LETTER **Communicated by Yoshifusa Ito**

An Integral Upper Bound for Neural Network Approximation

Paul C. Kainen
kainen@georgetown.edu
Department of Mathematics, Georgetown University, Washington, D.C. 20057-1233, U.S.A.

Věra Kůrková
vera@cs.cas.cz
Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, Czech Republic

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 as the number of network units increases. These bounds are obtained for various norms using the framework of Bochner integration. Results are applied to perceptron networks.

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 an 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 a function by a network with a given type of units tells us which combinations of properties of the function and the computational units guarantee fast rates of approximation.

A suitable tool for estimating rates of neural network approximation is a result from nonlinear approximation theory that 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, it gives an upper bound on the square of the error in approximation by convex combinations of n elements of G. The estimate was first proved by Maurey (see Pisier, 1981) using a probabilistic argument. Jones (1992) derived a slightly weaker estimate constructively with an iterative algorithm. Barron (1993) refined Jones's constructive argument to obtain the same estimate as Maurey. Various extensions to approximation errors measured by $\mathcal{L}^p$ and

supremum norms were also derived (Darken, Donahue, Gurvits, & Sontag, 1993; Donahue, Gurvits, Darken, & Sontag, 1997; Gurvits & Koiran, 1997).

The Maurey-Jones-Barron estimate can be reformulated in terms of a certain norm (called G-variation) of the function to be approximated. This norm was defined in Barron (1992) for sets of characteristic functions and extended in Kůrková (1997) to general sets. Investigation of properties of variational norms for G corresponding to various types of network units (or, alternatively, characterization of functions belonging to convex hulls of such sets of hidden unit functions) can provide some insight into the impact of a choice of the type of units on network complexity.

Jones (1992) suggested applying his estimate to functions with suitable integral representations by rewriting them as infinite convex combinations of elements from a trigonometric dictionary. Barron (1993, theorem 2) rigorously proved that such functions belong to the closure of the convex hull of the trigonometric dictionary using the law of large numbers and Fubini's theorem. He also applied the estimate to dictionaries formed by sigmoidal perceptrons by approximating sines by sigmoidals.

Girosi and Anzellotti (1993) applied the Maurey-Jones-Barron estimate to convolutions with gaussian and Bessel kernels, sketching an argument based on application of the concept of Bochner integration to functions representable as infinite networks. The Bochner integral extends the concept of Lebesgue integral to mappings into Banach spaces; when the Banach space consists of the set of all real-valued functions (or equivalence classes of functions) with certain properties, the value of a Bochner integral is such a function, not just a number. Bochner integration can be applied to mappings assigning to parameter (weights, biases, centroids) functions computable by perceptron or radial basis function units determined by these parameters.

Explicitly in terms of an upper bound on variational norm, in Kůrková, Kainen, and Kreinovich (1997), an estimate of rates of approximation is derived for compactly supported functions representable as infinite networks with any continuous hidden unit function. The estimate was shown to hold for the Heaviside function as well. Also, the variational norm was bounded by the $\mathcal{L}^1$-norm of the output weight function from the infinite network. For networks of Heaviside perceptrons, Kainen, Kůrková, and Vogt (2007) extend the upper bound to the case of a noncompact parameter set. This allows one to apply the Maurey-Jones-Barron estimate to a wide class of functions representable as integrals of Heaviside plane waves. Such representations were first derived in Ito (1991) using the Radon transform. For d odd, these representations were rediscovered in Kůrková et al. (1997) using an integral representation of the d-dimensional Dirac delta function under weaker smoothness conditions.

In this letter, we develop the idea of Girosi and Anzelotti (1993) of applying the concept of the Bochner integral to approximation from a dictionary. Using properties of the Bochner integral, we obtain a framework for investigation of functions having integral representations as infinite networks

with many kinds of hidden unit functions and sets of parameters. We prove under mild assumptions that the size of the $\mathcal{L}^1$-norm of the output-weight function from the infinite network is an important factor in network complexity. This gives some theoretical justification for regularization based on output-weight decay.

We illustrate our results on perceptron networks. Combining a representation of a smooth function as an integral combination of Heaviside perceptrons (Ito, 1991; Kůrková et al., 1997) with the estimate of variational norm in terms of the $\mathcal{L}^1$-norm of the output weight function, we obtain an upper bound on rates of approximation by perceptron networks for a wide class of functions. A preliminary version of some results appeared in conference proceedings (Kainen & Kůrková, 2008).

The letter is organized as follows. Section 2 introduces our approach and notation. Section 3 recalls the Maurey-Jones-Barron theorem and derives some properties of 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 integrals are summarized in the appendix.

2 Outline of the Approach

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

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

where G is a set of functions and is sometimes called a *dictionary*. To avoid trivialities, we assume that G contains nonzero elements. 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

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

where $\phi : \Omega \times Y \to \mathbb{R}$, Y is the set of parameters, and Ω is the set of input 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

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

In this letter, we consider parameterized sets of functions belonging to either an $\mathcal{L}^q$-space with $q \in [1, \infty)$ or a Banach space $(\mathcal{X}, \|.\|_{\mathcal{X}})$ of pointwise-defined functions on which all evaluation functionals are bounded.

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 write merely $\mathcal{L}^q(\Omega)$.

For $x \in \Omega$, we denote by $T_x : \mathcal{X} \to \mathbb{R}$ the evaluation functional at x defined for every $f \in \mathcal{X}$ as

$$T_x(f) := f(x).$$

The class of spaces with bounded evaluation functionals contains all spaces of bounded functions with the supremum norm. It also contains all reproducing kernel Hilbert spaces (RKHS), which are defined as Hilbert spaces of point-wise defined real-valued functions on which all evaluation functionals are bounded (Aronszajn, 1950).

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

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

We investigate the speed of decrease of distances $\|f - \text{span}_n \Phi(Y))\|_{\mathcal{X}}$ 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 f, which can be expressed for a suitable measure μ on Y and almost all $x \in \Omega$ as the Lebesgue integrals of the form

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

where $w : Y \to \mathbb{R}$ is the weight function.

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:

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

To derive this upper bound, we use the previously mentioned result of Maurey, Jones, and Barron on a variable-basis approximation, reformulating it 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 the Lebesgue integral allowing the 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

$$w\Phi(y) := w(y)\Phi(y) = w(y)\phi(., y). \tag{2.4}$$

Using the relationship between the Lebesgue integral, equation 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 the upper bound, equation 2.3, on rate of approximation by $\text{span}_n \Phi(Y)$.

3 Rates of Variable-Basis Approximation and Variational Norm

The following theorem is an upper bound on approximation by

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

derived by Maurey (see Pisier, 1981), Jones (1992), and Barron (1993).

Theorem 1 (Maurey-Jones-Barron). *Let G be a bounded nonempty subset of a Hilbert space $(\mathcal{X}, \|.\|_{\mathcal{X}})$ and $s_G = sup_{g\in G}\|g\|_{\mathcal{X}}$. Then for every $f \in cl\ conv\, G$ and for every positive integer n,*

$$\|f - conv_n G\|_{\mathcal{X}}^2 \leq \frac{s_G^2 - \|f\|_{\mathcal{X}}^2}{n}.$$

Theorem 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 $(\mathcal{X}, \|.\|_{\mathcal{X}})$ as the Minkowski functional of the closed convex symmetric hull of G, that is,

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

where the closure cl is taken with respect to the topology generated by the norm $\|.\|_{\mathcal{X}}$ and conv denotes the convex hull. Note that G-variation can be infinite. It is a norm (so it is subadditive, i.e., $\|f + g\|_G \leq \|f\|_G + \|g\|_G$) 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., $d = 1$), the 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 1 from Kůrková (1997; see also Kůrková, 2003).

Theorem 2. *Let $(\mathcal{X}, \|.\|_{\mathcal{X}})$ be a Hilbert space and 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,*

$$\|f - span_n G\|_{\mathcal{X}}^2 \leq \frac{s_G^2\|f\|_G^2 - \|f\|_{\mathcal{X}}^2}{n}.$$

This reformulation of theorem 1 in terms of the 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 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 & Sanguineti, 2005). For the definition of Radon measure see section 4.

Theorem 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 $f \in$ cl conv G and every positive integer n,*

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

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 continuous sigmoidal activation function (Kůrková et al., 1997). So to obtain from theorem 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., 2007). Thus, investigation of variational norms can provide some insight into properties of multivariable functions, which can be efficiently approximated by various computational models.

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 1. *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 $\|h\|_G \leq b$.*

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, $i > i_0$, $\|\frac{h}{b+\eta} - \frac{h_i}{b_i+\eta}\|_{\mathcal{X}} \leq \|\frac{h}{b+\eta} - \frac{h}{b_i+\eta}\|_{\mathcal{X}} + \|\frac{h}{b_i+\eta} - \frac{h_i}{b_i+\eta}\|_{\mathcal{X}} \leq \frac{\eta\|h\|_{\mathcal{X}}}{(b+\eta)(b_i+\eta)} + \frac{\eta}{b_i+\eta} \leq \frac{\eta\|h\|_{\mathcal{X}}}{b^2} + \frac{\eta}{b} < \frac{\varepsilon}{2}$.

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

4 Upper Bound on Variation with Respect to a Parameterized Family

It is easy to see that for $f \in \mathcal{X}$ representable as $f = \sum_{i=1}^{k} w_i g_i$ with all $g_i \in G$ and $w_i \in \mathbb{R}$, $\|f\|_G \leq \sum_{i=1}^{k} |w_i|$. By analogy, for f representable as

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

one should expect

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

Various special cases of integral representations of the form 4.1 have been investigated. Jones (1992) and Barron (1993) used weighted Fourier transform, and Girosi and Anzellotti (1993) used convolutions to prove that functions with such representations belong to the convex hulls of corresponding dictionaries. Explicitly as an upper bound on variation, the

estimate in terms of the $\mathcal{L}^1(Y, \mu)$-norm of the weight function w was derived in Kůrková et al. (1997) for an integral representation 4.1 of a function f defined on a compact domain $\Omega \subset \mathbb{R}^d$, the set of parameters Y compact and the hidden-unit function ϕ either continuous in both variables or ϕ corresponding to Heaviside perceptrons.

However, the functions of interest may be defined on noncompact domains, their integral representations may have parameters in noncompact sets Y such as $\mathbb{R}^d$, 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 the Lebesgue integral to functions with values in Banach spaces. For a mapping $h : Y \to \mathcal{X}$ from a measure space (Y, μ) to a Banach space $\mathcal{X}$, we denote by $I(h)$ the Bochner integral of h (if it exists), which is an element of $\mathcal{X}$ (see the appendix for a brief review, including definitions and basic properties). In the proofs of the next theorems, we consider the Bochner integral of the mapping $h = w\Phi$ defined in equation 2.4. Girosi and Anzellotti (1993) originally sketched such an approach for the case of integral representations in the form of convolutions.

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 noncompact sets of parameters. We assume that the functions from the family $\Phi(Y)$ are either in $\mathcal{L}^q(\Omega, \rho)$-space, with $q \in (1, \infty)$ and ρ a Radon measure, or in a Banach space, on which all evaluation functionals are bounded (this class includes all reproducing kernel Hilbert spaces and the space of bounded continuous functions with the supremum norm). 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. *Let $(\mathcal{X}, \|.\|_{\mathcal{X}})$ be a Banach space of real-valued functions on a set $\Omega \subseteq \mathbb{R}^d$ such that all evaluation functionals on $\mathcal{X}$ are bounded, and suppose that $f \in \mathcal{X}$ is 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 $(Y, \mathcal{S}, \mu)$ 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

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

Moreover, if, in addition, $(\mathcal{X}, \|.\|_{\mathcal{X}})$ *is a Hilbert space, then for all positive integers* n,

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

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 $\mathcal{S}$, that is, for every open subset U of Ω, $\rho(U) = \sup\{\rho(K) \mid K \subset U,\ K \text{ compact}\}$ and for every $A \in \mathcal{S}$, $\mu(A) = \inf\{\mu(U) \mid A \subset U \subseteq Y,\ U \text{open}\}$. 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 $\mu(Y_0) = 0$. The requirements on the measures and on ϕ and w enable the application of Fubini's theorem.

Theorem 5. *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* $f \in \mathcal{X}$ *is 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

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

and for all positive integers n, *when* $q \in (1, \infty)$ *and* q' *satisfies* $1/q + 1/q' = 1$, $r = min(q, q')$, $s = max(q, q')$,

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

and when $q = 2$,

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

Proof of Theorems 4 and 5. 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 $\mathcal{P}_k = \{P_{kj} \mid j = 1, \dots, m_k\}$, 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 $\max\{\operatorname{diam}(P_{kj}) \mid j = 1, \dots, m_k\}$, 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

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

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

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

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

To show that $w\Phi \in \mathcal{I}(Y, \mu; \mathcal{X})$, we use Lebesgue-dominated convergence (see proposition 1). 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

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

Thus, to apply proposition 1, 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 1,

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

By the choice of y_{kj}^ζ, for all k

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

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

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.

For theorem 4, we apply proposition to evaluation functionals. Thus, we get by proposition 2, $I(w\Phi)(x) = T_x(I(w\Phi)) = \int_Y T_x(w\Phi(y))\,d\mu(y) = \int_Y (w\Phi(y))(x)\,d\mu(y) = \int_Y w(y)\phi(x,y)\,d\mu(y) = f(x)$. Hence, $I(w\Phi) = f$. For theorem 5, the equality $I(w\Phi) = f$ is proved below at the end of the proof of theorem 7.

The upper bound on $\|f - \mathrm{span}_n \Phi(Y)\|_{\mathcal{X}}$ then follows by theorem 2 (in the Hilbert space case) and theorem 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.

Theorem 6. *Let $(\mathcal{X}, \|.\|_{\mathcal{X}})$ be a Banach space of real-valued functions on a set $\Omega \subseteq \mathbb{R}^d$ such that all evaluation functionals on $\mathcal{X}$ are bounded and suppose that $f \in \mathcal{X}$ is 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 the following conditions:

- *(i) Y is an open 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.*

Then

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

If, in addition, $(\mathcal{X}, \|.\|_{\mathcal{X}})$ is a Hilbert space, then for all positive integers n,

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

Theorem 7. *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 $f \in \mathcal{X}$ 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 is an open 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 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 $1/q + 1/q' = 1$, $r = min(q, q')$, $s = max(q, q')$,

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

and for $q = 2$,

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

As most steps of the proofs of theorems 6 and 7 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 6 and 7. Since Y is an open subset of $\mathbb{R}^p$, it is well known (and easy to check) that Y is the union of a countable family of compact subsets Y_m, which may be taken to be nested, so

$$Y = \cup_{m=1}^{\infty} Y_m$$

with $Y_m \subseteq Y_{m+1}$. This condition could replace the stronger requirement that Y is open in theorems 6 and 7.

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 1 and 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 1, $w\Phi \in \mathcal{I}(Y, \mu; \mathcal{X})$ and

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

To establish equation 4.5, there are two cases.

For spaces with bounded evaluation functionals (see theorem 6), by proposition 2, $I(w\Phi)(x) = T_x(I(w\Phi)) = \int_Y T_x(w\Phi(y))\, d\mu(y) = \int_Y (w\Phi(y))(x)\, d\mu(y) = \int_Y w(y)\phi(x, y)\, d\mu(y) = f(x)$. So equation 4.5 holds.

For $\mathcal{X} = \mathcal{L}^q(\Omega, \rho)$ (see theorem 6), we must show that for ρ-a.e. $x \in \Omega$, $f(x) = I(w\Phi)(x)$. It is equivalent to showing that for each bounded linear functional F on $\mathcal{X}$, $F(I(w\Phi)) = F(f)$. By the Riesz representation theorem (Martínez & Sanz, 2001), for any such F, there exists a $g_F \in \mathcal{L}^{q'}(\Omega, \rho)$ such that for all $g \in \mathcal{L}^q(\Omega, \rho)$, $F(g) = \int_\Omega g_F(x)g(x)d\rho(x)$, where $1/q + 1/q' = 1$.

As F is a bounded linear functional, by proposition 2, we have

$$\begin{aligned} F(I(w\Phi)) &= \int_Y w(y)F\,(\Phi(y))\; d\mu(y) \\ &= \int_Y \int_\Omega w(y)g_F(x)\phi(x, y)\, d\rho(x)\, d\mu(y). \end{aligned}$$

On the other hand,

$$F(f) = \int_\Omega g_F(x)f(x)\, d\rho(x) = \int_\Omega \int_Y w(y)g_F(x)\phi(x, y)\, d\mu(y)\, d\rho(x).$$

Since μ and ρ are σ-finite, we can apply Fubini's theorem (Hewitt & Stromberg, 1965) to show that the two iterated integrals are equal. For the verification that the absolute integral $I = \int_Y |w(y)F\,(\Phi(y))|d\mu(y)$ is indeed finite, replace μ by the finite measure $S \mapsto \int_S |w(y)|d\mu(y)$ and use Hölder's inequality (Martínez & Sanz, 2001) to see that

$$I \leq \|w\|_{\mathcal{L}^1(Y,\mu)} \|g_F\|_{\mathcal{L}^{q'}(\Omega,\rho)} \sup_{y\in Y} \|\Phi(y)\|_{\mathcal{L}^q(\Omega,\rho)}.$$

In both the Hilbert space and $\mathcal{L}^q$ cases, $\lim_{m\to\infty} \|f - f_m\|_{\mathcal{X}} = 0$ and thus by lemma 1, $\|f\|_{\Phi(Y)} \leq \|w\|_{\mathcal{L}^1(Y)}$. The upper bound on $\|f - \text{span}_n \Phi(Y)\|_{\mathcal{X}}$ then follows by theorem 2 (in the Hilbert space case) and theorem 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

In this section, we use the following specialization of theorem 7 about parameterized families in $\mathcal{L}^2(\Omega) = \mathcal{L}^2(\Omega, \lambda)$, where λ denotes the Lebesgue measure.

Corollary 1. *Let $\Omega \subseteq \mathbb{R}^d$, $d \geq 1$, be Lebesgue measurable, and suppose that $f \in \mathcal{L}^2(\Omega)$ is such that for λ-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 is an open subset of $\mathbb{R}^p$ and p a positive integer.*
- *(ii) $\Phi(Y)$ is a bounded subset of $\mathcal{L}^2(\Omega)$, $w \in \mathcal{L}^1(Y)$, and $w\Phi : Y \to \mathcal{L}^2(\Omega)$ 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,

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

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 $\Phi_\sigma(v,b)(x) := \phi_\sigma(v,b)(x) = \sigma(v \cdot x + b)$, 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, that is, $\vartheta(t) = 0$ 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 $\mathbb{R}^d$, $\Phi_\vartheta(S^{d-1} \times \mathbb{R}) = \Phi_\vartheta(\mathbb{R}^d \times \mathbb{R})$. Kůrková et al. (1997) showed 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 2, upper bounds on variation with respect to Heaviside perceptrons give estimates on 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 $\phi_\vartheta : \Omega \times S^{d-1} \times \mathbb{R} \to \mathbb{R}$ 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 1 for function $f \in \mathcal{L}^2(\Omega)$ representable for all $x \in \Omega$ as $f(x) = \int_{S^{d-1}\times\mathbb{R}} w(e,b)\vartheta(e \cdot x + b)\, de\, db$ 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 that are either compactly supported or have sufficiently rapid decay at infinity (along with their derivatives) can be expressed as networks with infinitely many Heaviside perceptrons. For d odd, such functions have a representation of the form

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

with

$$w_f(e,b) = a(d) \int_{H_{e,b}} \left(D_e^{(d)}(f)\right)(y)\, dy, \tag{5.2}$$

where $H_{e,b} = \{x \in \mathbb{R}^d \mid x \cdot e + b = 0\}$ and $a(d) = (-1)^{(d-1)/2}(1/2)(2\pi)^{1-d}$; $D_e^{(d)}$ denotes the directional derivative of order d in the direction e. The weight function $w_f(e,b)$ is thus a flow of order d through the hyperplane, scaled by $a(d)$, which goes to zero exponentially fast as $d \to \infty$.

Representation 5.1 was first derived in Ito (1991) (see theorem 3.1, proposition 2.2, and an equation on p. 387 of his paper). Ito used the Radon transform (see, e.g., Adams & Fournier, 2003) to prove that all functions from the Schwartz class have such a representation. In Kůrková et al. (1997), the same formula was derived for all compactly supported functions from $\mathcal{C}^d(\mathbb{R}^d)$, d odd, via an integral formula for the Dirac delta function. Equation 5.1 was extended to functions of *weakly controlled decay* in Kainen, Kůrková, and Vogt (in press). These are the functions that 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 of weakly controlled decay contains both above-mentioned classes (the Schwartz class and the class of all compactly supported functions from $\mathcal{C}^d(\mathbb{R}^d)$). In particular, it contains the gaussian function $\gamma_d(x) = \exp(-\|x\|^2)$.

Thus, applying corollary 1 and the remark at the beginning of the proof of theorems 6 and 7 to the integral representation 5.1 and taking advantage of the equality of $\Phi_\sigma(\mathbb{R}^{d+1})$-variation and $\Phi_\vartheta(S^{d-1} \times \mathbb{R})$-variation (Kůrková et al., 1997), 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 suitable functions are restricted to the set Ω.

Theorem 8. *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 $\Omega = supp(f)$ or f be of weakly controlled decay, and Ω be any compact subset of $\mathbb{R}^d$. Then for all positive integers n,*

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

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 8 by combining it with an upper bound on the $\mathcal{L}^1$-norm of the weighting function w_f from Kainen et al. (2007). This bound is formulated in terms of a Sobolev seminorm $\|.\|_{d,1,\infty}$, which is defined as

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

Kainen et al. (2007) showed that for all d odd and all f of 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 2. *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 $\Omega = supp(f)$ or f be of weakly controlled decay, and Ω be any compact subset of $\mathbb{R}^d$. Then for all positive integers n,*

$$\|f - span_n \Phi_\sigma(\Omega)\|^2_{\mathcal{L}^2(\Omega)} \leq \frac{k(d)^2 \lambda(\Omega)^2 \|f\|^2_{d,1,\infty} - \|f\|^2_{\mathcal{L}^2(\Omega)}}{n},$$

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 theorem and its extensions) in investigating model complexity of neural networks, we developed a unifying framework for the estimation of variational norms. Our proof technique is based on the idea of Girosi and Anzellotti (1993) of using the 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 function spaces and computational models of variable-basis or "dictionary" type. In fact, we believe that the hypothesis of continuity is too strong; w and ϕ measurable (in all variables) should be sufficient. But the formulation here is enough for our applications.

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. This leads to estimates of rates of approximation by sigmoidal perceptron networks. However, our estimates can also be combined with many other integral representations, for example, convolutions with gaussian and Bessel kernels, which were studied in Girosi and Anzellotti (1993) and Kainen, Kůrková, and Sanguineti (2009).

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 letter (for more details see, e.g., Zaanen, 1961).

Let $(Y, \mathcal{S}, \mu)$ be a measure space and $(\mathcal{X}, \|.\|_\mathcal{X})$ be a Banach space. A function $s : Y \to \mathcal{X}$ is called *simple* if it achieves a finite set of values; that

is, there exist $m \geq 1$, $f_1, \ldots, f_m \in \mathcal{X}$; $P_1, \ldots, P_m \in \mathcal{S}$ 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).

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

$$\lim_{k \to \infty} \int_Y \|s_k(y) - h(y)\|_{\mathcal{X}} d\mu(y) = 0. \tag{A.1}$$

If equation A.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 $\mathcal{X}$ that 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 9 (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 in Zaanen (1961) and Martínez and Sanz (2001), are used in proofs in section 4. The first one generalizes

Lebesgue-dominated convergence, while the second one describes a key linearity property.

Proposition 1. *Let $(Y, \mathcal{S}, \mu)$ be a measure space and $(\mathcal{X}, \|.\|_{\mathcal{X}})$ 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 $\|h_n(y)\|_{\mathcal{X}} \leq g(y)$ 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 2. *Let $(Y, \mathcal{S}, \mu)$ be a measure space, $(\mathcal{X}, \|.\|_{\mathcal{X}})$ a Banach space, $h \in \mathcal{I}(Y, \mu; \mathcal{X})$, and let T be a bounded linear functional on $\mathcal{X}$. Then*

$$T(I(h)) = \int_Y T\,(h(y))\; d\mu(y).$$

Acknowledgments

V. K. was partially supported by the Ministry of Education of the Czech Republic, project Center of Applied Cybernetics 1M684077004 (1M0567) and the Institutional Research Plan AV0Z10300504. P.C.K. was partially supported by Georgetown University.

References

Adams, R. A., & Fournier, J. J. F. (2003). *Sobolev spaces*. Orlando, FL: Academic Press.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of AMS, 68*, 337–404.

Barron, A. R. (1992). Neural net approximation. In *Proc. 7th Yale Workshop on Adaptive and Learning Systems* (pp. 69–72). New Haven, CT: Yale University Press.

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

Darken, C., Donahue, M., Gurvits, L., & 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* (pp. 303–309). New York: ACM.

Donahue, M., Gurvits, L., Darken, C., & Sontag, E. (1997). Rates of convex approximation in non-Hilbert spaces. *Constructive Approximation, 13*, 187–220.

Girosi, F., & Anzellotti, G. (1993). Rates of convergence for radial basis functions and neural networks. In R. J. Mammone (Ed.), *Artificial neural networks for speech and vision* (pp. 97–113). London: Chapman & Hall.

Gurvits, L., & Koiran, P. (1997). Approximation and learning of convex superpositions. *J. Computer and System Sciences, 55*, 161–170.

Hewitt, E., & Stromberg, K. (1965). *Real and abstract analysis*. New York: Springer-Verlag.

Ito, Y. (1991). Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory. *Neural Networks, 4*, 385–394.

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, 20*, 608–613.

Kainen, P. C., & Kůrková, V. (2008). Estimates of network complexity and integral representations. In V. Kůrková, R. Neruda, & J. Koutník (Eds.), *Artificial neural networks—ICANN 2008* (pp. 31–40). Berlin: Springer-Verlag.

Kainen, P. C., Kůrková, V., & Sanguineti, M. (2009). Estimates of approximation rates by gaussian radial-basis functions. *Journal of Complexity, 25*, 63–74.

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

Kainen, P. C., Kůrková, V., & Vogt, A. (forthcoming). Integral combinations of Heavisides. *Mathematische Nachrichten*.

Kůrková, V. (1997). Dimension–independent rates of approximation by neural networks. In K. Warwick & M. Kárný (Eds.), *Computer-intensive methods in control and signal processing: Curse of dimensionality* (pp. 261–270). Boston: Birkhauser.

Kůrková, V. (2003). High-dimensional approximation and optimization by neural networks. In J. Suykens, G. Horváth, S. Basu, C. Micchelli, & J. Vandewalle (Eds.), *Advances in learning theory: Methods, models and applications* (pp. 69–88). Amsterdam: IOS Press.

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

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

Martínez, C., & Sanz, M. (2001). *The theory of fractional powers of operators*. Amsterdam: Elsevier.

Pisier, G. (1981). Remarques sur un resultat non publié de B. Maurey. *Seminaire d'Analyse Fonctionelle 1980–81*, no. 5, 1–12.

Zaanen, A. C. (1961). *An Introduction to the Theory of Integration*. Amsterdam: North-Holland.

Received April 1, 2008; accepted February 7, 2009.