Acta Numerica (1999), pp. 143–195

Approximation theory of the MLP model in neural networks

Allan Pinkus
Department of Mathematics,
Technion – Israel Institute of Technology,
Haifa 32000, Israel
E-mail: pinkus@tx.technion.ac.il

In this survey we discuss various approximation-theoretic problems that arise in the multilayer feedforward perceptron (MLP) model in neural networks. The MLP model is one of the more popular and practical of the many neural network models. Mathematically it is also one of the simpler models. Nonetheless the mathematics of this model is not well understood, and many of these problems are approximation-theoretic in character. Most of the research we will discuss is of very recent vintage. We will report on what has been done and on various unanswered questions. We will not be presenting practical (algorithmic) methods. We will, however, be exploring the capabilities and limitations of this model.

In the first two sections we present a brief introduction and overview of neural networks and the multilayer feedforward perceptron model. In Section 3 we discuss in great detail the question of density. When does this model have the theoretical ability to approximate any reasonable function arbritrarily well? In Section 4 we present conditions for simultaneously approximating a function and its derivatives. Section 5 considers the interpolation capability of this model. In Section 6 we study upper and lower bounds on the order of approximation of this model. The material presented in Sections 3–6 treats the single hidden layer MLP model. In Section 7 we discuss some of the differences that arise when considering more than one hidden layer. The lengthy list of references includes many papers not cited in the text, but relevant to the subject matter of this survey.

CONTENTS

1	On neural networks	144
2	The MLP model	146
3	Density	150
4	Derivative approximation	162
5	Interpolation	165
6	Degree of approximation	167
7	Two hidden layers	182
References		187

1. On neural networks

It will be assumed that most readers are pure and/or applied mathematicians who are less than conversant with the theory of neural networks. As such we begin this survey with a very brief, and thus inadequate, introduction.

The question 'What is a neural network?' is ill-posed. From a quick glance through the literature one quickly realizes that there is no universally accepted definition of what the theory of neural networks is, or what it should be. It is generally agreed that neural network theory is a collection of models of computation very, very loosely based on biological motivations. According to Haykin (1994, p. 2):

'A neural network is a massively parallel distributed processor that has a natural propensity for storing experiential knowledge and making it available for use. It resembles the brain in two respects:

1. Knowledge is acquired by the network through a learning process.
2. Interneuron connection strengths known as synaptic weights are used to store the knowledge.'

This is a highly nonmathematical formulation. Let us try to be a bit less heuristic. Neural network models have certain common characteristics. In all these models we are given a set of inputs $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and some process that results in a corresponding set of outputs $\mathbf{y} = (y_1, \ldots, y_m) \in \mathbb{R}^m$. The basic underlying assumption of our models is that the process is given by some mathematical function, that is,

$$\mathbf{y} = G(\mathbf{x})$$

for some function G. The function G may be very complicated. More importantly, we cannot expect to be able to compute exactly the unknown G. What we do is choose our 'candidate' F (for G) from some parametrized set of functions using a given set of examples, that is, some inputs $\mathbf{x}$ and associated 'correct' outputs $\mathbf{y} = G(\mathbf{x})$, which we assume will help us to choose the parameters. This is a very general framework. In fact it is

still too general. Neural network models may be considered as particular choices of classes of functions $F(\mathbf{x}, \mathbf{w})$ where the $\mathbf{w}$ are the parameters, together with various rules and regulations as well as specific procedures for optimizing the choice of parameters. Most people would also agree that a neural network is an input/output system with many simple processors, each having a small amount of local memory. These units are connected by communication channels carrying data. Most neural network models have some sort of *training rule*, that is, they learn or are trained from a set of examples. There are many, many different models of neural network. (Sarle (1998) lists over 40 different recognized neural network models, and there are a plethora of additional candidates.)

Neural networks have emerged, or are emerging, as a practical technology, that is, they are being successfully applied to real world problems. Many of their applications have to do with pattern recognition, pattern classification, or function approximation, which are all based on a large set of available examples (training set). According to Bishop (1995, p. 5):

'The importance of neural networks in this context is that they offer a very powerful and very general framework for representing non-linear mappings from several input variables to several output variables, where the form of the mapping is governed by a number of adjustable parameters.'

The nonlinearity of the neural network models presents advantages and disadvantages. The price (and there always is a cost) is that the procedure for determining the values of the parameters is now a problem in nonlinear optimization which tends to be computationally intensive and complicated. The problem of finding efficient algorithms is of vital importance and the true utility of any model crucially depends upon its efficiency. (However, this is not an issue we will consider in this survey.)

The theory of neural nets has become increasing popular in the fields of computer science, statistics, engineering (especially electrical engineering), physics, and many more directly applicable areas. There are now four major journals in the field, as well as numerous more minor journals. These leading journals are *IEEE Transactions on Neural Networks*, *Neural Computation*, *Neural Networks* and *Neurocomputing*. Similarly, there are now dozens of textbooks on the theory. In the references of this paper are listed only five books, namely Haykin (1994), Bishop (1995), Ripley (1996), Devroye, Gyorfi and Lugosi (1996), and Ellacott and Bos (1996), all of which have appeared in the last five years. The IEEE has generally sponsored (since 1987) two annual conferences on neural networks. Their proceedings run to over 2000 pages and each contains a few hundred articles and abstracts. A quick search of Mathematical Reviews (MathSciNet) turned up a mere 1800 entries when the phrase 'neural network' was entered (and you should realize that much of the neural network literature, including all the above-mentioned journals, is

not written for or by mathematicians and is not reviewed by Mathematical Reviews). In other words, this is an explosively active research area and deserves the attention of the readership of *Acta Numerica*. Initially there was a definite lack of mathematical sophistication to the theory. It tended to be more a collection of *ad hoc* techniques with debatable justifications. To a pure mathematician, such as the author, reading through some of the early literature in the field was an alien experience. In recent years the professionals (especially statisticians) have established a more organized framework for the theory.

The reader who would like to acquire a more balanced and enlarged view of the theory of neural networks is urged to peruse a few of the above-mentioned texts. An additional excellent source of information about neural networks and its literature is the 'frequently asked questions' (FAQs) of the Usenet newsgroup `comp.ai.neural-nets`: see Sarle (1998).

This survey is not about neural networks *per se*, but about the approximation theory of the multilayer feedforward perceptron (MLP) model in neural networks. We will consider certain mathematical, rather than computational or statistical, problems associated with this widely used neural net model. More explicitly, we shall concern ourselves with problems of density (when the models have at least the theoretical capability of providing good approximations), degree of approximation (the extent to which they can approximate, as a function of the number of parameters), interpolation, and related issues. Theoretical results, such as those we will survey, do not usually have direct applications. In fact they are often far removed from practical considerations. Rather they are meant to tell us what is possible and, sometimes equally importantly, what is not. They are also meant to explain why certain things can or cannot occur, by highlighting their salient characteristics, and this can be very useful. As such we have tried to provide proofs of many of the results surveyed.

The 1994 issue of *Acta Numerica* contained a detailed survey: 'Aspects of the numerical analysis of neural networks' by S. W. Ellacott (1994). Only five years have since elapsed, but the editors have again opted to solicit a survey (this time albeit with a slightly altered emphasis) related to neural networks. This is not unwarranted. While almost half of that survey was devoted to approximation-theoretic results in neural networks, almost every one of those results has been superseded. It is to be hoped that the same will be said about this paper five years hence.

2. The MLP model

One of the more conceptually attractive of the neural network models is the multilayer feedforward perceptron (MLP) model. In its most basic form this is a model consisting of a finite number of successive layers. Each layer

consists of a finite number of *units* (often called *neurons*). Each unit of each layer is connected to each unit of the subsequent (and thus previous) layer. These connections are generally called *links* or *synapses*. Information flows from one layer to the subsequent layer (thus the term *feedforward*). The first layer, called the *input* layer, consists of the input. There are then intermediate layers, called *hidden* layers. The resulting output is obtained in the last layer, not surprisingly called the *output* layer. The rules and regulations governing this model are the following.

1. The input layer has as output of its jth unit the (input) value x_{0j}.
2. The kth unit of the ith layer receives the output x_{ij} from each jth unit of the $(i-1)$st layer. The values x_{ij} are then multiplied by some constants (called *weights*) w_{ijk} and these products are summed.
3. A shift θ_{ik} (called a *threshold* or *bias*) and then a fixed mapping σ (called an *activation function*) are applied to the above sum and the resulting value represents the output $x_{i+1,k}$ of this kth unit of the ith layer, that is,

$$x_{i+1,k} = \sigma\left(\sum_j w_{ikj}x_{ij} - \theta_{ik}\right).$$

A priori one typically fixes, for whatever reasons, the activation function, the number of layers and the number of units in each layer. The next step is to choose, in some way, the values of the weights w_{ijk} and thresholds θ_{ik}. These latter values are generally chosen so that the model behaves well on some given set of inputs and associated outputs. (These are called the *training set*.) The process of determining the weights and thresholds is called *learning* or *training*. In the multilayer feedforward perceptron model, the basic *learning algorithm* is called *backpropagation*. Backpropagation is a gradient descent method. It is extremely important in this model and in neural network theory. We shall not detail this algorithm nor the numerous numerical difficulties involved.

We will classify multilayer feedforward perceptron models not by their number of layers, but by their number of *hidden* layers, that is, the number of layers excluding the input and output layer. As is evident, neural network theory has its own terminology. Unfortunately it is also true that this terminology is not always consistent or logical. For example, the term *multilayer perceptron* is generically applied to the above model with at least one hidden layer. On the other hand the word *perceptron* was coined by F. Rosenblatt for the no hidden layer model with the specific activation function given by the Heaviside function

$$\sigma_o(t) = \begin{cases} 1, & t \geq 0, \\ 0, & t < 0. \end{cases}$$

Thus σ_o either fires or does not fire and the breakpoint is some threshold θ. (With this activation function the model is sometimes also referred to as the *McCulloch–Pitts* model.)

Mathematically a no (or zero) hidden layer perceptron network (sometimes confusingly termed a single layer feedforward network) is given as follows. Assume there are n inputs $\mathbf{x} = (x_{01}, \ldots, x_{0n})$, and m outputs $\mathbf{x} = (x_{11}, \ldots, x_{1m})$; then each output is given by

$$x_{1k} = \sigma\left(\sum_{j=1}^{n} w_{jk} x_{0j} - \theta_k\right), \qquad k = 1, \ldots, m, \tag{2.1}$$

for some choice of σ, w_{jk} and θ_k, $j = 1, \ldots, n$, $k = 1, \ldots, m$. This no hidden layer perceptron network is generally no longer used, except in problems of linear separation. There is a simple mathematical rationale for this. A function of the form (2.1) is constant along certain parallel hyperplanes and thus is limited in what it can do. For example, assume $m = 1$ (one output), $n = 2$, and σ is any increasing function. If the input is $\mathbf{x} = (x_1, x_2)$ and the output is y, then

$$y = \sigma\left(w_1 x_1 + w_2 x_2 - \theta\right).$$

Assume we are given four inputs $\mathbf{x}^1$, $\mathbf{x}^2$, $\mathbf{x}^3$ and $\mathbf{x}^4$, no three of which lie on a straight line. Then, as is easily seen, there are output values which cannot be interpolated or approximated well. For example, assume $\mathbf{x}^1$ and $\mathbf{x}^2$ lie on opposite sides of the line through $\mathbf{x}^3$ and $\mathbf{x}^4$. Set $y_1 = y_2 = 1$, $y_3 = y_4 = 0$. Then we cannot solve

$$y_i = \sigma\left(w_1 x_1^i + w_2 x_2^i - \theta\right), \qquad i = 1, \ldots, 4,$$

for any choice of w_1, w_2 and θ. In fact the difference between at least one of the y_i and the associated output will be at least $1/2$. This is totally unacceptable if one wishes to build a network that can approximate well any reasonable function, or classify points according to different criteria. With the Heaviside activation function and no hidden layer, two sets of points can be separated (*classified*) by this model if and only if they are linearly separable. To do more, hidden layers are necessary. The problem of being able to arbitrarily separate N generic points in $\mathbb{R}^n$ into two sets by use of a one hidden layer perceptron model with Heaviside activation function (and one output) was considered by Baum (1988). He showed that the problem is solvable if one uses at least $[N/n]$ units in the hidden layer. This model can be used with both continuously valued and discrete inputs. Baum considers the latter; we will consider the former. We will prove that hidden layers and nonlinearity (or, to be more precise, nonpolynomiality) of the activation function make for models that have the capability of approximating (and interpolating) arbitrarily well.

The model presented above permits generalization, and this can and often is done in a number of ways. The activation function may change from layer to layer (or from unit to unit). We can replace the simple linearity at each unit (*i.e.*, $\sum_j w_{ijk}x_{ij}$) by some more complicated function of the x_{ij}. The architecture may be altered to allow for different links between units of different layers (and perhaps also of the same layer). These are just a few of the many, many possible generalizations. As the mathematical analysis of the multilayer perceptron model is far from being well understood, we will consider only this basic model, with minor modifications. For example, while it is usual in the multilayer perceptron model to apply the same activation function at each hidden layer, it is often the case, and we will follow this convention here, that there be no activation function or threshold applied at the output layer. There may be various reasons for this, from a practical point of view, depending on the problem considered. From a mathematical perspective, applying an activation function to the output layer, especially if the activation function is bounded, is unnecessarily restrictive. Another simplification we will make is to consider models with only one output (unless otherwise noted). This is no real restriction and will tremendously simplify our notation.

With the above modifications (no activation function or threshold applied to the output layer and only one output), we write the output y of a single hidden layer perceptron model with r units in the hidden layer and input $\mathbf{x} = (x_1, \ldots, x_n)$ as

$$y = \sum_{i=1}^{r} c_i \sigma\left(\sum_{j=1}^{n} w_{ij}x_j - \theta_i \right).$$

Here w_{ij} is the weight between the jth unit of the input and the ith unit in the hidden layer, θ_i is the threshold at the ith unit of the hidden layer, and c_i is the weight between the ith unit of the hidden layer and the output. We will generally write this more succinctly as

$$y = \sum_{i=1}^{r} c_i \sigma(\mathbf{w}^i \cdot \mathbf{x} - \theta_i),$$

where $\mathbf{w} \cdot \mathbf{x} = \sum_{j=1}^{n} w_j x_j$ is the standard inner product. We can also express the output y of a two hidden layer perceptron model with r units in the first hidden layer, s units in the second hidden layer, and input $\mathbf{x} = (x_1, \ldots, x_n)$. It is

$$y = \sum_{k=1}^{s} d_k \sigma\left(\sum_{i=1}^{r} c_{ik} \sigma(\mathbf{w}^{ik} \cdot \mathbf{x} - \theta_{ik}) - \gamma_k \right).$$

That is, we iterate the one hidden layer model. We will not write out the exact formula for the output of this model with more hidden layers.

Some common choices for activation functions σ (all may be found in the literature) are the following.

1. The Heaviside function mentioned above, that is, $\sigma(t) = \chi_{[0,\infty)}(t)$. This is sometimes referred to in the neural network literature as the *threshold function.*
2. The *logistic sigmoid* given by
$$\sigma(t) = \frac{1}{1+e^{-t}}.$$
3. $\sigma(t) = \tanh(t/2)$, which is, up to a constant, just a shift of the logistic sigmoid.
4. The piecewise linear function of the form
$$\sigma(t) = \begin{cases} 0, & t \le -1, \\ (t+1)/2, & -1 \le t \le 1, \\ 1, & 1 \le t. \end{cases}$$
5. The *Gaussian sigmoid* given by
$$\sigma(t) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{t} e^{-y^2/2}\, dy.$$
6. The *arctan sigmoid* given by
$$\sigma(t) = \frac{1}{\pi} \arctan(t) + \frac{1}{2}.$$

The logistic sigmoid is often used because it is well suited to the demands of backpropagation. It is a C^2 function whose derivative is easily calculated.

Note that all the above functions are bounded (generally increasing from 0 to 1). The term *sigmoidal* is used for the class of activation functions satisfying $\lim_{t\to-\infty} \sigma(t) = 0$ and $\lim_{t\to\infty} \sigma(t) = 1$. However, there is a certain lack of consistency in the terminology. Some authors also demand that σ be continuous and/or monotonic (or even strictly monotonic) on all of $\mathbb{R}$. Others make no such demands. We shall try to be explicit in what we mean when we use the term.

3. Density

In this section we will consider density questions associated with the single hidden layer perceptron model. That is, we consider the set
$$\mathcal{M}(\sigma) = \operatorname{span}\{\sigma(\mathbf{w}\cdot\mathbf{x} - \theta) : \theta \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^n\},$$
and ask the following question. For which σ is it true that, for any $f \in C(\mathbb{R}^n)$, any compact subset K of $\mathbb{R}^n$, and any $\varepsilon > 0$, there exists a $g \in \mathcal{M}(\sigma)$ such that
$$\max_{\mathbf{x}\in K} |f(\mathbf{x}) - g(\mathbf{x})| < \varepsilon\,?$$

In other words, when do we have density of the linear space $\mathcal{M}(\sigma)$ in the space $C(\mathbb{R}^n)$, in the topology of uniform convergence on compacta (compact sets)? In fact we shall also restrict the permissible set of weights $\mathbf{w}$ and thresholds θ. To set terminology, we shall say that σ *has the density property* if $\mathcal{M}(\sigma)$ is dense in $C(\mathbb{R}^n)$ in the above topology. It should be noted that this norm is very strong. If μ is any nonnegative finite Borel measure, with support in some compact set K, then $C(K)$ is dense in $L^p(K,\mu)$ for any $1 \leq p < \infty$. Thus the results of this section extend also to these spaces.

In the renaissance of neural net theory that started in the mid-1980s, it was clearly understood that this density question, whether for the single hidden or any number of hidden layer perceptron model, was of fundamental importance to the theory. Density is the theoretical ability to approximate well. Density does not imply a good, efficient scheme for approximation. However, a lack of density means that it is impossible to approximate a large class of functions, and this effectively precludes any scheme based thereon from being in the least useful. This is what killed off the efficacy of the no hidden layer model. Nonetheless it should be understood that density does not imply that one can approximate well to every function from

$$\mathcal{M}_r(\sigma) = \left\{ \sum_{i=1}^{r} c_i \sigma(\mathbf{w}^i \cdot \mathbf{x} - \theta_i) : c_i, \theta_i \in \mathbb{R}, \mathbf{w}^i \in \mathbb{R}^n \right\},$$

for some fixed r. On the contrary, there is generally a lower bound (for any reasonable set of functions) on the degree to which one can approximate using $\mathcal{M}_r(\sigma)$, independent of the choice of σ. (We consider this at some length in Section 6.) This is to be expected and is natural. It is, in a sense, similar to the situation with approximation by polynomials. Polynomials are dense in $C[0,1]$ but polynomials of any fixed degree are rather sparse. (Note also that the sets $\mathcal{M}_r(\sigma)$ are not subspaces. However, they do have the important property that $\mathcal{M}_r(\sigma) + \mathcal{M}_s(\sigma) = \mathcal{M}_{r+s}(\sigma)$.)

Hecht-Nielsen (1987) was perhaps the first to consider the density problem for the single hidden layer perceptron model. He premised his observations on work based on the Kolmogorov Superposition Theorem (see Section 7). While many researchers subsequently questioned the exact relevance of this result to the above model, it is certainly true that this paper very much stimulated interest in this problem. In one of the first proceedings of the IEEE on the topic of neural networks, two papers appeared which discussed the density problem. Gallant and White (1988) constructed a specific continuous, nondecreasing sigmoidal function from which it was possible to obtain any trigonometric (Fourier) series. As such their activation function, which they called a *cosine squasher*, had the density property. Irie and Miyake (1988) constructed an integral representation for any $f \in L^1(\mathbb{R}^n)$ using a kernel of the form $\sigma(\mathbf{w} \cdot \mathbf{x} - \theta)$ where σ was an arbitrary function in $L^1(\mathbb{R})$. This

allowed an interpretation in the above framework (but of course restricted to $\sigma \in L^1(\mathbb{R})$).

In 1989 there appeared four much cited papers which considered the density problem for general classes of activation functions. They are Carroll and Dickinson (1989), Cybenko (1989), Funahashi (1989), and Hornik, Stinchcombe and White (1989). Carroll and Dickinson (1989) used a discretized inverse Radon transform to approximate L^2 functions with compact support in the L^2 norm, using any continuous sigmoidal function as an activation function. The main result of Cybenko (1989) is the density property, in the uniform norm on compacta, for any continuous sigmoidal function. (Cybenko does not demand monotonicity in his definition of sigmoidality.) His method of proof uses the Hahn–Banach Theorem and the representation (Riesz Representation Theorem) of continuous linear functionals on the space of continuous functions on a compact set. Funahashi (1989) (independently of Cybenko (1989)) proves the density property, in the uniform norm on compacta, for any continuous monotone sigmoidal function. He notes that, for σ continuous, monotone and bounded, it follows that $\sigma(\cdot + \alpha) - \sigma(\cdot + \beta) \in L^1(\mathbb{R})$ for any α, β. He then applies the previously mentioned result of Irie and Miyake (1988). Hornik, Stinchcombe and White (1989), unaware of Funahashi's paper, prove very much the same result. However, they demand that their activation function be only monotone and bounded, that is, they permit noncontinuous activation functions. Their method of proof is also totally different, but somewhat circuitous. They first allow sums and products of activation functions. This permits them to apply the Stone–Weierstrass Theorem to obtain density. They then prove the desired result, without products, using cosine functions and the ability to write products of cosines as linear combinations of cosines.

There were many subsequent papers which dealt with the density problem and some related issues. We quickly review some, but not all, of them.

Stinchcombe and White (1989) prove that σ has the density property for every $\sigma \in L^1(\mathbb{R})$ with $\int_{-\infty}^{\infty} \sigma(t)\,dt \neq 0$. Cotter (1990) considers different types of models and activation functions (non-sigmoidal) for which the Stone–Weierstrass Theorem can be employed to obtain density, for instance $\sigma(t) = e^t$, and others. Jones (1990) shows, using ridge functions (which we shall soon define), that to answer the question of density it suffices to consider only the univariate problem. He then proves, by constructive methods, that a bounded (not necessarily monotone or continuous) sigmoidal activation function suffices. Stinchcombe and White (1990) also reduce the question of density to the univariate case and then consider various activation functions (not necessarily sigmoidal) such as piecewise linear (with at least one knot), a subset of polynomial splines, and a subset of analytic functions. They also consider the density question when bounding the set of permissible weights and thresholds. Hornik (1991) proves density for any

continuous bounded and nonconstant activation function, and also in other norms. Itô, in a series of papers (Itô 1991*a*, 1991*b* and 1992) studies the problem of density using monotone sigmoidal functions, with only weights of norm 1. He also considers conditions under which one obtains uniform convergence on all of $\mathbb{R}^n$. Chui and Li (1992) constructively prove density where the activation function is continuous and sigmoidal, with weights and thresholds taking only integer values. Mhaskar and Micchelli (1992) extend the density result to what they call kth *degree sigmoidal functions.* They prove that if σ is continuous, bounded by some polynomial of degree k on all of $\mathbb{R}$, and

$$\lim_{t\to-\infty}\frac{\sigma(t)}{t^k}=0, \qquad \lim_{t\to\infty}\frac{\sigma(t)}{t^k}=1,$$

then density holds if and only if σ is not a polynomial. Other results may be found in Light (1993), Chen and Chen (1993, 1995), Chen, Chen and Liu (1995), Attali and Pagès (1997) and Burton and Dehling (1998).

As we have noted, a variety of techniques were used to attack a problem which many considered important and difficult. The solution to this problem, however, turns out to be surprisingly simple. Leshno, Lin, Pinkus and Schocken (1993) prove that the necessary and sufficient condition for any continuous activation function to have the density property is that it not be a polynomial. Also considered in that paper are some sufficient conditions on noncontinuous activation functions which also imply density. For some reason the publication of this article was delayed and the submission date incorrectly reported. In a subsequent issue there appeared a paper by Hornik (1993) which references Leshno, Lin, Pinkus and Schocken (1993) and restates and reproves their results in a slightly altered form. In Pinkus (1996) a somewhat different proof is given and it is also noted that the characterization of continuous activation functions with the density property can be essentially found in Schwartz (1944) (see also Edwards (1965, pp. 130–133)). The problem is in fact very much related to that of characterizing translation (and dilation) invariant subspaces of $C(\mathbb{R})$, in the topology of uniform convergence on compacta.

As we have said, the main theorem we will prove is the following.

Theorem 3.1 Let $\sigma\in C(\mathbb{R})$. Then $\mathcal{M}(\sigma)$ is dense in $C(\mathbb{R}^n)$, in the topology of uniform convergence on compacta, if and only if σ is not a polynomial.

If σ is a polynomial, then density cannot possibly hold. This is immediate. If σ is a polynomial of degree m, then, for every choice of $\mathbf{w}\in\mathbb{R}^n$ and $\theta\in\mathbb{R}$, $\sigma(\mathbf{w}\cdot\mathbf{x}-\theta)$ is a (multivariate) polynomial of total degree at most m, and thus $\mathcal{M}(\sigma)$ is the space of all polynomials of total degree m and does not span $C(\mathbb{R}^n)$. The main content of this theorem is the converse result.

We shall prove considerably more than is stated in Theorem 3.1. We shall show that we can, in diverse cases, restrict the permissible weights and thresholds, and also enlarge the class of eligible σ, while still obtaining the desired density. The next few propositions are amalgamations of results and techniques in Leshno, Lin, Pinkus and Schocken (1993) and Schwartz (1944).

We start the analysis by defining *ridge functions*. Ridge functions are multivariate functions of the form

$$g(a_1x_1 + \cdots + a_nx_n) = g(\mathbf{a} \cdot \mathbf{x})$$

where $g : \mathbb{R} \to \mathbb{R}$ and $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ is a fixed *direction*. In other words, they are multivariate functions constant on the parallel hyperplanes $\mathbf{a} \cdot \mathbf{x} = c$, $c \in \mathbb{R}$. Ridge functions have been considered in the study of hyperbolic partial differential equations (where they go under the name of *plane waves*), computerized tomography, projection pursuit, approximation theory, and neural networks (see, for instance, Pinkus (1997) for further details).

Set

$$\mathcal{R} = \operatorname{span}\{g(\mathbf{a} \cdot \mathbf{x}) : \mathbf{a} \in \mathbb{R}^n, g : \mathbb{R} \to \mathbb{R}\}.$$

Ridge functions are relevant in the theory of the single hidden layer perceptron model since each factor $\sigma(\mathbf{w} \cdot \mathbf{x} - \theta)$ is a ridge function for every choice of σ, $\mathbf{w}$ and θ. It therefore immediately follows that a lower bound on the extent to which this model with r units in the single hidden layer can approximate any function is given by the order of approximation from the manifold

$$\mathcal{R}_r = \left\{ \sum_{i=1}^{r} g_i(\mathbf{a}^i \cdot \mathbf{x}) : \mathbf{a}^i \in \mathbb{R}^n, g_i : \mathbb{R} \to \mathbb{R}, i = 1, \ldots, r \right\}.$$

(We return to this fact in Section 6.) In addition, if ridge functions are not dense in $C(\mathbb{R}^n)$, in the above topology, then it would not be possible for $\mathcal{M}(\sigma)$ to be dense in $C(\mathbb{R}^n)$ for any choice of σ. But ridge functions do have the density property. This is easily seen. $\mathcal{R}$ contains all functions of the form $\cos(\mathbf{a} \cdot \mathbf{x})$ and $\sin(\mathbf{a} \cdot \mathbf{x})$. These functions can be shown to be dense on any compact subset of $C(\mathbb{R}^n)$. Another dense subset of ridge functions is given by $e^{\mathbf{a} \cdot \mathbf{x}}$. Moreover, the set

$$\operatorname{span}\{(\mathbf{a} \cdot \mathbf{x})^k : \mathbf{a} \in \mathbb{R}^n, k = 0, 1, \ldots\}$$

contains all polynomials and thus is dense. In fact we have the following result due to Vostrecov and Kreines (1961) (see also Lin and Pinkus (1993)), which tells us exactly which sets of directions are both sufficient and necessary for density. We will use this result.

Theorem 3.2. (Vostrecov and Kreines 1961) The set of ridge functions

$$\mathcal{R}(A) = \text{span}\{g(\mathbf{a} \cdot \mathbf{x}) : g \in C(\mathbb{R}), \mathbf{a} \in A\}$$

in dense in $C(\mathbb{R}^n)$, in the topology of uniform convergence on compacta, if and only if there is no nontrivial homogeneous polynomial that vanishes on A.

Because of the homogeneity of the directions (allowing a direction $\mathbf{a}$ is equivalent to allowing all directions $\mu\mathbf{a}$ for every real μ, since we vary over all $g \in C(\mathbb{R})$), it in fact suffices to consider directions normalized to lie on the unit ball

$$S^{n-1} = \{\mathbf{y} : \|\mathbf{y}\|_2 = (y_1^2 + \cdots + y_n^2)^{1/2} = 1\}.$$

Theorem 3.2 says that $\mathcal{R}(A)$ is dense in $C(\mathbb{R}^n)$, for $A \subseteq S^{n-1}$, if no nontrivial homogeneous polynomial has a zero set containing A. For example, if A contains an open subset of S^{n-1} then no nontrivial homogeneous polynomial vanishes on A. In what follows we will always assume that $A \subseteq S^{n-1}$.

The next proposition is a simple consequence of the ridge function form of our problem, and immediately reduces our discussion from $\mathbb{R}^n$ to the more tractable univariate $\mathbb{R}$.

In what follows, Λ, Θ will be subsets of $\mathbb{R}$. By $\Lambda \times A$ we mean the subset of $\mathbb{R}^n$ given by

$$\Lambda \times A = \{\lambda\mathbf{a} : \lambda \in \Lambda, \mathbf{a} \in A\}.$$

Proposition 3.3 Assume Λ, Θ are subsets of $\mathbb{R}$ for which

$$\mathcal{N}(\sigma; \Lambda, \Theta) = \text{span}\{\sigma(\lambda t - \theta) : \lambda \in \Lambda, \theta \in \Theta\}$$

is dense in $C(\mathbb{R})$, in the topology of uniform convergence on compacta. Assume in addition that $A \subseteq S^{n-1}$ is such that $\mathcal{R}(A)$ is dense in $C(\mathbb{R}^n)$, in the topology of uniform convergence on compacta. Then

$$\mathcal{M}(\sigma; \Lambda \times A, \Theta) = \text{span}\{\sigma(\mathbf{w} \cdot \mathbf{x} - \theta) : \mathbf{w} \in \Lambda \times A, \theta \in \Theta\}$$

is dense in $C(\mathbb{R}^n)$, in the topology of uniform convergence on compacta.

Proof. Let $f \in C(K)$ for some compact set K in $\mathbb{R}^n$. Since $\mathcal{R}(A)$ is dense in $C(K)$, given $\varepsilon > 0$ there exist $g_i \in C(\mathbb{R})$ and $\mathbf{a}^i \in A$, $i = 1, \ldots, r$ (some r) such that

$$\left| f(\mathbf{x}) - \sum_{i=1}^{r} g_i(\mathbf{a}^i \cdot \mathbf{x}) \right| < \frac{\varepsilon}{2}$$

for all $\mathbf{x} \in K$. Since K is compact, $\{\mathbf{a}^i \cdot \mathbf{x} : \mathbf{x} \in K\} \subseteq [\alpha_i, \beta_i]$ for some finite interval $[\alpha_i, \beta_i]$, $i = 1, \ldots, r$. Because $\mathcal{N}(\sigma; \Lambda, \Theta)$ is dense in $C[\alpha_i, \beta_i]$, $i = 1, \ldots, r$, there exist constants $c_{ij} \in \mathbb{R}$, $\lambda_{ij} \in \Lambda$ and $\theta_{ij} \in \Theta$, $j = 1, \ldots, m_i$,

$i = 1, \dots, r$, for which

$$\left| g_i(t) - \sum_{j=1}^{m_i} c_{ij}\sigma(\lambda_{ij}t - \theta_{ij}) \right| < \frac{\varepsilon}{2r}$$

for all $t \in [\alpha_i, \beta_i]$ and $i = 1, \dots, r$. Thus

$$\left| f(\mathbf{x}) - \sum_{i=1}^{r}\sum_{j=1}^{m_i} c_{ij}\sigma(\lambda_{ij}\mathbf{a}^i \cdot \mathbf{x} - \theta_{ij}) \right| < \varepsilon$$

for all $\mathbf{x} \in K$. □

Proposition 3.3 permits us to focus on $\mathbb{R}$. We first prove density for a restricted class of activation functions.

Proposition 3.4 Let $\sigma \in C^\infty(\mathbb{R})$ and assume σ is not a polynomial. Then $\mathcal{N}(\sigma; \mathbb{R}, \mathbb{R})$ is dense in $C(\mathbb{R})$.

Proof. It is well known (in fact it is a well-known problem given to advanced math students) that, if $\sigma \in C^\infty$ on any open interval and is not a polynomial thereon, then there exists a point $-\theta_o$ in that interval for which $\sigma^{(k)}(-\theta_o) \neq 0$ for all $k = 0, 1, 2, \dots$. The earliest reference we have found to this result is Corominas and Sunyer Balaguer (1954). It also appears in the more accessible Donoghue (1969, p. 53), but there exist simpler proofs than that which appears there.

Since $\sigma \in C^\infty(\mathbb{R})$, and $[\sigma((\lambda + h)t - \theta_o) - \sigma(\lambda t - \theta_o)]/h \in \mathcal{N}(\sigma; \mathbb{R}, \mathbb{R})$ for all $h \neq 0$, it follows that

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\sigma(\lambda t - \theta_o)\Big|_{\lambda=0} = t\sigma'(-\theta_o)$$

is contained in $\overline{\mathcal{N}(\sigma; \mathbb{R}, \mathbb{R})}$, the closure of $\mathcal{N}(\sigma; \mathbb{R}, \mathbb{R})$. By the same argument

$$\frac{\mathrm{d}^k}{\mathrm{d}\lambda^k}\sigma(\lambda t - \theta_o)\Big|_{\lambda=0} = t^k\sigma^{(k)}(-\theta_o)$$

is contained in $\overline{\mathcal{N}(\sigma; \mathbb{R}, \mathbb{R})}$ for any k. Since $\sigma^{(k)}(-\theta_o) \neq 0$, $k = 0, 1, 2, \dots$, the set $\overline{\mathcal{N}(\sigma; \mathbb{R}, \mathbb{R})}$ contains all monomials and thus all polynomials. By the Weierstrass Theorem this implies that $\mathcal{N}(\sigma; \mathbb{R}, \mathbb{R})$ is dense in $C(K)$ for every compact $K \subset \mathbb{R}$. □

Let us consider this elementary proof in more detail. What properties of the function σ and of the sets Λ and Θ of weights and thresholds, respectively, did we use? In fact we only really needed to show that

$$\frac{\mathrm{d}^k}{\mathrm{d}\lambda^k}\sigma(\lambda t - \theta_o)\Big|_{\lambda=0} = t^k\sigma^{(k)}(-\theta_o)$$

is contained in $\overline{\mathcal{N}(\sigma; \Lambda, \Theta)}$ for every k, and that $\sigma^{(k)}(-\theta_o) \neq 0$ for all k. It

therefore suffices that Λ be any set containing a sequence of values tending to zero, and $\sigma \in C^\infty(\Theta)$, where Θ contains an open interval on which σ is not a polynomial. Let us restate Proposition 3.4 in this more general form.

Corollary 3.5 Let Λ be any set containing a sequence of values tending to zero, and let Θ be any open interval. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be such that $\sigma \in C^\infty(\Theta)$, and σ is not a polynomial on Θ. Then $\mathcal{N}(\sigma; \Lambda, \Theta)$ is dense in $C(\mathbb{R})$.

We also note that the method of proof of Proposition 3.4 shows that, under these conditions, in the closure of the linear combination of $k+1$ shifts and dilations of σ are the space of polynomials of degree k. We will use this fact in Section 6. As such we state it formally here.

Corollary 3.6 Let

$$\mathcal{N}_r(\sigma) = \left\{ \sum_{i=1}^{r} c_i \sigma(\lambda_i t - \theta_i) : c_i, \lambda_i, \theta_i \in \mathbb{R} \right\}.$$

If Θ is any open interval and $\sigma \in C^\infty(\Theta)$ is not a polynomial on Θ, then $\overline{\mathcal{N}_r(\sigma)}$ contains π_{r-1}, the linear space of algebraic polynomials of degree at most $r-1$.

We now consider how to weaken our smoothness demands on σ. We do this in two steps. We again assume that $\Lambda = \Theta = \mathbb{R}$. However, this is not necessary and, following the proof of Proposition 3.8, we state the appropriate analogue of Corollary 3.5.

Proposition 3.7 Let $\sigma \in C(\mathbb{R})$ and assume σ is not a polynomial. Then $\mathcal{N}(\sigma; \mathbb{R}, \mathbb{R})$ is dense in $C(\mathbb{R})$.

Proof. Let $\phi \in C_0^\infty(\mathbb{R})$, that is, $C^\infty(\mathbb{R})$ with compact support. For each such ϕ set

$$\sigma_\phi(t) = \int_{-\infty}^{\infty} \sigma(t-y)\phi(y)\,dy,$$

that is, $\sigma_\phi = \sigma * \phi$ is the convolution of σ and ϕ. Since $\sigma, \phi \in C(\mathbb{R})$ and ϕ has compact support, the above integral converges for all t, and as is easily seen (taking Riemann sums) σ_ϕ is contained in the closure of $\mathcal{N}(\sigma; \{1\}, \mathbb{R})$. Furthermore, $\sigma_\phi \in C^\infty(\mathbb{R})$.

It also follows that $\overline{\mathcal{N}(\sigma_\phi; \mathbb{R}, \mathbb{R})}$ is contained in $\overline{\mathcal{N}(\sigma; \mathbb{R}, \mathbb{R})}$ since

$$\sigma_\phi(\lambda t - \theta) = \int_{-\infty}^{\infty} \sigma(\lambda t - \theta - y)\phi(y)\,dy,$$

for each $\lambda \in \mathbb{R}$. Because $\sigma_\phi \in C^\infty(\mathbb{R})$ we have, from the method of proof of Proposition 3.4, that $t^k \sigma_\phi^{(k)}(-\theta)$ is in $\overline{\mathcal{N}(\sigma_\phi; \mathbb{R}, \mathbb{R})}$ for all $\theta \in \mathbb{R}$ and all k.

Now if $\mathcal{N}(\sigma;\mathbb{R},\mathbb{R})$ is not dense in $C(\mathbb{R})$ then t^k is not in $\overline{\mathcal{N}(\sigma;\mathbb{R},\mathbb{R})}$ for some k. Thus t^k is not in $\overline{\mathcal{N}(\sigma_\phi;\mathbb{R},\mathbb{R})}$ for each $\phi \in C_0^\infty(\mathbb{R})$. This implies that $\sigma_\phi^{(k)}(-\theta) = 0$ for all $\theta \in \mathbb{R}$ and each $\phi \in C_0^\infty(\mathbb{R})$. Thus σ_ϕ is a polynomial of degree at most $k-1$ for each $\phi \in C_0^\infty(\mathbb{R})$.

It is well known that there exist sequences of $\phi_n \in C_0^\infty(\mathbb{R})$ for which σ_{ϕ_n} converges to σ uniformly on any compact set in $\mathbb{R}$. We can, for example, take what are called *mollifiers* (see, for instance, Adams (1975, p. 29)). Polynomials of a fixed degree form a (closed) finite-dimensional linear subspace. Since σ_{ϕ_n} is a polynomial of degree at most $k-1$ for every ϕ_n, it therefore follows that σ is a polynomial of degree at most $k-1$. This contradicts our assumption. □

We first assumed $\sigma \in C^\infty(\mathbb{R})$ and then showed how to obtain the same result for $\sigma \in C(\mathbb{R})$. We now consider a class of discontinuous σ. We prove that the same result (density) holds for any σ that is bounded and Riemann-integrable on every finite interval. (By a theorem of Lebesgue, the property of Riemann-integrability for such functions is equivalent to demanding that the set of discontinuities of σ has Lebesgue measure zero: see, for instance, Royden (1963, p. 70).) It is not true that, for arbitrary σ, the space $\mathcal{N}(\sigma;\mathbb{R},\mathbb{R})$ is dense in $C(\mathbb{R})$ if σ is not a polynomial, without some smoothness conditions on σ.

Proposition 3.8 Assume $\sigma : \mathbb{R} \to \mathbb{R}$ is bounded and Riemann-integrable on every finite interval. Assume σ is not a polynomial (almost everywhere). Then $\mathcal{N}(\sigma;\mathbb{R},\mathbb{R})$ is dense in $C(\mathbb{R})$.

Proof. It remains true that, for each $\phi \in C_0^\infty(\mathbb{R})$,

$$\sigma_\phi(t) = \int_{-\infty}^{\infty} \sigma(t-y)\phi(y)\,dy$$

is in $C^\infty(\mathbb{R})$. Furthermore, for the σ_{ϕ_n} as defined in Proposition 3.7 we have that

$$\lim_{n\to\infty} \|\sigma - \sigma_{\phi_n}\|_{L^p(K)} = 0$$

for every $1 \le p < \infty$ and any compact K (see, for instance, Adams (1975, p. 30)). As such, if σ_{ϕ_n} is a polynomial of degree at most $k-1$ for each n, then σ is (almost everywhere) also a polynomial of degree at most $k-1$.

Thus the proof of this proposition exactly follows the method of proof of Proposition 3.7 if we can show that σ_ϕ is in the closure of $\mathcal{N}(\sigma;\{1\},\mathbb{R})$ for each $\phi \in C_0^\infty(\mathbb{R})$. This is what we now prove.

Let $\phi \in C_0^\infty(\mathbb{R})$ and assume that ϕ has support in $[-\alpha,\alpha]$. Set

$$y_i = -\alpha + \frac{2i\alpha}{m}, \qquad i = 0,1,\ldots,m,$$

$\Delta_i = [y_{i-1}, y_i]$, and $\Delta y_i = y_i - y_{i-1} = 2\alpha/m$, $i = 1, \ldots, m$. By definition,

$$\sum_{i=1}^{m} \sigma(t - y_i)\phi(y_i)\Delta y_i \in \mathcal{N}(\sigma; \{1\}, \mathbb{R})$$

for each m. We will prove that the above sum uniformly converges to σ_ϕ on every compact subset K of $\mathbb{R}$.

By definition,

$$\begin{aligned}
&\left|\sigma_\phi(t) - \sum_{i=1}^{m} \sigma(t - y_i)\phi(y_i)\Delta y_i\right| \\
&\quad = \sum_{i=1}^{m} \int_{\Delta_i} [\sigma(t - y)\phi(y) - \sigma(t - y_i)\phi(y_i)]\,\mathrm{d}y \\
&\quad = \sum_{i=1}^{m} \int_{\Delta_i} [\sigma(t - y) - \sigma(t - y_i)]\phi(y)\,\mathrm{d}y \\
&\qquad + \sum_{i=1}^{m} \int_{\Delta_i} \sigma(t - y_i)[\phi(y) - \phi(y_i)]\,\mathrm{d}y.
\end{aligned}$$

Since σ is bounded on $K - [-\alpha, \alpha]$, and ϕ is uniformly continuous on $[-\alpha, \alpha]$, it easily follows that

$$\lim_{m\to\infty} \sum_{i=1}^{m} \int_{\Delta_i} \sigma(t - y_i)[\phi(y) - \phi(y_i)]\,\mathrm{d}y = 0.$$

Now

$$\begin{aligned}
&\left|\sum_{i=1}^{m} \int_{\Delta_i} [\sigma(t - y) - \sigma(t - y_i)]\phi(y)\,\mathrm{d}y\right| \\
&\qquad \leq \|\phi\|_{L^\infty[-\alpha,\alpha]} \sum_{i=1}^{m} \left[\sup_{y\in\Delta_i} \sigma(t - y) - \inf_{y\in\Delta_i} \sigma(t - y)\right] \frac{2\alpha}{m}.
\end{aligned}$$

Since σ is Riemann-integrable on $K - [-\alpha, \alpha]$, it follows that

$$\lim_{m\to\infty} \sum_{i=1}^{m} \left[\sup_{y\in\Delta_i} \sigma(t - y) - \inf_{y\in\Delta_i} \sigma(t - y)\right] \frac{2\alpha}{m} = 0.$$

This proves the result. □

It is not difficult to check that the above conditions only need to hold locally, as in Corollary 3.5.

Corollary 3.9 Let Λ be any set containing a sequence of values tending to zero, and let Θ be any open interval. Assume $\sigma : \mathbb{R} \to \mathbb{R}$ is such that

σ is bounded and Riemann-integrable on Θ and not a polynomial (almost everywhere) on Θ. Then $\mathcal{N}(\sigma; \Lambda, \Theta)$ is dense in $C(\mathbb{R})$.

The above results should not be taken to mean that we recommend using only a minimal set of weights and thresholds. Such a strategy would be wrong.

In the cases thus far considered it was necessary, because of the method of proof, that we allow dilations (*i.e.*, the set Λ) containing a sequence tending to zero. This is in fact not necessary. We have, for example, the following simple result, which is proven by classical methods.

Proposition 3.10 Assume $\sigma \in C(\mathbb{R}) \cap L^1(\mathbb{R})$, or σ is continuous, nondecreasing and bounded (but not the constant function). Then $\mathcal{N}(\sigma; \{1\}, \mathbb{R})$ is dense in $C(\mathbb{R})$.

Proof. Assume $\sigma \in C(\mathbb{R}) \cap L^1(\mathbb{R})$. Continuous linear functionals on $C(\mathbb{R})$ are represented by Borel measures of finite total variation and compact support. If $\mathcal{N}(\sigma; \{1\}, \mathbb{R})$ is not dense in $C(\mathbb{R})$, then there exists such a nontrivial measure μ satisfying

$$\int_{-\infty}^{\infty} \sigma(t - \theta)\, \mathrm{d}\mu(t) = 0$$

for all $\theta \in \mathbb{R}$. Both σ and μ have 'nice' Fourier transforms. Since the above is a convolution this implies

$$\widehat{\sigma}(\omega)\widehat{\mu}(\omega) = 0$$

for all $\omega \in \mathbb{R}$. Now $\widehat{\mu}$ is an entire function (of exponential type), while $\widehat{\sigma}$ is continuous. Since $\widehat{\sigma}$ must vanish where $\widehat{\mu} \neq 0$, it follows that $\widehat{\sigma} = 0$ and thus $\sigma = 0$. This is a contradiction and proves the result.

If σ is continuous, nondecreasing and bounded (but not the constant function), then $\sigma(\cdot + a) - \sigma(\cdot)$ is in $C(\mathbb{R}) \cap L^1(\mathbb{R})$ (and not the zero function) for any fixed $a \neq 0$. We can now apply the result of the previous paragraph to obtain the desired result. □

The above proposition does not begin to tell the full story. A more formal study of $\mathcal{N}(\sigma; \{1\}, \mathbb{R})$ was made by Schwartz (1947), where he introduced the following definition of the class of *mean-periodic* functions.

Definition. A function $f \in C(\mathbb{R}^n)$ is said to be *mean-periodic* if

$$\operatorname{span}\{f(\mathbf{x} - \mathbf{a}) : \mathbf{a} \in \mathbb{R}^n\}$$

is *not* dense in $C(\mathbb{R}^n)$, in the topology of uniform convergence on compacta.

Translation-invariant subspaces (such as the above space) have been much studied in various norms (more especially L^2 and L^1). The study of mean-periodic functions was an attempt to provide a parallel analysis for the space

$C(\mathbb{R}^n)$. Unfortunately this subject is still not well understood for $n > 1$. Luckily we are interested in the univariate case and Schwartz (1947) provided a thorough analysis of such spaces (see also Kahane (1959)). The theory of mean-periodic functions is, unfortunately, too complicated to present here with proofs. The central result is that subspaces

$$\text{span}\{f(t-a) : a \in \mathbb{R}\}$$

spanned by mean-periodic functions in $C(\mathbb{R})$ are totally characterized by the functions of the form $t^m e^{\gamma t}$ which are contained in their closure, where $\gamma \in \mathbb{C}$. (These values γ determine the spectrum of f. Note that if γ is in the spectrum, then so is $\overline{\gamma}$.) From this fact follows this next result.

Proposition 3.11 Let $\sigma \in C(\mathbb{R})$, and assume that σ is not a polynomial. For any Λ that contains a sequence tending to a finite limit point, the set $\mathcal{N}(\sigma; \Lambda, \mathbb{R})$ is dense in $C(\mathbb{R})$.

Proof. Let $\delta \in \Lambda \backslash \{0\}$. If $\sigma(\delta t)$ is not mean-periodic then

$$\text{span}\{\sigma(\delta t - \theta) : \theta \in \mathbb{R}\}$$

is dense in $C(\mathbb{R})$, and we are finished. Assume not. Since σ is not a polynomial the above span contains, in its closure, $t^m e^{\gamma t}$ for some nonnegative integer m and $\gamma \in \mathbb{C}\backslash\{0\}$. (We may assume $m = 0$ since, by taking a finite linear combination of shifts, it follows that $e^{\gamma t}$ is also contained in the above closure.) Thus the closure of

$$\text{span}\{\sigma(\lambda t - \theta) : \theta \in \mathbb{R},\ \lambda \in \Lambda\}$$

contains $e^{(\gamma\lambda/\delta)t}$ for every $\lambda \in \Lambda$.

We claim that

$$\text{span}\{e^{(\gamma\lambda/\delta)t} : \lambda \in \Lambda\}$$

is dense in $C(\mathbb{R})$ if Λ has a finite limit point. This is a well-known result. One can prove it by the method of proof of Proposition 3.4. Alternatively, if the above span is not dense then

$$\int_{-\infty}^{\infty} e^{(\gamma\lambda/\delta)t}\, \mathrm{d}\mu(t) = 0, \qquad \lambda \in \Lambda,$$

for some nontrivial Borel measure μ of finite total variation and compact support. Now

$$g(z) = \int_{-\infty}^{\infty} e^{zt}\, \mathrm{d}\mu(t)$$

is an entire function on $\mathbb{C}$. But g vanishes on the set $\{\gamma\lambda/\delta : \lambda \in \Lambda\}$, and this set contains a sequence tending to a finite limit point. This implies that g is identically zero, which in turn implies that μ is the zero measure. This contradiction proves the density. □

Remark. As may be noted from the method of proof of Proposition 3.11, the condition on Λ can be replaced by the demand that Λ not be contained in the zero set of a nontrivial entire function.

We should also mention that Schwartz (1947, p. 907) proved the following result.

Proposition 3.12 Let $\sigma \in C(\mathbb{R})$. If $\sigma \in L^p(\mathbb{R})$, $1 \leq p < \infty$, or σ is bounded and has a limit at infinity or minus infinity, but is not the constant function, then σ is not mean-periodic.

Thus, in the above cases $\mathcal{N}(\sigma; \{\lambda\}, \mathbb{R})$ is dense in $C(\mathbb{R})$ for any $\lambda \neq 0$.

Remark. If the input is preprocessed, then, rather than working directly with the input $\mathbf{x} = (x_1, \ldots, x_n)$, this data is first converted to $\mathbf{h}(\mathbf{x}) = (h_1(\mathbf{x}), \ldots, h_m(\mathbf{x}))$ for some given fixed continuous functions $h_j \in C(\mathbb{R}^n)$, $j = 1, \ldots, m$. Set

$$\mathcal{M}_{\mathbf{h}}(\sigma) = \operatorname{span}\{\sigma(\mathbf{w} \cdot \mathbf{h}(\mathbf{x}) - \theta) : \mathbf{w} \in \mathbb{R}^m, \theta \in \mathbb{R}\}.$$

Theorem 3.1 is still valid in this setting if and only if $\mathbf{h}$ separates points, that is, $\mathbf{x}^i \neq \mathbf{x}^j$ implies $\mathbf{h}(\mathbf{x}^i) \neq \mathbf{h}(\mathbf{x}^j)$ (see Lin and Pinkus (1994)). Analogues of the other results of this section depend upon the explicit form of $\mathbf{h}$.

4. Derivative approximation

In this section we consider conditions under which a neural network in the single hidden layer perceptron model can simultaneously and uniformly approximate a function and various of its partial derivatives. This fact is requisite in several algorithms.

We first introduce some standard multivariate notation. We let $\mathbb{Z}_+^n$ denote the lattice of nonnegative multi-integers in $\mathbb{R}^n$. For $\mathbf{m} = (m_1, \ldots, m_n) \in \mathbb{Z}_+^n$, we set $|\mathbf{m}| = m_1 + \cdots + m_n$, $\mathbf{x}^{\mathbf{m}} = x_1^{m_1} \cdots x_n^{m_n}$, and

$$D^{\mathbf{m}} = \frac{\partial^{|\mathbf{m}|}}{\partial x_1^{m_1} \cdots \partial x_n^{m_n}}.$$

If q is a polynomial, then by $q(D)$ we mean the differential operator given by

$$q\left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}\right).$$

We also have the usual ordering on $\mathbb{Z}_+^n$, namely $\mathbf{m}^1 \leq \mathbf{m}^2$ if $m_i^1 \leq m_i^2$, $i = 1, \ldots, n$.

We say $f \in C^{\mathbf{m}}(\mathbb{R}^n)$ if $D^{\mathbf{k}} f \in C(\mathbb{R}^n)$ for all $\mathbf{k} \leq \mathbf{m}$, $\mathbf{k} \in \mathbb{Z}_+^n$. We set

$$C^{\mathbf{m}^1, \ldots, \mathbf{m}^s}(\mathbb{R}^n) = \bigcap_{j=1}^{s} C^{\mathbf{m}^j}(\mathbb{R}^n),$$

and, as a special case,

$$C^m(\mathbb{R}^n) = \bigcap_{|\mathbf{m}|=m} C^{\mathbf{m}}(\mathbb{R}^n) = \{f : D^{\mathbf{k}} f \in C(\mathbb{R}^n) \text{ for all } |\mathbf{k}| \leq m\}.$$

We recall that

$$\mathcal{M}(\sigma) = \text{span}\{\sigma(\mathbf{w} \cdot \mathbf{x} - \theta) : \mathbf{w} \in \mathbb{R}^n, \theta \in \mathbb{R}\}.$$

We say that $\mathcal{M}(\sigma)$ is dense in $C^{\mathbf{m}^1,\ldots,\mathbf{m}^s}(\mathbb{R}^n)$ if, for any $f \in C^{\mathbf{m}^1,\ldots,\mathbf{m}^s}(\mathbb{R}^n)$, any compact K of $\mathbb{R}^n$, and any $\varepsilon > 0$, there exists a $g \in \mathcal{M}(\sigma)$ satisfying

$$\max_{\mathbf{x}\in K} |D^{\mathbf{k}} f(\mathbf{x}) - D^{\mathbf{k}} g(\mathbf{x})| < \varepsilon,$$

for all $\mathbf{k} \in \mathbb{Z}_+^n$ for which $\mathbf{k} \leq \mathbf{m}^i$ for some i.

We will outline a proof (skipping over various details) of the following result.

Theorem 4.1 Let $\mathbf{m}^i \in \mathbb{Z}_+^n$, $i = 1, \ldots, s$, and set $m = \max\{|\mathbf{m}^i| : i = 1, \ldots, s\}$. Assume $\sigma \in C^m(\mathbb{R})$ and σ is not a polynomial. Then $\mathcal{M}(\sigma)$ is dense in $C^{\mathbf{m}^1,\ldots,\mathbf{m}^s}(\mathbb{R}^n)$.

This density question was first considered by Hornik, Stinchcombe and White (1990). They showed that, if $\sigma^{(m)} \in C(\mathbb{R}) \cap L^1(\mathbb{R})$, then $\mathcal{M}(\sigma)$ is dense in $C^m(\mathbb{R}^n)$. Subsequently Hornik (1991) generalized this to $\sigma \in C^m(\mathbb{R})$ which is bounded, but not the constant function. Hornik uses a functional analytic method of proof. With suitable modifications his method of proof can be applied to prove Theorem 4.1. Itô (1993) reproves Hornik's result, but for $\sigma \in C^\infty(\mathbb{R})$ which is not a polynomial. His method of proof is different. We essentially follow it here. This approach is very similar to the approach taken in Li (1996) where Theorem 4.1 can effectively be found. Other papers concerned with this problem are Cardaliaguet and Euvrard (1992), Gallant and White (1992), Itô (1994*b*), Mhaskar and Micchelli (1995) and Attali and Pagès (1997). Some of these papers contain generalizations to density in other norms, and related questions.

Proof. Polynomials are dense in $C^{\mathbf{m}^1,\ldots,\mathbf{m}^s}(\mathbb{R}^n)$. This may be shown in a number of ways. One proof of this fact is to be found in Li (1996). It therefore suffices to prove that one can approximate polynomials in the appropriate norm.

If h is any polynomial on $\mathbb{R}^n$, then h can be represented in the form

$$h(\mathbf{x}) = \sum_{i=1}^{r} p_i(\mathbf{a}^i \cdot \mathbf{x}) \tag{4.1}$$

for some choice of r, $\mathbf{a}^i \in \mathbb{R}^n$, and univariate polynomials p_i, $i = 1, \ldots, r$.

A precise proof of this fact is the following. (This result will be used again in Section 6, so we detail its proof here.) Let H_k denote the linear space of homogeneous polynomials of degree k (in $\mathbb{R}^n$), and $P_k = \cup_{s=0}^k H_s$ the linear space of polynomials of degree at most k. Set $r = \binom{n-1+k}{k} = \dim H_k$. Let $\mathbf{m}^1, \mathbf{m}^2 \in \mathbb{Z}_+^n$, $|\mathbf{m}^1| = |\mathbf{m}^2| = k$. Then $D^{\mathbf{m}^1}\mathbf{x}^{\mathbf{m}^2} = C_{\mathbf{m}^1}\delta_{\mathbf{m}^1,\mathbf{m}^2}$, for some easily calculated $C_{\mathbf{m}^1}$. This implies that each linear functional L on H_k may be represented by some $q \in H_k$ via

$$L(p) = q(D)p$$

for each $p \in H_k$. Now $(\mathbf{a}\cdot\mathbf{x})^k \in H_k$ and $D^{\mathbf{m}}(\mathbf{a}\cdot\mathbf{x})^k = k!\mathbf{a}^{\mathbf{m}}$ if $|\mathbf{m}| = k$. Thus $q(D)(\mathbf{a}\cdot\mathbf{x})^k = k!q(\mathbf{a})$.

Since $r = \dim H_k$, there exist r points $\mathbf{a}^1, \ldots, \mathbf{a}^r$ such that $\dim H_k|_A = r$ for $A = \{\mathbf{a}^1, \ldots, \mathbf{a}^r\}$. We claim that $\{(\mathbf{a}^i \cdot \mathbf{x})^k\}_{i=1}^r$ span H_k. If not, there exists a nontrivial linear functional that annihilates each $(\mathbf{a}^i \cdot \mathbf{x})^k$. Thus some nontrivial $q \in H_k$ satisfies

$$0 = q(D)(\mathbf{a}^i \cdot \mathbf{x})^k = k!q(\mathbf{a}^i), \qquad i = 1, \ldots, r.$$

This contradicts our choice of A, hence $\{(\mathbf{a}^i\cdot\mathbf{x})^k\}_{i=1}^r$ span H_k. It also follows that $\{(\mathbf{a}^i\cdot\mathbf{x})^s\}_{i=1}^r$ spans H_s for each $s = 0, 1, \ldots, k$. If not, then there exists a nontrivial $q \in H_s$ that vanishes on A. But, for any $p \in H_{k-s}$, the function $pq \in H_k$ vanishes on A, which is a contradiction. Thus

$$P_k = \operatorname{span}\{(\mathbf{a}^i \cdot \mathbf{x})^s : i = 1, \ldots, r,\ s = 0, 1, \ldots, k\}.$$

Let π_k denote the linear space of univariate polynomials of degree at most k. It therefore follows that

$$P_k = \left\{ \sum_{i=1}^r p_i(\mathbf{a}^i \cdot \mathbf{x}) : p_i \in \pi_k,\ i = 1, \ldots, r \right\}.$$

Thus h may be written in the form (4.1). Hence it follows that it suffices (see the proof of Proposition 3.3) to prove that one can approximate each univariate polynomial p on any finite interval $[\alpha, \beta]$ from

$$\mathcal{N}(\sigma; \mathbb{R}, \mathbb{R}) = \operatorname{span}\{\sigma(\lambda t - \theta) : \lambda, \theta \in \mathbb{R}\}$$

in the norm

$$\|f\|_{C^m[\alpha,\beta]} = \max_{k=0,1,\ldots,m} \max_{t\in[\alpha,\beta]} |f^{(k)}(t)|.$$

Since $\sigma \in C^m(\mathbb{R})$ is not a polynomial we have, from the results of Section 3, that $\mathcal{N}(\sigma^{(m)}; \mathbb{R}, \mathbb{R})$ is dense in $C(\mathbb{R})$. Let $f \in C^m(\mathbb{R})$. Then, given $\varepsilon > 0$, there exists a $g \in \mathcal{N}(\sigma; \mathbb{R}, \mathbb{R})$ such that

$$\|f^{(m)} - g^{(m)}\|_{C^m[\alpha,\beta]} < \varepsilon.$$

If every polynomial of degree at most $m-1$ is in the closure of $\mathcal{N}(\sigma;\mathbb{R},\mathbb{R})$ with respect to the norm $\|\cdot\|_{C^m[\alpha,\beta]}$, then, by choosing a polynomial p satisfying

$$p^{(k)}(\alpha) = (f-g)^{(k)}(\alpha), \qquad k = 0, 1, \ldots, m-1,$$

it follows, integrating m times, that $g+p$ is close to f in the norm $\|\cdot\|_{C^m[\alpha,\beta]}$. This follows by iterating the inequality

$$\begin{aligned} |f^{(k-1)}(x) - (g+p)^{(k-1)}(x)| &= \left| \int_\alpha^x [f^{(k)}(t) - (g+p)^{(k)}(t)]\, \mathrm{d}t \right| \\ &\leq (\beta - \alpha) \max_{\alpha \leq t \leq \beta} |f^{(k)}(t) - (g+p)^{(k)}(t)|. \end{aligned}$$

We have thus reduced our problem to proving that each of $1, t, \ldots, t^{m-1}$ is in the closure of $\mathcal{N}(\sigma;\mathbb{R},\mathbb{R})$ with respect to the norm $\|\cdot\|_{C^m[\alpha,\beta]}$.

Because $\sigma \in C^m(\mathbb{R})$ it follows from the method of proof of Proposition 3.4 that for $k \leq m-1$, the function $t^k \sigma^{(k)}(-\theta_o)$ is contained in the closure of $\mathcal{N}(\sigma;\mathbb{R},\mathbb{R})$ with respect to the usual uniform norm $\|\cdot\|_{C[\alpha,\beta]}$ on any $[\alpha,\beta]$ (and since σ is not a polynomial there exists a θ_o for which $\sigma^{(k)}(-\theta_o) \neq 0$). A detailed analysis, which we will skip, proves that t^k, $k \leq m-1$, is contained in the closure of $\mathcal{N}(\sigma,\mathbb{R},\mathbb{R})$ with respect to the more stringent norm $\|\cdot\|_{C^m[\alpha,\beta]}$. □

In the above we have neglected the numerous possible nuances which parallel those contained in Section 3 (see, for instance, Corollary 3.5, Propositions 3.10 and 3.11).

5. Interpolation

The ability to approximate well is related to the ability to interpolate. If one can approximate well, then one expects to be able to interpolate (the inverse need not, in general, hold). Let us pose this problem more precisely in our setting.

Assume we are given $\sigma \in C(\mathbb{R})$. For k distinct points $\{\mathbf{x}^i\}_{i=1}^k \subset \mathbb{R}^n$, and associated data $\{\alpha_i\}_{i=1}^k \subset \mathbb{R}$, can we always find m, $\{\mathbf{w}^j\}_{j=1}^m \subset \mathbb{R}^n$, and $\{c_j\}_{j=1}^m, \{\theta_j\}_{j=1}^m \subset \mathbb{R}$ for which

$$\sum_{j=1}^m c_j \sigma(\mathbf{w}^j \cdot \mathbf{x}^i - \theta_j) = \alpha_i, \qquad \text{for } i = 1, \ldots, k?$$

Furthermore, what is the relationship between k and m?

This problem has been considered, for example, in Sartori and Antsaklis (1991), Itô (1996), Itô and Saito (1996), and Huang and Babri (1998). In Itô and Saito (1996) it is proven that, if σ is sigmoidal, continuous and nondecreasing, one can always interpolate with $m = k$ and some $\{\mathbf{w}^j\}_{j=1}^m \subset S^{n-1}$.

Huang and Babri (1998) extend this result to any bounded, continuous, nonlinear σ which has a limit at infinity or minus infinity (but their $\mathbf{w}^j$ are not restricted in any way).

We will use a technique from Section 3 to prove the following result.

Theorem 5.1 Let $\sigma \in C(\mathbb{R})$ and assume σ is not a polynomial. For any k distinct points $\{\mathbf{x}^i\}_{i=1}^k \subset \mathbb{R}^n$ and associated data $\{\alpha_i\}_{i=1}^k \subset \mathbb{R}$, there exist $\{\mathbf{w}^j\}_{j=1}^k \subset \mathbb{R}^n$, and $\{c_j\}_{j=1}^k, \{\theta_j\}_{j=1}^k \subset \mathbb{R}$ such that

$$\sum_{j=1}^k c_j \sigma(\mathbf{w}^j \cdot \mathbf{x}^i - \theta_j) = \alpha_i, \qquad i = 1, \ldots, k. \tag{5.1}$$

Furthermore, if σ is not mean-periodic, then we may choose $\{\mathbf{w}^j\}_{j=1}^k \subset S^{n-1}$.

Proof. Let $\mathbf{w} \in \mathbb{R}^n$ be any vector for which the $\mathbf{w} \cdot \mathbf{x}^i = t_i$ are distinct, $i = 1, \ldots, k$. Set $\mathbf{w}^j = \lambda_j \mathbf{w}$ for $\lambda_j \in \mathbb{R}$, $j = 1, \ldots, k$. We fix the above $\mathbf{w}$ and vary the λ_j. We will have proven (5.1) if we can show the existence of $\{c_j\}_{j=1}^k$, $\{\lambda_j\}_{j=1}^k$ and $\{\theta_j\}_{j=1}^k$ satisfying

$$\sum_{j=1}^k c_j \sigma(\lambda_j t_i - \theta_j) = \alpha_i, \qquad i = 1, \ldots, k. \tag{5.2}$$

Solving (5.2) is equivalent to proving the linear independence (over λ and θ) of the k continuous functions $\sigma(\lambda t_i - \theta)$, $i = 1, \ldots, k$. If these functions are linearly independent there exist λ_j, θ_j, $j = 1, \ldots, k$, for which

$$\det\left(\sigma(\lambda_j t_i - \theta_j)\right)_{i,j=1}^k \neq 0$$

and then (5.2) can be solved, with these $\{\lambda_j\}_{j=1}^k$ and $\{\theta_j\}_{j=1}^k$, for any choice of $\{\alpha_i\}_{i=1}^k$. If, on the other hand, they are linearly dependent then there exist nontrivial coefficients $\{d_i\}_{i=1}^k$ for which

$$\sum_{i=1}^k d_i \sigma(\lambda t_i - \theta) = 0, \tag{5.3}$$

for all $\lambda, \theta \in \mathbb{R}$.

We rewrite (5.3) in the form

$$\int_{-\infty}^{\infty} \sigma(\lambda t - \theta)\, \mathrm{d}\widetilde{\mu}(t) = 0 \tag{5.4}$$

for all $\lambda, \theta \in \mathbb{R}$ with the measure

$$\mathrm{d}\widetilde{\mu} = \sum_{i=1}^k d_i \delta_{t_i}$$

(δ_{t_i} is the measure with point mass 1 at t_i). The measure $d\tilde{\mu}$ is a nontrivial Borel measure of finite total variation and compact support. In other words, it represents a nontrivial linear functional on $C(\mathbb{R})$. We have constructed, in (5.4), a nontrivial linear functional annihilating $\sigma(\lambda t - \theta)$ for all $\lambda, \theta \in \mathbb{R}$. This implies that

$$\text{span}\{\sigma(\lambda t - \theta) : \lambda, \theta \in \mathbb{R}\}$$

is not dense in $C(\mathbb{R})$, which contradicts Proposition 3.7. This proves Theorem 5.1 in this case.

If σ is not mean-periodic, then

$$\text{span}\{\sigma(t - \theta) : \theta \in \mathbb{R}\}$$

is dense in $C(\mathbb{R})$. As above this implies that the $\{\sigma(t_i - \theta)\}_{i=1}^k$ are linearly independent for every choice of distinct $\{t_i\}_{i=1}^k$. Thus, for any $\mathbf{w} \in S^{n-1}$ for which the $\mathbf{w} \cdot \mathbf{x}^i = t_i$ are distinct, $i = 1, \ldots, k$, there exist $\{\theta_j\}_{j=1}^k$ such that

$$\det\left(\sigma(\mathbf{w} \cdot \mathbf{x}^i - \theta_j)\right)_{i,j=1}^k \neq 0.$$

Choosing $\mathbf{w}^j = \mathbf{w}$, $j = 1, \ldots, k$, and the above $\{\theta_j\}_{j=1}^k$, we can solve (5.1). □

If σ is a polynomial, then whether we can or cannot interpolate depends upon the choice of the points $\{\mathbf{x}^i\}_{i=1}^k$ and on the degree of σ. If σ is a polynomial of exact degree r, then

$$\text{span}\{\sigma(\mathbf{w} \cdot \mathbf{x} - \theta) : \mathbf{w} \in S^{n-1}, \theta \in \mathbb{R}\}$$

is precisely the space of multivariate polynomials of total degree at most r.

6. Degree of approximation

For a given activation function σ we set, for each r,

$$\mathcal{M}_r(\sigma) = \left\{ \sum_{i=1}^r c_i \sigma(\mathbf{w}^i \cdot \mathbf{x} - \theta_i) : c_i, \theta_i \in \mathbb{R},\ \mathbf{w}^i \in \mathbb{R}^n \right\}.$$

We know, based on the results of Section 3, that if $\sigma \in C(\mathbb{R})$ is not a polynomial then to each $f \in C(K)$ (K a compact subset of $\mathbb{R}^n$) there exist $g_r \in \mathcal{M}_r(\sigma)$ for which

$$\lim_{r \to \infty} \max_{\mathbf{x} \in K} |f(\mathbf{x}) - g_r(\mathbf{x})| = 0.$$

However, this tells us nothing about the rate of approximation. Nor does it tell us if there is a method, reasonable or otherwise, for finding 'good' approximants. It is these questions, and more especially the first, which we will address in this section.

We first fix some additional notation. Let B^n denote the unit ball in $\mathbb{R}^n$, that is,

$$B^n = \{\mathbf{x} : \|\mathbf{x}\|_2 = (x_1^2 + \cdots + x_n^2)^{1/2} \leq 1\}.$$

In this section we approximate functions defined on B^n. $C^m(B^n)$ will denote the set of all functions f defined on B^n for which $D^{\mathbf{k}}f$ is defined and continuous on B^n for all $\mathbf{k} \in \mathbb{Z}_+^n$ satisfying $|\mathbf{k}| \leq m$ (see Section 4). The Sobolev space $\mathcal{W}_p^m = \mathcal{W}_p^m(B^n)$ may be defined as the completion of $C^m(B^n)$ with respect to the norm

$$\|f\|_{m,p} = \begin{cases} (\sum_{0\leq|\mathbf{k}|\leq m} \|D^{\mathbf{k}}f\|_p^p)^{1/p}, & 1 \leq p < \infty, \\ \max_{0\leq|\mathbf{k}|\leq m} \|D^{\mathbf{k}}f\|_\infty, & p = \infty \end{cases}$$

or some equivalent norm thereon. Here

$$\|g\|_p = \begin{cases} (\int_{B^n} |g(\mathbf{x})|^p \, d\mathbf{x})^{1/p}, & 1 \leq p < \infty, \\ \text{ess sup}_{\mathbf{x}\in B^n} |g(\mathbf{x})|, & p = \infty. \end{cases}$$

We set $\mathcal{B}_p^m = \mathcal{B}_p^m(B^n) = \{f : f \in \mathcal{W}_p^m, \|f\|_{m,p} \leq 1\}$. Since B^n is compact and $C(B^n)$ is dense in $L_p = L_p(B^n)$, we have that $\mathcal{M}(\sigma)$ is dense in L_p for each $\sigma \in C(\mathbb{R})$ that is not a polynomial.

We will first consider some lower bounds on the degree to which one can approximate from $\mathcal{M}_r(\sigma)$. As mentioned in Section 3, for any choice of $\mathbf{w} \in \mathbb{R}^n$, $\theta \in \mathbb{R}$, and function σ, each factor

$$\sigma(\mathbf{w} \cdot \mathbf{x} - \theta)$$

is a ridge function. Set

$$\mathcal{R}_r = \left\{ \sum_{i=1}^r g_i(\mathbf{a}^i \cdot \mathbf{x}) : \mathbf{a}^i \in \mathbb{R}^n, g_i \in C(\mathbb{R}), i = 1, \ldots, r \right\}.$$

Since $\mathcal{M}_r(\sigma) \subseteq \mathcal{R}_r$ for any $\sigma \in C(\mathbb{R})$, it therefore follows that, for every norm $\|\cdot\|_X$ on a normed linear space X containing $\mathcal{R}_r$,

$$E(f; \mathcal{M}_r(\sigma); X) = \inf_{g\in\mathcal{M}_r(\sigma)} \|f - g\|_X \geq \inf_{g\in\mathcal{R}_r} \|f - g\|_X = E(f; \mathcal{R}_r; X). \quad (6.1)$$

Can we estimate the right-hand side of (6.1) from below in some reasonable way? And if so, how relevant is this lower bound?

Maiorov (1999) has proved the following lower bound. Assume $m \geq 1$ and $n \geq 2$. Then for each r there exists an $f \in \mathcal{B}_2^m$ for which

$$E(f; \mathcal{R}_r; L_2) \geq Cr^{-m/(n-1)}. \quad (6.2)$$

Here, and throughout, C is some generic positive constant independent of the things it should be independent of! (In this case, C is independent of f and r.) The case $n = 2$ may be found in Oskolkov (1997). Maiorov also

proves that for each $f \in \mathcal{B}_2^m$

$$E(f; \mathcal{R}_r; L_2) \leq Cr^{-m/(n-1)}. \tag{6.3}$$

Thus he obtains the following result.

Theorem 6.1. (Maiorov 1999) For each $n \geq 2$ and $m \geq 1$,

$$E(\mathcal{B}_2^m; \mathcal{R}_r; L_2) = \sup_{f \in \mathcal{B}_2^m} E(f; \mathcal{R}_r; L_2) \asymp r^{-m/(n-1)}.$$

To be somewhat more precise, Maiorov (1999) proves the above result for $\mathcal{B}_2^m$ for all $m > 0$, and not only integer m (the definition of $\mathcal{B}_2^m$ for such m is then somewhat different). In addition, Maiorov, Meir and Ratsaby (1999) show that the set of functions for which the lower bound (6.2) holds is of large measure. In other words, this is not simply a worst case result.

The proof of this lower bound is too difficult and complicated to be presented here. However, the proof of the upper bound is more elementary and standard, and will be used again in what follows. As such we exhibit it here. It is also valid for every $p \in [1, \infty]$.

Theorem 6.2 For each $p \in [1, \infty]$ and every $m \geq 1$ and $n \geq 2$,

$$E(\mathcal{B}_p^m; \mathcal{R}_r; L_p) \leq Cr^{-m/(n-1)},$$

where C is some constant independent of r.

Proof. As in the proof of Theorem 4.1, let H_k denote the linear space of homogeneous polynomials of degree k (in $\mathbb{R}^n$), and $P_k = \cup_{s=0}^k H_s$ the linear space of polynomials of degree at most k. Set $r = \binom{n-1+k}{k} = \dim H_k$. Note that $r \asymp k^{n-1}$. We first claim that $P_k \subseteq \mathcal{R}_r$. This follows from the proof of Theorem 4.1 where it is proven that if π_k is the linear space of univariate polynomials of degree at most k, then for any set of $\mathbf{a}^1, \ldots, \mathbf{a}^r$ satisfying $\dim H_k|_A = r$, where $A = \{\mathbf{a}^1, \ldots, \mathbf{a}^r\}$, we have

$$P_k = \left\{ \sum_{i=1}^r g_i(\mathbf{a}^i \cdot \mathbf{x}) : \; g_i \in \pi_k, \; i = 1, \ldots, r \right\}.$$

Thus $P_k \subseteq \mathcal{R}_r$, and therefore

$$E(\mathcal{B}_p^m; \mathcal{R}_r; L_p) \leq E(\mathcal{B}_p^m; P_k; L_p).$$

It is a classical result that

$$E(\mathcal{B}_p^m; P_k; L_p) \leq Ck^{-m}.$$

Since $r \asymp k^{n-1}$ it therefore follows that

$$E(\mathcal{B}_p^m; P_k; L_p) \leq Cr^{-m/(n-1)}$$

for some appropriate C. □

Remark. Not only is it true that $E(\mathcal{B}_p^m; P_k; L_p) \leq Ck^{-m}$, but there also exist, for each p, m and k, linear operators $L : \mathcal{W}_p^m \to P_k$ for which

$$\sup_{f \in \mathcal{B}_p^m} \|f - L(f)\|_p \leq Ck^{-m}.$$

This metatheorem has been around for years. For a proof, see Mhaskar (1996).

Theorem 6.2 is not a very strong result. It simply says that we can, using ridge functions, approximate at least as well as we can approximate with any polynomial space contained therein. Unfortunately the lower bound (6.2), currently only proven for the case $p = 2$, says that we can do no better, at least for the given Sobolev spaces. This lower bound is also, as was stated, a lower bound for the approximation error from $\mathcal{M}_r(\sigma)$ (for every $\sigma \in C(\mathbb{R})$). But how relevant is it? Given $p \in [1, \infty]$ and m, is it true that for all $\sigma \in C(\mathbb{R})$ we have

$$E(\mathcal{B}_p^m; \mathcal{M}_r(\sigma); L_p) \leq Cr^{-m/(n-1)}$$

for some C? No, not for all $\sigma \in C(\mathbb{R})$ (see, for example, Theorem 6.7). Does there exist $\sigma \in C(\mathbb{R})$ for which

$$E(\mathcal{B}_p^m; \mathcal{M}_r(\sigma); L_p) \leq Cr^{-m/(n-1)}$$

for some C? The answer is yes. There exist activation functions for which this lower bound is attained. This in itself is hardly surprising. It is a simple consequence of the separability of $C[-1, 1]$. (As such the σ exhibited are rather pathological.) What is perhaps somewhat more surprising, at first glance, is the fact that there exist activation functions for which this lower bound is attained which are sigmoidal, strictly increasing and belong to $C^\infty(\mathbb{R})$.

Proposition 6.3. (Maiorov and Pinkus 1999) There exist $\sigma \in C^\infty(\mathbb{R})$ that are sigmoidal and strictly increasing, and have the property that for every $g \in \mathcal{R}_r$ and $\varepsilon > 0$ there exist $c_i, \theta_i \in \mathbb{R}$ and $\mathbf{w}^i \in \mathbb{R}^n$, $i = 1, \ldots, r+n+1$, satisfying

$$\left| g(\mathbf{x}) - \sum_{i=1}^{r+n+1} c_i \sigma(\mathbf{w}^i \cdot \mathbf{x} - \theta_i) \right| < \varepsilon$$

for all $\mathbf{x} \in B^n$.

This result and Theorem 6.2 immediately imply the following result.

Corollary 6.4 There exist $\sigma \in C^\infty(\mathbb{R})$ which are sigmoidal and strictly increasing, and for which

$$E(\mathcal{B}_p^m; \mathcal{M}_r(\sigma); L_p) \leq Cr^{-m/(n-1)} \tag{6.4}$$

for each $p \in [1, \infty]$, $m \geq 1$ and $n \geq 2$.

Proof of Proposition 6.3. The space $C[-1,1]$ is separable. That is, it contains a countable dense subset. Let $\{u_m\}_{m=1}^{\infty}$ be such a subset. Thus, to each $h \in C[-1,1]$ and each $\varepsilon > 0$ there exists a k (dependent upon h and ε) for which

$$|h(t) - u_k(t)| < \varepsilon$$

for all $t \in [-1,1]$. Assume each u_m is in $C^{\infty}[-1,1]$. (We can, for example, choose the $\{u_m\}_{m=1}^{\infty}$ from among the set of all polynomials with rational coefficients.)

We will now construct a strictly increasing C^{∞} sigmoidal function σ, that is, $\lim_{t\to-\infty}\sigma(t) = 0$ and $\lim_{t\to\infty}\sigma(t) = 1$, such that, for each $h \in C[-1,1]$ and $\varepsilon > 0$, there exists an integer m and real coefficients a_1^m, a_2^m, and a_3^m (all dependent upon h and ε) such that

$$|h(t) - (a_1^m\sigma(t-3) + a_2^m\sigma(t+1) + a_3^m\sigma(t+4m+1))| < \varepsilon$$

for all $t \in [-1,1]$. We do this by constructing σ so that $a_1^k\sigma(t-3) + a_2^k\sigma(t+1) + a_3^k\sigma(t+4k+1) = u_k(t)$, for each k, and $t \in [-1,1]$.

Let f be any C^{∞}, strictly monotone, sigmoidal function. There are many, for instance $f(t) = 1/(1+e^{-t})$. We define σ on $[4m, 4m+2]$, $m = 1, 2, \ldots$, in the following way. Set $\sigma(t+4m+1) = b_m + c_m t + d_m u_m(t)$ for $t \in [-1,1]$ where we choose the constants b_m, c_m and d_m so that

1. $\sigma(4m) = f(4m)$
2. $0 < \sigma'(t) \le f'(t)$ on $[4m, 4m+2]$.

This is easily done. We make one further assumption. On the intervals $[-4,-2]$ and $[0,2]$ we demand that σ again satisfy conditions 1 and 2, as above, and be linear, and that $\sigma(t-3)$ and $\sigma(t+1)$ be linearly independent on $[-1,1]$. To finish the construction, simply fill in the gaps in the domain of definition of σ (including all of $(-\infty, 4)$) in such a way that $\lim_{t\to-\infty}\sigma(t) = 0$. From the construction there exists, for each $k \ge 1$, reals a_1^k, a_2^k, a_3^k, for which

$$a_1^k\sigma(t-3) + a_2^k\sigma(t+1) + a_3^k\sigma(t+4k+1) = u_k(t).$$

Let $g \in \mathcal{R}_r$ and $\varepsilon > 0$ be given. We may write

$$g(\mathbf{x}) = \sum_{j=1}^{r} g_j(\mathbf{a}^j \cdot \mathbf{x})$$

for some $g_j \in C[-1,1]$ and $\mathbf{a}^j \in S^{n-1}$, $j = 1, \ldots, r$. From the above construction of σ there exist constants b_1^j, b_2^j, b_3^j and an integers k_j such that

$$|g_j(t) - (b_1^j\sigma(t-3) + b_2^j\sigma(t+1) + b_3^j\sigma(t+k_j))| < \varepsilon/r$$

for all $t \in [-1,1]$ and $j = 1, \ldots, r$.

Thus

$$|g_j(\mathbf{a}^j \cdot \mathbf{x}) - (b_1^j \sigma(\mathbf{a}^j \cdot \mathbf{x} - 3) + b_2^j \sigma(\mathbf{a}^j \cdot \mathbf{x} + 1) + b_3^j \sigma(\mathbf{a}^j \cdot \mathbf{x} + k_j))| < \varepsilon/r$$

for all $\mathbf{x} \in B^n$, and hence

$$\left| g(\mathbf{x}) - \sum_{j=1}^{r} \left(b_1^j \sigma(\mathbf{a}^j \cdot \mathbf{x} - 3) + b_2^j \sigma(\mathbf{a}^j \cdot \mathbf{x} + 1) + b_3^j \sigma(\mathbf{a}^j \cdot \mathbf{x} + k_j) \right) \right| < \varepsilon$$

for all $\mathbf{x} \in B^n$. Now each $\sigma(\mathbf{a}^j \cdot \mathbf{x} - 3)$, $\sigma(\mathbf{a}^j \cdot \mathbf{x} + 1)$, $j = 1, \ldots, r$, is a linear function, that is, a linear combination of $1, x_1, \ldots, x_n$. As such, the

$$\sum_{j=1}^{r} b_1^j \sigma(\mathbf{a}^j \cdot \mathbf{x} - 3) + b_2^j \sigma(\mathbf{a}^j \cdot \mathbf{x} + 1)$$

may be rewritten using at most $n + 1$ terms from the sum. This proves the proposition. □

Remark. The implications of Proposition 6.3 (and its proof) and Corollary 6.4 seem to be twofold. Firstly, sigmoidality, monotonicity and smoothness (C^∞) are not impediments to optimal degrees of approximation. Secondly, and perhaps more surprisingly, these excellent properties are not sufficient to deter the construction of 'pathological' activation functions. In fact there exist real (entire) analytic, sigmoidal, strictly increasing σ for which these same optimal error estimates hold (except that $3r$ replaces $r + n + 1$ in Proposition 6.3). For further details, see Maiorov and Pinkus (1999). In practice any approximation process depends not only on the degree (order) of approximation, but also on the possibility, complexity and cost of finding good approximants. The above activation functions are very smooth and give the best degree of approximation. However, they are unacceptably complex.

We now know something about what is possible, at least theoretically. However, there is another interesting lower bound which is larger than that given above. How can that be? It has to do with the 'method of approximation'. We will show that if the choice of coefficients, weights and thresholds depend continuously on the function being approximated (a not totally unreasonable assumption), then a lower bound on the error of approximation to functions in $\mathcal{B}_p^m$ from $\mathcal{M}_r(\sigma)$ is of the order of $r^{-m/n}$ (rather than the $r^{-m/(n-1)}$ proven above). We will also show that for all $\sigma \in C^\infty(\mathbb{R})$ (σ not a polynomial), and for many other σ, this bound is attained.

DeVore, Howard and Micchelli (1989) have introduced what they call a *continuous nonlinear d-width.* It is defined as follows.

Let K be a compact set in a normed linear space X. Let P_d be any continuous map from K to $\mathbb{R}^d$, and M_d any map whatsoever from $\mathbb{R}^d$ to X.

Thus $M_d(P_d(\cdot))$ is a map from K to X that has a particular (and perhaps peculiar) factorization. For each such P_d and M_d set

$$E(K; P_d, M_d; X) = \sup_{f \in K} \|f - M_d(P_d(f))\|_X,$$

and now define the *continuous nonlinear d-width*

$$h_d(K; X) = \inf_{P_d, M_d} E(K; P_d, M_d; X)$$

of K in X, where the infimum is taken over all P_d and M_d as above.

DeVore, Howard and Micchelli prove, among other facts, the asymptotic estimate

$$h_d(\mathcal{B}_p^m; L_p) \asymp d^{-m/n}.$$

In our context we are interested in the lower bound. As such, we provide a proof of the following.

Theorem 6.5. (DeVore, Howard and Micchelli 1989) For each fixed $p \in [1, \infty]$, $m \geq 1$ and $n \geq 1$

$$h_d(\mathcal{B}_p^m; L_p) \geq C d^{-m/n}$$

for some constant C independent of d.

Proof. The Bernstein d-width, $b_d(K; X)$, of a compact, convex, centrally symmetric set K in X is the term which has been applied to a codification of one of the standard methods of providing lower bounds for many of the more common d-width concepts. This lower bound is also valid in this setting, as we now show. For K and X, as above, set

$$b_d(K; X) = \sup_{X_{d+1}} \sup\{\lambda : \lambda S(X_{d+1}) \subseteq K\},$$

where X_{d+1} is any $(d+1)$-dimensional subspace of X, and $S(X_{d+1})$ is the unit ball of X_{d+1}.

Let P_d be any continuous map from K into R^d. Set

$$\widetilde{P}_d(f) = P_d(f) - P_d(-f).$$

Thus $\widetilde{P}_d$ is an odd, continuous map from K into R^d, *i.e.*, $\widetilde{P}_d(-f) = -\widetilde{P}_d(f)$. Assume $\lambda S(X_{d+1}) \subseteq K$. $\widetilde{P}_d$ is an odd, continuous map of $\partial(\lambda S(X_{d+1}))$ (the boundary of $S(X_{d+1})$) into $\mathbb{R}^d$. By the Borsuk Antipodality Theorem there exists an $f^* \in \partial(\lambda S(X_{d+1}))$ for which $\widetilde{P}_d(f^*) = 0$, *i.e.*, $P_d(f^*) = P_d(-f^*)$. As a consequence, for any map M_d from R^d to X,

$$2f^* = [f^* - M_d(P_d(f^*))] - [-f^* - M_d(P_d(-f^*))]$$

and therefore

$$\max\{\|f^* - M_d(P_d(f^*))\|_X, \| - f^* - M_d(P_d(-f^*))\|_X\} \geq \|f^*\|_X = \lambda.$$

Since $f^* \in K$, this implies that

$$E(K; P_d, M_d; X) \geq \lambda.$$

This inequality is valid for every choice of eligible P_d and M_d, and $\lambda \leq b_d(K; X)$. Thus $h_d(K; X) \geq b_d(K; X)$, and in particular $h_d(\mathcal{B}_p^m; L_p) \geq b_d(\mathcal{B}_p^m; L_p)$.

It remains to prove the bound $b_d(\mathcal{B}_p^m; L_p) \geq Cd^{-m/n}$. This proof is quite standard. Let ϕ be any nonzero function in $C^\infty(\mathbb{R}^n)$ with support in $[0,1]^n$. For $\ell > 0$ and any $\mathbf{j} \in \mathbb{Z}^n$, set

$$\phi_{\mathbf{j},\ell}(x_1, \ldots, x_n) = \phi(x_1\ell - j_1, \ldots, x_n\ell - j_n).$$

Thus the support of $\phi_{\mathbf{j},\ell}$ lies in $\prod_{i=1}^n [j_i/\ell, (j_i+1)/\ell]$. For ℓ large, the number of $\mathbf{j} \in \mathbb{Z}^n$ for which the support of $\phi_{\mathbf{j},\ell}$ lies totally in B^n is of the order of ℓ^n.

A simple change of variable argument shows that, for every $p \in [1, \infty]$ and $\mathbf{k} \in \mathbb{Z}_+^n$,

$$\|\phi_{\mathbf{j},\ell}\|_p = \ell^{-n/p}\|\phi\|_p,$$

and

$$\|D^{\mathbf{k}}\phi_{\mathbf{j},\ell}\|_p = \ell^{|\mathbf{k}|-n/p}\|D^{\mathbf{k}}\phi\|_p.$$

Furthermore, since the $\phi_{\mathbf{j},\ell}$ have distinct support (for fixed ℓ),

$$\left\| \sum_{\mathbf{j}} c_{\mathbf{j}}\phi_{\mathbf{j},\ell} \right\|_p = \ell^{-n/p}\|c\|_p\|\phi\|_p$$

and

$$\left\| D^{\mathbf{k}} \left(\sum_{\mathbf{j}} c_{\mathbf{j}}\phi_{\mathbf{j},\ell} \right) \right\|_p = \ell^{|\mathbf{k}|-n/p}\|c\|_p\|D^{\mathbf{k}}\phi\|_p$$

where $\|c\|_p$ is the ℓ_p-norm of the $\{c_{\mathbf{j}}\}$. Thus

$$\left\| \sum_{\mathbf{j}} c_{\mathbf{j}}\phi_{\mathbf{j},\ell} \right\|_{m,p} \asymp \ell^m \left\| \sum_{\mathbf{j}} c_{\mathbf{j}}\phi_{\mathbf{j},\ell} \right\|_p,$$

where we have restricted the $\mathbf{j}$ in the above summands to those $\mathbf{j}$ for which the support of $\phi_{\mathbf{j},\ell}$ lies totally in B^n.

We have therefore obtained a linear subspace of dimension of order ℓ^n with the property that, if

$$\left\| \sum_{\mathbf{j}} c_{\mathbf{j}}\phi_{\mathbf{j},\ell} \right\|_p \leq 1,$$

then

$$C\ell^{-m} \left\| \sum_{\mathbf{j}} c_{\mathbf{j}}\phi_{\mathbf{j},\ell} \right\|_{m,p} \leq 1$$

for some constant C independent of ℓ. This exactly implies that

$$b_d(\mathcal{B}_p^m; L_p) \geq C\ell^{-m}$$

where $d \asymp \ell^n$. Thus

$$h_d(\mathcal{B}_p^m; L_p) \geq b_d(\mathcal{B}_p^m; L_p) \geq Cd^{-m/n},$$

which proves the theorem. □

This theorem is useful in what it tells us about approximating from $\mathcal{M}_r(\sigma)$ by certain continuous methods. However, two things should be noted and understood. Firstly, these permissible 'methods of approximation' do not necessarily include all continuous methods of approximation. Secondly, some of the approximation methods being developed and used today in this setting are iterative and are not necessarily continuous.

Any element $g \in \mathcal{M}_r(\sigma)$ has the form

$$g(\mathbf{x}) = \sum_{i=1}^{r} c_i \sigma(\mathbf{w}^i \cdot \mathbf{x} - \theta_i)$$

for some constants $c_i, \theta_i \in \mathbb{R}$ and $\mathbf{w}^i \in \mathbb{R}^n$, $i = 1, \ldots, r$. In general, when approximating $f \in L_p$, our choice of g will depend upon these $(n+2)r$ parameters. (Some of these parameters may be fixed independent of the function being approximated.) For any method of approximation which continuously depends on these parameters, the lower bound of Theorem 6.5 holds.

Theorem 6.6 Let $Q_r : L_p \to \mathcal{M}_r(\sigma)$ be any method of approximation where the parameters c_i, θ_i and $\mathbf{w}^i$, $i = 1, \ldots, r$, are continuously dependent on the function being approximated (some may of course be fixed independent of the function). Then

$$\sup_{f \in \mathcal{B}_p^m} \|f - Q_r f\|_p \geq Cr^{-m/n}$$

for some C independent of r.

Additional upper and lower bound estimates appear in Maiorov and Meir (1999). Particular cases of their lower bounds for specific σ improve upon the lower bound for $E(\mathcal{B}_2^m; \mathcal{M}_r(\sigma); L_2)$ given in Theorem 6.1, without any assumption about the continuity of the approximating procedure. We only state this next result. Its proof is too complicated to be presented here.

Theorem 6.7. (Maiorov and Meir 1999) Let $p \in [1, \infty]$, $m \geq 1$ and $n \geq 2$. Let σ be the logistic sigmoid, that is,

$$\sigma(t) = \frac{1}{1 + e^{-t}},$$

or a (polynomial) spline of a fixed degree with a finite number of knots. Then

$$E(\mathcal{B}_p^m; \mathcal{M}_r(\sigma); L_p) \geq C(r \log r)^{-m/n}$$

for some C independent of r.

We now consider upper bounds. The next theorem may, with minor modifications, be found in Mhaskar (1996) (see also Ellacott and Bos (1996, p. 352)). Note that the logistic sigmoid satisfies the conditions of Theorem 6.8.

Theorem 6.8 Assume $\sigma : \mathbb{R} \to \mathbb{R}$ is such that $\sigma \in C^\infty(\Theta)$ on some open interval Θ, and σ is not a polynomial on Θ. Then, for each $p \in [1, \infty]$, $m \geq 1$ and $n \geq 2$,

$$E(\mathcal{B}_p^m; \mathcal{M}_r(\sigma); L_p) \leq Cr^{-m/n} \tag{6.5}$$

for some constant C independent of r.

Proof. The conditions of Theorem 6.8 imply, by Corollary 3.6, that $\overline{\mathcal{N}_{k+1}(\sigma)}$, the closure of $\mathcal{N}_{k+1}(\sigma)$, contains π_k, the linear space of univariate algebraic polynomials of degree at most k.

From the proof of Theorem 4.1 (see also Theorem 6.2), for $s = \dim H_k \asymp k^{n-1}$ there exist $\mathbf{a}^1, \ldots, \mathbf{a}^s$ in S^{n-1} such that

$$P_k = \left\{ \sum_{i=1}^{s} g_i(\mathbf{a}^i \cdot \mathbf{x}) : g_i \in \pi_k, \, i = 1, \ldots, s \right\},$$

where P_k is the linear space of n-variate algebraic polynomials of degree at most k.

Since each $g_i \in \overline{\mathcal{N}_{k+1}(\sigma)}$, and $\mathcal{M}_p(\sigma) + \mathcal{M}_q(\sigma) = \mathcal{M}_{p+q}(\sigma)$, it follows that

$$P_k \subseteq \overline{\mathcal{M}_{s(k+1)}(\sigma)}.$$

Set $r = s(k+1)$. Then

$$E(\mathcal{B}_p^m; \mathcal{M}_r(\sigma); L_p) = E(\mathcal{B}_p^m; \overline{\mathcal{M}_r(\sigma)}; L_p) \leq E(\mathcal{B}_p^m; P_k; L_p) \leq Ck^{-m}$$

for some constant C independent of r. Since $r \asymp k^n$, we have

$$E(\mathcal{B}_p^m; \mathcal{M}_r(\sigma); L_p) \leq Cr^{-m/n},$$

which proves the theorem. □

Remark. It is important to note that the upper bound of Theorem 6.8 can be attained by continuous (and in fact linear) methods in the sense of Theorem 6.6. The thresholds θ_i can all be chosen to equal θ_o where $\sigma^{(k)}(-\theta_o) \neq 0$, $k = 0, 1, 2, \ldots$ (see Proposition 3.4). The weights are also chosen independent of the function being approximated. The dependence on the function is only in the choice of the g_i and, as previously noted (see the

remark after Theorem 6.2), this can in fact be done in a linear manner. (For each $p \in (1, \infty)$, the operator of best approximation from P_k is continuous.)

Remark. For functions analytic in a neighbourhood of B^n, there are better order of approximation estimates, again based on polynomial approximation: see Mhaskar (1996).

If the optimal order of approximation from $\mathcal{M}_r(\sigma)$ is really no better than that obtained by approximating from the polynomial space P_k of dimension $r \asymp k^n$, then one cannot but wonder if it is really worthwhile using this model (at least in the case of a single hidden layer). It is not yet clear, from this perspective, what the mathematical or computational justifications are for choosing this model over other models. Some researchers, however, would be more than content if they could construct neural networks that algorithmically achieve this order of approximation.

Petrushev (1998) proves some general estimates concerning ridge and neural network approximation. These results are valid only for $p = 2$. However, they generalize Theorem 6.8 within that setting.

Let $L_2^1 = L_2[-1, 1]$ with the usual norm

$$\|g\|_{L_2^1} = \left(\int_{-1}^{1} |g(t)|^2 \, \mathrm{d}t \right)^{1/2}.$$

Similarly $\mathcal{H}_{m,2}$ will denote the Sobolev space on $[-1, 1]$ with norm

$$\|g\|_{\mathcal{H}_{m,2}} = \left(\sum_{j=0}^{m} \|g^{(j)}\|_{L_2^1}^2 \right)^{1/2}.$$

Set

$$E(\mathcal{H}_{m,2}; \mathcal{N}_k(\sigma); L_2^1) = \sup_{\|h\|_{\mathcal{H}_{m,2}} \leq 1} \inf_{g \in \mathcal{N}_k(\sigma)} \|h - g\|_{L_2^1}.$$

The point of the above is that this is all taking place in $\mathbb{R}^1$ rather than in $\mathbb{R}^n$.

Theorem 6.9. (Petrushev 1998) Let $m \geq 1$ and $n \geq 2$. Assume σ has the property that

$$E(\mathcal{H}_{m,2}; \mathcal{N}_k(\sigma); L_2^1) \leq Ck^{-m}, \tag{6.6}$$

for some C independent of k. Then

$$E(\mathcal{B}_2^{m+(n-1)/2}; \mathcal{M}_r(\sigma) : L_2) \leq Cr^{-(m+\frac{(n-1)}{2})/n}, \tag{6.7}$$

for some other C independent of r.

Remark. It follows from general 'interpolation' properties of spaces that, if (6.6) or (6.7) hold for a specific m, then they also hold for every positive value less than m.

The proof of Theorem 6.9 is too complicated to be presented here. The underlying idea is similar to that used in the proof of Theorem 6.8. One uses multivariate polynomials to approximate functions in $\mathcal{B}_2^{m+(n-1)/2}$, decomposes these multivariate polynomials into 'ridge' polynomials (the g_i in the proof of Theorem 6.8), and then approximates these univariate 'ridge' polynomials from $\mathcal{M}_k(\sigma)$.

One consequence of Theorem 6.9 which we wish to highlight (as it is not directly covered by Theorem 6.8) is the following.

Corollary 6.10. (Petrushev 1998) For each $k \in \mathbb{Z}_+$, let

$$\sigma_k(t) = t_+^k = \begin{cases} t^k, & t \geq 0, \\ 0, & t < 0. \end{cases}$$

Then

$$E(\mathcal{B}_2^m; \mathcal{M}_r(\sigma_k); L_2) \leq Cr^{-m/n}$$

for $m = 1, \ldots, k + 1 + \frac{(n-1)}{2}$, and some constant C independent of r.

A variation on a result of Petrushev (1998) proves this corollary for $m = k + 1 + (n-1)/2$. The other cases follow by taking differences (really just differentiating), or as a consequence of the above remark. Note that $\sigma_o(t)$ is the Heaviside function.

For given $k \in \mathbb{Z}_+$ assume σ is continuous, or piecewise continuous, and satisfies

$$\lim_{t \to -\infty} \frac{\sigma(t)}{t^k} = 0, \qquad \lim_{t \to \infty} \frac{\sigma(t)}{t^k} = 1.$$

(This is essentially what Mhaskar and Micchelli (1992) call kth degree sigmoidal.) Then $\lim_{\lambda \to \infty} \sigma(\lambda t)/\lambda^k = \sigma_k(t)$ uniformly off $[-c, c]$, any $c > 0$, and converges in $L_p[-1, 1]$ for any $p \in [1, \infty)$. Let σ_k be as defined in Corollary 6.10. Thus $\mathcal{M}_r(\sigma_k) \subseteq \overline{\mathcal{M}_r(\sigma)}$. In addition, if σ is a spline of degree k with at least one simple knot, then by taking (a finite number of) shifts and dilates we can again approximate σ_k in the $L_p[-1, 1]$ norm, $p \in [1, \infty)$. Thus, applying Corollary 6.10 we obtain the following.

Corollary 6.11 For given $k \in \mathbb{Z}_+$, let σ be as defined in the previous paragraph. Then

$$E(\mathcal{B}_2^m; \mathcal{M}_r(\sigma); L_2) \leq Cr^{-m/n}$$

for $m = 1, \ldots, k + 1 + \frac{(n-1)}{2}$, and some constant C independent of r.

Note that the error of approximation in all these results has exactly the same form as that given by (6.5). If $\sigma \in C^\infty(\Theta)$ as in Theorem 6.8, then (6.6) holds since $\overline{\mathcal{N}_k(\sigma)}$ contains π_{k-1}.

A different and very interesting approach to the problem of determining (or at least bounding) the order of approximation from the set $\mathcal{M}_r(\sigma)$ was initiated by Barron (1993). Until now we have considered certain standard smoothness classes (the W_p^m), and then tried to estimate the worst case error of approximation from functions in this class. Another approach is, given $\mathcal{M}_r(\sigma)$, to try to find classes of functions which are well approximated by $\mathcal{M}_r(\sigma)$. This is generally a more difficult problem, but one well worth pursuing. Barron does this, in a sense, in a specific but interesting setting.

What we present here is based on work of Barron (1993), and generalizations due to Makovoz (1996). We start with a general result which is a generalization, due to Makovoz (1996), of a result of Barron (1993) and Maurey (Pisier 1981) (see also Jones (1992)). (Their result does not contain the factor $\varepsilon_r(K)$.) It should be mentioned that, unlike the previously discussed upper bounds, these upper bounds are obtained by strictly nonlinear (and not necessarily continuous) methods.

Let H be a Hilbert space and K a bounded set therein. Let $\operatorname{co} K$ denote the convex hull of K. Set

$$\varepsilon_r(K) = \inf\{\varepsilon > 0 : K \text{ can be covered by } r \text{ sets of diameter } \leq \varepsilon\}.$$

Theorem 6.12. (Makovoz 1996) Let K be a bounded subset of a Hilbert space H. Let $f \in \operatorname{co} K$. Then there is an f_r of the form

$$f_r = \sum_{i=1}^{r} a_i g_i$$

for some $g_i \in K$, $a_i \geq 0$, $i = 1, \ldots, r$, and $\sum_{i=1}^r a_i \leq 1$, satisfying

$$\|f - f_r\|_H \leq \frac{2\varepsilon_r(K)}{\sqrt{r}}.$$

Letting K be the set of our approximants we may have here a very reasonable approximation-theoretic result. The problem, however, is to identify $\operatorname{co} K$, or at least some significant subset of $\operatorname{co} K$ in other than tautological terms. Otherwise the result could be rather sterile.

Barron (1993) considered σ which are bounded, measurable, and sigmoidal, and set

$$K(\sigma) = \{\pm\sigma(\mathbf{w} \cdot \mathbf{x} - \theta) : \mathbf{w} \in \mathbb{R}^n, \theta \in \mathbb{R}\}.$$

(Recall that $\mathbf{x} \in B^n$.) He then proved that $\overline{\operatorname{co} K(\sigma)}$ contains the set $\mathcal{B}$ of all functions f defined on B^n which can be extended to all of $\mathbb{R}^n$ such that

some shift of f by a constant has a Fourier transform $\widehat{f}$ satisfying

$$\int_{\mathbb{R}^n} \|\mathbf{s}\|_2 |\widehat{f}(\mathbf{s})| \, d\mathbf{s} \leq \gamma,$$

for some $\gamma > 0$.

Let us quickly explain, in general terms, why this result holds. As we mentioned earlier, at least for continuous, sigmoidal σ (see the comment after Corollary 6.10), $\sigma(\lambda \cdot)$ approaches $\sigma_o(\cdot)$ in norm as $\lambda \to \infty$, where σ_o is the Heaviside function. As such, $\overline{\text{co}}\, K(\sigma_o) \subseteq \overline{\text{co}}\, K(\sigma)$ (and, equally important in what will follow, we essentially have $K(\sigma_o) \subseteq K(\sigma)$, *i.e.*, we can replace each $\sigma_o(\mathbf{w} \cdot \mathbf{x} - \theta)$ by only the one term $\sigma(\lambda(\mathbf{w} \cdot \mathbf{x} - \theta))$ for some sufficiently large λ). So it suffices to prove that the above set of functions $\mathcal{B}$ is in fact contained in $\overline{\text{co}}\, K(\sigma_o)$.

Set

$$L_o = \{\pm \sigma_o(t - \theta) : \theta \in \mathbb{R}\},$$

for $t \in [-1, 1]$. (L_o is simply $K(\sigma_o)$ in $\mathbb{R}^1$.) Up to a constant (the 'shift' previously mentioned) h is contained in $\overline{\text{co}}\, L_o$ if and only if h is a function of bounded variation with total variation bounded by 1. If h is continuously differentiable, this just means that

$$\int_{-1}^{1} |h'(t)| \, dt \leq 1.$$

Applying this result to $K(\sigma_o)$, this implies that, for each $\mathbf{s} \in \mathbb{R}^n$, $\mathbf{s} \neq \mathbf{0}$,

$$\frac{\gamma e^{i\mathbf{s} \cdot \mathbf{x}}}{\|\mathbf{s}\|_2} \in \overline{\text{co}}\, K(\sigma_o)$$

for some γ (dependent on B^n). Thus, if

$$\int_{\mathbb{R}^n} \|\mathbf{s}\|_2 |\widehat{f}(\mathbf{s})| \, d\mathbf{s} \leq \gamma,$$

then

$$f(\mathbf{x}) = \int_{\mathbb{R}^n} \left(\frac{\gamma e^{i\mathbf{s} \cdot \mathbf{x}}}{\|\mathbf{s}\|_2} \right) \left(\frac{\|\mathbf{s}\|_2 \widehat{f}(\mathbf{s})}{\gamma} \right) d\mathbf{s} \in \overline{\text{co}}\, K(\sigma_o).$$

To apply Theorem 6.12 we should also obtain a good estimate for $\varepsilon_r(K(\sigma))$. This quantity is generally impossible to estimate. However, since $K(\sigma_o) \subseteq K(\sigma)$ we have $\mathcal{M}_r(\sigma_o) \subseteq \overline{\mathcal{M}_r(\sigma)}$, and it thus suffices to consider $\varepsilon_r(K(\sigma_o))$. Since we are approximating on B^n,

$$K(\sigma_o) = \{\pm \sigma_o(\mathbf{w} \cdot \mathbf{x} - \theta) : \|\mathbf{w}\|_2 = 1, \, |\theta| \leq 1\}.$$

(For any other $\mathbf{w}$ or θ we add no additional function to the set $K(\sigma_o)$.)

Now, if $\|\mathbf{w}^1\|_2 = \|\mathbf{w}^2\|_2 = 1$, $\|\mathbf{w}^1 - \mathbf{w}^2\|_2 \leq \varepsilon^2$, and $|\theta_1|, |\theta_2| \leq 1$,

$|\theta_1 - \theta_2| \leq \varepsilon^2$, then

$$\left(\int_{B^n} |\sigma_o(\mathbf{w}^1 \cdot \mathbf{x} - \theta_1) - \sigma_o(\mathbf{w}^2 \cdot \mathbf{x} - \theta_2)|^2 \, d\mathbf{x}\right)^{1/2} \leq C\varepsilon$$

for some constant C. Thus to estimate $\varepsilon_r(K(\sigma_o))$ we must find an ε^2-net for

$$\{(\mathbf{w}, \theta) : \|\mathbf{w}\|_2 = 1, |\theta| \leq 1\}.$$

It is easily shown that for this we need $(\varepsilon^2)^{-n}$ elements. Thus $\varepsilon_r(K(\sigma_o)) \leq Cr^{-1/2n}$.

We can now summarize.

Theorem 6.13. (Makovoz 1996) Let $\mathcal{B}$ be as defined above. Then, for any bounded, measurable, sigmoidal function σ,

$$E(\mathcal{B}; \mathcal{M}_r(\sigma); L_2) \leq E(\mathcal{B}; \mathcal{M}_r(\sigma) \cap \mathcal{B}; L_2) \leq Cr^{-(n+1)/2n} \tag{6.8}$$

for some constant C independent of r.

If σ is a piecewise continuous sigmoidal function, then from Corollary 6.11 we have

$$E(\mathcal{B}_2^{(n+1)/2}; \mathcal{M}_r(\sigma); L_2) \leq Cr^{-(n+1)/2n}.$$

This is the same error bound, with the same activation function, as appears in (6.8). As such it is natural to ask which, if either, is the stronger result. In fact the results are not comparable. The condition defining $\mathcal{B}$ cannot be restated in terms of conditions on the derivatives. What is known (see Barron (1993)) is that on B^n we essentially have

$$\mathcal{W}_\infty^{[n/2]+2} \subseteq \operatorname{span} \mathcal{B} \subseteq \mathcal{W}_\infty^1 \subseteq \mathcal{W}_2^1.$$

(The leftmost inclusion is almost, but not quite, correct: see Barron (1993).)

The error estimate of Barron (1993) did not originally contain the term $\varepsilon_r(K)$ and thus was of the form $Cr^{-1/2}$ (for some constant C). This initiated an unfortunate discussion concerning these results having 'defeated the curse of dimensionality'.

The literature contains various generalizations of the above results, and we expect more to follow. Makovoz (1996) generalizes Theorems 6.12 and 6.13 to $L_q(B, \mu)$, where μ is a probability measure on some set B in $\mathbb{R}^n$, $1 \leq q < \infty$. (For a discussion of an analogous problem in the uniform norm, see Barron (1992) and Makovoz (1998).) Donahue, Gurvits, Darken and Sontag (1997) consider different generalizations of Theorem 6.12 and they provide a general perspective on this type of problem. Hornik, Stinchcombe, White and Auer (1994) (see also Chen and White (1999)) consider generalizations of the Barron (1993) results to where the function and some of its derivatives are simultaneously approximated. Lower bounds on the error of

approximation are to be found in Barron (1992) and Makovoz (1996). However, these lower bounds essentially apply to approximating from $\mathcal{M}_r(\sigma_o) \cap \mathcal{B}$ (a restricted set of approximants and a particular activation function) and do not apply to approximation from all of $\mathcal{M}_r(\sigma)$. Other related results may be found in Mhaskar and Micchelli (1994), Yukich, Stinchcombe and White (1995) and Kurkova, Kainen and Kreinovich (1997).

For $f \in \mathcal{B}$ the following algorithm of approximation was introduced by Jones (1992) to obtain an iterative sequence $\{h_r\}$ of approximants ($h_r \in \mathcal{M}_r(\sigma)$) where σ is sigmoidal (as above). These approximants satisfy

$$\|f - h_r\|_2 \leq Cr^{-1/2},$$

for some constant C independent of f and r. The sequence is constructed as follows. We initialize the process by setting $h_0 = 0$, and then consider

$$\min_{0 \leq \alpha \leq 1} \min_{g \in \overline{K(\sigma)}} \|f - (\alpha h_{r-1} + (1-\alpha)g)\|_2.$$

Assume that these minima are attained for $\alpha_r \in [0,1]$ and $g_r \in \overline{K(\sigma)}$. Set

$$h_r = \alpha_r h_{r-1} + (1 - \alpha_r) g_r.$$

(In the above we assume that $\overline{K(\sigma)}$ is compact.) In fact, as mentioned by Jones (1992), improved upon by Barron (1993), and further improved by Jones (1999) (see also Donahue, Gurvits, Darken and Sontag (1997)), the α_r and g_r need not be chosen to attain the above minima exactly and yet the same convergence rate will hold.

We end this section by pointing out that much remains to be done in finding good upper bounds, constructing reasonable methods of approximation, and identifying classes of functions which are well approximated using this model. It is also worth noting that very few of the results we have surveyed used intrinsic properties of the activation functions. In Theorem 6.8 only the C^∞ property was used. Corollary 6.11 depends solely on the approximation properties of σ_k. Theorem 6.13 is a result concerning the Heaviside activation function.

7. Two hidden layers

Relatively little is known concerning the advantages and disadvantages of using a single hidden layer with many units (neurons) over many hidden layers with fewer units. The mathematics and approximation theory of the MLP model with more than one hidden layer is not well understood. Some authors see little theoretical gain in considering more than one hidden layer since a single hidden layer model suffices for density. Most authors, however, do allow for the possibility of certain other benefits to be gained from using more than one hidden layer. (See de Villiers and Barnard (1992) for a comparison of these two models.)

One important advantage of the multiple (rather than single) hidden layer model has to do with the existence of locally supported, or at least 'localized', functions in the two hidden layer model (see Lapedes and Farber (1988), Blum and Li (1991), Geva and Sitte (1992), Chui, Li and Mhaskar (1994)). For any activation function σ, every $g \in \mathcal{M}_r(\sigma)$, $g \neq 0$, has

$$\int_{\mathbb{R}^n} |g(\mathbf{x})|^p \, d\mathbf{x} = \infty$$

for every $p \in [1, \infty)$, and no $g \in \mathcal{M}_r(\sigma)$ has compact support. This is no longer true in the two hidden layer model. For example, let σ_o be the Heaviside function. Then

$$\sigma_o\left(\sum_{i=1}^{m} \sigma_o(\mathbf{w}^i \cdot \mathbf{x} - \theta_i) - \left(m - \frac{1}{2}\right)\right) = \begin{cases} 1, & \mathbf{w}^i \cdot \mathbf{x} \geq \theta_i, \ i = 1, \ldots, m, \\ 0, & \text{otherwise.} \end{cases} \tag{7.1}$$

Thus the two hidden layer model with activation function σ_o, and only one unit in the second hidden layer, can represent the characteristic function of any closed convex polygonal domain. For example, for $a_i < b_i$, $i = 1, \ldots, n$,

$$\sigma_o\left(\sum_{i=1}^{n} (\sigma_o(x_i - a_i) + \sigma_o(-x_i + b_i)) - \left(2n - \frac{1}{2}\right)\right)$$

is the characteristic function of the rectangle $\prod_{i=1}^{n} [a_i, b_i]$. (Up to boundary values, this function also has the representation

$$\sigma_o\left(\sum_{i=1}^{n} (\sigma_o(x_i - a_i) - \sigma_o(x_i - b_i)) - \left(n - \frac{1}{2}\right)\right)$$

since $\sigma_o(-t) = 1 - \sigma_o(t)$ for all $t \neq 0$.) If σ is a continuous or piecewise continuous sigmoidal function, then a similar result holds for such functions since $\sigma(\lambda \cdot)$ approaches $\sigma_o(\cdot)$ as $\lambda \to \infty$ in, say, $L^p[-1, 1]$ for every $p \in [1, \infty)$. The function

$$\sigma\left(\lambda\left(\sum_{i=1}^{m} \sigma(\lambda(\mathbf{w}^i \cdot \mathbf{x} - \theta_i)) - \left(m - \frac{1}{2}\right)\right)\right)$$

thus approximates the function given in (7.1) as $\lambda \to \infty$. Approximating by such localized functions has many, many advantages.

Another advantage of the multiple hidden layer model is the following. As was noted in Section 6, there is a lower bound on the degree to which the single hidden layer model with r units in the hidden layer can approximate any function. It is given by the extent to which a linear combination of r ridge functions can approximate this same function. This lower bound was shown to be attainable (Proposition 6.3 and Corollary 6.4), and, more importantly, ridge function approximation itself is bounded below (away

from zero) with some non-trifling dependence on r and on the set to be approximated.

In the single hidden layer model there is an intrinsic lower bound on the degree of approximation, depending on the number of units used. This is not the case in the two hidden layer model. We will prove, using the same activation function as in Proposition 6.3, that there is no theoretical lower bound on the error of approximation if we permit two hidden layers.

To be precise, we will prove the following theorem.

Theorem 7.1. (Maiorov and Pinkus 1999) There exists an activation function σ which is C^∞, strictly increasing, and sigmoidal, and has the following property. For any $f \in C[0,1]^n$ and $\varepsilon > 0$, there exist constants d_i, c_{ij}, θ_{ij}, γ_i, and vectors $\mathbf{w}^{ij} \in \mathbb{R}^n$ for which

$$\left| f(\mathbf{x}) - \sum_{i=1}^{4n+3} d_i \sigma \left(\sum_{j=1}^{2n+1} c_{ij} \sigma(\mathbf{w}^{ij} \cdot \mathbf{x} + \theta_{ij}) + \gamma_i \right) \right| < \varepsilon,$$

for all $\mathbf{x} \in [0,1]^n$.

In other words, for this specific activation function, any continuous function on the unit cube in $\mathbb{R}^n$ can be uniformly approximated to within any error by a two hidden layer neural network with $2n+1$ units in the first hidden layer and $4n+3$ units in the second hidden layer. (We recall that the constructed activation function is nonetheless rather pathological.)

In the proof of Theorem 7.1 we use the Kolmogorov Superposition Theorem. This theorem has been much quoted and discussed in the neural network literature: see Hecht-Nielsen (1987), Girosi and Poggio (1989), Kurkova (1991, 1992, 1995*b*), Lin and Unbehauen (1993). In fact Kurkova (1992) uses the Kolmogorov Superposition Theorem to construct approximations in the two hidden layer model with an arbitrary sigmoidal function. However, the number of units needed is exceedingly large, and does not provide for good error bounds or, in our opinion, a reasonably efficient method of approximation. Better error bounds follow by using localized functions (see, for instance, Blum and Li (1991), Itô (1994*a*), and especially Chui, Li and Mhaskar (1994)). Kurkova (1992) and others (see Frisch, Borzi, Ord, Percus and Williams (1989), Sprecher (1993, 1997), Katsuura and Sprecher (1994), Nees (1994, 1996)) are interested in using the Kolmogorov Superposition Theorem to find good algorithms for approximation. This is not our aim. We are using the Kolmogorov Superposition Theorem to prove that there is no theoretical lower bound on the degree of approximation common to all activation functions, as is the case in the single hidden layer model. In fact, we are showing that there exists an activation function with very 'nice' properties for which a fixed finite number of units in both hidden layers is

sufficient to approximate arbitrarily well any continuous function. We do not, however, advocate using this activation function.

The Kolmogorov Superposition Theorem answers (in the negative) Hilbert's 13th problem. It was proven by Kolmogorov in a series of papers in the late 1950s. We quote below an improved version of this theorem (see Lorentz, von Golitschek and Makovoz (1996, p. 553) for a more detailed discussion).

Theorem 7.2 There exist n constants $\lambda_j > 0$, $j = 1, \ldots, n$, $\sum_{j=1}^{n} \lambda_j \leq 1$, and $2n+1$ strictly increasing continuous functions ϕ_i, $i = 1, \ldots, 2n+1$, which map $[0,1]$ to itself, such that every continuous function f of n variables on $[0,1]^n$ can be represented in the form

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{2n+1} g\left(\sum_{j=1}^{n} \lambda_j \phi_i(x_j) \right) \tag{7.2}$$

for some $g \in C[0,1]$ depending on f.

Note that this is a theorem about representing (and not approximating) functions. There have been numerous generalizations of this theorem. Attempts to understand the nature of this theorem have led to interesting concepts related to the complexity of functions. Nonetheless the theorem itself has had few, if any, direct applications.

Proof of Theorem 7.1. We are given $f \in C[0,1]^n$ and $\varepsilon > 0$. Let g and the ϕ_i be as in (7.2). We will use the σ constructed in Proposition 6.3. Recall that to any $h \in C[-1,1]$ and $\eta > 0$ we can find constants a_1, a_2, a_3 and an integer m for which

$$|h(t) - (a_1\sigma(t-3) + a_2\sigma(t+1) + a_3\sigma(t+m))| < \eta$$

for all $t \in [-1,1]$. This result is certainly valid when we restrict ourselves to the interval $[0,1]$ and functions continuous thereon. As such, for the above g there exist constants a_1, a_2, a_3 and an integer m such that

$$|g(t) - (a_1\sigma(t-3) + a_2\sigma(t+1) + a_3\sigma(t+m))| < \frac{\varepsilon}{2(2n+1)} \tag{7.3}$$

for all $t \in [0,1]$. Further, recall that $\sigma(t-3)$ and $\sigma(t+1)$ are linear polynomials on $[0,1]$.

Substituting (7.3) in (7.2), we obtain

$$\left| f(x_1, \ldots, x_n) - \sum_{i=1}^{2n+1} \left[a_1\sigma\left(\sum_{j=1}^{n} \lambda_j \phi_i(x_j) - 3 \right) + a_2\sigma\left(\sum_{j=1}^{n} \lambda_j \phi_i(x_j) + 1 \right) \right. \right.$$
$$\left. \left. + a_3\sigma\left(\sum_{j=1}^{n} \lambda_j \phi_i(x_j) + m \right) \right] \right| < \frac{\varepsilon}{2} \tag{7.4}$$

for all $(x_1, \dots, x_n) \in [0,1]^n$. Since

$$\sigma\left(\sum_{j=1}^{n} \lambda_j \phi_i(x_j) - 3\right) \qquad \text{and} \qquad \sigma\left(\sum_{j=1}^{n} \lambda_j \phi_i(x_j) + 1\right)$$

are linear polynomials in $\sum_{j=1}^{n} \lambda_j \phi_i(x_j)$, for each i, we can in fact rewrite

$$\sum_{i=1}^{2n+1} a_1 \sigma\left(\sum_{j=1}^{n} \lambda_j \phi_i(x_j) - 3\right) + a_2 \sigma\left(\sum_{j=1}^{n} \lambda_j \phi_i(x_j) + 1\right)$$

as

$$\sum_{i=1}^{2n+2} d_i \sigma\left(\sum_{j=1}^{n} \lambda_j \phi_i(x_j) + \gamma_i\right)$$

where ϕ_{2n+2} is ϕ_k for some $k \in \{1, \dots, 2n+1\}$ (and γ_i is either -3 or 1 for each i).

Thus we may rewrite (7.4) as

$$\left| f(x_1, \dots, x_n) - \sum_{i=1}^{2n+2} d_i \sigma\left(\sum_{j=1}^{n} \lambda_j \phi_i(x_j) + \gamma_i\right) - \sum_{i=1}^{2n+1} a_3 \sigma\left(\sum_{j=1}^{n} \lambda_j \phi_i(x_j) + m\right) \right| < \frac{\varepsilon}{2} \tag{7.5}$$

for all $(x_1, \dots, x_n) \in [0,1]^n$.

For each $i \in \{1, \dots, 2n+1\}$, and $\delta > 0$ there exist constants b_{i1}, b_{2i}, b_{i3} and integers r_i such that

$$\left| \phi_i(x_j) - \Big(b_{i1} \sigma(x_j - 3) + b_{i2} \sigma(x_j + 1) + b_{i3} \sigma(x_j + r_i) \Big) \right| < \delta$$

for all $x_j \in [0,1]$. Thus, since $\lambda_j > 0$, $\sum_{j=1}^{n} \lambda_j \leq 1$,

$$\left| \sum_{j=1}^{n} \lambda_j \phi_i(x_j) - \sum_{j=1}^{n} \lambda_j \left(b_{i1} \sigma(x_j - 3) + b_{i2} \sigma(x_j + 1) + b_{i3} \sigma(x_j + r_i) \right) \right| < \delta$$

for all $(x_1, \dots, x_n) \in [0,1]^n$.

Again we use the fact that the $\sigma(x_j - 3)$ and $\sigma(x_j + 1)$ are linear polynomials on $[0,1]$ to rewrite the above as

$$\left| \sum_{j=1}^{n} \lambda_j \phi_i(x_j) - \sum_{j=1}^{2n+1} c_{ij} \sigma(\mathbf{w}^{ij} \cdot \mathbf{x} + \theta_{ij}) \right| < \delta \tag{7.6}$$

for all $(x_1, \dots, x_n) \in [0,1]^n$ for some constants c_{ij} and θ_{ij} and vectors $\mathbf{w}^{ij}$ (in fact the $\mathbf{w}^{ij}$ are all unit vectors).

We now substitute (7.6) into (7.5). As σ is uniformly continuous on every closed interval, we can choose $\delta > 0$ sufficiently small so that

$$\left| \sum_{i=1}^{2n+2} d_i \sigma\left(\sum_{j=1}^{n} \lambda_j \phi_i(x_j) + \gamma_i \right) + \sum_{i=1}^{2n+1} a_3 \sigma\left(\sum_{j=1}^{n} \lambda_j \phi_i(x_j) + m \right) - \sum_{i=1}^{2n+2} d_i \sigma\left(\sum_{j=1}^{2n+1} c_{ij} \sigma(\mathbf{w}^{ij} \cdot \mathbf{x} + \theta_{ij}) + \gamma_i \right) - \sum_{i=1}^{2n+1} a_3 \sigma\left(\sum_{j=1}^{2n+1} c_{ij} \sigma(\mathbf{w}^{ij} \cdot \mathbf{x} + \theta_{ij}) + m \right) \right| < \frac{\varepsilon}{2}. \tag{7.7}$$

From (7.5), (7.7), renumbering and renaming, the theorem follows. □

As a consequence of what was stated in the remark following the proof of Proposition 6.3, we can in fact prove Theorem 7.1 with a σ which is analytic (and not only C^∞), strictly increasing, and sigmoidal (see Maiorov and Pinkus (1999)). The difference is that we must then use $3n$ units in the first layer and $6n + 3$ units in the second layer. The restriction of Theorem 7.1 to the unit cube is for convenience only. The same result holds over any compact subset of $\mathbb{R}^n$.

We have established only two facts in this section. We have shown that there exist localized functions, and that there is no theoretical lower bound on the degree of approximation common to all activation functions (contrary to the situation in the single hidden layer model). Nonetheless there seems to be reason to conjecture that the two hidden layer model may be significantly more promising than the single hidden layer model, at least from a purely approximation-theoretic point of view. This problem certainly warrants further study.

Acknowledgement

The author is indebted to Lee Jones, Moshe Leshno, Vitaly Maiorov, Yuly Makovoz, and Pencho Petrushev for reading various parts of this paper. All errors, omissions and other transgressions are the author's responsibility.

REFERENCES

R. A. Adams (1975), *Sobolev Spaces*, Academic Press, New York.

F. Albertini, E. D. Sontag and V. Maillot (1993), 'Uniqueness of weights for neural networks', in *Artificial Neural Networks for Speech and Vision* (R. J. Mammone, ed.), Chapman and Hall, London, pp. 113–125.

J.-G. Attali and G. Pagès (1997), 'Approximations of functions by a multilayer perceptron: a new approach', *Neural Networks* **10**, 1069–1081.

A. R. Barron (1992), 'Neural net approximation', in *Proc. Seventh Yale Workshop*

on Adaptive and Learning Systems, 1992 (K. S. Narendra, ed.), Yale University, New Haven, pp. 69–72.

A. R. Barron (1993), 'Universal approximation bounds for superpositions of a sigmoidal function', *IEEE Trans. Inform. Theory* **39**, 930–945.

A. R. Barron (1994), 'Approximation and estimation bounds for artificial neural networks', *Machine Learning* **14**, 115–133.

P. L. Bartlett, V. Maiorov and R. Meir (1998), 'Almost linear VC dimension bounds for piecewise polynomial networks', *Neural Computation* **10**, 2159–2173.

E. B. Baum (1988), 'On the capabilities of multilayer perceptrons', *J. Complexity* **4**, 193–215.

C. M. Bishop (1995), *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford.

E. K. Blum and L. K. Li (1991), 'Approximation theory and feedforward networks', *Neural Networks* **4**, 511–515.

M. D. Buhmann and A. Pinkus (1999), 'Identifying linear combinations of ridge functions', *Adv. Appl. Math.* **22**, 103–118.

R. M. Burton and H. G. Dehling (1998), 'Universal approximation in p-mean by neural networks', *Neural Networks* **11**, 661–667.

P. Cardaliaguet and G. Euvrard (1992), 'Approximation of a function and its derivatives with a neural network', *Neural Networks* **5**, 207–220.

S. M. Carroll and B. W. Dickinson (1989), 'Construction of neural nets using the Radon transform', in *Proceedings of the IEEE 1989 International Joint Conference on Neural Networks*, Vol. 1, IEEE, New York, pp. 607–611.

T. Chen and H. Chen (1993), 'Approximations of continuous functionals by neural networks with application to dynamic systems', *IEEE Trans. Neural Networks* **4**, 910–918.

T. Chen and H. Chen (1995), 'Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems', *IEEE Trans. Neural Networks* **6**, 911–917.

T. Chen, H. Chen and R. Liu (1995), 'Approximation capability in $C(\mathbb{R}^n)$ by multilayer feedforward networks and related problems', *IEEE Trans. Neural Networks* **6**, 25–30.

X. Chen and H. White (1999), 'Improved rates and asymptotic normality for nonparametric neural network estimators', preprint.

C. H. Choi and J. Y. Choi (1994), 'Constructive neural networks with piecewise interpolation capabilities for function approximations', *IEEE Trans. Neural Networks* **5**, 936–944.

C. K. Chui and X. Li (1992), 'Approximation by ridge functions and neural networks with one hidden layer', *J. Approx. Theory* **70**, 131–141.

C. K. Chui and X. Li (1993), 'Realization of neural networks with one hidden layer', in *Multivariate Approximations: From CAGD to Wavelets* (K. Jetter and F. Utreras, eds), World Scientific, Singapore, pp. 77–89.

C. K. Chui, X. Li and H. N. Mhaskar (1994), 'Neural networks for localized approximation', *Math. Comp.* **63**, 607–623.

C. K. Chui, X. Li and H. N. Mhaskar (1996), 'Limitations of the approximation capabilities of neural networks with one hidden layer', *Adv. Comput. Math.* **5**, 233–243.

E. Corominas and F. Sunyer Balaguer (1954), 'Condiciones para que una foncion infinitamente derivable sea un polinomo', *Rev. Mat. Hisp. Amer.* **14**, 26–43.

N. E. Cotter (1990), 'The Stone–Weierstrass theorem and its application to neural networks', *IEEE Trans. Neural Networks* **1**, 290–295.

G. Cybenko (1989), 'Approximation by superpositions of a sigmoidal function', *Math. Control, Signals, and Systems* **2**, 303–314.

R. A. DeVore, R. Howard and C. Micchelli (1989), 'Optimal nonlinear approximation', *Manuscripta Math.* **63**, 469–478.

R. A. DeVore, K. I. Oskolkov and P. P. Petrushev (1997), 'Approximation by feedforward neural networks', *Ann. Numer. Math.* **4**, 261–287.

L. Devroye, L. Gyorfi and G. Lugosi (1996), *A Probabilistic Theory of Pattern Recognition*, Springer, New York.

M. J. Donahue, L. Gurvits, C. Darken and E. Sontag (1997), 'Rates of convex approximation in non-Hilbert spaces', *Const. Approx.* **13**, 187–220.

W. F. Donoghue (1969), *Distributions and Fourier Transforms*, Academic Press, New York.

T. Draelos and D. Hush (1996), 'A constructive neural network algorithm for function approximation', in *Proceedings of the IEEE 1996 International Conference on Neural Networks*, Vol. 1, IEEE, New York, pp. 50–55.

R. E. Edwards (1965), *Functional Analysis, Theory and Applications*, Holt, Rinehart and Winston, New York.

S. W. Ellacott (1994), 'Aspects of the numerical analysis of neural networks', in Vol. 3 of *Acta Numerica*, Cambridge University Press, pp. 145–202.

S. W. Ellacott and D. Bos (1996), *Neural Networks: Deterministic Methods of Analysis*, International Thomson Computer Press, London.

C. Fefferman (1994), 'Reconstructing a neural net from its output', *Revista Mat. Iberoamer.* **10**, 507–555.

R. A. Finan, A. T. Sapeluk and R. I. Damper (1996), 'Comparison of multilayer and radial basis function neural networks for text-dependent speaker recognition', in *Proceedings of the IEEE 1996 International Conference on Neural Networks*, Vol. 4, IEEE, New York, pp. 1992–1997.

H. L. Frisch, C. Borzi, D. Ord, J. K. Percus and G. O. Williams (1989), 'Approximate representation of functions of several variables in terms of functions of one variable', *Phys. Review Letters* **63**, 927–929.

K. Funahashi (1989), 'On the approximate realization of continuous mappings by neural networks', *Neural Networks* **2**, 183–192.

A. R. Gallant and H. White (1988), 'There exists a neural network that does not make avoidable mistakes', in *Proceedings of the IEEE 1988 International Conference on Neural Networks*, Vol. 1, IEEE, New York, pp. 657–664.

A. R. Gallant and H. White (1992), 'On learning the derivatives of an unknown mapping with multilayer feedforward networks', *Neural Networks* **5**, 129–138.

S. Geva and J. Sitte (1992), 'A constructive method for multivariate function approximation by multilayer perceptrons', *IEEE Trans. Neural Networks* **3**, 621–624.

F. Girosi and T. Poggio (1989), 'Representation properties of networks: Kolmogorov's theorem is irrelevant', *Neural Computation* **1**, 465–469.

F. Girosi and T. Poggio (1990), 'Networks and the best approximation property', *Biol. Cybern.* **63**, 169–176.

M. Gori, F. Scarselli and A. C. Tsoi (1996), 'Which classes of functions can a given multilayer perceptron approximate?', in *Proceedings of the IEEE 1996 International Conference on Neural Networks*, Vol. 4, IEEE, New York, pp. 2226–2231.

S. Haykin (1994), *Neural Networks*, MacMillan, New York.

R. Hecht-Nielsen (1987), 'Kolmogorov's mapping neural network existence theorem', in *Proceedings of the IEEE 1987 International Conference on Neural Networks*, Vol. 3, IEEE, New York, pp. 11–14.

R. Hecht-Nielsen (1989), 'Theory of the backpropagation neural network', in *Proceedings of the IEEE 1989 International Joint Conference on Neural Networks*, Vol. 1, IEEE, New York, pp. 593–605.

K. Hornik (1991), 'Approximation capabilities of multilayer feedforward networks', *Neural Networks* **4**, 251–257.

K. Hornik (1993), 'Some new results on neural network approximation', *Neural Networks* **6**, 1069–1072.

K. Hornik, M. Stinchcombe and H. White (1989), 'Multilayer feedforward networks are universal approximators', *Neural Networks* **2**, 359–366.

K. Hornik, M. Stinchcombe and H. White (1990), 'Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks', *Neural Networks* **3**, 551–560.

K. Hornik, M. Stinchcombe, H. White and P. Auer (1994), 'Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives', *Neural Computation* **6**, 1262–1275.

G. B. Huang and H. A. Babri (1998), 'Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions', *IEEE Trans. Neural Networks* **9**, 224–229.

S. C. Huang and Y. F. Huang (1991), 'Bounds on the number of hidden neurons in multilayer perceptrons', *IEEE Trans. Neural Networks* **2**, 47–55.

B. Irie and S. Miyake (1988), 'Capability of three-layered perceptrons', in *Proceedings of the IEEE 1988 International Conference on Neural Networks*, Vol. 1, IEEE, New York, pp. 641–648.

Y. Itô (1991*a*), 'Representation of functions by superpositions of a step or a sigmoid function and their applications to neural network theory', *Neural Networks* **4**, 385–394.

Y. Itô (1991*b*), 'Approximation of functions on a compact set by finite sums of a sigmoid function without scaling', *Neural Networks* **4**, 817–826.

Y. Itô (1992), 'Approximation of continuous functions on $\mathbb{R}^d$ by linear combinations of shifted rotations of a sigmoid function with and without scaling', *Neural Networks* **5**, 105–115.

Y. Itô (1993), 'Approximations of differentiable functions and their derivatives on compact sets by neural networks', *Math. Scient.* **18**, 11–19.

Y. Itô (1994*a*), 'Approximation capabilities of layered neural networks with sigmoidal units on two layers', *Neural Computation* **6**, 1233–1243.

Y. Itô (1994*b*), 'Differentiable approximation by means of the Radon transformation and its applications to neural networks', *J. Comput. Appl. Math.* **55**, 31–50.

Y. Itô (1996), 'Nonlinearity creates linear independence', *Adv. Comput. Math.* **5**, 189–203.

Y. Itô and K. Saito (1996), 'Superposition of linearly independent functions and finite mappings by neural networks', *Math. Scient.* **21**, 27-33.

L. K. Jones (1990), 'Constructive approximations for neural networks by sigmoidal functions', *Proc. IEEE* **78**, 1586–1589. Correction and addition, *Proc. IEEE* (1991) **79**, 243.

L. K. Jones (1992), 'A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training', *Ann. Stat.* **20**, 608–613.

L. K. Jones (1994), 'Good weights and hyperbolic kernels for neural networks, projection pursuit, and pattern classification: Fourier strategies for extracting information from high-dimensional data', *IEEE Trans. Inform. Theory* **40**, 439–454.

L. K. Jones (1997), 'The computational intractability of training sigmoidal neural networks', *IEEE Trans. Inform. Theory* **43**, 167–173.

L. K. Jones (1999), 'Local greedy approximation for nonlinear regression and neural network training', preprint.

J. P. Kahane (1959), *Lectures on Mean Periodic Functions*, Tata Institute, Bombay.

P. C. Kainen, V. Kurkova and A. Vogt (1999), 'Approximation by neural networks is not continuous', preprint.

H. Katsuura and D. A. Sprecher (1994), 'Computational aspects of Kolmogorov's superposition theorem', *Neural Networks* **7**, 455–461.

V. Y. Kreinovich (1991), 'Arbitrary nonlinearity is sufficient to represent all functions by neural networks: a theorem', *Neural Networks* **4**, 381–383.

V. Kurkova (1991), 'Kolmogorov's theorem is relevant', *Neural Computation* **3**, 617–622.

V. Kurkova (1992), 'Kolmogorov's theorem and multilayer neural networks', *Neural Networks* **5**, 501–506.

V. Kurkova (1995*a*), 'Approximation of functions by perceptron networks with bounded number of hidden units', *Neural Networks* **8**, 745–750.

V. Kurkova (1995*b*), 'Kolmogorov's theorem', in *The Handbook of Brain Theory and Neural Networks*, (M. Arbib, ed.), MIT Press, Cambridge, pp. 501–502.

V. Kurkova (1996), 'Trade-off between the size of weights and the number of hidden units in feedforward networks', *Neural Network World* **2**, 191–200.

V. Kurkova and P. C. Kainen (1994), 'Functionally equivalent feedforward neural networks', *Neural Computation* **6**, 543–558.

V. Kurkova, P. C. Kainen and V. Kreinovich (1997), 'Estimates of the number of hidden units and variation with respect to half-spaces', *Neural Networks* **10**, 1061–1068.

A. Lapedes and R. Farber (1988), 'How neural nets work', in *Neural Information Processing Systems* (D. Z. Anderson, ed.), American Institute of Physics, New York, pp. 442–456.

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

X. Li (1996), 'Simultaneous approximations of multivariate functions and their derivatives by neural networks with one hidden layer', *Neurocomputing* **12**, 327–343.

W. A. Light (1993), 'Ridge functions, sigmoidal functions and neural networks', in *Approximation Theory VII* (E. W. Cheney, C. K. Chui and L. L. Schumaker, eds), Academic Press, New York, pp. 163–206.

J. N. Lin and R. Unbehauen (1993), 'On realization of a Kolmogorov network', *Neural Computation* **5**, 18–20.

V. Ya. Lin and A. Pinkus (1993), 'Fundamentality of ridge functions', *J. Approx. Theory* **75**, 295–311.

V. Ya. Lin and A. Pinkus (1994), 'Approximation of multivariate functions', in *Advances in Computational Mathematics: New Delhi, India*, (H. P. Dikshit and C. A. Micchelli, eds), World Scientific, Singapore, pp. 257–265.

R. P. Lippman (1987), 'An introduction to computing with neural nets', *IEEE Magazine* **4**, 4–22.

G. G. Lorentz, M. von Golitschek and Y. Makovoz (1996), *Constructive Approximation: Advanced Problems*, Vol. 304 of *Grundlehren*, Springer, Berlin.

V. E. Maiorov (1999), 'On best approximation by ridge functions', to appear in *J. Approx. Theory*

V. E. Maiorov and R. Meir (1999), 'On the near optimality of the stochastic approximation of smooth functions by neural networks', to appear in *Adv. Comput. Math.*

V. Maiorov, R. Meir and J. Ratsaby (1999), 'On the approximation of functional classes equipped with a uniform measure using ridge functions', to appear in *J. Approx. Theory.*

V. Maiorov and A. Pinkus (1999), 'Lower bounds for approximation by MLP neural networks', *Neurocomputing* **25**, 81–91.

Y. Makovoz (1996), 'Random approximants and neural networks', *J. Approx. Theory* **85**, 98–109.

Y. Makovoz (1998), 'Uniform approximation by neural networks', *J. Approx. Theory* **95**, 215–228.

M. Meltser, M. Shoham and L. M. Manevitz (1996), 'Approximating functions by neural networks: a constructive solution in the uniform norm', *Neural Networks* **9**, 965–978.

H. N. Mhaskar (1993), 'Approximation properties of a multilayered feedforward artificial neural network', *Adv. Comput. Math.* **1**, 61–80.

H. N. Mhaskar (1994), 'Approximation of real functions using neural networks', in *Advances in Computational Mathematics: New Delhi, India*, (H. P. Dikshit and C. A. Micchelli, eds), World Scientific, Singapore, pp. 267–278.

H. N. Mhaskar (1996), 'Neural networks for optimal approximation of smooth and analytic functions', *Neural Computation* **8**, 164–177.

H. N. Mhaskar and N. Hahm (1997), 'Neural networks for functional approximation and system identification', *Neural Computation* **9**, 143–159.

H. N. Mhaskar and C. A. Micchelli (1992), 'Approximation by superposition of a sigmoidal function and radial basis functions', *Adv. Appl. Math.* **13**, 350–373.

H. N. Mhaskar and C. A. Micchelli (1993), 'How to choose an activation function',

in Vol. 6 of *Neural Information Processing Systems* (J. D. Cowan, G. Tesauro and J. Alspector, eds), Morgan Kaufman, San Francisco, pp. 319–326.

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

H. N. Mhaskar and C. A. Micchelli (1995), 'Degree of approximation by neural and translation networks with a single hidden layer', *Adv. Appl. Math.* **16**, 151–183.

H. N. Mhaskar and J. Prestin (1999), 'On a choice of sampling nodes for optimal approximation of smooth functions by generalized translation networks', to appear in *Proceedings of International Conference on Artificial Neural Networks*, Cambridge, England.

M. Nees (1994), 'Approximative versions of Kolmogorov's superposition theorem, proved constructively', *J. Comput. Appl. Anal.* **54**, 239–250.

M. Nees (1996), 'Chebyshev approximation by discrete superposition: Application to neural networks', *Adv. Comput. Math.* **5**, 137–151.

K. I. Oskolkov (1997), 'Ridge approximation, Chebyshev–Fourier analysis and optimal quadrature formulas', *Tr. Mat. Inst. Steklova* **219** *Teor. Priblizh. Garmon. Anal.*, 269–285.

P. P. Petrushev (1998), 'Approximation by ridge functions and neural networks', *SIAM J. Math. Anal.* **30**, 155–189.

A. Pinkus (1995), 'Some density problems in multivariate approximation', in *Approximation Theory: Proceedings of the International Dortmund Meeting IDoMAT 95*, (M. W. Müller, M. Felten and D. H. Mache, eds), Akademie Verlag, Berlin, pp. 277–284.

A. Pinkus (1996), 'TDI-Subspaces of $C(\mathbb{R}^d)$ and some density problems from neural networks', *J. Approx. Theory* **85**, 269–287.

A. Pinkus (1997), 'Approximating by ridge functions', in *Surface Fitting and Multiresolution Methods*, (A. Le Méhauté, C. Rabut and L. L. Schumaker, eds), Vanderbilt University Press, Nashville, pp. 279–292.

G. Pisier (1981), 'Remarques sur un resultat non publié de B. Maurey', in *Seminaire D'Analyse Fonctionnelle, 1980–1981*, École Polytechnique, Centre de Mathématiques, Palaiseau, France.

B. D. Ripley (1994), 'Neural networks and related methods for classification', *J. Royal Statist. Soc., B* **56**, 409–456.

B. D. Ripley (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge.

H. L. Royden (1963), *Real Analysis*, MacMillan, New York.

W. S. Sarle (1998), editor of *Neural Network, FAQ, parts 1 to 7*, Usenet newsgroup `comp.ai.neural-nets`, `ftp://ftp.sas.com/pub/neural/FAQ.html`

M.A. Sartori and P. J. Antsaklis (1991), 'A simple method to derive bounds on the size and to train multilayer neural networks', *IEEE Trans. Neural Networks* **2**, 467–471.

F. Scarselli and A. C. Tsoi (1998), 'Universal approximation using feedforward neural networks: a survey of some existing methods, and some new results', *Neural Networks* **11**, 15–37.

L. Schwartz (1944), 'Sur certaines familles non fondamentales de fonctions continues', *Bull. Soc. Math. France* **72**, 141–145.

L. Schwartz (1947), 'Théorie générale des fonctions moyenne-périodiques', *Ann. Math.* **48**, 857–928.

K. Y. Siu, V. P. Roychowdhury and T. Kailath (1994), 'Rational approximation techniques for analysis of neural networks', *IEEE Trans. Inform. Theory* **40**, 455–46.

E. D. Sontag (1992), 'Feedforward nets for interpolation and classification', *J. Comput. System Sci.* **45**, 20–48.

D. A. Sprecher (1993), 'A universal mapping for Kolmogorov's superposition theorem', *Neural Networks* **6**, 1089–1094.

D. A. Sprecher (1997), 'A numerical implementation of Kolmogorov's superpositions II', *Neural Networks* **10**, 447–457.

M. Stinchcombe (1995), 'Precision and approximate flatness in artificial neural networks', *Neural Computation* **7**, 1021–1039.

M. Stinchcombe and H. White (1989), 'Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions', in *Proceedings of the IEEE 1989 International Joint Conference on Neural Networks*, Vol. 1, IEEE, New York, pp. 613–618.

M. Stinchcombe and H. White (1990), 'Approximating and learning unknown mappings using multilayer feedforward networks with bounded weights', in *Proceedings of the IEEE 1990 International Joint Conference on Neural Networks*, Vol. 3, IEEE, New York, pp. 7–16.

B. G. Sumpter, C. Getino and D. W. Noid (1994), 'Theory and applications of neural computing in chemical science', *Annual Rev. Phys. Chem.* **45**, 439–481.

H. J. Sussmann (1992), 'Uniqueness of the weights for minimal feedforward nets with a given input–output map', *Neural Networks* **5**, 589–593.

Y. Takahashi (1993), 'Generalization and approximation capabilities of multilayer networks', *Neural Computation* **5**, 132–139.

J. de Villiers and E. Barnard (1992), 'Backpropagation neural nets with one and two hidden layers', *IEEE Trans. Neural Networks* **4**, 136–141.

B. A. Vostrecov and M. A. Kreines (1961), 'Approximation of continuous functions by superpositions of plane waves', *Dokl. Akad. Nauk SSSR* **140**, 1237–1240 = *Soviet Math. Dokl.* **2**, 1326–1329.

Z. Wang, M. T. Tham and A. J. Morris (1992), 'Multilayer feedforward neural networks: a canonical form approximation of nonlinearity', *Internat. J. Control* **56**, 655–672.

S. Watanabe (1996), 'Solvable models of layered neural networks based on their differential structure', *Adv. Comput. Math.* **5**, 205–231.

R. C. Williamson and U. Helmke (1995), 'Existence and uniqueness results for neural network approximations', *IEEE Trans. Neural Networks* **6**, 2–13.

J. Wray and G. G. Green (1995), 'Neural networks, approximation theory and finite precision computation', *Neural Networks* **8**, 31–37.

Y. Xu, W. A. Light and E. W. Cheney (1993), 'Constructive methods of approximation by ridge functions and radial functions', *Numerical Alg.* **4**, 205–223.

J. E. Yukich, M. B. Stinchcombe and H. White (1995), 'Sup-norm approximation bounds for networks through probabilistic methods', *IEEE Trans. Inform. Theory* **41**, 1021–1027.