Proof: First, let us observe that, under H_0, it follows that

$$\|r_i^2 - \tilde{r}_i^2\|_H^2 = 3(\lambda_i^2 + \tilde{\lambda}_i^2) - 2\left[2\lambda_i^2\left(\langle\varphi_i, \tilde{\varphi}_i\rangle\right)^2 + \lambda_i\tilde{\lambda}_i\right] \xrightarrow{k\uparrow\infty} 0. \tag{22}$$

Next, proceeding as in the proof of Theorem 7 we have

$$\begin{aligned}&\left\|\log\frac{dP_1}{dP_0}(r) - \log\tilde{g}_N(r)\right\|_H \\ &\quad\le \left\|\log\frac{dP_1}{dP_0}(r) - \frac{1}{2}\sum_{i=1}^{N} r_i^2\left(\frac{1}{\lambda_i} - 1\right) + \log\lambda_i\right\|_H \\ &\qquad + K_2\sum_{i=1}^{N}\|r_i^2 - \tilde{r}_i^2\|_H + \sqrt{3}\sum_{i=1}^{N}\left|1 - \frac{\tilde{\lambda}_i}{\lambda_i}\right| + \sum_{i=1}^{N}\left|\log\frac{\lambda_i}{\tilde{\lambda}_i}\right|\end{aligned}$$

where we have applied

$$\left|1 - \frac{1}{\lambda_i}\right| \le \left(\sum_{j=1}^{\infty}\left(1 - \frac{1}{\lambda_j}\right)^2\right)^{1/2} \le K_2 < \infty.$$

Finally, by letting $N \to \infty$ and taking (22) into account, the result follows. The proof is similar under H_1. □

REFERENCES

[1] S. Watanabe, "Karhunen–Loève expansion and factor analysis. Theoretical remarks and applications," in *Proc. Trans. 2th Prague Conf. Information Theory*, 1965.

[2] H. L. Van Trees, *Detection, Estimation, and Modulation Theory—Part I*. New York: Wiley, 1968.

[3] W. A. Gardner and L. E. Franks, "An alternative approach to linear least squares estimation of continuous random processes," in *Proc. 5th Annu. Princeton Conf. Information Sciences and Systems*, 1971.

[4] H. L. Van Trees, *Detection, Estimation, and Modulation Theory—Part III*. New York: Wiley, 1971.

[5] T. E. Fortmann and B. D. O. Anderson, "On the approximation of optimal realizable linear filters using a Karhunen–Loève expansion," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 561–564, July 1973.

[6] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. New York: Springer-Verlag, 1994.

[7] C. W. Helstrom, *Elements of Signal Detection & Estimation*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[8] R. Gutiérrez, J. C. Ruiz-Molina, and M. J. Valderrama, "On the numerical expansion of a second order stochastic process," *Appl. Stoch. Models Data Anal.*, vol. 8, no. 2, pp. 67–77, 1992.

[9] J. C. Ruiz-Molina, J. Navarro-Moreno, and M. J. Valderrama, "Differentiation of the modified approximative Karhunen–Loève expansion of a stochastic process," *Statist. Probab. Lett.*, vol. 42, pp. 91–98, 1999.

[10] C. T. H. Baker, *The Numerical Treatment of Integral Equations*. Oxford, U.K.: Oxford Univ. Press, 1977.

[11] J. C. Ruiz-Molina and M. J. Valderrama, "On the derivation of a suboptimal filter for signal estimation," *Statist. Probab. Lett.*, vol. 28, pp. 239–243, 1996.

[12] J. Navarro-Moreno, J. C. Ruiz-Molina, and M. J. Valderrama, "A solution to linear estimation problems using approximate Karhunen–Loève expansions," *IEEE Trans. Inform. Theory*, vol. 46, pp. 1677–1682, July 2000.

[13] J. C. Ruiz-Molina, J. Navarro-Moreno, and A. Oya, "Signal detection using approximative Karhunen–Loève expansions," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1672–1680, May 2001.

[14] R. T. Huang and R. A. Johnson, "Information transmission with time-continuous random processes," *IEEE, Trans. Inform. Theory*, vol. IT-9, pp. 84–94, Apr. 1963.

[15] W. L. Root, "Singular Gaussian measures in detection theory," in *Proc. Symp. Time Series Analysis*. New York: Wiley, 1963, pp. 292–315.

[16] T. T. Kadota, "Simultaneous diagonalization of two covariance kernels and application to second order stochastic processes," *SIAM J. Appl. Math.*, vol. 15, no. 6, pp. 1470–1480, 1967.

[17] ——, "Optimum estimation of nonstationary Gaussian signals in noise," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 253–257, Mar. 1969.

[18] ——, "Examples of optimum detection of Gaussian signals and interpretation of white noise," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 725–734, Sept. 1968.

[19] ——, "Simultaneously orthogonal expansion of two stationary Gaussian processes—Examples," *Bell Syst. Tech. J.*, vol. 45, pp. 1071–1096, 1966.

[20] F. Smithies, *Integral Equations*. New York: Cambridge Univ. Press, 1965.

[21] J. Hajek, "On linear statistical problems in stochastic processes," *Czech. Math. J.*, vol. 12, pp. 404–444, 1962.

[22] T. S. Pitcher, "An integral expression for the log likelihood ratio of two Gaussian processes," *SIAM J. Appl. Math.*, vol. 14, pp. 228–233, 1966.

[23] T. Kailath and H. L. Weinert, "An RKHS approach to detection and estimation problems. Part II: Gaussian signal detection," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 15–23, Jan. 1975.

Bounds on Rates of Variable-Basis and Neural-Network Approximation

Věra Kůrková and Marcello Sanguineti

***Abstract*—Tightness of bounds on rates of approximation by feedforward neural networks is investigated in a more general context of nonlinear approximation by variable-basis functions. Tight bounds on the worst case error in approximation by linear combinations of n elements of an orthonormal variable basis are derived.**

***Index Terms*—Approximation by variable-basis functions, bounds on rates of approximation, complexity of neural networks, high-dimensional optimal decision problems.**

I. INTRODUCTION

Feedforward neural networks have been successfully applied to the approximate solution of a large variety of high-dimensional problems ranging from the design of controllers acting on strongly nonlinear dynamic systems characterized by a large number of state variables to the identification of industrial processes and pattern recognition. All these problems share a common aspect: a certain network architecture is used to approximate multivariable input/output mappings (see, e.g., [1]–[3]).

Experience has shown that simple architectures with relatively few computational units can achieve surprisingly good performances. However, as the number of variables is scaled up, the feasibility of network design may become critical. Nevertheless, some network architectures can be mathematically proved to have desirable computational capabilities, guaranteeing that the number of computational units does not

Manuscript received June 26, 2000; revised November 24, 2000. 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 under Grant 201/00/1482. 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").

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 Genova, Italy (e-mail: marcello@dist.unige.it).

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

Publisher Item Identifier S 0018-9448(01)05455-4.

increase too fast with the dimensionality of certain tasks. Such theoretical results supplement experience with design criteria based on mathematical models.

Feedforward networks are often implemented on classical computers; for such implementations, one of the crucial issues is the *number of hidden units* needed to guarantee a desired accuracy. The dependence of such accuracy on the number of hidden units can be theoretically studied in the context of approximation theory in terms of *rates of approximation*.

Some insight into the reason why many high-dimensional tasks can be performed quite efficiently by neural networks with a moderate number of hidden units has been gained by Jones [4], who has constructed incremental approximants with a rate of convergence of the order of $O(1/\sqrt{n})$. The same estimates of rates of approximation had earlier been proved by Maurey using a probabilistic argument (it has been quoted by Pisier [5]; see also Barron [6]). Barron [6] has improved the constant in Jones's [4] upper bound and has applied such a bound to neural networks. Using a weighted Fourier transform, he has described sets of multivariable functions approximable by perceptron networks with n hidden units to an accuracy of the order of $O(1/\sqrt{n})$. However, it should be stressed that the sets of multivariable functions to which such estimates apply may become more and more constrained as the number d of variables increases (see, e.g., Girosi, Jones, and Poggio [7]), and some quantities not appearing in the notation $O(1/\sqrt{n})$ may depend on d (see, e.g., Kůrková, Savický, and Hlaváčková [8]).

Several authors have further improved or extended these upper bounds. An extension to $\mathcal{L}_p$ spaces, with $p \in (1, \infty)$, has been derived by Darken *et al.* [9] (with a rate of approximation of the order of $O(n^{-1/q})$, where $q = \max(p, p/(p-1))$. Similar upper bounds for the $\mathcal{L}_\infty$ space have been obtained by an extension of Maurey's probabilistic argument (see, e.g., Barron [10], Girosi [11], Gurvits and Koiran [12], Makovoz [13], Kůrková, Savický, and Hlaváčková [8]). An interesting improvement has been derived by Makovoz [14], who has combined a concept from metric entropy theory with a probabilistic argument. Possibilities of simplifying Jones's construction and/or modifying its parameters have been investigated by Dingankar and Sandberg [15], Docampo, Hush, and Abdallah [16], and Docampo [17]. Barron [6] and Kůrková and Sanguineti [18] have described sets of multivariable functions for which the worst case errors in linear approximation are larger than those in neural-network approximation. Using an argument much simpler than the proof techniques employed by Maurey, Jones, and Barron, Mhaskar and Micchelli [19] have obtained similar upper bounds for orthonormal approximating sets. For finite-dimensional spaces, Kůrková, Savický, and Hlaváčková [8] have improved Mhaskar and Micchelli's bounds up to tight ones. For perceptron networks, the tightness of Maurey–Jones–Barron's bound has been studied by Barron [10], Makovoz [14], and Kůrková and Sanguineti [20].

This work is motivated by recent papers by Dingankar [21], [31], and Levretsky [32], which explore the possibility of improving Maurey–Jones–Barron's upper bound of the order of $O(1/\sqrt{n})$ up to a bound of $O(1/n^2)$. We investigate the limitations of improvements of Maurey–Jones–Barron's upper bound in the general context of nonlinear approximation of the variable-basis type, i.e., approximation by linear combinations of n-tuples of elements of a given set of basis functions. This approximation scheme has been widely studied: it includes free-nodes splines (see, e.g., Petrushev [22] and DeVore and Lorentz [23, Ch. 13]), nonlinear trigonometric approximation (i.e., approximation by trigonometric polynomials with free frequencies; see, e.g., Maiorov [24], Belinskiĭ [25] and DeVore and Temlyakov [26]), sums of wavelets (see, e.g., DeVore, Jawerth, and Popov [27]), as well as feedforward multilayer neural networks with a single linear output unit. In the case of one-hidden-layer networks, a variable basis corresponds to computational units in the hidden layer. For a larger number of layers, such a basis becomes more complex, as it depends on the number of units in the previous hidden layers.

In the variable-basis approximation framework, Maurey–Jones–Barron's upper bound can be expressed in terms of *two* norms of the function to be approximated: 1) a norm in which the accuracy of approximation is measured, and 2) a norm tailored to the given basis (e.g., the computational units of a neural network).

We demonstrate the limitations of improvements of Maurey–Jones–Barron's upper bound in the case of an orthonormal basis, for which the norm 2) equals l_1 norm with respect to this basis. We derive tight upper bounds in terms of both norms 1) and 2). Our results are extensions to estimates derived by Kůrková, Savický, and Hlaváčková [8] for finite-dimensional spaces. From our estimates it follows that for a general variable basis, Maurey–Jones–Barron's bound cannot be substantially improved. In the orthonormal case, it can be improved at most by a factor dependent on the ratio between the above-mentioned norms, but the term $1/\sqrt{n}$ remains essentially unchanged (it is only replaced by $1/(2\sqrt{n-1})$).

The correspondence is organized as follows. Section II contains basic concepts and notations concerning feedforward neural networks and approximation in normed linear spaces. Section III presents tight bounds on rates of approximation for orthonormal bases. In Section IV, our results are discussed. All proofs are deferred to Section V.

II. Preliminaries

A. Feedforward Neural Networks

Feedforward neural networks compute parametrized sets of functions dependent on the type of computational units as well as on the type of their interconnections. 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

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

where $a_i \in A \subseteq \mathbb{R}^p$ ($\mathbb{R}$ denotes the set of real numbers), $\phi: \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ corresponds to a *computational unit*, and p and d are the dimensions of the *parameter space* and the *input space*, respectively. We denote by

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

the parametrized set of functions corresponding to the computational unit ϕ.

Perceptrons are the most widespread type of hidden units. A *perceptron* with an *activation function* $\psi: \mathbb{R} \to \mathbb{R}$ computes functions of the form

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

where $v \in \mathbb{R}^d$ is an *input weight vector* and $b \in \mathbb{R}$ is a *bias*.

Estimates of rates of approximation by feedforward neural networks as a function of the number of hidden units can be formulated in a more general context of *variable-basis approximation*.

B. Approximation by Variable-Basis Functions

By *linear space* we mean a linear space over real numbers. If X is a linear space, then the *dimension* of X is denoted by $\dim X$.

Let $(X, \|\cdot\|)$ be a Banach space (i.e., a complete normed linear space) with norm $\|\cdot\|$, then $B_r(\|\cdot\|)$ denotes the ball of radius r with respect to the norm $\|\cdot\|$, i.e.,

$$B_r(\|\cdot\|) = \{f \in X: \|f\| \leq r\}.$$

When it is clear from the context which norm is considered, we shall simply write X instead of $(X, \|\cdot\|)$. Recall that a Hilbert space is a complete normed linear space with the norm induced by an inner product.

If G is a subset of X and $c \in \mathbb{R}$, then we define

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

and, for c positive

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

The *closure* of G is denoted by $cl\, G$ and defined as

$$cl\, G = \{f \in X : \quad \forall \varepsilon > 0 \quad \exists g \in G \quad \|f - g\| < \varepsilon\}.$$

G is *dense* in $(X, \|\cdot\|)$ when $cl\, G = X$. $(X, \|\cdot\|)$ is *separable* when it has a countable dense subset.

The *linear span* of G, which we denote by $\operatorname{span} G$, consists of all linear combinations of elements of G, i.e.,

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

where $\mathbb{N}_+$ denotes the set of positive integers. The set of all linear combinations of at most n elements of G is denoted by $\operatorname{span}_n G$, and defined as

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

$\operatorname{conv} G$ denotes the *convex hull* of G, consisting of all convex combinations of elements of G, i.e.,

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

$\operatorname{conv}_n G$ is the set of all convex combinations of at most n elements of G, i.e.,

$$\operatorname{conv}_n G = \left\{\sum_{i=1}^{n} a_i g_i\colon a_i \in [0, 1],\, \sum_{i=1}^{n} a_i = 1,\, g_i \in G\right\}.$$

If M is a subset of a Banach space $(X, \|\cdot\|)$ and $f \in X$, then

$$\|f - M\| = \inf_{g \in M} \|f - g\|$$

denotes the *distance* of f from M. Approximation of functions from a set Y by elements of an approximating set M can be investigated in terms of the *worst case error*, which is formalized by the concept of *deviation of Y from M* defined as

$$\begin{aligned}\delta(Y, M) &= \delta(Y, M, (X, \|\cdot\|)) \\ &= \sup_{f \in Y} \|f - M\| = \sup_{f \in Y} \inf_{g \in M} \|f - g\|.\end{aligned}$$

Linear approximation theory investigates approximation by linear subspaces, which are often generated by the first n elements of a linearly independent subset G of X with a *fixed ordering*. For example, when G is the set of powers $\{x^{i-1}\colon i \in \mathbb{N}_+\}$, then the linear space generated by its first n elements is the set of all polynomials of order at most $n - 1$.

We call *nonlinear approximation by variable-basis functions* the approximation by linear combinations of all n-tuples of elements of a given set G. This corresponds to approximation by the set $\operatorname{span}_n G$ of all linear combinations of at most n elements of G, i.e., approximation by the *union of at most n-dimensional subspaces* generated by elements of G. One-hidden-layer feedforward networks with a linear output unit and n units computing the function ϕ in the hidden layer belong to this approximation scheme. The set

$$\operatorname{span}_n G_\phi = \left\{\sum_{i=1}^{n} w_i \phi(a_i, \cdot)\colon w_i \in \mathbb{R},\, a_i \in A \subseteq \mathbb{R}^p\right\}$$

consists of all linear combinations of at most n parametrized functions $\phi(a_i, \cdot)$ (the variable basis is obtained by varying the parameter vector of the computational unit). Also multilayer feedforward networks with a single linear output unit and n units in the last hidden layer belong to this approximation scheme, but the corresponding sets G are more complex and depend on the number of units in the previous hidden layers.

For elements of the convex closure of a bounded subset G of a Hilbert space, Maurey (see [5]), Jones [4], and Barron [6] have derived an upper bound of the order of $O(1/\sqrt{n})$ on the rate of approximation by $\operatorname{conv}_n G$. The following theorem presents this upper bound (see Barron [6, Lemma 1]) in a slightly reformulated way.

Theorem 1: Let $(X, \|\cdot\|)$ be a Hilbert space, G its subset, and b a positive real number such that for every $g \in G$, $\|g\| \leq b$. Then, for every $f \in cl\, \operatorname{conv} G$ and for every positive integer n

$$\|f - \operatorname{conv}_n G\| \leq \sqrt{\frac{b^2 - \|f\|^2}{n}}.$$

As $\operatorname{conv}_n G \subseteq \operatorname{span}_n G$, the upper bound given in Theorem 1 also applies to rates of approximation by $\operatorname{span}_n G$. However, when G is not closed up to multiplication by scalars, $\operatorname{conv} G$ is a proper subset of $\operatorname{span} G$, and hence also $cl\, \operatorname{conv} G$ is a proper subset of $cl\, \operatorname{span} G$. Thus, the density of $\operatorname{span} G$ in $(X, \|\cdot\|)$ does not guarantee that Theorem 1 can be applied to all elements of X.

As $\operatorname{conv}_n G(c) \subseteq \operatorname{span}_n G(c) = \operatorname{span}_n G$ for any positive c, replacing the set G with $G(c) = \{wg\colon g \in G,\, w \in \mathbb{R}, |w| \leq c\}$, we can apply Theorem 1 to all elements of $\cup_{c \in \mathbb{R}_+} cl\, \operatorname{conv} G(c)$. This approach can be mathematically formulated in terms of a norm tailored to a set G (in particular, to sets G_ϕ corresponding to various computational units ϕ in feedforward networks), called *G-variation* (variation with respect to G) and defined as the Minkowski functional of the set $cl\, \operatorname{conv}\, (G \cup -G)$, i.e.,

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

G-variation has been introduced by Kůrková [28] as an extension of Barron's [10] concept of variation with respect to half-spaces. It is a norm on the subspace $\{f \in X\colon \|f\|_G < \infty\} \subseteq X$. The closure in the definition depends on the topology induced on X by the norm $\|\cdot\|$. When X is finite-dimensional, all norms are topologically equivalent and, thus, G-variation does not depend on the choice of a norm on X. From the definition of G-variation it follows that, for every $f \in X$

$$\|f\| \leq s_G \|f\|_G, \qquad \text{where } s_G = \sup_{g \in G} \|g\|.$$

Intuitively, $\|f\|_G$ shows us how much the set G should be "dilated," so that f is in the closure of the convex symmetric hull of the "dilated" set. G-variation is a generalization of two concepts: total variation (see, e.g., Kolmogorov and Fomin [29, p. 328]) and l_1 norm. When G is an orthonormal basis of a separable Hilbert space $(X, \|\cdot\|)$, then l_1 *norm with respect to* G, denoted by $\|\cdot\|_{1,G}$, is defined as

$$\|f\|_{1,G} = \sum_{g \in G} |f \cdot g|.$$

So $\|\cdot\|_{1,G}$ is a norm on $\{f \in X\colon \|f\|_{1,G} < \infty\}$.

Proposition 1: Let $(X, \|\cdot\|)$ be a separable Hilbert space and G be its orthonormal basis. Then $\|\cdot\|_G = \|\cdot\|_{1,G}$.

Using Proposition 1, we obtain the following upper bound as a special case of Kůrková's [28] (see also [8]) reformulation of Maurey–Jones–Barron's theorem in terms of G-variation.

Theorem 2: Let $(X, \|\cdot\|)$ be a separable Hilbert space and G its orthonormal basis. Then, for every $f \in X$ and every positive integer n

$$\|f-\mathrm{span}_n\, G\| \leq \sqrt{\frac{\|f\|_{1,G}^2-\|f\|^2}{n}} = \frac{\|f\|_{1,G}}{\sqrt{n}}\sqrt{1-\frac{\|f\|^2}{\|f\|_{1,G}^2}}.$$

As $0 \in \mathrm{span}_n\, G$, we have, for all $f \in X$, $\|f - \mathrm{span}_n\, G\| \leq \|f\|$, which implies the trivial upper bound $\|f\|$ on $\|f - \mathrm{span}_n\, G\|$. Hence, when $\|f\|$ is such that

$$\|f\|^2 < \frac{\|f\|_{1,G}^2 - \|f\|^2}{n}$$

or, equivalently,

$$\frac{\|f\|}{\|f\|_{1,G}} < \frac{1}{\sqrt{n+1}}$$

the trivial upper bound, $\|f\|$, is better than the upper bound given in Theorem 2. For example, if $\|f\|_{1,G} = 1$, then the trivial upper bound is better when $\|f\| < 1/\sqrt{n+1}$.

The upper bound in Theorem 2 can also be formulated in terms of deviation. We denote the *deviation from* $\mathrm{span}_n\, G$ by $\delta_{G,n}$, i.e.,

$$\delta_{G,n}(Y) = \delta(Y, \mathrm{span}_n\, G).$$

The following properties of $\delta_{G,n}$ can be easily derived from its definition.

Proposition 2: Let $(X, \|\cdot\|)$ be a Banach space and G and Y its subsets. Then, for all positive integers n

i) for any $c \in \mathbb{R}$, $\delta_{G,n}(cY) = |c|\delta_{G,n}(Y)$;

ii) $\delta_{G,n}(cl\, Y) = \delta_{G,n}(Y)$;

iii) $\delta_{G,n+1}(Y) \leq \delta_{G,n}(Y)$;

iv) if $Y' \subseteq Y$, then $\delta_{G,n}(Y') \leq \delta_{G,n}(Y)$.

Theorem 2 implies the following upper bounds on $\delta_{G,n}$ of balls in l_1 norm with respect to G and on $\delta_{G,n}$ of their subsets defined by a constraint on the value of the norm $\|\cdot\|$.

Corollary 1: Let $(X, \|\cdot\|)$ be a separable Hilbert space, G its orthonormal basis, and r, b real numbers such that $0 \leq r \leq b$. Then, for every positive integer n

i) $\delta_{G,n}(B_b(\|\cdot\|_{1,G})) \leq \frac{b}{\sqrt{n}}$;

ii) $\delta_{G,n}(\{f \in B_b(\|\cdot\|_{1,G}): \|f\| \geq r\}) \leq \frac{b}{\sqrt{n}}\sqrt{1-\frac{r^2}{b^2}}$.

Thus, balls of radius b in l_1 norm with respect to G can be approximated by $\mathrm{span}_n\, G$ with accuracy $b/\sqrt{n}$, independently of the number of variables of the functions in the space X. However, it should be noted that the condition of being in the unit ball in l_1 norm with respect to G may become more and more constraining with an increasing number of variables [8].

In the next section, we shall investigate how tight are the upper bounds given in Theorem 2 and Corollary 1.

III. Tight Bounds on Rates of Approximation for Orthonormal Bases

Let G be an orthonormal basis of an infinite-dimensional separable Hilbert space. We shall show that, in this case, the maximum possible improvement of the bound in Theorem 2 lies in the replacements of the factor

$$\sqrt{1-\frac{\|f\|^2}{\|f\|_{1,G}^2}} \quad \text{with} \quad 1-\frac{\|f\|^2}{\|f\|_{1,G}^2}$$

and of the factor

$$\frac{\|f\|_{1,G}}{\sqrt{n}} \quad \text{with} \quad \frac{\|f\|_{1,G}}{2\sqrt{n-1}}.$$

When f is a finite linear combination of elements of an orthonormal subset G of a Hilbert space, then $\|f - \mathrm{span}_n\, G\|$ can be easily calculated. The proof of the following lemma is a straightforward consequence of the definition of $\mathrm{span}_n\, G$ and the orthonormality of G.

Lemma 1: Let $(X, \|\cdot\|)$ be a Hilbert space, G its orthonormal subset, and $f = \sum_{i=1}^k w_i g_i$, where, for all $i = 1, \ldots, k$, $w_i \in \mathbb{R}$ and $g_i \in G$. Then, for all positive integers $n < k$

$$\|f - \mathrm{span}_n\, G\| = \min\left\{\left\|\sum_{i\in I} w_i g_i\right\| : \mathrm{card}\, I = k-n\right\}.$$

Using a simple proof technique based on a rearrangement of an orthonormal basis of a separable Hilbert space, Mhaskar and Micchelli [19] have obtained bounds on $\delta_{G,n}$ in terms of l_1 norm with respect to G. They have shown that, for all positive integers n

$$\frac{b}{2\sqrt{n}} \leq \delta_{G,n}(B_b(\|\cdot\|_{1,G})) \leq \frac{b}{\sqrt{n+1}}.$$

Mhaskar and Michelli's estimates have been derived using simple arguments, but they are weaker than the estimates obtained by Maurey, Jones, and Barron, as they are formulated only in terms of $\|\cdot\|_{1,G}$ without taking into account the value of $\|\cdot\|$.

For finite-dimensional Hilbert spaces, Kůrková, Savický, and Hlaváčková [8] have improved Mhaskar and Michelli's upper bound up to $b/(2\sqrt{n})$, and have shown that this bound is tight when $\dim X \geq 2n$. Moreover, they have derived a tight estimate of the deviation $\delta_{G,n}$ from $\mathrm{span}_n\, G$ of sets defined by constraints on both norms, l_1 norm with respect to G, and $\|\cdot\|$.

The following theorems extend the results in [8] to infinite-dimensional separable Hilbert spaces. Their proofs exploit ideas contained in the papers by Mhaskar and Micchelli [19] and by Kůrková, Savický, and by Hlaváčková [8].

Theorem 3: Let $(X, \|\cdot\|)$ be an infinite-dimensional separable Hilbert space and G its orthonormal basis. Then, for every positive real number b and every positive integer n

$$\delta_{G,n}(B_b(\|\cdot\|_{1,G})) = \frac{b}{2\sqrt{n}}.$$

When G is a countable infinite orthonormal basis, Theorem 3 improves the upper bound i) in Corollary 1 up to an exact value of the deviation from $\mathrm{span}_n G$ of balls in l_1 norm with respect to G. In contrast to Corollary 1 ii), which expresses an upper bound in terms of both $\|f\|_{1,G}$, and $\|f\|$, Theorem 3 does not take $\|f\|$ into account. However, even without using the value of $\|f\|$, it gives a better bound than Corollary 1 ii) when the ratio $\|f\|/\|f\|_{1,G}$ is sufficiently small

$$\text{if } \frac{\|f\|}{\|f\|_{1,G}} < \frac{\sqrt{3}}{2}, \qquad \text{then } \frac{\|f\|_{1,G}}{2\sqrt{n}} < \sqrt{\frac{\|f\|_{1,G}^2 - \|f\|^2}{n}}.$$

The following theorem gives, for G orthonormal, the maximum improvement of Corollary 1 ii) achievable in terms of $\|\cdot\|_{1,G}$ and $\|\cdot\|$.

Theorem 4: Let $(X, \|\cdot\|)$ be an infinite-dimensional separable Hilbert space, G its orthonormal basis, and b, r real numbers such that $0 \leq r \leq b$. Then, for every positive integer $n \geq 2$

i) if $\frac{r}{b} \geq \frac{1}{\sqrt{2n-1}}$, then

$$\frac{b}{4\sqrt{n-1}}\left(1-\frac{r^2}{b^2}\right) \leq \delta_{G,n}(\{f \in X: \|f\|_{1,G} = b, \|f\| = r\})$$
$$\leq \frac{b}{2\sqrt{n-1}}\left(1-\frac{r^2}{b^2}\right);$$

ii) if $\sqrt{n} - \sqrt{n-1} \le \frac{r}{b} < \frac{1}{\sqrt{2n-1}}$, then

$$\frac{r}{2\sqrt{2}} < \delta_{G,n}(\{f \in X: \|f\|_{1,G} = b, \|f\| = r\}) \le \frac{b}{2\sqrt{n-1}}\left(1 - \frac{r^2}{b^2}\right);$$

iii) if $\frac{r}{b} < \sqrt{n} - \sqrt{n-1}$, then

$$\frac{r}{2\sqrt{2}} < \delta_{G,n}(\{f \in X: \|f\|_{1,G} = b, \|f\| = r\}) \le r.$$

Theorem 4 i) and iii) give, up to a constant factor, the best possible upper bounds on $\|f - \mathrm{span}_n G\|$ that can be obtained in terms of $\|f\|_{1,G}$ and $\|f\|$.

IV. Discussion

Our results contribute to investigation of tightness of the upper bounds on rates of approximation by neural networks derived by Maurey (see [5]), Jones [4], and Barron [6]. We have studied such bounds in the context of nonlinear approximation of the variable-basis type, which includes feedforward neural networks with a single linear output unit. For approximation by a variable orthonormal basis G of an infinite-dimensional separable Hilbert space we have derived tight bounds in terms of two norms: the norm in which the accuracy of the approximation error is measured, and the l_1 norm with respect to G.

When the set G of variable-basis functions is not orthonormal, the technique used to prove the results of Section III cannot be applied. In the special case of a variable basis corresponding to perceptrons with a sigmoidal activation function, Barron [10] and Makovoz [14] have derived tight bounds. Both have used probabilistic arguments, in [14] combined with concepts from metric entropy theory. The application of metric entropy tools to derive tight bounds for perceptron networks with a sigmoidal activation function has been further developed by Kůrková and Sanguineti [20].

From our results it follows that the upper bound obtained by Maurey, Jones, and Barron (stated here in the form of Theorem 1) cannot be essentially improved, unless some additional properties of the set of basis functions are guaranteed (see, e.g., Makovoz [14]). Thus, our results contribute to clarifying the issues recently discussed in [21], [31], and [32].

V. Proofs

A. Proof of Proposition 1

We first check that $\|\cdot\|_G \le \|\cdot\|_{1,G}$. Let $G = \{g_i: i \in \mathbb{N}_+\}$. Then every $f \in X$ can be expressed as $\sum_{i=1}^{\infty}(f \cdot g_i)g_i$. For $m \in \mathbb{N}_+$, set

$$f_m = \sum_{i=1}^{m}(f \cdot g_i)g_i.$$

If $b = \|f\|_{1,G}$, then, for all $m \in \mathbb{N}_+$, $f_m \in \mathrm{conv}\, G(b)$. $f = \lim_{m\to\infty} f_m$ in $\|\cdot\|$, and so f is in the closure of $\mathrm{conv}\, G(b)$ with respect to $\|\cdot\|$. Hence, $\|f\|_G \le b = \|f\|_{1,G}$.

We now verify that $\|\cdot\|_G \ge \|\cdot\|_{1,G}$. Let $b_\varepsilon < \|f\|_G + \varepsilon$ for some $\varepsilon > 0$, then, by the definition of $\|f\|_G$, there exists a sequence $\{f_i: i \in \mathbb{N}_+\}$ such that $f_i \in \mathrm{conv}\, G(b_\varepsilon)$ for all $i \in \mathbb{N}_+$, and $f = \lim_{i\to\infty} f_i$ in $\|\cdot\|$. For $m \in \mathbb{N}_+$, set

$$f_{m,i} = \sum_{j=1}^{m}(f_i \cdot g_j)g_j \quad \text{and} \quad f_m = \sum_{j=1}^{m}(f \cdot g_j)g_j.$$

Since the projection onto the m-dimensional subspace $\mathrm{span}\{g_1, \ldots, g_m\}$ is continuous (see, e.g., [30, p. 145]), we have $\lim_{i\to\infty} f_{m,i} = f_m$ in $\|\cdot\|$. As all norms on a finite-dimensional space are topologically equivalent, we also have $\lim_{i\to\infty} f_{m,i} = f_m$ in $\|\cdot\|_{1,G}$. Using the triangle inequality, we get

$$\sum_{j=1}^{m}(|f_{m,i} \cdot g_j| - |f_m \cdot g_j|) \le \sum_{j=1}^{m}|(f_{m,i} - f_m) \cdot g_j|$$

hence

$$\lim_{i\to\infty}\sum_{j=1}^{m}|f_{m,i} \cdot g_j| = \sum_{j=1}^{m}|f_m \cdot g_j|$$

for all $m \in \mathbb{N}_+$. Thus, for all $\varepsilon > 0$, we have

$$\|f\|_{1,G} \le b_\varepsilon < \|f\|_G + \varepsilon$$

which implies $\|\cdot\|_{1,G} \le \|\cdot\|_G$. □

B. Proof of Theorem 3

By Proposition 2 i)

$$b\delta_{G,n}(B_1(\|\cdot\|_{1,G})) = \delta_{G,n}(B_b(\|\cdot\|_{1,G})).$$

So it is sufficient to verify that $\delta_{G,n}(B_1(\|\cdot\|_{1,G})) = 1/(2\sqrt{n})$.

To derive the upper bound, let $f \in B_1(\|\cdot\|_{1,G})$ and, using the same trick as Mhaskar and Micchelli [19], reorder G in such a way that $f = \sum_{i=1}^{\infty} w_i g_i$, where, for all $i \in \mathbb{N}_+$, $|w_i| \ge |w_{i+1}|$ and $g_i \in G$. Set

$$f_n = \sum_{i=1}^{n} w_i g_i.$$

By Lemma 1

$$\|f - \mathrm{span}_n G\|^2 \le \|f - f_n\|^2 = \sum_{i=n+1}^{\infty} w_i^2 \le |w_{n+1}| \sum_{i=n+1}^{\infty} |w_i|.$$

As $\sum_{i=1}^{\infty}|w_i| = 1$, we have

$$\sum_{i=n+1}^{\infty}|w_i| \le 1 - n|w_{n+1}|$$

and hence

$$\|f - \mathrm{span}_n G\|^2 \le |w_{n+1}|(1 - n|w_{n+1}|).$$

Setting $t = |w_{n+1}|$, we get

$$\delta_{G,n}^2(B_1(\|\cdot\|_{1,G})) \le t(1 - nt).$$

The right-hand side of this inequality achieves its maximum, equal to $1/(4n)$, for $t = 1/(2n)$. Thus,

$$\delta_{G,n}(B_1(\|\cdot\|_{1,G})) \le \frac{1}{2\sqrt{n}}.$$

To verify the lower bound, let

$$f_n = \frac{1}{2n}\sum_{i=1}^{2n} g_i.$$

Then $\|f_n\|_{1,G} = 1$ and $\delta_{G,n}(B_1(\|\cdot\|_{1,G})) \ge \|f_n - \mathrm{span}_n G\|$. By Lemma 1 and the orthonormality of G

$$\|f_n - \mathrm{span}_n G\| = \left\|\frac{1}{2n}\sum_{i\in I} g_i\right\|$$

where $\mathrm{card}\, I = n$. So

$$\|f_n - \mathrm{span}_n G\| = \frac{1}{2\sqrt{n}}. \qquad \square$$

C. Proof of Theorem 4

By Proposition 2 i)

$$\delta_{G,n}(\{f \in X \colon \|f\|_{1,G} = b,\ \|f\| = r\}) = b\delta_{G,n}(\{f \in X \colon \|f\|_{1,G} = 1,\ \|f\| = r/b\}).$$

Thus, it is sufficient to verify the statement of the theorem for $b = 1$.

First, we derive the upper bound. Let $f \in X$ be such that $\|f\|_{1,G} = 1$ and $\|f\| = r$. Using the same trick as Mhaskar and Micchelli [19], reorder G in such a way that f can be represented as $f = \sum_{i=1}^{\infty} w_i g_i$, where, for all $i \in \mathbb{N}_+$, $|w_i| \geq |w_{i+1}|$ and $g_i \in G$. As

$$\|f\|^2 = \sum_{i=1}^{\infty} w_i^2 \leq |w_1| \|f\|_{1,G} = |w_1|$$

we have $|w_1| \geq \|f\|^2 = r^2$.

Set

$$f' = \sum_{i=2}^{\infty} w_i g_i$$
$$G' = G - \{g_1\}$$

and

$$h' = \sum_{i=2}^{n} w_i g_i.$$

By Lemma 1 and Theorem 3

$$\|f' - \mathrm{span}_{n-1} G'\| \leq \|f' - h'\| \leq \frac{\|f'\|_{1,G'}}{2\sqrt{n-1}} = \frac{\|f\|_{1,G} - |w_1|}{2\sqrt{n-1}} = \frac{1 - |w_1|}{2\sqrt{n-1}}.$$

Setting $h = w_1 g_1 + h'$, we get $h \in \mathrm{span}_n\, G$ and $\|f - h\| = \|f' - h'\|$. Hence,

$$\|f - \mathrm{span}_n G\| \leq \frac{1 - |w_1|}{2\sqrt{n-1}} \leq \frac{1 - r^2}{2\sqrt{n-1}}.$$

As $0 \in \mathrm{span}_n G$, for every $f \in X$ we have the trivial upper bound $\|f - \mathrm{span}_n G\| \leq \|f\| = r$. It is easy to check that

$$r < \frac{1 - r^2}{2\sqrt{n-1}} \qquad \text{if and only if } 0 \leq r < \sqrt{n} - \sqrt{n-1}.$$

So in both the cases i) and ii), we have the upper bound $(1 - r^2)/(2\sqrt{n-1})$, whereas in the case iii), we have the upper bound r.

To derive the lower bound, for every positive integer k and every real number c, consider

$$f_k^c = \frac{1}{k}\left((1 + (k-1)c)g_1 + \sum_{i=2}^{k}(1 - c)g_i\right).$$

It is easy to check that, for any $c \in [0, 1]$, $\|f_k^c\|_{1,G} = 1$, and for $c^2 = (kr^2 - 1)/(k - 1)$, $\|f_k^c\| = r$. For any $k \geq 1/r^2$, define

$$f_k = f_k^c, \qquad \text{where } c = \sqrt{\frac{kr^2 - 1}{k - 1}}$$

(as $r^2 \leq 1$, we have $c \in [0, 1]$). Then $\|f_k\|_{1,G} = 1$ and $\|f_k\| = r$.

By Lemma 1 and the definition of f_k, we have, for every $2 \leq n < k$

$$\|f_k - \mathrm{span}_n G\| = \frac{\sqrt{k-n}}{k}(1 - c) = \frac{\sqrt{k-n}}{k}\left(1 - \sqrt{\frac{kr^2 - 1}{k-1}}\right) = \frac{\sqrt{k-n}}{k-1}\,\frac{1 - r^2}{1 + \sqrt{\dfrac{kr^2 - 1}{k - 1}}}.$$

It is easy to check that the expression $\sqrt{k-n}/(k-1)$ achieves its maximum, equal to $1/(2\sqrt{n-1})$, for $k = 2n - 1$.

Now, we need to distinguish between two cases: $r \geq 1/\sqrt{2n-1}$ (corresponding to item i)) and $r < 1/\sqrt{2n-1}$ (corresponding to the items ii) and iii)). In the first case, set $k = 2n - 1$. As $2n - 1 \geq 1/r^2$, we have $k \geq 1/r^2$; thus, f_k is properly defined. As $r^2 \leq 1$, we have $(kr^2 - 1)/(k - 1) \leq 1$, and hence

$$\|f_k - \mathrm{span}_n G\| = \frac{\sqrt{k-n}}{k-1}\,\frac{1 - r^2}{1 + \sqrt{\frac{kr^2-1}{k-1}}} \geq \frac{1}{2\sqrt{n-1}}\,\frac{1 - r^2}{2} \geq \frac{1 - r^2}{4\sqrt{n-1}}.$$

In the second case, when $2n - 1 < 1/r^2$, set $k = \lceil 1/r^2 \rceil$. Then $k \geq 1/r^2$; thus, f_k is properly defined and, moreover, $2n \leq k$. Now

$$\|f_k - \mathrm{span}_n G\| = \frac{\sqrt{k-n}}{k}\left(1 - \sqrt{\frac{kr^2 - 1}{k - 1}}\right) = r\sqrt{\frac{k-n}{k}}\left(\frac{1}{r\sqrt{k}} - \sqrt{\frac{k - 1/r^2}{k(k-1)}}\right).$$

As $1/r^2 > k - 1$ and $2n \leq k$, we have

$$\|f_k - \mathrm{span}_n G\| > r\sqrt{1 - \frac{n}{k}}\left(\sqrt{\frac{k-1}{k}} - \frac{1}{\sqrt{k(k-1)}}\right) = r\sqrt{1 - \frac{n}{k}}\,\frac{k-2}{\sqrt{k(k-1)}} > \frac{r}{\sqrt{2}}\left(1 - \frac{2}{k}\right).$$

As $n \geq 2$, we have $k \geq 4$, and so

$$\frac{r}{\sqrt{2}}\left(1 - \frac{2}{k}\right) \geq \frac{r}{2\sqrt{2}}. \qquad \square$$

ACKNOWLEDGMENT

The authors wish to thank Prof. S. Giulini (University of Genoa) and Prof. P. C. Kainen (Georgetown University) for helpful discussions, and the reviewers for many useful comments and references (especially Reviewer 1 for the proof of the second part of Proposition 1).

REFERENCES

[1] R. Zoppoli, M. Sanguineti, and T. Parisini, "Approximating networks and extended Ritz method for the solution of functional optimization problems," *J. Optimiz. Theory Applic.*, to be published.

[2] T. Parisini and R. Zoppoli, "Neural approximations for infinite-horizon optimal control of nonlinear stochastic systems," *IEEE Trans. Neural Networks*, vol. 9, pp. 1388–1408, Nov. 1998.

[3] 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.

[4] 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.

[5] G. Pisier, “Remarques sur un resultat non publié de B. Maurey,” in *Séminaire d’Analyse Fonctionelle*. Palaiseau, France: École Polytechnique, Centre de Mathématiques, 1980–1981, vol. I, no. 12.

[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] F. Girosi, M. Jones, and T. Poggio, “Regularization theory and neural networks architectures,” *Neural Comput.*, vol. 7, no. 2, pp. 219–269, 1995.

[8] 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.

[9] 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.

[10] A. R. Barron, “Neural net approximation,” in *Proc. 7th Yale Worksh. Adaptive and Learning Systems*, K. S. Narendra, Ed., 1992, pp. 69–72.

[11] F. Girosi, “Approximation error bounds that use VC-bounds,” in *Proc. Int. Conf. Artificial Neural Networks*, vol. 1, Paris, France, 1995, pp. 295–302.

[12] L. Gurvits and P. Koiran, “Approximation and learning of convex superpositions,” *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 161–170, 1997.

[13] Y. Makovoz, “Uniform approximation by neural networks,” *J. Approx. Theory*, vol. 95, no. 2, pp. 215–228, 1998.

[14] ——, “Random approximants and neural networks,” *J. Approx. Theory*, vol. 85, no. 1, pp. 98–109, 1996.

[15] A. T. Dingankar and W. Sandberg, “A note on error bounds for approximation in inner product spaces,” *Circuits, Syst. Signal Processing*, vol. 15, no. 4, pp. 515–518, 1996.

[16] D. Docampo, R. Hush, and C. T. Abdallah, “Constructive function approximation: Theory and practice,” in *Intelligent Methods in Signal Processing and Communications*, D. Docampo, A. Figueiras, and F. Pérez, Eds. Boston, MA: Birkhäuser, 1997, pp. 199–219.

[17] D. Docampo, “New results on convex constructive function approximation,” in *Proc. Europ. Conf. Signal Analysis and Prediction*, Prague, Czech Republic, 1997, pp. 125–128.

[18] V. Kůrková and M. Sanguineti, “Comparison of worst-case errors in linear and neural netsork approximation,” *IEEE Trans. Inform. Theory*, to be published.

[19] 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.

[20] 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., Prague, Czech Republic, 2001, pp. 41–45.

[21] A. T. Dingankar, “The unreasonable effectiveness of neural network approximation,” *IEEE Trans. Automat. Contr.*, vol. 44, pp. 2043–2044, Nov. 1999.

[22] P. Petrushev, “Direct and converse theorems for spline and rational approximation and Besov spaces,” in *Function Spaces and Applications (Lecture Notes in Mathematics)*, M. Cwikel, J. Peetre, Y. Sagher, and H. Wallin, Eds. Berlin, Germany: Springer-Verlag, 1988, vol. 1302, pp. 363–377.

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

[24] V. E. Maiorov, “Trigonometric diameters of the Sobolev classes W_p^r in the space L_q,” *Math. Notes Acad. Sci. U.S.S.R.*, pp. 590–597, Jan. 1987. Translated from *Mat. Zametki*, vol. 40, no. 2, pp. 161–173, 1986.

[25] E. S. Belinskiĭ, “Approximation of functions of several variables by trigonometric polynomials with given number of harmonics, and estimates of ε-entropy,” *Anal. Math.*, vol. 15, no. 2, pp. 67–74, 1989.

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

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

[28] 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. Basel, Switzerland: Birkhäuser, 1997, pp. 261–270.

[29] A. N. Kolmogorov and S. V. Fomin, *Introductory Real Analysis*. New York: Dover, 1975.

[30] I. Singer, *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*. Berlin, Heidelberg, Germany: Springer-Verlag, 1970.

[31] A. T. Dingankar, “Author’s reply: Comments on ‘The unreasonable effectiveness of neural networks approximation’,” *IEEE Trans. Automat. Contr.*, vol. 46, pp. 512–513, Mar. 2001.

[32] E. Lavretsky, “Comments on ‘The unreasonable effectiveness of neural networks approximation’,” *IEEE Trans. Automat. Contr.*, vol. 46, pp. 511–512, Mar. 2001.