ESANN'1996 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-25-26 April 1996, D-Facto public., ISBN 2-9600049-6-5, pp. 167-172

Rates of approximation of real-valued boolean functions by neural networks

Kateřina Hlaváčková, Věra Kůrková

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, 182 00, Prague 8, Czech Republic
fax: +42 2 85 85 789, e-mail: katka@uivt.cas.cz *

Abstract. We investigate the approximation of real-valued functions of d boolean variables by one-hidden-layer perceptron networks. We show that each function $f : \{0,1\}^d \to \mathcal{R}$ can be approximated within an error ε by a network having $\lceil \frac{(2d+1)^2 H}{\varepsilon^2} \rceil$ perceptrons with any sigmoidal activation function, where $H > B_f^2 - \|f\|^2$ and B_f is a constant which depends on the Fourier transform of f. We derive a rate of approximation for $f : \{0,1\}^d \to [0,1]$ with a finite support that is only quadratical in d.

1 Introduction

In recent years, the approximation of functions of several real variables by feed-forward neural networks has been widely studied. The existence of an arbitrarily close approximation of any continuous or $\mathcal{L}_p$ function has been proved for perceptron type and radial-basis-function networks with quite general activation and kernel functions (see e.g. Mhaskar and Micchelli [5] and Park and Sandberg [6]). However, estimates of the number of hidden units that guarantee a given approximation error have remained less understood. Most upper bounds of this number grow exponentially with the number of input units (i.e. the dimension d of the input space). Jones [3] introduced a recursive construction of approximants with "dimension-independent" rates of convergence to functions in convex closures of bounded subsets of a Hilbert space and together with Barron proposed to apply it to sets of functions computable by one-hidden-layer neural networks. Applying Jones' estimate, several authors (e.g., Barron [1], Girosi and Anzellotti [2], Kůrková et al. [4]) characterized sets of functions of d real variables that can be approximated within an error of $\mathcal{O}(\frac{1}{\sqrt{n}})$ by networks with n hidden units of various types (perceptron or radial-basis-function).

In some applications, input data are represented using only binary values. When computational units used in the hidden layer are continuous sigmoidal perceptrons, the input/output functions of such network is a real-valued function of several boolean variables. A typical example of such an application is Sejnowski and Rosenberg's NETtalk [7], where a real-valued function of about two hundred boolean variables is approximated sufficiently well by a neural network with only moderately many hidden units.

* This work was partly supported by GACR grant 201/93/0427 and 201/96/0917.

ESANN'1996 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-25-26 April 1996, D-Facto public., ISBN 2-9600049-6-5, pp. 167-172

Motivated by these experimental results we investigate the approximation of real-valued functions of d boolean variables by one-hidden-layer perceptron-type networks. Extending Jones' theorem [3] to finite dimensional vector spaces, we characterize sets of boolean functions for which the approximation error of order $\mathcal{O}(\frac{1}{\sqrt{n}})$ is achievable by networks with n hidden units.

We extend Barron's [1] estimate of approximation error for networks with trigonometric perceptrons to more general activation functions (which includes all continuous sigmoidals and the Heaviside discontinuous threshold function). Following Barron's technique based on discrete Fourier transform we show that each function $f : \{0,1\}^d \to \mathcal{R}$ can be approximated by a network having $\lceil \frac{(2d+1)^2 H}{\varepsilon^2} \rceil$ sigmoidal perceptrons in the hidden layer within the error ε, where $H > B_f^2 - \|f\|^2$ and $B_f = \sum_{u \in \{0,1\}^d} |\tilde{f}(u)|$ and $\tilde{f}(u)$ are Fourier coefficients.

For functions with $B_f \leq B$ for some fixed value B the number of hidden units needed to guarantee approximation within ε is of order $\mathcal{O}(\frac{1}{\varepsilon^2})$. However with d increasing, the condition $B_f \leq B$ becomes increasingly restrictive. To illustrate this restriction, we show that B_f is bounded by the size of the support of f. In this case, the number of hidden units in a perceptron-type network approximating function f depends on the dimension d only quadratically.

2 The universal approximation property

$\mathcal{R}$ denotes the set of real numbers. For a positive integer d and $X \subseteq \mathcal{R}^d$ the set of all real-valued functions on X is denoted by $\mathcal{F}(X)$. In this paper we will consider the set $\mathcal{F}(\{0,1\}^d)$ of all boolean real-valued functions and the set $\mathcal{F}(\{j\pi; j = 0, \ldots, d\})$ of all real-valued functions on the discrete 1-dimensional set $\{j\pi; j = 0, \ldots, d\}$. For two vectors $v, x \in \mathcal{R}^d$ $v \cdot x$ denotes the standard Euclidean inner product of v and x. For any function $\psi : \mathcal{R} \to \mathcal{R}$ we define

$$\mathcal{P}_d(\psi) = \{f \in \mathcal{F}(\{0,1\}^d); f(x) = \sum_{i=1}^{m} w_i \psi(v_i \cdot x + b_i),\ w_i, b_i \in \mathcal{R}\ \&\ v_i \in \mathcal{R}^d\}.$$

So $\mathcal{P}_d(\psi)$ is the set of all real-valued functions of d boolean variables that can be computed by a one-hidden layer network with ψ-perceptrons and with a single linear output unit. Since $\{0,1\}^d$ as a subspace of $\mathcal{R}^d$ is discrete and compact and each function on a discrete topological space is continuous, all functions in $\mathcal{F}(\{0,1\}^d)$ are continuous. Thus the results which prove that sets of functions, computable by neural networks are dense in spaces of continuous functions on compact subsets of $\mathcal{R}^d$, can be employed to derive the "universal approximation property" for boolean real-valued functions. The following theorem is a direct corollary of a result by Mhaskar and Micchelli [5].

Theorem 2.1 *Let $\psi : \mathcal{R} \to \mathcal{R}$ be a non-polynomial function which is locally Riemann integrable. Then for every positive integer d, every function $f \in \mathcal{F}(\{0,1\}^d)$ and every $\varepsilon > 0$ there exists a function $g \in \mathcal{P}_d(\psi)$ such that for all $x \in \{0,1\}^d$ $|f(x) - g(x)| < \varepsilon$.*

ESANN'1996 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-25-26 April 1996, D-Facto public., ISBN 2-9600049-6-5, pp. 167-172

3 Jones Theorem

For $f, g \in \mathcal{F}(\{0,1\}^d)$ denote $< f, g >= \sum_{x \in \{0,1\}^d} f(x)g(x)$. It is easy to check that $< f, g >$ is an inner product and that $(\mathcal{F}(\{0,1\}^d), <>)$ is isomorphic to $\mathcal{R}^{2^d}$. Since the space $\mathcal{F}(\{0,1\}^d)$ is finite dimensional, in order to estimate the approximation error for boolean real-valued functions, we need to extend the result proved by Jones [3] for approximation of functions in Hilbert spaces also to the finite dimensional vector spaces. Let F be a real vector space with a norm $\|.\|$ generated by an inner product. For a subset $G \subseteq F$ denote by *cl conv* G the closure of the convex hull of G, where closure is taken with respect to the topology generated by the norm $\|.\|$. We denote by $\mathcal{N}$ the set of positive integers. The following theorem can be easily verified by inspection of the proof of theorem by Jones [3].

Theorem 3.1 (Jones) *Let F be a real vector space with a norm $\|.\|$ generated by an inner product on F, $B > 0$ and $G \subseteq F$, such that for every $g \in G$ $\|g\| \leq B$. Then for every function $f \in$ cl conv G, every $H > B^2 - \|f\|^2$ and for every $n \in \mathcal{N}$ there exists f_n in the convex hull of n members of G such that*

$$\|f - f_n\| \leq \sqrt{\frac{H}{n}}.$$

In the remainder of the paper $\|.\|$ denotes the norm generated by the inner product $<>$ and *cl* denotes the closure with respect to the topology generated by this norm. We will apply Theorem 3.1 to sets of functions computable by neural networks with one ψ-perceptron with an output weight bounded in absolute value by some bound B:

$$\mathcal{G}_d(\psi, B) = \{f \in \mathcal{F}(\{0,1\}^d); f(x) = w\psi(v \cdot x + b), w, b \in \mathcal{R}, v \in \mathcal{R}^d, |w| \leq B\},$$

where B is a positive constant and $\psi : \mathcal{R} \to \mathcal{R}$ is an activation function.

It is easy to verify that *conv* $\mathcal{G}_d(\psi, B) = \mathcal{P}_d(\psi, B)$, where

$$\mathcal{P}_d(\psi, B) = \{f \in \mathcal{F}(\{0,1\}^d);$$

$$f(x) = \sum_{i=1}^{m} w_i \psi(v_i.x + b_i); w_i, b_i \in \mathcal{R}, v_i \in \mathcal{R}^d, \sum_{i=1}^{m} |w_i| \leq B\}.$$

So $\mathcal{P}_d(\psi, B)$ is the set of real-valued functions of d boolean variables that can be computed by one-hidden-layer ψ-perceptron networks with the sum of absolute values of output weights bounded by B. In order to derive estimates of rates of approximation of real-valued boolean functions by one-hidden-layer networks from Jones' theorem we need a characterization of sets *cl* $\mathcal{P}_d(\psi, B)$ for various activation functions ψ.

ESANN'1996 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-25-26 April 1996, D-Facto public., ISBN 2-9600049-6-5, pp. 167-172

4 Rates of approximation of boolean real-valued functions by one-hidden-layer perceptron networks

First, we will extend an estimate obtained by Barron [1] for the case that $\psi = \cos$. Recall that a Fourier representation of a complex-valued function f on $\{0,1\}^d$ can be written as (see e.g [8, p.91] or [1]) $f(x) = \sum_{u\in\{0,1\}^d} e^{i\pi u\cdot x}\tilde{f}(u)$, where the Fourier coefficients are $\tilde{f}(u) = \frac{1}{2^d}\sum_{x\in\{0,1\}^d} e^{-i\pi u\cdot x} f(x)$.
For real-valued f we have $f(x) = \sum_{u\in\{0,1\}^d} \cos(\pi u\cdot x)\tilde{f}(u)$, where $\tilde{f}(u) = \frac{1}{2^d}\sum_{x\in\{0,1\}^d} \cos(\pi u\cdot x) f(x)$. Let $B_f = \sum_{u\in\{0,1\}^d} |\tilde{f}(u)|$.

To formulate a condition on an activation function ψ which is able to describe a bound B for a boolean function f for which $f \in cl\ conv\ \mathcal{G}_d(\psi, B) = cl\ \mathcal{P}_d(\psi, B)$ we introduce the following notation:

$$\mathcal{T}_d(\psi, C) = \{f \in \mathcal{F}(\{j\pi; j = 0,\dots,d\});$$
$$f(t) = \sum_{i=1}^{m} w_i\psi(v_i t + b_i), w_i, v_i, b_i \in \mathcal{R}, \sum_{i=1}^{m} |w_i| \le C\},$$

where C is a positive constant and ψ is an activation function. By $cl_{sup}\mathcal{T}_d(\psi, d)$ we denote the closure with respect to supremum (maximum) norm.

Proposition 4.1 *Let C be a positive real number, d a positive integer and $\psi : \mathcal{R} \to \mathcal{R}$ be any function for which there exists $\hat{\psi} \in cl_{sup}\mathcal{T}_d(\psi, C)$ such that for all $j = 0,\dots,d$ $\hat{\psi}(j\pi) = (-1)^j$. Then for every $f \in \mathcal{F}(\{0,1\}^d)$ $f \in cl\ \mathcal{P}_d(\psi, CB_f)$.*

Proof: For $\varepsilon > 0$ put $\delta = \frac{\varepsilon}{B_f}$. Since $\hat{\psi} \in cl_{sup}\mathcal{T}_d(\psi, C)$ there exists $m \in \mathcal{N}$, $w_i, v_i, b_i \in \mathcal{R}$ $(i = 1,\dots,m)$ such that for every $j = 0,\dots,d$ $|\hat{\psi}(j\pi) - \sum_{i=1}^{m} w_i\psi(v_i j\pi + b_i)| < \delta$. Note that for every $u, x \in \{0,1\}^d$ $u.x \in \{j\pi; j = 0,\dots,d\}$. So we have

$$\left|f(x) - \sum_{u\in\{0,1\}^d} \tilde{f}(u)\sum_{i=1}^{m} w_i\psi(v_i\pi u\cdot x + b_i)\right| =$$
$$\left|\sum_{x\in\{0,1\}^d} \tilde{f}(u)\cos(\pi u\cdot x) - \sum_{u\in\{0,1\}^d} \tilde{f}(u)\sum_{i=1}^{m} w_i\psi(v_i\pi u\cdot x + b_i)\right| =$$
$$\left|\sum_{u\in\{0,1\}^d} \tilde{f}(u)\left(\hat{\psi}(\pi u\cdot x) - \sum_{i=1}^{m}\psi(v_i\pi u\cdot x + b_i)\right)\right| < B_f\delta = \varepsilon.$$

Since $\sum_{u\in\{0,1\}^d} |\tilde{f}(u)| \sum_{i=1}^{m} |w_i| \le CB_f$, we have $f \in cl\ \mathcal{P}_d(\psi, CB_f)$. □

Recall that a function $\sigma : \mathcal{R} \to \mathcal{R}$ is called *sigmoidal* if $\lim_{t\to-\infty}\sigma(t) = 0, \lim_{t\to\infty}\sigma(t) = 1$ and $\sigma(\mathcal{R}) \subseteq [0,1]$ and that the *Heaviside function* $\vartheta : \mathcal{R} \to \mathcal{R}$ is the function satisfying $\vartheta(t) = 0$, for t < 0 and $\vartheta(t) = 1$ for t $\ge$ 0.

Theorem 4.2 *Let $\sigma : \mathcal{R} \to \mathcal{R}$ be any sigmoidal function and d be a positive integer. Then for every $f \in \mathcal{F}(\{0,1\}^d)$ $f \in cl\ \mathcal{P}_d(\sigma, (2d+1)B_f)$.*

ESANN'1996 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-25-26 April 1996, D-Facto public., ISBN 2-9600049-6-5, pp. 167-172

Proof: Put

$$\hat{\vartheta}(t) = 1 + 2\sum_{j=0}^{d-1}(-1)^{j+1}\vartheta\left(t - (2j+1)\frac{\pi}{2}\right).$$

It is easy to verify that for every $j = 0, \ldots, d$ $\hat{\vartheta}(j\pi) = (-1)^j$. For a sigmoidal function σ define $\sigma_r(t) = \sigma(rt)$. Put $\hat{\sigma_r}(t) = 1+2\sum_{j=0}^{d-1}(-1)^{j+1}\sigma_r(t-(2j+1)\frac{\pi}{2})$. It is easy to verify that for every $t \in \{j\pi; j = 0, \ldots, d\}$ $\lim_{r\to\infty}\hat{\sigma_r}(t) = \hat{\vartheta}(t)$. Since for every $r \in \mathcal{N}$ $\hat{\sigma_r} \in \mathcal{T}_d(\sigma, 2d+1)$ we have $\hat{\vartheta} \in cl_{sup}\mathcal{T}_d(\sigma, 2d+1)$.

So Proposition 4.1 implies that $f \in cl\ \mathcal{P}_d(\sigma, (2d+1)B_f)$. □

The following corollary is an immediate consequence of Theorems 3.1 and 4.2.

Corollary 4.3 *Let $\sigma : \mathcal{R} \to \mathcal{R}$ be any sigmoidal function and d be a positive integer. Then for every $f \in \mathcal{F}(\{0,1\}^d)$, every $H > B^2 - \|f\|^2$ and for every positive integer n there exists a function f_n that is a convex combination of n functions from $\mathcal{P}_d(\sigma, (2d+1)B_f)$ such that $\|f - f_n\| \leq \sqrt{\frac{H}{n}}$.*

We can reformulate this statement in neurocomputing terminology as follows:

Corollary 4.4 *Let $\sigma : \mathcal{R} \to \mathcal{R}$ be any sigmoidal function and d be a positive integer. Then for every $f \in \mathcal{F}(\{0,1\}^d)$ and for every $\varepsilon > 0$ there exists a function f_ε computable by a neural network with single linear output unit and $\lceil\frac{(2d+1)^2H}{\varepsilon^2}\rceil$ σ-perceptrons in the hidden layer such that $\|f - f_\varepsilon\| \leq \varepsilon$, where $H > B_f^2 - \|f\|^2$.*

So to estimate the rate of approximation of a given function $f \in \mathcal{F}(\{0,1\}^d)$ we need to find a bound on B_f. The following proposition gives such a bound in terms of the size of support. For a function $f \in \mathcal{F}(\{0,1\}^d)$ we denote by A_f its support, i.e. $A_f = \{x \in \{0,1\}^d; f(x) \neq 0\}$.

Proposition 4.5 *For every positive integer d and for every $f : \{0,1\}^d \to [0,1]$ $B_f \leq$ card A_f.*

Proof: We will proceed by induction. It is easy to check that when *card* $A_f = 0$ then $B_f = 0$, and when *card* $A_f = 1$ then $B_f = 1$.

Suppose that the proposition holds for k. Let $A_f = \{x_1, \ldots, x_{k+1}\}$. Denote $A'_f = A_f - \{x_{k+1}\}$ and let $f' : \{0,1\}^d \to [0,1]$ be such that $f(x) = f'(x)$ for all $x \in A'_f$ and $f'(x_{k+1}) = 0$. By our assumption $B_{f'} \leq k$. Let $u \in \{0,1\}^d$ be arbitrary. Suppose first that $u \cdot x_{k+1}$ is an even number. Then $\tilde{f}(u) = \frac{1}{2^d}\left(\sum_{x\in\{0,1\}^d - \{x_{k+1}\}} cos(\pi u \cdot x)f(x) + f(x_{k+1})\right) =$
$= \frac{1}{2^d}\left(\sum_{x\in\{0,1\}^d} cos(\pi u \cdot x)f'(x) + f(x_{k+1})\right) \leq \frac{k}{2^d} + \frac{1}{2^d} \leq \frac{k+1}{2^d}$. Let $u \cdot x_{k+1}$ be odd. Then $\tilde{f}(u) = \frac{1}{2^d}\left(\sum_{x\in\{0,1\}^d} cos(\pi u \cdot x)f'(x) - f(x_{k+1})\right) \leq \frac{k}{2^d} - \frac{f(x_{k+1})}{2^d} \leq \frac{k+1}{2^d}$. Thus $B_f \leq k+1$. □

ESANN'1996 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-25-26 April 1996, D-Facto public., ISBN 2-9600049-6-5, pp. 167-172

Corollary 4.6 *Let d be a positive integer, σ a sigmoidal function. Let $\mathcal{B}(k) = \{f : \{0,1\}^d \to [0,1]; card\ A_f \leq k\}$. Then for every $\varepsilon > 0$ there exists a function f_ε computable by a neural network with single linear output unit and $\lceil \frac{(2d+1)^2 H}{\varepsilon^2} \rceil$ σ-perceptrons in the hidden layer such that $\|f - f_\varepsilon\| \leq \varepsilon$, where $H > k^2 - \|f\|^2$.*

Note that the rate of approximation for the functions in $\mathcal{B}(k)$ depends only quadratically on the dimension d.

5 Conclusion

We derived an estimate of rates of approximation of real-valued functions of d boolean variables by one-hidden-layer perceptron networks with any sigmoidal activation function. We showed that the error of approximation of a function f achievable by a network with n hidden units is bounded from above by $\frac{(2d+1)H}{\sqrt{n}}$, where B_f depends on the Fourier transform of f. This bound can be only called "dimension-independent" when we restrict ourselves to classes of functions having B_f bounded by a fixed B for all d. We gave an example of such class - functions with support bounded by a fixed integer k.

References

1. Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* 39, 930-945, 1993.
2. Girosi, F., Anzelotti, G.: Rates of convergence for radial basis functions and neural networks. In *Artificial Neural Networks for Speech and Vision* (pp. 97-113). London:Chapman & Hall, 1993.
3. Jones, L.K.: A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Annals of Statistics* 20, 601-613, 1992.
4. Kůrková, V., Kainen, P.C., Kreinovich, V.: Dimension-independent rates of approximation by neural networks and variation with respect to half-spaces. In *Proceedings of WCNN'95* (pp. I. 54-57). INNS Press, 1995.
5. Mhaskar, H.N. , Micchelli, C.A.: Approximation by superposition of a sigmoidal function. *Advances in Applied Mathematics* 13, 350-373 (1992).
6. Park, J., Sandberg, I.W.: Universal approximation using radial-basis-function networks. *Neural Computation* 3, 246-257, 1991.
7. Sejnowski, T.J., Rosenberg, C:: Parallel networks that learn to pronounce English text. *Complex Systems* 1, 145-168, 1987.
8. Weaver, H.J.: *Applications of discrete and continuous Fourier analysis.* New York: John Willey, 1983.