Neural Networks, Vol. 4, pp. 251–257, 1991
Printed in the USA.

0893-6080/91 $3.00 + .00

ORIGINAL CONTRIBUTION

Approximation Capabilities of Multilayer Feedforward Networks

KURT HORNIK

Technische Universität Wien, Vienna, Austria

(*Received* 30 *January* 1990; *revised and accepted* 25 *October* 1990)

Abstract—*We show that standard multilayer feedforward networks with as few as a single hidden layer and arbitrary bounded and nonconstant activation function are universal approximators with respect to $L^p(\mu)$ performance criteria, for arbitrary finite input environment measures μ, provided only that sufficiently many hidden units are available. If the activation function is continuous, bounded and nonconstant, then continuous mappings can be learned uniformly over compact input sets. We also give very general conditions ensuring that networks with sufficiently smooth activation functions are capable of arbitrarily accurate approximation to a function and its derivatives.*

Keywords—Multilayer feedforward networks, Activation function, Universal approximation capabilities, Input environment measure, $L^p(\mu)$ approximation, Uniform approximation, Sobolev spaces, Smooth approximation.

1. INTRODUCTION

The approximation capabilities of neural network architectures have recently been investigated by many authors, including Carroll and Dickinson (1989), Cybenko (1989), Funahashi (1989), Gallant and White (1988), Hecht-Nielsen (1989), Hornik, Stinchcombe, and White (1989, 1990), Irie and Miyake (1988), Lapedes and Farber (1988), Stinchcombe and White (1989, 1990). (This list is by no means complete.)

If we think of the network architecture as a rule for computing values at l output units given values at k input units, hence implementing a class of mappings from R^k to R^l, we can ask how well arbitrary mappings from R^k to R^l can be approximated by the network, in particular, if as many hidden units as required for internal representation and computation may be employed.

How to measure the accuracy of approximation depends on how we measure closeness between functions, which in turn varies significantly with the specific problem to be dealt with. In many applications, it is necessary to have the network perform *simultaneously* well on all input samples taken from some compact input set X in R^k. In this case, closeness is measured by the uniform distance between functions on X, that is,

$$\rho_{\mu,X}(f, g) = \sup_{x\in X} |f(x) - g(x)|.$$

In other applications, we think of the inputs as random variables and are interested in the *average performance* where the average is taken with respect to the input environment measure μ, where $\mu(R^k) < \infty$. In this case, closeness is measured by the $L^p(\mu)$ distances

$$\rho_{p,\mu}(f, g) = \left[\int_{R^k} |f(x) - g(x)|^p \, d\mu(x)\right]^{1/p},$$

$1 \le p < \infty$, the most popular choice being $p = 2$, corresponding to mean square error.

Of course, there are many more ways of measuring closeness of functions. In particular, in many applications, it is also necessary that the *derivatives* of the approximating function implemented by the network closely resemble those of the function to be approximated, up to some order. This issue was first taken up in Hornik et al. (1990), who discuss the sources of need of smooth functional approximation in more detail. Typical examples arise in robotics (learning of smooth movements) and signal processing (analysis of chaotic time series); for a recent application to problems of nonparametric inference in statistics and econometrics, see Gallant and White (1989).

All papers establishing certain approximation ca-

Requests for reprints should be sent to Kurt Hornik, Institut für Statistik und Wahrscheinlichkeitstheorie, Technische Universität Wien, Wiedner Hauptstraße 8-10/107, A-1040 Wien, Austria.

pabilities of multilayer perceptrons thus far have been successful only by making more or less explicit assumptions on the activation function ψ, for example, by assuming ψ to be integrable, or sigmoidal respectively squashing (sigmoidal and monotone), etc. In this article, we shall demonstrate that these assumptions are unnecessary. We shall show that whenever ψ is *bounded and nonconstant*, then, for arbitrary input environment measures μ, standard multilayer feedforward networks with activation function ψ can approximate any function in $L^p(\mu)$ (the space of all functions on R^k such that $\int_{R^k} |f(x)|^p \, d\mu(x) < \infty$) arbitrarily well if closeness is measured by $\rho_{p,\mu}$, provided that sufficiently many hidden units are available.

Similarly, we shall establish that whenever ψ is *continuous*, *bounded and nonconstant*, then, for arbitrary compact subsets X of R^k, standard multilayer feedforward networks with activation function ψ can approximate any continuous function on X arbitrarily well with respect to uniform distance $\rho_{\mu,X}$, provided that sufficiently many hidden units are available. Hence, we conclude that it is not the specific choice of the activation function, but rather the *multilayer feedforward architecture* itself which gives neural networks the potential of being universal learning machines.

In addition to that, we significantly improve the results on smooth approximation capabilities of neural nets given in Hornik et al. (1990) by simultaneously relaxing the conditions to be imposed on the activation function and providing results for the previously uncovered cases of weighted Sobolev approximation with respect to finite input environment measures which do not have compact support, for example, Gaussian input distributions.

2. RESULTS

For notational convenience we shall explicitly formulate our results only for the case where there is only one hidden layer and one output unit. The corresponding results for the general multiple hidden layer multioutput case can easily be deduced from the simple case, cf. corollary 2.6 and 2.7 in Hornik et al. (1989).

If there is only one hidden layer and only one output unit, then the set of all functions implemented by such a network with n hidden units is

$$\mathfrak{N}_k^{(n)}(\psi) = \left\{h: R^k \to R \,|\, h(x) = \sum_{j=1}^{n} \beta_j \psi(a_j' x - \theta_j)\right\},$$

where ψ is the common activation function of the hidden units and $'$ denotes transpose so that if a has components $\alpha_1, \ldots, \alpha_k$ and x has components $\xi_1, \ldots, \xi_k$, $a'x$ is the dot product $\alpha_1\xi_1 + \cdots + \alpha_k\xi_k$. (Output units are always assumed to be linear.) The set of all functions implemented by such a network with an arbitrarily large number of hidden units is

$$\mathfrak{N}_k(\psi) = \bigcup_{n=1}^{\infty} \mathfrak{N}_k^{(n)}(\psi).$$

In what follows, some concepts from modern analysis will be needed. As a reference, we recommend Friedman (1982). For $1 \le p < \infty$, we write

$$\|f\|_{p,\mu} = \left[\int_{R^k} |f(x)|^p \, d\mu(x)\right]^{1/p}$$

so that $\rho_{p,\mu}(f, g) = \|f - g\|_{p,\mu}$. $L^p(\mu)$ is the space of all functions f such that $\|f\|_{p,\mu} < \infty$. A subset S of $L^p(\mu)$ is *dense* in $L^p(\mu)$ if for arbitrary $f \in L^p(\mu)$ and $\varepsilon > 0$ there is a function $g \in S$ such that $\rho_{p,\mu}(f, g) < \varepsilon$.

Theorem 1: *If ψ is unbounded and nonconstant, then $\mathfrak{N}_k(\psi)$ is dense in $L^p(\mu)$ for all finite measures μ on R^k.*

$C(X)$ is the space of all continuous functions on X. A subset S of $C(X)$ is *dense* in $C(X)$ if for arbitrary $f \in C(X)$ and $\varepsilon > 0$ there is a function $g \in S$ such that $\rho_{u,X}(f, g) < \varepsilon$.

Theorem 2: *If ψ is continuous, bounded and nonconstant, then $\mathfrak{N}_k(\psi)$ is dense in $C(X)$ for all compact subsets X of R^k.*

A k-tuple $\alpha = (\alpha_1, \ldots, \alpha_k)$ of nonnegative integers is called a multiindex. We then write $|\alpha| = \alpha_1 + \cdots + \alpha_k$ for the order of the multiindex α and

$$D^\alpha f(x) = \frac{\partial^{\alpha_1 + \cdots + \alpha_k} f}{\partial \xi_1^{\alpha_1} \cdots \partial \xi_k^{\alpha_k}}(x)$$

for the corresponding partial derivative of a sufficiently smooth function f of $x = (\xi_1, \ldots, \xi_k)' \in R^k$.

$C^m(R^k)$ is the space of all functions f which, together with all their partial derivatives $D^\alpha f$ of order $|\alpha| \le m$, are continuous on R^k. For all subsets X of R^k and $f \in C^m(R^k)$, let

$$\|f\|_{m,u,X} := \max_{|\alpha| \le m} \sup_{x \in X} |D^\alpha f(x)|.$$

A subset S of $C^m(R^k)$ is *uniformly* m *dense on compacta* in $C^m(R^k)$ if for all $f \in C^m(R^k)$, for all compact subsets X of R^k, and for all $\varepsilon > 0$ there is a function $g = g(f, X, \varepsilon) \in S$ such that $\|f - g\|_{m,u,X} < \varepsilon$.

For $f \in C^m(R^k)$, μ a finite measure on R^k and $1 \le p < \infty$, let

$$\|f\|_{m,p,u} := \left[\sum_{|\alpha| \le m} \int_{R^k} |D^\alpha f|^p \, d\mu\right]^{1/p},$$

and let the *weighted Sobolev space* $C^{m,p}(\mu)$ be defined by

$$C^{m,p}(\mu) = \{f \in C^m(R^k) : \|f\|_{m,p,u} < \infty\}.$$

Observe that $C^{m,p}(\mu) = C^m(R^k)$ if μ has compact support. A subset S of $C^{m,p}(\mu)$ is *dense* in $C^{m,p}(\mu)$, if for all $f \in C^{m,p}(\mu)$ and $\varepsilon > 0$ there is a function $g = g(f, \varepsilon) \in S$ such that $\|f - g\|_{m,p,\mu} < \varepsilon$.

We then have the following results.

Theorem 3: *If $\psi \in C^m(R^k)$ is nonconstant and bounded, then $\mathfrak{N}_k(\psi)$ is uniformly m-dense on compacta in $C^m(R^k)$ and dense in $C^{m,p}(\mu)$ for all finite measures μ on R^k with compact support.*

Theorem 4: *If $\psi \in C^m(R^k)$ is nonconstant and all its derivatives up to order m are bounded, then $\mathfrak{N}_k(\psi)$ is dense in $C^{m,p}(\mu)$ for all finite measures μ on R^k.*

3. DISCUSSION

The conditions imposed on ψ in our theorems are very general. In particular, they are satisfied by all smooth squashing activation functions—such as the logistic squasher or the arctangent squasher—that have become popular in neural network applications.

A lot of corollaries can be deduced from our theorems. In particular, as convergence in $L^p(\mu)$ implies convergence in μ measure, we conclude from Theorem 1 that whenever ψ is bounded and nonconstant, all measurable functions on R^k can be approximated by functions in $\mathfrak{N}_k(\psi)$ in μ measure. It follows that (cf. Lemma 2.1 in Hornik et al. [1989]) for arbitrary measurable functions f and $\varepsilon > 0$, we can find a compact subset X_ε of R^k and a function $g \in \mathfrak{N}_k(\psi)$ such that

$$\rho_{u,X_\varepsilon}(f, g) < \varepsilon, \qquad \mu(R^k \backslash X_\varepsilon) < \varepsilon.$$

This substantially improves Theorems 3 and 5 in Cybenko (1989) and Corollary 2.1 in Hornik et al. (1989), and is of basic importance for the use of artificial neural networks in classification and decision problems, cf. Cybenko (1989), Sections 3 and 4.

If the activation function is constant, only constant mappings can be learned, which is definitely not a very interesting case. The continuity assumption in Theorem 2 can be weakened. For example, Theorem 2.4 in Hornik et al. (1989) shows that whenever ψ is a squashing function, then $\mathfrak{N}_k(\psi)$ is dense in $C(X)$ for all compact subsets of R^k. In fact, their method can easily be modified to deliver the same uniform approximation capability whenever ψ has distinct finite limits at $\pm\infty$. Whether or not the continuity assumption can entirely be dropped is still an open (and quite challenging) problem.

There are, of course, unbounded functions which are capable of uniform approximation. For example, a simple application of the Stone–Weierstraß theorem (cf. Hornik et al. [1989]) implies that $\mathfrak{N}_k(\exp)$ is dense in $C(X)$, where of course exp is the standard exponential function. However, our theorems do definitely not remain valid for *all* unbounded activation functions. If ψ is a polynomial of degree d ($d \geq 1$), then $\mathfrak{N}_k(\psi)$ is just the space P_d of all polynomials in k variables of degree less than or equal to d. Hence, for all reasonably rich input spaces X or input environment measures μ, $\mathfrak{N}_k(\psi)$ cannot be dense in $C(X)$ or $L^p(\mu)$, respectively. Also, if the tail behavior of an unbounded function ψ is not compatible with the tail behavior of μ, then $x \mapsto \psi(a'x - \theta)$ may not be an element of $L^p(\mu)$ for most or all nonzero $\alpha \in R^k$.

By allowing for a much larger class of activation functions, Theorem 3 significantly improves the results in Hornik et al. (1990), where the conclusions of Theorem 3 are established under the assumption that there exists some $l \geq m$ such that $\psi \in C^l(R)$ and $0 < \int_R |D^l\psi| \, dt < \infty$ (l-finiteness). However, many interesting functions, such as all nonconstant *periodic* functions, are not l finite. Using Theorem 3 we easily infer that if ψ is a nonconstant finite linear combination of periodic functions in $C^m(R)$ (in particular, if ψ is a nonconstant trigonometric polynomial), then $\mathfrak{N}_k(\psi)$ is uniformly m dense on compacta in $C^m(R^k)$. Other interesting examples that can now be dealt with are functions such as $\psi(t) = \sin(t)/t$ (which is not l finite for any l), or more generally, all functions which are the Fourier transform of some finite signed measure which has finite absolute moments up to order m (such functions are usually not l finite).

Theorem 4 gives weighted Sobolev type approximation results for the previously uncovered case of finite input environment measures which are not compactly supported. Using Theorem 4 we may conclude that if ψ is the logistic or arctangent squasher, or a nonconstant trigonometric polynomial, then $\mathfrak{N}_k(\psi)$ is dense in $C^{m,p}(\mu)$, for all finite measures μ. In particular, we now have a result for inputs that follows a multivariate Gaussian distribution.

The following generalization of our results is immediate: suppose that ψ is unbounded, but that there is a nonconstant and bounded function $\phi \in \mathfrak{N}_1(\psi)$. Then, by Theorem 1, $\mathfrak{N}_k(\phi)$ is dense in $L^p(\mu)$. As $\mathfrak{N}_k(\phi) \subset \mathfrak{N}_k(\psi)$, we can state that in this case, $\mathfrak{N}_k(\psi)$ contains a subset which is dense in $L^p(\mu)$. (Observe that if the support of μ is not compact and ψ is unbounded, we do not necessarily have $\mathfrak{N}_k(\psi) \subset L^p(\mu)$; hence, we cannot simply state that $\mathfrak{N}_k(\psi)$ itself is dense in $L^p(\mu)$.) Similar considerations apply for the other theorems.

If Ω is an open subset of R^k, let $C^m(\Omega)$ be the space of all functions f which, together with all their partial derivatives $D^\alpha f$ of order $|\alpha| \leq m$, are continuous on Ω. Let us say that a subset S of $C^m(\Omega)$ is *uniformly* m *dense on compacta* in $C^m(\Omega)$ if for all $f \in C^m(\Omega)$, for all compact subsets X of Ω, and for all $\varepsilon > 0$ there is a function $g = g(f, X, \varepsilon) \in S$ such that $\|f - g\|_{m,u,X} < \varepsilon$.

It is easily seen that under the conditions of Theorem 3, $\mathfrak{N}_k(\psi)$ is uniformly m dense on compacta in $C^m(\Omega)$ for all open subsets Ω of R^k. In fact, it suffices to show that whenever $f \in C^m(\Omega)$ and X is a compact subset of Ω, then we can find a function $E_X f \in C^m(R^k)$ satisfying $E_X f(x) = f(x)$ for all $x \in X$. Now, by Problem 3.3.1 in Friedman (1982), we can find a function $h \in C^\infty(R^k)$ such that $h = 1$ on X, $0 \le h \le 1$ on $\Omega \backslash X$, and $h = 0$ outside Ω. Take $E_X f = hf$ on Ω and $E_X f = 0$ outside Ω.

Suppose that Ω is bounded. Functions f in $C^m(\Omega)$ do not necessarily satisfy $\|f\|_{m,u,\Omega} < \infty$. On the other hand, all functions in $C^m(R^k)$, and hence in particular all functions in $\mathfrak{N}_k(\psi)$ if $\psi \in C^m(R^k)$, satisfy $\|g\|_{m,u,X} < \infty$ for *each* compact subset X of R^k. Hence in general, it is not possible to approximate functions in $C^m(\Omega)$ by functions $\mathfrak{N}_k(\psi)$ arbitrarily well with respect to $\|\cdot\|_{m,u,\Omega}$.

However, one might ask whether such approximation is possible for at least all functions in the space $C_b^m(\Omega)$ which consists of all functions $f \in C^m(\Omega)$ for which $D^\alpha f$ is bounded and uniformly continuous on Ω for $0 \le |\alpha| \le m$. The following prominent counterexample shows that this is not always possible. Let $k = 1$, $\Omega = (-1,0) \cup (0,1)$ and let $f = 0$ on $(-1,0)$ and $f = 1$ on $(0,1)$. Then $f \in C_b^\infty(\Omega)$, but it is obviously impossible to approximate f by continuous functions on R uniformly over Ω. In fact, we always have $\|f - g\|_{0,u,\Omega} \ge 1/2$ for all $g \in C(R)$. Roughly speaking, if Ω is bounded, then $\mathfrak{N}_k(\psi)$ approximates all functions in $C_b^m(\Omega)$ arbitrarily well with respect to $\|\cdot\|_{m,u,\Omega}$ if the geometry of Ω is such that functions $f \in C_b^m(\Omega)$ can be extended to functions in $C^m(R^k)$. (Cf. also the next paragraph.)

Classical (nonweighted) Sobolev spaces are defined as follows. Let Ω be an open set in R^k, let the input environment measure μ be standard Lebesgue measure on Ω, for functions $f \in C^m(\Omega)$ let

$$\|f\|_{m,p,\Omega} = \left[\sum_{|\alpha| \le m} \int_\Omega |D^\alpha f|^p \, dx\right]^{1/p},$$

and let

$$H^{m,p}(\Omega) = \{f \in C^m(\Omega): \|f\|_{m,p,\Omega} < \infty\}.$$

(More precisely, standard Sobolev spaces are defined as the completions of the above $H^{m,p}(\Omega)$ with respect to $\|\cdot\|_{m,p,\Omega}$. The elements of these spaces are not necessarily classically smooth functions, but have generalized derivatives. See, for example, the discussion in Hornik et al. (1990).)

It is easily seen that globally smooth functions on R^k are not dense in $H^{m,p}(\Omega)$ (with respect to $\|\cdot\|_{m,p,\Omega}$) for most domains Ω. In the above example, no function in $C^1(R)$ can approximate f in $H^{1,1}(\Omega)$. Arbitrarily close approximations by globally smooth functions on R^k are only possible under certain conditions on the geometry of Ω that somehow exclude the possibility that Ω lies on both sides of part of its boundary. Such conditions are, for example, that Ω has the *segment property* (Adams, 1975, Theorem 3.18) or that Ω is *starshaped with respect to a point* (Maz'ja, 1985, Theorem 1.1.6.1). In both cases, it can be shown that $C_0^\infty(R^k)$, the space of all functions on R^k with compact support which are infinitely often continuously differentiable, is dense in $H^{m,p}(\Omega)$. Hence, if in addition Ω is bounded, $\mathfrak{N}_k(\psi)$ is dense in $H^{m,p}(\Omega)$ under the conditions of Theorem 3.

If the underlying input environment measure μ is not finite, but is regular in the sense that $\mu(X) < \infty$ for all compact subsets X of R^k (as an example we may take standard Lebesgue measure on R^k), then $\mathfrak{N}_k(\psi)$ is dense in all $L^p_{\text{loc}}(\mu)$ spaces, $1 \le p < \infty$, whenever ψ is bounded and nonconstant, improving results in Stinchcombe and White (1989).

Similarly, we can measure closeness of functions in $C^m(R^k)$ by the *local* weighted Sobolev space distance measure

$$\rho_{m,p,\text{loc},\mu}(f, g) := \sum_{n=1}^\infty 2^{-n} \min(\|f - g\|_{m,p,\mu_n}, 1),$$

where $1 \le p \le \infty$, μ_n is the restriction of μ to some bounded set X_n and the X_n exhaust all of R^k, that is, $\cup_{n=1}^\infty X_n = R^k$. It follows straightforwardly that, under the conditions of Theorem 3, $\mathfrak{N}_k(\psi)$ is dense in $C^m(R^k)$ with respect to $\rho_{m,p,\text{loc},\mu}$.

Concluding Remark

In this article, we established that multilayer feedforward networks are, under very general conditions on the hidden unit activation function, universal approximators provided that sufficiently many hidden units are available. However, it should be emphasized that our results do not mean that all activation functions ψ will *perform* equally well in specific learning problems. In applications, additional issues as, for example, minimal redundancy or computational efficiency, have to be taken into account as well.

4. PROOFS

In order to establish our theorems, we follow an approach first utilized by Cybenko (1989) that is based on an application of the Hahn–Banach theorem combined with representation theorems for continuous linear functionals on the function spaces under consideration.

Proof of Theorems 1 and 2: As ψ is bounded, $\mathfrak{N}_k(\psi)$ is a linear subspace of $L^p(\mu)$ for all finite measures μ on R^k. If, for some μ, $\mathfrak{N}_k(\psi)$ is not dense in $L^p(\mu)$, Corollary 4.8.7 in Friedman (1982) yields that there is a nonzero continuous linear functional Λ on $L^p(\mu)$ that vanishes on $\mathfrak{N}_k(\psi)$.

As well known (Friedman, 1982, Corollary 4.14.4 and Theorem 4.14.6), Λ is of the form $f \mapsto \Lambda(f) = \int_{R^k} f\, g\, d\mu$ with some g in $L^q(\mu)$, where q is the conjugate exponent $q = p/(p-1)$. (For $p = 1$ we obtain $q = \infty$; $L^\infty(\mu)$ is the space of all functions f for which the μ essential supremum

$$\|f\|_{\infty,\mu} = \inf\{N > 0 : \mu\{x \in R^k : |f(x)| > N\} = 0\}$$

is finite, that is, the space of all μ essentially bounded functions.)

If we write $\sigma(B) = \int_B g\, d\mu$, we find by Hölder's inequality that for all B,

$$|\sigma(B)| = \left| \int_{R^k} \mathbf{1}_B\, g\, d\mu \right| \le \|\mathbf{1}_B\|_{p,\mu} \|g\|_{q,\mu} \le (\mu(R^k))^{1/p} \|g\|_{q,\mu} < \infty,$$

hence σ is a nonzero finite signed measure on R^k such that $\Lambda(f) = \int_{R^k} f\, g\, d\mu = \int_{R^k} f\, d\sigma$. As Λ vanishes on $\mathfrak{N}_k(\psi)$, we conclude that in particular

$$\int_{R^k} \psi(a'x - \theta)\, d\sigma(x) = 0$$

for all $a \in R^k$ and $\theta \in R$.

Similarly, suppose that ψ is continuous and that for some compact subset X of R^k, $\mathfrak{N}_k(\psi)$ is not dense in $C(X)$. Proceeding as in the proof of Theorem 1 in Cybenko (1989), we find that in this case there exists a nonzero finite signed measure σ on R^k (σ is actually concentrated on X) such that

$$\int_{R^k} \psi(a'x - \theta)\, d\sigma(x) = 0$$

for all $a \in R^k$, $\theta \in R$.

Summing up, in either case we arrive at the following question. Can there exist a *nonzero* finite signed measure σ on R^k such that $\int_{R^k} \psi(a'x - \theta)\, d\sigma(x)$ vanishes for all $a \in R^k$ and $\theta \in R$? This question was first asked and investigated by Cybenko (1989) who basically gave the following definition.

Definition. A bounded function ψ is called *discriminatory* if no nonzero finite signed measure σ on R^k exists such that

$$\int_{R^k} \psi(a'x - \theta)\, d\sigma(x) = 0 \qquad \text{for all } a \in R^k,\ \theta \in R.$$

In Cybenko (1989), it is shown that if ψ is sigmoidal, then ψ is discriminatory. (The proof can trivially be generalized to the case where ψ has distinct and finite limits at $\pm\infty$.) However, the following much stronger result is true, which, upon combination with the above arguments, establishes Theorem 1 and 2.

Theorem 5: Whenever ψ is bounded and nonconstant, it is discriminatory.

Proof: Throughout the proof, certain techniques and results from Fourier analysis will be used. As a reference we recommend the excellent book by Rudin (1967).

Suppose that ψ is bounded and nonconstant and that σ is a finite signed measure on R^k such that $\int_{R^k} \psi(a'x - \theta)\, d\sigma(x) = 0$ for all $a \in R^k$ and $\theta \in R$. Fix $u \in R^k$ and let σ_u be the finite signed measure on R induced by the transformation $x \mapsto u'x$, that is, for all Borel sets of R we have

$$\sigma_u(B) = \sigma\{x \in R^k : u'x \in B\}.$$

Then at least for all bounded functions χ on R,

$$\int_{R^k} \chi(u'x)\, d\sigma(x) = \int_R \chi(t)\, d\sigma_u(t).$$

Hence by assumption,

$$\int_{R^k} \psi(\lambda u'x - \theta)\, d\sigma(x) = \int_R \psi(\lambda t - \theta)\, d\sigma_u(t) = 0$$

for all λ, $\theta \in R$.

To simplify notations, let us write $L = L^1(R)$ for the space of integrable functions on R (with respect to Lebesgue measure) and $M = M(R)$ for the space of finite signed measures on R. For $f \in L$, $\|f\|_L$ denotes the usual L^1 norm and $\hat{f}$ the Fourier transform. Similarly, for $\tau \in M$, $\|\tau\|_M$ denotes the total variation of τ on R and $\hat{\tau}$ the Fourier transform.

By choosing θ such that $\psi(-\theta) \neq 0$ and setting λ to zero, we find that in particular $\int_R d\sigma_u(t) = \hat{\sigma}_u(0) = 0$. For $u = 0$, σ_0 is concentrated at $t = 0$ and $\sigma_0\{0\} = \hat{\sigma}_0 = 0$, hence $\sigma_0 = 0$. Now suppose $u \neq 0$. Pick a function $w \in L$ whose Fourier transform has no zero (e.g., take $w(t) = \exp(-t^2)$). Consider the integral

$$\int_R \int_R \psi(\lambda(s + t) - \theta)\, w(s)\, ds\, d\sigma_u(t).$$

As

$$\int_R \int_R |\psi(\lambda(s + t) - \theta)| |w(s)|\, ds\, d|\sigma_u|(t) \le \|w\|_L \|\sigma_u\|_M \sup_{t \in R} |\psi(t)| < \infty,$$

we may apply Fubini's theorem to obtain

$$\begin{aligned} 0 &= \int_R \left[\int_R \psi(\lambda t - (\theta - \lambda s))\, d\sigma_u(t) \right] w(s)\, ds \\ &= \int_R \int_R \psi(\lambda(s + t) - \theta)\, w(s)\, ds\, d\sigma_u(t) \\ &= \int_R \psi(\lambda t - \theta)\, d(w * \sigma_u)(t), \end{aligned}$$

where $w * \sigma_u$ denotes the convolution of w and σ_u. By Theorem 1.3.5 in Rudin (1967), L is a closed ideal in M, hence in particular $w * \sigma_u$ is absolutely continuous with respect to Lebesgue measure. Let $h \in L$ be the corresponding Radon–Nikodym derivative. Then $\hat{h} = \hat{w}\hat{\sigma}_u$, hence in particular $\hat{h}(0) = 0$.

The above equation is then equivalent to $\int_R \psi(\lambda t - \theta)\, h(t)\, dt = 0$. Let $\alpha \neq 0$ and $\gamma \in R$. By first replacing λ by $1/\alpha$ and θ by $-\gamma/\alpha$ and then performing the change of variables $t \mapsto \alpha t - \gamma$, we obtain that for all $\gamma \in R$ and for all nonzero real α,

$$\int_R \psi(t) h(\alpha t - \gamma)\, dt = 0.$$

Let us write $M_\alpha h(t)$ for $h(\alpha t)$. The above equation implies that $\int_R \psi(t) f(t)\, dt$ vanishes for all f contained in the *closed translation invariant subspace I* spanned by the family $M_\alpha h$, $\alpha \neq 0$. By Theorem 7.1.2 in Rudin (1967), I is an ideal in L.

Following the notation in Rudin (1967), let us write $Z(f)$ for the set of all $\omega \in R$ where the Fourier transform $\hat{f}(\omega)$ of $f \in L$ vanishes, and if I is an ideal, define $Z(I)$, the *zero set* of I, as the set of ω where the Fourier transforms of *all* functions in I vanish.

Suppose that h is nonzero. As $M_\alpha h(\omega) = \hat{h}(\omega/\alpha)/\alpha$, we find that $Z(I) = \{0\}$ and in fact, I is precisely the set of all integrable functions f with $\int_R f(t)\, dt = \hat{f}(0) = 0$. To see this, let us first note that for all functions $f \in I$, we trivially have $\{0\} = Z(I) \subseteq Z(f)$. Conversely, suppose that f has zero integral. As the intersection of the boundaries of $Z(I)$ and $Z(f)$ (again trivially) equals $\{0\}$ and hence contains no perfect set, Theorem 7.2.4 in Rudin (1967) implies that $f \in I$.

Hence, if h is nonzero, the integral $\int_R \psi(t) f(t)\, dt$ vanishes for all integrable functions which have zero integral. It is easily seen that this implies that ψ is constant which was ruled out by assumption. Hence $h = 0$ and thus $\hat{h} = \hat{w}\hat{\sigma}_u$ is identically zero, which in turn yields that $\hat{\sigma}_u$ vanishes identically, because $\hat{w}$ has no zeros. By the uniqueness Theorem 1.3.7(b) in Rudin (1967), $\sigma_u = 0$.

Summing up, we find that $\sigma_u = 0$ for all $u \in R^k$. To complete the proof, let $\hat{\sigma}(u) = \int_{R^k} \exp(iu'x)\, d\sigma(x)$ be the Fourier transform of σ at u. Then

$$\begin{aligned} \hat{\sigma}(u) &= \int_{R^k} \exp(iu'x)\, d\sigma(x) \\ &= \int_R \exp(it)\, d\sigma_u(t) \\ &= 0, \end{aligned}$$

that is, $\hat{\sigma} = 0$. Again invoking the uniqueness Theorem 1.3.7(b) in Rudin (1967), $\sigma = 0$ and the proof of Theorem 5 is complete.

The proofs of the remaining theorems require some additional preparation. For functions f defined on R^k, let $\|f\|_u := \sup_{R^k} |f(x)|$. Let w be the familiar function in $C^\infty(R^k)$ with support in the unit sphere given by

$$w(x) = \begin{cases} c \exp(-1/(1 - |x|^2)), & \text{if } |x| < 1, \\ 0, & \text{if } |x| \geq 1, \end{cases}$$

where $|x|$ is the euclidean length of x and c is a constant chosen in a way that $\int_{R^k} w(x)\, dx = 1$. For $\varepsilon > 0$, let us write $w_\varepsilon(x) = \varepsilon^{-k} w(x/\varepsilon)$.

If f is a locally integrable function on R^k, let $J_\varepsilon f$ be the convolution $w_\varepsilon * f$. The following facts are well known (Adams, 1975, pp. 29ff.).

- $J_\varepsilon f \in C^\infty(R^k)$ with derivatives $D^\alpha J_\varepsilon f = D^\alpha w_\varepsilon * f$.
- $\|J_\varepsilon f\|_u \leq \|f\|_u$. Thus, if f is bounded, then $J_\varepsilon f(x)$ is uniformly bounded in x and ε.
- If f is continuous, then $J_\varepsilon f \to f$ uniformly on compacta as $\varepsilon \to 0$.

Similarly, if σ is a locally finite signed measure on R^k, let $J_\varepsilon \sigma$ be the convolution $w_\varepsilon * \sigma$, that is,

$$J_\varepsilon \sigma(x) = \int_{R^k} w_\varepsilon(x - y)\, d\sigma(y).$$

Then again, $J_\varepsilon \sigma \in C^\infty(R^k)$. If σ has compact support, $J_\varepsilon \sigma$ has compact support.

Finally, the following result can easily be established. (The first assertion is a straightforward application of Fubini's theorem using the symmetry of w_ε, and the second one follows by Lebesgue's bounded convergence theorem.)

Lemma. Suppose that f and σ satisfy one of the two following conditions: (a) f is continuous and σ is a finite signed measure with compact support; (b) f is bounded and continuous and σ is a finite signed measure. Then, if T_y denotes translation by y, that is, $T_y f(x) = f(x + y)$,

$$\int_{R^k} f\, J_\varepsilon \sigma\, x = \int_{R^k} \left[\int_{R^k} T_y f\, d\sigma \right] w_\varepsilon(y)\, dy = \int_{R^k} J_\varepsilon f\, d\sigma,$$

and

$$\lim_{\varepsilon \to 0} \int_{R^k} f\, J_\varepsilon \sigma\, dx = \int_{R^k} f\, d\sigma.$$

Proof of Theorem 3: If $\mathfrak{N}_k(\psi)$ is not uniformly m dense on compacta in $C^m(R^k)$, then by the usual dual space argument there exists a collection σ_α, $|\alpha| \leq m$ of finite signed measures with support in some compact subset X of R^k such that the functional

$$\Lambda(f) = \sum_{|\alpha| \leq m} \int_{R^k} D^\alpha f\, d\sigma_\alpha$$

vanishes on $\mathfrak{N}_k(\psi)$, but not identically on $C^m(R^k)$.

For $\varepsilon > 0$, define functionals Λ_ε by

$$\Lambda_\varepsilon(f) := \sum_{|\alpha| \leq m} \int_{R^k} D^\alpha f\, J_\varepsilon \sigma_\alpha\, dx.$$

(All integrals exist because all $J_\varepsilon \sigma_\alpha$ have compact support.) By part (a) of the above lemma, we conclude that

$$\Lambda_\varepsilon(f) = \int_{R^k} \left[\sum_{|\alpha| \le m} \int_{R^k} D^\alpha T_y f \, d\sigma_\alpha \right] w_\varepsilon(y) \, dy$$

$$= \int_{R^k} \Lambda(T_y f) \, w_\varepsilon(y) \, dy$$

and that

$$\lim_{\varepsilon \to 0} \Lambda_\varepsilon(f) = \Lambda(f)$$

for all $f \in C^m(R^k)$. Finally, integration by parts yields that

$$\Lambda_\varepsilon(f) = \int_{R^k} \underbrace{\left[\sum_{|\alpha| \le m} (-1)^{|\alpha|} D^\alpha J_\varepsilon \sigma_\alpha \right]}_{:= h_\varepsilon} f \, dx.$$

Let us write $\psi_{a,\theta}(x) = \psi(a'x - \theta)$. Suppose that Λ vanishes on $\mathfrak{N}_k(\psi)$. As $\psi_{a,\theta} \in \mathfrak{N}_k(\psi)$ for all $a \in R^k$ and $\theta \in R$, we infer that $\Lambda(\psi_{a,\theta}) = 0$. Observing that $T_y \psi_{a,\theta} = \psi_{a,\theta - a'y}$, we see that $\Lambda(T_y \psi_{a,\theta}) = 0$ for all $a, y \in R^k$, $\theta \in R$. It follows that

$$\int_{R^k} \psi_{a,\theta} \, h_\varepsilon \, dx = \Lambda_\varepsilon(\psi_{a,\theta}) = \int_{R^k} \Lambda(T_y \psi_{a,\theta}) \, w_\varepsilon(y) \, dy = 0$$

for all $a \in R^k$ and $\theta \in R$. As, by assumption, ψ is bounded and nonconstant, Theorem 5 implies that $h_\varepsilon \equiv 0$. Hence $\Lambda_\varepsilon(f) = \int_{R^k} f \, h_\varepsilon \, dx$ vanishes for all functions $f \in C^m(R^k)$ which in turn yields that

$$\Lambda(f) = \lim_{\varepsilon \to 0} \Lambda_\varepsilon(f) = 0$$

for all $f \in C^m(R^k)$, which was ruled out by assumption. We conclude that, under the conditions of Theorem 3, $\mathfrak{N}_k(\psi)$ is uniformly m dense on compacta in $C^m(R^k)$, establishing the first half of Theorem 3.

The second half of Theorem 3 now follows easily. We have to show that for all $f \in C^m(R^k)$ and $\varepsilon > 0$, there is a function $g \in \mathfrak{N}_k(\psi)$ such that $\|f - g\|_{m,p,\mu} < \varepsilon$. Let X be a compact set containing the support of μ. We find that

$$\|f - g\|_{m,p,\mu} \le \gamma \, \|f - g\|_{m,u,X},$$

where $\gamma^p = \mu(R^k) \, \# \{\alpha : |\alpha| \le m\}$. Hence, if we take $g \in \mathfrak{N}_k(\psi)$ such that $\|f - g\|_{m,u,X} < \varepsilon/\gamma$, which is possible by the first half of Theorem 3 that we just established, we find that $\|f - g\|_{m,p,\mu} < \varepsilon$ and the proof of Theorem 3 is complete.

Proof of Theorem 4: The proof of Theorem 4 parallels the one of Theorem 3. Let us write $C^{m,u}(R^k)$ for the space of all functions $f \in C^m(R^k)$ which, along with their derivatives up to order m, are bounded, that is,

$$C^{m,\mu}(R^k) = \{f \in C^m(R^k) : \|D^\alpha f\|_u < \infty, \, |\alpha| \le m\}.$$

It is easily seen that $C^{m,u}(R^k)$ is a dense subset of $C^{m,p}(\mu)$. By assumption, $\psi \in C^{m,u}(R)$, hence $\mathfrak{N}_k(\psi) \subset C^{m,u}(R^k) \subset C^{m,p}(\mu)$.

If $\mathfrak{N}_k(\psi)$ is not dense in $C^{m,p}(\mu)$, the usual dual space argument yields the existence of a suitable collection of functions $g_\alpha \in L^q(\mu)$, $|\alpha| \le m$, where q is the conjugate exponent $p/(p - 1)$, such that the functional

$$\Lambda(f) = \sum_{|\alpha| \le m} \int_{R^k} D^\alpha f \, g_\alpha \, d\mu$$

vanishes on $\mathfrak{N}_k(\psi)$, but not identically on $C^{m,u}(R^k)$. Now proceed as in the proof of Theorem 3 with the finite signed measures σ_α given by $d\sigma_\alpha = g_\alpha \, d\mu$, $C^{m,u}(R^k)$ replacing $C^m(R^k)$, and using part (b) of the lemma.

REFERENCES

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

Carroll, S. M., & Dickinson, B. W. (1989). Construction of neural nets using the Radon transform. In *Proceedings of the International Joint Conference on Neural Networks* (pp. I:607–611). San Diego: SOS Printing.

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

Friedman, A. (1982). *Foundation of modern analysis*. New York: Dover Publications.

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

Gallant, A. R., & White, H. (1988). There exists a neural network that does not make avoidable mistakes. In *IEEE Second International Conference on Neural Networks* (pp. I:657–664). San Diego: SOS Printing.

Gallant, A. R., & White, H. (1989). On learning the derivatives of an unknown mapping with multilayer feedforward networks. Preprint.

Hecht-Nielsen, R. (1989). Theory of the back propagation neural network. In *Proceedings of the International Joint Conference on Neural Networks* (pp. I:593–606). San Diego: SOS Printing.

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

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

Irie, B., & Miyake, S. (1988). Capabilities of three layer perceptrons. In *IEEE Second International Conference on Neural Networks* (pp. I:641–648). San Diego: SOS Printing.

Lapedes, A., & Farber, R. (1988). *How neural networks work*. Technical Report LA-UR-88-418. Los Alamos, NM: Los Alamos National Laboratory.

Maz'ja, V. G. (1985). *Sobolev spaces*. New York: Springer Verlag.

Rudin, W. (1967). *Fourier analysis on groups. Interscience Tracts in Pure and Applied Mathematics*, Vol. 12. New York: Interscience Publishers.

Stinchcombe, M., & White, H. (1989). Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions. In *Proceedings of the International Joint Conference on Neural Networks* (pp. I:613–618). San Diego: SOS Printing.

Stinchcombe, M., & White, H. (1990). Approximating and learning unknown mappings using multilayer feedforward networks with bounded weights. Preprint. San Diego: Department of Economics, University of California.