An integral upper bound for neural-network approximation

Kainen, P.C.
2008

Institute of Computer Science
Academy of Sciences of the Czech Republic

An integral upper bound for neural-network approximation

Paul C. Kainen and Věra Kůrková

Technical report No. 1023

May 2008

Pod Vodárenskou věží 2, 182 07 Prague 8, phone: +420 266 051 111, fax: +420 286 585 789,
e-mail:ics@cs.cas.cz

Institute of Computer Science
Academy of Sciences of the Czech Republic

An integral upper bound for neural-network approximation

Paul C. Kainen[1] and Věra Kůrková [2]

Technical report No. 1023

May 2008

Abstract:

Complexity of one-hidden-layer networks is studied using tools from nonlinear approximation and integration theory. For functions with suitable integral representations in the form of networks with infinitely many hidden units, upper bounds are derived on the speed of decrease of approximation error with an increasing number of network units. A unifying framework for derivation of such bounds is obtained using properties of Bochner integral. The results are applied to perceptron networks.

Keywords:
Model complexity of neural networks, integral representation in the form of network with infinitely many hidden units, rates of variable-basis approximation, variational norm, Bochner integral, perceptron networks.

[1]Department of Mathematics, Georgetown University, Washington, D. C. 20057-1233, USA, kainen@georgetown.edu
[2]Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07, Prague 8, Czech Republic, vera@cs.cas.cz

1 Introduction

Some understanding of the dependence of model complexity of neural networks on type of computational units and properties of training data can be derived by inspection of estimates of rates of decrease of approximation errors with increasing number of network units. Assuming that training data are chosen from a given multivariable function, the form of an estimate of error in approximation of such function by a network with a given type of units tells us which combinations of properties of the function and of the computational units guarantee fast rates of approximation. With such units, good accuracy can be achieved by a reasonably small network.

A suitable tool for obtaining estimates of rates of neural network approximation is a result by Maurey (see [Pisier, 1981]), Jones (1992) and Barron (1993). It belongs to nonlinear approximation theory and applies to approximation by so called variable-basis functions or "dictionaries". For functions from the convex hull of a bounded subset G of a Hilbert space, Maurey-Jones-Barron's theorem gives an upper bound on the square of the error in approximation by convex combinations of n elements of G. It implies an upper bound on rates of approximation by linear combinations of n elements G, which has the form $n^{-1/2}$ times a certain norm (called G-variation) of the function to be approximated [Kůrková, 2003]. Investigation of properties of variational norms for G corresponding to various types of network units can provide some insight into the impact of a choice of the type of units on model complexity.

Several authors applied Maurey-Jones-Barron's theorem to functions, which can be represented as networks with infinitely many units. Barron (1993) considered functions representable as weighted Fourier transforms, Girosi and Anzelotti (1993) convolutions with suitable kernels. Explicitly in terms of an upper bound on variation, Kůrková et al.(1997) derived an estimate of rates of approximation for perceptron networks proving that smooth compactly supported functions can be expressed as networks with infinitely many Heaviside perceptrons.

In this paper, we develop a rather general framework for investigation of rates of approximation of functions representable as integrals of the form of networks with infinitely many units of various types. Our approach is based on the Bochner integral. A special case of this method was sketched by Girosi and Anzelloti (1993) for convolutions with certain kernels. The Bochner integral extends the concept of Lebesgue integral to mappings into function spaces (the value of a Bochner integral is a function, not a number). Bochner integral can be applied to mappings assigning to parameters (such as weights, biases or centroids) functions computable by units (such as perceptrons or radial-basis-functions) with such parameters.

Using properties of the Bochner integral and a theorem on the relationship of its evaluations to corresponding Lebesgue integrals, we derive an upper bound on variational norm and hence on rates of approximation. For functions representable as networks with infinitely many hidden units, we show that the size of the $\mathcal{L}^1$-norm of the output weight function is a factor in network complexity.

We illustrate our results on perceptron networks. Combining a representation of smooth functions as an integral combination of Heaviside perceptrons [Kůrková et al., 1997] with estimates on variational norm, we obtain an upper bound on rates of approximation by perceptron networks for a wide class of functions.

The paper is organized as follows. Section 2 introduces our approach and states the main results. Section 3 recalls Maurey-Jones-Barron's theorem and variational norms. Section 4 gives upper bounds on variational norms for functions representable as integrals of the form of networks with infinitely many hidden units. In Section 5, we apply these estimates to perceptron networks. Section 6 is a brief discussion. Properties of Bochner integral are summarized in the Appendix.

2 Outline of approach and main results

One-hidden-layer feedforward networks belong to a class of computational models, which can mathematically be described as *variable-basis* schemas. Such models compute functions from sets of the form

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

where G is a set of functions, which is sometimes called a *dictionary*. For example, G can be the set of functions computable by perceptrons, radial-basis functions, kernel functions, or trigonometric polynomials. The number n expresses the *model complexity* (in the case of one-hidden-layer neural networks, it is the number of units in the hidden layer).

Often, sets G are parameterized; that is they are of the form

$$G_\phi = \{\phi(.,y) \,|\, y \in Y\},$$

where $\phi : \Omega \times Y \to \mathbb{R}$, Y is the set of parameters and Ω is the set of variables. Such a parameterized set of functions can be represented by a mapping

$$\Phi : Y \to \mathcal{X},$$

where $\mathcal{X}$ is a suitable function space. Φ is defined for all $y \in Y$ as

$$\Phi(y)(x) = \phi(x,y).$$

For example, the set of functions computable by perceptrons with an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ can be described by a mapping Φ_σ on $\mathbb{R}^{d+1}$ defined for $(v,b) \in \mathbb{R}^d \times \mathbb{R} = \mathbb{R}^{d+1}$ as $\Phi_\sigma(v,b)(x) = \sigma(v \cdot x + b)$.

For parameterized sets we use the notation

$$\Phi(Y) = G_\phi = \{\phi(.,y) \,|\, y \in Y\} \qquad \text{and} \qquad s_\Phi = \sup_{y \in Y} \|\phi(.,y)\|_{\mathcal{X}}. \tag{2.1}$$

In this paper, we consider parameterized sets of functions belonging either to an $\mathcal{L}^q$-space with $q \in [1,\infty)$ or to a reproducing kernel Hilbert space. For $\Omega \subseteq \mathbb{R}^d$, ρ a measure on Ω and $q \in [1,\infty)$, we denote by $\mathcal{L}^q(\Omega,\rho)$ the space of all real-valued functions h satisfying $\int_\Omega |h(y)|^q d\rho < \infty$. When ρ is the Lebesgue measure, we sometimes write merely $\mathcal{L}^q(\Omega)$. A Hilbert space $\mathcal{X}$ of point-wise defined real-valued functions on an arbitrary set Ω is called a *reproducing kernel Hilbert space (RKHS)* when all evaluation functionals on $\mathcal{X}$ are bounded [Aronszajn, 1950].

The distance of an element f of a normed linear space $(\mathcal{X}, \|.\|_{\mathcal{X}})$ from its subset is denoted

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

We investigate speed of decrease of distances $\|f - \mathrm{span}_n \Phi(Y)\|$ with n increasing for functions f representable as one-hidden-layer networks with infinitely many hidden units from $\Phi(Y)$. More precisely, we consider functions, which can be expressed for a suitable measure μ on Y and almost all $x \in \Omega$ as the Lebesgue integrals of the form

$$f(x) = \int_Y w(y)\phi(x,y)d\mu(y). \tag{2.2}$$

Such functions are images of the corresponding weight functions w under the integral operator L_ϕ defined as $L_\phi(w)(x) = \int_Y w(y)\phi(x,y)d\mu(y)$. We show that the "size" of the output weight function w is critical for the speed of decrease of approximation errors. In Section 4, with rather mild assumptions on μ, w and ϕ, we prove that this speed depends on the $\mathcal{L}^1(Y,\mu)$-norm of the weight function w:

$$\|f - \mathrm{span}_n \Phi(Y)\|_{\mathcal{X}}^2 \leq \frac{(s_\Phi \|w\|_{\mathcal{L}^1(Y,\mu)})^2 - \|f\|_{\mathcal{X}}^2}{n}. \tag{2.3}$$

To derive this upper bound, we use a result by Maurey, Jones and Barron on variable-basis approximation reformulated in terms of a norm called $\Phi(Y)$-variation. To estimate this norm, we take advantage of properties of the Bochner integral, which is an extension of the concept of Lebesgue integral allowing integration of mappings with values in function spaces. We consider the Bochner integral of the mapping $w\Phi : Y \to \mathcal{X}$, which is defined for all $y \in Y$ via scalar multiplication in $\mathcal{X}$ as

$$w\Phi(y) = w(y)\Phi(y) = w(y)\phi(.,y).$$

Using the relationship between the Lebesgue integral (2.2) which represents values of the function f and the Bochner integral of the mapping $w\Phi$, we obtain an estimate of $\Phi(Y)$-variation of f in terms of the $\mathcal{L}^1$-norm of the weight function w. This gives an upper bound (2.3) on rate of approximation by $\mathrm{span}_n\Phi(Y)$. Combining this estimate with a representation of smooth functions as Heaviside perceptron networks with infinitely many units from [Kůrková et al., 1997], we estimate rate of approximation by perceptron networks.

3 Rates of variable-basis approximation and variational norm

An upper bound on approximation by

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

was derived by Maurey (see [Pisier, 1981], rediscovered by Jones (1992) and refined by Barron (1993).

Theorem 3.1 (Maurey-Jones-Barron) *Let G be a bounded nonempty subset of a Hilbert space $(X, \|.\|)$ and $s_G = \sup_{g\in G} \|g\|$. Then for every $f \in \mathrm{cl}\,\mathrm{conv}\, G$ and for every positive integer n,*

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

Theorem 3.1 can be reformulated in terms of a norm called *G-variation*. This variational norm is defined for any bounded nonempty subset G of any normed linear space $(X, \|.\|)$ as the Minkowski functional of the closed convex symmetric hull of G, i.e.,

$$\|f\|_G = \inf \left\{ c > 0 \mid c^{-1} f \in \mathrm{cl}\,\mathrm{conv}\,(G \cup -G) \right\}, \tag{3.1}$$

where the closure cl is taken with respect to the topology generated by the norm $\|.\|$ and conv denotes the convex hull. Note that G-variation can be infinite (when the set on the right-hand side is empty). It is a norm on the subspace of $\mathcal{X}$ formed by those $f \in \mathcal{X}$, for which $\|f\|_G < \infty$. G-variation depends on the norm on the ambient space, but as this is implicit, we omit it in the notation. Variational norms were introduced by Barron (1992) for characteristic functions of certain families of subsets of $\mathbb{R}^d$, in particular, for the set of characteristic functions of closed half-spaces corresponding to the set of functions computable by Heaviside perceptrons. For functions of one variable (i.e., $d = 1$), variation with respect to half-spaces coincides, up to a constant, with the notion of total variation. The general concept was defined by Kůrková (1997). The following upper bound is a corollary of Theorem 3.1 from [Kůrková, 1997] (see also [Kůrková, 2003]).

Theorem 3.2 *Let $(\mathcal{X}, \|.\|)$ be a Hilbert space, G its bounded nonempty subset, $s_G = \sup_{g\in G} \|g\|$. Then for every $f \in \mathcal{X}$ and every positive integer n,*

$$\|f - \mathrm{span}_n G\|^2 \leq \frac{s_G^2 \|f\|_G^2 - \|f\|^2}{n}.$$

This reformulation of Theorem 3.1 in terms of variational norm allows one to formulate an upper bound on variable-basis approximation for all functions in a Hilbert space. A similar result to Theorem 3.2 can be obtained in the $\mathcal{L}^q$-spaces with $q \in (1, \infty)$ using a result by Darken et al. (1993); for a slightly simplified argument see also [Kůrková and Sanguineti, 2005]. For the definition of Radon measure see Section 4.

Theorem 3.3 *Let G be a bounded subset of $\mathcal{L}^q(\Omega,\rho)$, $q\in(1,\infty)$, and ρ a Radon measure. Then for every $f\in \operatorname{cl}\operatorname{conv} G$ and every positive integer n,*

$$\|f-\operatorname{conv}_n G\|_{\mathcal{L}^q(\Omega,\rho)}\le \frac{2^{1+1/r}\,s_G\|f\|_G}{n^{1/s}},$$

where $1/q+1/p=1$, $r=\min(p,q)$, $s=\max(p,q)$.

In some cases, variational norms with respect to two different sets are the same. For example, in $\mathcal{L}^q$-spaces with $q\in(1,\infty)$, variation with respect to Heaviside perceptrons equals variation with respect to perceptrons with any sigmoidal activation function [Kůrková et al., 1997]. So to obtain from Theorem 3.2 rates of approximation by perceptron networks, it suffices to study variation with respect to half-spaces for which estimates in terms of Sobolev seminorms are known [Kůrková et al., 1997, Kainen et al., 2007b].

Thus investigation of variational norms can provide some insight into properties of multivariable functions, which can be efficiently approximated by various computational models. The following proposition summarizes basic properties of variation: (i) and (ii) follow directly from the definition, for (iii) see [Kůrková and Sanguineti, 2002, Proposition 3(iii)].

Proposition 3.4 *Let $(\mathcal{X},\|.\|)$ be a normed linear space, G,H its nonempty bounded subsets and $s_{G,H}:=\sup_{g\in G}\|g\|_H$. Then their variational norms satisfy the following:*
(i) for $f\in\mathcal{X}$ representable as $f=\sum_{i=1}^k w_i g_i$ with all $g_i\in G$ and $w_i\in\mathbb{R}$, $\|f\|_G\le\sum_{i=1}^k|w_i|$;
(ii) for any linear isometry ψ of $\mathcal{X}$ and for every $f\in\mathcal{X}$, $\|f\|_G=\|\psi(f)\|_{\psi(G)}$;
(iii) for every $f\in\mathcal{X}$, $\|f\|_H\le s_{G,H}\|f\|_G$.

The next lemma shows that variation of the limit of a sequence of functions is bounded from above by the limit of their variations.

Lemma 3.5 *Let G be a nonempty, nonzero bounded subset of a normed linear space $(\mathcal{X},\|\cdot\|_{\mathcal{X}})$, $h\in\mathcal{X}$, $\{h_i\}_{i=1}^\infty\subset\mathcal{X}$ with $b_i=\|h_i\|_G<\infty$ for all i. If $\lim_{i\to\infty}\|h_i-h\|_{\mathcal{X}}=0$ and there exists a finite $b=\lim_{i\to\infty}b_i$, then $\|h\|_G\le b$.*

Proof. For all $\varepsilon>0$ choose some $\eta>0$ such that $\eta<\frac{\varepsilon b^2}{2(b+\|h\|_{\mathcal{X}})}$. By the convergence assumptions, there exists i_0 such that for all $i>i_0$, $\|h-h_i\|_{\mathcal{X}}<\eta$ and $|b-b_i|<\eta$. Then by the triangle inequality for all $i>i_0$, $\left\|\frac{h}{b+\eta}-\frac{h_i}{b_i+\eta}\right\|_{\mathcal{X}}\le\left\|\frac{h}{b+\eta}-\frac{h}{b_i+\eta}\right\|_{\mathcal{X}}+\left\|\frac{h}{b_i+\eta}-\frac{h_i}{b_i+\eta}\right\|_{\mathcal{X}}\le\frac{\eta\|h\|_{\mathcal{X}}}{(b+\eta)(b_i+\eta)}+\frac{\eta}{b_i+\eta}\le\frac{\eta\|h\|_{\mathcal{X}}}{b^2}+\frac{\eta}{b}<\frac{\varepsilon}{2}$.

By the definition of variation, $\|h_i\|_G=b_i$ implies that there exists $\delta_i<\eta$ such that $\frac{h_i}{b_i+\delta_i}\in\operatorname{cl}\operatorname{conv}(G\cup-G)$. As $\operatorname{conv}(G\cup-G)$ is symmetric and convex, also $\frac{h_i}{b_i+\eta}\in\operatorname{cl}\operatorname{conv}(G\cup-G)$.

Then $\left\|\frac{h}{b}-\operatorname{cl}\operatorname{conv}(G\cup-G)\right\|_{\mathcal{X}}\le\left\|\frac{h}{b}-\frac{h_i}{b_i+\eta}\right\|_{\mathcal{X}}\le\left\|\frac{h}{b}-\frac{h}{b+\eta}\right\|_{\mathcal{X}}+\left\|\frac{h}{b+\eta}-\frac{h_i}{b_i+\eta}\right\|_{\mathcal{X}}$
$\le\frac{\eta\|h\|_{\mathcal{X}}}{b^2}+\frac{\varepsilon}{2}<\varepsilon$. Infimizing over ε, we get $\frac{h}{b}\in\operatorname{cl}\operatorname{conv}(G\cup-G)$ and thus $\|h\|_G\le b$. □

4 Upper bound on variation with respect to a parameterized family

Analogy with Proposition 3.4 (i) suggests that for f representable as

$$f(x)=\int_Y w(y)\phi(x,y)d\mu(y) \tag{4.1}$$

one should expect

$$\|f\|_{\Phi(Y)}\le\int_Y|w(y)|d\mu. \tag{4.2}$$

Various special cases of integral representations of the form (4.1) have been investigated. E.g., Barron (1993) proved that a function f representable as a weighted Fourier transform belongs to

a convex hull of trigonometric perceptrons and thus Theorem 3.1 can be used to estimate rates of approximation of f by networks with trigonometrics perceptrons. Girosi and Anzelotti (1993) proved a similar estimate for convolutions with suitable kernels. Explicitly as an upper bound on variation, the estimate in terms of the $\mathcal{L}^1(Y)$-norm of the weight function w was derived by Kůrkovaá et al. (1997) for integral representations $\int_Y w(y)\phi(x,y)dy$ with both Ω and Y compact and ϕ continuous in both variables.

However, the functions of interest may be defined on non-compact domains, their integral representations may have parameters in non-compact sets Y such as $\mathbb{R}$, and some computational units (such as Heaviside perceptrons) are not continuous. The following theorems include these cases.

Arguments are based on Bochner's extension of Lebesgue's integral to functions with values in Banach spaces. A sketch of such an approach was given by Girosi and Anzelotti (1993) for the case of convolutions.

We use the *Bochner integral* $I(w\Phi)$ of the mapping $w\Phi : Y \to \mathcal{X}$. For the definition of the Bochner integral, related notation, properties of the Bochner integral and the relationship of its evaluations to the Lebesgue integral see the Appendix.

We first prove upper bounds for parameterized sets $\Phi(Y)$ with the set of the parameters Y compact and the dependence Φ on parameters continuous and then we extend these bounds to the case of non-compact sets of parameters. We assume that the functions from the family $\Phi(Y)$ are either in a reproducing kernel Hilbert space or in $\mathcal{L}^q(\Omega,\rho)$-space, with $q \in (1,\infty)$ and ρ a Radon measure.

Recall that a triple $(Y,\mathcal{S},\mu)$ is called a *measure space* if Y is a set, $\mathcal{S}$ is a σ-algebra of subsets of Y, and μ is a measure on $\mathcal{S}$.

Theorem 4.1 *Let $(\mathcal{X},\|.\|_{\mathcal{X}})$ be a reproducing kernel Hilbert space of real-valued functions on a set Ω and $f \in \mathcal{X}$ can be represented for all $x \in \Omega$ as*

$$f(x) = \int_Y w(y)\phi(x,y)d\mu,$$

where Y, w, ϕ and μ satisfy both following conditions:
(i) Y is a compact subset of $\mathbb{R}^p$, p a positive integer, and $(Y,\mathcal{S},\mu)$ is a measure space,
(ii) $\Phi(Y)$ is a bounded subset of $\mathcal{X}$, $w \in \mathcal{L}^1(Y,\mu)$ and $w\Phi : Y \to \mathcal{X}$ is continuous.
Then $\|f\|_{\Phi(Y)} \leq \|w\|_{\mathcal{L}^1(Y,\mu)}$ and all positive integers n,

$$\|f - \mathrm{span}_n\Phi(Y)\|^2_{\mathcal{X}} \leq \frac{s^2_\Phi\|w\|^2_{\mathcal{L}^1(Y,\mu)} - \|f\|^2_{\mathcal{X}}}{n}.$$

Before proving this theorem, we state a similar result for $\mathcal{L}^q$-spaces and then give a joint proof, which splits at its last step. Our second theorem holds for $\mathcal{L}^q(\Omega,\rho)$ spaces where ρ is *σ-finite*, which means that there exists a family $\{M_i\}$ of sets of finite measure such that $\cup_{i=1}^\infty M_i = \Omega$. For example, the Lebesgue measure on $\mathbb{R}^d$ is σ-finite. The second theorem also requires a slightly stronger assumption on μ. A triple $(Y,\mathcal{S},\mu)$ is called a *Radon measure space* if Y is a topological space, $\mathcal{S}$ is a σ-algebra which includes all Borel sets, and μ is a Radon measure on $\mathcal{S}$, i.e., for every open subset U of Ω, $\rho(U) = \sup\{\rho(K) \,|\, K \subset U,\ K \text{ compact}\}$ and for every $A \in \mathcal{S}$, $\mu(A) = \inf\{\mu(U) \,|\, A \subset U \subseteq Y,\ U\text{open}\}$. Note that if μ is Radon and $K \subseteq Y$ is compact, then $\mu(K) < \infty$. A property is said to hold for μ-a.e. $y \in Y$ if it holds for all $y \in Y \setminus Y_0$, where $\mu(Y_0) = 0$.

Theorem 4.2 *Let $\mathcal{X} = \mathcal{L}^q(\Omega,\rho)$, $q \in [1,\infty)$, where $\Omega \subseteq \mathbb{R}^d$ and ρ is a σ-finite measure. Let $f \in \mathcal{X}$ can be represented for ρ-a.e. $x \in \Omega$ as*

$$f(x) = \int_Y w(y)\phi(x,y)d\mu,$$

where Y, w, ϕ, and μ satisfy all of the following three conditions:
(i) Y is a compact subset of $\mathbb{R}^p$, p a positive integer, and $(Y,\mathcal{S},\mu)$ is a Radon measure space,
(ii) $\Phi(Y)$ is a bounded subset of $\mathcal{X}$, $w \in \mathcal{L}^1(Y,\mu)$ and $w\Phi : Y \to \mathcal{X}$ is continuous,

(iii) $\phi : \Omega \times Y \to \mathbb{R}$ *is* $\rho \times \mu$*-measurable.*
Then $\|f\|_{\Phi(Y)} \le \|w\|_{\mathcal{L}^1(Y,\mu)}$ *and all positive integers* n*, when* $q \in (1,\infty)$ *and* q' *satisfies* $1/q+1/q' = 1$*,* $r = \min(q,q')$*,* $s = \max(q,q')$*,*

$$\|f - \operatorname{span}_n \Phi(Y)\|_{\mathcal{X}} \le \frac{2^{1/r} 2 s_\Phi \|w\|_{\mathcal{L}^1(Y,\mu)}}{n^{1/s}},$$

and when $q = 2$*,*

$$\|f - \operatorname{span}_n \Phi(Y)\|_{\mathcal{X}}^2 \le \frac{s_\Phi^2 \|w\|_{\mathcal{L}^1(Y,\mu)}^2 - \|f\|_{\mathcal{X}}^2}{n}.$$

Proof of Theorems 4.1 and 4.2.
Let $\zeta > 0$ be arbitrary. We will show that $\|f\|_{\Phi(Y)} \le \|w\|_{\mathcal{L}^1(Y,\mu)} + \zeta$.

Consider a sequence $\{\mathcal{P}_k\}$ of partitions of Y into μ-measurable sets $\mathcal{P}_k = \{P_{kj} \,|\, j = 1,\dots,m_k\}$, such that for each k, $\mathcal{P}_{k+1}$ is a refinement of $\mathcal{P}_k$ and the mesh of $\mathcal{P}_k$ is at most $1/k$ (the *mesh* of $\mathcal{P}_k$ is defined as $\max\{\operatorname{diam}(P_{kj}) \,|\, j = 1,\dots,m_k\}$, where $\operatorname{diam}(A) = \sup_{a,b\in A} d(a,b)$, and $d(a,b)$ denotes the Euclidean distance on $\mathbb{R}^p$).

For each $k \ge 1$ and each $j = 1,\dots,m_k$, choose $y_{kj}^\zeta \in P_{kj}$ such that

$$|w(y_{kj}^\zeta)| \le \frac{\zeta}{m_k}\mu(P_{kj}) + \inf_{y\in P_{kj}} |w(y)|.$$

Define a simple function $s_k^\zeta = s_k$ by

$$s_k(y) = \sum_{j=1}^{m_k} \chi_{P_{kj}}(y) w(y_{kj}^\zeta)\Phi(y_{kj}^\zeta).$$

By the definition of the Bochner integral, each $s_k \in \mathcal{I}(Y,\mu;\mathcal{X})$.

To show that also $w\Phi \in \mathcal{I}(Y,\mu;\mathcal{X})$, we use Lebesgue dominated convergence (Proposition 7.2). By compactness of Y and continuity of $w\Phi : Y \to \mathcal{X}$, $c = \sup_{y\in Y} |w(y)|\, \|\Phi(y)\|_{\mathcal{X}} < \infty$. Set $g(y) = c$ for all $y \in Y$, then $g \in \mathcal{L}^1(Y,\mu)$. For every $y \in Y$ and $k \ge 1$, there is at most one P_{kj} with $y \in P_{kj}$. Thus we have either $s_k(y) = 0 \le c$ or

$$\|s_k(y)\|_{\mathcal{X}} \le |w(y_{kj}^\zeta)|\, \|\Phi(y_{kj}^\zeta)\|_{\mathcal{X}} \le c = g(y).$$

Thus to apply Proposition 7.2 it remains to check that for μ-a.e. $y \in Y$, $\lim_{k\to\infty} \|s_k(y) - w\Phi(y)\|_{\mathcal{X}} = 0$.

As Y is compact, the continuous map $w\Phi : Y \to \mathcal{X}$ is uniformly continuous. Hence, for all $\varepsilon > 0$, there exists $\delta > 0$ such that for all $y_1, y_2 \in Y$, whenever $d(y_1,y_2) < \delta$, we have $\|w(y_1)\Phi(y_1) - w(y_2)\Phi(y_2)\|_{\mathcal{X}} < \varepsilon$, where $d(y_1,y_2)$ denotes the Euclidean distance on $\mathbb{R}^p$. For all $k > 1/\delta$, the mesh of $\mathcal{P}_k$ is smaller than δ and thus for μ-a.e. $y \in Y$, $\|s_k(y) - w(y)\Phi(y)\|_{\mathcal{X}} < \varepsilon$.

Therefore, according to Proposition 7.2,

$$w\Phi \in \mathcal{I}(Y,\mu;\mathcal{X}) \quad \text{and} \quad \lim_{k\to\infty} \|I(s_k) - I(w\Phi)\|_{\mathcal{X}} = 0. \tag{4.3}$$

By Proposition 3.4(i) and the choice of y_{kj}^ζ, for all k

$$\|I(s_k)\|_{\Phi(Y)} \le \sum_{j=1}^{m_k} \mu(P_{kj})|w(y_{kj}^\zeta)| \le \|w\|_{\mathcal{L}^1(Y,\mu)} + \zeta. \tag{4.4}$$

Since the sequence $\{\|I(s_k)\|_{\Phi(Y)}\}$ is bounded, replacing it with a subsequence if necessary, we get by Lemma 3.5, $\|I(w\Phi)\|_{\Phi(Y)} \le \lim_{k\to\infty} \|I(s_k)\|_{\Phi(Y)} \le \|w\|_{\mathcal{L}^1(Y,\mu)} + \zeta$. Infimizing over $\zeta > 0$, we obtain $\|I(w\Phi)\|_{\Phi(Y)} \le \|w\|_{\mathcal{L}^1(Y,\mu)}$.

Thus to get an upper bound on $\|f\|_{\Phi(Y)}$ it remains to show that the Bochner integral $I(w\Phi)$ is equal to f. Here the proofs of the two theorems split.

In the case of Theorem 4.1, we apply Proposition 7.3 to evaluation functionals denoted for each $x \in \Omega$ by T_x. By the definition of a RKHS, all evaluation functionals are bounded. Thus we get by Proposition 7.3, $I(w\Phi)(x) = T_x(I(w\Phi)) = \int_Y T_x(w\Phi(y))d\mu(y) = \int_Y (w\Phi(y))(x)d\mu(y) = \int_Y w(y)\phi(x,y)d\mu(y) = f(x)$. Hence $I(w\Phi) = f$.

In the case of Theorem 4.2, the equality $I(w\Phi) = f$ follows from Theorem 7.4 from the Appendix with $\Psi = w\Phi$.

The upper bound on $\|f - \mathrm{span}_n \Phi(Y)\|_{\mathcal{X}}$ then follows by Theorem 3.2 (in the Hilbert space case) and Theorem 3.3 (in the $\mathcal{L}^q$-space case). □

The next two theorems extend the upper bounds on rates of approximation also to the case when the parameter set Y is not compact and continuity of $w\Phi$ holds merely μ-a.e. on Y.

Theorem 4.3 *Let $(\mathcal{X}, \|.\|_{\mathcal{X}})$ be a reproducing kernel Hilbert space of real-valued functions on a set Ω and suppose that $f \in \mathcal{X}$ can be represented for all $x \in \Omega$ as*

$$f(x) = \int_Y w(y)\phi(x,y)d\mu(y),$$

where Y, w, ϕ and μ satisfy both following conditions:
(i) $Y \subseteq \mathbb{R}^p$, p a positive integer, $Y \setminus Y_0 = \cup_{m=1}^{\infty} Y_m$, where $\mu(Y_0) = 0$ and for all $m \geq 1$, Y_m is compact and $Y_m \subseteq Y_{m+1}$, and $(Y, \mathcal{S}, \mu)$ is a Radon measure space,
(ii) $\Phi(Y)$ is a bounded subset of $\mathcal{X}$, $w \in \mathcal{L}^1(Y,\mu)$, and $w\Phi : Y \setminus Y_0 \to \mathcal{X}$ is continuous.
Then $\|f\|_{\Phi(Y)} \leq \|w\|_{\mathcal{L}^1(Y,\mu)}$ and all positive integers n,

$$\|f - \mathrm{span}_n \Phi(Y)\|_{\mathcal{X}}^2 \leq \frac{s_\Phi^2 \|w\|_{\mathcal{L}^1(Y,\mu)}^2 - \|f\|_{\mathcal{X}}^2}{n}.$$

Theorem 4.4 *Let $\mathcal{X} = \mathcal{L}^q(\Omega, \rho)$, $q \in [1, \infty)$, where $\Omega \subseteq \mathbb{R}^d$ and ρ is σ-finite measure. Let $f \in \mathcal{X}$ satisfy for ρ-a.e. $x \in \Omega$*

$$f(x) = \int_Y w(y)\phi(x,y)d\mu(y),$$

where Y, w, ϕ, and μ satisfy the following three conditions:
(i) $Y \subseteq \mathbb{R}^p$, p a positive integer, $Y \setminus Y_0 = \cup_{m=1}^{\infty} Y_m$, where $\mu(Y_0) = 0$ and for all $m \geq 1$, Y_m is compact and $Y_m \subseteq Y_{m+1}$,
(ii) $\Phi(Y)$ is a bounded subset of $\mathcal{X}$, $w \in \mathcal{L}^1(Y,\mu)$, and $w\Phi : Y \setminus Y_0 \to \mathcal{X}$ is continuous,
(iii) $(Y, \mathcal{S}, \mu)$ is a Radon measure space and $\phi : \Omega \times Y \to \mathbb{R}$ is $\rho \times \mu$-measurable.
Then for all positive integers n, for all $q \in [1, \infty)$

$$\|f\|_{\Phi(Y)} \leq \|w\|_{\mathcal{L}^1(Y,\mu)},$$

for all $q \in (1, \infty)$ and q' satisfying $1/q + 1/q' = 1$, $r = \min(q, q')$, $s = \max(q, q')$,

$$\|f - \mathrm{span}_n \Phi(Y)\|_{\mathcal{X}} \leq \frac{2^{1/r} 2 s_\Phi \|w\|_{\mathcal{L}^1(Y,\mu)}}{n^{1/s}},$$

and for $q = 2$,

$$\|f - \mathrm{span}_n \Phi(Y)\|_{\mathcal{X}}^2 \leq \frac{s_\Phi^2 \|w\|_{\mathcal{L}^1(Y,\mu)}^2 - \|f\|_{\mathcal{X}}^2}{n}.$$

As most steps of the proofs of Theorems 4.3 and 4.4 are the same, we give a joint proof, which splits only at the step verifying the equality of evaluations of the Bochner integral $I(w\Phi)$ at ρ-a.e. $x \in \Omega$ to Lebesgue integrals $\int_Y w(y)\phi(x,y)d\mu(y)$.

Proof of Theorems 4.3 and 4.4.
For all $m \geq 1$ and all $x \in \Omega$, let $w_m : Y \to \mathbb{R}$, $\phi_m(x,.) : Y \to \mathbb{R}$, and $\Phi_m : Y \to \mathcal{X}$, resp., be defined as w, $\phi(x,.)$, and Φ on Y_m and as 0 on $Y \setminus Y_m$. As μ is a Radon measure, all compact sets Y_m have finite measures, and so

$$f_m(x) := \int_Y w_m(y)\phi_m(x,y)d\mu(y) = \int_{Y_m} w(y)\phi(x,y)d\mu(y)$$

are finite for all m. Thus by Theorems 4.1 and 4.2, $I(w_m\Phi_m) = f_m$ and $\|f_m\|_{\Phi(Y_m)} \leq \|w_{|Y_m}\|_{\mathcal{L}^1(Y_m)} \leq \|w\|_{\mathcal{L}^1(Y)}$. As $\Phi(Y_m) \subset \Phi(Y)$, we get $\|f_m\|_{\Phi(Y)} \leq \|f_m\|_{\Phi(Y_m)} \leq \|w\|_{\mathcal{L}^1(Y)}$.

We show that $\lim_{m\to\infty} \|f - f_m\|_{\mathcal{X}} = 0$ by first using Lebesgue dominated convergence to verify that $w\Phi$ is Bochner integrable with $\lim_{m\to\infty} \|I(w\Phi) - I(w_m\Phi_m)\|_{\mathcal{X}} = 0$ and then by showing that

$$I(w\Phi) = f. \tag{4.5}$$

By definition of w_m and Φ_m, for every $y \in Y \setminus Y_0$, there exists m_y such that for all $m \geq m_y$, $w_m(y)\Phi_m(y) = w(y)\Phi(y)$ and so for μ-a.e. $y \in Y$, $\lim_{m\to\infty} \|w_m(y)\Phi_m(y) - w(y)\Phi(y)\|_{\mathcal{X}} = 0$. For all $y \in Y$, $\|w_m(y)\Phi_m(y)\|_{\mathcal{X}} \leq s_\Phi w(y)$. As $s_\Phi w \in \mathcal{L}^1(Y,\mu)$, by Proposition 7.2 $w\Phi \in \mathcal{I}(Y,\mu;\mathcal{X})$ and

$$\lim_{m\to\infty} \|I(w\Phi) - I(w_m\Phi_m)\|_{\mathcal{X}} = 0.$$

To establish (4.5), we distinguish two cases. When $\mathcal{X}$ is a RKHS (Theorem 4.3), we use boundedness of evaluation functionals. For each $x \in \Omega$, let T_x denote the evaluation functional at x. By Proposition 7.3, $I(w\Phi)(x) = T_x(I(w\Phi)) = \int_Y T_x(w\Phi(y))d\mu(y) = \int_Y (w\Phi(y))(x)d\mu(y) = \int_Y w(y)\phi(x,y)d\mu(y) = f(x)$. So (4.5) holds. When $\mathcal{X} = \mathcal{L}^q(\Omega,\rho)$ (Theorem 4.4), the equation follows from Theorem 7.4 from the Appendix with $\Psi = w\Phi$. In both cases $\lim_{m\to\infty} \|f - f_m\|_{\mathcal{X}} = 0$ and thus by Lemma 3.5, $\|f\|_{\Phi(Y)} \leq \|w\|_{\mathcal{L}^1(Y)}$.

The upper bound on $\|f - \mathrm{span}_n\Phi(Y)\|_{\mathcal{X}}$ then follows by Theorems 3.2 (in the Hilbert space case) and Theorem 3.3 (in the $\mathcal{L}^q$-space case). □

Thus for functions representable as networks with infinitely many units, the growth of model complexity with increasing accuracy depends on the $\mathcal{L}^1$-norm of the output weight function.

5 Approximation by perceptron networks

To apply results from Section 4 to neural networks, we use the following straightforward corollary of Theorem 4.4 about parameterized families in $\mathcal{L}^2(\Omega) = \mathcal{L}^2(\Omega,\lambda)$, where λ denotes the Lebesgue measure.

Corollary 5.1 *Let $\Omega \subseteq \mathbb{R}^d$ be Lebesgue measurable, $f \in \mathcal{L}^2(\Omega)$ be such that for a.e. $x \in \Omega$,*

$$f(x) = \int_Y w(y)\phi(x,y)dy,$$

where Y, w, and ϕ satisfy the following three conditions:
(i) $Y \subseteq \mathbb{R}^p$ is Lebesgue measurable, p is a positive integer, $Y \setminus Y_0 = \cup_{m=1}^\infty Y_m$, where $\lambda(Y_0) = 0$ and for all positive integers m, Y_m is compact and $Y_m \subseteq Y_{m+1}$,
(ii) $\Phi(Y)$ is a bounded subset of $\mathcal{L}^2(\Omega)$, $w \in \mathcal{L}^1(Y)$, and $w\Phi : Y \setminus Y_0 \to \mathcal{X}$ is continuous,
(iii) $\phi : \Omega \times Y \to \mathbb{R}$ is Lebesgue measurable.
Then $\|f\|_{\Phi(Y)} \leq \|w\|_{\mathcal{L}^1(Y)}$ and for all positive integers n,

$$\|f - \mathrm{span}_n\Phi(Y)\|^2_{\mathcal{L}^2(\Omega)} \leq \frac{s_\Phi^2\|w\|^2_{\mathcal{L}^1(Y)} - \|f\|^2_{\mathcal{L}^2(\Omega)}}{n}.$$

A function $\sigma : \mathbb{R} \to \mathbb{R}$ is called *sigmoidal* when it is nondecreasing and $\lim_{t\to-\infty} \sigma(t) = 0$ and $\lim_{t\to\infty} \sigma(t) = 1$. For every compact $\Omega \subset \mathbb{R}^d$, the mapping

$$\Phi_\sigma : \mathbb{R}^d \times \mathbb{R} \to \mathcal{L}^2(\Omega),$$

which is defined for all $x \in \Omega$ as $\Phi_\sigma(v, b)(x) = \sigma(v \cdot x + b)$, maps parameters (input weights v and biases b) of perceptrons with the activation function σ to functions computable by such perceptrons.

Let $\vartheta : \mathbb{R} \to \mathbb{R}$ denote the *Heaviside function*, i.e., $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geq 0$, and S^{d-1} denote the unit sphere in $\mathbb{R}^d$. It is easy to see that for any bounded subset Ω of $\mathbb{R}^d$, $\Phi_\vartheta(S^{d-1} \times \mathbb{R}) = \Phi_\vartheta(\mathbb{R}^d \times \mathbb{R})$. It was shown by Kůrková et al. (1997) that for every $\Omega \subset \mathbb{R}^d$ compact and every continuous sigmoidal function σ, $\Phi_\sigma(\mathbb{R}^d \times \mathbb{R})$-variation in $\mathcal{L}^2(\Omega)$ is equal to $\Phi_\vartheta(S^{d-1} \times \mathbb{R})$-variation. Thus by Theorem 3.2, upper bounds on variation with respect to Heaviside perceptrons give estimates of rates of approximation by perceptron networks with an arbitrary continuous sigmoidal activation function.

It is easy to check that for Ω compact, $\Phi_\vartheta : S^{d-1} \times \mathbb{R} \to \mathcal{L}^2(\Omega)$ is continuous, $\Phi_\vartheta(S^{d-1} \times \mathbb{R})$ is a bounded subset of $\mathcal{L}^2(\Omega)$ and $\vartheta : \Omega \times S^{d-1} \times \mathbb{R} \to \mathbb{R}$ is Lebesgue measurable. Moreover, $S^{d-1} \times \mathbb{R}$ can be expressed as a union of a nested family of compact sets. Thus by Corollary 5.1 for function $f \in \mathcal{L}^2(\Omega)$ representable for all $x \in \Omega$ as $f(x) = \int_{S^{d-1}\times\mathbb{R}} w(e, b)\vartheta(e \cdot x + b)dedb$ with $w \in \mathcal{L}^1(S^{d-1} \times \mathbb{R})$, $\Phi_\vartheta(S^{d-1} \times \mathbb{R})$-variation of f is bounded from above by $\|w\|_{\mathcal{L}^1(S^{d-1}\times\mathbb{R})}$.

Sufficiently smooth functions which are either compactly supported or decay fast at infinity can be expressed as networks with infinitely many Heaviside perceptrons. It was shown by Kůrková et al. (1997) that, for d odd, compactly supported d-variable real-valued functions which are sufficiently differentiable have a representation of the form

$$f(x) = \int_{S^{d-1}\times\mathbb{R}} w_f(e, b)\vartheta(e \cdot x + b)dedb, \tag{5.1}$$

where the weight function $w_f(e, b)$ is a product of a function $a(d)$ of the number of variables d converging with d increasing exponentially fast to zero and a "flow of the order d through the hyperplane" $H_{e,b} = \{x \in \mathbb{R}^d \,|\, x \cdot e + b = 0\}$. More precisely,

$$w_f(e, b) = a(d) \int_{H_{e,b}} (D_e^{(d)}(f))(y)dy,$$

where

$$a(d) = (-1)^{(d-1)/2}(1/2)(2\pi)^{1-d}$$

and $D_e^{(d)}$ denotes the directional derivative of the order d in the direction e.

The integral representation (5.1) was extended in [Kainen et al., 2007b] to functions of a *weakly controlled decay*. Such functions have to satisfy for all multi-indexes α with $0 \leq |\alpha| = \alpha_1 + \ldots \alpha_d < d$, $\lim_{\|x\|\to\infty}(D^\alpha f)(x) = 0$ (where $D^\alpha = (\partial/\partial x_1)^{\alpha_1} \ldots (\partial/\partial x_d)^{\alpha_d}$) and there exists $\varepsilon > 0$ such that for each multi-index α with $|\alpha| = d$, $\lim_{\|x\|\to\infty}(D^\alpha f)(x)\|x\|^{d+1+\varepsilon} = 0$. The class of functions with weakly controlled decay contains all d-times continuously differentiable functions with compact support as well as all functions from the Schwartz class $\mathcal{S}(\mathbb{R}^d)$ (for the definition see [Adams and Fournier, 2003, p.251]). In particular, it contains the Gaussian function $\gamma_d(x) = \exp(-\|x\|^2)$.

Thus applying Corollary 5.1 to the integral representation (5.1) we get for a large class of functions the following upper bound on rates of approximation by perceptron networks. To avoid complicated notation, in the upper bound in $\mathcal{L}^2(\Omega)$-norm in the next theorem we assume that all functions are restricted to the set Ω.

Theorem 5.2 *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous sigmoidal function or σ be the Heaviside function, d be an odd positive integer, $f \in \mathcal{C}^d(\mathbb{R}^d)$ be either compactly supported with $\Omega = \mathrm{supp}(f)$ or f be of a weakly controlled decay and Ω be any compact subset of $\mathbb{R}^d$. Then for all positive integers n,*

$$\|f - \mathrm{span}_n\Phi_\sigma(\mathbb{R}^{d+1})\|^2_{\mathcal{L}^2(\Omega)} \leq \frac{\lambda(\Omega)^2\|w_f\|^2_{\mathcal{L}^1(S^{d-1}\times\mathbb{R})} - \|f\|^2_{\mathcal{L}^2(\Omega)}}{n},$$

where $w_f(e, b) = a(d)\int_{H_{e,b}}(D_e^{(d)}(f))(y)dy$, and $a(d) = (-1)^{(d-1)/2}(1/2)(2\pi)^{1-d}$.

An estimate in terms of the maximal value of the $\mathcal{L}^1$-norms of the partial derivatives of the function to be approximated can be derived from Theorem 5.2 by combining it with an upper bound on the $\mathcal{L}^1$-norm of the weighting function w_f from [Kainen et al., 2007b]. This bound is formulated in terms of a *Sobolev seminorm* $\|.\|_{d,1,\infty}$, which is defined as

$$\|f\|_{d,1,\infty} = \max_{|\alpha|=d} \|D^\alpha f\|_{\mathcal{L}^1(\mathbb{R}^d)}.$$

It was shown by Kainen et al. (2007b) that for all d odd and all f of a weakly controlled decay

$$\|w_f\|_{\mathcal{L}^1(S^{d-1}\times\mathbb{R})} \leq k(d)\|f\|_{d,1,\infty},$$

where $k(d) \sim \left(\frac{4\pi}{d}\right)^{1/2}\left(\frac{e}{2\pi}\right)^{d/2}$.

Corollary 5.3 *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous sigmoidal function or the Heaviside function, d be an odd positive integer, $f \in \mathcal{C}^d(\mathbb{R}^d)$ be either compactly supported with $\Omega = \mathrm{supp}(f)$ or f be of a weakly controlled decay and Ω be any compact subset of $\mathbb{R}^d$. Then for all positive integers n,*

$$\|f - \mathrm{span}_n \Phi_\sigma(\mathbb{R}^{d+1})\|^2_{\mathcal{L}^2(\Omega)} \leq \frac{k(d)^2\lambda(\Omega)^2\|f\|^2_{d,1,\infty} - \|f\|^2_{\mathcal{L}^2(\Omega)}}{n},$$

where $k(d) \sim \left(\frac{4\pi}{d}\right)^{1/2}\left(\frac{e}{2\pi}\right)^{d/2}$.

6 Conclusion

To apply tools from nonlinear approximation theory (the Maurey-Jones-Barron's theorem) to investigation of model complexity of neural networks, we developed a unifying framework for estimation of variational norms. Our proof technique is based on the idea of Girosi and Anzelotti (1993) of utilization of Bochner integral of mappings of parameters to functions computable by hidden units. Our estimates hold under mild assumptions on hidden units and output-weight functions and can be applied to a wide range of computational models of variable-basis type or "dictionaries".

We have shown that for functions representable as networks with infinitely many units, the growth of model complexity with increasing accuracy depends on the $\mathcal{L}^1$-norms of the output weight functions. Applying these estimates to integral representations in the form of networks with infinitely many Heaviside perceptrons, we derived estimates of rates of approximation by sigmoidal perceptron networks. Our estimates can be combined with many other integral representations, such as convolutions with Gaussian and Bessel kernels [Girosi and Anzellotti, 1993], [Kainen et al., 2007a].

7 Appendix: Properties of Bochner integral

The Bochner integral is a generalization of the Lebesgue integral to functions with values in a Banach space. Here, we recall the definition of the Bochner integral and some related concepts, notations, results and techniques needed in the proofs in our paperto understand following theorems and proofs (for more details see, e.g., [Zaanen, 1961]).

Let $(Y, \mathcal{S}, \mu)$ be a measure space. Let $\mathcal{X}$ be a Banach space with norm $\|\cdot\|_{\mathcal{X}}$. Call $s : Y \to \mathcal{X}$ a *simple function* if for $m \geq 1$, $f_1, \ldots, f_m \in \mathcal{X}$; $P_1, \ldots, P_m \in \mathcal{S}$ such that for all $j = 1, \ldots, m$, $\mu(P_j) < \infty$, for all distinct pairs $i, j = 1, 2, \ldots, m$, $P_i \cap P_j = \emptyset$, and

$$s = \sum_{j=1}^m f_j \chi_{P_j},$$

where χ_P denotes the characteristic function of the subset P of Y.

Let

$$I(s) := \sum_{j=1}^m \mu(P_j) f_j \in \mathcal{X}.$$

Then $I(s)$ is independent of the representation of s as a linear combination of characteristic functions [Zaanen, 1961, pp.130–132].

A function $h : Y \to \mathcal{X}$ is called *strongly measurable* (with respect to μ) provided there exists a sequence $\{s_k\}$ of simple functions such that, for μ-a.e. $y \in Y$,

$$\lim_{k\to\infty} \|s_k(y) - h(y)\|_{\mathcal{X}} = 0.$$

A function $h : Y \to \mathcal{X}$ is *Bochner integrable* (with respect to μ) if it is strongly measurable and there exists a sequence $\{s_k\}$ of simple functions $s_k : Y \to \mathcal{X}$ such that

$$\lim_{k\to\infty} \int_Y \|s_k(y) - h(y)\|_{\mathcal{X}} d\mu(y) = 0. \tag{7.1}$$

If (7.1) holds, the sequence $\{I(s_k)\}$ converges to an element $I(h) \in \mathcal{X}$, independent of the sequence of simple functions, called the *Bochner integral of* h (with respect to μ).

Let $\mathcal{I}(Y, \mu; \mathcal{X})$ denote the family of all functions from Y to $\mathcal{X}$ which are *Bochner integrable with respect to* μ.

The following theorem asserts that, for h strongly measurable, Bochner integrability of a mapping $h : Y \to \mathcal{X}$ is equivalent to Lebesgue integrability of $\|h\| : Y \to \mathbb{R}$.

Theorem 7.1 (Bochner) *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a Banach space and $(Y, \mathcal{S}, \mu)$ a measure space. Let $h : Y \to \mathcal{X}$ be strongly measurable. Then*

$$h \in \mathcal{I}(Y, \mu; \mathcal{X}) \text{ if and only if } \int_Y \|h(y)\|_{\mathcal{X}} d\mu(y) < \infty.$$

The next two results, which can be found, e.g., in [Zaanen, 1961, p. 132], [Martínez and Sanz, 2001, p. 324], are used in proofs in Section 4. The first one generalizes Lebesgue dominated convergence, while the second one describes a key linearity property.

Proposition 7.2 *Let $(Y, \mathcal{S}, \mu)$ be a measure space and $\mathcal{X}$ a Banach space. If $\{h_n\}_{n=1}^{\infty} \subset \mathcal{I}(Y, \mu; \mathcal{X})$ and $h : Y \to \mathcal{X}$ satisfies*

$$\lim_{n\to\infty} \|h_n(y) - h(y)\|_{\mathcal{X}} = 0$$

for μ-a.e. $y \in Y$, and if there exists $g \in \mathcal{L}^1(Y, \mu)$ with $\|h_n(y)\|_{\mathcal{X}} \le g(y)$ for μ-a.e. y in Y, then

$$h \in \mathcal{I}(Y, \mu; \mathcal{X}) \quad \text{and} \quad \lim_{n\to\infty} \|I(h) - I(h_n)\|_{\mathcal{X}} = 0.$$

Proposition 7.3 *Let $(Y, \mathcal{S}, \mu)$ be a measure space and let $\mathcal{X}$ be a Banach space. Let $h \in \mathcal{I}(Y, \mu; \mathcal{X})$ and let T be a bounded linear functional on $\mathcal{X}$. Then*

$$T(I(h)) = \int_Y T\left(h(y)\right) d\mu(y).$$

The following theorem on evaluation of Bochner integrals of mappings to $\mathcal{L}^q$-spaces was proven in [Kainen, 2007].

Theorem 7.4 *Suppose $(Y, \mathcal{S}, \mu)$ is a Radon measure space with $Y \subseteq \mathbb{R}^p$, $p \ge 1$, ρ a σ-finite measure on Ω, and $\psi : \Omega \times Y \to \mathbb{R}$ is $\rho \times \mu$-measurable, and for some $Y_0 \in \mathcal{S}$ with $\mu(Y_0) = 0$, and $Y \setminus Y_0$ is a countable union of compacta. Also suppose that for some q, $1 \le q < \infty$, $\Psi : Y \setminus Y_0 \to \mathcal{L}^q(\Omega, \rho)$ defined by $\Psi(y)(x) = \psi(x, y)$ is continuous and $\Psi(Y \setminus Y_0)$ is bounded. If for ρ-a.e. $x \in \Omega$, $f(x) = \int_Y \psi(x, y) d\mu(y)$, then Ψ is Bochner integrable and $I(\Psi) = f$.*

Acknowledgements

V. K. was partially supported by the Ministry of Education of the Czech Republic, project Center of Applied Cybernetics 1M684077004 (1M0567).

Bibliography

[Adams and Fournier, 2003] Adams, R. A. and Fournier, J. J. F. (2003). *Sobolev Spaces*. Academic Press, Amsterdam.

[Aronszajn, 1950] Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of AMS*, (68):33–404.

[Barron, 1992] Barron, A. R. (1992). Neural net approximation. In *Proc. 7th Yale Workshop on Adaptive and Learning Systems*, pages 69–72. Yale University Press.

[Barron, 1993] Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, (39):930–945.

[Darken et al., 1993] Darken, C., Donahue, M., Gurvits, L., and Sontag, E. (1993). Rate of approximation results motivated by robust neural network learning. In *Proceedings of the 6th Annual ACM Conference on Computational Learning Theory*, pages 303–309, New York. ACM.

[Girosi and Anzellotti, 1993] Girosi, F. and Anzellotti, G. (1993). Rates of convergence for radial basis functions and neural networks. In *Artificial Neural Networks for Speech and Vision*, pages 97–113, London. R. J. Mammone (Ed.), Chapman & Hall.

[Jones, 1992] Jones, L. K. (1992). A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Annals of Statistics*, (24):608–613.

[Kainen, 2007] Kainen, P. C. (2007). On evaluation of Bochner integrals. *Submitted for publication*.

[Kainen et al., 2002] Kainen, P. C., Kůrková, V., and Vogt, A. (2002). An integral formula for Heaviside neural networks. *Neural Network World*, (10):313–320.

[Kainen et al., 2007a] Kainen, P. C., Kůrková, V., and Sanguineti, M. (2007a). Estimates of approximation rates by Gaussian radial-basis functions. In *Adaptive and Natural Computing Algorithms – ICANNGA'07, (Eds. B. Beliczynski, A. Dzielinski, M. Iwanowski, B. Ribeiro), Part II, LNCS 4432*, pages 11–18, Berlin, Heidelberg. Springer-Verlag.

[Kainen et al., 2007b] Kainen, P. C., Kůrková, V., and Vogt, A. (2007b). A Sobolev-type upper bound for rates of approximation by linear combinations of Heaviside plane waves. *Journal of Approximation Theory*, (147):1–10.

[Kůrková, 1997] Kůrková, V. (1997). Dimension–independent rates of approximation by neural networks. In Warwick, K. and Kárný, M., editors, *Computer–Intensive Methods in Control and Signal Processing: Curse of Dimensionality*, pages 261–270. Birkhauser, Boston.

[Kůrková, 2003] Kůrková, V. (2003). High-dimensional approximation and optimization by neural networks, chapter 4. In Suykens, J., Horváth, G., Basu, S., Micchelli, C., and Vandewalle, J., editors, *Advances in Learning Theory: Methods, Models and Applications*, pages 69–88. IOS Press, Amsterdam.

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

[Kůrková and Sanguineti, 2002] Kůrková, V. and Sanguineti, M. (2002). Comparison of worst case errors in linear and neural network approximation. *IEEE Trans. on Information Theory*, (48):264–275.

[Kůrková and Sanguineti, 2005] Kůrková, V. and Sanguineti, M. (2005). Error estimates for approximate optimization by the extended Ritz method. *SIAM J. Optimization*, 2(15):461–487.

[Pisier, 1981] Pisier, G. (1981). Remarques sur un resultat non publié de B. Maurey. *Seminaire d'Analyse Fonctionelle 1980-81*, I(12).

[Martínez and Sanz, 2001] Martínez, C. and Sanz, M. (2001). *The Theory of Fractional Powers of Operators*. Elsevier, Amsterdam.

[Zaanen, 1961] Zaanen, A. C. (1961). *An Introduction to the Theory of Integration*. N. Holland, Amsterdam.