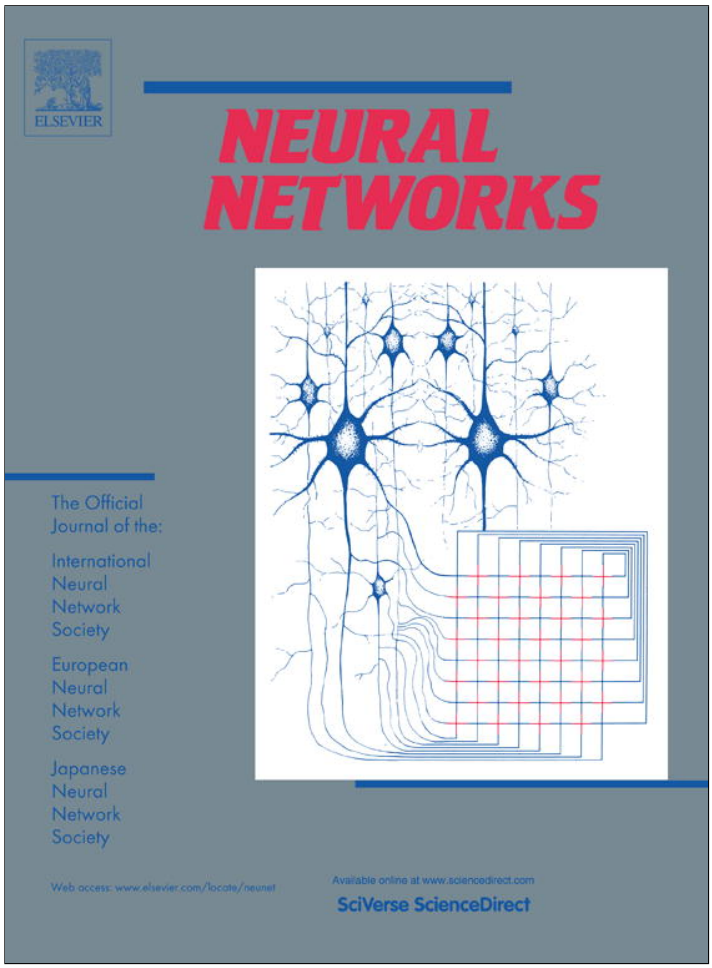

(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

Neural Networks 33 (2012) 160–167

Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Complexity estimates based on integral transforms induced by computational units

Věra Kůrková *

Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží2, 182 07 Prague, Czech Republic

ARTICLE INFO

Article history:
Received 11 August 2011
Received in revised form 20 March 2012
Accepted 9 May 2012

Keywords:
Neural networks
Estimates of model complexity
Approximation from a dictionary
Integral transforms
Norms induced by computational units

ABSTRACT

Integral transforms with kernels corresponding to computational units are exploited to derive estimates of network complexity. The estimates are obtained by combining tools from nonlinear approximation theory and functional analysis together with representations of functions in the form of infinite neural networks. The results are applied to perceptron networks.

1. Introduction

Integral transformations play an important role in many branches of applied science. A large class of such transformations has the form

$$T_K(f)(x) = \int f(y)K(x,y)dy,$$

where the function of two variables K is called a kernel of the integral operator T_K (the term "kernel" is derived from the German term "kern" introduced by Hilbert in 1904 Pietch (1987, p. 291)). Also functions computable by units used in neurocomputing depend on two vector variables, an input and a parameter, and thus they can be considered as kernels. An integral transformation with a kernel corresponding to a computational unit computing a function $\phi : \Omega \times A \to \mathbb{R}$, where Ω is a set of inputs and A is a set of parameters, can be viewed as a mapping T_ϕ assigning to an output weight function $w : A \to \mathbb{R}$ an input–output function $T_\phi(w) : \Omega \to \mathbb{R}$ in the form

$$T_\phi(w)(x) = \int w(y)\phi(x,y)dy$$

of a network with one linear output and one hidden layer with infinitely many computational units.

Integral transformations have been used in the mathematical theory of neurocomputing since the early 1990s. First, they occurred in proofs of the universal approximation property. Carroll and Dickinson (1989) and Ito (1991) used the Radon transform to show that functions satisfying various smoothness assumptions can be represented as integrals in the form of networks with infinitely many sigmoidal perceptrons. Discretizing these integral representations they proved the universal approximation property of perceptron networks. Park and Sandberg (1991, 1993) derived the universal approximation property of radial-basis function networks in $\mathcal{L}^p$-spaces using convolutions with properly scaled kernels. Similar ideas were used by Mhaskar (2004, 2006), see also Schaback and Wendland (2006) and references therein. Note that the use of integral transforms for approximation of functions is very common in approximation theory. The book (DeVore & Lorentz, 1993) gives many examples (the best constants in the Favard inequality in the trigonometric polynomial approximation are obtained in terms of an integral of the target function derivative against a Bernoulli spline kernel). The idea of discretizing integral transforms to obtain approximation as a discrete sum is also very old (see, e.g., Bernstein, 1931; Szabados, 1974).

Later, integral transforms with kernels corresponding to computational units were employed to obtain estimates of network complexity. Such estimates can be derived from inspection of upper bounds on speed of decrease of errors in approximation of multivariable functions by networks with increasing number of units. Jones (1992) proved an upper bound on rates of approximation of functions from certain convex sets and suggested applying the bound to functions with representations as infinite networks with trigonometric perceptrons. Barron (1993) refined Jones' result and used it to derive an estimate of model complexity for sigmoidal perceptron networks based on an integral representation in the form of a weighted Fourier transform. Girosi and Anzellotti

* Tel.: +420 723365028.
E-mail address: vera@cs.cas.cz.

0893-6080/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2012.05.002

(1993) combined the estimates by Jones and Barron with convolutions with suitable kernels and Girosi (1995) proposed an alternative method for estimation of rates of neural network approximation based on a result from machine learning. Kainen, Kůrková, and Vogt (2007) and Kůrková, Kainen, and Kreinovich (1997) applied the estimates of rates of approximation by Jones and Barron to representations of sufficiently differentiable functions in the form of networks with infinitely many Heaviside perceptrons.

In this paper, we present a unifying framework for estimation of model complexity of neural networks based on representations of multivariable functions as images of integral transforms with kernels corresponding to network units. We combine upper bounds on rates of approximation by convex combinations of functions from "dictionaries" of computational units reformulated in terms of "variational" norms tailored to these units together with upper bounds on these norms derived using integral transforms with kernels corresponding to the units. Using a geometric characterization of variational norms, we prove that $\mathcal{L}^1$-norms of output-weight functions in representations of functions as infinite networks with units from a variety of dictionaries are crucial factors in estimates of growth of model complexity with increasing accuracy requirements. Various special cases of the latter estimate have been proven earlier using a variety of proof techniques requiring more complicated tools (such as a probabilistic argument Barron, 1993, an approximation of integrals by Riemann sums Kůrková et al., 1997, and interpretation of infinite networks as Bochner integrals Girosi & Anzellotti, 1993; Kainen & Kůrková, 2009). The results here are proven under minimal assumptions and thus they hold for quite general dictionaries of hidden units and ambient function spaces and so they allow applications to classes of networks to which previous results were not applicable. Our proof technique takes advantage of a version of the Hahn–Banach theorem. A preliminary version of some of the results appeared in conference proceedings (Kůrková, 2009).

The paper is organized as follows. In Section 2, basic concepts and notations concerning computational units and integral operators defined by such units are introduced. In Section 3, variational norms induced by computational units are defined and estimates of rates of approximation are reformulated in terms of these norms. In Section 4, the geometric characterization of the variational norm is proven and employed to derive its properties. In Section 5, a short argument proving the relationship between the variational norm of a function representable as an infinite network and the $\mathcal{L}^1$-norm of the output-weight function of this network is given. In Section 6, the results are applied to integral representations of smooth functions in the form of infinite networks with Heaviside perceptrons. Section 7 is a brief discussion. For the readers' convenience, some mathematical concepts and results used in the paper are recalled in the Appendix.

2. Integral transforms induced by computational units

Computational units (such as perceptrons, radial or kernel units) compute functions of two vector variables representing *inputs* and *parameters* (e.g., weights, biases, centroids). So formally computational units can be described as mappings

$$\phi : \Omega \times A \to \mathbb{R},$$

where $\Omega \subseteq \mathbb{R}^d$ is a set of variables and $A \subseteq \mathbb{R}^s$ is a set of parameters. We denote by

$$G_\phi = G_\phi(A) = G_\phi(\Omega, A) := \{\phi(\cdot, a) \mid a \in A\}$$

the parameterized set of functions on Ω determined by ϕ. The set G_ϕ is sometimes called a *dictionary*. We use the shorter notation G_ϕ or $G_\phi(A)$ when the sets Ω or A are clear from the context.

For example, a *perceptron with an activation function* $\sigma : \mathbb{R} \to \mathbb{R}$ can be described by a mapping $\phi_\sigma : \mathbb{R}^d \times \mathbb{R}^{d+1}$ defined for $(v, b) \in \mathbb{R}^d \times \mathbb{R} = \mathbb{R}^{d+1}$ as

$$\phi_\sigma(x, (v, b)) := \sigma(v \cdot x + b). \tag{1}$$

An important type of activation function is the *Heaviside function* $\vartheta : \mathbb{R} \to \mathbb{R}$ defined as $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geq 0$. An *RBF unit* with an even function $\beta : \mathbb{R} \to \mathbb{R}$ can be described by a mapping $\phi_\beta : \mathbb{R}^d \times (\mathbb{R}^d \times \mathbb{R}_+) \to \mathbb{R}$ defined as

$$\phi_\beta(x, (v, b)) := \beta(b\|x - v\|) \tag{2}$$

and a *kernel unit* with a symmetric positive semidefinite kernel $K : \Omega \times \Omega \to \mathbb{R}$ computes functions of the form

$$\phi(x, a) = K(x, a).$$

A widely used network architecture is a *one-hidden-layer network with a single linear output*. Such a network with n units computing ϕ can compute input–output functions from the set

$$\operatorname{span}_n G_\phi(A) := \left\{ \sum_{i=1}^n w_i \phi(\cdot, a_i) \,\middle|\, w_i \in \mathbb{R},\ a_i \in A \right\}.$$

A network unit computing a function $\phi : \Omega \times A \to \mathbb{R}$ can also induce an integral operator. The operator depends on a measure μ on A. For a function $w : A \to \mathbb{R}$ in a suitable space of functions on A such that for all $x \in \Omega$ the integral (3) exists, we denote by $T_{\phi,\mu}$ the operator defined as

$$T_{\phi,\mu}(w)(x) := \int_A w(a)\phi(x, a)d\mu(a). \tag{3}$$

When μ is the Lebesgue measure, we write for short T_ϕ and da. Metaphorically, the integral on the right-hand side of the Eq. (3) can be interpreted as a *one-hidden-layer neural network with infinitely many units* computing functions from a dictionary $G_\phi = \{\phi(\cdot, a) \mid a \in A\}$. So the operator $T_{\phi,\mu}$ transforms output-weight functions $w : A \to \mathbb{R}$ of infinite networks with units from the dictionary G_ϕ to input–output functions $T_{\phi,\mu}(w) : \Omega \to \mathbb{R}$.

Recall that when $\phi \in \mathcal{L}^p(\Omega \times A, \rho \times \mu)$, then $T_{\phi,\mu} : \mathcal{L}^q(A, \mu) \to \mathcal{L}^p(\Omega, \rho)$, where $\frac{1}{p} + \frac{1}{q} = 1$, is a continuous operator (Friedman, 1982, p. 138). When in addition Ω and A are compact subsets of $\mathbb{R}^d$ and ρ and μ are Lebesgue measures, then $T_\phi : \mathcal{L}^q(A) \to \mathcal{L}^p(\Omega)$ is compact (Friedman, 1982, p. 188).

Note that classes of functions which can be expressed as integrals in the form (3) representing infinite neural networks with typical computational units such as perceptrons or RBF are quite large. For example, all sufficiently smooth compactly supported functions or functions decreasing sufficiently rapidly at infinity (in particular, the Gaussian function) can be expressed as networks with infinitely many Heaviside perceptrons (Ito, 1991; Kainen et al., 2007; Kainen, Kůrková, & Vogt, 2010; Kůrková et al., 1997). Functions from various Sobolev spaces can be represented as infinite networks with Gaussian RBF units (Kainen, Kůrková, & Sanguineti, 2009). Other large classes of functions can be obtained as limits of sequences of input–output functions of infinite networks with quite general radial or kernel functions (Park & Sandberg, 1991, 1993).

3. Norms induced by computational units

An importance of the role of integral transforms induced by computational units in investigation of model complexity of neural networks follows from the role of such transforms in estimation of norms induced by computational units. In this section, we introduce these norms and survey some estimates of rates of approximation in which these norms play an important role.

For G a bounded nonempty subset of a normed linear space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$, the norm G-*variation*, denoted $\|\cdot\|_G$, is defined for all $f \in \mathcal{X}$ as

$$\|f\|_{G,\mathcal{X}} := \inf\{c > 0 \mid f/c \in \mathrm{cl}_{\mathcal{X}} \operatorname{conv}(G \cup -G)\},$$

where the closure $\mathrm{cl}_{\mathcal{X}}$ is taken with respect to the topology generated by the norm $\|\cdot\|_{\mathcal{X}}$ and conv denotes the *convex hull*. So G-variation depends on the ambient space norm, but when it is clear from the context, we write merely $\|f\|_G$ instead of $\|f\|_{G,\mathcal{X}}$. Note that G-variation is the Minkowski functional of the closed convex symmetric hull of G. It is easy to check that G-variation is a norm on the subspace of $\mathcal{X}$ formed by those f for which $\|f\|_G$ is finite.

The concept of variation with respect to a set of functions was introduced by Barron (1992) for sets of characteristic functions. In particular, *variation with respect to half-spaces* has been used in neurocomputing as it is induced by the set of functions computable by Heaviside perceptrons (see Section 6 for more details). Barron's concept was generalized in Kůrková (1997, 2003) to a variation with respect to an arbitrary bounded set of functions and applied to various dictionaries of computational units. Typically, such dictionaries are neither balanced nor convex.

We recall some upper bounds on rates of approximation by sets of the form $\mathrm{span}_n G$ in various ambient function spaces. Typically, such bounds are of the form

$$\|f - \mathrm{span}_n G\|_{\mathcal{X}} \leq n^{-1/s}\xi(d),$$

where ξ is a function of the number of variables d which often involves G-variation $\|f\|_G$ of the function f to be approximated. Inspection of these bounds shows that a network with

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

units can approximate f within ε. Thus it is important to estimate G-variation for wide classes of multivariable functions.

The following theorem from Kůrková (2003) and Kůrková and Sanguineti (2005) is a reformulation of results by Barron (1993), Darken, Donahue, Gurvits, and Sontag (1993), Jones (1992) and Pisier (1981) in terms of G-variation.

Theorem 3.1. *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a normed linear space, G its bounded nonempty subset, $s_G = \sup_{g\in G}\|g\|_{\mathcal{X}}$, $f \in \mathcal{X}$, and n be a positive integer. Then*

(i) *for $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ a Hilbert space,*

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

(ii) *for $(\mathcal{X}, \|\cdot\|_{\mathcal{X}}) = (\mathcal{L}^p(\Omega), \|\cdot\|_{\mathcal{L}^p})$, $p \in (1, \infty)$, and $\Omega \subseteq \mathbb{R}^d$ Lebesgue measurable,*

$$\|f - \mathrm{span}_n G\|_{\mathcal{L}^p} \leq \frac{2^{1+1/r}\, s_G\|f\|_G}{n^{1/s}},$$

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

The estimates by Barron (1993), Darken et al. (1993), Jones (1992) and Pisier (1981) were formulated for approximation of functions f from $\mathrm{cl}_{\mathcal{X}} \operatorname{conv} G$ by elements of

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

As for all $c > 0$,

$$\|f - \mathrm{span}_n G\|_{\mathcal{X}} \leq \|f - \mathrm{conv}_n(c(G \cup -G))\|_{\mathcal{X}},$$

the concept of G-variation enables extension of upper bounds from Theorem 3.1 to all functions in $\mathcal{X}$ with finite G-variations. Note that better upper bounds on approximation by $\mathrm{conv}_n G$ in Hilbert spaces were derived in Kůrková and Sanguineti (2008) and Lavretsky (2002), but these results are existential only and it seems difficult to interpret them in terms of characterizations of functions to be approximated which can be estimated. More precisely, in Kůrková and Sanguineti (2008) it was shown that for every $f \in \operatorname{conv} G$ there exists $\alpha_f \in [0, 1)$ such that $\|f - \mathrm{conv}_n G\|_{\mathcal{X}}^2 \leq \alpha_f^{n-1}(s_G^2 - \|f\|_{\mathcal{X}}^2)$, but estimates of such α_f are not known.

Theorem 3.1 gives estimates of approximation errors in $\mathcal{L}^p$-spaces with $p \in (1, \infty)$. It was shown in Donahue, Gurvits, Darken, and Sontag (1997) that the proof method used in Barron (1993), Donahue et al. (1997) and Jones (1992) based on construction of incremental approximants cannot be extended to approximation in $\mathcal{L}^1$ and $\mathcal{L}^\infty$-spaces. However, for special cases of sets G, e.g., sets of characteristic functions with finite coVC-dimension (see Appendix for the definition), some estimates in the supremum norm were obtained by probabilistic proof techniques. The following theorem is a reformulation of an upper bound from Gurvits and Koiran (1997, Theorem 3) in terms of G-variation. By $(\mathcal{F}(\Omega), \|\cdot\|_{\sup})$ is denoted the space of all bounded functions on Ω with the supremum norm.

Theorem 3.2. *Let $\Omega \subseteq \mathbb{R}^d$, G be a subset of the set of characteristic functions on Ω such that the coVC-dimension $h^*(G)$ is finite, then for all $f \in \mathcal{F}(\Omega)$,*

$$\|f - \mathrm{span}_n G\|_{\sup} \leq 6\sqrt{3}\,\|f\|_{G,\sup}\, h^*(G)^{1/2}\ (\log n)^{1/2}\ n^{-1/2}.$$

4. Properties of variational norm

To apply results from Section 3 to neurocomputing we need estimates of variational norms tailored to various computational units. As large classes of functions can be represented as infinite networks with perceptrons and Gaussian radial units (Girosi, 1995; Ito, 1991; Kainen et al., 2009; Kůrková et al., 1997), estimates of variational norms of functions from these classes can lead to useful insights about network complexity. In this section, we derive properties of the variational norm which will be used in the next section to estimate G_ϕ-variation for functions representable as integrals in the form of infinite networks with units computing ϕ.

First, we prove a characterization of the variational norm in terms of bounded linear functionals using a version of the Hahn–Banach theorem. Although in general normed linear spaces this characterization (Theorem 4.1) is rather abstract, in Hilbert spaces it has an interpretation in terms of angles between functions and thus we call it "geometric".

The main advantage of the next characterization of G-variation is that it leads to a simple proof of its estimate for functions representable as infinite networks in the form (3).

By $\mathcal{X}^*$ is denoted the *dual* of $\mathcal{X}$ (the space of all bounded linear functionals on $\mathcal{X}$) and $S_G = \{l \in \mathcal{X}^* \mid (\exists g \in G)\ (l(g) \neq 0)\}$.

Theorem 4.1. *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a normed linear space, G be its nonempty bounded subset and $f \in \mathcal{X}$ be such that $\|f\|_G < \infty$. Then*

$$\|f\|_G = \sup_{l\in S_G} \frac{|l(f)|}{\sup_{g\in G}|l(g)|}.$$

Proof. First, we show that for all $c > 0$ and all $f \in \mathcal{X}$

$$f/c \in \mathrm{cl}_{\mathcal{X}} \operatorname{conv}(G \cup -G)$$
$$\Longrightarrow \left((\forall l \in \mathcal{X}^*)(\,|l(f)| \leq c \sup_{g\in G}|l(g)|\,)\right). \tag{4}$$

If $f/c \in \mathrm{cl}_{\mathcal{X}} \operatorname{conv}(G \cup -G)$ then there exists a sequence $\{f_k\}$ such that

$\lim_{k\to\infty} \|f/c - f_k\|_{\mathcal{X}} = 0$ and all f_k can be represented as $f_k = \sum_{i=1}^{m_k} w_{k,i} g_{k,i}$, where $\sum_{i=1}^{m_k} |w_{k,i}| = 1$ and all $g_{k,i} \in G$. Then for all $l \in \mathcal{X}^*$, $l(f_k) = \sum_{i=1}^{m_k} w_{k,i} l(g_{k,i})$ and so $|l(f_k)| \leq \sup_{g\in G} |l(g)|$. Since l is continuous, also $|l(f/c)| \leq \sup_{g\in G} |l(g)|$ and thus $|l(f)| \leq c \sup_{g\in G} |l(g)|$.

Now, we prove that for all $c > 0$ and all $f \in \mathcal{X}$ with $\|f\|_G < \infty$ the following implication holds:

$$\left((\forall\, l \in S_G)(\, |l(f)| \leq c \sup_{g\in G} |l(g)|\,)\right)$$
$$\Longrightarrow f/c \in \mathrm{cl}_{\mathcal{X}} \operatorname{conv}(G \cup -G). \tag{5}$$

Assume by contradiction that $f/c \notin \mathrm{cl}_{\mathcal{X}} \operatorname{conv}(G \cup -G)$. Then by Mazur's theorem (Yoshida, 1965, p. 108) (see Theorem A.1 in the Appendix), there exists $l \in \mathcal{X}^*$ such that $l(f/c) > 1$ and for all $h \in \mathrm{cl}_{\mathcal{X}} \operatorname{conv}(G \cup -G)$, $|l(h)| \leq 1$. Thus in particular for all $g \in G$, $|l(g)| \leq 1$. Hence $|l(f)| = l(f) > c \geq c \sup_{g\in G} |l(g)|$. It remains to show that $l \in S_G$. As $\|f\|_G$ is finite, there exists some $b > 0$ such that $f/b \in \mathrm{cl}_{\mathcal{X}} \operatorname{conv}(G \cup -G)$, and hence by (4), $|l(f)| \leq b \sup_{g\in G} |l(g)|$. If l were in $\mathcal{X}^* \setminus S_G$, this would imply $|l(f)| \leq b \sup_{g\in G} |l(g)| = 0$. But $l(f) = 0$ is in contradiction with $l(f) > c > 0$.

It follows from the implications (4) and (5) that

$$f/c \in \mathrm{cl}_{\mathcal{X}} \operatorname{conv}(G \cup -G) \Longleftrightarrow c \geq \sup_{l\in S_G} \frac{|l(f)|}{\sup_{g\in G} |l(g)|}. \tag{6}$$

Thus $\|f\|_G = \inf\{c > 0 \mid f/c \in \mathrm{cl}_{\mathcal{X}} \operatorname{conv}(G \cup -G)\} = \sup_{l\in S_G} \frac{|l(f)|}{\sup_{g\in G} |l(g)|}$. □

Theorem 4.1 is an extension of a characterization of the variational norm in Hilbert spaces proven in Kůrková, Savický, and Hlaváčková (1998), which was used there and in Kůrková (2008) to prove existence of functions with variations growing with the input dimension d exponentially.

When $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ is a Hilbert space, then all bounded linear functionals are inner products (Friedman, 1982, p. 206). Denoting by $G^\perp$ the *orthogonal complement* of a subset G of $\mathcal{X}$, i.e., $G^\perp = \{h \in \mathcal{X} \mid (\forall g \in G)\,(h \cdot g = 0)\}$, we get by Theorem 4.1 for all $f \in \mathcal{X}$

$$\|f\|_G = \sup_{h \in \mathcal{X} \setminus G^\perp} \frac{|f \cdot h|}{\sup_{g\in G} |g \cdot h|}.$$

In particular, for all $f \in \mathcal{X} \setminus G^\perp$

$$\|f\|_G \geq \frac{\|f\|_{\mathcal{X}}^2}{\sup_{g\in G} |f \cdot g|}. \tag{7}$$

The inequality (7) shows that the closer a function f is to orthogonality to all elements of the set G, the larger the value of G-variation f has.

To illustrate Theorem 4.1, consider the finite dimensional space $\mathbb{R}^m$ with the Euclidean norm denoted $\|\cdot\|_2$. Let $G = \{e_1, \ldots, e_m\}$ be an orthonormal basis of $\mathbb{R}^m$. It is easy to see that for all $f = \sum_{i=1}^m w_i e_i$, $\|f\|_G = \|f\|_1 = \sum_{i=1}^m |w_i|$. Let $u = (1, \ldots, 1)$. Then by Theorem 4.1 for all $f \in \mathbb{R}^d$,

$$\|f\|_G \geq \frac{|f \cdot u|}{\sup_{i=1,\ldots,m} |e_i \cdot u|} = \frac{\sum_{i=1}^m |w_i|}{1}.$$

As in this case $\|f\|_G = \|f\|_1$, the supremum from Theorem 4.1 is the maximum. Moreover for all $f \in \mathbb{R}^m$, the maximum is achieved for the same linear functional, which is the inner product with $u = (1, \ldots, 1)$.

It was shown in Kůrková and Sanguineti (2002) that for an infinite orthonormal basis $G = \{e_i\}$ of $(\ell_2, \|\cdot\|_2)$, $\|\cdot\|_G = \|\cdot\|_1$. In this case, for $f \in \ell_2 \cap \ell_1$ with a representation $f = \sum_{i=1}^\infty w_i e_i$, we have

$$\|f\|_G = \|f\|_1 = \sup_k \frac{|f \cdot h_k|}{\sup_i |e_i \cdot h_k|},$$

where $h_k = \sum_{i=1}^k \operatorname{sign}(w_i) e_i$, with $\operatorname{sign}(x) = 1$ for $x > 0$, $\operatorname{sign}(x) = -1$ for $x < 0$, and $\operatorname{sign}(0) = 0$.

Theorem 4.1 implies that in the definition of G-variation, infimum can be replaced with minimum.

Proposition 4.2. *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a normed linear space, G its nonempty bounded subset, and $f \in \mathcal{X}$ be such that $\|f\|_G < \infty$. Then*

$$\|f\|_G = \min\{c > 0 \mid f/c \in \mathrm{cl}_{\mathcal{X}} \operatorname{conv}(G \cup -G)\}.$$

Proof. By (4), for all $l \in \mathcal{X}^*$ and all $c > 0$ such that $f/c \in \mathrm{cl}_{\mathcal{X}} \operatorname{conv}(G \cup -G)$, $|l(f)| \leq c \sup_{g\in G} |l(g)|$. Hence also for $b = \|f\|_G = \inf\{c > 0 \mid f/c \in \mathrm{cl}_{\mathcal{X}} \operatorname{conv}(G \cup -G)\}$, $|l(f)| \leq b \sup_{g\in G} |l(g)|$. Thus by (5) also $f/b \in \mathrm{cl}_{\mathcal{X}} \operatorname{conv}(G \cup -G)$. □

Another useful property of the variational norm following from Theorem 4.1 is a bound on variation of the limit of a sequence of functions. Although this property can also be derived directly from the definition of variation (see Kainen & Kůrková, 2009; Kůrková et al., 1997), application of Theorem 4.1 gives a shorter proof.

Proposition 4.3. *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a normed linear space, G its nonempty bounded subset, $f \in \mathcal{X}$, and $\{f_k\}_{k=1}^\infty \subset \mathcal{X}$ be such that $\lim_{k\to\infty} \|f_k - f\|_{\mathcal{X}} = 0$, $b_k = \|f_k\|_G < \infty$ for all k, and $b = \lim_{k\to\infty} b_k < \infty$. Then $\|f\|_G \leq b$.*

Proof. By Theorem 4.1,

$$\lim_{k\to\infty} \|f_k\|_G = \lim_{k\to\infty} \sup_{l\in S_G} \frac{|l(f_k)|}{\sup_{g\in G} |l(g)|} \geq \sup_{l\in S_G} \lim_{k\to\infty} \frac{|l(f_k)|}{\sup_{g\in G} |l(g)|}.$$

As all $l \in \mathcal{X}^*$ are continuous, $\lim_{k\to\infty} \frac{|l(f_k)|}{\sup_{g\in G} |l(g)|} = \frac{|l(f)|}{\sup_{g\in G} |l(g)|}$. Because f/b is a limit of a sequence of elements of $\operatorname{conv}(G \cup -G)$, $\|f\|_G$ is finite. Thus $\|f\|_G$ satisfies (5) and so we obtain $\lim_{k\to\infty} \|f_k\|_G \geq \|f\|_G$. □

Proposition 4.3 implies that balls in G-variation are closed in the ambient space norm, but it does not imply that the linear subspace $\mathcal{X}_G = \{f \in \mathcal{X} \mid \|f\|_G < \infty\}$ of functions with finite values of G-variation is closed as a subspace of $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$. This can be shown by the following example. Consider the space $(\ell_2, \|\cdot\|_2)$ and let $G = \{e_i\}$ be its orthonormal basis. It was shown in Kůrková and Sanguineti (2002) that $\|\cdot\|_G = \|\cdot\|_1$, i.e., for $f = \sum_{i=1}^\infty w_i e_i$ with $\|f\|_1 = \sum_{i=1}^\infty |w_i| < \infty$, $\|f\|_G = \|f\|_1$. So for any $f = \sum_{i=1}^\infty w_i e_i \in \ell_2$ which is not in ℓ_1, we have $\lim_{k\to\infty} \|f - f_k\|_2 = 0$, where all $f_k = \sum_{i=1}^k w_i e_i$ have finite G-variations equal to $\sum_{i=1}^k |w_i|$.

5. Variation of functions in images of transforms induced by computational units

In this section, we use the characterization of variational norm from Theorem 4.1 to derive an upper bound on variation of functions representable as integrals in the form of networks with infinitely many units.

First, consider the case of a finite set A of parameters. Each ordering of A as $A = \{a_1, \ldots, a_m\}$ determines a linear operator $T_\phi : \mathbb{R}^m \to \mathcal{X}$ defined for all $w = (w_1, \ldots, w_m) \in \mathbb{R}^m$ as

$$T_\phi(w)(x) = \sum_{i=1}^m w_i \phi(x, a_i).$$

It follows easily from the definition of variational norm that for each f which can be represented as $f = T_\phi(w)$ for some $w \in \mathbb{R}^m$,

$$\|f\|_{G_\phi(A)} = \min\left\{ \|w\|_1 \,\middle|\, f = \sum_{i=1}^{m} w_i \phi(\cdot, a_i) \right\}. \tag{8}$$

Note that for some functions which can be exactly represented as input–output functions of finite neural networks, the networks might be too large to be implementable. In such cases, Theorem 3.1 and the upper bound (8) can be used to obtain estimates of rates of approximation of f by input–output functions of smaller networks. Note that the value of ℓ_1 or ℓ_2-norm of output weight vector $w = (w_1, \ldots, w_m)$ plays a role of a stabilizer to be minimized in output-weight regularization (Fine, 1999).

When the set A of parameters is infinite, analogy with (8) suggests that for f representable as

$$f(x) = T_{\phi,\mu}(w) = \int_A w(a)\phi(x, a)d\mu(a),$$

the estimate

$$\|f\|_{G_{\phi,\mu}(A)} \leq \|w\|_{\mathcal{L}^1(A,\mu)} \tag{9}$$

might hold. The inequality (9) can only be considered when quantities on both its sides are well defined, i.e., when

(i) $G_\phi(A)$ is a bounded subset of $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ and
(ii) $w \in \mathcal{L}^1(A, \mu)$.

Our main result (Theorem 5.1) shows that in a wide class of function spaces, the assumptions (i) and (ii) are sufficient to guarantee the relationship (9) between $G_\phi(A)$-variation and $\mathcal{L}^1$-norm. We show that this relationship follows easily from the geometric characterization of G-variation given in Theorem 4.1 provided that in the ambient function space a certain commutativity property of bounded linear functionals holds. A linear space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ of functions on $\Omega \subseteq \mathbb{R}^d$ has a *commutativity property of linear functionals with kernel operators* if for every integral operator $T_\phi : (\mathcal{L}^1_\mu(A), \|\cdot\|_{\mathcal{L}^1}) \to (\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ with a kernel $\phi : \Omega \times A \to \mathbb{R}$ such that $G_\phi = \{\phi(\cdot, a) \mid a \in A\}$ is a bounded subset of $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$, and every linear functional $l \in \mathcal{X}^*$ and every $g \in \mathcal{L}^1(A)$

$$l(T_\phi(f)) = \int_A f(a)\, l(\phi(\cdot, a))\, d\mu(a).$$

This property holds, for example, in spaces $(\mathcal{L}^p(\Omega, \rho), \|\cdot\|_{\mathcal{L}^p})$ with $p \in [1, \infty)$, $(\mathcal{C}_c(\Omega), \|\cdot\|_{\sup})$, and $(\mathcal{C}_0(\mathbb{R}^d), \|\cdot\|_{\sup})$ as it is shown in Theorem 5.2.

The next theorem on the relationship (9) between G_ϕ-variation of an input–output function of an infinite network and the $\mathcal{L}^1$-norm of its output-weight function has a short proof based on the geometric characterization of the variational norm from Theorem 4.1.

Theorem 5.1. *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a space of functions on $\Omega \subseteq \mathbb{R}^d$ satisfying the commutativity property of linear functionals with kernel operators, μ be a σ-finite measure on $A \subseteq \mathbb{R}^s$, $w \in \mathcal{L}^1(A, \mu)$, $\phi : \Omega \times A \to \mathbb{R}$ be such that $G_\phi(A) = \{\phi(\cdot, a) \mid a \in A\}$ is a bounded subset of $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$, and $f \in \mathcal{X}$ be such that for all $x \in \Omega$, $f(x) = \int_A w(a)\phi(x, a)d\mu(a)$. Then*

$$\|f\|_{G_\phi(A)} \leq \|w\|_{\mathcal{L}^1(A,\mu)}.$$

Proof. By the commutativity property, for all $l \in \mathcal{X}^*$, $l(f) = \int_A w(a) l(\phi(\cdot, a))d\mu(a)$. Thus $|l(f)| \leq \sup_{a\in A} |l(\phi(\cdot, a))| \int_A |w(a)| d\mu(a) = \sup_{a\in A} |l(\phi(\cdot, a))|\ \|w\|_{\mathcal{L}^1(A,\mu)}$. So by Theorem 4.1,

$$\|f\|_{G_\phi(A)} = \sup_{l \in \mathcal{X}^* \setminus G_\phi(A)^\perp} \frac{|l(f)|}{\sup_{a\in A} |l(\phi(\cdot, a))|} \leq \|w\|_{\mathcal{L}^1(A,\mu)}. \quad \square$$

The next theorem describes some function spaces with the commutativity property. For $\Omega \subseteq \mathbb{R}^d$, by $(\mathcal{C}_c(\Omega), \|\cdot\|_{\sup})$ is denoted the *space of all continuous compactly supported functions on Ω with the supremum norm* and by $(\mathcal{C}_0(\mathbb{R}^d), \|\cdot\|_{\sup})$ the *space of all continuous functions on $\mathbb{R}^d$ vanishing at infinity* (i.e., functions f for which $\lim_{\|x\|\to\infty} f(x) = 0$).

Theorem 5.2. *Each of the following spaces satisfies the commutativity property of linear functionals with kernel operators:*

(i) $(\mathcal{L}^p(\Omega, \rho), \|\cdot\|_{\mathcal{L}^p})$ *with $p \in [1, \infty)$, and ρ a measure on $\Omega \subseteq \mathbb{R}^d$;*
(ii) $(\mathcal{C}_c(\Omega), \|\cdot\|_{\sup})$ *with Ω a locally compact subset of $\mathbb{R}^d$;*
(iii) $(\mathcal{C}_0(\Omega), \|\cdot\|_{\sup})$ *with $\Omega = \mathbb{R}^d$.*

Proof. First, we prove the statement for case (i). By the properties of the duals of $\mathcal{L}^p$-spaces with $p \in [1, \infty)$ (Friedman, 1982, pp. 176, 180), for every $l \in \mathcal{X}^*$ there exists $h \in \mathcal{L}^q(\Omega, \rho)$ (where for $p > 1$, q satisfies $1/q + 1/p = 1$, while for $p = 1$, $q = \infty$), such that for all $f \in \mathcal{L}^p(\Omega, \rho)$,

$$l(f) = \int_\Omega f(x)h(x)d\rho(x).$$

By Hölder's inequality (Friedman, 1982, p. 96) for all $a \in A$, $\phi(\cdot, a)\, h \in \mathcal{L}^1(\Omega, \rho)$ and

$$\|\phi(\cdot, a)\, h\|_{\mathcal{L}^1} \leq \|\phi(\cdot, a)\|_{\mathcal{L}^p} \|h\|_{\mathcal{L}^q}.$$

Thus for all $a \in A$, $\int_\Omega |\phi(x, a)h(x)|d\rho(x) \leq \|\phi(\cdot, a)\|_{\mathcal{L}^p} \|h\|_{\mathcal{L}^q}$.

By the assumption $G_\phi(A)$ is bounded and so $\sup_{a\in A} \|\phi(\cdot, a)\|_{\mathcal{L}^p} = s_\phi$ is finite. Thus also

$$\int_\Omega \int_A |w(y)\phi(x, y)h(x)|d\mu(y)d\rho(x) \leq s_\phi \|w\|_{\mathcal{L}^1}$$

is finite. So we can use Fubini's theorem (Friedman, 1982, p. 86) to obtain

$$\begin{aligned} l(f) &= \int_\Omega \left(\int_A w(a)\phi(x, a)d\mu(a) \right) h(x)d\rho(x) \\ &= \int_A w(a) \left(\int_\Omega \phi(x, a)h(x)d\rho(x) \right) d\mu(a) \\ &= \int_A w(a) l(\phi(\cdot, a))d\mu(a). \end{aligned}$$

The proof of cases (ii) and (iii) is analogous to case (i). The only difference is in the characterization of bounded linear functionals. By the Riesz representation theorem (Rudin, 1974), for every $l \in \mathcal{X}^*$, there exists a signed measure ν on Ω such that for all $f \in \mathcal{C}_c(\Omega)$ or $f \in \mathcal{C}_0(\Omega)$, $l(f) = \int_\Omega f(x)d\nu(x)$ and $|\nu|(\Omega) = \|l\|_{\mathcal{X}^*}$, where $|\nu|$ denotes the total variation of ν. Thus for all $a \in A$, $\int_\Omega |\phi(x, a)|d\nu(x) \leq \|\phi(\cdot, a)\|_{\sup} |\nu|(\Omega)$. As $\sup_{a\in A} \|\phi(\cdot, a)\|_{\sup} = s_\phi$ is finite, also

$$\begin{aligned} &\int_A |w(a)| \int_\Omega |\phi(x, a)|d\nu(x)d\mu(a) \\ &\quad \leq s_\phi \|l\|_{\mathcal{X}^*} \int_A |w(a)|d\mu(a) \leq s_\phi \|l\|_{\mathcal{X}^*} \|w\|_{\mathcal{L}^1} \end{aligned}$$

is finite. Thus we can use Fubini's theorem to obtain

$$\begin{aligned} l(f) &= \int_\Omega \left(\int_A w(a)\phi(x, a)d\mu(a) \right) d\nu(x) \\ &= \int_A w(a) \left(\int_\Omega \phi(x, a)d\nu(x) \right) d\mu(a) \\ &= \int_A w(a) l(\phi(\cdot, a))d\mu(a). \quad \square \end{aligned}$$

Combining Theorems 4.1 and 5.2 we get the next corollary.

Corollary 5.3. *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be one of the following spaces:*

(i) $(\mathcal{L}^p(\Omega, \rho), \|\cdot\|_{\mathcal{L}^p})$ *with* $q \in [1, \infty)$, *and* ρ *a* σ*-finite measure;*
(ii) $(\mathcal{C}_c(\Omega), \|\cdot\|_{\sup})$ *with* Ω *a locally compact subset of* $\mathbb{R}^d$;
(iii) $(\mathcal{C}_0(\Omega), \|\cdot\|_{\sup})$ *with* $\Omega = \mathbb{R}^d$.

Let μ *be a* σ*-finite measure on* $A \subseteq \mathbb{R}^s$, $w \in \mathcal{L}^1(A, \mu)$, $\phi : \Omega \times A \to \mathbb{R}$ *be such that* $G_\phi(A) = \{\phi(\cdot, a) \mid a \in A\}$ *is a bounded subset of* $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ *and* $f \in \mathcal{X}$ *be such that for all* $x \in \Omega$, $f(x) = \int_A w(a)\phi(x, a)d\mu(a)$. *Then*

$$\|f\|_{G_\phi(A)} \leq \|w\|_{\mathcal{L}^1(A,\mu)}.$$

Applying the upper bound on G_ϕ-variation from Corollary 5.3 to estimates of rates of approximation given in Theorem 3.1 we get the following upper bounds on rates of approximation from the dictionary G_ϕ.

Corollary 5.4. *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a space of functions on $\Omega \subseteq \mathbb{R}^d$, $A \subseteq \mathbb{R}^s$, μ be a measure on A, $\phi : \Omega \times A \to \mathbb{R}$ be such that $G_\phi(A) = \{\phi(\cdot, a) \mid a \in A\}$ is a bounded subset of $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$, and $s_\phi = \sup_{a\in A} \|\phi(\cdot, a)\|_{\mathcal{X}}$. Let $f \in (\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be such that for some $w \in \mathcal{L}^1(A, \mu)$, $f(x) = \int_A w(a)\phi(x, a)d\mu(a)$. Then for all n*

(i) *for* $(\mathcal{X}, \|\cdot\|_{\mathcal{X}}) = (\mathcal{L}^2(\Omega, \rho), \|\cdot\|_{\mathcal{L}^2})$,

$$\|f - \operatorname{span}_n G_\phi(A)\|_{\mathcal{X}}^2 \leq \frac{s_\phi^2 \|w\|_{\mathcal{L}^1(A,\mu)}^2 - \|f\|_{\mathcal{X}}^2}{n};$$

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

$$\|f - \operatorname{span}_n G_\phi(A)\|_{\mathcal{L}^p} \leq \frac{2^{1+1/r} s_\phi \|w\|_{\mathcal{L}^1}}{n^{1/s}},$$

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

6. Variation with respect to perceptrons

In this section, we apply our results to perceptron networks. We consider the dictionary formed by functions computable by perceptron networks with the *Heaviside activation functions* $\vartheta : \mathbb{R} \to \mathbb{R}$ defined as $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geq 1$. We denote this dictionary

$$G_{\phi_\vartheta} = G_{\phi_\vartheta}(S^{d-1} \times \mathbb{R}, \Omega) := \{\vartheta(e \cdot . + b) : \Omega \to \mathbb{R} \mid e \in S^{d-1},\ b \in \mathbb{R}\},$$

where S^{d-1} denotes the unit sphere in $\mathbb{R}^d$. Recall that G_{ϕ_ϑ}-variation has been called *variation with respect to half-spaces* (Barron, 1992) as the dictionary G_{ϕ_ϑ} consists of characteristic functions of half-spaces of $\mathbb{R}^d$. Note that for every continuous *sigmoidal function* (i.e., a non-decreasing $\sigma : \mathbb{R} \to \mathbb{R}$ with $\lim_{t\to-\infty} \sigma(t) = 0$ and $\lim_{t\to\infty} \sigma(t) = 1$)

$$\|\cdot\|_{G_{\phi_\vartheta}} = \|\cdot\|_{G_{\phi_\sigma}},$$

in $\mathcal{L}^p(\Omega)$ with $p \in (1, \infty)$ and Ω compact (Kůrková et al., 1997). So estimates of variation with respect to half-spaces apply also to G_{ϕ_σ}-variation with any continuous sigmoidal function.

We use an integral representation of a sufficiently smooth function in terms of an infinite network with Heaviside perceptrons. Such a representation was derived for all compactly supported functions from $\mathcal{C}^\infty(\mathbb{R}^d)$ (space of continuous functions with continuous derivatives of all orders) by Ito (1991) who used the Radon transform. Kůrková et al. (1997) derived the same formula for all compactly supported functions from $\mathcal{C}^d(\mathbb{R}^d)$ (space of all continuous functions on $\mathbb{R}^d$ with continuous derivatives up to the order d) by a different proof technique based on an expression of the Dirac delta function as the derivative of the Heaviside function and a representation of the d-dimensional Dirac delta function δ_d as an integral of derivatives of the one-dimensional delta function

$$\delta_d(x) = a_d \int_{S^{d-1}} \delta_1^{(d-1)}(e \cdot x)\, de,$$

where $a_d = (-1)^{\frac{d-1}{2}}/(2(2\pi)^{d-1})$. Kainen et al. (2007) extended the representation as an infinite network with Heaviside perceptrons to functions with so called *weakly controlled decay* (see the Appendix for the definition). This class contains all compactly supported functions from $\mathcal{C}^d(\mathbb{R}^d)$ and the Schwartz class $\mathcal{S}(\mathbb{R}^d)$ (all functions from $\mathcal{C}^\infty(\mathbb{R}^d)$ which are together with all their derivatives rapidly decreasing (Adams & Fournier, 2003, p. 251)). In particular, the Gaussian function belongs to the class of functions of a weakly controlled decay. The next theorem from Kainen et al. (2007) describes this representation. By $D_e^{(d)}$ is denoted the *directional derivative* of the order d in the direction of the unit d-dimensional vector e and by $H_{e,b}$ the hyperplane $\{x \in \mathbb{R}^d \mid \vartheta(e \cdot x + b) = 0\}$.

Theorem 6.1. *Let d be an odd integer and $f \in \mathcal{C}^d(\mathbb{R}^d)$ be of a weakly controlled decay, then for all $x \in \mathbb{R}^d$*

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

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

Combining this integral representation with Theorem 4.1 we obtain the next corollary.

Corollary 6.2. *Let d be an odd positive integer, $\Omega \subset \mathbb{R}^d$ has finite Lebesgue measure $\lambda(\Omega)$, $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous sigmoidal function, and $f \in \mathcal{C}^d(\mathbb{R}^d)$ be a function with a weakly controlled decay. Then for all n,*

$$\|f_{|\Omega} - \operatorname{span}_n G_{\phi_\sigma}(\Omega)\|_{\mathcal{L}^2(\Omega)} \leq \frac{\lambda(\Omega)\|w_f\|_{\mathcal{L}^1(S^{d-1}\times\mathbb{R})}}{\sqrt{n}},$$

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

Proof. By Theorem 6.1, for all $x \in \Omega$, $f(x) = \int_{S^{d-1}\times\mathbb{R}} w_f(e, b)\, \vartheta(e \cdot x + b)\, de\, db$. As $G_{\phi_\vartheta}(\omega)$ is a bounded subset of $\mathcal{L}^2(\Omega)$ with $s_{\phi_\vartheta} \leq \lambda(\Omega)$ and $w_f \in \mathcal{L}^1(S^{d-1} \times \mathbb{R})$, we can apply Theorem 4.1 to obtain $\|f\|_{G_{\phi_\theta}} \leq \|w_f\|_{\mathcal{L}^1(S^{d-1}\times\mathbb{R})}$. Then the statement follows from the equality $\|f\|_{G_{\phi_\sigma}} = \|f\|_{G_{\phi_\theta}}$ (Kůrková et al., 1997) and Corollary 5.3(i). □

The upper bound from Corollary 6.2 provides some insight into the impact of the input dimension d on rates of approximation by perceptron networks. The factor $|a(d)|$ decreases to zero exponentially fast with d increasing and thus it can compensate increase of the factor $\|\int_{H_{e,b}} (D_e^{(d)}(f))(y) dy\|_{\mathcal{L}^1(S^{d-1}\times\mathbb{R})}$, which is bounded from above by the maximum of the $\mathcal{L}^1$-norms of all iterated partial derivatives of the order d of f. Note that $\lambda(\Omega)$ depends on the shape of the d-dimensional domain $\Omega \subset \mathbb{R}^d$. When Ω is the Euclidean d-dimensional ball, then $\lambda(\Omega)$ goes to zero exponentially fast with d increasing, while when Ω is a cube with the side larger than one, $\lambda(\Omega)$ increases exponentially fast.

Here, we have stated the estimate only for d odd, for which the output-weight function w_f in the representation of a compactly supported smooth function f as an integral in the form of infinite perceptron network has a simpler form than w_f in the case of d even given in Ito (1991).

7. Discussion

We have shown that the $\mathcal{L}^1$-norm of an output-weight function in an integral representation of a smooth function as an "infinite network" is an important factor in estimates of model complexity of networks approximating such functions. This result is interesting in connection with the usefulness of output-weight regularization minimizing ℓ_1 or ℓ_2-norms of output weights (Fine, 1999).

Various special cases of Theorem 5.1 have been derived by a variety of proof techniques, but they all required some restrictions on the domain Ω and the set of parameters A (compactness in Kůrková et al., 1997), and on ϕ and w (continuity in Kainen & Kůrková, 2009), or a special choice of ϕ (a trigonometric function in Barron, 1993). Our approach uses only minimal assumptions necessary for existence of the quantities which are compared: the set G_ϕ has to be bounded so that G_ϕ-variation can be defined and the output weight w has to be in $\mathcal{L}^1(A, \mu)$ so that its $\mathcal{L}_1$-norm is finite.

The essential part of our proof of Theorem 5.1 is the characterization of G-variation in terms of linear functionals given in Theorem 4.1. This characterization is based on the Mazur theorem (Theorem A.1 in the Appendix) which is a version of the Hahn–Banach theorem. Thus we avoided the technicalities of Bochner integration which were used in Kainen and Kůrková (2009). The argument there used dominated convergence to prove the existence of the Bochner integral. It also needed the Fubini theorem which we also used in the proof of Theorem 5.2. For the Bochner integral applied to neural networks see also Kainen and Vogt (in press).

Acknowledgments

The author thanks the reviewers for their helpful comments. This work was partially supported by the Institutional Research Plan AV0Z10300504, RVO 67985807, and GAČR grant P202/11/1368.

Appendix

For the reader's convenience we include several concepts and tools used in the paper.

By $\chi_S : \Omega \to \{0, 1\}$ is denoted the *characteristic function* of $S \subseteq \Omega$, i.e., $\chi_S(x) = 1$ if $x \in S$, otherwise $\chi_S(x) = 0$. Let $\mathcal{F}$ be any family of characteristic functions of subsets of Ω and $\mathcal{S}_\mathcal{F} = \{S \subseteq \Omega \mid \chi_S \in \mathcal{F}\}$ be the family of the corresponding subsets of Ω. Then a subset A of Ω is said to be *shattered* by $\mathcal{F}$ if $\{S \cap A \mid S \in \mathcal{S}_\mathcal{F}\}$ is the whole power set of A. The *VC-dimension* of $\mathcal{F}$ is the largest cardinality of any subset A which is shattered by $\mathcal{F}$.

The *coVC-dimension* of $\mathcal{F}$ is the VC-dimension of the set $\mathcal{F}' := \{ev_x \mid x \in \Omega\}$, where the *evaluation* $ev_x : \mathcal{F} \to \{0, 1\}$ is defined for every $\chi_S \in \mathcal{F}$ as $ev_x(\chi_S) = \chi_S(x)$.

The concept of VC-dimension was also extended to real-valued functions. Let $\mathcal{F}$ be a family of real-valued functions on Ω with range in the interval (a_1, a_2), where $-\infty \le a_1 < a_2 \le +\infty$. Then the *VC-dimension* of $\mathcal{F}$ is defined as the VC-dimension of the set $\mathcal{I}_\mathcal{F} = \{\vartheta(f(t) - c) \mid f \in \mathcal{F},\ c \in (a_1, a_2),\ t \in \Omega\}$ of characteristic functions, where $\vartheta : \mathbb{R} \to \mathbb{R}$ is the *Heaviside function*.

Recall that for a positive integer s and $q \in [1, \infty)$, the *Sobolev space* $\mathcal{W}^{q,s}(\mathbb{R}^d)$ is formed by all functions having t-th order partial derivatives in $\mathcal{L}^q(\mathbb{R}^d)$ for all $t \le s$ and the norm $\|\cdot\|_{\mathcal{W}^{q,s}}$ is defined as

$$\|f\|_{\mathcal{W}^{q,s}} = \left(\sum_{|\alpha| \le s} \|D^\alpha f\|^q_{\mathcal{L}^2}\right)^{1/q},$$

where α denotes a multi-index (i.e., a vector of non-negative integers), $|\alpha| = \alpha_1 + \cdots + \alpha_d$, and D^α is the corresponding partial derivative operator.

A function $f : \mathbb{R}^d \to \mathbb{R}$ is of a *weakly controlled decay* when it satisfies for all multi-indexes α with $0 \le |\alpha| = \alpha_1 + \cdots \alpha_d < d$, $\lim_{\|x\| \to \infty}(D^\alpha f)(x) = 0$ (where $D^\alpha = (\partial/\partial x_1)^{\alpha_1} \cdots (\partial/\partial x_d)^{\alpha_d}$) and for some $\varepsilon > 0$, all multi-indexes α with $|\alpha| = d$ satisfy

$$\lim_{\|x\| \to \infty} (D^\alpha f)(x)\|x\|^{d+1+\varepsilon} = 0.$$

The following theorem on separation of a function from a closed convex balanced set is from Yoshida (1965, p. 106).

Theorem A.1 (*Mazur*)**.** *Let $\mathcal{X}$ be a real locally convex linear topological space, M a closed convex balanced subset of $\mathcal{X}$. Then for any $f \notin M$ there exists a continuous linear functional l on $\mathcal{X}$ such that $l(f) > 1$ and for all $h \in M$, $|l(h)| \le 1$.*

References

Adams, R. A., & Fournier, J. J. F. (2003). *Sobolev spaces*. Amsterdam: Academic Press.
Barron, A. R. (1992). Neural net approximation. In K. Narendra (Ed.), *Proc. 7th Yale workshop on adaptive and learning systems*. Yale University Press.
Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, *39*, 930–945.
Bernstein, S. N. (1931). Sur le maximum absolu d'une somme trigonométrique. *Comptes Rendus de l'Academie des Sciences, Serie I (Mathematique)*, *193*, 433–436.
Carroll, S. M., & Dickinson, B. W. (1989). Construction of neural net using the Radon transform. In *Proceedings of IJCN. Vol. I* (pp. 607–611).
Darken, C., Donahue, M., Gurvits, L., & Sontag, E. (1993). Rate of approximation results motivated by robust neural network learning. In *Proceedings of the sixth annual ACM conference on computational learning theory* (pp. 303–309). New York: The Association for Computing Machinery.
DeVore, R. A., & Lorentz, G. G. (1993). *Constructive approximation*. Berlin: Springer-Verlag.
Donahue, M., Gurvits, L., Darken, C., & Sontag, E. (1997). Rates of convex approximation in non-Hilbert spaces. *Constructive Approximation*, *13*, 187–220.
Fine, T. (1999). *Feedforward neural network methodology*. New York: Springer.
Friedman, A. (1982). *Modern analysis*. New York: Dover.
Girosi, F. (1995). Approximation error bounds that use VC-bounds. In *Proceedings of ICANN 1995* (pp. 295–302).
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). Chapman & Hall.
Gurvits, L., & Koiran, P. (1997). Approximation and learning of convex superpositions. *Journal of Computer and System Sciences*, *55*, 161–170.
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. (2009). An integral upper bound for neural network approximation. *Neural Computation*, *21*(10), 2970–2989.
Kainen, P. C., Kůrková, V., & Sanguineti, M. (2009). Complexity of Gaussian radial basis networks approximating smooth 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. (2010). Integral combinations of Heavisides. *Mathematische Nachrichten*, *283*(6), 854–878.
Kainen, P. C., & Vogt, A. (2012). Bochner interals and neural networks. In: Bianchini, M. Jain, L. Maggini, M. (Eds.), *Handbook on neural information processing. Berlin, Heidelberg: Springer-Verlag* (in press).
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. The curse of dimensionality* (pp. 261–270). Boston, MA: Birkhäuser.
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 (Chapter 4).
Kůrková, V. (2008). Minimization of error functionals over perceptron networks. *Neural Computation*, *20*(1), 252–270.
Kůrková, V. (2009). Model complexity of neural networks and integral transforms. In M. Polycarpou, C. Panayiotou, C. Alippi, & G. Ellinas (Eds.), *LNCS: Vol. 5768. Artificial neural networks—ICANN 2009* (pp. 708–718). Springer-Verlag.

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. (2002). Comparison of worst-case errors in linear and neural network approximation. *IEEE Transactions on Information Theory*, *28*, 264–275.
Kůrková, V., & Sanguineti, M. (2005). Error estimates for approximate optimization by the extended Ritz method. *SIAM Journal on Optimization*, *15*, 461–487.
Kůrková, V., & Sanguineti, M. (2008). Geometric upper bounds on rates of variable-basis approximation. *IEEE Transactions on Information Theory*, *54*, 5618–5688.
Kůrková, V., Savický, P., & Hlaváčková, K. (1998). Representations and rates of approximation of real–valued Boolean functions by neural networks. *Neural Networks*, *11*, 651–659.
Lavretsky, E. (2002). On the geometric convergence of neural approximations. *IEEE Transactions on Neural Networks*, *13*, 274–282.
Mhaskar, H. N. (2004). On the tractability of multivariate integration and approximation by neural networks. *Journal of Complexity*, *20*, 274–282.
Mhaskar, H. N. (2006). Weighted quadrature formulas and approximation by zonal function networks on the sphere. *Journal of Complexity*, *22*, 348–370.
Park, J., & Sandberg, I. (1991). Universal approximation using radial-basis-function networks. *Neural Computation*, *3*, 246–257.
Park, J., & Sandberg, I. (1993). Approximation and radial-basis-function networks. *Neural Computation*, *5*, 305–316.
Pietch, A. (1987). *Eigenvalues and s-numbers*. Cambridge: Cambridge University Press.
Pisier, G. (1981). Remarques sur un résultat non publié de B. Maurey. In *École polytechnique*: *Vol. 12. Séminaire d'analyse fonctionnelle 1980–1981, vol. I*. Palaiseau, France: Centre de Mathématiques.
Rudin, W. (1974). *Real and complex analysis*. New York: MacGraw-Hill.
Schaback, R., & Wendland, H. (2006). Kernel techniques: from machine learning to meshless methods. *Acta Numerica*, *15*, 543–639.
Szabados, J. (1974). On an interpolatory analogon of the de la vallée poussin means. *Studia Scientiarum Mathematicarum Hungarica*, *9*, 187–190.
Yoshida, K. (1965). *Functional analysis*. Berlin: Springer-Verlag.