This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier’s archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright
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 věží 2, 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.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

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

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

where the function of two variables $K$ is called a kernel of the integral operator $T_\phi$ (the term “kernel” is derived from the German term “kern” introduced by Hilbert in 1904 Petch (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 $\Omega$ is a set of inputs and $A$ is a set of parameters, can be viewed as a mapping $T_\phi$ assigning to an output weight function $w : A \to \mathbb{R}$ an input–output function $T_\phi(w)(x) : \Omega \to \mathbb{R}$ 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 $L^p$-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

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 \(L_p\)-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 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 \(L_p\)-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 \( A \subseteq \mathbb{R}^n \) 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 \( \Omega \) determined by \( \phi \). The set \( G_\phi \) is sometimes called a dictionary. We use the shorter notation \( G_\phi \) or \( G_\phi(A) \) when the sets \( \Omega \) 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 \( \phi_\sigma : \mathbb{R}^d \times \mathbb{R}^{d+1} \to \mathbb{R}^d \) defined for \((v, b) \in \mathbb{R}^d \times \mathbb{R} = \mathbb{R}^{d+1}\)

\[ \phi_\sigma(v, b) := \sigma(v \cdot x + b). \]

(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

\[ \phi_\beta(v, b) := \beta(b|x - v|) \]

(2)

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

\[ \text{span}_\mu G_\phi(A) := \left\{ \sum_{i=1}^n w_i \phi(\cdot, a_i) \mid w_i \in \mathbb{R}, a_i \in A \right\}. \]

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 \( \mu \) 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

\[ T_{\phi, \mu}(w)(x) := \int_A w(a)\phi(x, a)\,d\mu(a). \]

(3)

When \( \mu \) is the Lebesgue measure, we write for short \( T_{\phi, \mu} \) 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 \( G_\phi \) to input–output functions \( T_{\phi, \mu}(w) : \Omega \to \mathbb{R} \).

Recall that when \( \phi \in L^p(\Omega \times A, \rho \times \mu) \), then \( T_{\phi, \mu} : L^p(\Omega, \mu) \to L^p(\Omega, \rho) \), where \( \frac{1}{p} + \frac{1}{q} = 1 \), is a continuous operator (Friedman, 1982, p. 138). When in addition \( \Omega \) and \( A \) are compact subsets of \( \mathbb{R}^d \) and \( \rho \) and \( \mu \) are Lebesgue measures, then \( T_{\phi} : L^q(\Omega) \to L^p(\Omega) \) 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, Kúrková, & Vogt, 2007; Kúrková et al., 1997). Functions from various Sobolev spaces can be represented as infinite networks with Gaussian RBF units (Kainen, Kúrková, & 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 $(X, \| \cdot \|_X)$, the norm $G$-\textit{variation}, denoted $\| \cdot \|_G$, is defined for all $f \in X$ as
\[ \|f\|_{G,X} := \inf \{ c > 0 \mid \|f/c\|_{cl_X \text{ conv} (G \cup -G)} \}, \]
where the closure $cl_X$ is taken with respect to the topology generated by the norm $\| \cdot \|_X$ and $\text{conv}$ 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 $\|f\|_{G,X}$. 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 $X$ 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 $\text{span}_c G$ in various ambient function spaces. Typically, such bounds are of the form
\[ \|f - \text{span}_c G\|_X \leq n^{-1/2} \xi(d), \]
where $\xi$ 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
\[ n \geq \left( \frac{\xi(d)}{\varepsilon} \right)^3 \]
units can approximate $f$ within $\varepsilon$. 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), Donahue et al. (1997) and Jones (1992) based on construction of incremental approximants cannot be extended to approximation in $L^p$- and $L^\infty$-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 $(F(\Omega), \| \cdot \|_{\sup})$ is denoted the space of all bounded functions on $\Omega$ with the supremum norm.

**Theorem 3.2.** Let $\Omega \subseteq \mathbb{R}^d$, $G$ be a subset of the set of characteristic functions on $\Omega$ such that the coVC-dimension $h^*(G)$ is finite, then for all $f \in F(\Omega)$,
\[ \|f - \text{span}_c G\|_{\sup} \leq 6 \sqrt{3} \|f\|_{\sup} h^*(G)^{1/2} (\log n)^{1/2} n^{-1/2}. \]

### 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 $G$-variation for functions representable as integrals in the form of infinite networks with units computing $\varphi$.

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 $X^*$ is denoted the dual of $X$ (the space of all bounded linear functionals on $X$) and $S_c = \{ f \in X^* \mid \|f\|_G = 0 \}$.

**Theorem 4.1.** Let $(X, \| \cdot \|_X)$ be a normed linear space, $G$ be its nonempty bounded subset and $f \in X$ be such that $\|f\|_X < \infty$. Then
\[ \|f\|_G = \sup_{h \in S_c} \left( \frac{\|f\|_G}{\sup_{g \in G} \|l(g)\|} \right). \]

**Proof.** First, we show that for all $c > 0$ and all $f \in X$
\[ f/c \in cl_X \text{ conv}(G \cup -G) \implies (\forall l \in \mathcal{X}^*) (\|l(f)\| \leq c \sup_{g \in G} \|l(g)\|). \]

If $f/c \in cl_X \text{ conv}(G \cup -G)$ then there exists a sequence $\{f_n\}$ such that
lim_{k \to \infty} \|f/c - f_k\|_X = 0 and all $f_k$ can be represented as $f_k = \sum_{i=1}^{m_k} w_{ki} g_i$, where $\sum_{i=1}^{m_k} w_{ki} g_i = 1$ and all $g_i \in G$. Then for all $l \in X^*$, $f_k(l) = \sum_{i=1}^{m_k} w_{ki} (l g_i)$ and so $|l f_k| \leq \sup_{x \in G} |l g_i|$. Since $l$ is continuous, also $|l f/c| \leq \sup_{x \in G} |l g_i|$ and thus $|l f| \leq \sup_{x \in G} |l g_i|$. 

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

$$\left( \forall l \in S_c \right) (|l f| \leq c \sup_{g \in G} |l g_i|) \implies f/c \in cl_X conv(G \cup -G). \quad (5)$$

Assume by contradiction that $f/c \notin cl_X conv(G \cup -G)$ and let $h \in cl_X conv(G \cup -G)$ such that $\|l h\| \leq 1$. Thus in particular for all $g \in G$, $|l h| \leq 1$. Hence $|l f/c| = |l f| > c \geq \sup_{x \in G} |l g_i|$. It remains to show that $l \in S_c$. As $\|f\|_c$ is finite, there exist some $b > 0$ such that $|l f| > b \sup_{x \in G} |l g_i|$. Hence by (4), $|l f| \leq b \sup_{x \in G} |l g_i|$. If $l \notin S_c$, this would imply $|l f| \leq b \sup_{x \in G} |l g_i| = 0$. But $|l f| > 0$ is in contradiction with $|l f| > c \geq \sup_{x \in G} |l g_i|$. It follows from the implications (4) and (5) that $f/c \in cl_X conv(G \cup -G) \iff c \geq \sup_{l \in S_c} \frac{|l f|}{\sup_{g \in G} |l g_i|}$. \[
\text{Theorem 4.1 is an extension of the characterization of the variational norm in Hilbert spaces proven in Kůrková, Savický, and Hlaváčková (1998), which was used there and in Kůrková (2008) to prove existence of functions with variations growing with the input dimension d exponentially.}
\]

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

$$\|f\|_c = \sup_{h \in cl_X c^\bot} \frac{|\langle h, f \rangle|}{\sup_{g \in G} |\langle g, h \rangle|}. \quad (6)$$

In particular, for all $f \in X \setminus G^\bot$

$$\|f\|_c \geq \frac{\|f\|_X^2}{\sup_{g \in G} |\langle g, f \rangle|}. \quad (7)$$

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 $f = \sum_{i=1}^{m} w_i e_i$, $\|f\|_2 = \|f\|_1 = \sum_{i=1}^{m} |w_i|$. Let $u = (1, \ldots, 1)$. Then by Theorem 4.1 for all $f \in \mathbb{R}^d$

$$\|f\|_c \geq \frac{\|f \cdot u\|_2}{\sup_{i=1,\ldots,m} |e_i \cdot u|} = \frac{\sum_{i=1}^{m} |w_i|}{m}. \quad (8)$$

As in this case $\|f\|_c = \|f\|_1$, 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 follows easily from the definition of variational norm that for each $f$ which can be represented as $f = T_\phi(w)$ for some $w \in \mathbb{R}^m$, 
\[
\|f\|_{C_0(A)} = \min \left\{ \|w\|_1 \bigg| f = \sum_{i=1}^{m} w_i \phi(\cdot, a_i) \right\}.
\] (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 $\ell_1$ or $\ell_2$-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 
\[
\|f\|_{C_0(A)} \leq \|w\|_{L^1(A, \mu)}
\] might hold. The inequality (9) can only be considered when quantities on both its sides are well-defined, i.e., when:

(i) $C_0(A)$ is a bounded subset of $(\mathcal{X}, \|\cdot\|_X)$ and
(ii) $w \in L^1(A, \mu)$.

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 $C_0(A)$-variation and $L^1$-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 $(\mathcal{X}, \|\cdot\|_X)$ of functions on $\Omega \subseteq \mathbb{R}^d$ has a commutativity property of linear functionals with kernel operators if for every integral operator $T_\phi$ on $(L^p(A), \|\cdot\|_{L^p}) \rightarrow (\mathcal{X}, \|\cdot\|_X)$ with a kernel $\phi : \Omega \times A \rightarrow \mathbb{R}$ such that $C_0(A) = \{\phi(\cdot, a) | a \in A\}$ is a bounded subset of $(\mathcal{X}, \|\cdot\|_X)$, and every linear functional $l$ on $\mathcal{X}$ and every $g \in L^1(A)$ 
\[
l(T_\phi(f)) = \int_A f(a) l(\phi(\cdot, a)) d\mu(a).
\]
This property holds, for example, in spaces $(L^p(\Omega, \rho), \|\cdot\|_{L^p})$ with $p \in [1, \infty)$, $(C(\Omega), \|\cdot\|_{sup})$, and $(C_0(\mathbb{R}^d), \|\cdot\|_{sup})$ as it is shown in Theorem 5.2.

The next theorem on the relationship (9) between $C_0(A)$-variation of an input–output function of an infinite network and the $L^1$-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 $(\mathcal{X}, \|\cdot\|_X)$ be a space of functions on $\Omega \subseteq \mathbb{R}^d$ satisfying the commutativity property of linear functionals with kernel operators, $\mu$ be a $\sigma$-finite measure on $\Omega \subseteq \mathbb{R}^d$, $w \in L^1(A, \mu)$, $\phi : \Omega \times A \rightarrow \mathbb{R}$ be such that $C_0(A) = \{\phi(\cdot, a) | a \in A\}$ is a bounded subset of $(\mathcal{X}, \|\cdot\|_X)$, 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 
\[
\|f\|_{C_0(A)} \leq \|w\|_{L^1(A, \mu)}.
\]

**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 $|l(f)| \leq \sup_{a \in A} l(\phi(\cdot, a)) \int_A |w(a)| d\mu(a) = \sup_{a \in A} l(\phi(\cdot, a)) \int_A w(a) d\mu(a)$. By Theorem 4.1, 
\[
\|f\|_{C_0(A)} = \sup_{l \in \mathcal{X}^*} \left\{ \sup_{a \in A} l(\phi(\cdot, a)) \right\} \leq \|w\|_{L^1(A, \mu)}. \quad \Box
\]

The next theorem describes some function spaces with the commutativity property. For $\Omega \subseteq \mathbb{R}^d$, by $(C_0(\Omega), \|\cdot\|_{sup})$ denotes the space of all continuous compactly supported functions on $\Omega$ with the supremum norm and by $(C_0(\mathbb{R}^d), \|\cdot\|_{sup})$ the space of all continuous functions on $\mathbb{R}^d$ vanishing at infinity (i.e., functions $f$ such that $\lim_{|x| \rightarrow \infty} f(x) = 0$).

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

(i) $(L^p(\Omega, \rho), \|\cdot\|_{L^p})$ with $p \in [1, \infty)$, and $\rho$ a measure on $\Omega \subseteq \mathbb{R}^d$;
(ii) $(C(\Omega), \|\cdot\|_{sup})$ with $\Omega$ a locally compact subset of $\mathbb{R}^d$;
(iii) $(C_0(\mathbb{R}^d), \|\cdot\|_{sup})$ with $\mathbb{R}^d = \mathbb{R}^d$.

**Proof.** First, we prove the statement for case (i). By the properties of the duals of $L^p$-spaces with $p \in [1, \infty)$ (Friedman, 1982, pp. 176, 180), for every $l \in \mathcal{X}^*$ there exists $h \in L^q(\Omega, \rho)$ (where for $p > 1$, $q$ satisfies $1/q + 1/p = 1$, while for $p = 1$, $q = \infty$), such that for all $f \in L^p(\Omega, \rho)$, 
\[
l(f) = \int_{\Omega} (f(x) h(x)) d\rho(x).
\]
By Hölder’s inequality (Friedman, 1982, p. 96) for all $a \in A$, $\phi(\cdot, a) h \in L^q(\Omega, \rho)$ and 
\[
\|\phi(\cdot, a) h\|_{L^q} = \|\phi(\cdot, a)\|_{L^p} \|h\|_{L^q}.
\]
Thus for all $a \in A$, $\int_{\Omega} |\phi(\cdot, a) h(x)| d\rho(x) \leq \|\phi(\cdot, a)\|_{L^p} \|h\|_{L^q}$.
By the assumption $C_0(A)$ is bounded and so $\|\phi(\cdot, a)\|_{L^p}$ is finite. Thus also 
\[
\int_{\Omega} \left| \int_{\Omega} (f(x) h(x) d\rho(x)) \right| d\mu(a) = \int_{\Omega} (f(x) h(x) d\rho(x)) d\mu(a).
\]

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 $\nu$ on $\Omega$ such that for all $f \in C(\Omega)$ or $f \in C_0(\mathbb{R}^d)$, $l(f) = \int_{\Omega} f(x) d\nu(x)$ and $|l(\nu)| = \|l\|_{\mathcal{X}^*}$, where $|\nu|$ denotes the total variation of $\nu$. Thus for all $a \in A$, $\int_{\Omega} \phi(x, a) d\nu(x) \leq \|\phi(\cdot, a)\|_{\mathcal{X}^*}$.

The suprema $\|\phi(\cdot, a)\|_{\mathcal{X}^*}$ is finite, also 
\[
\int_{\Omega} |w(a)| \int_{\Omega} |\phi(x, a)| d\nu(x) d\mu(a) \leq \|w(a)\|_{L^1} \|\nu\|_{\mathcal{X}^*} \|\phi(\cdot, a)\|_{\mathcal{X}^*}.
\]
is finite. Thus we can use Fubini's theorem (Friedman, 1982, p. 86) to obtain 
\[
l(f) = \int_{\Omega} \left( \int_{\Omega} w(a) \phi(x, a) d\mu(a) \right) h(x) d\rho(x) = \int_A (w(a) h(x) d\rho(x)) d\mu(a) = \int_{\Omega} (w(a) h(x) d\rho(x)) d\mu(a) = \int_{\Omega} (w(a) l(\phi(a)) d\mu(a).
\]

\[
\Box
\]
Corollary 5.3. Let \((X, \| \cdot \|_X)\) be one of the following spaces:

(i) \(L^p(\Omega, \rho, \| \cdot \|_{L^p})\) with \(p \in [1, \infty]\) and \(\rho\) a \(\sigma\)-finite measure;
(ii) \((C(\Omega), \| \cdot \|_{sup})\) with \(\Omega\) a locally compact subset of \(\mathbb{R}^d\);
(iii) \((G_0(\Omega), \| \cdot \|_{sup})\) with \(\Omega = \mathbb{R}^d\).

Let \(\mu\) be a \(\sigma\)-finite measure on \(A \subseteq \mathbb{R}^d, w \in L^1(A, \mu), \phi : \Omega \times A \to \mathbb{R}\) be such that \(G_0(A) = \{\phi(a) | a \in A\}\) is a bounded subset of \((X, \| \cdot \|_X)\) and \(f \in X\) be such that for all \(x \in \Omega, f(x) = \int_A w(a)\phi(x, a)d\mu(a)\).

Then

\[\|f\|_{G_0(A)} \leq \frac{\|w\|_{L^1(A, \mu)}^2}{n}\|

Applying the upper bound on \(G_0\)-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 \(G_0\).

Corollary 5.4. Let \((X, \| \cdot \|_X)\) be a space of functions on \(\mathbb{R}^d, A \subseteq \mathbb{R}^d, \mu\) be a measure on \(A, \phi : \Omega \times A \to \mathbb{R}\) be such that \(G_0(A) = \{\phi(a) | a \in A\}\) is a bounded subset of \((X, \| \cdot \|_X)\) and \(s_\phi = sup_{a \in A} |\phi(a)|_X\). Let \(f \in (X, \| \cdot \|_X)\) be such that for some \(w \in L^1(A, \mu), f(x) = \int_A w(a)\phi(x, a)d\mu(a)\). Then for all \(n\)

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

\[\|f - \text{span}_n G_0(\Omega)\|_{L^2}^2 \leq \frac{\|s_\phi\|^2}{n} \]

(ii) for \((X, \| \cdot \|_X) = (L^2(\Omega), \| \cdot \|_{L^2}), p \in (1, \infty),\)

\[\|f - \text{span}_n G_0(\Omega)\|_{L^p}^{2/p} \leq \frac{\|s_\phi\|^{2/p}}{n^{2/p}} \]

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

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 \(G_{\vartheta} = G_{\vartheta}(S^{d-1}) \in \mathbb{R}^d\). Recall that \(G_{\vartheta}\) has been called variation with respect to half-spaces (Barron, 1992) as the dictionary \(G_{\vartheta}\) 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 \(\lim_{t \to \infty} \sigma(t) = 1\))

\[\|\cdot\|_{G_{\vartheta}} = \|\cdot\|_{G_{\vartheta}}\]

in \(L^p(\Omega)\) with \(p \in (1, \infty)\) and \(\Omega\) compact (Kůrková et al., 1997). So estimates of variation with respect to half-spaces apply also to \(G_{\vartheta}\)-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 \(C^\infty(\mathbb{R}^d)\) (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 \(C^0(\mathbb{R}^d)\) (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 \(\delta_d\) as an integral of derivatives of the one-dimensional delta function

\[\delta_d(x) = ad \int_{S^{d-1}} \delta^{(d-1)}(e \cdot x) de,\]

where \(a_\vartheta = (\frac{-1}{d})^{d/2}(2\pi)^{-d}\). Kainen et al. (2007) extended the representation of 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 \(C^0(\mathbb{R}^d)\) and the Schwartz class \(\mathcal{S}(\mathbb{R}^d)\) (all functions from \(C^\infty(\mathbb{R}^d)\) 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 \(D^\infty_{\vartheta}\) is denoted the directional derivative of the \(d\)-dimensional vector \(e\) and by \(H_{\vartheta}\), the hyperplane \(\{x \in \mathbb{R}^d | \vartheta(e \cdot x + b) = 0\}\).

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

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

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

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

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

\[\|f\|_{\Omega} - \text{span}_n G_{\vartheta}(\Omega)\|_{L^p(\Omega)} \leq \frac{\lambda(\Omega)\|w_f\|_{L^1(S^{d-1} \times \mathbb{R})}}{\sqrt{n}}.\]

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

Proof. By Theorem 6.1, for all \(x \in \Omega, f(x) = \int_{S^{d-1} \times \mathbb{R}} w_f(e, b) \vartheta(e \cdot x + b) \,db\). As \(G_{\vartheta}(\omega)\) is a bounded subset of \(L^2(\Omega)\) with \(s_{\vartheta} \leq \lambda(\Omega)\) and \(w_f \in L^2(S^{d-1} \times \mathbb{R})\), we can apply Theorem 4.1 to obtain \(\|f\|_{\Omega} - \text{span}_n G_{\vartheta}(\Omega)\|_{L^p(\Omega)} \leq \|w_f\|_{L^1(S^{d-1} \times \mathbb{R})}\). Then the statement follows from the equality \(\|f\|_{G_{\vartheta}} = \|f\|_{G_{\vartheta}}\)

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 \(\int_{H_{\vartheta}} D^d_{\vartheta}(f)(y) dy\) with \(\lambda(\Omega)\) depends on the shape of the \(d\)-dimensional domain \(\Omega \subseteq \mathbb{R}^d\). When \(\Omega\) is the Euclidean \(d\)-dimensional ball, then \(\lambda(\Omega)\) goes to zero exponentially fast with \(d\) increasing, while when \(\Omega\) 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 $L^1$-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 $\ell_1$ or $\ell_2$-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 $\Omega$ and the set of parameters $A$ (compactness in Kůrková et al., 1997), and on $\phi$ and $w$ (continuity in Kainen & Kůrková, 2009), or a special choice of $\phi$ (a trigonometric function in Barron, 1993). Our approach uses only minimal assumptions necessary for existence of the quantities which are compared: the $G$-type variation can be defined and the output weight $w$ has to be in $L^1(A, \mu)$ so that its $L^1$-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 GACR grant P202/11/1368.

Appendix

For the reader’s convenience we include several concepts and tools used in the paper. By $\chi_{\Omega} : \Omega \rightarrow [0,1]$ is denoted the characteristic function of $\Omega \subseteq \mathbb{R}^n$, i.e. $\chi_{\Omega}(x) = 1$ if $x \in \Omega$, otherwise $\chi_{\Omega}(x) = 0$. Let $\mathcal{F}$ be any family of characteristic functions of subsets of $\Omega$ and $\mathcal{F}_s = \{ \Omega \subseteq \Omega \mid \chi_{\Omega} \in \mathcal{F} \}$ be the family of the corresponding subsets of $\Omega$. Then a subset $A$ of $\Omega$ is said to be shattered by $\mathcal{F}$ if $\{ S \cap A \mid S \in \mathcal{F}_s \}$ 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 $\mathcal{F}^c := \{ e_{\chi_{\Omega}} \mid x \in \Omega \}$, where the evaluation $e_{\chi_{\Omega}} : \mathcal{F} \rightarrow [0,1]$ is defined for every $\chi_{\Omega} \in \mathcal{F}$ as $e_{\chi_{\Omega}}(\chi_{\Omega}) = \chi_{\Omega}(x)$.

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

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

$$\|f\|_{W^k_s} = \left( \sum_{|\alpha| \leq s} \|D^\alpha f\|_q^q \right)^{1/q} ,$$

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

A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is of a weakly controlled decay when it satisfies for all multi-indexes $\alpha$ with $0 \leq |\alpha| = \alpha_1 + \cdots + \alpha_d < d$,

$$(\partial^{\alpha_1} / \partial x_1^{\alpha_1}) \cdots (\partial^{\alpha_d} / \partial x_d^{\alpha_d}) (D^\alpha f)(x) = 0$$

where $D^\alpha = (\partial / \partial x_1)^{\alpha_1} \cdots (\partial / \partial x_d)^{\alpha_d}$ and for some $c > 0$, all multi-indexes $\alpha$ with $|\alpha| = d$ satisfy

$$\| (D^\alpha f)(x) \|_1 \leq c = 0.$$

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 $X$ be a real locally convex linear topological space, $M$ a closed convex balanced subset of $X$. Then for any $f \notin M$ there exists a continuous linear functional $l_1$ on $X$ such that $l_1(f) > 1$ and for all $h \in M, \|l_1(h)\| \leq 1$.

**References**