n(A) \geq 1$. □

Note that, for a countable orthonormal set A, Proposition 8 gives a lower bound on the Kolmogorov width of A larger than the bound, equal to $1/\sqrt{n+1}$, derived in the previous subsection for $n < m$ using the Bernstein width.

Proposition 8 implies a lower bound on the Kolmogorov width of any set G, for which there exists an orthonormal set A with a finite value of $\sup_{\alpha \in A} \|\alpha\|_G$.

Corollary 2: Let $(X, \|.\|)$ be a Hilbert space, G and A be its subsets, and let A be orthonormal with

$$0 \neq s_{A,G} = \sup_{\alpha \in A} \|\alpha\|_G < \infty.$$

If A is infinite, then, for all positive integers n

$$d_n(G) \geq \frac{1}{s_{A,G}}.$$

If A is finite of cardinality m, then, for all positive integers $n \leq m$

$$d_n(G) \geq \frac{1}{s_{A,G}} \sqrt{1 - \frac{n}{m}}.$$

Proof: By Proposition 3 iii), $\|.\|_G \leq s_{A,G}\|.\|_A$. Thus,

$$B_{1/s_{A,G}}(\|.\|_A) = \frac{1}{s_{A,G}} B_1(\|.\|_A) \subseteq B_1(\|.\|_G).$$

Hence, by Proposition 6 we have

$$d_n(G) = d_n(B_1(\|.\|_G)) \geq \frac{1}{s_{A,G}} d_n(B_1(\|.\|_A)) = \frac{1}{s_{A,G}} d_n(A)$$

and we conclude by Proposition 8. □

Corollary 2 implies that, whenever the unit ball in G-variation contains a ball of nonzero radius η in variation with respect to an infinite orthonormal set, G cannot be approximated with an error smaller than η using a linear approximation scheme. No increase at all in the dimension n of the linear approximating space can decrease the Kolmogorov n-width of G below η.

Even when $B_1(\|.\|_G)$ is not "large enough" to contain a ball of some nonzero radius in variation with respect to an infinite orthonormal set, it might contain a ball in variation with respect to some orthogonal set, the elements of which have norms going to zero "rather slowly" with respect to a positive integer d. The following definition formalizes the concept of such a slow decrease.

Let $(X, \|.\|)$ be a normed linear space, A its countable subset, and d a positive integer. We say that A is *not quickly vanishing with respect to* d if A can be linearly ordered as $A = \{\alpha_j : j \in \mathbb{N}_+\}$, so that the norms of its elements are nonincreasing and, for all $r \in \mathbb{N}_+$, $\|\alpha_{r^d}\| \geq 1/r$. Note that A is not quickly vanishing with respect to d if and only if it can be represented as $A = \bigcup_{r \in \mathbb{N}_+} A_r$, where, for all $r \in \mathbb{N}_+$, $card\, A_r \geq r^d$, $\|\alpha\| \geq 1/r$ for all $\alpha \in A_r$, and, for all $r' > r$ and $\alpha' \in A_{r'}$, $\|\alpha\| \geq \|\alpha'\|$.

The following proposition demonstrates the slow decrease of the Kolmogorov n-widths of orthogonal, not quickly vanishing sets: if an orthogonal set is not quickly vanishing with respect to d and $n = r^d/2$ for some integer r, then its n-width is bounded from below by $1/(\sqrt{2}\sqrt[d]{2n})$.

Proposition 9: Let $(X, \|.\|)$ be a Hilbert space and A be its countable orthogonal subset, not quickly vanishing with respect to a positive integer d. Then, for all positive integers n

$$d_n(A) \geq \frac{1}{\sqrt{2}\sqrt[d]{m_n}}$$

where

$$m_n = \min\left\{m \in \mathbb{N}_+ \colon (2n \le m)\,(\exists r \in \mathbb{N}_+)\,(m = r^d)\right\}.$$

If $2n = r^d$ for some integer r, then

$$d_n(A) \ge \frac{1}{\sqrt{2}\sqrt[d]{2n}}. \tag{1}$$

Proof: Let $A_r = \{\alpha_1, \ldots, \alpha_{r^d}\} \subset A$. Then $card\, A_r = r^d$ and, by Proposition 8, for all $r \in \mathbb{N}_+$ and all $n \le r^d$

$$d_n(A_r^0) \ge \sqrt{1 - \frac{n}{r^d}}$$

where A_r^0 denotes the set of normalized elements of A_r. By Proposition 5 vi)

$$d_n(A_r) \ge \frac{1}{r}\sqrt{1 - \frac{n}{r^d}}.$$

For any $n \in \mathbb{N}_+$, take the smallest $r \in \mathbb{N}_+$ for which $2n \le m_n = r^d$. By Proposition 5, we have

$$d_n(A) \ge d_n(A_r) \ge \frac{1}{r}\sqrt{1 - \frac{n}{r^d}} \ge \frac{1}{r\sqrt{2}}.$$

As $r = \sqrt[d]{m_n}$, we get $d_n(A) \ge 1/(\sqrt{2}\sqrt[d]{m_n})$. □

The lower bound (1) implies that, in linear approximation of an orthogonal set of functions of d variables that is not quickly vanishing with respect to d, the dimension of a linear subspace necessary to guarantee an accuracy ε has to be of the order of $O((1/\varepsilon)^d)$. Thus, this lower bound exhibits the curse of dimensionality (the term "dimensionality" referring to the number d of variables).

Let $\{G_d \colon d \in \mathbb{N}_+\}$ be a family of sets for which there exist orthogonal sets $\{A_d \colon d \in \mathbb{N}_+\}$ not quickly vanishing with respect to d and such that, for all d, $s_{A_d, G_d} = \sup_{\alpha \in A_d} \|\alpha\|_{G_d}$ is finite. Even when s_{A_d, G_d} does not grow too quickly with d, using Proposition 9 we might get useful lower bounds on the Kolmogorov widths of the sets $\{G_d \colon d \in \mathbb{N}_+\}$.

Corollary 3: Let $(X, \|.\|)$ be a Hilbert space, G_d and A_d be its subsets, A_d be orthogonal not quickly vanishing with respect to a positive integer d, and

$$0 \ne s_{A_d, G_d} = \sup_{\alpha \in A_d} \|\alpha\|_{G_d} < \infty.$$

Then, for any $n \in \mathbb{N}_+$

$$d_n(G_d) \ge \frac{1}{s_{A_d, G_d}\sqrt{2}\sqrt[d]{m_n}}$$

where

$$m_n = \min\left\{m \in \mathbb{N}_+ \colon (2n \le m)\,(\exists r \in \mathbb{N}_+)\,(m = r^d)\right\}.$$

In particular, if $2n = r^d$ for some integer r, then

$$d_n(G_d) \ge \frac{1}{s_{A_d, G_d}\sqrt{2}\sqrt[d]{2n}}. \tag{2}$$

Note that the dependence of s_{A_d, G_d} on d is crucial for the speed of decrease of the estimate of the Kolmogorov n-width of G_d. For example, if d is the number of variables and s_{A_d, G_d} is constant, then (2) implies the curse of dimensionality.

In the next section, we shall apply Corollary 3 to sets of d-variable functions computable by one-hidden-layer perceptron networks.

V. Comparison of Rates of Linear and Perceptron Network Approximation

A. Variation With Respect to Perceptrons

To apply the tools developed in the previous sections to perceptron networks, we shall first derive some basic properties of variation with respect to sets of functions computable by perceptrons with various types of activation functions.

The most common activation functions are *sigmoidals*, i.e., bounded measurable functions $\sigma \colon \mathbb{R} \to \mathbb{R}$ such that

$$\lim_{t \to -\infty} \sigma(t) = 0 \qquad \text{and} \qquad \lim_{t \to +\infty} \sigma(t) = 1.$$

One can use both continuous sigmoidals (like the *logistic sigmoid* $1/(1 + \exp(-t))$ or the *hyperbolic tangent*) and the discontinuous *Heaviside function* ϑ, defined as $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \ge 0$. Let J be a closed interval in $\mathbb{R}$. Note that the set $P_d(\vartheta, J)$ of functions computable by Heaviside perceptrons is equal to the *set of characteristic functions of half-spaces* of $\mathbb{R}^d$ restricted to J^d. Indeed, the function $\vartheta(v \cdot . + b)$ restricted to J^d is equal to the characteristic function of $\{x \in J^d \colon v \cdot x + b \ge 0\}$. We shall write $H_d(J)$ and H_d instead of $P_d(\vartheta, J)$ and $P_d(\vartheta)$, respectively, and we shall call variation with respect to H_d *variation with respect to half-spaces* and denote it by $\|.\|_{H_d}$.

Sometimes it is more convenient to use as an activation the *signum* function, defined as $\mathrm{sgn}(t) = -1$ for $t < 0$ and $\mathrm{sgn}(t) = 1$ for $t \ge 0$. We shall write S_d instead of $P_d(\mathrm{sgn})$ to denote variation with respect to signum perceptrons. Also other types of activation functions have been considered, like the *cosine* function [34] and the *ramp* function κ [14], defined as $\kappa(t) = t\vartheta(t)$, i.e., $\kappa(t) = 0$ for $t < 0$ and $\kappa(t) = t$ for $t \ge 0$. We shall write R_d instead of $P_d(\kappa)$.

The following proposition describes some elementary relationships among variations with respect to perceptrons with various kinds of activation functions.

Proposition 10: Let d be a positive integer and $p \in [1, \infty)$. Then, in $(\mathcal{L}_p(J^d), \|.\|_p)$, the following holds:

i) $\|.\|_{S_d} \le \|.\|_{H_d} \le 3\|.\|_{S_d}$;

ii) for any sigmoidal function σ

$$\|.\|_{P_d(\sigma)} \le \|.\|_{H_d};$$

iii) for any continuous nondecreasing sigmoidal σ

$$\|.\|_{P_d(\sigma)} = \|.\|_{H_d};$$

iv) for any continuous nondecreasing sigmoidal σ

$$\|.\|_{R_d} = 2\|.\|_{P_d(\sigma)}.$$

Proof:

i) $\|.\|_{S_d} \le \|.\|_{H_d}$ follows from Proposition 3 iii), noting that, for $t \in J = [t_1, t_2]$ and $b \ge \max\{|t_1|, |t_2|\}$

$$\vartheta(t) = \frac{1}{2}\mathrm{sgn}(t) + \frac{1}{2} = \frac{1}{2}\mathrm{sgn}(t) + \frac{1}{2}\mathrm{sgn}(t + b)$$

implies $\|\vartheta\|_{S_d} \le 1$. Similarly, $\|.\|_{H_d} \le 3\|.\|_{S_d}$ is obtained from Proposition 3 iii) noting that $\operatorname{sgn}(t) = 2\vartheta(t) - 1 = 2\vartheta(t) - \vartheta(t+b)$, implies $\|\operatorname{sgn}\|_{H_d} \le 3$.

For ii) and iii), see [16, Propositions 3.3 and 3.4].

To verify iv), consider the function ρ defined as $\rho(t) = 0$ for $t < 0$, $\rho(t) = t$ for $0 \le t < 1$, and $\rho(t) = 1$ for $t \ge 1$. Since $\rho(t) = \kappa(t) - \kappa(t-1)$, we have $\|.\|_{R_d} \le 2\|.\|_{P_d(\rho)}$. As ρ is a continuous nondecreasing sigmoidal, it follows from iii) that $\|.\|_{P_d(\rho)} = \|.\|_{H_d} = \|.\|_{P_d(\sigma)}$ for any continuous nondecreasing sigmoidal function σ. □

Thus, in $(\mathcal{L}_p(J^d), \|.\|_p)$, $p \in [1, \infty)$, variation with respect to half-spaces is equal to variation with respect to perceptrons with any continuous nondecreasing sigmoidal activation function or, up to a multiplicative constant, to variation with respect to signum or ramp perceptrons. In particular

$$B_1(\|.\|_{P_d(\sigma)}) = B_1(\|.\|_{H_d})$$

for any continuous nondecreasing sigmoidal σ. Thus, applying to perceptron networks with such sigmoidals Corollary 1 (as well as its various extensions to $\mathcal{L}_p$ spaces with $p \in (1, \infty)$, which can also be formulated in terms of variation), we can restrict our investigation to variation with respect to half-spaces. Moreover, since by Proposition 10 ii) for any sigmoidal function σ we have

$$B_1(\|.\|_{P_d(\sigma)}) \supseteq B_1(\|.\|_{H_d})$$

any lower bound on $d_n(B_1(\|.\|_{H_d}))$ can be applied to $d_n(B_1(\|.\|_{P_d(\sigma)}))$.

B. Variation With Respect to Half-Spaces of Plane Waves

To obtain a lower bound on the Kolmogorov width of the unit ball in variation with respect to half-spaces, we shall use not quickly vanishing orthogonal families containing plane waves. A function $f\colon \mathbb{R}^d \to \mathbb{R}$ is called a *plane wave* if it can be represented as $f(x) = \psi(v \cdot x)$, where $\psi\colon \mathbb{R} \to \mathbb{R}$ and $v \in \mathbb{R}^d$. Note that plane waves are constant along hyperplanes parallel to the cozero hyperplane $\{x \in \mathbb{R}^d\colon v \cdot x = 0\}$ of the linear function $v \cdot x$.

For $d = 1$, variation with respect to half-spaces coincides with total variation up to a constant [15], [16]. Recall that the *total variation* of a function $h\colon J \to \mathbb{R}$ of bounded variation on $J = [t_0, t_n]$ is defined as

$$V(h, J) = \sup \sum_{i=1}^{n} |h(t_i) - h(t_{i-1})|$$

where the supremum is taken over all finite partitions $t_0 < \cdots < t_n$ of J (see, e.g., [18, p. 328]). It follows directly from the definition of total variation that, for a periodic function $h\colon \mathbb{R} \to \mathbb{R}$ with a period τ and bounded variation on $[0, \tau]$

$$V(h, J) \le \left\lceil \frac{l}{\tau} \right\rceil V(h, [0, \tau]) \tag{3}$$

where l denotes the length of the interval J.

Variation with respect to half-spaces of a plane wave $f(x) = \psi(v \cdot x)$ can be estimated in terms of the total variation of ψ using the property (3) together with the following two lemmas.

Lemma 1: Let d be a positive integer and

$$f \in (\mathcal{L}_p([0, 1]^d), \|.\|_p), \qquad p \in [1, \infty)$$

be a plane wave $f(x) = \psi(v \cdot x)$, where $v = (v_1, \ldots, v_d) \in \mathbb{R}_+^d$ and $J = [0, \sum_{i=1}^d v_i]$. If $\|\psi\|_{H_1(J)} < \infty$, then

$$\|f\|_{H_d([0, 1]^d)} \le \|\psi\|_{H_1(J)}$$

where $H_d([0, 1]^d)$-variation is considered with respect to $(\mathcal{L}_p([0, 1]^d), \|.\|_p)$, and $H_1(J)$-variation with respect to $(\mathcal{L}_p(J), \|.\|_p)$.

Proof: Set $b = \|\psi\|_{H_1(J)}$. From the definition of $H_1(J)$-variation, we have

$$\psi(t) = \lim_{m \to \infty} \left(\sum_{j=1}^{n_m} w_{m,j} \vartheta(t + t_{m,j}) + \sum_{j=1}^{n_{m'}} w'_{m',j} \vartheta(-t + t'_{m',j}) \right)$$

in $(\mathcal{L}_p(J), \|.\|_p)$, where, for all $m \in \mathbb{N}_+$,

$$t_{m,0}, \ldots, t_{m,n_m}, t_{m',0}, \ldots, t_{m',n_{m'}} \in J$$

and

$$\sum_{j=1}^{n_m} |w_{m,j}| + \sum_{j=1}^{n_{m'}} |w'_{m',j}| \le b.$$

As, for all $x \in [0, 1]^d$, $v \cdot x \in J = [0, \sum_{i=1}^d v_i]$, we have

$$\psi(v \cdot x) = \lim_{m \to \infty} \left(\sum_{j=1}^{n_m} w_{m,j} \vartheta(v \cdot x + t_{m,j}) + \sum_{j=1}^{n_{m'}} w'_{m',j} \vartheta(-v \cdot x + t'_{m',j}) \right)$$

in $(\mathcal{L}_p([0, 1]^d), \|.\|_p)$. Hence,

$$\|f(x)\|_{H_d([0, 1]^d)} = \|\psi(v \cdot x)\|_{H_d([0, 1]^d)} \le b = \|\psi(t)\|_{H_1(J)}.$$

□

Lemma 2: Let $h\colon J \to \mathbb{R}$ be a function of bounded variation over a closed interval $J \subset \mathbb{R}$. Then

$$\|h\|_{H_1(J)} \le V(h, J) + |h(0)|$$

where $H_1(J)$-variation is considered with respect to $(\mathcal{L}_p(J), \|.\|_p)$, $p \in [1, \infty)$.

Proof: By [27, Theorem 6] (see also [15]) applied to $\overline{h}(t) = h(t) - h(0)$, we get

$$\|\overline{h}\|_{H_1(J)_{\sup}} = V(\overline{h}, J)$$

where $\|.\|_{H_1(J)_{\sup}}$ denotes $H_1(J)$-variation with respect to the topology of the uniform convergence on J. We conclude noting that

$$V(h, J) = V(\overline{h}, J)$$

$$\|h\|_{H_1(J)_{\sup}} \le \|\overline{h}\|_{H_1(J)_{\sup}} + |h(0)|$$

and

$$\|h\|_{H_1(J)} \le \|h\|_{H_1(J)_{\sup}}$$

where $\|.\|_{H_1(J)}$ denotes $H_1(J)$-variation with respect to $\|.\|_p$. □

From Lemma 1 and Lemma 2, we get the following upper bound on variation with respect to half-spaces of plane waves.

Proposition 11: Let d be a positive integer,

$$f \in (\mathcal{L}_p([0,1]^d), \|.\|_p), \qquad p \in [1, \infty)$$

be a plane wave such that $f(x) = \psi(v \cdot x)$, where $v = (v_1, \ldots, v_d) \in \mathbb{R}_+^d$, and ψ be a function of bounded variation on J, where $J = [0, \sum_{i=1}^d v_i]$. Then

$$\|f\|_{H_d([0,1]^d)} \le V(\psi, J) + |\psi(0)|.$$

C. Lower Bounds for Perceptrons With Periodic or Sigmoidal Activations

We shall derive estimates of the Kolmogorov widths of balls in variation with respect to perceptrons using embeddings of suitable orthogonal sets.

It is easy to check that the family

$$A_d(\sin) = \left\{\sqrt{2}\,\sin(\pi v \cdot x)\colon v \in \mathbb{N}_+^d\right\}$$

is orthonormal in $(\mathcal{L}_2([0,1]^d), \|.\|_2)$ for all positive integers d.

As $A_d(\sin)$ is a subset of $\sqrt{2}P_d(\sin)$ the following lower bound follows from Proposition 8.

Proposition 12: For all positive integers d, n, in $(\mathcal{L}_2([0,1]^d), \|.\|_2)$

$$d_n(P_d(\sin)) = d_n(B_1(\|.\|_{P_d(\sin)})) \ge \frac{1}{\sqrt{2}}.$$

Thus, there is no possibility of decreasing the $\mathcal{L}_2$ worst case error in linear approximation of $B_1(\|.\|_{P_d(\sin)})$ (even of $P_d(\sin)$) under $1/\sqrt{2}$ by increasing the dimension of a linear approximating space. Thus, perceptrons with sine activation cannot be efficiently approximated linearly.

On the other hand, from Corollary 1 it follows that functions in the unit ball $B_1(\|.\|_{P_d(\sin)})$ can be approximated by $span_k P_d(\sin)$ with a worst case error bounded from above by $O(1/\sqrt{k})$. More precisely

$$\delta_{P_d(\sin),k}\left(B_1(\|.\|_{P_d(\sin)})\right) \le \frac{1}{\sqrt{2k}}.$$

Barron [6, p. 942, Theorem 6] considered sets of functions Γ_c^d, $c > 0$, defined as

$$\Gamma_c^d = \left\{f\colon \mathbb{R}^d \to \mathbb{R}\colon \int_{\mathbb{R}^d} |\omega|\,|\tilde{f}(\omega)|\,d\omega \le c\right\}$$

where $\tilde{f}$ is the Fourier transform of f, and $|\omega| = \sqrt{\omega \cdot \omega}$ denotes the l_2 norm of the frequency ω (note that, with d increasing, the condition defining the sets Γ_c^d becomes more constraining). Barron proved that, in $(\mathcal{L}_2([0,1]^d), \|.\|_2)$

$$d_k(\overline{\Gamma}_c^d) \ge \frac{ca}{d\sqrt[d]{k}}$$

where $a \ge 1/(8\pi e^{\pi-1})$ and $\overline{\Gamma}_c^d = \{f|_{[0,1]^d}\colon f \in \Gamma_c^d\}$, while, in $(\mathcal{L}_2(B_1^d), \|.\|_2)$

$$\delta_{P_d(\sigma),k}(\tilde{\Gamma}_c^d) \le \frac{2c}{\sqrt{k}}$$

where $\tilde{\Gamma}_c^d = \{f|_{B_1^d}\colon f \in \Gamma_c^d\}$ and B_1^d denotes the unit ball in $\mathbb{R}^d$ with the l_2 norm. (The reader should note that there is an unfortunate misprint in the Discussion in [6] referring to the result of [6, Theorem 6] as $O(1/\sqrt[d]{n})$, omitting the factor d in the denominator.)

It should be noted that in [6] the Kolmogorov n-width $d_k(\overline{\Gamma}_c^d)$ is considered in $(\mathcal{L}_2([0,1]^d), \|.\|_2)$, while $\delta_{P_d(\sigma),k}(\tilde{\Gamma}_c^d)$ is considered in $(\mathcal{L}_2(B_1^d), \|.\|_2)$. The volume of the unit ball B_1^d in $\mathbb{R}^d$ with the l_2 norm is equal to $\pi^{d/2}/\Gamma((d+2)/2)$ [35, p. 304], where Γ denotes the gamma function. Thus, the Lebesgue measure of B_1^d goes to zero with the dimension d, in contrast to the behavior of the d-dimensional cube of side 1 (see also the remarks in [36, Sec. 18.2]).

Barron [6] compared linear approximation by k-dimensional subspaces with approximation by sigmoidal perceptron networks with k hidden units. Taking into account the number of free parameters, we shall instead compare approximation by $k(d+2)$-dimensional subspaces with networks having k perceptrons. We shall derive a lower bound of the form $1/(4d\sqrt[d]{2k(d+2)})$ on the Kolmogorov $k(d+2)$-width in $(\mathcal{L}_2([0,1]^d), \|.\|_2)$ of the set of functions computable by networks with k sigmoidal perceptrons. Further, we shall show that, for Heaviside activation functions, the worst case error is achieved.

To derive from Corollary 3 lower bounds on the Kolmogorov width of balls in variation with respect to half-spaces, we shall scale elements of $\{\sin(\pi v \cdot x)\colon v \in \mathbb{N}_+^d\}$ to obtain an orthogonal set not quickly vanishing with respect to d, which can be embedded in a ball in variation with respect to half-spaces.

Theorem 2: In $(\mathcal{L}_2([0,1]^d), \|.\|_2)$, for all positive integers d and n

$$d_n(H_d) = d_n\left(B_1(\|.\|_{H_d})\right) \ge d_{m_n}(H_d) \ge \frac{1}{4d\sqrt[d]{m_n}}$$

where

$$m_n = \min\left\{m \in \mathbb{N}_+\colon (2n \le m)\,(\exists r \in \mathbb{N}_+)\,(m = r^d)\right\}.$$

In particular, if $2n = r^d$ for some integer r, then

$$d_n(H_d) = d_n\left(B_1(\|.\|_{H_d})\right) \ge \frac{1}{4d\sqrt[d]{2n}}.$$

Proof: Using Corollary 3 and Proposition 11, we shall derive a lower bound on $d_n(H_d)$ using an orthogonal, not quickly vanishing set A_d obtained by a proper scaling of the elements of the set $A_d(\sin) = \{\sqrt{2}\,\sin(\pi v \cdot x)\colon v \in \mathbb{N}_+^d\}$. For $d, r \in \mathbb{N}_+$, set

$$A_{d,r} = \left\{\alpha_v(.)\colon v \in \{1, \ldots, r\}^d\right\} \subset \left(\mathcal{L}_2([0,1]^d), \|.\|_2\right)$$

where

$$\alpha_v(x) = c_v \sin(\pi v \cdot x)\colon [0,1]^d \to \mathbb{R}$$

$v = (v_1, \ldots, v_d) \in \mathbb{R}_+^d$, and

$$c_v = \frac{d\sqrt{2}}{\left\lceil \sum_{k=1}^d v_k \right\rceil}.$$

Let $A_d = \bigcup_{r \in \mathbb{N}_+} A_{d,r}$. We shall show that $A_d \subseteq B_{2d\sqrt{2}}(\|.\|_{H_d})$, and that A_d is not quickly vanishing with respect to d.

We first verify that $A_d \subseteq B_{2d\sqrt{2}}(\|.\|_{H_d})$. By Proposition 11

$$\|\sin(\pi v \cdot x)\|_{H_d} \le V\left(\sin(\pi t), \left[0, \sum_{k=1}^d v_k\right]\right) \le 2\left\lceil \sum_{k=1}^d v_k \right\rceil.$$

Thus, for any $\alpha_v \in A_d$, $\|\alpha_v\|_{H_d} \leq 2d\sqrt{2}$, and hence by Proposition 3 iii), $\|.\|_{H_d} \leq 2d\sqrt{2}\|.\|_{A_d}$. Finally, by Proposition 6 i) and iii), we obtain

$$d_n(H_d) = d_n\left(B_1(\|.\|_{H_d})\right) \geq \frac{1}{2d\sqrt{2}}\, d_n(A_d).$$

Reindex A_d as $\{\alpha_i : i \in \mathbb{N}_+\}$ using a linear ordering of $\mathbb{N}_+^d$ such that the sequence $\{\|\alpha_i\|_2 : i \in \mathbb{N}_+\}$ is nonincreasing and, for all $r \in \mathbb{N}_+$, α_{r^d} corresponds to $\alpha_{(r,\ldots,r)}$. As

$$\|\alpha_{r^d}\|_2 = \left\| \frac{d\sqrt{2}}{dr} \sin(\pi v \cdot x) \right\|_2 = \frac{\sqrt{2}}{r} \|\sin(\pi v \cdot x)\|_2 = \frac{1}{r}$$

A_d is not quickly vanishing with respect to d. Hence, by Proposition 9, for all positive integers n we get

$$d_n(H_d) = d_n\left(B_1(\|.\|_{H_d})\right) \geq \frac{1}{2d\sqrt{2}}\, d_n(A_d) \geq \frac{1}{4d\sqrt[d]{m_n}}$$

where

$$m_n = \min\left\{m \in \mathbb{N}_+ : (2n \leq m)\,(\exists r \in \mathbb{N}_+)\,(m = r^d)\right\}. \quad \square$$

The following corollary shows that, for $n = r^d/2$ for some integer r and for any n-dimensional subspace X_n of $(\mathcal{L}_2([0,1]^d), \|.\|_2)$, there exists a half-space of $[0,1]^d$ that achieves the lower bound from Theorem 2, i.e., its characteristic function cannot be approximated by an element from X_n within an error smaller than $1/(4d\sqrt[d]{2n})$.

Corollary 4: For all positive integers d, n, and any n-dimensional subspace X_n of $(\mathcal{L}_2([0,1]^d), \|.\|_2)$, there exists a characteristic function χ of a half-space of $[0,1]^d$ such that

$$\|\chi - X_n\|_2 \geq \frac{1}{4d\sqrt[d]{m_n}}$$

where

$$m_n = \min\left\{m \in \mathbb{N}_+ : (2n \leq m)\,(\exists r \in \mathbb{N}_+)\,(m = r^d)\right\}.$$

In particular, if $2n = r^d$ for some integer r, then

$$\|\chi - X_n\|_2 \geq \frac{1}{4d\sqrt[d]{2n}}.$$

Proof: By Theorem 2, for any n-dimensional subspace X_n of $(\mathcal{L}_2([0,1]^d), \|.\|_2)$, we have

$$\sup_{\zeta \in H_d} \|\zeta - X_n\| = \sup_{\zeta \in H_d} e_{X_n}(\zeta) \geq \frac{1}{4d\sqrt[d]{m_n}}.$$

As H_d is compact in $(\mathcal{L}_2([0,1]^d), \|.\|_2)$ (see, e.g., [29]) and the error functional e_{X_n} is continuous (see, e.g., [19, p. 391]), the supremum of e_{X_n} on H_d is achieved at some χ. $\square$

As variation with respect to half-spaces is bounded from below by variation with respect to perceptrons with any sigmoidal activation function σ (see Proposition 10 ii)), we have

$$d_n(B_1(\|.\|_{P_d(\sigma)})) \geq d_n(B_1(\|.\|_{H_d})).$$

Hence, Theorem 2 can be extended to include all sigmoidal perceptrons.

Corollary 5: Let d and n be positive integers and σ any sigmoidal function. Then in $(\mathcal{L}_2([0,1]^d), \|.\|_2)$

$$d_n\left(P_d(\sigma)\right) = d_n\left(B_1(\|.\|_{P_d(\sigma)})\right) \geq d_{m_n}\left(P_d(\sigma)\right) \geq \frac{1}{4d\sqrt[d]{m_n}}$$

where

$$m_n = \min\left\{m \in \mathbb{N}_+ : (2n \leq m)\,(\exists r \in \mathbb{N}_+)\,(m = r^d)\right\}.$$

In particular, for $2n = r^d$ for some integer r

$$d_n\left(P_d(\sigma)\right) = d_n\left(B_1(\|.\|_{P_d(\sigma)})\right) \geq \frac{1}{4d\sqrt[d]{2n}}.$$

As the number of free parameters in a perceptron network with k hidden units is $k(d+2)$, we have to compare $d_{k(d+2)}(B_1(\|.\|_{P_d(\sigma)}))$ with $\delta_{P_d(\sigma),k}(B_1(\|.\|_{P_d(\sigma)}))$. From Corollaries 1 and 5 we obtain, for σ nondecreasing sigmoidal, the following estimates:

$$\begin{aligned} d_{k(d+2)}\left(B_1(\|.\|_{P_d(\sigma)})\right) &= d_{k(d+2)}\left(P_d(\sigma)\right) \\ &\geq \frac{1}{4d\sqrt[d]{2k}}\,\frac{1}{\sqrt[d]{d+2}} \end{aligned}$$

$$\delta_{P_d(\sigma),k}(B_1(\|.\|_{P_d(\sigma)})) \leq \frac{1}{\sqrt{k}}.$$

ACKNOWLEDGMENT

The authors are grateful to the Associate Editor Prof. G. Lugosi (Pompeu Fabra University) for his many useful comments. They also wish to thank Prof. P. C. Kainen (Georgetown University) and Prof. S. Giulini (University of Genoa) for helpful discussions.

REFERENCES

[1] R. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton Univ. Press, 1957.

[2] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific, 1996.

[3] T. Parisini, M. Sanguineti, and R. Zoppoli, "Nonlinear stabilization by receding-horizon neural regulators," *Int. J. Contr.*, vol. 70, no. 3, pp. 341–362, 1998.

[4] T. J. Sejnowski and C. R. Rosenberg, "Parallel networks that learn to pronounce English text," *Complex Syst.*, vol. 1, no. 1, pp. 145–168, 1987.

[5] R. Zoppoli, M. Sanguineti, and T. Parisini, "Approximating networks and extended Ritz method for the solution of functional optimization problems," *J. Optimiz. Theory and Applic.*, vol. 112, no. 2, Feb. 2002.

[6] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inform. Theory*, vol. 39, pp. 930–945, May 1993.

[7] P. C. Kainen, V. Kůrková, and A. Vogt, "Approximation by neural networks is not continuous," *Neurocomput.*, vol. 29, no. 1–3, pp. 47–56, 1999.

[8] ——, "Geometry and topology of continuous best and near best approximations," *J. Approx. Theory*, vol. 105, no. 2, pp. 252–262, 2000.

[9] R. A. DeVore and G. G. Lorentz, "Constructive approximation," in *Grundlehren der Mathematischen Wissenschaften*. Berlin, Germany: Springer-Verlag, 1993, vol. 303.

[10] R. A. DeVore, B. Jawerth, and V. Popov, "Compression of wavelet decompositions," *Amer. J. Math.*, vol. 114, no. 4, pp. 737–785, 1992.

[11] R. A. DeVore and V. N. Temlyakov, "Nonlinear approximation by trigonometric sums," *J. Fourier Anal. Applic.*, vol. 2, no. 1, pp. 29–48, 1995.

[12] V. Kůrková and M. Sanguineti, "Bounds on rates of variable-basis and neural network approximation," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2659–2665, Sept. 2001.

[13] L. K. Jones, "A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training," *Ann. Statist.*, vol. 20, no. 1, pp. 608–613, 1992.

[14] L. Breiman, "Hinging hyperplanes for regression, classification, and function approximation," *IEEE Trans. Inform. Theory*, vol. 39, pp. 993–1013, May 1993.

[15] A. R. Barron, "Neural net approximation," in *Proc. 7th Yale Workshop on Adaptive and Learning Systems*, K. S. Narendra, Ed. New Haven, CT: Yale Univ. Press, 1992, pp. 69–72.

[16] V. Kůrková, P. C. Kainen, and V. Kreinovich, "Estimates of the number of hidden units and variation with respect to half-spaces," *Neural Networks*, vol. 10, no. 6, pp. 1061–1068, 1997.
[17] A. Friedman, *Foundations of Modern Analysis*. New York: Dover, 1982.
[18] A. N. Kolmogorov and S. V. Fomin, *Introductory Real Analysis*. New York: Dover, 1975.
[19] I. Singer, *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*. Berlin Heidelberg, Germany: Springer-Verlag, 1970.
[20] A. N. Kolmogorov, "Über die beste Annäherung von Funktionen einer gegebenen Funktionenklasse," *Ann. Math.*, vol. 37, no. 1, pp. 107–110, 1936. English translation: "On the best approximation of function of a given class," in *Selected Works of A. N. Kolmogorov*, vol. 1, V. M. Tikhomirov, Ed. Norwell, MA: Kluwer, 1991, pp. 202–205.
[21] A. Pinkus, *n-Widths in Approximation Theory*. Berlin Heidelberg, Germay: Springer-Verlag, 1985.
[22] 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. Cambridge, MA: Birkhäuser, 1997, pp. 261–270.
[23] V. Kůrková, P. Savický, and K. Hlaváčková, "Representations and rates of approximation of real-valued Boolean functions by neural networks," *Neural Networks*, vol. 11, no. 4, pp. 651–659, 1998.
[24] G. Pisier, "Remarques sur un resultat non publié de B. Maurey," in *Séminaire d'Analyse Fonctionelle*. Palaiseau, France: École Polytech., Centre de Mathématiques, 1980–1981, vol. I, no.12.
[25] V. Kůrková, "Incremental approximation by neural networks," in *Complexity: Neural Network Approach*, V. Kůrková, M. Kárný, and K. Warwick, Eds. London, U.K.: Springer-Verlag, 1998, pp. 177–188.
[26] H. N. Mhaskar and C. A. Micchelli, "Dimension-independent bounds on the degree of approximation by neural networks," *IBM J. Res. Devel.*, vol. 38, no. 3, pp. 277–283, 1994.
[27] C. Darken, M. Donahue, L. Gurvits, and E. Sontag, "Rate of approximation results motivated by robust neural network learning," in *Proc. 6th Annu. ACM Conf. Computational Learning Theory*, Santa Cruz, CA, 1993, pp. 303–309.
[28] F. Girosi, "Approximation error bounds that use VC-bounds," in *Proc. Int. Conf. Artificial Neural Networks ICANN'95*, vol. 1, Paris, France, 1995, pp. 295–302.
[29] L. Gurvits and P. Koiran, "Approximation and learning of convex superpositions," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 161–170, 1997.
[30] Y. Makovoz, "Uniform approximation by neural networks," *J. Approx. Theory*, vol. 95, no. 2, pp. 215–228, 1998.
[31] ——, "Random approximants and neural networks," *J. Approx. Theory*, vol. 85, no. 1, pp. 98–109, 1996.
[32] V. Kůrková and M. Sanguineti, "Tightness of upper bounds on rates of neural-network approximation," in *Proc. Int. Conf. Artificial Neural Networks and Genetic Algorithms*, V. Kůrková, R. Neruda, and M. Kárný, Eds. Vienna, Austria: Springer-Verlag, 2001, pp. 41–45.
[33] G. C. Lorentz, *Approximation of Functions*. New York: Chelsea , 1986.
[34] A. R. Gallant and H. White, "There exists a neural network that does not make avoidable mistakes," in *Proc. II IEEE Int. Conf. Neural Networks*. San Diego, CA: SOS, 1988, vol. I, pp. 657–664.
[35] R. Courant, *Differential and Integral Calculus*. New York: Wiley-Interscience, 1988, vol. II.
[36] P. C. Kainen, "Utilizing geometric anomalies of high dimension: When complexity makes computation easier," in *Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality*, K. Warwick and M. Kárný, Eds. Cambridge, MA: Birkhäuser, 1997, pp. 283–294.