Trade-off Between the Size of Weights and the Number of Hidden Units in Feedforward Networks

Kůrková, Věra

1996

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Trade-off between the size of weights and the number of hidden units in feedforward networks

Věra Kůrková

Technical report No. V-695

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic
phone: (+442) 66 05 32 31   fax: (+442) 8585789
e-mail: vera@uivt.cas.cz

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Trade-off between the size of weights and the number of hidden units in feedforward networks

Věra Kůrková [1]

Technical report No. V-695

Abstract

Abstract. We examine the effect of compensating a constraint on the number of hidden units in feedforward networks by increasing the size of their parameters. We describe functions that can be approximated with any accuracy by only changing parameters in perceptron type and radial-basis-function networks while the number of hidden units remains fixed. We show that unless the activation function satisfies a special type of recursion, only linear combinations of the functions exactly computable by such networks and their iterated partial derivatives can be approximated in this way.

Keywords

Approximation of functions, one-hidden-layer neural networks, sigmoidal perceptrons, radial-basis-functions.

[1]This work was by GA AV grant A2030602, GA CR grant 201/96/0917 and KBN grant 8T11A02311.

1 Introduction

Although neural networks of many types can approximate continuous or $\mathcal{L}_p$-functions (see e.g., [?], [?]), as the accuracy of approximation increases one may require an arbitrarily large number of hidden units and the size of the network parameters may also grow without bound. Thus, complexity of a network can be measured either by the number of hidden units or by the size of its parameters.

The question of whether this "universal approximation property" can be achieved even with bounded parameters was answered by Stinchcombe and White [?] for bounds depending on certain characteristics of the activation function and extended to arbitrarily small bounds by Hornik [?]. Hence, a constraint on the size of parameters can be compensated by an increase of the number of hidden units.

The complementary task is to characterize functions that can be approximated with any accuracy by varying only parameters in networks with a *fixed* number of hidden units. The first example illustrating the trade-off between the size of weights and the number of hidden units was given by Girosi and Poggio [?]. They gave an example of a function that can be approximated with any accuracy by changing parameters (increasing output weights and decreasing one bias) in a network with only two hidden sigmoidal perceptrons.

Besides being useful for comparison of complexity of networks measured by the number of hidden units and the size of weights, characterizing functions that can be approximated with any accuracy by networks with a fixed number of units can also be used to compare the approximation capabilities of networks of different types: either one-hidden-layer networks with different types of units or networks with the same types of hidden units but different numbers of layers.

The first step in this direction was taken by Chui et al. [?] who compared capabilities of one and two-hidden-layer Heaviside perceptron networks. They proved that, while characteristic functions of d-dimensional cubes for $d \geq 2$ are exactly computable by two-hidden-layer networks with $2d$ units in the first hidden layer and 1 unit in the second one, such characteristic functions cannot be approximated arbitrarily well by one-hidden-layer networks with a *bounded* number of hidden units. In [?], we extended their results by showing that sets of functions computable by one-hidden-layer networks with a constrained number of perceptrons with the Heaviside activation function are closed in $\mathcal{L}_p$-spaces.

Gori et al. [?] listed several examples of functions that can be approximated arbitrarily well by networks with fixed number of perceptrons with various activation functions. In the case of functions of one variable and inverse tangent activation they gave a complete characterization of the set of such functions.

In this paper, we characterize sets of multivariable functions that can be approximated with any accuracy by networks with a constrained number of hidden units. If a hidden unit function, as well as the ratio between change of output weights and hidden units parameters, are "reasonable", we show that the only functions that can be approximated by networks with a fixed number of units are linear combinations of iterated partial derivatives of the hidden unit function with respect to its parameters. We estimate complexity and rates of approximation of such functions. To illustrate what kind of functions can be achieved in this way for standard networks, linear combinations of iterated partial derivatives of perceptrons with hyperbolic tangent activation or of Gaussian radial-basis-functions are characterized.

The paper is organized as follows. In section 2, we recall basic concepts and results concerning the universal approximation property. Next in section 3, we introduce a new concept of complexity for a function with respect to a class of neural networks and estimate this complexity of iterated partial derivatives of a smooth hidden unit function with respect to its parameters. In section 4, we show that if we eliminate cases when we cannot infer something about a limit of a sequence of functions computable by networks with a fixed number of hidden units, then we only obtain the iterated partial derivative functions described in the preceding section. In section 5 we characterize functions that can be approximated arbitrarily well by networks with a fixed number of hyperbolic tangent perceptrons or with Gaussian radial-basis-functions. Section 6 is a brief discussion. All proofs are deferred to section 7.

2 The universal and the best approximation property

Let $\mathcal{R}$ denotes the set of real numbers, $\mathcal{N}$ the set of natural numbers and $\mathcal{N}_+$ the set of positive integers. In this paper we examine approximation of continuous functions by one-hidden-layer networks with a single linear output unit. Such networks compute functions of the form $\sum_{i=1}^{m} w_i \phi(\mathbf{y}_i, \mathbf{x})$, where $m \in \mathcal{N}_+$ corresponds to the number of hidden units, $w_i \in \mathcal{R}$, $i = 1, \ldots, m$, to *output weights* and $\phi : \mathcal{R}^{p+d} \to \mathcal{R}$ to the type of hidden units with $\mathbf{y}_i \in \mathcal{R}^p$ representing their parameters and $\mathbf{x} \in \mathcal{R}^d$ input vectors. We call such networks *ϕ-networks*.

For example, for *perceptrons* with an activation function $\psi : \mathcal{R} \to \mathcal{R}$ the number of parameters p equals to $d+1$ and $\phi(\mathbf{v}, b, \mathbf{x}) = P_\psi(\mathbf{v}, b, \mathbf{x}) = \psi(\mathbf{v} \cdot \mathbf{x} + b)$, where $\mathbf{v} \in \mathcal{R}^d$ is an *input weight* vector and $b \in \mathcal{R}$ is a *bias*. For *radial-basis-function* (RBF) units with a radial (even) function $\psi : \mathcal{R} \to \mathcal{R}$ $\phi(\mathbf{v}, b, \mathbf{x}) = B_\psi(\mathbf{v}, b, \mathbf{x}) = \psi(b\|\mathbf{x} - \mathbf{v}\|)$, where $\mathbf{v} \in \mathcal{R}^d$ is a *centroid*, $b \in \mathcal{R}$, $b > 0$, is a *width* and $\|.\|$ denotes the Euclidean norm on $\mathcal{R}^d$.

For $A \subseteq \mathcal{R}^d$, a function $\phi : \mathcal{R}^{p+d} \to \mathcal{R}$ representing a type of a computational unit, m positive integer and $B > 0$ we denote by $\mathcal{F}(\phi, A, m, B)$ the set of functions on A computable by ϕ-networks with at most m hidden units with all the network parameters bounded by B. Thus, $\mathcal{F}(\phi, A, m, B)$ denotes the set of all functions from A to $\mathcal{R}$ of the form $\sum_{i=1}^{m} w_i \phi(\mathbf{y}_i, \mathbf{x})$, where $w_i \in \mathcal{R}$ and $\mathbf{y}_i \in \mathcal{R}^p$ such that for all $i = 1, \ldots, m$ $\|\mathbf{y}_i\| \leq B$.

When either m or B or both m and B are not bounded we will use notation $\mathcal{F}(\phi, A, *, B)$, $\mathcal{F}(\phi, A, m, *)$ or $\mathcal{F}(\phi, A, *, *)$, resp. We will abbreviate $\mathcal{F}(\phi, A, *, *)$ by $\mathcal{F}(\phi, A)$.

Standard choices for a perceptron activation function include the Heaviside function ϑ satisfying $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geq 0$ and the hyperbolic tangent τ which is affinely equivalent to the logistic sigmoid $\lambda(t) = \frac{1}{1+\exp(-t)}$. The standard choice for a radial function is the Gaussian, denoted γ, with $\gamma(t) = \exp(-t^2)$.

Capabilities of networks to approximate functions are studied mathematically in terms of closures and dense subspaces; see, e.g. [?] for the basic definitions and theorems. For $A \subseteq \mathcal{R}^d$ we denote by $\mathcal{C}(A)$ the set of all continuous functions on A with the topology of uniform convergence. For $X \subseteq A$ we denote by $cl(X)$ the closure of X in this topology.

For any locally Riemann-integrable non-polynomial activation function ψ, for any positive integer d and any compact $A \subset \mathcal{R}^d$, the set $\mathcal{F}(P_\psi, A)$ is known to be dense in $\mathcal{C}(A)$, i.e. $cl(\mathcal{F}(P_\psi, A)) = \mathcal{C}(A)$
(see e.g. [?]). The set $\mathcal{F}(B_\psi, A)$ is dense in $\mathcal{C}(A)$ for any continuous function ψ with finite non-zero integral, any positive integer d and any compact $A \subset \mathcal{R}^d$ (see [?], [?]). In neural networks terminology this capability is called the *universal approximation property*.

Hornik [?] proved that for any analytic non-polynomial activation function ψ the universal approximation property can be achieved even using networks with parameters constrained by an arbitrarily small bound. More precisely, for any $B > 0$ $cl(\mathcal{F}(P_\psi, A, *, B)) = \mathcal{C}(A)$. Of course, the constraint on parameters has to be compensated by an increase of the number of hidden units.

In practical situations, the number of hidden units is bounded by some fixed positive integer. In addition, the parameters are also bounded. Under these conditions, we showed in [?] that for many types of feedforward networks, given any continuous function, there is a choice of network parameterization (not necessarily unique) producing an approximation with the minimum error. We call this the *best approximation property*. In fact we showed that for A compact such function spaces are compact too, which in particular implies that $\mathcal{F}(P_\psi, A, m, B)$ and $\mathcal{F}(B_\psi, A, m, B)$ are closed for any bounded continuous ψ. Hence no function that is not already contained in $\mathcal{F}(P_\psi, A, m, B)$ or $\mathcal{F}(B_\psi, A, m, B)$, resp., can be approximated with any accuracy by networks with bounds on both the size of parameters and the number of hidden units.

This suggests the question how quickly such best approximation error decreases as a function of either the number of hidden units or the size of parameters. Recently, dependence of the approximation error on the number of hidden units has became better understood. Following Jones [?] and Barron [?], several authors (e.g., [?], [?], [?]) characterized sets of functions that can be approximated by one-hidden-layer neural networks with "dimension-independent" rates of approximation, i.e. for which the number of hidden units needed for a given accuracy does not grow exponentially with the number of variables of the function to be approximated. However, these results estimate the number of hidden units without any constraint on the size of weights.

3 Iterated partial derivatives with respect to network parameters

Thus, the only case that remains to be investigated is the case when the number of hidden units is bounded. Functions that can be approximated arbitrarily well by such networks are in the closures of the sets $\mathcal{F}(\phi, A, m, *)$.

Each class of neural networks having the universal approximation property determines a hierarchy on $\mathcal{C}(A)$ ordered by complexity defined as the minimal number of hidden units needed for an arbitrarily close approximation of a given function. For $f \in \mathcal{C}(A)$ define ϕ-complexity

$$\nu(f, \phi, A) = \min\{m \in \mathcal{N}_+; f \in cl(\mathcal{F}(\phi, A, m, *))\}$$

if the set over which the minimum is taken is non-empty; otherwise set $\nu(f, \phi, A) = \infty$.

Girosi and Poggio's [?] gave an example of a low-complexity function with respect to sigmoidal perceptron networks, based on the equality

$$\frac{1}{2(1+\cosh(x))} = \lim_{n\to\infty} n\left(\frac{1}{1+\exp(-x)} - \frac{1}{1+\exp(-x-\frac{1}{n})}\right) = \lim_{n\to\infty} n\left(\lambda(x) - \lambda(x+\frac{1}{n})\right).$$

Hence the function $\frac{1}{1+\cosh(x)}$ can be approximated with any accuracy by a network with only two perceptrons having the logistic sigmoid λ as an activation function. One can easily verify that the convergence is uniform on any compact $A \subset \mathcal{R}$. Thus, we have an upper estimate $\nu(\frac{1}{1+\cosh(x)}, P_\lambda, A) \leq 2$ for any compact $A \subset \mathcal{R}$.

Following Girosi and Poggio's method, we can find an analogous example for RBF networks. For instance, the function $x^2\gamma(x)$ can be approximated with any accuracy by a Gaussian RBF network with only two hidden units. Indeed, for every $x \in \mathcal{R}$

$$-2x^2\gamma(x) = \frac{\partial\gamma(b(x-c))}{\partial b}\Big|_{b=1,c=0} = \lim_{n\to\infty} n\left(\gamma((1+\frac{1}{n})x) - \gamma(x)\right).$$

Thus $\nu(x^2\gamma(x), B_\gamma, A) \leq 2$ for any compact $A \subset \mathcal{R}$.

The technique of Leshno et al.'s proof [?] of the universal approximation property of one-hidden-layer networks with perceptrons with any non-polynomial analytic activation function ψ is based on an observation that powers are "simple" functions with respect to such networks. Since $\frac{\partial^k\psi(vx+b)}{\partial b^k} = x^k\psi^{(k)}(vx+b)$ and there exists a real number b_k such that $\psi^{(k)}(b_k) \neq 0$ (ψ is non-polynomial) we can approximate the k-th power x^k with any accuracy by a network with $k+1$ hidden ψ-perceptrons (the k-th derivative of any function f can be approximated within any accuracy by a linear combination of the terms $f(x+jh)$, $j = 0, \dots, k$). Hence $\nu(x^k, P_\psi, A) \leq k+1$ for any compact $A \subset \mathcal{R}$.

To generalize these examples we need some notation. Let $\mathcal{C}^q(\mathcal{R}^p)$ denote the set of all functions on $\mathcal{R}^p$ for which all iterated partial derivatives of order at most q exist and are continuous and let $\mathcal{C}^\infty(\mathcal{R}^p)$ denote the set of functions with continuous partial derivatives of all orders. For a function $f \in \mathcal{C}^q(\mathcal{R}^p)$, $s \in \{1,\dots,q\}$ and $j \in \{1,\dots,p\}$ denote by $D_j^{(s)} f$ the partial derivative of order s with respect to the j-th variable, and for $s=0$ define $D_j^{(0)} f = f$. We will write D_j instead of $D_j^{(1)}$. For a multiindex $\mathbf{s} = (s_1,\dots,s_p) \in \mathcal{N}^p$ let $|\mathbf{s}| = \sum_{j=1}^p s_j$ and for a finite set P let $|P|$ denotes the number of its elements.

For $\phi \in \mathcal{C}^\infty(\mathcal{R}^{p+d})$, $s \in \mathcal{N}$, $r \in \mathcal{N}_+$ denote by $\mathcal{D}(\phi, A, r, s)$ the set of all functions $f : \mathcal{R}^d \to \mathcal{R}$ of the form

$$f(\mathbf{x}) = \sum_{i=1}^{m} \sum_{\mathbf{s} \in P_i} a_{i\mathbf{s}} D_1^{(s_1)} \dots D_p^{(s_p)} \phi(\mathbf{y}_i, \mathbf{x}),$$

where $m \in \mathcal{N}_+$, for every $i = 1,\dots,m$ $\mathbf{y}_i \in \mathcal{R}^p$, $P_i \subset {\mathcal{N}_+}^p$ is finite and $\sum_{i=1}^m |P_i| \le r$, for every $\mathbf{s} \in P_i$ $a_{i\mathbf{s}} \in \mathcal{R}$ and $|\mathbf{s}| \le s$. Thus, $\mathcal{D}(\phi, A, r, s)$ contains linear combinations of r functions obtained using partial differential operators of order at most s acting on $\phi(\mathbf{y}, \mathbf{x}) = \phi(y_1,\dots,y_p,x_1,\dots,x_d)$ with respect to the first p variables $y_1,\dots,y_p$. Note that since we allow $\mathbf{s} = \mathbf{0}$, $\mathcal{F}(\phi, A, m, *) \subseteq \mathcal{D}(\phi, A, m, 0)$. Let $\mathcal{D}(\phi, A) = \cup\{\mathcal{D}(\phi, A, r, s); s \in \mathcal{N}, r \in \mathcal{N}_+\}$.

It follows from the definition of a derivative that $D_1^{(s_1)} \dots D_p^{(s_p)}\phi$ is a limit of a linear combination of the translates of ϕ of the form $\phi(y_1 + \frac{i_1}{n}, \dots, y_p + \frac{i_p}{n}, \mathbf{x})$, where $j \in \{1,\dots,p\}$ and $i_j \in \{0,\dots,s_j\}$. Thus each iterated partial derivative of ϕ can be approximated arbitrarily well by functions computable by ϕ-networks with a $\Pi_{j=1}^p(s_j+1)$ hidden units. Using the mean value theorem we can verify by induction that this convergence is uniform on any compact $A \subset \mathcal{R}^d$. The number of terms in the linear combination corresponding to the number of hidden units depends polynomially on the sum of orders $|\mathbf{s}|$ and exponentially on the dimension of the parameter space.

Theorem 3.1 *Let d, p, r be positive integers, s be a non-negative integer, $\phi \in \mathcal{C}^\infty(\mathcal{R}^{p+d})$ and $A \subset \mathcal{R}^d$ be compact. Then $\mathcal{D}(\phi, A, r, s) \subseteq cl(\mathcal{F}(\phi, A, r(s+1)^p))$ and so for every $f \in \mathcal{D}(\phi, A, r, s)$ $\nu(f, \phi, A) \le r(s+1)^p$.*

Thus, if a hidden unit function, ϕ, is smooth, then the set of functions computable by ϕ-networks with a fixed number of hidden units is not closed; it contains linear combinations of iterated partial derivatives of ϕ with respect to its parameters. However, networks approximating an iterated partial derivative $D_1^{(s_1)} \dots D_p^{(s_p)}\phi$ have output weights growing with $\mathcal{O}(n^{|\mathbf{s}|})$ and differences between hidden unit parameters of order $\mathcal{O}(\frac{1}{n})$. Implementation of such networks might not be feasible for large n. On the other hand, if n is small enough to allow implementation, then the achievable approximation error could not be sufficiently accurate.

We can estimate this error using the following proposition. Recall that a modulus of continuity of a function $g : \mathcal{R}^p \to \mathcal{R}$ is a function $\omega_g : (0, \infty) \to \mathcal{R}$ defined by $\omega_g(\delta) = \sup\{|g(\mathbf{y} - \mathbf{y}')|; \mathbf{y}, \mathbf{y}' \in \mathcal{R} \& (\forall i = 1,\dots,p)(|y_i - {y'}_i| \le \delta\}$. By $\|.\|_\infty$ is denoted the supremum norm.

Proposition 3.2 *Let d, s, n be positive integers, $g \in \mathcal{C}^\infty(\mathcal{R})$ and $\Delta_n^{(s)} g(y) = n(g(y + \frac{1}{n}) - g(y))$ for every $y \in \mathcal{R}$. Then $\|\Delta_n^{(s)} g - D^{(s)} g\|_\infty \le \sum_{i=1}^s (2n)^{s-i+1} \omega_{D^{(j)} g}(\frac{1}{n})$.*

4 Limits of sequences of functions computable by networks with a fixed number of hidden units

In the previous section we have shown that, if the hidden unit function, ϕ, is smooth, all linear combinations of iterated partial derivatives of ϕ with respect to its parameters have finite complexity measured by the number of ϕ hidden units. For a complete characterization of such finite complexity functions we have to investigate closures of sets of functions computable by networks with a fixed number of hidden units. Recall that all elements in the closure of a set in the topology of uniform convergence are limits of sequences of elements of this set. Hence, we have to study for fixed m limits of the form

$$\lim_{n\to\infty} \sum_{i=1}^{m} w_{in}\phi(\mathbf{y}_{in}, \mathbf{x}). \tag{4.1}$$

We will show that for a uniformly continuous hidden unit function, ϕ, such a limit is either a linear combination of iterated partial derivatives of ϕ with respect to its parameters or else we cannot infer anything about it. There are two types of situations when we cannot infer anything about a limit of the form (1): the first one is caused by an "unbalanced" ratio between growth of sequences of output weights $\{w_{in}; n \in \mathcal{N}_+\}$ and hidden unit parameters $\{\mathbf{y}_{in}; n \in \mathcal{N}_+\}$, while the second one is caused by a property of ϕ.

For $A \subseteq \mathcal{R}^d$, call a function $\phi \in \mathcal{C}^\infty(\mathcal{R}^{p+d})$ *derivative recursive* on A if the constant zero function can be represented as a function from $\mathcal{D}(\phi, A)$ in a non-trivial way, i.e. the functional equation

$$\sum_{i=1}^{m} \left(\sum_{\mathbf{s}\in P_i} a_{i\mathbf{s}} D_1^{(s_1)} \dots D_p^{(s_p)} \phi(\mathbf{y}_i, \mathbf{x}) \right) = 0 \tag{4.2}$$

is satisfied on A, where m is a positive integer, for every $i = 1, \dots, m$ $\emptyset \neq P_i \subset \mathcal{N}^p$, for every $\mathbf{s} = (s_1, \dots, s_p) \in P_i$ $a_{i\mathbf{s}}$ is a non-zero real number, for all pairs $i, j \in \{1, \dots, m\}$ such that $i \neq j$ also $\mathbf{y}_i \neq \mathbf{y}_j$, and if ϕ is odd or even in $\mathbf{y}$ moreover $\mathbf{y}_i \neq -\mathbf{y}_j$.

Any smooth function ϕ satisfying the negation of this condition is "reasonable " in the sense that it does generate cases when we cannot infer anything about a limit of the form (1). We will show that when ϕ is not derivative recursive then all limits of functions computable by ϕ-networks with fixed number of hidden units with a "balanced" ratio between growth of their output weights and input parameters converge to linear combinations of iterated partial derivatives of ϕ.

We call a sequence of functions computable by networks with a single linear output unit and a fixed number m of ϕ hidden units *balanced* when for each hidden unit the sequence of its innner parameters $\{\mathbf{y}_{in}; n \in \mathcal{N}_+\}$ is convergent and the sequence of its output weights $\{w_{in}; n \in \mathcal{N}_+\}$ grows only polynomially with the decrease of the distance $\mathbf{y}_{in}$ from its limit value $\mathbf{y}_i$. More precisely for every $i = 1, \dots, m$ there exists $\mathbf{y}_i \in \mathcal{R}$ such that $\lim_{n\to\infty} \mathbf{y}_{in} = \mathbf{y}_i$, $\{w_{in}; n \in \mathcal{N}_+\}$ is either convergent or divergent and when it is divergent, then there exists $k_i \in \mathcal{N}$ such that the sequence $\{w_{in}\|\mathbf{y}_{in} - \mathbf{y}_i\|_\infty^{k_i}; n \in \mathcal{N}\}$ is convergent, where the subscript denotes the supremum norm on $\mathcal{R}^p$ and the superscript means raising to the k_i power. Denote by $\hat{\mathcal{F}}(\phi, A, m, *)$ the subset of $cl(\mathcal{F}(\phi, A, m, *))$ containing only limits of balanced sequences. Since each $f \in \mathcal{F}(\phi, A)$ is trivially a limit of a balanced sequence, we have $\mathcal{F}(\phi, A, m, *) \subseteq \hat{\mathcal{F}}(\phi, A, m, *)$.

The following theorem shows that if the hidden unit function ϕ is not derivative recursive, then the only functions among limits of balanced sequences that have finite complexity measured by the number of ϕ-hidden units are the functions described in the previous section. Its proof is based on the Taylor formula for multivariable functions.

Theorem 4.1 *Let p, d be positive integers, $\phi \in \mathcal{C}^\infty(\mathcal{R}^{p+d})$ be uniformly continuous, $A \subseteq \mathcal{R}^d$. If ϕ is not derivative recursive on A then for every positive integer m $\hat{\mathcal{F}}(\phi, A, m, *) \subseteq \mathcal{D}(\phi, A)$.*

Generally, verifying that a function is not derivative recursive is a difficult task which requires us to find some special properties of the function ϕ that contradict the functional equation (2). In [**?**] we studied a stronger condition requiring that the functional equation (2) is satisfied with all P_i containing only the zero vector, i.e.

$$\sum_{i=1}^{m} a_i \phi(\mathbf{y}_i, \mathbf{x}) = 0, \tag{4.3}$$

non-trivially (all $a_i \neq 0$, all the vectors $\{\mathbf{y}_i; i = 1, \ldots, m\}$ are distinct and when ϕ is either odd or even in $\mathbf{y}$ then also $\mathbf{y}_i \neq -\mathbf{y}_j$ for $i \neq j$). Note that the negation of this condition guarantees that an input/output function of a ϕ-network determines the network parameterization uniquely up to a permutation of hidden units (see [**?**], [**?**]).

When a function of one variable satisfies this special-case condition then it can be expressed as a linear combination of its scaled and translated copies in a non-trivial way. Such functions were called "affinely recursive"in [**?**]. For example, singularities of complex extensions can play such a role (see [**?**]). It is shown in [**?**] that many analytic functions cannot be affinely recursive. On the other hand, with the exception of polynomials, all examples of affinely recursive functions known to us (such as the Daubechies scaling function) are non-smooth. Since the Gaussian has no poles, we used its asymptotic properties in [**?**] to verify that B_γ does not satisfy the functional equation (3) on $\mathcal{R}^d$. However, here we need a weaker condition (2) allowing also iterated partial derivative terms.

5 Local and non-local hidden units

To describe functions having finite complexity measured by the number of hidden units of the most popular types – perceptrons with the hyperbolic tangent τ as an activation function and radial-basis-function units with the Gausssian function γ as a radial function – we need to characterize $\mathcal{D}(P_\tau, A)$ and $\mathcal{D}(B_\gamma, A)$. For a polynomial Q of several variables, let $deg(Q)$ denotes its degree, i.e. the maximum of exponents of all of its variables.

Theorem 5.1 *Let d, r be positive integers, s be a non-negative integer, $A \subseteq \mathcal{R}^d$. Then every $f \in \mathcal{D}(P_\tau, A, r, s)$ can be represented as $f(\mathbf{x}) = \sum_{i=1}^m \sum_{\mathbf{s} \in P_i} a_{i\mathbf{s}} {x_1}^{s_1} \dots {x_d}^{s_d} Q_{\mathbf{s}}(\tau(\mathbf{v}_i \cdot \mathbf{x} + b_i))$, where m is a positive integer, $\sum_{i=1}^m |P_i| \leq r$ and for every $\mathbf{s} \in \cup_{i=1}^m P_i$ $Q_{\mathbf{s}} : \mathcal{R} \to \mathcal{R}$ is a polynomial with $deg(Q_{\mathbf{s}}) \leq s + 1$.*

Theorem 5.2 *Let d, r be positive integers, s be a non-negative integer, $A \subseteq \mathcal{R}^d$. Then every $f \in \mathcal{D}(B_\gamma, A, r, s)$ can be represented as $f(\mathbf{x}) = \sum_{i=1}^m \gamma(b_i \|\mathbf{x} - \mathbf{v}_i\|) Q_i(b_i, \|\mathbf{x} - \mathbf{v}_i\|, v_{i1}, \dots, v_{id}, x_1, \dots, x_d)$, where m is a positive integer and for every $i = 1, \dots, m$ $Q_i : \mathcal{R}^{2(d+1)} \to \mathcal{R}$ is a polynomial with $deg(Q_i) \leq 2s$.*

For $d \geq 2$ perceptrons and radial-basis-function units are geometrically opposite: perceptrons apply a sigmoidal activation function to a weighted sum of inputs plus a bias and so correspond to *non-localized* regions of the input space by partitioning it with fuzzy hyperplanes (or sharp ones if the sigmoid is Heaviside's step-function), while RBF units calculate the distance between an input vector and a centroid, multiply by a scale-factor called width and then apply a radial function – hence corresponding to *localized* regions. Thus, perceptron type networks that compute linear combinations of ridge functions and RBF networks that compute linear combinations of radial functions should be efficient in approximating different types of functions. Note that for $d \geq 2$ the functions with low complexity with respect to hyperbolic tangent perceptron networks described in Theorem 5.1 contain no linear combinations of radial functions. On the other side, there are no linear combinations of ridge functions among low complexity functions with respect to Gaussian RBF described in Theorem 5.2.

6 Discussion

We have introduced a new concept of complexity determined for sets of continuous functions of several variables by classes of neural networks possessing the universal approximation property. We have shown that if a hidden unit function, ϕ, does not satisfy a special type of recursion and if we restrict to cases when the approximation has to be achieved using networks with a polynomial ratio between output weights and differences between hidden unit parameters, then the only functions having finite complexity are linear combinations of iterated partial derivatives of ϕ with respect to its parameters.

Although theoretically for these finite complexity functions we can compensate a constraint on the number of hidden units by increasing parameters, practically this method of approximation is limited by precision bounds that do not allow us to implement two rapidly diverging scales of parameters.

7 Proofs

Lemma 7.1 *Let d, p, q be positive integers, $\phi \in \mathcal{C}^q(\mathcal{R}^{p+d})$, $A \subset \mathcal{R}^d$ be compact, $j \in \{1, \ldots, d\}$ and $s \in \{1, \ldots, q\}$. Then for every $\mathbf{y} \in \mathcal{R}^p$ $D_j^{(s)}(\mathbf{y}, .) = \lim_{n\to\infty} \Delta_{njs}\phi(\mathbf{y}, .)$ uniformly on A, where $\Delta_{jn}^{(s)}\phi(\mathbf{y}, .) : \mathcal{R}^d \to \mathcal{R}$ is defined by $\Delta_{jn}^{(s)}\phi(\mathbf{y}, \mathbf{x}) = n^s \Big(\sum_{i=0}^{s} (-1)^{i+1} \binom{s}{i} \phi(y_1, \ldots, y_j$*

Proof
Without loss of generality assume that $j = 1$. First, we will verify the statement for $s = 1$: Let U be a compact neighborhood of $\mathbf{y}$. Then both ϕ and $D_1^{(1)} = D_1\phi$ are uniformly continuous on $A \times U$. Hence for any $\varepsilon > 0$ there exists $\delta > 0$ such that for every $\mathbf{x}, \mathbf{x}' \in A$ with $\|\mathbf{x} - \mathbf{x}'\| < \delta$ and for every $\mathbf{y}' \in U$ with $\|\mathbf{y} - \mathbf{y}'\| < \delta$ we have $\mathbf{y}' \in U$, $|\phi(\mathbf{y}, \mathbf{x}) - \phi(\mathbf{y}', \mathbf{x}')| < \frac{\varepsilon}{3}$ and $\Big|D_1\phi(\mathbf{y}, \mathbf{x}) - D_1\phi(\mathbf{y}', \mathbf{x}')\Big| < \frac{\varepsilon}{3}$.

Compactness of A guarantees that the existence of a finite set $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \subseteq A$ such that for every $\mathbf{x} \in A$ there exists $i \in \{1, \ldots, k\}$ with $\|\mathbf{x} - \mathbf{x}_i\| < \delta$. Since $\lim_{n\to\infty} \Delta_{1n}^{(1)}\phi(\mathbf{y}, .) = D_1\phi(\mathbf{y}, .)$ on A pointwise, there exists n_0 such that $\frac{1}{n_0} < \delta$, $\frac{1}{n}$-neighbourhood of $\mathbf{y}$ is contained in U and for every $n \geq n_0$ and for every $i = 1, \ldots, k$ $|\Delta_{1n}^{(1)}\phi(\mathbf{y}, \mathbf{x}_i) - D_1\phi(\mathbf{y}, \mathbf{x}_i)| < \frac{\varepsilon}{3}$. Hence for every $\mathbf{x} \in A$

$$|\Delta_{1n}^{(1)}\phi(\mathbf{y}, \mathbf{x}) - D_1\phi(\mathbf{y}, \mathbf{x})| \leq$$

$$|\Delta_{1n}^{(1)}\phi(\mathbf{y}, \mathbf{x}) - \Delta_{1n}^{(1)}\phi(\mathbf{y}, \mathbf{x}_i)| + |\Delta_{1n}^{(1)}\phi(\mathbf{y}, \mathbf{x}_i) - D_1\phi(\mathbf{y}, \mathbf{x}_i)| + |D_1\phi(\mathbf{y}, \mathbf{x}_i) - D_1\phi(\mathbf{y}, \mathbf{x})| <$$

$$\frac{2\varepsilon}{3} + |\Delta_{1n}^{(1)}\phi(\mathbf{y}, \mathbf{x}) - \Delta_{1n}^{(1)}\phi(\mathbf{y}, \mathbf{x}_i)|.$$

By the mean value theorem for every $n \in \mathcal{N}_+$ there exist $y_{1n}, y_{1ni} \in [y_1, y_1 + \frac{1}{n}]$ such that $\Delta_{1n}^{(1)}\phi(\mathbf{y}, \mathbf{x}) = n(\phi(y_1 + \frac{1}{n}, y_2, \ldots, y_p, \mathbf{x}) - \phi(\mathbf{y}, \mathbf{x})) = D_1\phi(y_{1n}, y_2, \ldots, y_p, \mathbf{x})$ and $\Delta_{1n}^{(1)}\phi(\mathbf{y}, \mathbf{x}_i) = D_1\phi(y_{1ni}, y_2, \ldots, y_p, \mathbf{x}_i)$. Putting $\mathbf{y}_n = (y_{1n}, y_2, \ldots, y_p)$ and $\mathbf{y}_{ni} = (y_{1ni}, y_2, \ldots, y_p)$ we have $|\Delta_{1n}^{(1)}\phi(\mathbf{y}, \mathbf{x}) - \Delta_{1n}^{(1)}\phi(\mathbf{y}, \mathbf{x}_i)| = |D_1\phi(\mathbf{y}_n, \mathbf{x}) - D_1\phi(\mathbf{y}_{ni}, \mathbf{x}_i)| < \frac{\varepsilon}{3}$. Hence for every $\mathbf{x} \in X$ $|\Delta_{1n}^{(1)}\phi(\mathbf{y}, \mathbf{x}) - D_1\phi(\mathbf{y}, \mathbf{x})| < \varepsilon$.

Assume that the statement is true for s and that $D_1^{(s+1)}\phi$ is continuous. Then an analogous argument as in the first step shows that it is also true for $s + 1$. □

Proof of Theorem 3.1
Let $f(\mathbf{x}) = \sum_{i=1}^m \sum_{\mathbf{s}\in P_i} a_{i\mathbf{s}} D_1^{(s_1)} \ldots D_p^{(s_p)}\phi(\mathbf{y}_i, \mathbf{x})$, where for every $i = 1, \ldots, m$ $|P_i| \leq s$ and $\sum_{i=1}^m |P_i| \leq r$. Inspection of the proof of Lemma 7.1 shows that for each $i = 1, \ldots, m$ and $\mathbf{s} \in P_i$ $D_j^{(s_j)}\phi(\mathbf{y}_i, .)$ is a limit of a uniformly convergent sequence of functions from $\mathcal{F}(\phi, A, \Pi_{j=1}^p(s_j + 1), *)$. Thus a linear combination of r such functions is in $cl(\mathcal{F}(\phi, A, r\left(\Pi_{j=1}^p(s_j + 1)\right), *))$. $(s + 1)^p$ is an upper bound on $\Pi_{j=1}^p(s_j + 1)$ satisfying $\sum_{j=1}^p s_j = s$. □

Proof of Proposition 3.2
It follows from the mean value theorem that for every $n \in \mathcal{N}_+$ $\|\Delta_n^{(1)}g - D^{(1)}g\|_\infty \leq$

$\omega_{D^{(1)}g}(\frac{1}{n})$. Asume that the statement is true for s then $\|\Delta_n^{(s+1)}g - D^{(s+1)}g\|_\infty \leq$ $\|\Delta_n^{(1)}D^{(s)}g - D^{(1)}D^{(s)}g\|_\infty + \|\Delta_n^{(1)}\Delta_n^{(s)}g - \Delta_n^{(1)}D^{(s)}g\|_\infty$ $2n\omega_{D^{(s)}g}(\frac{1}{n}) + \omega_{D^{(s+1)}g}(\frac{1}{n})$ and so it is also true for $s+1$. □

Recall that a directional derivative of order k of a function g of p variables in the direction of a vector $\mathbf{h} \in \mathcal{R}^d$ is defined by

$$D_{\mathbf{h}}^{(k)}g(\mathbf{y}) = \sum_{\mathbf{s}\in P_k} \begin{pmatrix} k \\ s_1 \dots s_d \end{pmatrix} h_1^{s_1} \dots h_d^{s_d} D_1^{(s_1)} \dots D_d^{(s_p)} g(\mathbf{y}), \tag{7.1}$$

where $P_k = \{\mathbf{s} \in \{0, \dots, k\}^p; |\mathbf{s}| = k\}$. For $\mathbf{a}, \mathbf{b} \in \mathcal{R}^p$ we denote by $L(\mathbf{a}, \mathbf{b})$ the *line segment* connecting $\mathbf{a}$ and $\mathbf{b}$.

Lemma 7.2 *Let d, p be positive integers, $A \subseteq \mathcal{R}^d$, $\phi \in \mathcal{C}^\infty(\mathcal{R}^{p+d})$ be derivative recursive on A and let $\{f_n(\mathbf{x}) = \sum_{i=1}^m w_{in}\,(\phi(\mathbf{y}_i + \mathbf{h}_{in}, \mathbf{x}))\,; n \in \mathcal{N}_+\}$ be a balanced sequence converging on A pointwise to a function $f : A \to \mathcal{R}$. Then $f \in \mathcal{D}(\phi, A)$.*

Proof

For $k, n \in \mathcal{N}_+$, $\mathbf{s} \in P_k$ and $i \in \{1, \dots, m\}$ let $H_{in\mathbf{s}} = \frac{1}{k!}\begin{pmatrix} k \\ s_1 \dots s_p \end{pmatrix} h_{in1}^{s_1} \dots h_{inp}^{s_p}$. Let $\sim$ denotes an equivalence on $\{1, \dots, m\}$ defined in the case that ϕ is either odd or even with respect to $\mathbf{y}$ by $i \sim j$ if $\mathbf{y}_i = \pm\mathbf{y}_j$, and when ϕ is neither odd nor even by $i \sim j$ if $\mathbf{y}_i = \mathbf{y}_j$. Let J be a subset of $\{1, \dots, m\}$ containing exactly one representative of each class of $\sim$. For every $i \in J$ let $J_i = \{j \in \{1, \dots, m\}; j \sim i\}$, $J_{i+} = \{j \in \{1, \dots, m\}; \mathbf{y}_i = \mathbf{y}_j\}$ and $J_{i-} = \{j \in \{1, \dots, m\}; \mathbf{y}_i = -\mathbf{y}_j\}$. When ϕ that is neither odd nor even define $\hat{w}_{in} = \sum_{j\in J_i} w_{jn}$, while when ϕ is either odd or even define $\hat{w}_{in} = \sum_{j\in J_{i+}} w_{jn} - \sum_{j\in J_{i-}} w_{jn}$. For $i \in J$ let $K_i = \{k \in \mathcal{N}; (\forall \mathbf{s} \in P_k)(\{\hat{w}_{in}H_{in\mathbf{s}}; n \in \mathcal{M}\}$ is convergent $)\}$.

Since $\{f_n; n \in \mathcal{N}_+\}$ is a balanced sequence for every $i \in J$ $K_i \neq \emptyset$; let $k_i = \min K_i$ and for $\mathbf{s} \in P_{k_i}$ let $c_{i\mathbf{s}} = \lim_{n\to\infty} w_{in}H_{in\mathbf{s}}$. Let $\mathcal{M}$ be an infinite subset of $\mathcal{N}_+$ such that (i) for every $i \in J$, for every $r = 1, \dots, k_i - 1$ and for every $\mathbf{s} \in P_r$ the sequence $\{w_{in}H_{in\mathbf{s}}; n \in \mathcal{M}\}$ is either convergent or divergent.

By the Taylor formula for multivariable functions (see e.g., [?, p.130]) we have

$$\sum_{i=1}^m w_{in}\,(\phi(\mathbf{y}_i + \mathbf{h}_{in}, \mathbf{x}) - \phi(\mathbf{y}_i, \mathbf{x})) =$$

$$\sum_{i=1}^m w_{in} \left(\sum_{r=1}^{k_i - 1} \frac{D_{\mathbf{h}_{in}}^{(r)}\phi(\mathbf{y}_i, \mathbf{x})}{r!} + \frac{D_{\mathbf{h}_{in}}^{(k_i)}\phi(\mathbf{z}_{in}, \mathbf{x})}{k_i!} \right),$$

where $\mathbf{z}_{in} \in L(\mathbf{y}_i, \mathbf{y}_i + \mathbf{h}_{in})$ and $D_{\mathbf{h}_{in}}^{(r)}$ are directional derivatives.

Since for every $i = 1, \dots, m$ $\lim_{n\to\infty} \mathbf{h}_{in} = \mathbf{0}$, we have $\lim_{n\to\infty} \mathbf{z}_{in} = \mathbf{y}_i$. Hence by (4)

$$\lim_{\substack{n\to\infty \\ n\in\mathcal{M}}} \sum_{i\in J} w_{in} \frac{D_{\mathbf{h}_{in}}^{(k_i)}\phi(\mathbf{z}_{in}, \mathbf{x})}{k_i!} = \sum_{i\in J} \sum_{\mathbf{s}\in P_{k_i}} c_{i\mathbf{s}} D_1^{(s_1)} \dots D_p^{(s_p)} \phi(\mathbf{y}_i, \mathbf{x}).$$

Let

$$g(\mathbf{x}) = f(\mathbf{x}) - \sum_{i\in J}\sum_{\mathbf{s}\in P_{k_i}} c_{i\mathbf{s}} D_1^{(s_1)}\dots D_p^{(s_p)}\phi(\mathbf{y}_i,\mathbf{x}).$$

Then

$$g(\mathbf{x}) = \lim_{\substack{n\to\infty\\ n\in\mathcal{M}}} \sum_{i=1}^{m}\sum_{r=1}^{k_i-1}\sum_{\mathbf{s}\in P_r} w_{in} H_{in\mathbf{s}} D_1^{(s_1)}\dots D_p^{(s_p)}\phi(\mathbf{y}_i,\mathbf{x}).$$

For every $n\in\mathcal{M}$ put $v_n = \max\{|\hat{w}_{in}H_{in\mathbf{s}}|; i=1,\dots,m, \mathbf{s}\in P_r, r=1,\dots,k_i-1\}$ and let $u_{in\mathbf{s}} = \frac{\hat{w}_{in}H_{in\mathbf{s}}}{v_n}$. Note that for every $i\in J$ either $u_{i_0\mathbf{s}_0}=1$ or $u_{i_0\mathbf{s}_0}=-1$.

Let $\mathcal{M}'$ be an infinite subset of $\mathcal{M}$ such that
(i) for every $i\in J$, for every $r=1,\dots,k_i-1$ and for every $\mathbf{s}\in P_r$ the sequence $\{\hat{w}_{in}H_{in\mathbf{s}}; n\in\mathcal{M}'\}$ is either convergent or divergent
(ii) for every $n\in\mathcal{M}'$ there exists $u_{i\mathbf{s}} = \lim_{\substack{n\to\infty\\ n\in\mathcal{M}'}} \hat{w}_{i\mathbf{s}}$
(ii) either there exists $i_0\in J$, $r_0\in\{1,\dots,k_i-1\}$, $\mathbf{s}_0\in P_{r_0}$ such that for every $n\in\mathcal{M}'$ $v_n = w_{i_0n}H_{i_0n\mathbf{s}_0}$ or there exists $i_0, r_0, \mathbf{s}_0$ such that for every $n\in\mathcal{M}'$ $v_n = -w_{i_0n}H_{i_0n\mathbf{s}_0}$.

Then

$$0 = \lim_{\substack{n\to\infty\\ n\in\mathcal{M}'}} \sum_{i\in J}\sum_{r=1}^{k_i-1}\sum_{\mathbf{s}\in P_r} u_{in\mathbf{s}} D_1^{(s_1)}\dots D_p^{(\mathbf{s}_p)}\phi(\mathbf{y}_i,\mathbf{x}).$$

Let $I = \{i\in J; (\exists\mathbf{s}\in S_i)(u_{i\mathbf{s}}\neq 0)\}$. Then we have a functional equation

$$0 = \sum_{i\in I}\sum_{\mathbf{s}\in S_i} u_{i\mathbf{s}} D_1^{(s_1)}\dots D_p^{(s_p)}\phi(\mathbf{y}_i,\mathbf{x})$$

with $\hat{v}_{i_0\mathbf{s}_0}=1$ or $\hat{v}_{i_0\mathbf{s}_0}=-1$, which contradicts the assumption that ϕ is not derivative recursive on A. $\square$

Proof of Theorem 4.1

Let $f\in cl(\mathcal{F}(\phi,A))$ be a limit of a balanced sequence $\{\sum_{i=1}^m w_{in}\phi(\mathbf{y}_{in},\mathbf{x}); n\in\mathcal{N}_+\}$. Let $I=\{i\in\{1,\dots,m\}; \{w_{in};\in\mathcal{N}_+\}$ is divergent $\}$, $J=\{i\in\{1,\dots,m\};\{w_{in};\in\mathcal{N}\}$ is convergent $\}$, for every $i=1,\dots,m$ let $\mathbf{y}_i=\lim_{n\to\infty}\mathbf{y}_{in}$, for every $i\in J$ $w_i = \lim_{n\to\infty} w_{in}$ and $f_1(\mathbf{x}) = \sum_{i\in J} w_i\phi(\mathbf{y}_i,\mathbf{x})$. Since ϕ is uniformly continuous on $A\times\mathcal{R}^p$ $f_1(\mathbf{x}) = \lim_{n\to\infty}\sum_{i\in J} w_{in}\phi(\mathbf{y}_{in},\mathbf{x})$ uniformly on $\mathcal{R}^d$.

Let $\sim$ denotes an equivalence on I defined by $i\sim k$ if $\lim_{n\to\infty}\mathbf{y}_{in} = \lim_{n\to\infty}\mathbf{y}_{kn}$. Let $\tilde{I}$ be a subset of I containing exactly one representative of each class of $\sim$. For every $i\in\tilde{I}$ define $K_i=\{k\in I; k\sim i\}$, $\hat{w}_{in}=\sum_{k\in K_i} w_{kn}$ and $\mathbf{h}_{in} = \mathbf{y}_{in}-\mathbf{y}_i$.

Then

$$f(\mathbf{x}) - f_1(\mathbf{x}) = \lim_{n\to\infty}\left(\sum_{i\in\tilde{I}}\sum_{k\in K_i} w_{kn}(\phi(\mathbf{y}_k+\mathbf{h}_{kn},\mathbf{x}) - \phi(\mathbf{y}_i,\mathbf{x}) + \phi(\mathbf{y}_i,\mathbf{x}))\right) =$$

$$\lim_{n\to\infty}\left(\sum_{i\in I} w_{in}(\phi(\mathbf{y}_i+\mathbf{h}_{in},\mathbf{x}) - \phi(\mathbf{y}_i,\mathbf{x})) + \sum_{i\in\tilde{I}}\hat{w}_{in}\phi(\mathbf{y}_i,\mathbf{x})\right).$$

Let $\mathcal{M}$ be an infinite subset of $\mathcal{N}_+$ satisfying the following conditions:
(i) for each $i \in \tilde{I}$ the sequence $\{\hat{w}_{in}; n \in \mathcal{M}\}$ is either convergent or divergent
(ii) there exists $i_0 \in \tilde{I}$ such that either for every $n \in \mathcal{M}$ $1 \leq \hat{w}_{i_0 n} = \max\{|\hat{w}_{in}|; i \in \tilde{I}\}$ or $1 \leq -\hat{w}_{i_0 n} = \max\{|\hat{w}_{in}|; i \in \tilde{I}\}$
(iii) for every $i \in \tilde{I}$ $\{\frac{\hat{w}_{in}}{v_n}; n \in \mathcal{M}\}$ is convergent, where $v_n = |w_{i_0 n}|$.

Let $\hat{w}_i = \lim_{\substack{n\to\infty\\ n\in\mathcal{M}}} \hat{w}_{in}$, $u_i = \lim_{\substack{n\to\infty\\ n\in\mathcal{M}}} \frac{\hat{w}_{in}}{v_n}$, $\hat{I} = \{i \in \{1,\dots,m\}; \{\hat{w}_{in}; \in \mathcal{N}\}$ is divergent $\}$, $\hat{J} = \{i \in \{1,\dots,m\}; \{\hat{w}_{in}; \in \mathcal{N}\}$ is convergent $\}$ and $f_2(\mathbf{x}) = \sum_{i\in\hat{J}} \hat{w}_i \phi(\mathbf{y}_i, \mathbf{x})$. Note that for every $n \in \mathcal{M}$ $\frac{|\hat{w}_{in}|}{v_n} = 1$ and so either $u_{i_0} = 1$ or $u_{i_0} = -1$.

Uniform continuity of ϕ guarantees that $f_2(\mathbf{x}) = \lim_{\substack{n\to\infty\\ n\in\mathcal{M}}} \sum_{i\in\hat{J}} \hat{w}_{in}\phi(\mathbf{y}_{in}, \mathbf{x})$ uniformly on A.

Hence

$$f(\mathbf{x}) - f_1(\mathbf{x}) - f_2(\mathbf{x}) = \lim_{\substack{n\to\infty\\ n\in\mathcal{M}}} \left(\sum_{i\in I} w_{in}(\phi(\mathbf{y}_{in}, \mathbf{x}) - \phi(\mathbf{y}_i, \mathbf{x})) + \sum_{i\in\hat{I}} \hat{w}_{in}\phi(\mathbf{y}_i, \mathbf{x}) \right)$$

uniformly on A.

We will show by contradiction that $\hat{I} = \emptyset$. Assume that $\hat{I} \neq \emptyset$. Then

$$0 = \lim_{\substack{n\to\infty\\ n\in\mathcal{M}}} \frac{f(\mathbf{x}) - f_1(\mathbf{x}) - f_2(\mathbf{x})}{v_n} = \lim_{\substack{n\to\infty\\ n\in\mathcal{M}}} \sum_{i\in I} \left(\frac{w_{in}}{v_n}(\phi(\mathbf{y}_i + \mathbf{h}_{in}, \mathbf{x}) - \phi(\mathbf{y}_i, \mathbf{x})) + \sum_{i\in\hat{I}} \frac{\hat{w}_{in}}{v_n}\phi(\mathbf{y}_i, \mathbf{x}) \right).$$

Thus,

$$-\sum_{i\in\hat{I}} u_i\phi(\mathbf{y}_i, \mathbf{x}) = \lim_{\substack{n\to\infty\\ n\in\mathcal{M}}} \sum_{i\in I} \frac{\hat{w}_{in}}{v_n}(\phi(\mathbf{y}_i + \mathbf{h}_{in}, \mathbf{x}) - \phi(\mathbf{y}_i, \mathbf{x})). \tag{7.2}$$

The sequence on the right side of (5) is balanced since it is obtained from a subsequence of a balanced sequence by dividing each n-th member by v_n satisfying $v_n \geq 1$. Since this sequence converges to a finite function, by Lemma 7.2 its limit must be a function from $\mathcal{D}(\phi, A)$, i.e. a function of the form $\sum_{i\in I}\sum_{\mathbf{s}\in P_i} a_{\mathbf{s}i} D_1^{(s_1)} \dots D_p^{(s_d)}\phi(\mathbf{y}_i, \mathbf{x})$. Thus, we have a functional equation

$$\sum_{i\in\hat{I}} a_i\phi(\mathbf{y}_i, \mathbf{x}) + \sum_{i\in I}\sum_{\mathbf{s}\in P_i} a_{\mathbf{s}i} D_1^{(s_1)} \dots D_p^{(s_p)}\phi(\mathbf{y}_i, \mathbf{x}) = 0,$$

where for some $i_0 \in \tilde{I}$ either $u_{i_0} = 1$ or $u_{i_0} = -1$, which contradicts the assumption that ϕ is not derivative recursive on A.

Thus $\hat{I} = \emptyset$ and we have

$$f(\mathbf{x}) - f_1(\mathbf{x}) - f_2(\mathbf{x}) = \lim_{\substack{n\to\infty\\ n\in\mathcal{M}}} \left(\sum_{i\in I} w_{in}(\phi(\mathbf{y}_i + \mathbf{h}_{in}, \mathbf{x}) - \phi(\mathbf{y}_i, \mathbf{x})) \right)$$

uniformly on A. By Lemma 7.2 the limit on the right side is a function from $\mathcal{D}(\phi, A)$. Since $f_1, f_2 \in \mathcal{F}(\phi, A) \subseteq \mathcal{D}(\phi, A)$, we have $f \in \mathcal{D}(\phi, A)$. $\Box$

To prove Theorems 5.1 and 5.2 we need formulas for higher order partial derivatives of P_τ and B_γ with respect to the first $d+1$ variables. Recall that $\tau'(t) = 1 - (\tau(t))^2$ and that $\gamma'(t) = -2t\gamma(t)$.

Lemma 7.3 *There exists a sequence of polynomials $\{p_s : \mathcal{R} \to \mathcal{R}; s \in \mathcal{N}_+\}$ such that for all positive integers d, s and for every $\mathbf{x}, \mathbf{v} \in \mathcal{R}^d$, $b \in \mathcal{R}$*

$$D^{(s)}_{d+1} P_\tau(\mathbf{v}, b, \mathbf{x}) = \frac{\partial^s \tau(\mathbf{v}\cdot\mathbf{x} + b)}{\partial b^s} = p_s(\tau(\mathbf{v}\cdot\mathbf{x} + b)),$$

and for every $j = 1, \ldots, d$

$$D^{(s)}_j P_\tau(\mathbf{v}, b, \mathbf{x}) = \frac{\partial^s \tau(\mathbf{v}\cdot\mathbf{x} + b)}{\partial {v_j}^s} = {x_j}^s p_s(\tau(\mathbf{v}\cdot\mathbf{x} + b)),$$

and $\{p_s; s \in \mathcal{N}_+\}$ satisfies the following recursion: $p_1(t) = 1 - t^2$, $p_{s+1}(t) = {p_s}'(t)(1 - t^2)$.

Proof
The first part is true for $s = 1$ since

$$\frac{\partial \tau(\mathbf{v}\cdot\mathbf{x} + b)}{\partial b} = 1 - \tau(bv\cdot\mathbf{x}))^2 = p_1(\tau(\mathbf{v}\cdot\mathbf{x} + b)).$$

Suppose that it is true for s. Then

$$\frac{\partial^{s+1} \tau(\mathbf{v}\cdot\mathbf{x} + b)}{\partial b^{s+1}} = \frac{\partial p_s(\tau(\mathbf{v}\cdot\mathbf{x} + b)}{\partial b} =$$

$${p_s}'(\tau(\mathbf{v}\cdot\mathbf{x} + b))(1 - (\tau(\mathbf{v}\cdot\mathbf{x} + b))^2 = p_{s+1}(\tau(\mathbf{v}\cdot\mathbf{x} + b)).$$

When $p_s(t)$ is a polynomial, then $p_{s+1}(t) = {p_s}'(1 - t^2)$ is a polynomial, too. Thus, the first part holds also for $s + 1$.

The second part is true for $s = 1$ since

$$\frac{\partial \tau(\mathbf{v}\cdot\mathbf{x} + b)}{\partial v_j} = (1 - (\tau(\mathbf{v}\cdot\mathbf{x} + b))^2)x_j = x_j p_1(\tau(\mathbf{v}\cdot\mathbf{x} + b)).$$

Suppose that it holds for s. Then

$$\frac{\partial^{s+1} \tau(\mathbf{v}\cdot\mathbf{x} + b)}{\partial {v_j}^{s+1}} = x_j^s {p_s}'(\tau(\mathbf{v}\cdot\mathbf{x} + b))(1 - (\tau(\mathbf{v}\cdot\mathbf{x} + b))^2)x_j = x_j^{s+1} p_{s+1}(\tau(\mathbf{v}\cdot\mathbf{x} + b)).$$

When $p_s(t)$ is a polynomial, then $p_{s+1}(t) = {p_s}'(t)(1-t^2)$ is a polynomial, too. Thus, the second part is also true for $s + 1$. $\Box$

Lemma 7.4 *There exist two sequences of polynomials* $\{p_s : \mathcal{R}^2 \to \mathcal{R}; s \in \mathcal{N}_+\}$ *and* $\{q_s : \mathcal{R}^3 \to \mathcal{R}; s \in \mathcal{N}_+\}$ *such that for all positive integers* d, s, *for every* $\mathbf{x}, \mathbf{v} \in \mathcal{R}^d$

$$D^{(s)}_{d+1} B_\gamma(\mathbf{v}, b, \mathbf{x}) = \frac{\partial^s \gamma(b\|\mathbf{x}-\mathbf{v}\|)}{\partial b^s} = \gamma(b\|\mathbf{x}-\mathbf{v}\|)\ p_s(b, \|\mathbf{x}-\mathbf{v}\|),$$

for every $j = 1, \ldots, d$

$$D^{(s)}_{j} B_\gamma(\mathbf{v}, b, \mathbf{x}) = \frac{\partial^s \gamma(b\|\mathbf{x}-\mathbf{v}\|)}{\partial {v_j}^s} = \gamma(b\|\mathbf{x}-\mathbf{v}\|)\ q_s(b, v_j, x_j)$$

and $\{p_s; s \in \mathcal{N}_+\}$ *and* $\{q_s; s \in \mathcal{N}_+\}$ *satisfy the following recursions:* $p_1(t_1,t_2) = -2t_1{t_2}^2$, $p_{s+1}(t_1,t_2) = p_1(t_1,t_2)p_s(t_1,t_2) + \frac{\partial p_s(t_1,t_2)}{\partial t_1}$, $q_1(t_1,t_2,t_3) = 2{t_1}^2(t_3 - t_2)$, $q_{s+1}(t_1,t_2,t_3) = q_1(t_1,t_2,t_3)q_s(t_1,t_2,t_3) + \frac{\partial q_s(t_1,t_2,t_3)}{\partial t_2}$.

Proof.
The first part is true for $s = 1$ since

$$\frac{\partial \exp(-b^2\|\mathbf{x}-\mathbf{v}\|^2)}{\partial b} = \exp(-b^2\|\mathbf{x}-\mathbf{v}\|^2)(-2b\|\mathbf{x}-\mathbf{v}\|^2) = \gamma(b\|\mathbf{x}-\mathbf{v}\|)p_1(b, \|\mathbf{x}-\mathbf{v}\|),$$

where $p_1(t_1,t_2) = -2t_1{t_2}^2$.

Suppose that it is true for s. Then

$$\frac{\partial^{s+1} \exp(-b^2\|\mathbf{x}-\mathbf{v}\|^2)}{\partial b^{s+1}} = \frac{\partial \exp(-b^2\|\mathbf{x}-\mathbf{v}\|^2)p_s(b, \|\mathbf{x}-\mathbf{v}\|)}{\partial b} =$$

$$\exp(-b^2\|\mathbf{x}-\mathbf{v}\|^2)\left(p_1(b, \|\mathbf{x}-\mathbf{v}\|)p_s(b\|\mathbf{x}-\mathbf{v}\|) + \frac{\partial p_s(b, \|\mathbf{x}-\mathbf{v}\|)}{\partial b}\right).$$

When $p_s(t_1,t_2)$ is a polynomial, then $p_{s+1}(t_1,t_2) = p_1(t_1,t_2)p_s(t_1,t_2) + \frac{\partial p_s(t_1,t_2)}{\partial t_1}$ is a polynomial, too. Thus the first part also holds for $s+1$.

The second part is true for $s = 1$ since

$$\frac{\partial \exp(-b^2\|\mathbf{x}-\mathbf{v}\|^2)}{\partial v_j} = \exp(-b^2\|\mathbf{x}-\mathbf{v}\|^2)(-2b^2(v_j - x_j)) = \gamma(b\|\mathbf{x}-\mathbf{v}\|)q_1(b, v_j, x_j),$$

where $q_1(t_1,t_2,t_3) = 2{t_1}^2(t_3 - t_2)$.

Suppose that it holds for s. Then

$$\frac{\partial^{s+1} \exp(-b^2\|\mathbf{x}-\mathbf{v}\|^2)}{\partial {v_j}^{s+1}} = \exp(-b^2\|\mathbf{x}-\mathbf{v}\|^2)\left(q_1(b, v_j, x_j)q_s(b, v_j, x_j) + \frac{\partial q_s(b, v_j, x_j)}{\partial v_j}\right).$$

When $q_s(t_1,t_2,t_3)$ is a polynomial, then $q_{s+1}(t_1,t_2,t_3) = q_1(t_1,t_2,t_3)q_s(t_1,t_2,t_3) + \frac{\partial q_s(t_1,t_2,t_3)}{\partial t_2}$ is a polynomial, too. Thus, the second part is also true for $s+1$. $\Box$

Proof of Theorem 5.1

Extending the definition of p_s also to $s = 0$ by setting $p_0(t) = t$ we get from Lemma 7.3 for every $j = 1, \ldots, d+1$ and every $s_j \in \mathcal{N}$ $D_j^{(s_j)} P_\tau(\mathbf{v}, b, \mathbf{x}) = x_j{}^{s_j} p_{s_j}(P_\tau(\mathbf{v}, b, \mathbf{x}))$, where $deg(p_{s_j}) \leq s_j$. Since derivative of a polynomial is a polynomial, we get applying Lemma 7.3 repeatedly for any $\mathbf{s} = (s_1, \ldots, s_{d+1})$ a polynomial $Q_\mathbf{s}$ with $deg(Q_\mathbf{s}) \leq |s|+1$ such that for every $\mathbf{x}, \mathbf{v} \in \mathcal{R}^d$ and $b \in \mathcal{R}$ $D_1^{(s_1)} \ldots D_{d+1}^{(s_{d+1})} P_\tau(\mathbf{v}, b, \mathbf{x}) = x_1{}^{s_1} \ldots x_d{}^{s_d} Q_\mathbf{s}(P_\tau(\mathbf{v}, b, \mathbf{x}))$. Hence

$\sum_{i=1}^m \sum_{\mathbf{s} \in P_i} D_1^{(s_1)} \ldots D_{d+1}^{(s_{d+1})} P_\tau(\mathbf{v}_i, b_i, \mathbf{x}_i) = x_1{}^{s_1} \ldots x_d{}^{s_d} Q_\mathbf{s}(P_\tau(\mathbf{v}_i, b_i, \mathbf{x}))$. $\square$

Proof of Theorem 5.2

Extending the definition of p_s and q_s also to $s = 0$ by setting $p_0 = 1$ and $q_0 = 1$ we get from Lemma 7.4 for every $j = 1, \ldots, d+1$ and every $s_j \in \mathcal{N}$ $D_j^{(s_j)} B_\gamma(\mathbf{v}, b, \mathbf{x}) = B_\gamma(\mathbf{v}, b, \mathbf{x}) p_{s_j}(b, \|\mathbf{x}-\mathbf{v}\|)$ or $D_j^{(s_j)} B_\gamma(\mathbf{v}, b, \mathbf{x}) = B_\gamma(\mathbf{v}, b, \mathbf{x}) q_{s_j}(b, v_j, x_j)$, where $deg(p_{s_j}) \leq 2s_j$ and $deg(_{s_j}) \leq 2s_j$. Since derivative of a polynomial is a polynomial we get applying Lemma 7.4 repeatedly a polynomial $Q_\mathbf{s} : \mathcal{R}^{2(d+1)} \to \mathcal{R}$ such that $deg(Q_\mathbf{s}) \leq 2|\mathbf{s}|$ and for every $\mathbf{x}, \mathbf{v} \in \mathcal{R}^d$ and $b \in \mathcal{R}$ $D_1^{(s_1)} \ldots D_{d+1}^{(s_{d+1})} B_\gamma(\mathbf{v}, b, \mathbf{x}) = B_\gamma(\mathbf{v}, b, \mathbf{x}) Q_\mathbf{s}(b, \|\mathbf{x} - \mathbf{v}\|, v_1, \ldots, v_d, x_1, \ldots, x_d)$. Let $Q_i = \sum_{i=1}^m \sum_{\mathbf{s} \in P_i} a_{i\mathbf{s}} Q_\mathbf{s}$. Then

$\sum_{i=1}^m \sum_{\mathbf{s} \in P_i} a_{i\mathbf{s}} D_1^{(s_1)} \ldots D_{d+1}^{(s_{d+1})} B_\gamma(\mathbf{v}_i, b_i, \mathbf{x}) = \sum_{i=1}^m B_\gamma(\mathbf{v}_i, b_i, \mathbf{x}) Q_i(b_i, \|\mathbf{x}-\mathbf{v}_i\|, v_{i1}, \ldots, v_{id}, x_1, \ldots, x_d)$.

$\square$

Bibliography

[1] H. N. Mhaskar and C. A. Micchelli, "Approximation by superposition of sigmoidal and radial basis functions", *Advances in Applied Mathematics*, vol. 13, pp. 350–373, 1992.

[2] J. Park and I. W. Sandberg, "Approximation and radial-basis-function networks", *Neural Computation*, vol. 5, pp. 305–316, 1993.

[3] M. Stinchcombe and H. White, "Approximating and learning unknown mappings using multilayer feedforward networks with bounded weights", in *Proceedings of IJCNN*, vol. III, pp. 7–16. IEEE Press, New York, 1990.

[4] K. Hornik, "Some new results on neural network approximation", *Neural Networks*, vol. 6, pp. 1069–1072, 1993.

[5] F. Girosi and T. Poggio, "Networks and the best approximation property", *Biological Cybernetics*, vol. 63, pp. 169–176, 1990.

[6] C. K. Chui and X. Li and H. N. Mhaskar, "Neural networks for localized approximation", *Mathematics of Computation*, (in press).

[7] V. Kůrková, "Approximation of functions by perceptron networks with bounded number of hidden units", *Neural Networks*, vol. 8, pp. 745–750, 1995.

[8] M. Gori and F. Scarselli and A. C. Tsoi, "Which classes of functions can a given multilayer perceptron approximate?", in *Proccedings of the ICNN'96*, pp. 1481–1487. IEEE, 1996.

[9] F.G. Simmons, *Introduction to Topology and Modern Analysis*, McGraw-Hill, New York, 1963.

[10] V. Kůrková and K. Hlaváčková, "Approximation of continuous functions by rbf and kbf networks", in *Proceedings of ESANN'94*, pp. 167–174. D facto, Brussels, 1994.

[11] L.K. Jones, "A simple lemma on greedy approximation in hilbert space and convergence rates for projection pursuit regression and neural network training", *Annals of Statistics*, vol. 20, pp. 608–613, 1992.

[12] A.R. Barron, "Universal approximation bounds for superposition of a sigmoidal function", *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 930–945, 1993.

[13] H. N. Mhaskar and C. A. Micchelli, "Dimension-independent bounds on the degree and of approximation by neural networks", *IBM Journal of Research and Development*, vol. 38, pp. 277–284, 1994.

[14] F. Girosi and G. Anzellotti, "Rates of convergence for radial basis fucntion and neural networks", in *Artificial Neural Networks for Speech and Vision*, pp. 97–113. Chapman & Hall, London, 1993.

[15] V. Kůrková and P.C. Kainen and V. Kreinovich, "Estimates of the number of hidden units and variation with respect to half-spaces", *Neural Networks*, in press.

[16] M. Leshno and V. Lin and A. Pinkus and S. Schocken, "Multilayer feedforward networks with a non-polynomial activation function can approximate any function", *Neural Networks*, vol. 6, pp. 861–867, 1993.

[17] V. Kůrková and P.C. Kainen, "Singularities of finite scaling functions", *Applied Math. Letters*, vol. 9, pp. 33–37, 1996.

[18] V. Kůrková and P.C. Kainen, "Functionally equivalent feedforward neural networks", *Neural Computation*, vol. 6, pp. 543–558, 1994.

[19] V. Kůrková and R. Neruda, "Uniqueness of functional representations by gaussian basis function networks", in *Proceedings of ICANN'94*, pp. 471–474. Springer, London, 1994.

[20] P.C. Kainen and V. Kůrková and V. Kreinovitch and O. Sirisengtaksin, "Uniqueness of network parameterization and faster learning", *Neural, Parallel and Scientific Computations*, vol. 2, pp. 459–466, 1994.

[21] C. H. Edwards, *Advanced Calculus in Several Variables*, Dover, New York, 1994.