Chapter 5
Approximating Multivariable Functions by Feedforward Neural Nets

Paul C. Kainen, Věra Kůrková, and Marcello Sanguineti

Abstract. Theoretical results on approximation of multivariable functions by feedforward neural networks are surveyed. Some proofs of universal approximation capabilities of networks with perceptrons and radial units are sketched. Major tools for estimation of rates of decrease of approximation errors with increasing model complexity are proven. Properties of best approximation are discussed. Recent results on dependence of model complexity on input dimension are presented and some cases when multivariable functions can be tractably approximated are described.

Keywords: multivariable approximation, feedforward neural networks, network complexity, approximation rates, variational norm, best approximation, tractability of approximation.

1 Introduction

Many classification, pattern recognition, and regression tasks can be formulated as mappings between subsets of multidimensional vector spaces, using

Paul C. Kainen
Department of Mathematics and Statistics, Georgetown University
Washington, D.C. 20057-1233, USA
e-mail: kainen@georgetown.edu

Věra Kůrková
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2, 182 07, Prague 8, Czech Republic
e-mail: vera@cs.cas.cz

Marcello Sanguineti
DIBRIS, University of Genova,
via Opera Pia 13 - 16145 Genova, Italy
e-mail: marcello.sanguineti@unige.it

M. Bianchini et al. (Eds.): *Handbook on Neural Information Processing*, ISRL 49, pp. 143–181.
DOI: 10.1007/978-3-642-36657-4_5

a suitable encoding of inputs and outputs. The goal of a computer scientist modeling such tasks is to find a mathematical structure which operates by adjusting parameters in order to classify or recognize different patterns and to approximate the regression function associated to data.

Mathematical formalizations have shown that many types of feedforward networks (including all standard types that are popular in applications as well as many others that may not have been considered by experimentalists) satisfy these requirements. Provided they contain sufficiently many basic computational units, it is possible to adjust their parameters so that they approximate to any accuracy desired a wide variety of mappings between subsets of multidimensional spaces. In neural network terminology, such classes of networks are called *universal approximators*. When network parameters are adjusted using suitable learning methods (e.g., gradient-descent, incremental or genetic algorithms), there is no need for explicit formulation of rules or feature extraction. However, we remark that *implicitly* some estimate of the number and nature of features may guide the practitioners choice of architectural parameters such as the number of hidden units and their type.

The universal approximation property has been proven for one-hidden layer networks with almost all types of reasonable computational units. Various interesting proof techniques have been used such as integral representations (Carrol and Dickenson [7], Ito [25], Park and Sandberg [60]), the Hahn-Banach Theorem (Cybenko [11]), the Stone-Weierstrass theorem (Hornik, Stinchcombe and White [24]), and orthogonal polynomials (Mhaskar [56], Leshno et al. [54]).

But universality cannot be proved within reasonable bounds on complexity. Each kind of network implementation determines a different measure of complexity. Currently, feedforward networks are mostly simulated on classical computers. For such simulations, the limiting factor is the number of hidden units and the size of their parameters. Thus, a suitable type of computational units and a structure for their connections has to be found that can solve a given task within the feasible limits of network complexity. Some guidelines for the choice of a type of neural network can be derived from mathematically grounded complexity theory of neural networks. Its recent achievements include estimates of rates of approximation by various types of feedforward networks and comparisons of complexity requirements of approximation by neural networks with linear approximation.

In this chapter, we first briefly sketch some ideas used in proofs of universal approximation capabilities of networks with perceptrons and radial units. Then the focus changes to network complexity. Major tools are described for estimating rates of decrease of approximation errors with increasing model complexity. We start with the Maurey-Jones-Baron Theorem holding in Hilbert spaces, present its extension to $\mathcal{L}^p$-spaces, and finally give an improvement to geometric rate due to Kůrková and Sanguineti [51]. We discuss other improvements and their limitations and show how estimates of rates of approximation of multivariable functions can be reformulated in terms of worst-case errors in sets of functions defined as balls in norms tailored to computational units.

Further, we sketch the results of Kainen, Kůrková, and Vogt [32, 33, 34] that similarly to linear approximation, approximation by a variable basis of half-space characteristic functions always has a best approximant, but unlike linear approximation, neural network approximation does not have *continuous* best approximation. This leads to the initially surprising result that lower bounds on the possible rate of continuous methods for approximation do not apply to neural networks (cf. [13]).

Finally, recent results on dependence of model complexity on input dimension (so-called "tractability") are considered; see [31]. We focus on those which were found by Kainen, Kůrková, and Vogt [35], dealing with representing the Gaussian via half-space characteristic functions (i.e., by a perceptron network using the Heaviside function as its sigmoidal activation function), and on those found by the current authors, utilizing Gaussian networks and the Bessel Potential function; see [30].

The chapter is organized as follows. In Section 2, we introduce a general model of approximation from a dictionary which includes one-hidden-layer networks. Section 3 sketches some proofs of universal approximation property of radial-basis function and perceptron networks, while the next section, Section 4, presents quadratic estimates of model complexity following from the Maurey-Jones-Barron Theorem and its extension by Darken et. al. In Section 5 we give geometric estimates of model complexity. Then in Section 6 these estimates are reformulated in terms of norms tailored to computational units. Section 7 shows that neural network approximation does not have a continuous best approximation and Section 8 is devoted to tractability of neural-network approximation. Section 9 is a brief discussion. A summary of the main notations used in the paper is given in Section 10.

2 Dictionaries and Variable-Basis Approximation

Feedforward neural networks compute parametrized sets of functions depending both on the type of computational units and on the type of their interconnections. *Computational units* compute functions depending on two vector variables: an *input vector* and a *parameter vector*. Generally, such units compute functions of the form $\phi : X \times Y \to \mathbb{R}$, where ϕ is a function of two variables, an input vector $x \in X \subseteq \mathbb{R}^d$ and a parameter $y \in Y \subseteq \mathbb{R}^s$, where $\mathbb{R}$ denotes the set of real numbers.

Sets of input-output functions of one-hidden-layer networks with one linear output unit can be formally described as

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

where the set G is called a *dictionary* [20], n is the *number of hidden units*, and w_i, $i = 1, \dots, n$, are output weights. We write $\operatorname{span} G$ for the linear space consisting of all finite linear combinations of elements from G; that is,

$$\operatorname{span} G = \bigcup_{n=1}^{\infty} \operatorname{span}_n G.$$

Such a computational model $\operatorname{span}_n G$ is sometimes called a *variable-basis* scheme [45, 46, 51], in contrast to *fixed-basis schemes* consisting of linear combinations of first n elements from a set G with a fixed linear ordering. Note that also networks with several hidden layers and one linear unit belong to this scheme (however in this case, the set G depends on the number of units in the previous hidden layers). Kernel models, splines with free nodes, and trigonometric polynomials with variable frequencies and phases are also special cases of variable schemes (see the references in [46].

The number n of computational units is often interpreted as *model complexity.* Note that the set of input-output functions of networks with an arbitrary (finite) number of hidden units is just $\operatorname{span} G$.

Dictionaries are usually given as parameterized families of functions modelling computational units. They can be described as sets of the form

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

where $\phi : X \times Y \to \mathbb{R}$ is a function of two variables, an input vector $x \in X \subseteq \mathbb{R}^d$, where d is called *input dimension*, and a parameter $y \in Y \subseteq \mathbb{R}^s$. The most popular dictionaries used in neurocomputing include perceptrons, radial, and kernel units.

An element h of $\operatorname{span}_n G_\phi(X, Y)$ then has the form

$$h = \sum_{i=1}^{n} w_i \phi(\cdot, y_i), \ w_i \in \mathbb{R}, \ y_i \in Y,$$

so h is determined by the function ϕ and $n(s+1)$ real parameters.

In practical applications, inputs are bounded so one often studies networks with inputs in some compact (i.e., closed and bounded) subset X of $\mathbb{R}^d$. However, even if inputs are bounded, one may not know a priori what the bound is, so unbounded inputs are also considered, and the theoretical analysis is sometimes easier in this case.

For suitable choices of ϕ, sets $\operatorname{span}_n G_\phi$ model families of input-output functions implemented by one-hidden-layer neural networks of various types. For example, the *perceptron* with *activation function* $\psi : \mathbb{R} \to \mathbb{R}$ takes

$$\phi(x, (v, b)) = \psi(v \cdot x + b).$$

Geometrically, perceptrons compute functions which are constant on all hyperplanes parallel to the hyperplane $\{x \in \mathbb{R}^d \,|v \cdot x = -b\}$. The scalar b is called *bias* while the components of v are called *inner weights.*

The terms “weights” or “parameters” are used to refer to both inner weights and biases, as well as the *outer* weights, which are the linear coefficients of combination of multiple units. The units might be perceptrons or other types. Thus, the distinction between inner and outer weights for these one-hidden-layer neural networks is just the distinction between the inputs and the outputs of the hidden units.

The most common activation functions are *sigmoidals*, i.e., bounded and measurable functions $\sigma : \mathbb{R} \to \mathbb{R}$ with limits 0 and 1 at $-\infty$ and ∞, resp. In some literature, sigmoidals are also required to be non-decreasing. Widely-used sigmoidals are the *logistic sigmoid* $\sigma(t) := 1/(1 + \exp(-t))$ and the *Heaviside function*, defined as $\vartheta(t) := 0$ for $t < 0$ and $\vartheta(t) := 1$ for $t \geq 0$. For a sigmoid σ, we let

$$P_d^\sigma := \{x \mapsto \sigma(v \cdot x + b) \,|\, v \in \mathbb{R}^d, b \in \mathbb{R}\}, \tag{1}$$

We write H_d instead of P_d^ϑ. Since for $t \neq 0$, $\vartheta(t) = \vartheta(t/|t|)$,

$$H_d := \{x \mapsto \vartheta(e \cdot x + b) \,|\, e \in \mathcal{S}^{d-1}, b \in \mathbb{R}\}, \tag{2}$$

where $\mathcal{S}^{d-1} := \{x \in \mathbb{R}^d \,|\, \|x\| = 1\}$ denotes the sphere in $\mathbb{R}^d$. Thus, H_d is the set of characteristic functions of closed half-spaces of $\mathbb{R}^d$ parameterized by the pair (e, b), where e is the direction of the orthogonal vector to the hyperplane and b is the distance from 0 to the hyperplane along a perpendicular.

Radial-basis functions (*RBF*) are given by

$$\phi(x, (v, b)) := \psi(b\|x - v\|),$$

where $\psi : \mathbb{R} \to \mathbb{R}$, $v \in \mathbb{R}^d$, and $b \in \mathbb{R}$. For a radial unit, the parameter b is called the *width*, and v the *center*. An RBF unit is constant on the set of all x at each, fixed distance from its center. The corresponding sets $\operatorname{span}_n G_\phi$ are called *RBF networks*; a typical activation function for a radial unit is the Gaussian function $\psi(t) = \exp(-t^2) := e^{-t^2}$. This leads to the usual picture of “Gaussian hills” which makes plausible the density of their linear combination.

Note that sigmoidal perceptrons and RBF units are geometrically opposite: perceptrons apply a sigmoidal function to a weighted sum of inputs plus a bias, so they respond to *non-localized* regions of the input space by partitioning it with fuzzy hyperplanes (or sharp ones if the sigmoid is Heaviside’s step function). The functions computed by perceptrons belong to the class of *plane waves.* In contrast, RBF units calculate the distance to a center, multiply it by a width factor and finally apply an activation function which is often an even function – hence they respond to *localized* regions. The functions computed by radial units belong to the class of *spherical waves.*

Although perceptrons were inspired neurobiologically (e.g., [64]), plane waves have long been studied by mathematicians, motivated by various problems from physics (e.g., [10]).

3 The Universal Approximation Property

The first theoretical question concerning a given type of a feedforward network architecture is whether a sufficiently elaborate network of this type can approximate all reasonable functions encountered in applications. In neural network terminology, this capability of a class of neural networks is called the *universal approximation property*, while mathematically it is defined as density of the set of input-output functions of the given class of networks. Recall that a subset F of a normed linear space is *dense* if $\operatorname{cl} F = \mathcal{X}$, where the closure cl is defined by the topology induced by the norm $\|.\|_{\mathcal{X}}$, i.e.,

$$\operatorname{cl} G := \{f \in \mathcal{X} \mid (\forall \varepsilon > 0)(\exists g \in G)(\|f - g\|_{\mathcal{X}} < \varepsilon)\} .$$

Density of sets of input-output functions has been studied in both the sup-norm and $\mathcal{L}^p(X)$-cases. We write $(\mathcal{C}(X), \|\cdot\|_{\sup})$ for the space of all continuous bounded functions on a subset X of $\mathbb{R}^d$ with the supremum norm $\|\cdot\|_{\sup}$, defined for every continuous function on X as

$$\|f\|_{\sup} := \sup_{x \in X} |f(x)| .$$

For $p \in [1, \infty)$ let $(\mathcal{L}^p(X), \|\cdot\|_p)$ denote the set of all equivalence classes (w.r.t. equality up to sets of Lebesgue measure zero) of Lebesgue-measurable functions f on X such that the following $\mathcal{L}^p$-norm is finite:

$$\|f\|_p := \left(\int_X |f(x)|^p dx \right)^{1/p} < \infty .$$

Choice of norm is problem-dependent. Predicting the movement of a robotic welding tool might best utilize the supremum norm, while minimizing cost might be more likely over $\mathcal{L}^2$.

For RBF networks, the universal approximation property is intuitively quite clear - imagine the surface as a combination of Gaussian hills of various widths and heights. The classical method of approximation by convolutions with a suitable sequence of kernels enables one to prove this property for many types of radial functions. For $d \in \mathbb{N}_+$, where $\mathbb{N}_+$ denotes the set of positive integers, and ψ an even function, let $F_d^\psi(X)$ denote the dictionary

$$F_d^\psi(X) := \{f : X \to \mathbb{R} \mid f(x) = \psi(b\|x - v\|),\, v \in \mathbb{R}^d,\, b \in \mathbb{R}\}$$

of functions on $X \subseteq \mathbb{R}^d$ computable by RBF networks with the radial function ψ and the distance from centers measured by a norm $\|\cdot\|$ on $\mathbb{R}^d$. In the following, we shall consider the Euclidean norm.

First, Hartman et al. [22] proved density of RBF networks with Gaussian radial function in $(\mathcal{C}(X), \|.\|_{\sup})$ for X compact convex. This proof used the special property of the Gaussian function that a d-dimensional Gaussian is the product of d one-dimensional Gaussians. Later, Park and Sandberg [61] extended the density property to RBFs with fairly general radial functions in $(\mathcal{L}^p(\mathbb{R}^d), \|.\|_p)$. Their proof exploits classical results on approximation of functions by convolutions with a sequence of kernel functions converging in the distributional sense to the Dirac delta function δ (see, e.g., [73]). The next theorem is from [61]; we sketch the idea of the proof.

Theorem 1 (Park and Sandberg). *For every positive integer d, every $p \in (1, \infty)$, every integrable bounded function $\psi : \mathbb{R} \to \mathbb{R}$ with finite non-zero integral and such that $\psi \in \mathcal{L}^p(\mathbb{R})$, $\operatorname{span} F_d^{\psi}(\mathbb{R}^d)$ is dense in $(\mathcal{L}^p(\mathbb{R}^d), \| \cdot \|_p)$.*

Proof. When $\int_{\mathbb{R}} \psi(t)dt = c \neq 0$, by letting $\psi_0 = \frac{\psi}{c}$ we define a sequence $\{\psi_n(t) \,|\, n \in \mathbb{N}_+\}$ as $\psi_n(t) = n^d \psi_0(nt)$. By a classical result [4, p. 101] (see also [60, Lemma 1]), for every $f \in \mathcal{L}^p(\mathbb{R}^d)$ one has $f = \lim_{n\to\infty} f * \psi_n$ in $\|.\|_p$. Approximating the integrals

$$\int_{\mathbb{R}^d} f(x) \, \frac{n^d}{c} \psi(n(x-v))dv$$

by Riemann sums we get a sequence of functions of the form of RBF networks with ψ as a radial function.

Exploiting a similar classical result on approximation of functions in $(\mathcal{C}(X), \|.\|_{\sup})$ with X compact by convolutions with a sequence of bump functions, one gets an analogous proof of universal approximation property for RBF networks in $(\mathcal{C}(X), \|.\|_{\sup})$. Note that these arguments can be extended to other norms on $\mathbb{R}^d$ than the Euclidean one. Using a more sophisticated proof technique based on Hermite polynomials, Mhaskar [56] showed that for the Gaussian radial function, the universal approximation property (in sup norm) can even be achieved using networks with a given fixed width.

In one dimension, perceptron networks can also be localized as a pair of overlapping sigmoidal units with opposite-sign weights create a “bump” function. Hence, for $d = 1$, every “reasonable” function can be written as a limit of linear combinations of Heaviside perceptron units.

However, in contrast to localized Gaussian radial units and the one-dimensional case, for d greater than 1, the universal approximation property is far from obvious for perceptrons. But mathematics extends the range of visualization and offers tools that enable us to prove universal approximation for perceptrons with various types of activations.

One such tool is the Stone-Weierstrass theorem (Stone’s extension of the classical result of Weierstrass regarding density of polynomials on a compact interval). A family of real-valued functions on a set X *separates points* if for any two distinct points in X there is a function in the family which takes on distinct values at the two points (i.e., for each pair of distinct points $x, y \in X$ there exists f in the family such that $f(x) \neq f(y)$). A family $\mathcal{A}$ of real-valued functions on X is an *algebra* if it is closed with respect to scalar multiplication and with respect to pointwise addition and multiplication of functions, e.g., for $f, g \in \mathcal{A}$, and $r \in \mathbb{R}$, the functions rf, $x \mapsto f(x)+g(x)$, and $x \mapsto f(x)g(x)$ belong to $\mathcal{A}$. The Stone-Weiestrass Theorem (e.g., [65, pp. 146-153]) states that an algebra $\mathcal{A}$ of real-valued functions on any compact set X is dense in $(\mathcal{C}(X), \|\cdot\|_{\sup})$ if and only if $\mathcal{A}$ separates points and is nontrivial in the sense that it contains a nonzero constant function.

For a function $\psi : \mathbb{R} \to \mathbb{R}$ we denote by $P_d^{\psi}(X)$ the dictionary

$$P_d^{\psi}(X) := \{f : X \to \mathbb{R} \,|\, f(x) = \psi(v \cdot x + b),\ v \in \mathbb{R}^d,\ b \in \mathbb{R}\}$$

of functions on $X \subseteq \mathbb{R}^d$ computable by perceptron networks with the activation function ψ. The linear span of this dictionary, $\operatorname{span} P_d^{\psi}(X)$, is closed under addition and for reasonable ψ, it also separate points, but for many ψ it is not closed under multiplication. An exception is the function $\exp(t) = e^t$ which does produce a multiplicatively closed set $P^d_{\exp}(X)$ and so can serve as a tool to prove the universal approximation property for other activation functions.

The following "input-dimension-reduction theorem" by Stinchombe and White [69] exploits properties of $P^d_{\exp}(X)$ to show that for perceptron networks, it suffices to check the universal approximation property for networks with a single input.

Theorem 2 (Stinchombe and White). *Let $\psi : \mathbb{R} \to \mathbb{R}$, d be a positive integer, and X a compact subset of $\mathbb{R}^d$. Then $\operatorname{span} P^1_{\psi}(X)$ is dense in $(\mathcal{C}(X), \|\cdot\|_{\sup})$ if and only if $\operatorname{span} P^d_{\psi}(X)$ is dense in $(\mathcal{C}(X), \|\cdot\|_{\sup})$.*

Proof. By the Stone-Weierstrass Theorem, $\operatorname{span} P^d_{\exp}(X)$ is dense in $(\mathcal{C}(X), \|\cdot\|_{\sup})$. Using composition of two approximations, the first one approximating $f \in \mathcal{C}(X)$ by an element of $\operatorname{span} P^d_{\exp}(X)$, and the second one approximating exp on a suitable compact subset $Y \subset \mathbb{R}$ by an element of $P^1_{\psi}(Y)$, one gets density of $\operatorname{span} P^d_{\psi}(X)$ in $(\mathcal{C}(X), \|\cdot\|_{\sup})$.

The Stone-Weierstrass Theorem was first used by Hornik, Stinchcombe, and White [24] to prove universal approximation property for one-hidden-layer sigmoidal perceptron networks. Later, Leshno et al. [54] characterized activation functions which determine perceptron networks with the universal approximation property. They showed that the universal approximation property is not restricted to (biologically motivated) sigmoidals but, with the exception of polynomials, it is satisfied by any reasonable activation function.

Theorem 3 (Leshno, Lin, Pinkus, and Schocken). *Let $\psi : \mathbb{R} \to \mathbb{R}$ be a locally bounded piecewise continuous function, d be a positive integer, and X a compact subset of $\mathbb{R}^d$. Then $\operatorname{span} P^d_{\psi}(X)$ is dense in $(\mathcal{C}(X), \|.\|_{\sup})$ if and only if ψ is not an algebraic polynomial.*

Proof. We give only a sketch. The trick of Leshno et al.'s proof in [54] consists in expressing all powers as limits of higher-order partial derivatives with respect to the weight parameter v of the function $\psi(v \cdot x + b)$ (ψ being analytic guarantees that all the derivatives exist). It follows directly from the definition of iterated partial derivative that

$$\frac{\partial^k \psi(v \cdot x + b)}{\delta v^k}$$

can be expressed as a limit of functions computable by perceptron networks with activation ψ. More precisely,

$$\frac{\partial \psi(v \cdot x + b)}{\partial v} = \lim_{\eta \to 0} \frac{\psi((v+\eta)x + b) - \psi(v \cdot x + b)}{\eta},$$

and similarly for $k > 1$. Since $\frac{\partial^k \psi(v \cdot x + b)}{\partial v^k} = x^k \psi^{(k)}(v \cdot x + b)$, and for ψ non-polynomial, none of the $\psi^{(k)}$ is identically equal to zero for all k, setting $v = 0$ and choosing some b_k, for which $\psi^{(k)}(b_k) = c_k \neq 0$, one gets a sequence of functions from $\operatorname{span} P_\psi^d(X)$ converging to $c_k x^k$. As all polynomials are linear combinations of powers, they can be obtained as limits of sequences of functions from $\operatorname{span} P_\psi^d(X)$. So by Weierstrass' theorem and Theorem 2, $\operatorname{span} P_d^\psi(X)$ is dense in $(\mathcal{C}(X), \|\cdot\|_{\sup})$ for any ψ which is analytic and non-polynomial. The statement can be extended to nonanalytic functions satisfying the assumptions of the theorem using suitable convolutions with analytic functions.

Inspection of the proof given by Leshno et al. [54] shows that the theorem is valid even when input weights are bounded by an arbitrarily small upper bound. However to achieve density, the set of hidden unit parameters must have either a finite or an infinite accumulation point.

Another standard method for treating approximation problems is based on the Hahn-Banach Theorem: to verify density of a set of functions, it is sufficient to show that every bounded linear functional that vanishes on this set must be equal to zero on the whole linear space. This method was used by Cybenko [11] to establish universal approximation for sigmoidal perceptron networks.

Other proofs of universal approximation property of perceptron networks took advantage of integral representations based on Radon transform (Carroll and Dickinson [7] and Ito [25]) and on Kolmogorov's representation of continuous functions of several variables by means of superpositions of continuous one-variable functions (Kůrková [37]).

In practical applications the domain of the function to be computed by a network is finite and so one can apply results from interpolation theory, which show that for finite domain functions, one can replace *arbitrarily close approximation* by *exact representation.* A major result of interpolation theory, Micchelli's theorem [58], proves that any function on a finite subset of $\mathbb{R}^d$ can be exactly represented as a network with Gaussian RBF units. An analogous result for sigmoidal perceptron networks has been proven by Ito [26].

However, development of interpolation theory has been motivated by the need to construct surfaces with certain characteristics fitted to a small number of points. Although its results are valid for any number of pairs of data, the application to neurocomputing leads to networks with the same number of hidden units as the number of input-output pairs. For large sets of data, this requirement may prevent implementation. In addition, it is well-known that fitting input-output function to all the training data may produce "overfitting", in which characteristics of noise and other random artifacts mask the intrinsic nature of the function. In such cases, estimates of accuracy achievable using networks with fewer hidden units are needed. These can be derived from estimates of rates of approximation by variable basis schemes that apply to both finite and infinite domains.

4 Quadratic Rates of Approximation

The cost of universality is arbitrarily large model complexity and hence the set of network parameters has to also be sufficiently large. Dependence of accuracy of approximation on the number of hidden units can be studied in the framework of approximation theory in terms of *rates of approximation.* In other words, rates of approximation characterize the trade-off between accuracy of approximation and model complexity.

Jones [27] introduced a recursive construction of approximants with rates of order $\mathcal{O}(1/\sqrt{n})$. Together with Barron [2] he proposed to apply it to sets of functions computable by one-hidden-layer sigmoidal perceptron networks. The following theorem is a version of Jones' result as improved by Barron [2]. It was actually discovered and proved first by Maurey (see [63]) using a probabilistic argument to guarantee existence rather than the incremental algorithm given here.

The following theorem restates the Maurey-Jones-Barron's estimate. Its proof is from [41], where the argument of Barron [2, p. 934, Lemma 1] is simplified. Let $\operatorname{conv}_n G$ denote the set of all convex combinations of n elements from the set G, i.e.,

$$\operatorname{conv}_n G := \left\{ \sum_{i=1}^{n} a_i g_i \;\middle|\; a_i \in [0,1],\ \sum_{i=1}^{n} a_i = 1,\ g_i \in G \right\},$$

while $\operatorname{conv} G$ denotes the *convex hull* of G

$$\operatorname{conv} G := \bigcup_{i=1}^{n} \operatorname{conv}_n G.$$

For $\mathcal{X}$ a normed linear space, in approximating an element $f \in \mathcal{X}$ by elements from a subset A of $\mathcal{X}$, the error is the *distance from f to A*,

$$\|f - A\|_{\mathcal{X}} := \inf_{g \in A} \|f - g\|_{\mathcal{X}}.$$

Theorem 4 (Maurey-Jones-Barron). *Let $(\mathcal{X}, \|.\|_{\mathcal{X}})$ be a Hilbert space, G be a non empty bounded subset of $\mathcal{X}$, and $s_G := \sup_{g \in G} \|g\|_{\mathcal{X}}$. Then for every $f \in \mathrm{cl}\,\mathrm{conv}\, G$ and for every positive integer n,*

$$\|f - \mathrm{conv}_n G\|_{\mathcal{X}} \leq \sqrt{\frac{s_G^2 - \|f\|_{\mathcal{X}}^2}{n}}.$$

Proof. Since distance from $\mathrm{conv}_n G$ is continuous on $(\mathcal{X}, \|.\|_{\mathcal{X}})$, i.e., $f \mapsto \|f - \mathrm{conv}_n G\|_{\mathcal{X}}$ is continuous for $f \in \mathcal{X}$ [67, p. 391], it suffices to verify the statement for $f \in \mathrm{conv}\, G$. Let

$$f = \sum_{j=1}^{m} a_j h_j$$

be a representation of f as a convex combination of elements of G. Set

$$c := s_G^2 - \|f\|^2.$$

By induction we construct $\{g_n\}_{n \geq 1} \subseteq G$ such that for all n

$$e_n^2 = \|f - f_n\|^2 \leq \frac{c}{n}, \text{ where } f_n = \sum_{i=1}^{n} \frac{g_i}{n}.$$

Indeed, for the basis case of the induction, note that

$$\sum_{j=1}^{m} a_j \|f - h_j\|^2 = \|f\|^2 - 2\,\langle f, \sum_{j=1}^{m} a_j h_j \rangle + \sum_{j=1}^{m} a_j \|h_j\|^2 \leq s_G^2 - \|f\|^2 = c,$$

so there exists $j \in \{1, \ldots, m\}$ for which $\|f - h_j\|^2 \leq c$. Take $f_1 = g_1 := h_j$.

Suppose $g_1, \ldots, g_n$ satisfy the error-bound. Then

$$e_{n+1}^2 = \|f - f_{n+1}\|^2 = \|\frac{n}{n+1}(f - f_n) + \frac{1}{n+1}(f - g_{n+1})\|^2 =$$

$$= \frac{n^2}{(n+1)^2} e_n^2 + \frac{2n}{(n+1)^2} \langle f - f_n, f - g_{n+1} \rangle + \frac{1}{(n+1)^2} \|f - g_{n+1}\|^2.$$

As in the basis case,

$$\sum_{j=1}^{m} a_j \left(\frac{2n}{(n+1)^2} \langle f - f_n, f - h_j \rangle + \frac{1}{(n+1)^2} \|f - h_j\|^2 \right) =$$

$$= \frac{1}{(n+1)^2}(\sum_{j=1}^{m} a_j g_j - \|f\|^2) \leq \frac{1}{(n+1)^2}(s_G^2 - \|f\|^2) = \frac{c}{(n+1)^2}$$

So there must exist $j \in \{1, \dots, m\}$ such that

$$\frac{2n}{(n+1)^2}\,\langle f - f_n, f - h_j\rangle + \frac{1}{(n+1)^2}\|f - h_j\|^2 \leq \frac{c}{(n+1)^2}.$$

Setting $g_{n+1} := h_j$, we get $e_{n+1}^2 \leq \frac{n^2}{(n+1)^2} e_n^2 + \frac{c}{(n+1)^2} \leq c/(n+1)^2$.

Inspection of the proof shows that approximants in the convex hull of G can always be taken to be barycenters of simplices - i.e., with all parameters of the convex combination equal.

Theorem 4 actually gives an upper bound on *incremental* learning algorithms (i.e., algorithms which at each step add a new hidden unit but do not change the previously-determined inner parameters for the already chosen units, see e.g., Kůrková [40]). In principle, non-incremental algorithms might do better. Darken at al. [12] extended Maurey-Jones-Barron's estimate to $\mathcal{L}^p$-spaces, $p \in (1, \infty)$. They used a more sophisticated argument replacing inner products with peak functionals and taking advantage of Clarkson's inequalities (see [23, pp. 225, 227]). Recall that for a Banach space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ and $f \in \mathcal{X}$, we denote by Π_f a *peak functional for* f, i.e., a continuous linear functional such that $\| \Pi_f \|_{\mathcal{X}} = 1$ and $\Pi_f(f) = \|f\|_{\mathcal{X}}$ [5, p.1].

The next theorem is a slight reformulation of [12, Theorem 5] with a simplified proof.

Theorem 5 (Darken-Donahue-Gurvits-Sontag). *Let $\Omega \subseteq \mathbb{R}^d$ be open, G a subset of $(\mathcal{L}_p(\Omega), \|\cdot\|_p)$, $p \in (1, \infty)$, $f \in \mathrm{cl}\,\mathrm{conv}\, G$, and $r > 0$ such that $G \subseteq B_r(f, \|\cdot\|)$. Then for every positive integer n*

$$\|f - \mathrm{span}_n G\|_p \leq \frac{2^{1/a} r}{n^{1/b}}\,,$$

where $q := p/(p-1)$, $a := \min(p, q)$, and $b := \max(p, q)$.

Proof. As in the proof of Theorem 4, it is sufficient to verify the statement for $f \in \mathrm{conv}\, G$. Let $f = \sum_{j=1}^{m} w_j h_j$ be a representation of f as a convex combination of elements of G. We show by induction that there exist a sequence $\{g_i\}$ of elements of G such that the barycenters $f_n = \sum_{i=1}^{n} \frac{g_i}{n}$ satisfy $e_n := \|f - f_n\| \leq \frac{2^{1/a} r}{n^{1/b}}$.

First we check that there exists $g_1 \in G$ such that $f_1 = g_1$ satisfies $e_1 = \|f - f_1\|_p \leq 2^{1/a}\, r$. This holds trivially as $G \subseteq B_r(f, \|\cdot\|)$, so for any $g \in G$ we have $\|f - g\| \leq r < 2^{1/a}\, r$. Hence we can set $f_1 := g_1$ for any $g_1 \in G$.

Assume that we already have $g_1, \dots, g_n$. Then

$$f_{n+1} = \frac{n}{n+1}\, f_n + \frac{1}{n+1}\, g_{n+1} = \frac{1}{n+1} \sum_{i=1}^{n+1} g_i\,.$$

We shall express e_{n+1}^a in terms of e_n^a.

Let Π_n be a peak functional for $f - f_n$. Since $\sum_{j=1}^m w_j\,(f - h_j) = 0$, by linearity of Π_n we have $0 = \Pi_n\left(\sum_{j=1}^m w_j\,(f-h_j)\right) = \sum_{j=1}^m w_j\,\Pi_n(f-h_j)$. Thus, there must exist $j \in \{1,\dots,m\}$ such that $\Pi_n(f-h_j) \le 0$. Set $g_{n+1} = h_j$, so $\Pi_n(f - g_{n+1}) \le 0$. For every $p \in (1,\infty)$, $q =: p/(p-1)$, $a := \min(p,q)$, and $f, g \in \mathcal{L}^p(\Omega)$, Clarkson's inequalities imply

$$\|f+g\|_p^a + \|f-g\|_p^a \le 2\left(\|f\|_p^a + \|g\|_p^a\right)$$

(see, e.g., [47, Proposition A3 (iii)]. Hence we get

$$\begin{aligned} e_{n+1}^a = \|f - f_{n+1}\|_p^a &= \left\|\frac{n}{n+1}(f-f_n) + \frac{1}{n+1}(f-g_{n+1})\right\|_p^a \\ &\le 2\ \left(\left\|\frac{n}{n+1}\,(f-f_n)\right\|_p^{\bar{p}} + \left\|\frac{1}{n+1}\,(f-g_{n+1})\right\|_p^a\right) \\ &\quad \times \left\|\frac{n}{n+1}\,(f-f_n) - \frac{1}{n+1}\,(f-g_{n+1})\right\|_p^a . \end{aligned} \tag{3}$$

As $\|\Pi_n\| = 1$ and $\Pi_n(f - g_{n+1}) \le 0$, we have $\|\frac{n}{n+1}\,(f - f_n) - \frac{1}{n+1}\,(f - g_{n+1})\|_p \ge \left\|\Pi_n\left(\frac{n}{n+1}\,(f-f_n) - \frac{1}{n+1}\,(f-g_{n+1})\right)\right\|_p \ge \left\|\Pi_n\left(\frac{n}{n+1}\,(f-f_n)\right)\right\|_p = \frac{n}{n+1}\ \|\Pi_n(f-f_n)\|_p = \frac{n}{n+1}\ \|f-f_n\|_p$. Hence

$$-\left\|\frac{n}{n+1}\,(f-f_n) - \frac{1}{n+1}\,(f-g_{n+1})\right\|_p^a \le -\left(\frac{n}{n+1}\ \|f-f_n\|_p\right)^a . \tag{4}$$

By (3) and (4),

$$e_{n+1}^a = \|f - f_{n+1}\|_p^a \le$$

$$2\left(\|\frac{n}{n+1}\,(f-f_n)\|_p^a + \|\frac{1}{n+1}\,(f-g_{n+1})\|_p^a\right) - \left(\frac{n}{n+1}\ \|f-f_n\|_p\right)^a =$$

$$\frac{2}{(n+1)^a}\ \|f-g_{n+1}\|_p^a + \left(\frac{2}{n+1}\right)^a\ \|f-f_n\|_p^a =$$

$$\frac{2}{(n+1)^a}\ \|f-g_{n+1}\|_p^a + \left(\frac{2}{n+1}\right)^a\ e_n^a\,.$$

As $e_n = \|f - f_n\| \le \frac{2^{1/a}\,r}{n^{1/b}}$, we get $e_{n+1}^a \le \frac{2\,r^a}{(n+1)^a} + \left(\frac{n}{n+1}\right)^a\left(\frac{2^{1/a}\,r}{n^{1/b}}\right)^a = \frac{2\,r^a}{(n+1)^a}\ \left(1 + \frac{n^a}{n^{a/b}}\right) = \frac{2\,r^a}{(n+1)^a}\ \left(1 + n^{a-a/b}\right)$. It can easily be verified that $a - \frac{a}{b} = 1$ in both cases, $a = p$ (and so $b = q = \frac{p}{p-1}$) and $a = q$ (and so $b = p$). Thus $e_{n+1}^a \le \frac{2\,r^a}{(n+1)^a}\,(n+1)$. Hence, $e_{n+1}^a \le \frac{2\,r^a}{(n+1)^{a-1}} = \frac{2r^a}{(n+1)^{a/b}}$, i.e., $e_{n+1} \le \frac{2^{1/a}r}{(n+1)^{1/b}}$.

The Maurey-Jones-Barron Theorem and its extension to $\mathcal{L}_p$-spaces received a lot of attention because they imply an estimate of model complexity of the order $\mathcal{O}(\varepsilon^{-2})$. Several authors derived improvements and investigated tightness of the improved bounds for suitable sets G such as G orthogonal [45, 52], G formed by functions computable by sigmoidal perceptrons, and precompact G with certain covering number properties [49, 55]. In particular, for a dictionary G of functions computable by perceptrons with certain sigmoidal activations, impossibility of improving the exponent $-\frac{1}{2}$ in the upper bound from Theorem (4) over $-(\frac{1}{2}+\frac{1}{d})$ (where d is the number of variables) was proved in [1, 55, 49].

To illustrate an improvement of the Maurey-Jones-Barron Theorem, we state an upper bound by Makovoz [55, Theorem 1], which brings in the entropy numbers of G. Recall that the *n-th entropy number* $e_n(G)$ of a subset G of a normed linear space is defined as

$$e_n(G) := \inf\{\varepsilon > 0 \,|\, (G \subseteq \cup_{i=1}^{n} U_i) \,\&\, (\forall i = 1, \ldots, n)\, (\operatorname{diam}(U_i) \leq \varepsilon\,)\},$$

where $\operatorname{diam}(U) = \sup_{x,y \in U} \|x - y\|$.

Theorem 6 (Makovoz). *Let G be a bounded subset of a Hilbert space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$. Then for every $f \in \operatorname{span} G$ of the form $f = \sum_{i=1}^{\infty} c_i g_i$ such that $\sum_{i=1}^{\infty} |c_i| < \infty$ and every positive integer n there exists $g = \sum_{i=1}^{n} a_i\, g_i \in \operatorname{span}_n G$ such that*

$$\|f - g\|_{\mathcal{X}} \leq \frac{2\; e_n(G)\; \sum_{i=1}^{\infty} |c_i|}{\sqrt{n}}\,,$$

where $\sum_{i=1}^{n} |a_i| \leq \sum_{i=1}^{\infty} |c_i|$.

5 Geometric Rates of Approximation

Throughout this section let $(\mathcal{X}, \|.\|_{\mathcal{X}})$ be a Hilbert space with G a non empty bounded subset of $\mathcal{X}$. The Maurey-Jones-Barron Theorem and its improvements presented in the previous section are worst-case estimates (i.e., they give upper bounds holding for all functions from the closure of the symmetric convex hull of G). Thus, one can expect that for suitable subsets of this hull, better rates may hold.

Lavretsky [53] noticed that a certain geometric condition would allow substantial improvement in the Jones-Barron's iterative construction [2, 27]. More precisely, for $\delta > 0$, he defined the set

$$\begin{aligned} F_\delta(G) := \big\{ f \in \operatorname{cl}\operatorname{conv} G \,\big|\, \forall h \in \operatorname{conv} G,\ f \neq h\ \exists g \in G\ :\ \\ (f-g)\cdot(f-h) \leq -\delta\, \|f-g\|_{\mathcal{X}}\, \|f-h\|_{c}X \big\} \end{aligned} \qquad (5)$$

and proved the following result.

Theorem 7 (Lavretsky). *Let $(\mathcal{X}, \|.\|_{\mathcal{X}})$ be a Hilbert space with G any bounded symmetric subset containing 0, $s_G := \sup_{g \in G} \|g\|$, and $\delta > 0$. Then for every $f \in F_\delta(G)$ and every positive integer n,*

$$\|f - \mathrm{conv}_n\, G\|_{\mathcal{X}} \le \sqrt{(1-\delta^2)^{n-1}(s_G^2 - \|f\|_{\mathcal{X}}^2)}\,.$$

In [51], Kůrková and Sanguineti improved the idea of Lavretsky by showing that for every function f in the convex hull of a bounded subset G of a Hilbert space there exists $\tau_f \in [0,1)$ such that the rate of approximation of f by convex combinations of n functions from G is bounded from above by $\sqrt{\tau_f^{n-1}\,(s_G^2 - \|f\|_{\mathcal{X}}^2)}$.

Theorem 8 (Kůrková-Sanguineti). *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a Hilbert space, G its bounded nonempty subset, and $s_G := \sup_{g \in G} \|g\|_{\mathcal{X}}$. For every $f \in \mathrm{conv}\, G$ there exists $\tau_f \in [0,1)$ such that for every positive integer n*

$$\|f - \mathrm{conv}_n\, G\|_{\mathcal{X}} \le \sqrt{\tau_f^{n-1}\ (s_G^2 - \|f\|_{\mathcal{X}}^2\,)}.$$

Proof. Let $f = \sum_{j=1}^m a_j\, g_j$ be a representation of f as a convex combination of elements of G with all $a_j > 0$ (and $\sum_j a_j = 1$). Let G' be the set of elements combined; i.e.,

$$G' := \{g_1, \ldots, g_m\}\,.$$

For each $n = 1, \ldots, m$, we find $f_n \in \mathrm{conv}_n G$, and $\rho_n > 0$ such that

$$\|f - f_n\|_{\mathcal{X}}^2 \le (1-\rho_n^2)^{n-1}\ \left(s_G^2 - \|f\|_{\mathcal{X}}^2\right)\ . \tag{6}$$

Let $g_{j_1} \in G'$ be nearest to f, i.e.,

$$\|f - g_{j_1}\|_{\mathcal{X}} = \min_{g \in G'} \|f - g\|_{\mathcal{X}},$$

and set $f_1 := g_{j_1}$. As

$$\begin{aligned}\sum_{j=1}^m a_j \|f - g_j\|_{\mathcal{X}}^2 &= \|f\|_{\mathcal{X}}^2 - 2f \cdot \sum_{i=1}^m a_j g_j + \sum_{j=1}^m a_j \|g_j\|_{\mathcal{X}}^2 \\ &\le s_G^2 - \|f\|_{\mathcal{X}}^2\,,\end{aligned}$$

we get $\|f - f_1\|_{\mathcal{X}}^2 \le s_G^2 - \|f\|_{\mathcal{X}}^2$ and so (6) holds for $n = 1$ with any $\rho_1 \in (0,1)$.

Assuming that we have f_{n-1}, we define f_n. When $f_{n-1} = f$, we set $f_n := f_{n-1}$ and the estimate holds trivially.

When $f_{n-1} \neq f$, we define f_n as the convex combination

$$f_n := \alpha_n f_{n-1} + (1 - \alpha_n) g_{j_n}, \tag{7}$$

with $g_{j_n} \in G'$ and $\alpha_n \in [0,1]$ chosen in such a way that for some $\rho_n > 0$

$$\|f - f_n\|_{\mathcal{X}}^2 \leq (1 - \rho_n^2)^{n-1} \|f - f_{n-1}\|_{\mathcal{X}}^2 .$$

First, we choose a suitable g_{j_n} and then we find α_n depending on our choice of g_{j_n}. Denoting $e_n := \|f - f_n\|_{\mathcal{X}}$, by (7) we get

$$e_n^2 = \alpha_n^2 e_{n-1}^2 + 2\alpha_n(1-\alpha_n)(f - f_{n-1}) \cdot (f - g_{j_n}) + (1-\alpha_n)^2 \|f - g_{j_n}\|_{\mathcal{X}}^2 . \quad (8)$$

For all $j \in \{1, \dots, m\}$, set

$$\eta_j := -\frac{(f - f_{n-1}) \cdot (f - g_j)}{\|f - f_{n-1}\|_{\mathcal{X}} \, \|f - g_j\|_{\mathcal{X}}}$$

(both terms in the denominator are nonzero: the first one because we are considering the case when $f \neq f_{n-1}$ and the second one because we assume that for all j, $a_j > 0$ and thus $f \neq g_j$). Note that for all j, $\eta_j \in [-1, 1]$ as it is the cosine of the angle between the vectors $f - f_{n-1}$ and $f - g_j$.

As $f = \sum_{j=1}^m a_j \, g_j$, we have

$$\sum_{j=1}^m a_j (f - f_{n-1}) \cdot (f - g_j) = (f - f_{n-1}) \cdot (f - \sum_{j=1}^m a_j g_j) = 0.$$

Thus
(i) either there exists $g \in G'$, for which $(f - f_{n-1}) \cdot (f - g) < 0$
(ii) or for all $g \in G'$, $(f - f_{n-1}) \cdot (f - g) = 0$.

We show that case (ii) can't happen since it would imply that $f = f_{n-1}$. Indeed, $f_{n-1} \in \mathrm{conv}_{n-1} G'$ and thus can be expressed as

$$f_{n-1} = \sum_{k=1}^{n-1} b_k g_{j_k}$$

with all $b_k \in [0, 1]$ and $\sum_{k=1}^{n-1} b_k = 1$. If for all $g \in G'$, $(f - f_{n-1}) \cdot (f - g) = 0$, then $\|f - f_{n-1}\|_{\mathcal{X}}^2$ is equal to

$$(f - f_{n-1}) \cdot (f - \sum_{k=1}^{n-1} b_k g_{j_k}) = \sum_{k=1}^{n-1} b_k (f - f_{n-1}) \cdot (f - g_{j_k}) = 0 \, .$$

By assumption, $f \neq f_{n-1}$, so case (i) must hold. Therefore, the subset

$$G'' := \{ g \in G' \mid (f - f_{n-1}) \cdot (f - g) < 0 \}$$

is nonempty. Let $g_{j_n} \in G''$ be chosen so that

$$\eta_{j_n} = \max_{j=1,\dots,m} \eta_j$$

and set $\rho_n := \eta_{j_n}$. As $G'' \neq \emptyset$, we have $\rho_n > 0$. Let $r_n := \|f - g_{j_n}\|_{\mathcal{X}}$. By (8) we get

$$e_n^2 = \alpha_n^2 e_{n-1}^2 - 2\alpha_n(1-\alpha_n)\rho_n e_{n-1} r_n + (1-\alpha_n)^2 r_n^2. \tag{9}$$

To define f_n as a convex combination of f_{n-1} and g_{j_n}, it remains to find $\alpha_n \in [0,1]$ for which e_n^2 is minimal as a function of α_n. By (9) we have

$$e_n^2 = \alpha_n^2 \left(e_{n-1}^2 + 2\rho_n e_{n-1} r_n + r_n^2\right) - 2\alpha_n \left(\rho_n e_{n-1} r_n + r_n^2\right) + r_n^2. \tag{10}$$

Thus

$$\frac{\partial e_n^2}{\partial \alpha_n} = 2\alpha_n \left(e_{n-1}^2 + 2\rho_n e_{n-1} r_n + r_n^2\right) - 2\left(\rho_n e_{n-1} r_n + r_n^2\right)$$

and

$$\frac{\partial^2 e_n^2}{\partial^2 \alpha_n} = 2\left(e_{n-1}^2 + 2\rho_n e_{n-1} r_n + r_n^2\right).$$

As now we are considering the case when $f \neq f_{n-1}$, we have $e_{n-1} > 0$ and hence $\frac{\partial e_n^2}{\partial^2 \alpha_n} > 0$. So the minimum is achieved at

$$\alpha_n = \frac{\rho_n e_{n-1} r_n + r_n^2}{e_{n-1}^2 + 2\rho_n e_{n-1} r_n + r_n^2}. \tag{11}$$

Plugging (11) into (10) we get

$$e_n^2 = \frac{(1-\rho_n^2)e_{n-1}^2 r_n^2}{e_{n-1}^2 + 2\rho_n e_{n-1} r_n + r_n^2} < \frac{(1-\rho_n^2)e_{n-1}^2 r_n^2}{r_n^2} = (1-\rho_n^2)e_{n-1}^2. \tag{12}$$

Let

$$k := \max\{n \in \{1,\ldots,m\} \mid f_n \neq f_{n-1}\}.$$

Setting

$$\rho_f := \min\{\rho_n \mid n = 1,\ldots,k\},$$

by induction we get the upper bound

$$\|f - \mathrm{conv}_n\, G\|_{\mathcal{X}}^2 \leq (1-\rho_f^2)^{n-1} \left(s_G^2 - \|f\|_{\mathcal{X}}^2\right)$$

holding for all n (for $n > m$ it holds trivially with $f_n = f$). We conclude by setting $\tau_f := 1 - \rho_f^2$.

We illustrated Theorem 8 in [51] by estimating values of parameters of geometric rates when G is an orthonormal basis. We derived also insights into the structure of sets of functions with fixed values of parameters of such rates. As for Theorem 4, the proof of Theorem 8 is based on a constructive incremental procedure, described in [51]. For every function $f \in \operatorname{conv} G$ and its any representation $f = \sum_{j=1}^{m} a_j g_j$ as a convex combination of elements of G, the proof constructs a linear ordering

$$\{g_{j_1}, \dots, g_{j_m}\}$$

of the subset

$$G' := \{g_1, \dots, g_m\} .$$

Then it shows that for every positive integer $n \leq m$ and for some $\tau_f \in [0,1)$ one has

$$\| f - \operatorname{span}\{g_{j_1}, \dots, g_{j_m}\} \|_{\mathcal{X}}^2 \leq \tau_f^{n-1} \left(s_G^2 - \|f\|_{\mathcal{X}}^2 \right) .$$

Table 1 describes this procedure.

The geometric bound would be more useful if one could calculate τ_f as a functional of f. Nevertheless, this bound shows the possibility of substantial improvement for suitably nice functions f.

The speed of decrease of the estimate depends on $\tau_f \in [0,1)$ which is obtained as the smallest cosine of the angles between functions used in the construction of approximants. Inspection of the proof shows that the parameter τ_f is not defined uniquely. It depends on the choice of a representation of $f = \sum_{j=1}^{m} a_j g_j$ as a convex combination of elements of G and on the choice of g_{j_n} for those positive integers n, for which there exist more than one g_j with the same cosine ρ_n. However, the *minimal parameter*, for which the geometric upper bound from Theorem 8 holds, is unique.

Let

$$\tau(f) := \min\{\tau > 0 \mid \|f - \operatorname{conv}_n G\|_{\mathcal{X}}^2 \leq \tau^{n-1}(s_G^2 - \|f\|^2)\}. \qquad (13)$$

By Theorem 8, for every $f \in \operatorname{conv} G$ the set over which the minimum in (13) is taken is nonempty and bounded. It follows from the definition of this set that its infimum is achieved, i.e., it is a minimum. Therefore,

$$\|f - \operatorname{conv}_n G\|_{\mathcal{X}} \leq \sqrt{\tau(f)^{n-1}(s_G^2 - \|f\|_{\mathcal{X}}^2)} .$$

6 Approximation of Balls in Variational Norms

The Maurey-Jones-Barron Theorem is a useful tool for estimation of rates of variable-basis approximation. Since $\operatorname{conv}_n G \subseteq \operatorname{span}_n G$, the upper bound from Theorem 4 on approximation by $\operatorname{conv}_n G$ also applies to approximation by $\operatorname{span}_n G$. When G is bounded, $\operatorname{conv} G$ is a proper subset of $\operatorname{span} G$ and so $\operatorname{cl} \operatorname{conv} G$ is a proper subset of $\operatorname{cl} \operatorname{span} G$; thus, Theorem 4 cannot be applied to all elements of $\mathcal{X}$. However, its corollary on approximation by $\operatorname{span}_n G$ applies to all functions in $\mathcal{X}$. Indeed, replacing the set G by sets of the form

Table 1 The incremental construction used in the proof of Theorem 8

```
Choose $g_{j_1} \in \{g_j \,|\, j = 1, \dots, m\}$ such that
        $\|f - g_{j_1}\| = \min_{j=1,\dots,m} \|f - g_j\|$;
$f_1 := g_{j_1}$;
For $n = 2, \dots, m-1$:
  begin
    for $j = 1, \dots, m$,
        compute $\eta_j := -\frac{(f-f_{n-1})\cdot(f-g_j)}{\|f-f_{n-1}\| \, \|f-g_j\|}$
    if for $j = 1, \dots, m$ one has $\eta_j = 0$,   then
      end
    else
      begin
        $\rho_n := \max\{\eta_j > 0 \,|\, j = 1, \dots, m\}$;
        choose $g_{j_n}$ such that $\rho_n = \eta_{j_n}$;
        compute $e_{n-1} := \|f - f_{n-1}\|$;
        compute $r_n := \|f - g_{j_n}\|$;
        compute $\alpha_n := \frac{\rho_n e_{n-1} r_n + r_n^2}{e_{n-1}^2 + 2\rho_n e_{n-1} r_n + r_n^2}$;
        $f_n := \alpha_n \, f_{n-1} + (1 - \alpha_n) \, g_n$;
        $n := n + 1$.
      end
  end
Let
        $k := \max\{n \in \{1, \dots, m\} \mid f_n \neq f_{n-1}\}$
and
        $\rho_f := \min\{\rho_n \mid n = 1, \dots, k\}$

        $\tau_f := 1 - \rho_f^2$
```

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

with $c > 0$, we get $\operatorname{conv}_n G(c) \subset \operatorname{span}_n G(c) = \operatorname{span}_n G$ for any $c \in \mathbb{R}$. 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).

Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a normed linear space and G be its bounded non empty subset, then *G-variation* (variation with respect to G) is defined as the Minkowski functional of the set $\operatorname{cl}\operatorname{conv}(G \cup -G)$, where $-G := \{f \in \mathcal{X} \,|\, f = -g,\ g \in G\}$, i.e.,

$$\|f\|_G := \inf\{c > 0 \mid f/c \in \operatorname{cl}\operatorname{conv}(G \cup -G)\}.$$

Note that G-variation can be infinite and that it is a norm on the subspace of $\mathcal{X}$ formed by those $f \in \mathcal{X}$, for which $\|f\|_G$ is finite. The closure in its definition depends on the topology induced on $\mathcal{X}$ by the norm $\|\cdot\|_{\mathcal{X}}$. However, when $\mathcal{X}$ is finite dimensional, G-variation does not depend on the choice of a norm on $\mathcal{X}$, since all norms on a finite-dimensional space induce the same topology.

Intuitively, G-variation of f measures how much the set G needs to be inflated to contain f in the closure of its symmetric convex hull. It is easy to check that

$$\|\cdot\|_{\mathcal{X}} \leq s_G \|\cdot\|_G\,, \tag{14}$$

where $s_G := \sup_{g\in G} \|g\|_{\mathcal{X}}$. Indeed, if for $b > 0$, $f/b \in \operatorname{cl}\operatorname{conv}(G \cup -G)$, then $f/b = \lim_{\varepsilon\to 0} h_\varepsilon$, where $h_\varepsilon \in \operatorname{conv}(G \cup -G)$ and so $\|h_\varepsilon\| \leq s_G$. Thus, $\|f\|_{\mathcal{X}} \leq s_G\, b$. Hence, by the definition of $\|f\|_G$ we have $\|f\|_{\mathcal{X}} \leq s_G\, \|f\|_G$.

Variation with respect to G was introduced by Kůrková [39] as an extension of Barron's [1] concept of *variation with respect to half-spaces* corresponding to G formed by functions computable by Heaviside perceptrons, defined as

$$H_d := \{x \mapsto \vartheta(e \cdot x + b) \,|\, e \in \mathcal{S}^{d-1}, b \in \mathbb{R}\},$$

where $\mathcal{S}^{d-1} := \{x \in \mathbb{R}^d \,|\, \|x\| = 1\}$ denotes the sphere of radius 1 in $\mathbb{R}^d$ (recall that ϑ is the Heaviside function, defined as $\vartheta(t) := 0$ for $t < 0$ and $\vartheta(t) := 1$ for $t \geq 0$). In particular, if f is an input-output function of a one-hidden-layer network with Heaviside perceptrons, then variation with respect to half-spaces of f is equal to the sum of absolute values of output weights. For $d = 1$, variation with respect to half-spaces is up to a constant equal to total variation which plays an important role in integration theory. G-variation is also an extension of the notion of ℓ_1-norm. When G is an orthonormal basis of a separable Hilbert space, G-variation is equal to the *ℓ_1-norm with respect to G*, which is defined for every $f \in X$ as $\|f\|_{1,G} := \sum_{g\in G} |f \cdot g|$ [18, 19, 45, 52].

Approximation capabilities of sets of functions can be studied in terms of *worst-case errors*, formalized by the concept of *deviation*. For two subsets A and M of $\mathcal{X}$, the deviation of M from A is defined as

$$\delta(M,A) = \delta(M,A;\mathcal{X}) = \delta(M,A;(\mathcal{X},\|\cdot\|_{\mathcal{X}})) := \sup_{f\in M} \|f - A\|_{\mathcal{X}} \,. \tag{15}$$

We use the abbreviated notations when the ambient space and the norm are clear from the context. When the supremum in (15) is achieved, deviation is the *worst-case error* in approximation of functions from M by functions from A. In this section, we consider the case in which the set M of functions to be approximated is a ball in G-variation. By the definition, the unit ball in G-variation is the closure in the norm $\|.\|_{\mathcal{X}}$ of the symmetric convex hull of G, i.e.,

$$B_1(\|.\|_G) := \mathrm{cl}\left(\mathrm{conv}\,(G \cup -G)\right). \tag{16}$$

The estimates given in the next corollary follow by the Maurey-Jones-Barron Theorem (here Theorem 4) and its extension by Darken et al. (Theorem 5). These estimates give upper bounds on rates of approximation by $\mathrm{span}_n G$ for all functions in a Hilbert space and in $\mathcal{L}^p(X)$ spaces with $p \in (1,\infty)$.

Corollary 1. *Let $(\mathcal{X},\|.\|_{\mathcal{X}})$ be a Banach space, G its bounded nonempty subset, $s_G := \sup_{g\in G}\|g\|_{\mathcal{X}}$. Then for every $f \in \mathcal{X}$ and every positive integer n,*
(i) for $(\mathcal{X},\|.\|_{\mathcal{X}})$ a Hilbert space,

$$\|f - \mathrm{span}_n G\|_{\mathcal{X}} \le \sqrt{\frac{s_G^2\|f\|_G^2 - \|f\|_{\mathcal{X}}^2}{n}}\,,$$

so

$$\delta(B_r(\|.\|_G), \mathrm{span}_n\, G) \le \frac{r\,s(G)}{n^{1/2}}\,;$$

(ii) for $(\mathcal{X},\|.\|_{\mathcal{X}}) = (\mathcal{L}^p(X),\|.\|_{\mathcal{L}^p(X)})$ with $p\in(1,\infty)$,

$$\|f - \mathrm{span}_n G\|_{\mathcal{X}} \le \frac{2^{1+1/a}\, s_G\|f\|_G}{n^{1/b}}\,,$$

where $a := \min(p, \frac{p}{p-1})$ and $b := \max(p, \frac{p}{p-1})$, so

$$\delta(B_r(\|.\|_{G,\mathcal{L}^p(X)}), \mathrm{span}_n\, G_d)_{\mathcal{L}^p(X)} \le 2^{1+1/a}\,\frac{r\,s(G)}{n^{1/b}}\,.$$

By the upper bounds from Corollary 1, all functions from the unit ball in G-variation can be approximated within $s_G/\sqrt{n}$ or $2^{1+1/a}s_G/n^{1/b}$ by networks with n hidden units from the dictionary G, independently on the number d of variables. For this reason, such estimates are sometimes called "dimension-independent", which is misleading since with increasing number of variables, the condition of being in the unit ball in G-variation becomes more and more constraining.

Note that since $0 \in \mathrm{span}_n G$, we have for all $f \in \mathcal{X}$, $\|f - \mathrm{span}_n G\|_{\mathcal{X}} \le \|f\|$, thus the bound from Corollary 1 is nontrivial only when $\|f\|_{\mathcal{X}}^2 \ge \frac{(s_G\|f\|_G)^2 - \|f\|_{\mathcal{X}}^2}{n}$ or equivalently $\frac{\|f\|_{\mathcal{X}}}{s_G\|f\|_G} \ge \frac{1}{\sqrt{n+1}}$. For example, for $s_G = 1$ and $\|f\|_G = 1$, this implies that $\|f\|_{\mathcal{X}} \ge \frac{1}{\sqrt{n+1}}$.

Properties of balls in variation corresponding to standard hidden units are not yet well understood. Such balls can be described as subsets of images of balls in $\mathcal{L}_1$-norm under certain integral transformations (see, e.g., [2, 15, 43]).

In [52], examples of functions with variation with respect to half-spaces (i.e., with respect to Heaviside perceptrons) growing exponentially with the number of variables d were given. However, such exponentially-growing lower bounds on variation with respect to half-spaces are merely lower bounds on an upper bound on rates of approximation. They do not prove that such functions cannot be approximated by networks with n perceptrons with faster rates than $\frac{s_G \, \|f\|_G}{\sqrt{n}}$.

Combining Theorem 6 with estimates of entropy numbers, Makovoz [55] disproved the possibility of a substantial improvement of the upper bound from Corollary 1 (i) for the set G corresponding to perceptrons with certain activations. We say that a sigmoidal is *polynomially quickly approximating the Heaviside* if there exist η, $C > 0$ such that for all $t \in \mathbb{R}$ one has $|\sigma(t) - \vartheta(t)| \leq C\,|t|^{\eta}$. The following theorem states Makovoz' result in terms of variation norm.

Theorem 9 (Makovoz). *Let d be a positive integer, σ either the Heaviside function or a Lipschitz continuous sigmoidal polynomially quickly approximating the Heaviside, and $X \subset \mathbb{R}^d$ compact. If $\tau > 0$ is such that for some $c > 0$ and all positive integers n one has*

$$\delta\big(B_1(\|.\|_{P^d_\sigma(X)})), \mathrm{conv}_n\, P^d_\sigma(X)\big) \leq \frac{c}{n^\tau}\,,$$

then $\tau \leq \frac{1}{2} + \frac{1}{d}$.

Hence, for a wide family of sigmoidal perceptron networks the term $n^{-1/2}$ cannot be improved beyond $n^{-1/2-1/d}$, so in high dimension, $n^{-1/2}$ is essentially best possible.

In [49], Kůrková and Sanguineti extended this tightness result to more general approximating sets. Recall that the *ε-covering number* of a subset G of $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ is the cardinality of a minimal ε-net in G, i.e.,

$$\mathcal{N}(G,\varepsilon) := \min\Big\{m \in \mathbb{N}_+ \mid \exists f_1, \ldots, f_m \in G \text{ such that } G \subseteq \bigcup_{i=1}^{m} B_\varepsilon(f_i, \|\cdot\|_{\mathcal{X}})\Big\}.$$

If the set over which the minimum is taken is empty, then $\mathcal{N}(G,\varepsilon) = +\infty$. When there exists $\beta > 0$ such that $\mathcal{N}(G,\varepsilon) \leq \left(\frac{1}{\varepsilon}\right)^\beta$ for $\varepsilon \downarrow 0$, G is said to have *power-type covering numbers.*

For a subset A of a normed linear space $(\mathcal{X}, \|.\|_{\mathcal{X}})$ and a positive integer r, we denote

$$A_r := \left\{ f \in A \,\middle|\, \|f\|_{\mathcal{X}} \geq \frac{1}{r} \right\}.$$

The larger the sets A_r, the slower the decrease of the norms of the elements of A. When A_r is finite for all positive integers r, we call the function $\alpha_A : \mathbb{N}_+ \to \mathbb{N}_+$ defined as

$$\alpha_A(r) := \operatorname{card} A_r$$

the *decay function of* A, where $\operatorname{card} A_r$ denotes the number of elements of A_r. A set A such that A_r is finite for all positive integers r is *slowly decaying with respect to* γ if there exists $\gamma > 0$ such that $\alpha_A(r) = r^\gamma$. Note that if A is a precompact subset of a Hilbert space and A_r is orthogonal, then A_r must be finite. Thus decay functions are defined for all precompact orthogonal subsets of Hilbert spaces and also for subsets $A = \bigcup_{r=1}^{\infty} A_r$ with all A_r orthogonal but A not necessarily orthogonal. Finally, we call *slowly decaying* a set A formed by d-variable functions with the decay function $\alpha_A(r) = r^d$.

Theorem 10 (Kůrková-Sanguineti). *Let* $(\mathcal{X}, \|.\|_{\mathcal{X}})$ *be a Hilbert space,* G *its bounded precompact subset with* $s_G = \sup_{g\in G} \|g\|$ *and power-type covering numbers; let* $t > 0$ *and* $\gamma > 0$*, and* $B_1(\|.\|_G) \supseteq t\, A$*, where* A *is slowly decaying with respect to* γ*. If* $\tau > 0$ *is such that for some* $c > 0$ *and all positive integers* n *one has*

$$\delta\big(B_1(\|.\|_G), \operatorname{conv}_n(G \cup -G)\big) \leq \frac{c}{n^\tau}\,,$$

then $\tau \leq \frac{1}{2} + \frac{1}{\gamma}$.

The proof of Theorem 10 exploits characteristics of generalized Hadamard matrices. A Hadamard matrix is a $d \times d$ matrix of ± 1 entries such that the rows are pairwise-orthogonal; i.e., they have dot product of zero. An $r \times d$ matrix of ± 1s is called quasi-orthogonal if the dot product of any two distinct rows is small compared to d. When the dot product is bounded in absolute value by some constant t, then Kainen and Kůrková [29] showed that as d goes to infinity, the maximum number of rows in a quasi-orthogonal matrix grows exponentially.

It was proven in [49] that Theorem 6 follows by Theorem 10 applied to the set $P_d^\sigma(X)$ of functions computable by perceptrons, where σ is either the Heaviside function or a Lipschitz continuous sigmoidal polynomially quickly approximating the Heaviside.

7 Best Approximation and Non-continuity of Approximation

To estimate rates of variable-basis approximation, it is helpful to study properties like existence, uniqueness, and continuity of corresponding approximation operators.

Existence of a best approximation has been formalized in approximation theory by the concept of proximinal set (sometimes also called "existence" set). A subset M of a normed linear space $(\mathcal{X}, \|.\|_{\mathcal{X}})$ is called *proximinal* if for every $f \in X$ the distance $\|f - M\|_{\mathcal{X}} = \inf_{g \in M} \|f - g\|_{\mathcal{X}}$ is achieved for some element of M, i.e., $\|f - M\|_{\mathcal{X}} = \min_{g \in M} \|f - g\|_{\mathcal{X}}$ (Singer [67]). Clearly a proximinal subset must be closed. On the other hand, for every f in $\mathcal{X}$, the distance-from-f function is continuous on $\mathcal{X}$ [67, p. 391] and hence on any subset M. When M is compact, therefore, it is necessarily proximinal.

Two generalizations of compactness also imply proximinality. A set M is called *boundedly compact* if the closure of its intersection with any bounded set is compact. A set M is called *approximatively compact* if for each $f \in X$ and any sequence $\{g_i\}$ in M such that $\lim_{i \to \infty} \|f - g_i\|_{\mathcal{X}} = \|f - M\|_{\mathcal{X}}$, there exists $g \in M$ such that $\{g_i\}$ converges subsequentially to g [67, p. 368]. Any closed, boundedly compact set is approximatively compact, and any approximatively compact set is proximinal [67, p. 374].

We investigate the existence property for one-hidden-layer Heaviside perceptron networks. Gurvits and Koiran [21] have shown that for all positive integers d the set H_d of characteristic functions of closed half-spaces in $\mathbb{R}^d$ intersected with the unit cube $[0,1]^d$ is compact in $(\mathcal{L}^p([0,1]^d), \|.\|_p)$ with $p \in [1, \infty)$. This can be easily verified once the set H_d is reparametrized by elements of the unit sphere S^d in $\mathbb{R}^{d+1}$. Indeed, a function $\vartheta(v \cdot x + b)$, with the vector $(v_1, \ldots, v_d, b) \in \mathbb{R}^{d+1}$ nonzero, is equal to $\vartheta(\hat{v} \cdot x + \hat{b})$, where $(\hat{v}_1, \ldots, \hat{v}_d, \hat{b}) \in S^d$ is obtained from $(v_1, \ldots, v_d, b) \in \mathbb{R}^{d+1}$ by normalization. Since S^d is compact, so is H_d. However, $\mathrm{span}_n H_d$ is neither compact nor boundedly compact for any positive integers n, d.

The following theorem from [34] shows that $\mathrm{span}_n H_d$ is approximatively compact in $\mathcal{L}^p$-spaces. It extends a result of Kůrková [38], who showed that $\mathrm{span}_n H_d$ is closed in $\mathcal{L}^p$-spaces with $p \in (1, \infty)$.

Theorem 11 (Kainen-Kůrková-Vogt). *Let d be any positive integer. Then $\mathrm{span}_n H_d$ is an approximatively compact subset of $(\mathcal{L}^p([0,1]^d, \|.\|_p)$ for $n \geq 1$ and $p \in [1, \infty)$.*

Theorem 11 shows that for all positive integers n, d a function in $\mathcal{L}^p([0,1]^d)$ with $p \in [1, \infty)$ has a best approximation among functions computable by one-hidden-layer networks with a single linear output unit and n Heaviside perceptrons in the hidden layer. Thus for any p-integrable function on $[0,1]^d$ there exists a linear combination of n characteristic functions of closed half-spaces that is nearest in the $\mathcal{L}^p$-norm. In other words, in the space of parameters of networks of this type, there exists a global minimum of the error functional defined as $\mathcal{L}^p$-distance from the function to be approximated. A related proposition is proved by Chui, Li, and Mhaskar in [9], where certain sequences are shown to have subsequences that converge almost everywhere (a. e.). These authors work in $\mathbb{R}^d$ rather than $[0,1]^d$ and show a. e. convergence rather than $\mathcal{L}^p$-convergence.

Theorem 11 cannot be extended to perceptron networks with differentiable activation functions, e.g., the logistic sigmoid or hyperbolic tangent. For such functions, the sets

$$\mathrm{span}_n P_d(\psi),$$

where

$$P_d(\psi) := \{f : [0,1]^d \to \mathbb{R} \,|\, f(x) = \psi(v \cdot x + b), v \in \mathbb{R}^d, b \in \mathbb{R}\},$$

are not closed and hence cannot be proximinal. This was first observed by Girosi and Poggio [16] and later exploited by Leshno et al. [54] for a proof of the universal approximation property.

Cheang and Barron [8] showed that linear combinations of characteristic functions of closed half-spaces with relatively few terms can yield good approximations of such functions as the characteristic function χ_B of a ball. However, χ_B is not approximated by the linear combination itself but rather by the characteristic function of the set where the linear combination exceeds a certain threshold. This amounts to replacing a linear output in the corresponding neural network by a threshold unit.

Note that Theorem 11 does not offer any information on the error of the best approximation. Estimates on such errors available in the literature (e.g., DeVore, Howard, and Micchelli [13], Pinkus [62]) are based on the assumption that the best approximation operators are continuous. However, it turns out that continuity of such operators may not hold [33], [32] as we now explain.

Recall [67] that for a subset M of a normed linear space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ and $f \in \mathcal{X}$, the *(metric) projection* $\mathrm{Pr}_M(f)$ of f to M is the set of elements in M at minimum distance from f; i.e.,

$$\mathrm{Pr}_M(f) := \{g \in M \mid \|f - g\|_{\mathcal{X}} = \|f - M\|_{\mathcal{X}} = \inf_{h \in M} \|f - h\|_{\mathcal{X}}\}\,.$$

When f is in M, it is its own metric projection. An element of $\mathrm{Pr}_M(f)$ is called a *best approximation* to f *from* M. A mapping $\Psi : \mathcal{X} \to M$ is called a *best approximation mapping* (to elements of $\mathcal{X}$ from M) with respect to $\|\cdot\|_{\mathcal{X}}$ if it maps every element of $\mathcal{X}$ into its projection in M, i.e., for every $f \in \mathcal{X}$ one has $\Psi(f) \in \mathrm{Pr}_M(f)$, that is, $\|f - \Psi(f)\|_{\mathcal{X}} = \|f - M\|_{\mathcal{X}}$.

A classical result from approximation theory [67] states that when $\mathcal{X}$ is a uniformly convex Banach space (for example an $\mathcal{L}^2$-space), the best approximation mapping to a closed convex subset is unique and continuous. This has a basic consequence in linear approximation: it means that for every element f of such a space, there exists a unique linear combination of fixed basis functions (i.e., a unique element of a linear approximating subspace) that minimizes the distance from f and that such a best approximation varies continuously as f is varied.

The situation is different when one considers approximation by neural networks. This is mainly due to the fact that, instead of a finite-dimensional subspace, the approximating functions belong to the union $\mathrm{span}_n\, G$ of finite-dimensional subspaces spanned by all n-tuples of elements of G. The following result from [33, Theorem 2.2] (see also [32]) states the non-existence of continuous best approximation by $\mathrm{span}_n\, G$ in $\mathcal{L}^p$-spaces, $p \in (1, \infty)$. By $\mathrm{card}\, G$ we denote the number of elements of G.

Theorem 12 (Kainen-Kůrková-Vogt). *Let X be a measurable subset of $\mathbb{R}^d$, n a positive integer, and G a linearly independent subset of $\mathcal{L}^p(X)$, $p \in (1, \infty)$, with* $\operatorname{card} G > n$. *Then there exists no continuous best approximation of $\mathcal{L}^p(X)$ by* $\operatorname{span}_n G$.

According to Theorem 12, in $\mathcal{L}^p$-spaces with $p \in (1, \infty)$, for every positive integer n and every linearly independent subset G with $\operatorname{card} G > n$ there is no continuous best approximation mapping to $\operatorname{span}_n G$. As regards the requirement of linear independence of sets of functions representing neural networks, it was proved for the hyperbolic tangent as hidden unit, for certain Heaviside networks, and for Gaussian radial-basis functions. A characterization of linearly independent families for different types of activation functions was given in [44].

Combining Theorem 11 with Theorem 12, we conclude that while best approximation operators exist from $\mathcal{L}^p([0,1]^d)$ to $\operatorname{span}_n H_d$, they cannot be continuous for $p \in (1, \infty)$. This loss of continuity has important consequences on estimates of approximation rates by neural networks. In particular, the lack of continuous dependence in approximation by neural networks does not allow one to apply *a priori* the lower bounds available for linear approximators. In contrast to deviation from a single subspace, deviation from $\operatorname{span}_n G$ which is a union of many such subspaces is much more difficult to estimate since, as we have seen, with the exception of some marginal cases, best approximation mappings to such unions do not posess the good properties of best approximation mapping to a single linear subspace.

8 Tractability of Approximation

8.1 A Shift in Point-of-View: Complexity and Dimension

Only recently has the influence of input dimension on approximation accuracy and rate been studied. Input dimension d is the number of distinct one-dimensional input channels to the computational units. So if a chip-bearing structure like an airplane's wing is providing $400,000$ distinct channels of information, then $d = 400,000$. Some experimental results have shown that optimization over connectionistic models built from relatively few computational units with a simple structure can obtain surprisingly good performances in selected optimization tasks (seemingly high-dimensional); see, e.g., [17, 28, 48, 50, 51, 59, 68, 74, 75] and the references therein. Due to the fragility and lack of theoretical understanding even for these examples, together with the ever-growing amount of data provided by new technology, we believe it is important to explicitly consider the role of d in the theory. Algorithms might require an exponential growth in time and resources as d increases [3] and so even powerful computers would be unable to handle them - hence, they would not be feasible.

On the other hand, in applications, functions of hundreds of variables have been approximated quite well by networks with only a moderate number of hidden units (see, e.g., NETtalk in [66]). Estimates of rates of approximation by neural networks derived from constructive proofs of the universal approximation property have not been able to explain such successes since the arguments in these papers lead to networks with complexity growing exponentially with the number of input units, e.g., $\mathcal{O}(1/\sqrt[d]{n})$. Current theory only predicts that to achieve an accuracy within ε, approximating functions of complexity of order $(1/\varepsilon)^d$ are required.

Some insights into properties of sets of multivariable functions that can be approximated by neural networks with good rates can be derived from the results of the previous sections. The approximation rates that we presented typically include several factors, one of which involves the number n of terms in the linear combinations, while another involves the number d of inputs to computational units. Dependence on dimension d can be implicit; i.e., estimates involve parameters that are constant with respect to n but *do* depend on d and the manner of dependence is not specified; see, e.g., [1, 2, 6, 12, 14, 15, 21, 27, 55]. Terms depending on d are referred to as “constants” since these papers focus on the number n of computational units and assume a fixed value for the dimension d of the input space. Such estimates are often formulated as $O(\kappa(n))$, where dependence on d is hidden in the “big O” notation [36]. However, in some cases, such “constants” actually grow at an exponential rate in d [52, 42]. Moreover, the families of functions for which the estimates are valid may become negligibly small for large d [46].

In general dependence of approximation errors on d may be harder to estimate than dependence on n [71] and few such estimates are available. Deriving them can help to determine when machine-learning tasks are feasible. The role of d is considered explicitly in information-based complexity (see [70, 71, 72]) and more recently this situation has been studied in the context of functional approximation and neural networks [30, 57].

8.2 Measuring Worst-Case Error in Approximation

We focus on upper bounds on worst-case errors in approximation from dictionaries, formalized by the concept of deviation defined in equation (15). An important case is when the deviation of a set A_d of functions of d-variables from the set $\operatorname{span}_n G_d$ takes on the factorized form

$$\delta(A_d, \operatorname{span}_n G_d)_{\mathcal{X}_d} \leq \xi(d)\, \kappa(n)\,. \tag{17}$$

In the bound (17), dependence on the number d of variables and on model complexity n are separated and expressed by the functions $\xi : \mathbb{N}' \to \mathbb{R}_+$ and $\kappa : \mathbb{N}_+ \to \mathbb{R}$, respectively, with $\mathbb{N}'$ an infinite subset of the set $\mathbb{N}_+$ of positive integers and κ nonincreasing nonnegative. Such estimates have been derived, e.g., in [1, 2, 6, 12, 27, 35].

Definition 1. *The problem of approximating* A_d *by elements of* $\operatorname{span}_n G_d$ *is called* **tractable** *with respect to* d *in the worst case or simply tractable if in the upper bound (17) for every* $d \in \mathbb{N}'$ *one has* $\xi(d) \leq d^\nu$ *for some* $\nu > 0$ *and* κ *is a nonincreasing function. We call the problem* **hyper-tractable** *if the upper bound (17) holds with* $\lim_{d\to\infty} \xi(d) = 0$ *and* κ *is nonincreasing.*

Thus, if approximation of A_d by $\operatorname{span}_n G_d$ is hyper-tractable, then the scaled problem of approximating $r_d A_d$ by $\operatorname{span}_n G_d$ is tractable, unless r_d grows faster than $\xi(d)^{-1}$. If $\xi(d)$ goes to zero at an exponential rate, then the scaled problem is hyper-tractable if r_d grows polynomially.

When $\kappa(n) = n^{-1/s}$, the input-output functions of networks with

$$n \geq \left(\frac{\xi(d)}{\varepsilon}\right)^s$$

computational units can approximate given class of functions within ε. If $\xi(d)$ is polynomial, model complexity is a polynomial in d.

In [31], we derived conditions on sets of functions which can be tractably approximated by various dictionaries, including cases where such sets are large enough to include many smooth functions on $\mathbb{R}^d$ (for example, d-variable Gaussian functions on $\mathbb{R}^d$ in approximation by perceptron networks). Even if $\xi(d)$ is a polynomial in d, this may not provide sufficient control of model complexity unless the degree is quite small. For large dimension d of the input space, even quadratic approximation may not be sufficient. But there are situations where dependence on d is *linear* or better [31], and cases are highlighted in which the function $\xi(d)$ decreases exponentially fast with dimension.

As the arguments and proof techniques exploited to derive the results in [31] are quite technical, here we report only some results on tractability of approximation by Gaussian RBF networks and certain perceptron networks, presented in two tables. In the following, for a norm $\|\cdot\|$ in a space of d-variable functions and $r_d > 0$, we denote by $B_{r_d}(\|\cdot\|)$ the ball of radius r_d in such a norm.

8.3 Gaussian RBF Network Tractability

We first consider tractability of approximation by Gaussian RBF networks. The results are summarized in Table 2. To explain and frame the results, we need to discuss two functions - the Gaussian and the Bessel Potential. The former is involved since we choose it as activation function for the RBF network, the latter because it leads to a norm which is equivalent to the Sobolev norm (that is, both Sobolev and Bessel Potential norms are bounded by a multiple of the other as non-negative functionals).

Let $\gamma_{d,b} : \mathbb{R}^d \to \mathbb{R}$ denote the d-dimensional Gaussian function of *width* $b > 0$ and *center* $0 = (0, \ldots, 0)$ in $\mathbb{R}^d$:

$$\gamma_{d,b}(x) := e^{-b\|x\|^2} .$$

We write γ_d for $\gamma_{d,1}$. Width is parameterized inversely: larger values of b correspond to sharper peaks. For $b > 0$, let

$$G_d^\gamma(b) := \left\{\tau_y(\gamma_{d,b}) \,|\, y \in \mathbb{R}^d\right\} ,$$

denote the set of d-variable Gaussian RBFs with width $b > 0$ and all possible centers, where, for each vector y in $\mathbb{R}^d$, τ_y is the translation operator defined for any real-valued function $g : \mathbb{R}^d \to \mathbb{R}$ by

$$\tau_y(g)(x) := g(x - y) \,.\forall x \in \mathbb{R}^d$$

The set of Gaussians with varying widths is denoted

$$G_d^\gamma := \bigcup_{b>0} G_d^\gamma(b) .$$

For $m > 0$, the *Bessel potential of order m* on $\mathbb{R}^d$ is the function $\beta_{d,m}$ with the Fourier transform

$$\hat{\beta}_{d,m}(\omega) = (1 + \|\omega\|^2)^{-m/2} ,$$

where we consider the *Fourier transform* $\mathcal{F}(f) := \hat{f}$ as

$$\hat{f}(\omega) := (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(x) e^{ix\cdot\omega} dx.$$

For $m > 0$ and $q \in [1, \infty)$, let

$$L^{q,m}(\mathbb{R}^d) := \{f \mid f = w * \beta_d, m, \; w \in \mathcal{L}^q(\mathbb{R}^d)\}$$

be the *Bessel potential space* which is formed by convolutions of functions from $\mathcal{L}^q(\mathbb{R}^d)$ with $\beta_{d,m}$. The Bessel norm is defined as

$$\|f\|_{L^{q,m}(\mathbb{R}^d)} := \|w_f\|_{\mathcal{L}^q(\mathbb{R}^d)} \quad \text{for} \quad f = w_f * \beta_{d,m}.$$

In row 1 of Table 2, $\xi(d) = (\pi/2b)^{d/4} r_d$. Thus for $b = \pi/2$, the estimate implies tractability for r_d growing with d polynomially, while for $b > \pi/2$, it implies tractability even when r_d increases exponentially fast. Hence, the width b of Gaussians has a strong impact on the size of radii r_d of balls in $G_d^\gamma(b)$-variation for which $\xi(d)$ is a polynomial. The narrower the Gaussians, the larger the balls for which the estimate in row 1 of Table 2 implies tractability.

For every $m > d/2$, the upper bound from row 2 of Table 2 on the worst-case error in approximation by Gaussian-basis-function networks is of the factorized form $\xi(d)\kappa(n)$, where $\kappa(n) = n^{-1/2}$ and

Table 2 Factorized approximation rates for Gaussian RBF networks. In row 1, $b > 0$; in row 2, $m > d/2$.

ambient space	dictionary	approximated functions	$\xi(d)$	$\kappa(n)$
$(\mathcal{L}^2(\mathbb{R}^d), \|.\|_{\mathcal{L}^2(\mathbb{R}^d)})$	$G_d^\gamma(b)$	$B_{r_d}(\|.\|_{G_d^\gamma(b)})$	$r_d\,\left(\frac{\pi}{2b}\right)^{d/4}$	$n^{-1/2}$
$(\mathcal{L}^2(\mathbb{R}^d), \|.\|_{\mathcal{L}^2(\mathbb{R}^d)})$	G_d^γ	$B_{r_d}(\|.\|_{L^{1,m}}) \cap L^{2,m}$	$\left(\frac{\pi}{2}\right)^{d/4} \frac{\Gamma(m/2-d/4)}{\Gamma(m/2)}\, r_d$	$n^{-1/2}$

$$\xi(d) = r_d \left(\frac{\pi}{2}\right)^{d/4} \frac{\Gamma(m/2 - d/4)}{\Gamma(m/2)}.$$

Let $h > 0$ and put $m_d = d/2 + h$. Then $\xi(d)/r_d = \left(\frac{\pi}{2}\right)^{d/4} \frac{\Gamma(h/2)}{\Gamma(h/2+d/4)}$, which goes to zero exponentially fast with increasing d. So for $h > 0$ and $m_d \geq d/2 + h$, the approximation of functions in $B_{r_d}(\|.\|_{L^{1,m_d}(\mathbb{R}^d)}) \cap L^{2,m_d}(\mathbb{R}^d)$ by $\mathrm{span}_n\, G_d^\gamma$ is hyper-tractable in $\mathcal{L}^2(\mathbb{R}^d)$.

8.4 Perceptron Network Tractability

Let us now consider tractability of approximation by perceptron networks. The results are summarized in Table 3. This subsection also requires some technical machinery. We describe a class of real-valued functions on $\mathbb{R}^d$, the functions of weakly-controlled decay, defined by Kainen, Kůrková, and Vogt in [35], which have exactly the weakest possible constraints on their behavior at infinity to guarantee finiteness of a certain semi-norm and we show that functions in this class have a nice integral formula, leading to our results. This subsection provides an instance in which $\mathbb{N}'$, the domain of the dimensional complexity function ξ, is the odd positive integers.

The dictionary of functions on $X \subseteq \mathbb{R}^d$ computable by perceptron networks with the activation function ψ is denoted by

$$P_d^\psi(X) := \{f : X \to \mathbb{R} \,|\, f(x) = \psi(v \cdot x + b),\ v \in \mathbb{R}^d,\ b \in \mathbb{R}\},$$

so $P_d^\psi(\mathbb{R}^d) = P_d^\psi$ as defined in (1). For ϑ the Heaviside function, as in (2),

$$P_d^\vartheta(X) = H_d(X) := \{f : X \to \mathbb{R} \,|\, f(x) = \vartheta(e \cdot x + b),\ e \in S^{d-1}, b \in \mathbb{R}\},$$

where S^{d-1} is the sphere constituted by the unit-euclidean-norm vectors in $\mathbb{R}^d$ and $H_d(X)$ is the set of characteristic functions of closed half-spaces of $\mathbb{R}^d$ restricted to X. Of course, all these functions, and their finite linear combinations, have finite sup-norms. The integral formula we develop below shows that the nice functions (of weakly controlled decay) are tractably approximated by the linear span of $H_d(X)$.

For $\mathcal{F}$ any family of functions on $\mathbb{R}^d$ and $\Omega \subseteq \mathbb{R}^d$, let

$$\mathcal{F}|_\Omega := \{f|_\Omega \,|\, f \in \mathcal{F}\},$$

where $f|_\Omega$ is the restriction of f to Ω. We also use the phrase "variation with respect to half-spaces" for the restrictions of H_d. For simplicity, we may write H_d instead of $H_d|_\Omega$. When $\Omega_d \subset \mathbb{R}^d$ has finite Lebesgue measure, for each continuous nondecreasing sigmoid σ, variation with respect to half-spaces is equal to $P_d^\sigma|_{\Omega_d}$-variation in $\mathcal{L}^2(\Omega_d)$ [43]. Hence, investigating tractability of balls in variation with respect to half-spaces has implications for approximation by perceptron networks with arbitrary continuous nondecreasing sigmoids.

A real-valued function f on $\mathbb{R}^d$, d odd, is of *weakly-controlled decay* [35] if f is d-times continuously differentiable and for all multi-indices $\alpha \in \mathbb{N}^d$ with $|\alpha| = \sum_{i=1}^d \alpha_i$ and $D^\alpha = \partial^{\alpha_1} \cdot \ldots \cdot \partial^{\alpha_d}$, such that

$$\lim_{\|x\|\to\infty} D^\alpha f(x) = 0, \; \forall \alpha, \; |\alpha| < d \quad (i)$$

$$\exists \varepsilon > 0 \setminus \lim_{\|x\|\to\infty} D^\alpha f(x) \|x\|^{d+1+\varepsilon} = 0, \; \forall \alpha, \; |\alpha| = d. \; (ii)$$

Let $\mathcal{V}(\mathbb{R}^d)$ denote the set of functions of weakly controlled decay on $\mathbb{R}^d$. This set includes the Schwartz class of smooth functions rapidly decreasing at infinity as well as the class of d-times continuously differentiable functions with compact supports. In particular, it includes the Gaussian function. Also, if $f \in \mathcal{V}(\mathbb{R}^d)$, then $\|D^\alpha f\|_{\mathcal{L}^1(\mathbb{R}^d)} < \infty$ if $|\alpha| = d$. The maximum over all α with $|\alpha| = d$ is called the *Sobolev seminorm* of f and is denoted $\|f\|_{d,1,\infty}$. We denote by A_{r_d} the intersection of $\mathcal{V}(\mathbb{R}^d)$ with the ball $B_{r_d}(\|\cdot\|_{d,1,\infty})$ of radius r_d in the Sobolev seminorm $\|.\|_{d,1,\infty}$. Then

$$A_{r_d} := \mathcal{V}(\mathbb{R}^d) \cap B_{r_d}(\|\cdot\|_{d,1,\infty}) = r_d \, A_1.$$

The estimates in rows 1 and 2 of Table 3 imply that approximation of functions from balls of radii r_d in variation with respect to half-spaces is tractable in the space $\mathcal{M}(\mathbb{R}^d)$ of bounded measurable functions on $\mathbb{R}^d$ with respect to supremum norm, when the radius r_d grows polynomially. In $(\mathcal{L}^2(\Omega_d), \|.\|_{\mathcal{L}^2(\Omega_d)})$, this approximation is tractable when r_d times $\lambda(\Omega_d)$ grows polynomially with d. If for all $d \in \mathbb{N}'$, Ω_d is the unit ball in $\mathbb{R}^d$, then this approximation is hyper-tractable unless r_d is exponentially growing.

In row 5, we denote by $G_d^{\gamma,1} := \left\{\tau_y(\gamma_d) \,|\, y \in \mathbb{R}^d\right\}$ the set of d-variable Gaussians with widths equal to 1 and varying centers. Using a result of Kainen, Kůrková, and Vogt [35], we have $\xi(d) = (2\pi d^{3/4}) \, \lambda(\Omega_d)^{1/2}$. This implies that approximation of d-variable Gaussians on a domain Ω_d by perceptron

Table 3 Factorized approximation rates for perceptron networks. In rows 1 and 2, $B_{r_d}(\|.\|_{H_d,\mathcal{M}(\mathbb{R}^d)})$ and $B_{r_d}(\|.\|_{H_d|_{\Omega_d},\mathcal{L}^2(\Omega_d)})$ denote the balls of radius r_d in H_d-variations with respect to the ambient $\|.\|_{\mathcal{M}(\mathbb{R}^d)}$- and $\|.\|_{\mathcal{L}^2(\Omega_d)}$-norms, respectively. In rows 3 and 4, $k_d = 2^{1-d}\pi^{1-d/2}d^{d/2}/\Gamma(d/2) \sim (\pi d)^{1/2}(e/2\pi)^{d/2}$. We assume that $\lambda(\Omega_d) < \infty$ and $\Omega_d \neq \emptyset$, where λ denotes the Lebesgue measure.

ambient space	dictionary G_d	target set $\mathcal{F}$ to be approx.	$\xi(d)$	$\kappa(n)$
$(\mathcal{M}(\mathbb{R}^d), \|.\|_{\mathcal{M}(\mathbb{R}^d)})$	$H_d(\mathbb{R}^d)$	$B_{r_d}(\|.\|_{H_d(\mathbb{R}^d),\mathcal{M}(\mathbb{R}^d)})$	$6\sqrt{3}\, r_d\, d^{1/2}$	$(\log n)^{1/2}n^{-1/2}$
$(\mathcal{L}^2(\Omega_d), \|.\|_{\mathcal{L}^2(\Omega_d)})$	$H_d\|_{\Omega_d}$	$B_{r_d}(\|.\|_{H_d\|_{\Omega_d},\mathcal{L}^2(\Omega_d)})$	$\lambda(\Omega_d)\, r_d$	$n^{-1/2}$
$(\mathcal{L}^2(\Omega_d), \|.\|_{\mathcal{L}^2(\Omega_d)})$ $\Omega_d \subset \mathbb{R}^d$, d odd	$H_d\|_{\Omega_d}$	A_{r_d}	$r_d\, k_d \lambda(\Omega_d)^{1/2}$	$n^{-1/2}$
$(\mathcal{L}^2(\Omega_d), \|.\|_{\mathcal{L}^2(\Omega_d)})$ $\Omega_d \subset \mathbb{R}^d$, d odd	$P_d^{\sigma}(\Omega_d)$ σ continuous nondecr. sigmoid	A_{r_d}	$r_d\, k_d \lambda(\Omega_d)^{1/2}$	$n^{-1/2}$
$(\mathcal{L}^2(\Omega_d), \|.\|_{\mathcal{L}^2(\Omega_d)})$ $\Omega_d \subset \mathbb{R}^d$, d odd	$H_d\|_{\Omega_d}$	$G_d^{\gamma,1}\|_{\Omega_d}$	$(2\,\pi\, d)^{3/4}\, \lambda(\Omega_d)^{1/2}$	$n^{-1/2}$

networks is tractable when the Lebesgue measure $\lambda(\Omega_d)$ grows polynomially with d, while if the domains Ω_d are unit balls in $\mathbb{R}^d$, then the approximation is hyper-tractable.

9 Discussion

A number of years ago, Lotfi Zadeh remarked to two of the current authors that the developed countries must control their "under-coordinated technology" and this 21st century need is a driving force behind neural nets and other new computational approaches. It is hoped that modern methods will permit greater control and coordination of technology, and the techniques described in this article represent a key part of current understanding.

As we have shown, neural network theory unifies and embodies a substantial portion of the real and functional analysis which has been developed during the 19th and 20th centuries. This analysis is based on deep and hard-won knowledge and so presents the possibility of intellectual leverage in tackling the problems which arise in the large-scale practical application of computation. Thus, the abundance of powerful mathematical tools which are utilized gives modern approaches a possibility of overcoming previous obstacles.

Another good reason to consider the methodologies discussed in this article is their promise to enable much larger dimensionality to be considered in applications through a more sophisticated model capable of being put into special-purpose hardware. One of the obstructions to rapid and accurate neural computations may be that most work has dealt with static, rather than dynamic, problem instances. It is clear that human pattern recognition is strongly facilitated by dynamics. Through functional analysis, which has begun to be more strongly utilized in neural network theory, one can properly analyze high-dimensional phenomena in time as well as space. There will always be limitations due to hardware itself, but neural network approaches via integral formulae of the sort encountered in mathematical physics hold out the possibility of direct instantiation of neural networks by analog computations, optical or implemented in silicon.

Finally, the notion of hyper-tractable computations appears to show that for some problems, increase of dimension can improve computational performance. As lower bounds to neural network accuracy are not currently known, it may be that the heuristic successes discovered in a few well-chosen examples can be extended to a much broader domain of problem types.

10 Summary of Main Notations

$\mathbb{R}$	Set of real numbers.
d	Input dimension.
$\mathcal{S}^{d-1}$	Sphere of radius 1 in $\mathbb{R}^d$.
$\mathbb{N}_+$	Set of positive integers.
$(\mathcal{X}, \\|\cdot\\|_{\mathcal{X}})$	Normed linear space.
$B_r(\\|\cdot\\|_{\mathcal{X}})$	Ball of radius r in the norm $\\|\cdot\\|_{\mathcal{X}}$.
G	Subset of $(\mathcal{X}, \\|\cdot\\|_{\mathcal{X}})$, representing a generic dictionary.
$G(c)$	$\{wg \,\vert\, g \in G,\, w \in \mathbb{R},\, \vert w\vert \leq c\}$.
$\mathrm{card}(G)$	Cardinality of the set G.
$e_n(G)$	n-th entropy number of the set G.
$\mathrm{diam}(G)$	Diameter of the set G.
$\mathrm{span}\, G$	Linear span of G.

$\operatorname{span}_n G$	Set of all linear combinations of at most n elements of G.
$\operatorname{conv} G$	Convex hull of G.
$\operatorname{conv}_n G$	Set of all convex combinations of at most n elements of G.
$\operatorname{cl} G$	Closure of G in the norm $\|\cdot\|_{\mathcal{X}}$.
s_G	$\sup_{g \in G} \|g\|_{\mathcal{X}}$.
G_ϕ	Dictionary of functions computable by a unit of type ϕ.
$\psi(v \cdot x + b)$	Perceptron with an activation ψ, bias b, and weight vector v.
σ	Sigmoidal function.
ϑ	Heaviside function.
H_d	Set of characteristic functions of closed half-spaces of $\mathbb{R}^d$ (Heaviside perceptrons).
$P_d^\psi(X)$	Dictionary of functions on $X \subseteq \mathbb{R}^d$ computable by perceptron networks with activation ψ.
$\psi(b\|x - v\|)$	Radial-basis function with activation ψ, width b, and center v.
$F_d^\psi(X)$	Dictionary of functions on $X \subseteq \mathbb{R}^d$ computable by RBF networks with activation ψ.
$\gamma_{d,b}$	d-dimensional Gaussian of width b and center 0.
$G_d^\gamma(b)$	Set of d-variable Gaussian RBFs with width b and all possible centers.
G_d^γ	Set of Gaussians with varying widths.
$\|f - A\|_{\mathcal{X}}$	Distance from f to the set A in the norm $\|\cdot\|_{\mathcal{X}}$.
$\mathit{\Pi}_f$	Peak functional for f.
$\delta(M, A)$	Deviation of M from A in the norm $\|\cdot\|_{\mathcal{X}}$.
$\mathcal{N}(G, \varepsilon)$	ε-covering number of G in the norm $\|\cdot\|_{\mathcal{X}}$.

$\alpha_A(r)$	Decay function of the set A.
$\mathrm{Pr}_M(f)$	Projection of f to the set M.
$f\|_\Omega$	Restriction of f to the set Ω.
$\hat{f}$	Fourier transform of f.
$\beta_{d,m}$	Bessel potential of order m on $\mathbb{R}^d$.
$L^{q,m}(\mathbb{R}^d)$	Bessel potential space of order m in $\mathcal{L}^q(\mathbb{R}^d)$.
$\\|\cdot\\|_{1,A}$	ℓ_1-norm with respect to A.
$\\|\cdot\\|_G$	G-variation with respect to $\\|\cdot\\|_{\mathcal{X}}$.
$\\|\cdot\\|_{H_d}$	Variation with respect to half-spaces (Heaviside perceptrons).
$(\mathcal{C}(X), \\|\cdot\\|_{\sup})$	Space of all continuous functions on a subset $X \subseteq \mathbb{R}^d$, with the supremum norm.
λ	Lebesgue measure.
$(\mathcal{L}^p(X), \\|\cdot\\|_p)$, $p \in [1, \infty]$	Space of all Lebesgue-measurable and p-integrable functions f on X, with the $\mathcal{L}^p(X)$-norm.
$\mathcal{V}(\mathbb{R}^d)$	Set of functions of weakly controlled decay on $\mathbb{R}^d$.
$\\|\cdot\\|_{d,1,\infty}$	Sobolev seminorm.

Acknowledgement. V. Kůrková was partially supported by GA ČR grant P202/11/1368 and institutional support 67985807. M. Sanguineti was partially supported by a PRIN grant from the Italian Ministry for University and Research, project "Adaptive State Estimation and Optimal Control". Collaboration of V. Kůrková and P. C. Kainen was partially supported by MŠMT project KONTAKT ALNN ME10023 and collaboration of V. Kůrková and M. Sanguineti was partially supported by CNR - AVČR 2010-2012 project "Complexity of Neural-Network and Kernel Computational Models".

References

1. Barron, A.R.: Neural net approximation. In: Narendra, K. (ed.) Proc. 7th Yale Workshop on Adaptive and Learning Systems, pp. 69–72. Yale University Press (1992)

2. Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Trans. on Information Theory 39, 930–945 (1993)
3. Bellman, R.: Dynamic Programming. Princeton University Press, Princeton (1957)
4. Bochner, S., Chandrasekharan, K.: Fourier Transform. Princeton University Press, Princeton (1949)
5. Braess, D.: Nonlinear Approximation Theory. Springer (1986)
6. Breiman, L.: Hinging hyperplanes for regression, classification and function approximation. IEEE Trans. Inform. Theory 39(3), 999–1013 (1993)
7. Carrol, S.M., Dickinson, B.W.: Construction of neural nets using the Radon transform. In: Proc. the Int. Joint Conf. on Neural Networks, vol. 1, pp. 607–611 (1989)
8. Cheang, G.H.L., Barron, A.R.: A better approximation for balls. J. of Approximation Theory 104, 183–203 (2000)
9. Chui, C.K., Li, X., Mhaskar, H.N.: Neural networks for localized approximation. Math. of Computation 63, 607–623 (1994)
10. Courant, R., Hilbert, D.: Methods of Mathematical Physics, vol. II. Interscience, New York (1962)
11. Cybenko, G.: Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals, and Systems 2, 303–314 (1989)
12. Darken, C., Donahue, M., Gurvits, L., Sontag, E.: Rate of approximation results motivated by robust neural network learning. In: Proc. Sixth Annual ACM Conf. on Computational Learning Theory, pp. 303–309. ACM, New York (1993)
13. DeVore, R.A., Howard, R., Micchelli, C.: Optimal nonlinear approximation. Manuscripta Mathematica 63, 469–478 (1989)
14. Donahue, M., Gurvits, L., Darken, C., Sontag, E.: Rates of convex approximation in non-Hilbert spaces. Constructive Approximation 13, 187–220 (1997)
15. Girosi, F., Anzellotti, G.: Rates of convergence for Radial Basis Functions and neural networks. In: Mammone, R.J. (ed.) Artificial Neural Networks for Speech and Vision, pp. 97–113. Chapman & Hall (1993)
16. Girosi, F., Poggio, T.: Networks and the best approximation property. Biological Cybernetics 63, 169–176 (1990)
17. Giulini, S., Sanguineti, M.: Approximation schemes for functional optimization problems. J. of Optimization Theory and Applications 140, 33–54 (2009)
18. Gnecco, G., Sanguineti, M.: Estimates of variation with respect to a set and applications to optimization problems. J. of Optimization Theory and Applications 145, 53–75 (2010)
19. Gnecco, G., Sanguineti, M.: On a variational norm tailored to variable-basis approximation schemes. IEEE Trans. on Information Theory 57, 549–558 (2011)
20. Gribonval, R., Vandergheynst, P.: On the exponential convergence of matching pursuits in quasi-incoherent dictionaries. IEEE Trans. on Information Theory 52, 255–261 (2006)
21. Gurvits, L., Koiran, P.: Approximation and learning of convex superpositions. J. of Computer and System Sciences 55, 161–170 (1997)
22. Hartman, E.J., Keeler, J.D., Kowalski, J.M.: Layered neural networks with Gaussian hidden units as universal approximations. Neural Computation 2, 210–215 (1990)
23. Hewit, E., Stromberg, K.: Abstract Analysis. Springer, Berlin (1965)
24. Hornik, K., Stinchcombe, M., White, H.: Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. Neural Networks 3, 551–560 (1990)

25. Ito, Y.: Representation of functions by superpositions of a step or sigmoidal function and their applications to neural network theory. Neural Networks 4, 385–394 (1991)
26. Ito, Y.: Finite mapping by neural networks and truth functions. Mathematical Scientist 17, 69–77 (1992)
27. 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, 608–613 (1992)
28. Juditsky, A., Hjalmarsson, H., Benveniste, A., Delyon, B., Ljung, L., Sjöberg, J., Zhang, Q.: Nonlinear black-box models in system identification: Mathematical foundations. Automatica 31, 1725–1750 (1995)
29. Kainen, P.C., Kůrková, V.: Quasiorthogonal dimension of Euclidean spaces. Applied Mathematics Letters 6, 7–10 (1993)
30. Kainen, P.C., Kůrková, V., Sanguineti, M.: Complexity of Gaussian radial basis networks approximating smooth functions. J. of Complexity 25, 63–74 (2009)
31. Kainen, P.C., Kurkova, V., Sanguineti, M.: Dependence of Computational Models on Input Dimension: Tractability of Approximation and Optimization Tasks. IEEE Transactions on Information Theory 58, 1203–1214 (2012)
32. Kainen, P.C., Kůrková, V., Vogt, A.: Geometry and topology of continuous best and near best approximations. J. of Approximation Theory 105, 252–262 (2000)
33. Kainen, P.C., Kůrková, V., Vogt, A.: Continuity of approximation by neural networks in L_p-spaces. Annals of Operational Research 101, 143–147 (2001)
34. Kainen, P.C., Kůrková, V., Vogt, A.: Best approximation by linear combinations of characteristic functions of half-spaces. J. of Approximation Theory 122, 151–159 (2003)
35. Kainen, P.C., Kůrková, V., Vogt, A.: A Sobolev-type upper bound for rates of approximation by linear combinations of Heaviside plane waves. J. of Approximation Theory 147, 1–10 (2007)
36. Knuth, D.E.: Big omicron and big omega and big theta. SIGACT News 8, 18–24 (1976)
37. Kůrková, V.: Kolmogorov's theorem and multilayer neural networks. Neural Networks 5, 501–506 (1992)
38. Kůrková, V.: Approximation of functions by perceptron networks with bounded number of hidden units. Neural Networks 8, 745–750 (1995)
39. Kůrková, V.: Dimension-independent rates of approximation by neural networks. In: Warwick, K., Kárný, M. (eds.) Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality, Birkhäuser, Boston, MA, pp. 261–270 (1997)
40. Kůrková, V.: Incremental approximation by neural networks. In: Warwick, K., Kárný, M., Kůrková, V. (eds.) Complexity: Neural Network Approach, pp. 177–188. Springer, London (1998)
41. Kůrková, V.: High-dimensional approximation and optimization by neural networks. In: Suykens, J., et al. (eds.) Advances in Learning Theory: Methods, Models and Applications, ch. 4, pp. 69–88. IOS Press, Amsterdam (2003)
42. Kůrková, V.: Minimization of error functionals over perceptron networks. Neural Computation 20, 252–270 (2008)
43. Kůrková, V., Kainen, P.C., Kreinovich, V.: Estimates of the number of hidden units and variation with respect to half-spaces. Neural Networks 10, 1061–1068 (1997)

44. Kůrková, V., Neruda, R.: Uniqueness of functional representations by Gaussian basis function networks. In: Proceedings of ICANN 1994, pp. 471–474. Springer, London (1994)
45. Kůrková, V., Sanguineti, M.: Bounds on rates of variable-basis and neural-network approximation. IEEE Trans. on Information Theory 47, 2659–2665 (2001)
46. Kůrková, V., Sanguineti, M.: Comparison of worst case errors in linear and neural network approximation. IEEE Trans. on Information Theory 48, 264–275 (2002)
47. Kůrková, V., Sanguineti, M.: Error estimates for approximate optimization by the extended Ritz method. SIAM J. on Optimization 15, 261–287 (2005)
48. Kůrková, V., Sanguineti, M.: Learning with generalization capability by kernel methods of bounded complexity. J. of Complexity 21, 350–367 (2005)
49. Kůrková, V., Sanguineti, M.: Estimates of covering numbers of convex sets with slowly decaying orthogonal subsets. Discrete Applied Mathematics 155, 1930–1942 (2007)
50. Kůrková, V., Sanguineti, M.: Approximate minimization of the regularized expected error over kernel models. Mathematics of Operations Research 33, 747–756 (2008)
51. Kůrková, V., Sanguineti, M.: Geometric upper bounds on rates of variable-basis approximation. IEEE Trans. on Information Theory 54, 5681–5688 (2008)
52. Kůrková, V., Savický, P., Hlaváčková, K.: Representations and rates of approximation of real–valued Boolean functions by neural networks. Neural Networks 11, 651–659 (1998)
53. Lavretsky, E.: On the geometric convergence of neural approximations. IEEE Trans. on Neural Networks 13, 274–282 (2002)
54. Leshno, M., Lin, V.Y., Pinkus, A., Schocken, S.: Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Networks 6, 861–867 (1993)
55. Makovoz, Y.: Random approximants and neural networks. J. of Approximation Theory 85, 98–109 (1996)
56. Mhaskar, H.N.: Versatile Gaussian networks. In: Proc. of IEEE Workshop of Nonlinear Image Processing, pp. 70–73 (1995)
57. Mhaskar, H.N.: On the tractability of multivariate integration and approximation by neural networks. J. of Complexity 20, 561–590 (2004)
58. Micchelli, C.A.: Interpolation of scattered data: Distance matrices and conditionally positive definite functions. Constructive Approximation 2, 11–22 (1986)
59. Narendra, K.S., Mukhopadhyay, S.: Adaptive control using neural networks and approximate models. IEEE Trans. on Neural Networks 8, 475–485 (1997)
60. Park, J., Sandberg, I.W.: Universal approximation using radial–basis–function networks. Neural Computation 3, 246–257 (1991)
61. Park, J., Sandberg, I.W.: Approximation and radial-basis-function networks. Neural Computation 5, 305–316 (1993)
62. Pinkus, A.: Approximation theory of the MLP model in neural networks. Acta Numerica 8, 143–195 (1999)
63. Pisier, G.: Remarques sur un résultat non publié de B. Maurey. In: Séminaire d'Analyse Fonctionnelle 1980-1981, Palaiseau, France. École Polytechnique, Centre de Mathématiques, vol. I(12) (1981)
64. Rosenblatt, F.: The perceptron: A probabilistic model for information storage and organization of the brain. Psychological Review 65, 386–408 (1958)

65. Rudin, W.: Principles of Mathematical Analysis. McGraw-Hill (1964)
66. Sejnowski, T.J., Rosenberg, C.: Parallel networks that learn to pronounce English text. Complex Systems 1, 145–168 (1987)
67. Singer, I.: Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces. Springer, Heidelberg (1970)
68. Smith, K.A.: Neural networks for combinatorial optimization: A review of more than a decade of research. INFORMS J. on Computing 11, 15–34 (1999)
69. Stinchcombe, M., White, H.: Approximation and learning unknown mappings using multilayer feedforward networks with bounded weights. In: Proc. Int. Joint Conf. on Neural Networks IJCNN 1990, pp. III7–III16 (1990)
70. Traub, J.F., Werschulz, A.G.: Complexity and Information. Cambridge University Press (1999)
71. Wasilkowski, G.W., Woźniakowski, H.: Complexity of weighted approximation over $\mathbb{R}^d$. J. of Complexity 17, 722–740 (2001)
72. Woźniakowski, H.: Tractability and strong tractability of linear multivariate problems. J. of Complexity 10, 96–128 (1994)
73. Zemanian, A.H.: Distribution Theory and Transform Analysis. Dover, New York (1987)
74. Zoppoli, R., Parisini, T., Sanguineti, M., Baglietto, M.: Neural Approximations for Optimal Control and Decision. Springer, London (in preparation)
75. Zoppoli, R., Sanguineti, M., Parisini, T.: Approximating networks and extended Ritz method for the solution of functional optimization problems. J. of Optimization Theory and Applications 112, 403–439 (2002)