Dimension-Independent Approximation by Neural Networks and Its Comparison with Linear Approximation

Kůrková, Věra
1999

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Dimension-independent approximation by neural networks and its comparison with linear approximation

Věra Kůrková Marcello Sanguineti

Technical report No. 789

October 1999

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic
phone: +4202 6605 3231 fax: +4202 8585789
e-mail: vera@uivt.cas.cz, marcello@dist.unige.it

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Dimension-independent approximation by neural networks and its comparison with linear approximation

Věra Kůrková Marcello Sanguineti[1]

Technical report No. 789
October 1999

Abstract

In this paper, we compare rates of approximation achievable using any linear approximation method with rates of approximation by neural networks. We compare such rates in terms of the worst-case errors in approximation by n-dimensional linear subspaces and neural networks with n hidden units in the last hidden layer. We develop a general framework for such comparisons for sets of multivariate functions either computable by various types of network computational units or approximable with dimension-independent rates by networks with such units. Applying this approach to perceptron networks, we derive lower bounds on the worst-case errors in linear approximation of sets of functions computable by perceptrons with periodic and sigmoidal activation functions.

Keywords

linear and nonlinear approximation, Kolmogorov width, dimension-independent approximation, curse of dimensionality, variation with respect to a set of functions, complexity of neural networks, one-hidden-layer perceptron networks

[1]Department of Communications, Computer, and System Sciences, University of Genova, Via Opera Pia 13, 16145 Genova, Italy

1 Introduction

Since the work of Weierstrass and Chebyshev in the 19th century, approximation by polynomials and rational functions has developed into a unifying framework, and the difference between linear and rational approximation became apparent. Later, in addition to rational approximation, other types of nonlinear approximating families, like splines and exponential sums, were investigated. Renewal of interest in artificial neural networks in the early 1980s lead to many successful applications, equalling or exceeding other approaches. Since feedforward neural networks compute parametrized nonlinear families of functions of a different type than all previously studied families, they form a new branch in the field of nonlinear approximation.

The theoretical investigation of neural networks as nonlinear approximators has been mostly focused on questions of existence of an arbitrarily close approximation and estimates of its rates in dependence on network complexity, while the difference between linear and neural network approximation has remained less understood. The first result in this direction is Barron's [1] comparison of rates achievable using linear and neural network approximation. He described sets of multivariable functions, for which approximation by one-hidden layer sigmoidal perceptron networks is "dimension-independent" (is bounded from above by $\mathcal{O}(\frac{1}{\sqrt{n}})$, where n is the number of hidden units), while the accuracy of approximation achievable using any linear method depends on the dimension (is bounded from below by $\mathcal{O}(\frac{1}{d\sqrt[d]{n}})$, where d is the number of variables of the function to be approximated). Kainen, Kůrková and Vogt ([10], [11], [12]) have initiated comparison of properties of projections (best approximation operators) in linear and neural network approximation. They have shown that many useful properties of linear approximators like uniqueness, homogeneity and continuity are not satisfied by neural network approximators, suggesting that this loss might allow improved rates of approximation (since arguments proving slow rates of linear approximators are based on these properties).

In this paper, we extend the work of Barron [1] on comparisons of rates of approximations. We develop a general framework for such comparisons for two kinds of sets of multivariable functions: sets of functions computable by various kinds of computational units and sets approximable with dimension-independent rates by networks with such units. We compare the worst-case approximation errors, formalized in terms of the Kolmogorov n-width (infimum of deviations from n-dimensional linear subspaces) and the deviation from the union of n-dimensional subspaces spanned by computational units (corresponding to n-hidden-layer neural networks).

Applying this general approach to perceptron networks, we derive lower bounds on the worst-case error in linear approximation of sets of functions computable by perceptrons with various types of activations. We show that for some periodic activation functions such sets cannot be efficiently approximated using linear methods, since no increase of the dimension of the linear approximating sets can decrease the worst-case error under certain constant bound. For sigmoidal perceptrons, such error is bounded from below by $\mathcal{O}(\frac{1}{d\sqrt[d]{2n}})$, where n corresponds to the dimension of the linear approximating space and d is the number of variables.

The paper is organized as follows. Section 2 contains basic concepts and notations

concerning approximation in normed linear spaces and feedforward neural networks. Section 3 describes dimension-independent rates of approximation by such networks in terms of balls in certain norms tailored to computational units. To compare such rates with those achievable using linear approximation schemes, in Section 4 we investigate methods of estimation of Kolmogorov width of balls in norms of this type. In Section 5, the tools developed in previous sections are applied to perceptron networks. Section 6 contains a brief discussion.

2 Preliminaries

2.1 Approximation in normed linear spaces

Let $\mathcal{R}$ denote the set of real numbers, $\mathcal{R}_+$ the set of nonnegative real numbers, $\mathcal{N}$ the set of natural numbers and $\mathcal{N}_+$ the set of positive integers.

In this paper, we assume that approximating functions as well as functions to be approximated are from some *real normed linear space* $(X, \|.\|)$; for brevity, in the following the term "real" will be omitted. Whenever there is no ambiguity on the norm $\|.\|$, we will write X instead of $(X, \|.\|)$. When X is finite-dimensional, we denote by $dim\, X$ its dimension. By $B_r(\|.\|)$ is denoted the ball in X of radius r with respect to the norm $\|.\|$, i.e. $B_r(\|.\|) = \{f \in X; \|f\| \leq r\}$.

Standard choices of a normed linear space are the space $(\mathcal{C}(K), \|.\|_{\mathcal{C}})$ of all continuous functions on some compact subset K of $\mathcal{R}^d$ (often the d-dimensional cube $[0,1]^d$) with the supremum norm denoted by $\|.\|_{\mathcal{C}}$ and defined by $\|f\|_{\mathcal{C}} = \sup_{x \in K} |f(x)|$, and $(\mathcal{L}_p(K), \|.\|_p)$, where $\mathcal{L}_p(K) = \{f : K \to \mathcal{R}; (\int f^p d\lambda)^{\frac{1}{p}} \leq \infty\}$ for $p \in [1, \infty)$ (λ denotes the Lebesgue measure, but other measures may be used, too) with $\mathcal{L}_p$-norm defined by $\|f\|_p = (\int f^p d\lambda)^{\frac{1}{p}}$.

Some properties of approximation can be formulated for normed linear spaces satisfying certain conditions, e.g. for Banach or Hilbert spaces. Recall that a *Banach space* is a normed linear space that is complete and that a *Hilbert space* is a Banach space with a norm generated by an inner product, i.e. $\|f\| = \sqrt{f \cdot f}$ (see e.g. Friedman [6]).

If G is a subset of a normed linear space $(X, \|.\|)$, then G^0 denotes the set of its *normalized* elements, i.e. $G^0 = \{g^0 = \frac{g}{\|g\|}, g \in G\}$. The *closure* of G is denoted by $cl\, G$ and defined by $cl\, G = \{f \in X; (\forall \varepsilon > 0)(\exists g \in G)(\|f - g\| < \varepsilon)\}$. The *interior* of G is denoted by $int\, G$ and defined by $int\, G = \{g \in G; (\exists \epsilon > 0)\, (\forall f \in X)\, (\|f - g\| < \epsilon) \Rightarrow f \in G\}$, and the *boundary* as $\partial G = cl\, G - int\, G$. A normed linear space $(X, \|.\|)$ is called *separable* if it contains a countable dense subset.

For $c \in \mathcal{R}$ we denote $c\, G = \{c\, g; g \in G\}$ and $G(c) = \{wg; g \in G, w \in \mathcal{R}\, \&\, |w| \leq c\}$. G is called *homogeneous* if $c\, G = G$ for all $c \in \mathcal{R}$. If $G = G(1)$, then G is called *balanced*. $G(1)$ is called the *balanced hull of* G.

The *Minkowski functional* $\nu_G : X \to \mathcal{R}_+ \cup \{\infty\}$ of a subset G of a normed linear space $(X, \|.\|)$ is defined as $\nu_G(f) = \inf\{c \in \mathcal{R}_+; \frac{f}{c} \in G\}$. Recall that when G is balanced and convex, then ν_G is a norm on $\{f \in X; \nu_G(f) < \infty\}$. When, in addition to these two properties, G is also closed, then the unit ball in ν_G is closed in the topology induced on X by $\|.\|$.

The *linear span* of G, denoted by $span\, G$, is the set of all linear combinations of elements of G, i.e. $span\, G = \{\sum_{i=1}^{n} w_i g_i; w_i \in \mathcal{R}, g_i \in G, n \in \mathcal{N}_+\}$; $span_n G$ denotes the set of all linear combinations of at most n elements of G, i.e. $span_n G = \{\sum_{i=1}^{n} w_i g_i; w_i \in \mathcal{R}, g_i \in G\}$. *Convex hull* of G, denoted by $conv\, G$, is the set of all convex combinations of its elements, i.e. $conv\, G = \{\sum_{i=1}^{n} a_i g_i; a_i \in [0,1], \sum_{i=1}^{n} a_i = 1, n \in \mathcal{N}_+\}$; $conv_n G$ denotes the set of all convex combinations of at most n elements of G, i.e. $conv_n G = \{\sum_{i=1}^{n} a_i g_i; a_i \in [0,1], \sum_{i=1}^{n} a_i = 1, g_i \in G\}$. A set G is called *convex* if $G = conv\, G$.

The theory of approximation investigates properties of error functionals measuring the accuracy of approximation as a distance from a set of approximating functions. An error functional $e_M : X \to \mathcal{R}_+$ of a subset M of $(X, \|.\|)$ is defined as $e_M(f) = \|f - M\| = \inf_{g \in M} \|f - g\|$. Recall that, for each normed linear space $(X, \|.\|)$ and each of its subsets M, e_M is continuous but it does not need to be linear (even for M finite-dimensional subspace of X, see e.g. Singer [25]).

Suitability of an approximating set for approximation of functions from a given set (often defined in terms of a bound on some norm different from the one used to measure the accuracy of approximation) can be characterized by the *worst-case error*, which is mathematically formalized by the concept of deviation of a set Y of functions to be approximated from the approximating set M. The *deviation of Y from M* is defined as

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

whenever there is no ambiguity about the normed linear space under consideration, we will write $\delta(Y, M)$ instead of $\delta(Y, M, (X, \|.\|))$. Note that deviation describes the size of the smallest neighbourhood of M containing Y (if $\delta(Y, M) = \varepsilon$, then ε is the infimum of all the real numbers, for which $Y \subseteq U_\varepsilon(M) = \{f \in X; \|f - M\| \leq \varepsilon\}$).

The following properties of deviation follow directly from its definition and from continuity of e_M.

Proposition 2.1 *Let $(X, \|.\|)$ be a normed linear space, Y and M be its subsets. Then*
(i) $\delta(Y, M) = \delta(cl\, Y, M)$;
(ii) when M is homogeneous, then for every $c \in \mathcal{R}$ $\delta(cY, M) = |c|\delta(Y, M)$;
(iii) when M is convex, then $\delta(Y, M) = \delta(conv\, Y, M)$.

To describe a theoretical lower bound on linear approximation, Kolmogorov [13] investigated infimum of deviations over all n-dimensional subspaces of X. He introduced the concept of *n-width* (which became later called *Kolmogorov n-width*) of a set Y, defined by

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

where the infimum is taken over all n-dimensional subspaces X_n of X. The following proposition summarizes basic properties of Kolmogorov n-width that can be easily verified (see also Lorentz [20], Pinkus [23, p. 10]).

Proposition 2.2 *Let $(X, \|.\|)$ be a normed linear space and Y be its subset. Then for all positive integers n*
(i) $d_n(Y) \geq d_{n+1}(Y)$;
(ii) $d_n(cl\, Y) = d_n(Y)$;
(iii) for every $c \in \mathcal{R}$ $d_n(cY) = |c|\, d_n(Y)$;
(iv) $d_n(conv\, Y) = d_n(Y)$;
(iv) $d_n(Y^0) \inf_{f \in Y} \|f\| \leq d_n(Y) \leq d_n(Y^0) \sup_{f \in Y} \|f\|$;
(v) if $Y_1 \subseteq Y_2 \subseteq X$, *then* $d_n(Y_2) - \delta(Y_1, Y_2) \leq d_n(Y_1) \leq d_n(Y_2)$.

Thus the Kolmogorov width of a set is equal to the Kolmogorov width of its closed, convex, balanced hull. Since each convex balanced set determines a norm on X, in which it forms the unit ball (via the Minkowski functional), 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.

2.2 Rates of approximation

Rates of approximation characterize the trade-off between the accuracy of approximation and the complexity of the approximating function. When approximating functions are from a parametrized family, then their complexity can be measured by the length of a parameter vector (corresponding, for example, to the degree of a polynomial or a rational function, the number of knots in a spline, the number of hidden units in a neural network). Such a parametrized family can be represented as a sequence of sets of functions (often nested), with parameter vectors of increasing length. In traditional approximation schemes (like polynomials and fixed series expansions), these sets are finite-dimensional subspaces of increasing dimensionality.

Let $\{M_n; n \in \mathcal{N}_+\}$ be a sequence of nested subsets of a normed linear space $(X, \|.\|)$, then the *rate of approximation* of $f \in X$ by $\{M_n; n \in \mathcal{N}_+\}$ is measured by the speed of decrease of $e_{M_n}(f)$. The rate of approximation of a subset Y of X is characterized by the decrease of the worst-case error, corresponding to the deviation $\delta(Y, M_n) = \sup_{f \in Y} e_{M_n}(f) = \sup_{f \in Y} \|f - M_n\|$. When $\bigcup_{n \in \mathcal{N}_+} M_n$ is dense in X, then for each $f \in X$ the sequence $\{e_{M_n}(f); n \in \mathcal{N}_+\}$ converges to 0, but for practical applications this convergence has to be sufficiently fast to guarantee a desired accuracy of approximation for n small enough so that all functions from M_n are implementable.

In the case of functions of d variables, it might happen that the deviation is of order $\mathcal{O}\left(\frac{1}{\sqrt[d]{n}}\right)$. It means that to achieve an accuracy within ε, there are required approximating functions of the complexity of order $\left(\frac{1}{\varepsilon}\right)^d$. Such an exponentional dependence of complexity on the number of variables is called the *curse of dimensionality*. When the complexity of approximating functions does not depend on the number of variables d, then the approximation scheme is called *dimension-independent*. For example, in $(\mathcal{L}_p([0,1]^d), \|.\|_p)$ the Kolmogorov widths of balls in Sobolev norms of fixed order exhibit the curse of dimensionality, while when the order is appropriately increasing with d, the Kolmogorov width of such balls is dimension-independent (see e.g. Pinkus [23, pp. 232-233]).

2.3 One-hidden-layer neural networks

Feedforward neural networks compute parametrized sets of functions depending on both the type of computational units and the type of their interconnections. *Computational units* compute functions of two vector variables: an *input vector* and a *parameter vector*. Generally, they compute functions of the form $\phi : \mathcal{R}^p \times \mathcal{R}^d \to \mathcal{R}$, where ϕ corresponds to the type of the unit, p and d to the dimension of the *parameter* and the *input space*, respectively.

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

$$\sum_{i=1}^{n} w_i \phi(\mathbf{a}_i, .),$$

where $\mathbf{a}_i \in \mathcal{R}^p$.

Denote by $G_\phi = \{\phi(\mathbf{a}, \cdot); \mathbf{a} \in \mathcal{R}^p\}$ a parametrized set of functions corresponding to a type of computational unit ϕ. Then a ϕ-network with n hidden units can generate as its input/output functions all elements of $span_n\, G_\phi$, which is the union of all n-dimensional subspaces spanned by n-tuples of elements of G_ϕ. Thus in the case of neural networks, the approximating functions are members of *unions of finite dimensional subspaces* generated by hidden unit functions.

Standard types of hidden units are perceptrons. A *perceptron* with an *activation function* $\psi : \mathcal{R} \to \mathcal{R}$ computes functions of the form $\phi((\mathbf{v}, b), \mathbf{x}) = \psi(\mathbf{v} \cdot \mathbf{x} + b) : \mathcal{R}^{d+1} \times \mathcal{R}^d \to \mathcal{R}$, where $\mathbf{v} \in \mathcal{R}^d$ is an *input weight vector* and $b \in \mathcal{R}$ is a *bias*.

Let J be a compact (i.e. closed and bounded) subset of $\mathcal{R}$ (the standard choice is $J = [0, 1]$). By $P_d(\psi, J) = \{f : J^d \to \mathcal{R}; f(\mathbf{x}) = \psi(\mathbf{v} \cdot \mathbf{x} + b), \mathbf{v} \in \mathcal{R}^d, b \in \mathcal{R}\}$ is denoted the set of functions on J^d computable by ψ-perceptrons (when it is clear from the context which J is considered, we will write only $P_d(\psi)$ instead of $P_d(\psi, J)$).

So $span_n P_d(\psi, J)$ denotes the set of functions on J^d computable by ψ-perceptron networks with n hidden units, and $span\, P_d(\psi, J)$ represents the set of functions computable by such networks with any number of hidden units.

3 Dimension-independent approximation

3.1 Approximation from unions of finite dimensional subspaces

To derive tools for the estimation of rates of approximation by one-hidden-layer neural networks, we investigate approximation properties of sets of functions of the form $span_n\, G$, where G is any subset of a normed linear space $(X, \|.\|)$. This approximation scheme includes approximation by multilayer feedforward networks with a single linear output and n hidden units in the last hidden layer. In particular, it includes ϕ-networks.

Rates of approximation from $span_n G$ are measured by the *deviation from $span_n G$* that we denote by $\delta_{G,n}$, i.e.

$$\delta_{G,n}(Y) = \delta_{G,n}(Y, (X, \|.\|)) = \delta(Y, span_n G, (X, \|.\|)).$$

The following proposition states the basic properties of $\delta_{G,n}$ that follow directly from Proposition 2.1 (notice that $span_n G$ is homogeneous).

Proposition 3.1 *Let $(X, \|.\|)$ be a normed linear space, Y and G be its subsets. Then for any positive integer n*
(i) $\delta_{G,n}(Y) \geq \delta_{G,n+1}(Y)$;
(ii) $\delta_{G,n}(Y) = \delta_{G,n}(cl\, Y)$;
(iii) for every $c \in \mathcal{R}$ $\delta_{G,n}(c\, Y) = |c|\delta_{G,n}(Y)$.

Since $span_n G$ is not a linear subspace, many of the convenient properties (like uniqueness, continuity, homogeneity etc.) of best approximation operators used as tools in linear approximation theory are no longer valid (see Kainen, Kůrková and Vogt [10], [11], [12]). On the other hand, since the union of all linear subspaces spanned by n-tuples of elements of a given set G is often much bigger than any single n-dimensional subspace of X, for some sets of functions, rates of nonlinear approximation by $span_n G$ might be considerably better than rates achievable using any linear approximating family. Since we look for sets Y for which $d_n(Y$ is bigger than $\delta_{G,n}(Y)$, we will investigate upper bounds on $\delta_{G,n}$ and lower bounds on $d_n(Y)$.

3.2 Variation with respect to a set of functions

Sets of multivariable functions with dimension independent upper bounds on the deviations from sets of the form $span_n G$ can be described in terms of norms tailored to sets G.

Let G be a subset of a normed linear space $(X, \|.\|)$. Then *G-variation (variation with respect to the set G)* denoted by $\|.\|_G$ is defined as the Minkowski functional of the closed convex balanced set $cl\, conv\, G(1)$, i.e.

$$\|f\|_G = \inf\left\{c \in \mathcal{R}_+; \frac{f}{c} \in cl\, conv\, G(1)\right\} = \inf\{c \in \mathcal{R}_+; f \in cl\, conv\, G(c)\}.$$

Thus G-variation is a norm on the subspace $\{f \in X; \|f\|_G < \infty\} \subseteq X$. G-variation was defined by Kůrková [16] as an extension of Barron's [2] concept of variation with respect to half-spaces. The following proposition states the basic properties of G-variation.

Proposition 3.2 *Let $(X, \|.\|)$ be a normed linear space, G and F be its subsets. Then*
(i) for all $f \in X$ $\|f\| \leq \|f\|_G \sup_{g \in G} \|g\|$;
(ii) if $f \in span\, G$, then $\|f\|_G = \min\{\sum_{i=1}^m |w_i|; f = \sum_{i=1}^m w_i g_i, m \in \mathcal{N}_+, g_i \in G, w_i \in \mathcal{R}\}$;
(iii) $\|.\|_G \leq c\|.\|_F$ if and only if for all $h \in F$ $\|h\|_G \leq c$.

Proof. (i) and (ii) follow immediately from the definition of G-variation. To verify (iii), set $b = \|f\|_F$. Let $f = \lim_{m\to\infty} f_m$ in $\|.\|$, where for all $m \in \mathcal{N}_+$ $f_m \in conv\, F(b)$. Then $f_m = \sum_{i=1}^{n_m} w_{m,i} h_{m,i}$, where $\sum_{i=1}^{n_m} |w_{m,i}| \leq b$ and $h_{m,i} \in F$. Since all $h_{m,i} \in cl\, conv\, G(c)$, we have $f_{m,i} \in cl\, conv\, G(b\, c)$ and so $\|f\|_G \leq c\, b = c\|f\|_F$. ∎

When X is finite-dimensional, all norms on X are equivalent, i.e. they determine the same topology, and thus the concept of G-variation does not depend on the norm on X. In an infinite-dimensional case, when G-variation depends on the choice of a norm $\|.\|$ on X; we will assume that it is clear from the context with respect to which norm G-variation is considered.

Note that 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 up to a constant with total variation (see Barron [2], Kůrková, Kainen and Kreinovich [18]).

Let A be an orthogonal basis of a separable Hilbert space $(X, \|.\|)$, then l_1-*norm with respect to* A is defined as $\|f\|_{1,A} = \sum_{\alpha \in A} |f \cdot \alpha|$. The following proposition describes the relationship between A-variation and the l_1-norm with respect to A.

Proposition 3.3 *Let $(X, \|.\|)$ be a separable Hilbert space and A be its orthogonal basis, then $\|.\|_A \leq \|.\|_{1,A}$; when $\|.\|$ and $\|.\|_A$ are equivalent, then $\|.\|_A = \|.\|_{1,A}$.*

Proof. First, check that $\|.\|_A \leq \|.\|_{1,A}$. Let $A = \{\alpha_i; i \in \mathcal{N}_+\}$. Then every $f \in X$ can be represented as $\sum_{i=1}^{\infty} (f \cdot \alpha_i)\alpha_i$. For $m \in \mathcal{N}_+$ set $f_m = \sum_{i=1}^{m} (f \cdot \alpha_i)\, \alpha_i$. If $b = \|f\|_{1,A}$, then for all $m \in \mathcal{N}_+$ $f_m \in conv\, A(b)$. $f = \lim_{m \to \infty} f_m$ in $\|.\|$, and so f is in the closure of $conv\, A(b)$ with respect to $\|.\|$. Hence $\|f\|_A \leq b = \|f\|_{1,A}$.

To verify that $\|.\|_A \geq \|.\|_{1,A}$ for $\|.\|$ and $\|.\|_A$ equivalent, let $f_m = \sum_{i=1}^{m} (f \cdot \alpha_i)$, and $b = \lim_{m \to infty} \|f_m\|_{1,A}$. Since $\|f_m\|_{1,A} = \|f_m\|$ (see Kůrková, Savický and Hlaváčková [19, p.654]), $\lim_{m \to \infty} f_m = f$ in $\|.\|$ and $\lim_{m \to infty} \|f_m\|_A = b$. When $\|.\|$ and $\|.\|_A$ are equivalent, the set $U = \{h \in X; \|h\|_A < b\}$ is open in $\|.\|$, and so $\|f\|_A \geq b = \|f\|_{1,A}$. ■

Thus when A is an orthogonal basis of X, then the unit ball in A-variation contains the unit ball in the l_1-norm with respect to A. For example, the unit ball in variation with respect to the Fourier basis contains the unit ball in the l_1-norm with respect to this basis.

Some insight into properties of sets of multivariable functions that can be approximated by neural networks with dimension-independent rates were obtained by Jones [9] and Barron [1]. Using the concept of G-variation, Kůrková ([16], [17]) reformulated Barron's [1] improvement of Jones' result [9] in the following way.

Theorem 3.4 *Let $(X, \|.\|)$ be a Hilbert space and G be its subset. Then for every $f \in X$ and for every positive integer n*

$$\|f - span_n G\|^2 \leq \frac{(s_G \|f\|_G)^2 - \|f\|^2}{n},$$

where $s_G = \sup_{g \in G} \|g\|$.

Since $span_n G = span_n G^0$, Theorem 3.4 implies

$$\|f - span_n G\|^2 \leq \frac{\|f\|_{G^0}^2 - \|f\|^2}{n}.$$

As an immediate corollary we get a description of sets of multivariable functions that can be approximated by $span_n G$ with dimension-independent rates.

Corollary 3.5 *Let $(X, \|.\|)$ be a Hilbert space and G be its subset. Then for every positive integer n*

$$\delta_{G,n}(B_1(\|.\|_G)) \leq \frac{s_G}{\sqrt{n}},$$

where $s_G = \sup_{g\in G} \|g\|$. In particular,

$$\delta_{G,n}(B_1(\|.\|_{G^0})) \leq \frac{1}{\sqrt{n}}.$$

Thus balls in G^0-variation can be approximated by elements of $span_n G$ with a rate of approximation that does not depend on the number of variables of functions in X. However, with increasing number of variables the condition of being in the unit ball in G^0-variation becomes more and more constraining (see Kůrková, Savický and Hlaváčková [19] for examples of functions with variation depending exponentially on the number of variables).

Note that estimates of $\delta_{G,n}$ in terms of G-variation are not restricted to Hilbert spaces. Darken et al. [5] extended Jones-Barron's theorem to $\mathcal{L}_p$-spaces for $p \in (1, \infty)$ with a slightly worse rate of approximation (of order $\mathcal{O}(n^{-\frac{1}{q}})$, where $q = \max\left(p, \frac{p}{p-1}\right)$); there also exist extensions for supremum norm (see e.g. Barron [2], Girosi [7], Gurvits and Koiran [8], Kůrková, Savický and Hlaváčková [19]).

4 Kolmogorov width of balls in G-variation

4.1 Basic properties of the Kolmogorov width of balls in G-variation

When for some family of subsets $\{G_d; d \in \mathcal{N}_+\}$ of a family of normed linear spaces $\{(X_d, \|.\|); d \in \mathcal{N}_+\}$, where each X_d consists of functions of d variables, Kolmogorov n-width of balls in G_d-variation is considerably larger than their deviation from $span_n G_d$, then approximation by $span_n G_d$ outperforms any linear approximation scheme. Such better performance is especially remarkable when $d_n(B_1(\|.\|_{G_d^\circ}))$ is of order $\mathcal{O}\left(\frac{1}{\sqrt[d]{n}}\right)$, since by Corollary 3.5 $\delta_{G_d,n}(B_1(\|.\|_{G_d^\circ})) \leq \frac{1}{\sqrt{n}}$.

To describe sets G_d with $\delta_{G_d,n}(B_1(\|.\|_{G_d})$ smaller than $d_n(B_1(\|.\|_{G_d}))$, we will investigate lower bounds on Kolmogorov n-width of balls in variation with respect to a general set of functions. The following proposition summarizes basic properties of Kolmogorov width of such balls that follow from Proposition 2.2, Proposition 3.2 and from the fact that $B_1(\|.\|_G) = cl\, conv\, G(1)$.

Proposition 4.1 *Let $(X, \|.\|)$ be a normed linear space, G and F be its subsets. Then for any positive integer n*
(i) $d_n(B_1(\|.\|_G)) = d_n(G)$;
(ii) if $G \supseteq F$ then $d_n(B_1(\|.\|_G)) \geq d_n(B_1(\|.\|_F))$;
(iii) if $a = \sup_{h\in F} \|h\|_G \leq \infty$, then $d_n(B_1(\|.\|_G)) = d_n(G) \geq \frac{1}{a} d_n(F) = \frac{1}{a} d_n(B_1(\|.\|_F))$.

The first one of these elementary properties has important consequences. In particular, it implies that any estimate of the worst-case error in the linear approximation of the unit ball in G-variation applies also to G itself. Thus a speed of the decrease of $d_n(G)$ can be evaluated using $d_n(B_1(\|.\|_G))$ (to derive a lower bound on the Kolmogorov width of a larger set might be easier).

4.2 Lower bounds in terms of the Bernstein width

As pointed out in Proposition 3.2 (i), it follows directly from the definition of G-variation that, for any subset G of a normed linear space $(X, \|.\|)$, $\|.\| \leq s_G \|.\|_G$, where $s_G = \sup_{g \in G} \|g\|$. Thus the unit ball in $\|.\|$ contains the ball of radius $\frac{1}{s_G}$ in G-variation. When also the unit ball in G-variation contains a ball of some nonzero radius in $\|.\|$ (i.e., it has nonempty interior in the topology induced on X by $\|.\|$), then the norms $\|.\|_G$ and $\|.\|$ are equivalent. In such a case, we can estimate the Kolmogorov width of the unit ball in G-variation from below using the Bernstein width.

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

$$b_n(Y) = \sup_{X_{n+1}} \sup\{r \in \mathcal{R}_+;\ B_r(\|.\|^{X_{n+1}}) \subseteq Y\},$$

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. Pinkus [23]). Notice that when Y is closed, convex and balanced, then for all n $b_n(Y)$ is the diameter of the maximal ball in $\|.\|^{X_{n+1}}$ contained in Y. Thus we can extend the concept of the Bernstein width by defining

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

The following proposition is an easy modification of a lower bound on Kolmogorov width from Lorentz [20, p. 133].

Proposition 4.2 *Let $(X, \|.\|)$ be a Banach space, G be its subset such that $\{f \in X; \|f\|_G < \infty\} = X$. Then for any positive integer n such that $n < \dim X$ $d_n(G) = d_n(B_1(\|.\|) \geq b(B_1(\|.\|_G)) = \inf\{\|f\|; \|f\|_G = 1\}$.*

Proof. Let X_n be an n-dimensional subspace of X and let $h \in X - X_n$. Let $g \in X_n$ be the closest element to h, i.e. $\|h - X_n\| = \|h - g\|$ and set $f = \frac{h-g}{\|h-g\|_G}$. Then $0 \in X_n$ is the closest element to f: indeed, for all $g' \in X_n$ we have $\|f\| \leq \|f - g'\|$ because $\|h - g\| \leq \|h - g - ag'\|$, where $a = \|h - g\|_G$. Since $\|f\|_G = 1$, we have $e_{X_n}(f) = \|f\|$ and by Proposition 4.1 (i) $d_n(G) = d_n(B_1(\|.\|_G)) \geq \|f\| \geq \inf\{\|f\|; f \in B_1(\|.\|_G)\} = b(B_1(\|.\|_G))$. ∎

Since all norms on a finite-dimensional space are equivalent, we can apply Proposition 4.2 to $\mathcal{R}^m$ with the l_2-norm and A any of its orthonormal bases. Then we get $d_n(A) = d_n(B_1(\|.\|_A)) \geq \frac{1}{\sqrt{m}}$ for any $n < m$.

However, when $\|.\|$ and $\|.\|_G$ are not equivalent, this method of estimation of the Kolmogorov width of balls in G-variation gives a trivial lower bound (equal to zero). A more sophisticated method based on the Borsuk antipodality theorem (see e.g. Pinkus

[23]) shows that the Kolmogorov n-width is bounded from below by the Bernstein n-width. More precisely, for every closed, convex, balanced subset Y of a Banach space $(X, \|.\|)$ and for all positive integers n $d_n(Y) \geq b_n(Y)$.

To obtain from this estimate a lower bound on $d_n(G) = d_n(B_1(\|.\|_G))$ larger than the upper bound on $\delta_{G,n}$ guaranteed by Corollary 3.5, $b_n(G)$ must be larger than $\frac{s_G}{\sqrt{n}}$. For example for A countable orthonormal, it is easy to check that $b_n(B_1(\|.\|_A)) = \frac{1}{\sqrt{n+1}}$ when $n < card\, A$ (since $B_1(\|.\|_{1,A}) \cap X_{n+1} = B_1(\|.\|_A) \cap X_{n+1}$). Thus using the Bernstein width as a lower bound on the Kolmogorov width of balls in A-variation for A countable orthonormal, we get the same lower bound, $\frac{1}{\sqrt{n}}$, on $d_n(B_1(\|.\|_A))$ as the upper bound, $\frac{1}{\sqrt{n}}$, on $\delta_{A,n}(B_1(\|.\|_A))$ following from Corollary 3.5.

However in the next section, we will show that orthogonality of A enables to derive lower bounds on $d_n(B_1(\|.\|_A))$ larger than the Bernstein n-width.

4.3 Lower bounds on the Kolmogorov width of orthogonal sets

Even when the unit ball in G-variation does not contain a ball of any radius in the norm $\|.\|$, in which the approximation error is measured, it might happen that it contains a ball of a non-zero 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 lower bound, which is an improvement of Barron's estimate [1, Lemma 6, p. 942].

Proposition 4.3 *Let $(X, \|.\|)$ be a Hilbert space, A be its orthonormal subset. When A is infinite, then for all positive integers n $d_n(A) \geq 1$. When 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 and let $p_{X_n} : X \to X_n$ be the best approximation mapping (projection) from X to X_n (see e.g. Singer [25]). Then for every $f \in X$ we have $e_{X_n}(f)^2 = \|f - p_{X_n}(f)\|^2 = 1 - \|p_{X_n}(f)\|^2$ and $\|p_{X_n}(f)\|^2 = \sum_{j=1}^n (f \cdot h_j)^2$.

For each $m \in \mathcal{N}_+$ choose some subset $A_m \subseteq A$ of cardinality m. 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 some $\alpha_m \in A_m$ such that $\|p_{X_n}(\alpha_m)\|^2 \leq \frac{n}{m}$ and thus $e_{X_n}(\alpha_m) \geq \sqrt{1 - \frac{n}{m}}$. So if A is finite of cardinality m, we have $d_n(A) \geq \sqrt{1 - \frac{n}{m}}$. If A is infinite, then for all $m \in \mathcal{N}_+$ we have $d_n(A) \geq \sqrt{1 - \frac{n}{m}}$ and so $d_n(A) \geq 1$. ∎

Notice that for a countable orthonormal set A this proposition gives a better lower bound on the Kolmogorov width than the lower bound derived using the Bernstein width.

For A, G subsets of a normed linear space $(X, \|.\|)$ let a_G denotes the Bernstein radius of the set A with respect to teh norm $\|.\|_G$, i.e. $a_G = \sup_{\alpha \in A} \|\alpha\|_G$.

Proposition 4.3 implies a lower bound on the Kolmogorov width for any set G for which there exists an ortonormal set A with bounded supremum of G-variation of all its elements.

Corollary 4.4 *Let $(X, \|.\|)$ be a Hilbert space, G, A be its subsets, A be orthonormal with that a_G finite. If A is infinite, then for all positive integers n $d_n(G) \geq \frac{1}{a_G}$. When A is finite of cardinality m, then for all positive integers $n \leq m$ $d_n(G) \geq \frac{1}{a_G}\sqrt{1-\frac{n}{m}}$.*

Proof. By Proposition 3.2 (iii) $\|.\|_G \leq c\|.\|_A$. So $B_{\frac{1}{a_G}}(\|.\|_A) = \frac{1}{a_G}B_1(\|.\|_A) \subseteq B_1(\|.\|_G)$. Hence by Propositions 2.2 (iii) and 4.1 (iii) $d_n(G) = d_n(B_1(\|.\|_G)) \geq \frac{1}{a_G}\, d_n(B_1(\|.\|_A)) = \frac{1}{a_G}\, d_n(A)$. and hence the lower bound follows from Proposition 4.3. ∎

Corollary 4.4 implies that whenever the unit ball in G-variation contains a ball of a non-zero radius r in variation with respect to some infinite orthonormal set, then G cannot be approximated within an accuracy smaller than r using any linear approximation scheme. No increase at all of the dimension n of the linear approximating set can decrease the n-width of G below r.

Even when $B_1(\|.\|_G)$ is not "large enough" to contain a ball of some non-zero 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 that are going to zero rather slowly with respect to the dimension d. The following definition formalizes the concept of such a slow decrease.

Let $(X, \|.\|)$ be a normed linear space and $A \subset X$. We say that A is *slowly decaying with respect to d* if A can be linearly ordered as $A = \{\alpha_j;\, j \in \mathcal{N}_+\}$ so that the norms of its elements are non-increasing, and for all $r \in \mathcal{N}_+$ $\|\alpha_{r^d}\| \geq \frac{1}{r}$. The following lemma gives an equivalent formulation of the concept of a slowly decaying set.

Lemma 4.5 *Let $(X, \|.\|)$ be a normed linear space and A be a subset of X. Then A is slowly decaying with respect to d if and only if A can be represented as $A = \cup_{r\in\mathcal{N}_+} A_r$, where for all $r \in \mathcal{N}_+$ $card\, A_r \geq r^d$ and for all $\alpha \in A_r$ $\|\alpha\| \geq \frac{1}{r}$ and for all $r' > r$ and $\alpha' \in A_{r'}$ $\|\alpha\| \geq \|\alpha'\|$.*

The following proposition shows that the Kolmogorov width of any orthogonal slowly decaying set exhibits the curse of dimensionality.

Proposition 4.6 *Let $(X, \|.\|)$ be a Hilbert space and A be its orthogonal subset that is slowly decaying with respect to some positive integer d. Then for any positive integer n $d_n(A) \geq \frac{1}{\sqrt{2}\sqrt[d]{m_n}}$, where $m_n = \min\{m \in \mathcal{N}_+;\, (2n \leq m)\,\&\,(\exists r \in \mathcal{N}_+)(m = r^d)\}$. In particular, for $n = \frac{r^d}{2}$ for some integer r $d_n(A) \geq \frac{1}{\sqrt{2}\sqrt[d]{2n}}$.*

Proof. Let $A_r = \{\alpha_1, \ldots, \alpha_{r^d}\}$. Then $card\, A_r = r^d$ and, by Proposition 4.3, for all $r \in \mathcal{N}_+$ and all $n \leq r^d$ $d_n(A_r^0) \geq \sqrt{1-\frac{n}{r^d}}$. By Proposition 2.2 (iv) then $d_n(A_r) \geq \frac{1}{r}\sqrt{1-\frac{n}{r^d}}$.

For each $n \in \mathcal{N}_+$ take $r \in \mathcal{N}_+$ for which $m_n = r^d$. Since $n < 2n \leq m_n = r^d$, we have $d_n(A) \geq d_n(A_r) \geq \frac{1}{r}\sqrt{1-\frac{n}{r^d}}$. Setting $r = \sqrt[d]{m_n}$ we get $d_n(A) \geq \frac{1}{\sqrt{2}}\frac{1}{\sqrt[d]{m_n}}$. ∎

The curse of dimensionality also applies to the Kolmogorov width of any family of sets $\{G_d; d \in \mathcal{N}_+\}$, for which there exists a family of orthogonal sets $\{A_d; d \in \mathcal{N}_+\}$,

where each A_d is slowly decaying with respect to d and there exists an upper bound on G_d-variation of all elements of A_d and all d. Even when $\sup_{\alpha\in A_d}\|\alpha\|_G$ depend on d, but do not grow too quickly, we get useful lower bounds on the Kolmogorov width of the family $\{G_d; d\in\mathcal{N}_+\}$

Corollary 4.7 *Let $(X,\|.\|)$ be a Hilbert space, G, A be its subsets, A be orthogonal slowly decaying with respect to some positive integer d and $c=\sup_{\alpha\in A}\|\alpha\|_G$ be finite. Then for any $n\in\mathcal{N}_+$ $d_n(G)\geq\frac{1}{c\sqrt{2}\sqrt[d]{m_n}}$, where $m_n=\min\{m\in\mathcal{N}_+;\ (2n\leq m)\,\&\,(\exists r\in\mathcal{N}_+)(m=r^d)\}$. In particular, when $n=r^d$ for some integer r, then $d_n(G)\geq\frac{1}{c\sqrt{2}\sqrt[d]{2n}}$.*

Corollary 4.7 gives a method of deriving lower bounds on Kolmogorov width that we will apply in the next section to sets of functions computable by standard computational units of neural networks: Heaviside and sigmoidal perceptrons, and perceptrons with periodic activation function.

5 Kolmogorov width of sets of functions computable by perceptrons

5.1 Variation with respect to perceptrons

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

A one-hidden-layer perceptron network with an activation function ψ and n hidden units computes functions from the parametrized set $span_n P_d(\psi)$, where $P_d(\psi)=P_d(\psi,J)=\{f:J^d\to\mathcal{R}; f(\mathbf{x})=\psi(\mathbf{v}\cdot\mathbf{x}+b),\mathbf{v}\in\mathcal{R}^d,b\in\mathcal{R}\}$ for some compact $J\subset\mathcal{R}$, corresponding to the domain of network inputs.

The most common activation functions are *sigmoidals*, i.e. functions $\sigma:\mathcal{R}\to[0,1]$ such that $\lim_{t\to-\infty}\sigma(t)=0$ and $\lim_{t\to+\infty}\sigma(t)=1$. There are used both continuous sigmoidals like the *logistic sigmoid* $\frac{1}{1+e^{-t}}$ or the *hyperbolic tangent*, as well as the discontinuous *Heaviside function* ϑ, defined by $\vartheta(t)=0$ for $t<0$ and $\vartheta(t)=1$ for $t\geq 0$.

Notice 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 J: indeed, $\vartheta(\mathbf{v}\cdot.+b)$ restricted to J is equal to the characteristic function of the positive half-space $H^+_{\mathbf{v},b}=\{\mathbf{x}\in J^d;\mathbf{v}\cdot\mathbf{x}+b\geq 0\}$. We will write $H_d(J)$, H_d, resp., instead of $P_d(\vartheta,J)$, $P_d(\psi)$,resp., and call variation with respect to H_d *variation with respect to half-spaces*, denoted by $\|.\|_{H_d}$.

Sometimes it is more convenient to use *signum*, defined by $sgn(t)=-1$ for $t<0$ and $sgn(t)=1$ for $t\geq 0$. Since signum function can be obtained from Heaviside function by the linear transformation $sgn(t)=2\vartheta(t)-1$, any function computable by a network with n Heaviside perceptrons can be computed by a network consisting of $n+1$ signum perceptrons. For the sake of notational convenience, we will write $\|.\|_{S_d}$ to denote variation with respect to signum perceptrons.

There have been considered also other types of activation functions like *cosine* (Jones, [9]) and the *ramp function* $\kappa : \mathcal{R} \to \mathcal{R}$ (Breiman, [3]), defined as $\kappa(t) = t\,\vartheta(t)$, i.e. $\kappa(t) = 0$ for $t < 0$ and $\kappa(t) = t$ for $t \geq 0$. We will write R_d instead of $P_d(\kappa)$.

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

Proposition 5.1 *Let d be a positive integer and $p \in [1, \infty)$, then in $(\mathcal{L}_p([0,1]^d), \|.\|_p)$ the following holds:*
(i) for every sigmoidal function $\sigma : \mathcal{R} \to \mathcal{R}$ $\|.\|_{P_d(\sigma)} = \|.\|_{H_d}$;
(ii) $\|.\|_{S_d} \leq \|.\|_{H_d} \leq 3\|.\|_{S_d}$;
(iii) $\|.\|_{R_d} = 2\|.\|_{H_d}$.

Proof. For (i) and (ii) see [18], [16], [19]. To verify (iii) consider $\rho : \mathcal{R} \to \mathcal{R}$ defined as $\rho(t) = 0$ for $t < 0$, $\rho(t) = t$ for $0 \leq t < 1$, and $\rho(t) = 1$ for $t \geq 1$. Since $\rho(t) = \kappa(t) - \kappa(t-1)$, we have $\|.\|_{R_d} \leq 2\|.\|_{P_d(\rho)}$. Since ρ is sigmoidal, it follows from (i) that $\|.\|_{P_d(\rho)} = \|.\|_{H_d}$. ∎

So variation with respect to half-spaces is equal to variation with respect to perceptrons with any sigmoidal activation function or, up to a multiplicative constant, to variation with respect to signum or ramp perceptrons. Thus applying to perceptron networks Corollary 3.5 (or its various extensions to $\mathcal{L}_p$ spaces with $p \in (1, \infty]$ that can also be formulated in terms of variation), we can restrict ourselves to the variation with respect to half-spaces.

To construct a lower bound on Kolmogorov width of the unit ball in variation with respect to half-spaces, we will use orthogonal slowly decaying families containing plane waves. A function $f : \mathcal{R}^d \to \mathcal{R}$ is called a *plane wave*, if it can be represented as $f(\mathbf{x}) = \psi(\mathbf{v} \cdot \mathbf{x})$, where $\psi : \mathcal{R} \to \mathcal{R}$ is any function and $\mathbf{v} \in \mathcal{R}^d$. Notice that plane waves are constant along hyperplanes parallel to the cozero hyperplane $\{\mathbf{x} \in \mathcal{R}^d; \mathbf{v} \cdot \mathbf{x} = 0\}$ of the linear function $\mathbf{v} \cdot \mathbf{x}$.

We will use square waves and cosine plane waves. *Square waves* are plane waves obtained from the *Haar function*, denoted by $\xi : \mathcal{R} \to \mathcal{R}$ and defined by $\xi(t) = 1$ for $t \in [i, i + \frac{1}{2})$ and $\xi(t) = -1$ for $t \in [i - \frac{1}{2}, i)$ for all integers i. When the Haar function on an interval $J \subset \mathcal{R}$ is appropriately scaled, than it can be represented as a convex combination of characteristic functions of half-intervals (half-spaces) of J. More precisely, $c_J \xi \in conv\, H_1(J)$, where $c_J = \frac{1}{2\lceil l \rceil}$, l denotes the length of the interval J and $\lceil x \rceil$ the smallest integer that is greater or equal to a real number x.

Variation with respect to half-spaces of a general plane wave $f(\mathbf{x}) = \psi(\mathbf{v} \cdot \mathbf{x})$ can be computed from the total variation of ψ. Notice that for $d = 1$ variation with respect to half-spaces coincides up to a constant with total variation (see Barron [2], Kůrková, Kainen and Kreinovich [18]). Recall (see e.g. Kolmogorov and Fomin [14]) that *total variation* of a function $\psi : J \to \mathcal{R}$, where $J \subset \mathcal{R}$ is a closed interval, is defined as

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

where the supremum is taken over all finite partitions $t_0 < \ldots < t_n$ of $J = [t_0, t_n]$.

It follows directly from the definition of total variation that for a periodic function $\psi : \mathcal{R} \to \mathcal{R}$ with a period τ, $V(\psi, J) \leq \lceil \frac{l}{\tau} \rceil V(\psi, [0, \tau])$, where l denotes the length of the interval J. We will use this property together with the following two lemmas to estimate variation with respect to half-spaces of cosine plane waves.

Lemma 5.2 *Let d be a positive integer and $\psi(\mathbf{v} \cdot \mathbf{x}) \in (\mathcal{L}_p, ([0,1]^d), \|.\|_p)$, $p \in [1, \infty]$, be a plane wave. Then $\|\psi(\mathbf{v} \cdot \mathbf{x})\|_{H_d([0,1]^d)} = \|\psi\|_{H_1(J)}$, where $J = [0, \sum_{i=1}^d v_i]$ and $H_1(J)$ is considered with respect to $(\mathcal{L}_p(J), \|.\|_p)$.*

Proof. To check that $\|\psi(\mathbf{v} \cdot \mathbf{x})\|_{H_d([0,1]^d)} \leq \|\psi\|_{H_1(J)}$, let $b = \|\psi\|_{H_1(J)}$. It follows from the definition of H_d-variation that $\psi = \lim_{m \to \infty} \sum_{j=1}^{n_m} w_{m,j}\, \theta(t - b_{m,j})$ in $(\mathcal{L}_p(J), \|.\|_p)$, where for all $m \in \mathcal{N}_+$ $\sum_{j=1}^{n_m} |w_{m,j}| \leq b$. Then $\psi(\mathbf{v} \cdot \mathbf{x}) = \lim_{m \to \infty} \sum_{j=1}^{n_m} w_{m,j}\, \theta(\mathbf{v} \cdot \mathbf{x} - b_{m,j})$ in $(\mathcal{L}_p, ([0,1]^d), \|.\|_p)$, since for all $\mathbf{x} \in [0,1]^d$ we have $\mathbf{v} \cdot \mathbf{x} \in J$. So $\|\psi(\mathbf{v} \cdot \mathbf{x})\|_{H_d([0,1]^d)} \leq b$.

It is easy to see that $\|\psi(t)\|_{H_1(J)} = \|\psi(\|\mathbf{v}\| t)\|_{H_1(J^*)}$ where $J^* = [0, \sum_{i=1}^d v_i / \|\mathbf{v}\|]$. Thus to prove that $\|\psi(\mathbf{v} \cdot \mathbf{x})\|_{H_d([0,1]^d)} \geq \|\psi\|_{H_1(J)}$, it is sufficient to show that $\|\psi(\mathbf{v} \cdot \mathbf{x})\|_{H_d([0,1]^d)} \geq \|\psi(\|\mathbf{v}\| t)\|_{H_1(J^*)}$.

Let $\mathbf{u} \in [0,1]^d$ be such that $\|\mathbf{u}\| = \max\{\mathbf{x} \in [0,1]^d \& \mathbf{x}^0 = \mathbf{v}^0\}$. Then $\sum_{i=1}^d v_i / \|\mathbf{v}\| = \|\mathbf{u}\|$. Let $\hat{J} = \{t\mathbf{u}; t \in [0,1]\} \subseteq [0,1]^d$. Set $\|\psi(\mathbf{v} \cdot \mathbf{x})\|_{H_d([0,1]^d)} = b$, then $\psi(\mathbf{v} \cdot \mathbf{x}) = \lim_{m \to \infty} f_m$ in $(\mathcal{L}_p([0,1]^d), \|.\|_p)$, where for all $m \in \mathcal{N}_+$ $f_m \in conv\, H_d(b)$. Setting $\hat{f}_m = f_m / \hat{J}$ we get $\hat{f}_m \in conv\, H_1(\hat{J})(b)$ and $\lim_{m \to \infty} \hat{f}_m = \psi(\|\mathbf{v}\| t)$ in $(\mathcal{L}_p(J^*), \|.\|_p)$. Thus $\|\psi(\|\mathbf{v}\| t)\|_{H_1(J^*)} \leq b$. ∎

If $J \subset \mathcal{R}$ is a closed interval, we say that $\psi : J \to \mathcal{R}$ is *piecewise uniformly continuous* if there exist real numbers $s_1 < \ldots < s_k$ such that $J = [s_1, s_k]$ and for all $i = 1, \ldots, k-1$ $\psi/(s_i, s_{i+1})$ is uniformly continuous.

Lemma 5.3 *Let $J \subset \mathcal{R}$ be a closed interval, and $\psi : J \to \mathcal{R}$ be piecewise uniformly continuous. Then*

$$\|\psi\|_{H_1(J)} \leq V(\psi, J)$$

Proof. Let $s_1 < \ldots < s_k$ be a partition of J such that $\psi/(s_i, s_{i+1})$ is uniformly continuous for all $i = 1, \ldots, k-1$. Hence for every $m \in \mathcal{N}_+$ there exists a partition $t_{m,0} < \ldots < t_{m,n_m}$ of J refining $s_1 < \ldots < s_k$, such that setting $w_{m,i} = \psi(t_{m,i}) - \psi(t_{m,i-1})$ and $\psi_m(t) = \sum_{i=1}^{n_m} w_{m,i} \vartheta(t - t_{m,i})$, we have $\lim_{m \to \infty} \psi_m = \psi$ in $(\mathcal{L}_p(J), \|.\|_p)$.

Setting $\psi_m(t) = \sum_{i=1}^{n_m} \vartheta(t - t_{m,i})$, we get $\psi = \lim_{m \to \infty} \psi_m$ in $(\mathcal{L}_p(J), \|.\|_p)$. Since for all $m \in \mathcal{N}_+$ $\|\psi_m\|_{H_1(J)} \leq \sum_{i=1}^{n_m} |w_{m,i}| = \sum_{i=1}^{n_m} |\psi(t_{m,i}) - \psi(t_{m,i-1})| \leq V(\psi, J)$, we have $\|\psi\|_{H_1(J)} \leq V(\psi, J)$. ∎

From Lemma 5.2 and Lemma 5.3 we get the following upper bound on variation with respect to half-spaces of plane waves.

Proposition 5.4 *Let d be a positive integer and $\psi(\mathbf{v} \cdot \mathbf{x}) \in (\mathcal{L}_p([0,1]^d), \|.\|_p)$ be a plane wave such that $\psi : J \to \mathcal{R}$, where $J = [0, \sum_{i=1}^d v_i]^d$ and $\mathbf{v} = (v_1, \ldots, v_d)$, is piecewise uniformly continuous. Then $\|\psi(\mathbf{v} \cdot \mathbf{x})\|_{H_d([0,1]^d)} \leq V(\psi, J)$.*

It follows immediately from Proposition 5.4 that for any $\mathbf{v} \in \mathcal{R}^d$ $\|\xi(\mathbf{v} \cdot \mathbf{x})\|_{H_d([0,1]^d)} \leq 4\lceil \sum_{i=1}^d v_i \rceil$ and $\|\cos(2\pi \mathbf{v} \cdot \mathbf{x})\|_{H_d([0,1]^d)} \leq 4\lceil \sum_{i=1}^d v_i \rceil$. Moreover, it is easy to check that any square wave $\xi(\mathbf{v} \cdot \mathbf{x})$ is in the convex hull of $H_d([0,1]^d)(4\lceil \sum_{i=1}^d v_i \rceil)$.

5.2 Lower bounds for perceptrons with periodic activation functions

To derive lower bounds on the Kolmogorov width of the set of functions computable by a single perceptron with d inputs using the methods developed in the previous sections, we need to find suitable orthogonal sets of functions, for which variation with respect to such perceptrons does not grow too quickly with d. For some periodic activation functions ψ, there even exist orthogonal sets with $P_d(\psi)$-variation bounded by a constant independent on d.

It is well-known that the following two families of plane waves are for all positive integers d orthonormal in $(\mathcal{L}_2([0,1]^d), \|.\|_2)$:

$$A_d(cos) = \left\{ \frac{1}{\sqrt{2}} \cos(2\pi \mathbf{v} \cdot \mathbf{x}); \mathbf{v} \in \mathcal{N}_+^d \right\} \tag{5.1}$$

$$A_d(\xi) = \left\{ \xi(\mathbf{v} \cdot \mathbf{x}); \mathbf{v} \in \{2^j; j \in \mathcal{N}_+\}^d \right\}. \tag{5.2}$$

Since the first one is a subset of $\frac{1}{\sqrt{2}} P_d(\cos)$ and the second one of $P_d(\xi)$, the following lower bounds follow immediately from Proposition 4.3.

Proposition 5.5 *For all positive integers d, n in $(\mathcal{L}_2([0,1]^d), \|.\|_2)$*

$$d_n(P_d(\cos)) = d_n(B_1(\|.\|_{P_d(\cos)})) \geq \sqrt{2}$$

$$d_n(P_d(\xi)) = d_n(B_1(\|.\|_{P_d(\xi)})) \geq 1.$$

Thus there is no possibility of decreasing the worst-case error in linear approximation of $P_d(cos)$, $P_d(\xi)$, resp., under $\sqrt{2}$, 1, resp., by increasing the dimension of the linear approximating subspace. So perceptrons with either cosine or Haar function as activations cannot be efficiently approximated linearly. On the other hand, it follows from Corollary 3.5 that

$$\delta_{P_d(cos),n}(B_1(\|.\|_{P_d(cos)}) \leq \frac{\sqrt{2}}{\sqrt{n}},$$

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

5.3 Lower bounds for sigmoidal perceptrons

It was shown above, that for any sigmoidal activation function σ $P_d(\sigma)$-variation is equal to variation with respect to half-spaces. Estimating the total variation of elements of an orthogonal family of plane waves, we can find proper scalars that allow to decrease the norms of such a family so that it can be embeded into the unit ball in variation with respect to half-spaces.

Barron [1, Theorem 6, p. 942] used the above defined orthonormal family $A_d(\cos)$ to estimate the Kolmogorov width of sets Γ_c defined as $\Gamma_c = \{f \in (\mathcal{L}_2([0,1]^d), \|.\|_2); c_f \leq c\}$, where $c_f = \int_{\mathcal{R}^d} \|\omega\|_2 \, |\tilde{f}(\omega)| \, d\omega$, $\tilde{f}$ is the Fourier transform of f and $\|\omega\|_2 = \sqrt{\omega \cdot \omega}$

denotes the l_2-norm of the frequency ω. He proved that $d_n(\Gamma_c) \geq \kappa \frac{c}{d\sqrt[d]{n}}$, where $\kappa \geq \frac{1}{8\pi e^{\pi-1}}$, while $\delta_{P_d(\sigma),n}(\Gamma_c) \leq \frac{c}{\sqrt{n}}$.

His result shows that neural networks outperform linear approximation only for n large enough with respect to d. Indeed, consider approximation of Γ_c and assume that n is the maximal number of hidden units that is feasible using a given type of implementation. Since $\lim_{d\to\infty} \frac{c}{d\sqrt[d]{n}} = 0$, for large input dimension d $d_n(\Gamma_c)$ might be quite small. Only for n sufficiently larger than d, the upper bound, $\frac{c}{\sqrt{n}}$, on $\delta_{P_d(\sigma),n}(\Gamma_c)$ is smaller than the lower bound, $\frac{\kappa}{d\sqrt[d]{n}}$, on $d_n(\Gamma_1)$ (for example, for $c = 1$ and $d = 3$ n must be greater than $\frac{3^6}{\kappa}$). Only for such large n, Barron's result implies that approximation by perceptron networks with n hidden units outperforms approximation by elements of any n-dimensional linear subspace.

In the following, we improve this Barron's result in two directions. We improve the lower bound on the Kolmogorov width and show that it even applies to the set of characteristic functions of half-spaces and that the worst-case error is achieved.

The following theorem shows that even the set of characteristic functions of half-spaces H_d has a lower bound on its Kolmogorov n-width of order $\mathcal{O}\left(\frac{1}{d\sqrt[d]{n}}\right)$. Since all the elements of the set H_d, $P_d(\sigma)$, resp., have in approximation by $span_n H_d$, $span_n P_d(\sigma)$ error equal to zero, H_d, $P_d(\sigma)$, resp., is an example of a set of functions, for which neural network approximation outperforms any linear approximation method for all n (representing, respectively, the number of computational units and the dimension of the linear subspace).

Theorem 5.6 *For all positive integers d, n in $(\mathcal{L}_2([0,1]^d), \|.\|_2)$*

$$d_n(H_d) \geq d_{m_n}(H_d) \geq \frac{1}{4d\sqrt[d]{m_n}},$$

where $m_n = \min\{m \in \mathcal{N}_+; (2n \leq m) \& (\exists r \in \mathcal{N}_+)(m = r^d)\}$. In particular, for $n = \frac{r^d}{2}$ for some integer r

$$d_n(H_d) \geq \frac{1}{4d\sqrt[d]{2n}}.$$

Proof. Taking advantage of Proposition 4.7 we will derive a lower bound on $d_n(H_d)$ using an orthogonal slowly decaying set A_d obtained from $D-d(\cos)$ by proper scaling. For all $d, r \in \mathcal{N}_+$ set $A_{d,r} = \{\alpha_{\mathbf{v}}; \mathbf{v} \in \{1,\ldots,r\}^d\} \subset (\mathcal{L}_2([0,1]^d), \|.\|_2)$, where $\alpha_{\mathbf{v}}(\mathbf{x}) = c_{\mathbf{v}} \cos(2\pi \mathbf{v}\cdot\mathbf{x}) : [0,1]^d \to \mathcal{R}$ and $c_{\mathbf{v}} = \frac{1}{\sqrt{2}} \frac{d}{\lceil \sum_{k=1}^d v_k \rceil}$, where $\mathbf{v} = (v_1,\ldots,v_d)$. Let $A_d = \cup_{r\in\mathcal{N}_+} A_{d,r}$. We will show that $A_d \subseteq B_{d\sqrt{8}}(\|.\|_{H_d})$ and that A_d is slowly decaying with respect to d.

It follows from Proposition 5.4 that $\|\cos(2\pi \mathbf{v}\cdot\mathbf{x})\|_{H_d} \leq V(\cos(2\pi t), [0, \sum_{k=1}^d v_k]) \leq 4\lceil \sum_{k=1}^d v_k \rceil$. Thus for every $\alpha_{\mathbf{v}} \in A_d$ $\|\alpha_{\mathbf{v}}\|_{H_d} \leq d\sqrt{8}$ and hence by Proposition 3.2 (iii) $\|.\|_{H_d} \leq d\sqrt{8}\|.\|_{A_d}$. So by Proposition 4.1 (i) and (iii) $d_n(B_1(\|.\|_{H_d})) \geq \frac{1}{d\sqrt{8}} d_n(A_d)$.

For all d A_d is orthogonal and it can be ordered in such a way that it is slowly decaying with respect to d. Reindex A_d as $\{\alpha_i; i \in \mathcal{N}_+\}$ using any linear ordering of ${\mathcal{N}_+}^d$ such that $\{\|\alpha_i\|_2; i \in \mathcal{N}_+\}$ are nondecreasing and for all $r \in \mathcal{N}_+$ α_{r^d} corresponds

to $\alpha_{(r,\ldots,r)}$. Since $\|\alpha_{r^d}\| = \frac{1}{r}$, in such a linear ordering A_d is slowly decaying with respect to d.

By Proposition 4.7 for all positive integers n $d_n(H_d) = d_n(B_1(\|.\|_{H_d})) \geq \frac{1}{d\sqrt{8}} d_n(A_d) \geq \frac{1}{4d\sqrt[d]{m_n}}$, where $m_n = \min\{m \in \mathcal{N}_+; (2n \leq m) \& (\exists r \in \mathcal{N}_+)(m = r^d)\}$. ■

The following corollary shows that for each $n = \frac{r^d}{2}$ for some integer r and for each X_n n-dimensional subspace of $(\mathcal{L}_2([0,1]^d), \|.\|_2)$, there exists a half-space of $[0,1]^d$ such that its characteristic function χ_n has distance from X_n at least $\frac{1}{4d\sqrt[d]{2n}}$.

Corollary 5.7 *For all positive integers d, n and every n-dimensional subspace X_n of $(\mathcal{L}_2([0,1]^d), \|.\|_2)$ there exists a characteristic function χ_n of a half-space of $[0,1]^d$ such that*

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

where $m_n = \min\{m \in \mathcal{N}_+; (2n \leq m) \& (\exists r \in \mathcal{N}_+)(m = r^d)\}$. In particular, for $n = r^d 2$ for some integer r

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

Proof. By Theorem 5.6 for every n-dimensional subspace X_n of $(\mathcal{L}_2([0,1]^d), \|.\|_2)$ we have $\sup_{\chi \in H_d} \|\chi - X_n\| \geq \frac{1}{4d\sqrt[d]{m_n}}$. Since H_d is compact (see Chui, Li and Mhaskar [4], Gurvits and Koiran [8]) and the error functional e_{X_n} is continuous, the supremum of e_{X_n} on H_d is achieved at some χ_n. ■

Thus to approximate all characteristic functions of half-spaces of $[0,1]^d$ within a given accuracy ε, the dimension of any linear approximating space has to be larger than $\left(\frac{1}{4d\varepsilon}\right)^d$. When the desired accuracy of approximation ε_d is smaller than $\frac{1}{4d}$, then the required dimension of a linear approximating space might be too large to be feasible. For example, when $\varepsilon = \frac{1}{8d}$, the dimension of linear space must be at least 2^d. In such a case, the set of characteristic functions of half-spaces of d-dimensional unit cube (which is equal to the set of functions computable by Heaviside perceptrons) cannot be efficiently approximated using any linear approximating family. Note, however, that the value of ε which implies such an exponential growth of the dimension of the linear approximating space goes to zero with d.

Since variation with respect to half-spaces is equal to variation with respect to perceptrons with any sigmoidal activation function, Theorem 5.6 can be generalized to include all sigmoidal perceptrons.

Corollary 5.8 *Let d, n be positive integers and $\sigma : \mathcal{R} \to \mathcal{R}$ be any sigmoidal function. Then in $(\mathcal{L}_2([0,1]^d), \|.\|_2)$*

$$d_n(P_d(\sigma)) \geq d_{m_n}(P_d(\sigma)) \geq \frac{1}{4d\sqrt[d]{m_n}},$$

where $m_n = \min\{m \in \mathcal{N}_+; (2n \leq m) \& (\exists r \in \mathcal{N}_+)(m = r^d)\}$. In particular, for $n = r^d 2$ for some $r \in \mathcal{N}_+$

$$d_n(P_d(\sigma)) \geq \frac{1}{4d\sqrt[d]{2n}}.$$

Proof. By Proposition 5.1 (i) $\|.\|_{P_d(\sigma)} = \|.\|_{H_d}$, and, by Proposition 4.1 (i), $d_n(H_d) = d_n(B_1(\|.\|_{H_d})) = d_n(B_1(\|.\|_{P_d(\sigma)})) = d_n(P_d(\sigma))$. ∎

So to approximate within ε by an n-dimensional subspace the set of all functions computable by a sigmoidal perceptron with d inputs, n has to be at least $O\left(\left(\frac{1}{4d\varepsilon}\right)^d\right)$.

6 Discussion

We have studied the worst-case errors in approximation by a linear approximating set and by certain class of nonlinear sets that includes sets of functions computable by feedforward neural networks. Our aim was to describe sets of multivariable functions for which the worst-case errors in linear approximation are larger than those in approximation by neural networks.

Taking advantage of relatively small upper bounds on approximation errors of balls in certain norms tailored to a type of computational unit by such networks with units of such a type, we have explored possibilities of finding large lower bounds on the worst-case errors of such norms in linear approximation. We have considered various methods of estimation of the Kolmogorov width describing best accuracy achiavable using linear methods for such balls. Applying these methods to perceptron with various types of activation functions, we have obtained two types of sets on which neural networks outperform linear approximation. The first one includes the above mentioned balls for perceptrons with some periodic activation functions: they can be approximated by such networks with n hidden units within $\frac{1}{\sqrt{n}}$, while no increase of the dimension of a linear approximating set can decrease the worst case error under a certain constant. The second one includes sets of functions computable by sigmoidal perceptrons: the worst-case error in approximation by such networks is zero, while in linear approximation it is bounded from below by $\frac{1}{d\sqrt[d]{n}}$.

Note that an analogous argument as in the proof of Theorem 5.6 for scaled cosine plane waves, can be done for scaled square waves. The unit ball in variation with respect to half-spaces also contains scaled waves obtained from the Haar function. Moreover, appropriately scaled square waves are equal to convex combinations of characteristic functions of half-spaces. However, using scaled orthogonal square waves instead of cosine plane waves to derive a lower bound on the Kolmogorov width of the set of characteristic functions of half-spaces, we obtain a much smaller value for such a lower bound. The reason is that to guarantee orthogonality for a family of square waves, dyadic scalars have to be used. Such scaling, however, has to be compensated by much faster decrease of the norms of such waves, so that they remain in the unit ball in variation with respect to half-spaces. Using the same proof technique as for Theorem 5.6, by means of square waves we obtain for all positive integers d, n the following estimate of Kolmogorov n-width in $(\mathcal{L}_2([0,1]^d), \|.\|_2)$:

$$d_n(H_d) \geq d_{m_n}(H_d) \geq \frac{1}{4d\sqrt{2}} \frac{1}{2\sqrt[d]{m_n}},$$

where $m_n = \min\{m \in \mathcal{N}_+; (2n \leq m) \& (\exists r \in \mathcal{N}_+)(m = r^d)\}$.

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

$$d_n(H_d) \geq \frac{1}{4d\sqrt{2}} \frac{1}{2^{\sqrt[d]{2n}}}.$$

Acknowledgements

The authors whish to thank to Prof. P. C. Kainen from Georgetown University and Prof. S. Giulini from University of Genova for helpful discussions. V. Kůrková was partially supported by GAČR grant 201/99/0092 and her visits to the University of Genova were funded by Czech-Italian reciprocity program and by the Italian Ministry for the University and Research. The visits of M. Sanguineti to the Academy of Sciences of the Czech Republic in Prague were founded by grants CNR 96.02472.CT07, CNR 97.00048.PF42.

Bibliography

[1] Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. on Information Theory* 39, pp. 930-945, 1993.

[2] Barron, A. R.: Neural net approximation. *Proc. 7th Yale Workshop on Adaptive and Learning Systems*. K. Narendra Ed., pp. 69-72. Yale University Press, 1992.

[3] Breiman, L.: Hinging hyperplanes for regression, classification, and function approximation. *IEEE Trans. on Information Theory* 39, pp. 999-1013, 1993.

[4] Chui, C.K., Li, X., Mhaskar H.N.: Neural networks for localized approximation. *Math. Comput.* 63, pp. 607–623, 1994.

[5] Darken, C., Donahue, M., Gurvits, L., Sontag, E. Rate of approximation results motivated by robust neural network learning. *Proc. Sixth Annual ACM Conference on Computational Learning Theory*. The Association for Computing Machinery, New York, N.Y., pp. 303-309, 1993.

[6] Friedman, A. (1982): *Foundations of Modern Analysis*. Dover, New York, 1982 (originally published in 1970 by Holt, Rinehart and Winston, New York).

[7] Girosi, F.: Approximation error bounds that use VC-bounds. *Proc. Int. Conf. on Artificial Neural Networks ICANN'95*, pp. 295–302, 1995.

[8] Gurvits, L., Koiran, P. Approximation and learning of convex superpositions. *Journal of Computer and System Sciences* 55, pp. 161–170, 1997.

[9] Jones, L.K.: A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Annals of Statistics* 20, pp. 608-613, 1992.

[10] Kainen, P.C., Kůrková, V., Vogt, A: Approximation by neural networks is not continuous. *Neurocomputing*, 1999 (to appear).

[11] Kainen, P.C., Kůrková, V., Vogt, A: Continuity of approximation by neural networks in L_p spaces. *Annals of Operational Research*, 1999 (to appear).

[12] Kainen, P.C., Kůrková, V., Vogt, A: Geometry and topology of continuous best and near best approximations. Submitted to *Journal of Approximation Theory*, 1999.

[13] Kolmogorov, A. N.: Über die beste Annäherung von Funktionen einer gegebenen Funktionenklasse. *Annals of Math.* 37, pp. 107-110, 1936 (English translation On the best approximation of functions of a given class, in *Selected works on A. N. Kolmogorv*, vol. I (ed. V. M. Tikhomirov) (pp. 202–205). Kluwer, 1991.

[14] Kolmogorov, A. N., Fomin, S. V.: *Introductory Real Analysis.* Dover, New York, 1970.

[15] Kůrková, V.: Approximation of functions by neural networks. *Proc. ICSC/IFAC Symposium on Neural Computation '98*, pp. 29-35, 1998.

[16] Kůrková, V.: Dimension-independent rates of approximation by neural networks. *Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality* (Eds. K. Warwick, M. Kárný). Birkhauser, pp. 261-270, 1997.

[17] Kůrková, V.: Incremental approximation by neural networks. *Complexity: Neural Network Approach.* (Eds. M. Kárný, K. Warwick, V. Kůrková). Springer, London, pp. 177-188, 1998.

[18] Kůrková, V., Kainen, P.C., Kreinovich, V.: Estimates of the number of hidden units and variation with respect to half-spaces. *Neural Networks* 10, pp. 1061 -1068, 1997.

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

[20] Lorentz, G.G.: *Approximation of Functions.* Chelsea Publishing Company, New York, N.Y., 1966.

[21] Mhaskar, H.N.: Neural networks for optimal approximation of smooth and analytic functions. *Neural Computation* 8, pp. 164-177, 1996.

[22] Mhaskar, H.N., Micchelli, C.A.: Approximation by superposition of sigmoidal and radial basis functions. *Advances in Applied Mathematics* 13, pp. 350-373, 1992.

[23] Pinkus, A.: *$n-$Widths in Approximation Theory.* Springer-Verlag, New York, 1986.

[24] Sejnowski, T.J., Rosenberg, C.: Parallel networks that learn to pronounce English text. *Complex Systems* 1, pp. 145-168, 1987.

[25] Singer, I.: *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces.* Springer, Berlin, 1970.