Neural Networks 11: 651--659, 1998

Representations and rates of approximation of real-valued Boolean functions by neural networks

Věra Kůrková, Petr Savický, Kateřina Hlaváčková
Institute of Computer Science, Academy of Sciences of the Czech Republic

Acknowledgements: V. Kůrková and K. Hlaváčková were partially supported by GA AV grants A2030602 and A2075606 and P. Savický was partially supported by GA AV grant A2010976.

Requests for reprints should be sent to:
Věra Kůrková, Institute of Computer Science
Pod vodárenskou věží 2, 182 07 Prague 8, Czechia
phone +420 266053231, fax +420 286585789, e-mail vera@uivt.cas.cz

Running title: Representations and rates of approximation

Representations and rates of approximation of real-valued Boolean functions by neural networks

Abstract

We give upper bounds on rates of approximation of real-valued functions of d Boolean variables by one-hidden-layer perceptron networks. Our bounds are of the form $\frac{c}{\sqrt{n}}$, where c depends on certain norms of the function being approximated and n is the number of hidden units. We describe sets of functions where these norms grow either polynomially or exponentially with d.

Keywords. Real-valued Boolean function, perceptron network, rate of approximation, variation with respect to half-spaces, decision tree, Hadamard communication matrix.

1 Introduction

The existence of an arbitrarily close approximation has been proved for perceptron type and radial-basis-function networks with quite general activation and kernel functions (see e.g. Leshno et al., 1993, Mhaskar & Micchelli, 1992, Park & Sandberg, 1993). The dependence of approximation error upon the number of hidden units, i.e. the rate of approximation, has become better understood. Jones (1992) introduced a recursive construction of approximants with "dimension-independent" rates of convergence to functions in convex closures of bounded subsets in a Hilbert space and together with Barron he proposed to apply this method to the sets of functions computable by one hidden-layer neural networks. Applying Jones' estimate, several authors (e.g. Barron, 1993; Girosi & Anzellotti, 1993; Kůrková et al., 1997) characterized sets of functions with d real variables that can be approximated by networks with n hidden units of various types (perceptron or radial-basis-function) within an error $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$. Using the construction of approximants based on a rearrangement of a fixed basis of a separable Hilbert space, Mhaskar and Micchelli (1994) obtained characterizations of a different type also providing the approximation error within $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$.

In some applications, input data are represented using only binary values. When computational units used in the hidden layer are continuous sigmoidal perceptrons or when the output weights are real numbers, the input/output functions of such networks are real-valued functions of several Boolean variables. A typical example of an application of this type is Sejnowski and Rosenberg's NETtalk (1987) where a real-valued function with approximately two hundred Boolean variables is approximated sufficiently well by a neural network with only eighty hidden units.

Motivated by these experimental results, we investigate both representation and approximation of real-valued functions of several Boolean variables by one-hidden-layer perceptron networks. In contrast to the case of functions of several *real* variables which can be implemented by such networks only *approximately*, all real-valued functions of *d Boolean* variables can be computed *exactly* by perceptron networks with any sigmoidal activation function (Ito, 1992).

We consider two such exact representations obtained by expressing functions from two standard bases (the Euclidean and the Fourier one) as functions computable by one-hidden-layer perceptron networks. Since both of these representations require networks with the number of hidden units growing exponentially with the number of variables d, we examine the effect of reduction of the number of hidden units upon accuracy decrease. We estimate the rate of approximation in terms of various norms of the function to be approximated, namely the standard l_1, l_2-norms and the variation with respect to a set of functions (which is Kůrková's, 1997, generalization of Barron's, 1992, concept of variation with respect to half-spaces).

We derive our estimates using two methods: In the first one, we directly apply Jones-Barron's theorem (Jones, 1992, Barron, 1993) reformulated in terms of variation with respect to a set of functions; In the second one, we derive a strengthening of Mhaskar and Micchelli's (1994) bound on rate of approximation from an orthonormal approximating sets and apply it to the Fourier basis with elements represented as functions computable by perceptron networks.

We describe functions for which the second method gives considerably better estimates. Moreover, we show that if only the l_1 and l_2-norm of the function being approximated are known, then this method gives the best possible upper estimate up to a constant factor. To illustrate the strengths and weaknesses of our estimates, we give examples of functions with the norms involved growing both exponentially and polynomially with d.

The paper is organized as follows. In section 2, we recall and extend estimates of rates of approximation applicable to the approximation from a general subset of a finite dimensional Hilbert space (subsection 2.1) and their stronger version for approximation from an orthonormal subset (subsection 2.2). In section 3, we use these tools to obtain upper bounds on rates of approximation of real-valued functions of d Boolean variables by one-hidden-layer perceptron networks (subsection 3.1). Further we investigate tightness of these estimates by applying them to functions representable by polynomial size decision trees and to functions with Hadamard communication matrices (subsection 3.2). In section 4 we discuss our results and some open problems. All proofs are deferred to section 5.

2 Rates of approximation in finite dimensional linear spaces

The estimates of rates of approximation of real-valued functions of several Boolean variables by functions computable by perceptron networks presented in this paper are derived using quite general tools applicable to any finite dimensional Hilbert space. We recall and extend these tools in this section.

Let $\mathcal{R}$, $\mathcal{N}$ denote the set of real numbers, natural numbers, resp., and $\mathcal{R}_+$, $\mathcal{N}_+$ the set of positive reals, integers, resp.

In this paper, by a linear space we always mean a real linear space. For a subset $\mathcal{G}$ of a linear space $\mathcal{X}$ we denote by $\operatorname{span}\mathcal{G}$, $\operatorname{conv}\mathcal{G}$ the *linear span* of $\mathcal{G}$, the *convex hull* of $\mathcal{G}$, resp., and by $\operatorname{span}_n\mathcal{G}$, $\operatorname{conv}_n\mathcal{G}$ the set of all linear, convex, resp., combinations of n elements of $\mathcal{G}$.

For a subset $\mathcal{H}$ of a normed linear space $(\mathcal{X}, \|.\|)$ and $f \in \mathcal{X}$ we denote the distance of f from $\mathcal{H}$ by $\|f - \mathcal{H}\| = \inf_{h\in\mathcal{H}} \|f - h\|$ and call it *error of approximation of* f *by* $\mathcal{H}$; $\frac{\|f-\mathcal{H}\|}{\|f\|}$ is *relative error of approximation of* f *by* $\mathcal{H}$.

Rate of approximation of f *by* $\mathcal{G}$ is $\|f - span_n\mathcal{G}\|$ as a function of n; *relative rate of approximation of* f *by* $\mathcal{G}$ is $\frac{\|f-span_n\mathcal{G}\|}{\|f\|}$.

By a Hilbert space we mean a complete normed linear space with a norm induced by an inner product (also including the finite-dimensional case). When $(\mathcal{X}, \cdot)$ is a Hilbert space, we denote by $\|.\|_2$ the norm induced on $\mathcal{X}$ by the inner product, i.e. $\|f\|_2 = \sqrt{f \cdot f}$.

2.1 Upper bounds for general approximating sets

The first tool we use is a reformulation of Jones' (1992) estimate of $\|f - \operatorname{conv}_n\mathcal{G}\|_2$ in terms of $\mathcal{G}$-variation, a special case of Minkowski's functional introduced by Kůrková (1997). The following theorem is equivalent to Barron's (1993) improvement of Jones' result.

Theorem 2.1 (Jones-Barron) *Let* $\mathcal{X}$ *be a Hilbert space,* b *be a positive real number and* $\mathcal{G}$ *be a subset of* $\mathcal{X}$ *such that for every* $g \in \mathcal{G}$ $\|g\|_2 \le b$. *Then for every* $f \in \operatorname{conv}\mathcal{G}$ *and for every positive integer* n *there exists* $f_n \in \operatorname{conv}_n \mathcal{G}$ *such that*

$$\|f - f_n\|_2 \le \sqrt{\frac{b^2 - \|f\|_2^2}{n}}.$$

Notice that the upper bound on the distance from $\operatorname{conv}_n\mathcal{G}$ guaranteed by this theorem to all elements of $\operatorname{conv}\mathcal{G}$ can be extended to $\operatorname{cl}\operatorname{conv}\mathcal{G}$ (for which this theorem is formulated in Barron, 1993) if we add to the upper bound an arbitrarily small positive number.

To apply this theorem to functions in $\operatorname{span}\mathcal{G}$ consider for each $f = \sum_{i=1}^m w_i g_i$, where all $w_i \in \mathcal{R}$ and all $g_i \in \mathcal{G}$, a representation $f = \sum_{i=1}^m \frac{|w_i|}{a} \operatorname{sgn}(w_i) a g_i$, where

$a = \sum_{i=1}^m |w_i|$ and sgn denotes the signum function. Thus any $f \in \operatorname{span} \mathcal{G}$ is for a sufficiently large a in $\operatorname{conv} \mathcal{G}(a)$, where $\mathcal{G}(a) = \{wg; w \in \mathcal{R}, |w| \leq a, g \in \mathcal{G}\}$.

To derive estimates of rates of approximation by neural networks from Jones-Barron's theorem, Barron (1992) introduced a concept of variation of a function with respect to a set of characteristic functions, in particular *variation with respect to half-spaces.* Kůrková (1997) generalized this concept to variation with respect to a set of functions in a normed linear space. For a subset $\mathcal{G}$ of a normed linear space $(\mathcal{X}, \|.\|)$ she defined $\mathcal{G}$*-variation* of $f \in \mathcal{X}$ as

$$V(f, \mathcal{G}) = \inf\{a > 0; f \in \operatorname{cl} \operatorname{conv} \mathcal{G}(a)\}.$$

Using this notion and the considerations above, to each $f \in \cup_{a \in \mathcal{R}_+} \operatorname{cl} \operatorname{conv} \mathcal{G}(a)$ we can apply Jones-Barron's estimate with $b = V(f, \mathcal{G}) \sup_{g \in \mathcal{G}} \|g\|$.

It is straightforward to show that $\mathcal{G}$-variation is a norm on $\{f \in \mathcal{X}; V(f, \mathcal{G}) < \infty\}$ and that for every $f \in \mathcal{X}$ $\|f\| \leq V(f, \mathcal{G}) \sup_{g \in \mathcal{G}} \|g\|$; notice that $\mathcal{G}$-variation is the Minkowski functional of the set $\operatorname{cl} \operatorname{conv} \mathcal{G}(1)$.

For a nonzero $f \in \mathcal{X}$ let $f^0 = \frac{f}{\|f\|}$ be the normalization of f and let $\mathcal{G}^0$ denotes the set of normalized elements of $\mathcal{G}$, i.e. $\mathcal{G}^0 = \{g^0; g \in \mathcal{G}\}$. We call $V(f^0, \mathcal{G}^0)$ *normalized* $\mathcal{G}$*-variation of* f. For every $f \in \mathcal{X}$ $\|f\| \leq V(f, \mathcal{G}^0)$, i.e. the unit ball of $\mathcal{G}^0$-variation is contained in the unit ball of $\|.\|$.

Since, clearly, $V(f, \mathcal{G}^0) \leq V(f, \mathcal{G}) \sup_{g \in \mathcal{G}} \|g\|$, we use $\mathcal{G}^0$-variation in our estimates.

The following result gives a geometric characterization of $\mathcal{G}$-variation. Its proof is based on separation of a point from a closed convex set by a hyperplane (see e.g. Holmes, 1975). $\mathcal{G}^\perp$ denotes the orthogonal complement of $\mathcal{G}$.

Theorem 2.2 *Let* $(\mathcal{X}, \|.\|_2)$ *be a Hilbert space and* $\mathcal{G}$ *be its non-empty subset. Then for every* $f \in \mathcal{X}$

$$V(f, \mathcal{G}) = \sup_{h \in S} \frac{|f \cdot h|}{\sup\limits_{g \in \mathcal{G}} |g \cdot h|},$$

where $S = \{h \in \mathcal{X} - \mathcal{G}^\perp; \|h\|_2 = 1\}$.

Hence in particular, $V(f^0, \mathcal{G}^0) \geq \frac{1}{\sup_{g \in \mathcal{G}} |f^0 \cdot g^0|}$. Thus, functions that are "almost orthogonal" to $\mathcal{G}$ have a large normalized $\mathcal{G}$-variation.

When $\mathcal{G}$ is finite then the following simpler characterization of $\mathcal{G}$-variation is a straightforward consequence of the definition.

Proposition 2.3 *Let* $(\mathcal{X}, \|.\|)$ *be a normed linear space,* $\mathcal{G}$ *be its finite subset with card* $\mathcal{G} = n$ *and* $f \in \operatorname{span} \mathcal{G}$. *Then*

$$V(f, \mathcal{G}) = \min \left\{ \sum_{i=1}^n |w_i|; \; f = \sum_{i=1}^n w_i g_i, (\forall i = 1, \ldots, n)(w_i \in \mathcal{R}, g_i \in \mathcal{G}) \right\}.$$

Note that this characterization enables us to define $\mathcal{G}$-variation for finite $\mathcal{G}$ independently of the norm $\|.\|$. Characterization of variation with respect to sets that are larger than the dimension of the space can be further simplified. The following lemma, proved using a technique from linear programming, shows that it is sufficient to reduce the number of elements in a linear combination used to compute variation to the dimensionality of the space.

Lemma 2.4 *Let* $\mathcal{X}$ *be a finite dimensional linear space with* $\dim \mathcal{X} = m$, $\mathcal{G}$ *be its subset,* n *be a positive integer and* $f \in span_n\mathcal{G}$ *has a representation* $f = \sum_{i=1}^n w_i g_i$, *where for every* $i = 1, \ldots, n$ $w_i \in \mathcal{R}$ *and* $g_i \in \mathcal{G}$. *Then there exists a representation of* f *of the form* $f = \sum_{i=1}^n v_i g_i$ *such that at most* m *of the coefficients* $v_1, \ldots, v_n$ *are non-zero and* $\sum_{i=1}^n |v_i| \leq \sum_{i=1}^n |w_i|$.

The following corollary is Jones-Barron's theorem reformulated in terms of $\mathcal{G}^0$-variation.

Corollary 2.5 *Let $(\mathcal{X}, \|.\|_2)$ be a Hilbert space and $\mathcal{G}$ be its subset. Then for every $f \in \operatorname{span} \mathcal{G}$ and for every positive integer n there exists $f_n \in \operatorname{span}_n \mathcal{G}$ such that*

$$\|f - f_n\|_2 \leq \frac{V(f, \mathcal{G}^0)}{\sqrt{n}} \sqrt{1 - \frac{1}{V(f^0, \mathcal{G}^0)^2}}.$$

In the next section, we will show that when $\mathcal{G}$ is an orthonormal set of functions then the upper bound guaranteed by Corollary 2.5 can be slightly improved for $n \geq 2$.

Corollary 2.5 gives an upper bound on relative rate of approximation of the form

$$\frac{\|f - f_n\|_2}{\|f\|_2} \leq \frac{V(f^0, \mathcal{G}^0)}{\sqrt{n}} \sqrt{1 - \frac{1}{V(f^0, \mathcal{G}^0)^2}}.$$

Thus, the number of elements of $\mathcal{G}$ needed to guarantee a given relative error depends only on $V(f^0, \mathcal{G}^0)$.

If $V(f^0, \mathcal{G}^0)$ is large, then the factor $\sqrt{1 - \frac{1}{V(f^0,\mathcal{G}^0)^2}}$ is close to 1 and so its role becomes negligible. Neglecting this factor, we get a relative error less than 1 only for $n > V(f^0, \mathcal{G}^0)^2$. For $n \leq V(f^0, \mathcal{G}^0)^2$ the upper bound implied by Corollary 2.5 becomes trivial.

On the other hand, if $V(f^0, \mathcal{G}^0)$ is close to 1 then the factor $\sqrt{1 - \frac{1}{V(f^0,\mathcal{G}^0)}}$ is close to zero and it might outweigh the first factor $\frac{V(f^0,\mathcal{G}^0)}{\sqrt{n}}$. In such a case, even approximation by only one element of $\mathcal{G}$ might be quite good. For example, if $V(f^0, \mathcal{G}^0) \leq 1 + \delta$ then there exists $f_1 \in \operatorname{span}_1 \mathcal{G}$ such that $\frac{\|f - f_1\|_2}{\|f\|_2} \leq \sqrt{\delta(2 + \delta)}$.

Both Jones' proof and its Barron's modification are constructive – they are based on an upper estimate of $\|f - f_n\|_2$ expressed by a recursive formula. The same upper bound on $\|f - \operatorname{span}_n \mathcal{G}\|_2$ as is implied by Jones-Barron's theorem was obtained by Maurey using a probabilistic argument (see Barron, 1993).

Darken et al. (1993) extended Jones-Barron's theorem to $\mathcal{L}_p$-norms for $p \in (1, \infty)$ with a slightly worse rate of approximation – of the order of only $\mathcal{O}(n^{-\frac{1}{q}})$, where $q = \max(p, \frac{p}{p-1})$. They also described counter-examples showing that the sequence of incremental approximants constructed by Jones may fail to converge in normed linear spaces in which the unit ball has a sharp corner. In particular, Jones' technique does not work for l_1 and l_∞-norms.

However, the probabilistic argument used by Maurey combined with Chernoff bound (see e.g. Alon & Spencer, 1992) for estimating the probability of large deviations from the expected value gives estimates even for rates of approximation measured in l_∞-norm (i.e. maximum norm). The following upper bound on uniform approximation error is a straightforward generalization of Bruck and Smolensky's (1992) bound on rate of approximation by elements of the Fourier basis, see also Siu and Bruck (1991).

Theorem 2.6 *Let m be a positive integer and let $\mathcal{G} \subseteq \{-1, 1\}^m$. Then for every $f \in \operatorname{span} \mathcal{G} \subseteq \mathcal{R}^m$ and for every positive integer n there exists $f_n \in \operatorname{span}_n \mathcal{G}$ such that*

$$\|f - f_n\|_\infty \leq \frac{V(f, \mathcal{G}^0)}{\sqrt{n}} \sqrt{\frac{2 \ln(2m)}{m}}.$$

Notice that Theorem 2.6 implies that $\|f - f_n\|_2 \leq \frac{V(f,\mathcal{G}^0)}{\sqrt{n}} \sqrt{2 \ln(2m)}$. For spaces of moderate dimension this estimate is only slightly worse than the direct estimate of l_2-error in Corollary 2.5.

2.2 Upper bounds for orthonormal approximating sets

Mhaskar and Micchelli (1994) showed that when the set of approximating functions is an orthonormal basis, then upper bounds on rates of approximation can be improved and proofs can be simplified. We will strengthen their results to achieve tight estimates for finite dimensional Hilbert spaces.

For an orthonormal basis $\mathcal{A}$ of a finite dimensional Hilbert space $\mathcal{X}$, we denote by $\|.\|_{1,\mathcal{A}}$, the l_1-*norm with respect to* $\mathcal{A}$, i.e. $\|f\|_{1,\mathcal{A}} = \sum_{i=1}^m |w_i|$, where $f = \sum_{i=1}^m w_i g_i$. It is easy to see that for every $f \in \mathcal{X}$ $V(f,\mathcal{A}) = \|f\|_{1,\mathcal{A}}$, i.e. $\mathcal{A}$-variation is the l_1-norm with respect to $\mathcal{A}$.

So Corollary 2.5 gives an upper bound on the rate of approximation from an orthonormal set $\mathcal{A}$ in terms of the l_1-norm with respect to $\mathcal{A}$ and the l_2-norm, namely $\frac{\|f\|_{1,\mathcal{A}}}{\sqrt{n}}\sqrt{1-\frac{\|f\|_2^2}{\|f\|_{1,\mathcal{A}}^2}}$. Mhaskar and Micchelli (1994) obtained an upper bound on rate of approximation from an orthonormal basis of a separable Hilbert space of the form $\frac{\|f\|_{1,\mathcal{A}}}{\sqrt{n+1}}$. The following theorem shows that in finite dimensional case their bound can be improved.

Theorem 2.7 *Let* $(\mathcal{X},\|.\|_2)$ *be a finite dimensional Hilbert space and let* $\mathcal{A}$ *be its orthonormal basis. Then for every* $f \in \mathcal{X}$ *and for every positive integer* n *there exists* $f_n \in \mathrm{span}_n\mathcal{A}$ *such that*

$$\|f-f_n\|_2 \le \frac{\|f\|_{1,\mathcal{A}}}{2\sqrt{n}}.$$

The following theorem shows that when the only information available about f is the value of its $\mathcal{A}$-variation then this upper bound cannot be further improved.

Theorem 2.8 *Let* $(\mathcal{X},\|.\|_2)$ *be a finite dimensional Hilbert space,* n *be a positive integer such that* $2n \le \dim\mathcal{X}$ *and let* $b \ge 0$ *be an arbitrary real number. Then for every orthonormal basis* $\mathcal{A}$ *of* $\mathcal{X}$ *there exists* $f \in \mathcal{X}$ *with* $\|f\|_{1,\mathcal{A}} = b$ *such that for every* $f_n \in \mathrm{span}_n\mathcal{A}$

$$\|f-f_n\|_2 \ge \frac{\|f\|_{1,\mathcal{A}}}{2\sqrt{n}}.$$

However, if in addition to $\|f\|_{1,\mathcal{A}}$ also $\|f\|_2$ is known, then the upper bound given in Theorem 2.7 can be improved.

Theorem 2.9 *Let* $(\mathcal{X},\|.\|_2)$ *be a finite dimensional Hilbert space and* $\mathcal{A}$ *be its orthonormal basis. Then for every* $f \in \mathcal{X}$ *and for every positive integer* n *there exists* $f_n \in \mathrm{span}_n\mathcal{A}$ *such that*

$$\|f-f_n\|_2 \le \frac{\|f\|_{1,\mathcal{A}}}{2\sqrt{n-1}}\left(1-\frac{\|f\|_2^2}{\|f\|_{1,\mathcal{A}}^2}\right).$$

Note that the upper bound implied by Jones-Barron's theorem in this case is

$$\|f-f_n\|_2 \le \frac{\|f\|_{1,\mathcal{A}}}{\sqrt{n}}\sqrt{1-\frac{\|f\|_2^2}{\|f\|_{1,\mathcal{A}}^2}}.$$

If $\|f\|_{1,\mathcal{A}}$ is close to $\|f\|_2$ then the upper bound folowing from Theorem 2.9 may be better by an arbitrarily large factor.

Theorem 2.9 yields a non-trivial upper bound on $\|f-\mathrm{span}_n\mathcal{A}\|$ only if

$$\frac{\|f\|_{1,\mathcal{A}}}{2\sqrt{n-1}}\left(1-\frac{\|f\|_2^2}{\|f\|_{1,\mathcal{A}}^2}\right) < \|f\|_2.$$

This is equivalent to $\frac{\|f\|_{1,\mathcal{A}}}{\|f\|_2} < \sqrt{n} + \sqrt{n-1}$. Otherwise, the trivial upper bound that is equal to $\|f\|_2$ (which corresponds to the approximation of f by the constant zero function) is better.

The following theorem shows that these two bounds together, i.e. the minimum of $\|f\|_2$ and the bound from Theorem 2.9, yield a bound on $\|f - \mathrm{span}_n\mathcal{A}\|$ that is, up to a constant factor, the best possible upper bound expressed in terms of only $\|f\|_{1,\mathcal{A}}$ and $\|f\|_2$.

Theorem 2.10 *Let $(\mathcal{X}, \|.\|_2)$ be a finite dimensional Hilbert space, n be a positive integer and b, r be positive real numbers such that $r \leq b$ and $\max\{2n-1, b^2/r^2\} \leq \dim \mathcal{X}$. Then for every orthonormal basis $\mathcal{A}$ of $\mathcal{X}$ there exists $f \in \mathcal{X}$ such that $\|f\|_{1,\mathcal{A}} = b$, $\|f\|_2 = r$ and for every $f_n \in \mathrm{span}_n\mathcal{A}$*

$$\|f - f_n\|_2 \geq \frac{1}{2} \min \left\{ \frac{\|f\|_{1,\mathcal{A}}}{2\sqrt{n-1}} \left(1 - \frac{\|f\|_2^2}{\|f\|_{1,\mathcal{A}}^2} \right), \|f\|_2 \right\}.$$

3 Approximation of real-valued Boolean functions by perceptron networks

Using tools derived in the previous section, we will estimate rates of approximation of real functions of d Boolean variables by perceptron networks.

For a positive integer d denote by $\mathcal{B}(\{0,1\}^d)$ the *linear space of all real-valued functions of d Boolean variables.* It is easy to see that $\mathcal{B}(\{0,1\}^d)$ is isomorphic to $\mathcal{R}^{2^d}$. For any two functions $f, g \in \mathcal{B}(\{0,1\}^d)$, the standard Euclidean inner product is $f \cdot g = \sum_{x \in \{0,1\}^d} f(x)g(x)$ and $\|f\|_2 = \sqrt{f \cdot f}$.

We study representations and rates of approximation of functions from $\mathcal{B}(\{0,1\}^d)$ by functions computable by networks with a single linear output unit and one-hidden-layer containing perceptrons with signum activation function. *Signum* (defined by $\mathrm{sgn}(t) = -1$ for $t < 0$ and $\mathrm{sgn}(t) = 1$ for $t \geq 0$) can be obtained from more common Heaviside activation function ϑ (defined by $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geq 0$) using a simple linear transformation: $\mathrm{sgn}(t) = 2\vartheta(t) - 1$. Thus any function computable by a network with n Heaviside perceptrons can be computed by a network with $n + 1$ signum perceptrons.

We use signum for technical reasons: since the absolute value of sgn is a constant equal to 1, any function from $\mathcal{B}(\{0,1\}^d)$ computable by a perceptron with signum activation function, i.e. a function of the form $\mathrm{sgn}(v \cdot x + b)$, has the l_2-norm equal to $\sqrt{2^d}$, while the l_2-norm of a function computable by a perceptron with Heaviside activation function, i.e. a function of the form $\vartheta(v \cdot x + b)$, depends on the size of the half-space determined by the inequality $v \cdot x + b \geq 0$.

Let $\mathcal{H}_d$ denote the *set of functions from $\mathcal{B}(\{0,1\}^d)$ computable by* sgn *perceptrons*, i.e.

$$\mathcal{H}_d = \{f \in \mathcal{B}(\{0,1\}^d); (\exists v \in \mathcal{R}^d, b \in \mathcal{R})(\forall x \in \{0,1\}^d)(f(x) = \mathrm{sgn}(v \cdot x + b)\}.$$

Since $\vartheta(t) = \frac{\mathrm{sgn}(t)+1}{2}$ and $\mathrm{sgn}(t) = 2\vartheta(t) - 1$, it is easy to see that the variation of any f with respect to halfspaces (Heaviside perceptrons) is at least $V(f, \mathcal{H}_d)$ and at most $3V(f, \mathcal{H}_d)$.

3.1 Upper bounds on rates of approximation

Both Corollary 2.5 and Theorem 2.6 imply upper bounds on rates of approximation (measured in l_2 and l_∞-norm) of functions from $\mathcal{B}(\{0,1\}^d)$ by signum perceptron networks in terms of $\mathcal{H}_d$-variation.

Corollary 3.1 *For every positive integer d, for every $f \in \mathcal{B}(\{0,1\}^d)$ and for every positive integer n there exist functions $f_n, f'_n \in \mathrm{span}_n \mathcal{H}_d$ such that*

$$\|f - f_n\|_2 \leq \frac{V(f, \mathcal{H}_d^0)}{\sqrt{n}}$$

and

$$\|f - f'_n\|_\infty \leq \frac{V(f, \mathcal{H}_d^0)}{\sqrt{n}} \sqrt{\frac{(d+1) 2 \ln 2}{2^d}}.$$

To take advantage of this corollary we need to estimate $\mathcal{H}_d^0$-variation. Barron (1993) used Fourier representation to estimate variation with respect to half-spaces by a spectral norm for real domain functions and suggested its use also in the Boolean case. Here we derive an estimate of $\|f - \mathrm{span}_n \mathcal{H}_d\|$ based on spectral norm of f and for comparison also an estimate based on l_1-norm.

Let $\mathcal{F}_d$ denote the *Fourier orthonormal basis* of $\mathcal{B}(\{0,1\}^d)$ (see e.g. Weaver, 1983) defined by $\mathcal{F}_d = \left\{ \frac{1}{\sqrt{2^d}} (-1)^{u \cdot x}; u \in \{0,1\}^d \right\}$. Every $f \in \mathcal{B}(\{0,1\}^d)$ can be represented as $f(x) = \frac{1}{\sqrt{2^d}} \sum\limits_{u \in \{0,1\}^d} \tilde{f}(u)(-1)^{u \cdot x}$, where the Fourier coefficients $\tilde{f}(u)$ are given by the formula $\tilde{f}(u) = \frac{1}{\sqrt{2^d}} \sum\limits_{x \in \{0,1\}^d} f(x)(-1)^{u \cdot x}$. The l_1-norm with respect to Fourier basis, $\|f\|_{1,\mathcal{F}_d} = \|\tilde{f}\|_1 = \sum_{u \in \{0,1\}^d} |\tilde{f}(u)|$, is called the *spectral norm.*

The generalized parity functions are defined as follows. For a subset $I \subset \{0,1\}^d$, *I-parity* is defined by $p_I(u) = 1$ if $\sum_{i \in I} u_i$ is odd, and $p_I(u) = 0$ otherwise. Notice that the elements of the Fourier basis $\mathcal{F}_d$ are exactly the *generalized parity functions*, if we interpret the output 1 as -1 and 0 as 1.

It is easy to verify that every function from the Fourier basis $\mathcal{F}_d$ can be expressed as a linear combination of at most $d+1$ signum perceptrons. Indeed, it is easy to verify that for every $u, x \in \{0,1\}^d$ $(-1)^{u \cdot x} = \frac{1+(-1)^d}{2} + \sum_{j=1}^d (-1)^j \mathrm{sgn}(u \cdot x - j + \frac{1}{2})$. Moreover, any linear combination of n elements of $\mathcal{F}_d$ belongs to $\mathrm{span}_{dn+1} \mathcal{H}_d$, since all of the n occurrences of the constant function may be expressed by a single perceptron.

Combining this representation with Theorem 2.7, we obtain the following upper bound.

Corollary 3.2 *Let d be a positive integer and $f \in \mathcal{B}(\{0,1\}^d)$. Then for every positive integer there exists $f_{dn+1} \in \mathrm{span}_{dn+1} \mathcal{H}_d$ such that*

$$\|f - f_{dn+1}\|_2 \leq \frac{\|\tilde{f}\|_1}{2\sqrt{n}}.$$

Let $\mathcal{E}_d$ be the *Euclidean orthonormal basis* of $\mathcal{B}(\{0,1\}^d)$, i.e. $\mathcal{E}_d = \{e_u; u \in \{0,1\}^d\}$, where $e_u(u) = 1$ and for every $x \in \{0,1\}^d$ with $x \neq u$ $e_u(x) = 0$.

It is easy to verify that for any $u \in \{0,1\}^d$ $e_u(x)$ is expressible as $\frac{\mathrm{sgn}(v \cdot x + b) + 1}{2}$ for an appropriate v and b. Analogously as above, adding several occurrences of the constant function together, we obtain a representation of every linear combination of n functions of the Euclidean basis as an element of $\mathrm{span}_{n+1} \mathcal{H}_d$. This implies the following corollary of Theorem 2.7.

Corollary 3.3 *Let d be a positive integer and $f \in \mathcal{B}(\{0,1\}^d)$. Then for every positive integer n there exists $f_{n+1} \in \mathrm{span}_{n+1} \mathcal{H}_d$ such that*

$$\|f - f_{n+1}\|_2 \leq \frac{\|f\|_1}{2\sqrt{n}}.$$

Since it is usually easier to estimate the spectral norm of the function being approximated than its $\mathcal{H}_d$-variation, Corollary 3.2 gives a more feasible method of estimation than Corollary 3.1. However, the set of characteristic functions of half-spaces is much bigger than the set of parities and thus, for many functions, $\mathcal{H}_d$-variation might be considerably smaller than their spectral norm. For such functions upper bounds derived using Corollary 3.2 might be too large. In the next section, we will give examples illustrating the relationship between $\mathcal{H}_d^0$-variation and the spectral norm.

3.2 Polynomial and exponential upper bounds

In this section, we discuss the strength and weakness of the above described method of estimating rates of approximation of real-valued functions of several Boolean variables by perceptron networks using Fourier representations of functions from $\mathcal{H}_d$. For this purpose, it is natural to compare relative errors of approximation. Hence, we formulate our estimates in terms of normalized variation.

We describe functions for which the methods from the previous section give tight estimates as well as functions for which they do not give good results. Finally, we describe a set of functions with normalized $\mathcal{H}_d$-variation growing exponentially with d.

An easy example of a set of functions for which Corollary 3.2 guarantees small relative error of approximation is the set of linear combinations of a "small" number of generalized parities. Let $f = \sum_{i=1}^{k} w_i g_i$, where $g_i \in \mathcal{F}_d$. By the Cauchy inequality $\sum_{i=1}^{k} |w_i| \leq \sqrt{k \sum_{i=1}^{k} w_i^2}$ and hence $\frac{\|\tilde{f}\|_1}{\|f\|_2} \leq \sqrt{k}$. Thus for $n \geq \frac{k}{4\varepsilon^2}$ Corollary 3.2 guarantees approximation within a relative error not exceeding ε. Note however that the guaranteed error for these functions is far from the true approximation error, which is zero for $n \geq k$.

A more interesting example of functions for which Corollary 3.2 yields a non-trivial estimate of relative error of approximation are functions representable by decision trees of certain type.

Recall that a *decision tree* that represents a function $f : \{0,1\}^d \to \mathcal{R}$ is a binary tree with labeled nodes and edges. Every internal node is labeled by one of the variables $x_1, \dots, x_d$, and two outgoing edges are labeled by 0 and 1. The leaves of the tree are labeled by real numbers. The computation starts at the root. If the computation reaches an internal node labeled by x_i, then the computation continues along the edge, whose label coincides with the actual value of the variable x_i. Finally, a leaf is reached. Its label determines the value of the function. The *size* of a decision tree is the number of its leaves. The following theorem extends a result of Kushilevicz and Mansour (1991) (Lemma 5.1).

Theorem 3.4 *Let d, s be positive integers, $f \in \mathcal{B}(\{0,1\}^d)$ be expressible by a decision tree of size s such that for all $x \in \{0,1\}^d$ $f(x) \neq 0$. Then*

$$\frac{\|\tilde{f}\|_1}{\|f\|_2} \leq s \frac{\max_{x\in\{0,1\}^d} |f(x)|}{\min_{x\in\{0,1\}^d} |f(x)|}.$$

Thus for $n \geq \left(\frac{s \ \max_{x\in\{0,1\}^d} |f(x)|}{2\varepsilon \ \min_{x\in\{0,1\}^d} |f(x)|} \right)^2$ Corollary 3.2 guarantees relative error of approximation at most ε. Hence for a function representable by a decision tree with both the size and the ratio of the maximum and the minimum value of $|f(x)|$ for any $x \in \{0,1\}^d$ bounded by a polynomial in d we can achieve approximation within relative error ε by network with a polynomial number of hidden units.

To show limitations of the method of deriving estimates of rates of approximation by perceptron networks via an orthonormal approximating set, we will describe sets of functions for which normalized $\mathcal{H}_d$-variation grows linearly with d, while both normalized $\mathcal{F}_d$-variation and $\mathcal{E}_d$-variation grow exponentially.

Recall that a *bent function* is a function $f \in \mathcal{B}(\{0,1\}^d)$ such that for all $x, u \in \{0,1\}^d$ $|f(x)| = 1$ and $|\tilde{f}(u)| = 1$. For any bent function f, we have $V(f^0, \mathcal{F}_d) = \frac{\|\tilde{f}\|_1}{\|f\|_2} = \sqrt{2^d}$. Hence, relative approximation error $\frac{\|f - \mathrm{span}_n \mathcal{H}_d\|}{\|f\|_2}$ less than 1 is guaranteed by Corollary 3.2 only if the number n grows exponentially with d. Since $V(f^0, \mathcal{E}_d) = \frac{\|f\|_1}{\|f\|_2} = \sqrt{2^d}$, we do not obtain a good approximation error using the Euclidean basis as well. Moreover, the approximation error for f using $\mathcal{F}_d$ or $\mathcal{E}_d$ is indeed large. For any bent function f and any $h \in \mathcal{B}(\{0,1\}^d)$ such that $h \in \mathrm{span}_n \mathcal{E}_d$ or $h \in \mathrm{span}_n \mathcal{F}_d$ we have $\frac{\|f-h\|_2}{\|f\|_2} \geq \sqrt{1 - \frac{n}{2^d}}$, which is close to 1 unless n is exponential.

In order to demonstrate an example of a bent function with small normalized $\mathcal{H}_d$-variation, we use symmetric bent functions. Recall that a function $f \in \mathcal{B}(\{0,1\}^d)$ is called *symmetric* if it does not depend on the order of input variables; more precisely, for every $x, y \in \{0,1\}^d$ $w(x) = w(y)$ implies $f(x) = f(y)$, where $w(x)$ is the weight of x defined by $w(x) = \sum_{i=1}^d x_i$.

Any symmetric function f can be represented by a vector $(c_0, c_1, \ldots, c_d) \in \mathcal{R}^{d+1}$ in such a way that for every $x \in \{0,1\}^d$ $f(x) = c_{w(x)}$. Hence f can be represented by a linear combination of functions computable by signum perceptrons as $f(x) = \frac{c_d + c_0}{2} + \sum_{j=1}^d \frac{c_j - c_{j-1}}{2} \mathrm{sgn}(\sum_{i=1}^d x_i - j + 1/2)$. It follows that $f \in \mathrm{span}_{d+1} \mathcal{H}_d$. Moreover, if $|c_j| = 1$ for all $j = 0, 1, \ldots, d$, we have $V(f^0, \mathcal{H}^0) = V(f, \mathcal{H}_d) \leq d + 1$.

Examples of symmetric bent functions were given by Bruck (1990); Savický (1994) gave a complete characterization of symmetric bent functions. An example of a function of this type is $\phi_d : \{0,1\}^d \to \{-1,1\}$ defined for d even by $\phi_d(x) = -1$ if $w(x) \equiv 0 \pmod 4$ or $w(x) \equiv 1 \pmod 4$, and $\phi_d(x) = 1$ otherwise.

Since normalized $\mathcal{F}_d$-variation (spectral norm) of any bent function grows exponentially, symmetric bent functions are examples of functions for which upper bounds from Corollary 3.2 differ from upper bounds from Corollary 3.1 exponentially.

Finally, we will describe functions for which normalized $\mathcal{H}_d$-variation grows exponentially with d. Since all the values of the functions are 1 or -1, their $\mathcal{H}_d$-variation and their normalized $\mathcal{H}_d$-variation coincide.

For every $a, b \in \{0,1\}^{d/2}$, where d is even, let $a * b$ denote their *concatenation*. Then for every even positive integer d every vector $x \in \{0,1\}^d$ can be represented in a unique way as $x = x_l * x_r$, where $x_l, x_r \in \{0,1\}^{d/2}$.

Recall that a *communication matrix of a function* $f : \{0,1\}^d \to \{-1,1\}$ is $2^{d/2}$ by $2^{d/2}$ matrix M_f with rows and columns indexed by vectors $a, b \in \{0,1\}^{d/2}$ in such a way that $M_f(a,b) = f(a * b)$.

A matrix M with entries from $\{-1,1\}$ is called *Hadamard matrix* if every two distinct columns (or equivalently rows) of M are orthogonal.

An example of a function with a Hadamard communication matrix is the "inner product mod 2" $\beta_d : \{0,1\} \to \{-1,1\}$ defined by $\beta_d(x_l * x_r) = (-1)^{x_l \cdot x_r}$ for all $x = x_l * x_r \in \{0,1\}^d$.

The proof of the fact that the $\mathcal{H}_d$-variation of any function of d Boolean variables with a Hadamard communication matrix grows exponentially is based on Lindsey's lemma that estimates the difference between the number of 1s and -1s in submatrices of a Hadamard matrix.

Lemma 3.5 (Lindsey) *Let n be a positive integer and let M be an n by n Hadamard matrix. Let A, B be subsets of the set of indices of rows, columns, resp., of M. Then*

$$\left| \sum_{a \in A} \sum_{b \in B} M(a,b) \right| \leq \sqrt{n \, card\, A \; card\, B}.$$

Hajnal et al. (1987) used a special case of this lemma to derive an exponential lower bound on the number of hidden units needed to compute β_d by a one-hidden-layer perceptron network with a single output Heaviside perceptron with all the weights between hidden units and the output unit being *integers bounded by a polynomial in d.* The following lemma is based on inspection of their proof.

Lemma 3.6 (Hajnal et al.) *Let d be an even positive integer and $f : \{0,1\}^d \to \{-1,1\}$ be a function with Hadamard communication matrix M_f. Then for every $g \in \mathcal{H}_d$ $|f \cdot g| \leq \mathcal{O}(2^{5d/6})$.*

Combining this lemma with the geometric characterization of variation (Theorem 2.2) we get an exponential lower bound on $\mathcal{H}_d$-variation for functions with Hadamard communication matrix.

Theorem 3.7 *Let d be an even positive integer and $f : \{0,1\}^d \to \{-1,1\}$ be a function with Hadamard communication matrix M_f. Then $V(f, \mathcal{H}_d) \geq \Omega(2^{d/6})$.*

4 Discussion

In this paper, we considered two exact representations (obtained using the Euclidean and the Fourier bases) of real-valued Boolean functions of d variables as functions computable by one-hidden-layer perceptron networks. Since both of these representations require networks with the number of hidden units growing exponentially with the number of variables d, we examined the effect of the reduction of the number of hidden units upon decrease of accuracy. We estimated rates of approximation in terms of various norms of the function to be approximated, namely standard l_1, l_2-norms and variation with respect to a set of functions. To illustrate the strength and weakness of our estimates we gave examples of functions with these norms growing both exponentially and polynomially.

It is an open question whether one-hidden-layer perceptron networks with a single linear output unit can represent the "inner product mod 2" using less than an exponential number of hidden units. It is also not known whether this function can be approximated within a small error by a network with linear output unit with arbitrary real weights and with n Heaviside perceptrons for n bounded by a polynomial in d. Since Theorem 3.7 gives an exponential lower bound on an upper bound given in Corollary 3.1, our results show that this question cannot be solved using Jones-Barron theorem.

5 Proofs

Proof of Theorem 2.2.

Let $b = V(f, \mathcal{G})$. It is easy to see that $f \in \operatorname{cl}\operatorname{conv}\mathcal{G}(b)$. Hence, for every $\varepsilon > 0$ there exists f_ε such that $\|f - f_\varepsilon\| \leq \varepsilon$ and $f_\varepsilon = \sum_{i=1}^m w_i g_i$, where $\sum_{i=1}^m |w_i| \leq b$ and $g_i \in \mathcal{G}$ for every $i = 1, \dots, m$. For every $h \in S$ we have $|f \cdot h| - \varepsilon \leq |f_\varepsilon \cdot h| = |\sum_{i=1}^m w_i g_i \cdot h| \leq \sum_{i=1}^m |w_i||g_i \cdot h| \leq b \sup_{g \in \mathcal{G}} |g \cdot h|$. Since this holds for every $\varepsilon > 0$, we have $b \geq \sup_{h \in S} \frac{|f \cdot h|}{\sup_{g \in \mathcal{G}} |g.h|}$.

For the other direction it is sufficient to show that for every $0 < b < V(f, \mathcal{G})$ there exists $h \in S$ such that $\frac{|f \cdot h|}{\sup_{g \in \mathcal{G}} |g \cdot h|} \geq b$. Since $f \notin \operatorname{cl}\operatorname{conv}\mathcal{G}(b)$, f can be separated from $\operatorname{cl}\operatorname{conv}\mathcal{G}(b)$ by a hyperplane (see e.g. Holmes, 1975). It follows that there exists $h \in S$ such that for every $f' \in \operatorname{cl}\operatorname{conv}\mathcal{G}(b)$ we have $f \cdot h > f' \cdot h$. Since $\mathcal{G}(b)$ is closed under multiplication by -1, we have for every $g \in \mathcal{G}$ $f \cdot h > b\,|g \cdot h|$. Hence $\frac{|f \cdot h|}{|g \cdot h|} \geq b$. □

Proof of Lemma 2.4.

Let us prove the lemma by induction with respect to n. If $n \leq m$, the conclusion is trivial. For the induction step, it is sufficient to verify that any linear combination of $n > m$ coefficients may be replaced by a linear combination with at most $n-1$ nonzero coefficients satisfying the requirements of the lemma.

Assume that $f = \sum_{i=1}^{n} \alpha_i g_i$ for $n > m$, where $\alpha_i \neq 0$ for all $i = 1, \ldots, n$. Since $n > \dim \mathcal{X}$, there are numbers β_i such that $\sum_{i=1}^{n} \beta_i g_i = 0$ and $\beta_i \neq 0$ for at least one i. Then, for every t, we have $f = \sum_{i=1}^{n}(\alpha_i - t\beta_i) g_i$. Moreover, the function $\phi(t) = \sum_{i=1}^{n} |\alpha_i - t\beta_i|$ is a piecewise linear function. It is easy to verify that there exists i_0 such that $\beta_{i_0} \neq 0$ and ϕ achieves its global minimum at $t_0 = \alpha_{i_0}/\beta_{i_0}$. This implies that f is a linear combination of $\{g_i; i \in \{1, \ldots, n\} - \{i_0\}\}$ with the sum of absolute values of the coefficients equal to $\phi(t_0) \leq \phi(0) = \sum_{i=1}^{n} |\alpha_i|$. □

Proof of Theorem 2.7.

Let $\mathcal{A} = \{g_1, \ldots, g_m\}$, let $f \in \mathcal{X}$ and let $b = \|f\|_{1,\mathcal{A}}$. Using the same trick as Mhaskar and Micchelli (1994), we assume, without loss of generality, that $f = \sum_{i=1}^{m} w_i g_i$, where $|w_1| \geq |w_2| \geq \ldots \geq |w_m|$ (otherwise we reorder the basis $\mathcal{A}$). Let $h_n = \sum_{i=1}^{n} w_i g_i$. Then, $\|f - h_n\|_2^2 = \sum_{i=n+1}^{m} w_i^2 \leq |w_{n+1}| \sum_{i=n+1}^{m} |w_i|$. Moreover, $\sum_{i=n+1}^{m} |w_i| \leq b - n|w_{n+1}|$. Denoting $t = |w_{n+1}|$, we obtain $\|f - h_n\|_2^2 \leq t(b - nt)$. The last expression achieves its maximum for $t = b/(2n)$. Since the maximum is equal to $b^2/(4n)$, the required estimate follows. □

Proof of Theorem 2.8.

Let $f = b/(2n) \sum_{i=1}^{2n} g_i$. It is easy to see that $\|f\|_{1,\mathcal{A}} = b$ and for every $h \in \operatorname{span}_n \mathcal{A}$, $\|f - h\|_2^2 \geq b^2/(4n)$. □

Proof of Theorem 2.9.

Let $\mathcal{A} = \{g_1, \ldots, g_m\}$ and let $f \in \mathcal{X}$. Assume, without loss of generality, that $f = \sum_{i=1}^{m} w_i g_i$, where $|w_1| \geq |w_2| \geq \ldots \geq |w_m|$ (otherwise we reorder the basis $\mathcal{A}$). Let $f' = \sum_{i=2}^{m} w_i g_i$ and $\mathcal{A}' = \{g_2, \ldots, g_m\}$. By Theorem 2.7, there exists a function $h' \in \operatorname{span}_{n-1} \mathcal{A}'$ such that $\|f' - h'\|_2 \leq \frac{\|f'\|_{1,\mathcal{A}'}}{2\sqrt{n-1}} = \frac{\|f\|_{1,\mathcal{A}} - |w_1|}{2\sqrt{n-1}}$. Since $\|f\|_2^2 = \sum_{i=1}^{m} w_i^2 \leq |w_1| \|f\|_{1,\mathcal{A}}$, we have $|w_1| \geq \|f\|_2^2 / \|f\|_{1,\mathcal{A}}$. Denoting $h = w_1 g_1 + h' \in \operatorname{span}_n \mathcal{A}$ and using the fact that $\|f - h\|_2 = \|f' - h'\|_2$, we obtain the theorem. □

Lemma 5.1 *Let $b \geq r > 0$, let $b^2/r^2 \leq k \leq \dim \mathcal{X}$ and let $1 \leq n \leq k$. Then, there exists a function $f \in \mathcal{X}$ such that $\|f\|_1 = b$, $\|f\|_2 = r$ and for every $f_n \in \operatorname{span}_n \mathcal{A}$, we have*

$$\|f - f_n\|_2 \geq \frac{\sqrt{k-n}}{k} \left(b - \sqrt{\frac{kr^2 - b^2}{k-1}} \right).$$

Proof.

Let $\mathcal{A} = \{g_1, \ldots, g_m\}$ and let $k \leq m$. Let

$$f = \left(b + (k-1)\sqrt{\frac{kr^2 - b^2}{k-1}} \right) \frac{g_1}{k} + \left(b - \sqrt{\frac{kr^2 - b^2}{k-1}} \right) \frac{g_2 + \ldots + g_k}{k}.$$

By a direct calculation one can verify that $\|f\|_1 = b$ and $\|f\|_2 = r$. Moreover, it is easy to see that if h is a linear combination of at most n elements of $\mathcal{A}$, then at least $k - n$ coordinates of $f - h$ coincide with coordinates of f and hence

$$\|f - h\|_2^2 \geq (k-n) \frac{1}{k^2} \left(b - \sqrt{\frac{kr^2 - b^2}{k-1}} \right)^2.$$

□

Proof of Theorem 2.10.

First, assume that $b^2/r^2 \leq 2n-1$. The lower bound from Lemma 5.1 may be equivalently formulated as

$$\|f-h\|_2 \geq \frac{\sqrt{k-n}}{k-1}\,\frac{b^2-r^2}{b+\sqrt{\frac{kr^2-b^2}{k-1}}}.$$

Since $(kr^2-b^2)/(k-1) \leq b^2$, this implies that

$$\|f-h\|_2 \geq \frac{\sqrt{k-n}}{k-1}\left(\frac{b^2-r^2}{2b}\right).$$

Substituting $k = 2n-1$, we obtain

$$\|f-h\|_2 \geq \frac{b}{4\sqrt{n-1}}\left(1-\frac{r^2}{b^2}\right).$$

Now, let $b^2/r^2 > 2n-1$. The lower bound from Lemma 5.1 may be equivalently formulated as

$$\|f-h\|_2 \geq \sqrt{\frac{k-n}{k}}\left(\frac{b/r}{\sqrt{k}} - \sqrt{\frac{k-b^2/r^2}{k(k-1)}}\right)r.$$

Assume that $b^2/r^2 \geq n$. Let $k = \lceil b^2/r^2 \rceil$. Since $b^2/r^2 \geq k-1$, we obtain

$$\|f-h\|_2 \geq \sqrt{1-\frac{n}{k}}\left(\sqrt{\frac{k-1}{k}} - \frac{1}{\sqrt{k(k-1)}}\right)r \geq (1-\frac{2}{k})r\sqrt{1-\frac{n}{k}}.$$

Since $b^2/r^2 > 2n-1$, we have $k \geq 2n$. Hence, for every $h \in \mathrm{span}_n\mathcal{A}$, we have

$$\|f-h\|_2 \geq \frac{1}{\sqrt{2}}(1-1/n)r \geq r/2.$$

□

Proof of Theorem 3.4.

Denote the leaves of the tree by $v_1,\ldots,v_s$ and for $j=1,\ldots,s$ let D_j be the set of inputs $x \in \{0,1\}^d$ for which the computation reaches the leaf v_j and let $val(v_j)$ be the value of f assigned to v_j. Using the notation D_j also for the characteristic function of the set D_j, we obtain $f = \sum_{j=1}^s val(v_j)D_j$. It follows that $\tilde{f} = \sum_{j=1}^s val(v_j)\tilde{D}_j$.

By the definition of the Fourier transform, we have for any u

$$\tilde{D}_j(u) = \frac{1}{\sqrt{2^d}} \sum_{x\in\{0,1\}^d} (-1)^{u\cdot x}$$

In order to calculate $\tilde{D}_j(u)$ for an arbitrary $u \in \{0,1\}^d$, consider the restriction of the function $(-1)^{u\cdot x}$ to the set D_j. Let us distinguish two possibilities. First, assume that there is a variable x_i that is not tested on the path from the root of the tree to v_j and, moreover, $u_i = 1$. Then the Boolean vectors in D_j may be partitioned into set of pairs in such a way that the vectors in each pair differ only in the i-th coordinate. Since $u_i = 1$, the sum of the values of the function $(-1)^{u\cdot x}$ at elements of any of the pairs is zero. It follows that $\tilde{D}_j(u) = 0$.

Second, assume that for all $i = 1, \ldots, d$, if $u_i = 1$, then x_i is tested on the path to v_j and hence x_i has the same value for all $x \in D_j$. This means that $(-1)^{u \cdot x}$ is constant on D_j. In other words, $|\tilde{D}_j(u)| = card\, D_j / \sqrt{2^d}$.

Let the number of variables tested on the path to v_j be ℓ_j. Then, $card\, D_j = 2^{d-\ell_j}$ and there are exactly 2^{ℓ_j} vectors u for which the second possibility takes place. For these vectors u, we have $|\tilde{D}_j(u)| = card\, D_j / \sqrt{2^d} = 2^{d-\ell_j}/\sqrt{2^d}$. Altogether, $\|\tilde{D}_j\|_1 = \sum_u |\tilde{D}_j(u)| = 2^{\ell_j} 2^{d-\ell_j}/\sqrt{2^d} = \sqrt{2^d}$.

By combining the previous paragraph with the fact that for all $j = 1, \ldots, s$ we have $\tilde{f} = \sum_{j=1}^{s} val(v_j)\tilde{D}_j$, we obtain $\|\tilde{f}\|_1 \leq \max_x |f(x)| s \sqrt{2^d}$. On the other hand, $\|f\|_2 \geq \min_x |f(x)| \sqrt{2^d}$. Combining these two estimates, the theorem follows. □

Proof of Lemma 3.6.

Consider the communication matrices M_f and M_g. The function g is expressible as $g = \mathrm{sgn}(h_1 + h_2 + b)$, where h_1, h_2, resp., is a weighted sum of the first, second, resp., $d/2$ variables and b is a real number. Consider any ordering of $2^{d/2}$ assignments to the first $d/2$ variables in which the value of h_1 does not decrease. Reorder the rows in both M_f and M_g according to this ordering. Analogously, reorder the columns according to an ordering of the assignments of the second $d/2$ variables in which h_2 is non-decreasing. After this reordering, each row and also each column of M_g starts with a (possibly empty) initial segment of -1's followed by a (possibly empty) segment of 1's.

Consider a matrix M^* defined by $M^*(a, b) = M_f(a, b)M_g(a, b)$ (where M_f and M_g are reordered as described above). It is easy to verify that $f \cdot g = \sum_{a,b \in 2^{d/2}} M^*(a, b)$. Let $k = \lceil d/3 \rceil$ and consider a partition of the matrix M^* into $2^{d/2-k}$ times $2^{d/2-k}$ square submatrices of size 2^k by 2^k. Also, decompose both matrices M_f and M_g into submatrices in the same way.

Call a submatrix of M^* a positive, negative, resp., submatrix if all the corresponding entries in M_g are $1, -1$, resp. Call the submatrices that are neither positive nor negative mixed submatrices. Let us consider separately the contribution of positive, negative and mixed submatrices into the absolute value of the inner product $f \cdot g$. Because of the monotone structure of M_g, there are at most $2 \cdot 2^{d/2-k}$ mixed submatrices. Hence, the contribution of the mixed matrices is at most $2 \cdot 2^{d/2-k} \cdot 2^{2k}$. In order to estimate the contribution of negative submatrices, we use the Lindsey's lemma. To this end, we combine all negative submatrices in one column into one rectangle of size 2^k times at most $2^{d/2}$. The contribution of all of these rectangles together is at most $2^{d/2-k}\sqrt{2^{d/2} \cdot 2^k \cdot 2^{d/2}}$. Similarly, we get the same bound for the positive submatrices. Altogether, we have $|f \cdot g| \leq 2(2^{d/2+k} + 2^{d-k/2})$. Since $k = \lceil d/3 \rceil$, we obtain $|f \cdot g| = \mathcal{O}(2^{5d/6})$. □

References

Alon, N., & Spencer J. H. (1992). The Probabilistic Method. New Yourk: John Wiley.

Barron, A. R. (1992). Neural net approximation. In *Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems* (pp. 69–72).

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

Bruck, J. (1990). Harmonic analysis of polynomial threshold functions. *SIAM Journal Discrete Mathematics*, **3**, 168–177.

Bruck, J., & Smolensky, R. (1992). Polynomial threshold functions, AC^0 functions and spectral norms. *SIAM Journal of Computing*, **21**, 33–42.

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).

Girosi, F., & Anzellotti, G. (1993). Rates of convergence for radial basis function and neural networks. In *Artificial Neural Networks for Speech and Vision* (pp.97–113). London: Chapman & Hall.

Hajnal, A., Maas, W., Pudlák, P., Szegedy, M., & Turán, G. (1987). Threshold circuits of Bounded depth. In *Proceedings of the 28th Annual Symposium on Foundation of Computer Science* (pp. 99–110).

Holmes, R. B. (1975). Geometrical Functional Analysis and Its Applications, New York: Springer.

Ito, Y. (1992). Finite mapping by neural networks and truth functions. *Mathematical Scientist*, **17**, 69–77.

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.

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

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.

Kushilevicz, E., Mansour, Y. (1991). Learning decision trees using the Fourier spectrum. In *Proceedings of 23rd ACM STOC* (pp. 455-464). Montreal: ACM Press.

Leshno M., Lin V. Y, Pinkus A. & Schocken S.(1993). Multilayer feedforward networks with a non-polynomial activation can approximate any function. *Neural Networks*, **6**, 861–867.

Mhaskar H. N., & Micchelli, C. A. (1992). Approximation by superposition of sigmoidal and radial basis functions. *Advances in Applied Mathematics*, **13**, 350–373.

Mhaskar, H. N., & Micchelli, C. A. (1994). Dimension-independent bounds on the degree of approximation by neural networks. *IBM Journal of Research and Development*, **38**, 277–284.

Park J., & Sandberg, I. W. (1993). Approximation and radial-basis-function networks. *Neural Computation*, **5**, 305–316.

Savický, P. (1994). On the bent Boolean functions that are symmetric. *Europ. J. Combinatorics*, **15**, 407–410.

Sejnowski, T. J., & Rosenberg, C. (1987). Parallel networks that learn to pronounce English text. *Complex Systems*, **1**, 145–168.

Siu, K. Y., & Bruck, J. (1991). On the power of threshold circuits with small weights. *SIAM Journal Discrete Mathematics*, **4**, 423–435.

Weaver, H. J. (1983). Applications of discrete and continuous Fourier analysis. New York: John Wiley.