Neural Networks, Vol. 3, pp. 551–560, 1990
Printed in the USA.

0893-6080/90 $3.00 + .00

ORIGINAL CONTRIBUTION

Universal Approximation of an Unknown Mapping and Its Derivatives Using Multilayer Feedforward Networks

KURT HORNIK, MAXWELL STINCHCOMBE, AND HALBERT WHITE

University of California, San Diego

(*Received* 11 *August 1989; revised and accepted* 31 *January* 1990)

Abstract—*We give conditions ensuring that multilayer feedforward networks with as few as a single hidden layer and an appropriately smooth hidden layer activation function are capable of arbitrarily accurate approximation to an arbitrary function and its derivatives. In fact, these networks can approximate functions that are not differentiable in the classical sense, but possess only a generalized derivative, as is the case for certain piecewise differentiable functions. The conditions imposed on the hidden layer activation function are relatively mild; the conditions imposed on the domain of the function to be approximated have practical implications. Our approximation results provide a previously missing theoretical justification for the use of multilayer feedforward networks in applications requiring simultaneous approximation of a function and its derivatives.*

Keywords—Approximation, Derivatives, Sobolev space, Feedforward networks.

1. INTRODUCTION

The capability of sufficiently complex multilayer feedforward networks to approximate an unknown mapping f: $\mathbb{R}^r \to \mathbb{R}$ arbitrarily well has been recently investigated by Cybenko (1989), Funahashi (1989), Hecht-Nielsen (1989), Hornik, Stinchcombe, and White (1989) (HSW) (all for sigmoid hidden layer activation functions) and Stinchcombe and White (1989) (SW) (non-sigmoid hidden layer activation functions). In applications, it may be desirable to approximate not only the unknown mapping, but also its unknown derivatives. This is the case in Jordan's (1989) recent investigation of robot learning of smooth movement. Jordan states:

> The Jacobian matrix $\partial z/\partial x$ . . . is the matrix that relates small changes in the controller output to small changes in the task space results and cannot be assumed to be available a priori, or provided by the environment. However, all of the derivatives in the matrix are *forward* derivatives. They are easily obtained by differentiation if a forward model is available. The forward model itself must be learned, but this can be achieved directly by system identification. Once the model is accurate over a particular domain, its derivatives provide a learning operator that allows the system to convert errors in task space into errors in articulatory space and thereby change the controller.

Thus, learning an adequate approximation to the Jacobian matrix of an unknown mapping is a key component of Jordan's approach to robot learning of smooth movement.

Despite the success of Jordan's experiments, there is no existing theoretical guarantee that multilayer feedforward networks generally have the capability to approximate an unknown mapping and its derivatives simultaneously. For example, a network with hard limiting hidden layer activations approximates unknown mappings with a piecewise-constant function, the first derivatives of which exist and are zero almost everywhere. Obviously, the derivatives of such a network output function cannot approximate the derivatives of an arbitrary function.

Intuition suggests that networks having smooth hidden layer activation functions ought to have output function derivatives that will approximate the derivatives of an unknown mapping. However, the justification for this intuition is not obvious. Consider the class of single hidden layer feedforward networks having network output functions belonging to the set

$$\Sigma(G) \equiv \{g : \mathbb{R}^r \to \mathbb{R} \mid g(x) = \sum_{j=1}^{q} \beta_j G(x'\gamma_j);$$
$$x \in \mathbb{R}^r, \beta_j \in \mathbb{R}, \gamma_j \in \mathbb{R}^{r+1}, j = 1, \ldots, q, q \in \mathbb{N}\},$$

Acknowledgements: We are indebted to Angelo Melino for pressing us on the issue addressed here and to the referees for numerous helpful suggestions. White's participation was supported by NSF Grant SES-8806990.

Requests for reprints should be sent to Halbert White, Department of Economics, D-008, University of California, San Diego, La Jolla, CA 92093.

where x represents an r vector of network inputs ($r \in \mathbb{N} \equiv \{1, 2, \ldots\}$), $\bar{x} \equiv (1, x^T)^T$ (the superscript T denotes transposition), β_j represents hidden to output layer weights and γ_j represents input to hidden layer weights, $j = 1, \ldots, q$, where q is the number of hidden units, and G is a given hidden unit activation function. The first partial derivatives of the network output function are given by

$$\partial g(x)/\partial x_i = \sum_{j=1}^{q} \beta_j \gamma_{ji} DG(\bar{x}^T \gamma_j), \qquad i = 1, \ldots, r,$$

where x_i is the ith component of x, γ_{ji} is the ith component of γ_j, $i = 1, \ldots, r$ (γ_{j0} is the input layer bias to hidden unit j), and DG denotes the first derivative of G. Available results ensure that there exist choices for β_j and γ_j, $j = 1, \ldots, q$ for which $\partial g/\partial x_i$ can well approximate $\partial f/\partial x_i$, the derivative of the unknown mapping. (Note that if G is sigmoid, then DG is non-sigmoid, so that the results of *SW* are relevant.) The problem is that these choices for β_j and γ_j are not necessarily the choices for which g adequately approximates f or for which $\partial g/\partial x_h$ approximates $\partial f/\partial x_h$ for $h \neq i$. Nor is it obvious that a single set of weights exists that simultaneously ensures an adequate approximation to f and its derivatives.

Our purpose here is to establish rigorously that such a set of weights does indeed exist, and that multilayer feedforward networks with as few as a single hidden layer and fairly arbitrary hidden layer activation functions are in fact capable of arbitrarily accurate approximation to an unknown mapping and its derivatives, to as many orders as desired.

This fact not only justifies corresponding aspects of Jordan's (1989) approach to network learning of smooth movements, but generally supports use of multilayer feedforward networks in any application requiring approximation of an unknown mapping and its derivatives. For example, a net appropriately trained to approximate the transfer function of a (perfectly measured) deterministic chaos (e.g., as in Lapedes & Farber, 1987) could be used to obtain information on the Lyapounov exponents of the underlying chaos. (The Lyapounov exponents are defined in terms of the first derivatives of the transfer function.)

Another potential application area is economics, where theoretical considerations lead to hypotheses about the derivative properties (e.g., "elasticities," $\partial \ln f/\partial \ln x_i = (\partial f/\partial x_i)(x_i/f)$), of certain functions arising in the theory of the firm and of the consumer (production functions, cost functions, utility functions and expenditure functions). (See, e.g., Varian, 1978.) Approximation of these functions and their derivatives can aid in confirmation or refutation of particular theories of the firm or the consumer. Such analyses have been conducted by Elbadawi, Gallant, and Souza (1983) using Fourier series. An approach based on kernel regression is described by Vinod and Ullah (1985) (see also Ullah, 1988). Our results establish neural network models as providing an alternative framework for studying the theory of the firm and of the consumer.

Approximation of derivatives also permits sensitivity analyses in which the relative effects on output of small changes in input variables in different regions of input space can be investigated. Gilstrap and Dominy (1989) have proposed such analyses as the basis on which network knowledge can be explicated.

Finally, we note that any network suitably trained to approximate a mapping satisfying some nonlinear partial differential equations (pde) will have an output function that itself approximately satisfies the pde by virtue of its approximation of the mapping's derivatives.

Formally, our results are obtained by showing that for broad classes of multilayer feedforward networks, the set $\Sigma(G)$ is dense in general spaces of functions where distance between functions is measured taking into account differences between the derivatives of the functions (including derivatives of order zero).

Because the mathematical background regarding these spaces may be somewhat unfamiliar, we provide a synopsis of the relevant material in section 2. Section 3 contains our main results. Section 4 provides a brief discussion on implementation of a feedforward net that yields the desired derivatives as outputs, together with some brief remarks concerning learning of the representations shown here to be possible. Section 5 provides a summary and some concluding remarks. Mathematical proofs are gathered into the Mathematical Appendix.

2. BACKGROUND ON FUNCTION SPACES

This section reviews relevant basic concepts for the theoretical results of the following section. For additional detail, see for example, Adams (1975) and Showalter (1977).

We are concerned here with how well the collection of network output functions $\Sigma(G)$ can approximate certain spaces of functions. Given a function space, say S, we can measure the distance between two elements of S using a metric ρ. Formally, ρ is a mapping with the properties: (1) for all $f, g \in S$, $\rho(f, g) \geq 0$; (2) for $f, g, h \in S$, $\rho(f, h) \leq \rho(f, g) + \rho(g, h)$; (3) $\rho(f, g) = 0$ if and only if $f = g$. The pair (S, ρ) is called a metric space. To describe the ability of the set $\Sigma(G)$ to approximate the space S, the concept of ρ-denseness applies.

DEFINITION 2.1: *Let U be a subset of $\mathbb{R}^r$, let S be a collection of functions $f : U \to \mathbb{R}$ and let ρ be a metric on S. For any g in $\Sigma(G)$ (recall $g : \mathbb{R}^r \to \mathbb{R}$) define*

the restriction of g to U, $g_{|U}$, as $g_{|U}(x) = g(x)$ for x in U, $g_{|U}(x)$ unspecified for x not in U.

Suppose that for any f in S and $\varepsilon > 0$ there exists g in $\Sigma(G)$ such that $\rho(f, g_{|U}) < \varepsilon$. Then we say that $\Sigma(G)$ contains a subset ρ-dense in S. If in addition $g_{|U}$ belongs to S for every g in $\Sigma(G)$, we say that $\Sigma(G)$ is ρ-dense in S. □

The first part of this definition allows for the possibility that $\Sigma(G)$ may contain functions g for which $g_{|U}$ does not belong to S. Even so, when $\Sigma(G)$ has this denseness property, it always contains a single hidden layer feedforward network output function capable of arbitrarily accurate approximation to any member of S in terms of the metric ρ.

We shall consider approximating elements of a variety of metric spaces (S, ρ) using feedforward networks. For all of what follows, we let U be an open subset of $\mathbb{R}^r$. (We could have $U = \mathbb{R}^r$.) To specify the first function space of interest of interest, let $C(U)$ be the set of all functions continuous on U. Let α be an r-tuple $\alpha = (\alpha_1, \ldots, \alpha_r)^t$ of nonnegative integers (a "multi-index"). If x belongs to $\mathbb{R}^r$, let $x^\alpha \equiv x_1^{\alpha_1} \cdot x_2^{\alpha_2} \cdot \ldots \cdot x_r^{\alpha_r}$. Denote by D^α the partial derivative

$$\partial^{|\alpha|}/\partial x^\alpha = \partial^{\alpha_1 + \cdots + \alpha_r}/(\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \ldots \partial x_r^{\alpha_r})$$

of order $|\alpha| \equiv \alpha_1 + \alpha_2 + \cdots + \alpha_r$. For nonnegative integers m, we define $C^m(U) \equiv \{f \in C(U) | D^\alpha f \in C(U)$ for all α, $|\alpha| \leq m\}$ and $C^\infty(U) \equiv \cap_{m=1}^\infty C^m(U)$. We let D^0 be the identity, so that $C^0(U) \equiv C(U)$. Thus, the functions in $C^m(U)$ have continuous derivatives up to order m on U, while the functions in $C^\infty(U)$ have continuous derivatives on U of every order.

For these spaces, we adopt the following terminology.

DEFINITION 2.2: *Let $m, l \in \{0\} \cup \mathbb{N}$, $0 \leq m \leq l$, and $U \subset \mathbb{R}^r$ be given, and let $S \subset C^l(U)$. Suppose that for any f in S, compact $K \subset U$ and $\varepsilon > 0$ there exists g in $\Sigma(G)$ such that $\max_{|\alpha| \leq m} \sup_{x \in K} |D^\alpha f(x) - D^\alpha g(x)| < \varepsilon$. Then we say that $\Sigma(G)$ is m-uniformly dense on compacta in S.* □

There are many metrics inducing m-uniform convergence on compacta. For example, see Dugundji (1966, p. 172). We denote any such metric ρ_K^m. This metric depends on U, but we suppress this for notational convenience.

When $\Sigma(G)$ is m-uniformly dense on compacta in S, then no matter how we choose an f in S, a compact subset K of U, or the accuracy of approximation $\varepsilon > 0$, we can always find a single hidden layer feedforward network having output function g (in $\Sigma(G)$) with all derivatives of $g_{|U}$ on K up to order m lying within ε of those of f on K. This is a strong and very desirable approximation property. In the next section, we impose conditions on G and U that ensure that $\Sigma(G)$ is indeed m-uniformly dense on compacta in particular useful subsets S of $C^m(U)$. Thus, such networks can be used to approximate any unknown mapping and its derivatives to any desired degree of accuracy in this precise sense.

Another useful approach to measuring distances between functions taking into account differences between derivatives is based on metrics defined for collections of functions known as Sobolev spaces. To define these spaces, we must first introduce the spaces $L_p(U, \mu)$.

For any open subset U of $\mathbb{R}^r$, let $\mathfrak{B}(U)$ be the Borel σ-field generated by the open subsets of U (i.e., $\mathfrak{B}(U)$ is the smallest collection of subsets of U that contains U and all open subsets of U and is closed under complements and countable unions).

A function $f : U \to \mathbb{R}$ is said to be (*Borel*) *measurable* if for every open subset A of $\mathbb{R}$, the set $\{x \in U | f(x) \in A\}$ belongs to $\mathfrak{B}(U)$. Continuous functions on U are measurable, as are piecewise continuous functions. Nonmeasurable functions exist (see, e.g., Halmos, 1974, p. 69), but they are pathological, and generally not relevant in applications.

A measure μ assigns a number in $[0, \infty]$ to every set in $\mathfrak{B}(U)$, with $\mu(\varnothing) = 0$ and $\mu(B) = \Sigma_{i=1}^\infty \mu(B_i)$ whenever $B = \cap_{i=1}^\infty B_i$, $B_i \cup B_j = \varnothing$ for all $i \neq j$. When $\mu(U) < \infty$, μ is called a finite measure. An important measure is the Lebesgue measure λ on $(\mathbb{R}^r, \mathfrak{B}(\mathbb{R}^r))$. When $r = 1$, λ measures the length of intervals $B = (a, b)$ as $\lambda(B) = b - a$. For $r = 2$, λ measures the area of rectangles $B = (a, b) \times (c, d)$ as $\lambda(B) = (b - a)(d - c)$. When $r = 3$, λ measures volumes in a similar manner. Generally (i.e., for all r), λ provides a measure of the generalized volume of a set.

The space $L_p(U, \mu)$ is the collection of all measurable functions f such that $\|f\|_{p,U,\mu} \equiv [\int_U |f|^p d\mu]^{1/p} < \infty$, $1 \leq p < \infty$, where the integral is defined in the sense of Lebesgue. When $\mu = \lambda$ we may write either $\int_U f d\lambda$ or $\int_U f(x) dx$ to denote the same integral. We measure the distance between two functions f and g belonging to $L_p(U, \mu)$ in terms of the metric $\rho_{p,U,\mu}(f, g) \equiv \|f - g\|_{p,U,\mu}$. Two functions that differ only on sets of μ-measure zero have $\rho_{p,U,\mu}(f, g) = 0$. We shall not distinguish between such functions. Thus, $f \in L_p(U, \mu)$ represents an equivalence class of functions, all of which differ from each other only on sets of μ-measure zero. Functions in $L_p(U, \mu)$ need not have derivatives, and the distance measure $\rho_{p,U,\mu}$ takes no account of relationships between any derivatives that do exist.

The first Sobolev space we consider is denoted $S_p^m(U, \mu)$, defined as the collection of all functions f in $C^m(U)$ such that $\|D^\alpha f\|_{p,U,\mu} < \infty$ for all $|\alpha| \leq m$. We define the Sobolev norm $\|f\|_{m,p,U,\mu} \in (\Sigma_{|\alpha| \leq m}$

$D^\alpha f\|_{p,U,\mu}^p)^{1/p}$. The Sobolev metric is

$$\rho_{p,\mu}^m(f, g) \equiv \|f - g\|_{m,p,U,\mu} \qquad f, g \in S_p^m(U, \mu).$$

Note that $\rho_{p,\mu}^m$ depends implicitly on U, but we suppress this dependence for notational convenience. The Sobolev metric explicitly takes into account distances between derivatives. Two functions in $S_p^m(U, \mu)$ are close in the Sobolev metric $\rho_{p,\mu}^m$ when all derivatives of order $0 \leq |\alpha| \leq m$ are close in L_p metric.

For many interesting choices for G, $\Sigma(g)$ need not be a subset of $S_p^m(U, \mu)$. However, we shall generally be able to find H in $\Sigma(G)$ such that for every h in $\Sigma(H)$, $h_{|U}$ belongs to $S_p^m(U, \mu)$ and $\Sigma(H)$ is $\rho_{p,\mu}^m$-dense in $S_p^m(U, \mu)$.

In the next section, we give conditions on G, U, and μ ensuring that single hidden layer feedforward networks can be used to approximate any unknown mapping and its derivatives to any desired degree of accuracy in the metric $\rho_{p,\mu}^m$. In particular, we take $U = \mathbb{R}^r$ and assume that μ is finite and compactly supported, that is, there is a compact subset K of $\mathbb{R}^r$ such that $\mu(K) = \mu(\mathbb{R}^r)$.

Next we consider the Sobolev space $S_p^m(\text{loc})$ defined as the collection of all functions f in $C^m(\mathbb{R}^r)$ such that for every open bounded subset U of $\mathbb{R}^r$ the function f belongs to $S_p^m(U, \lambda)$. To define a metric on this space of functions, let $U_n \equiv \{x \in \mathbb{R}^r : |x_i| < n, i = 1, \ldots, r\}$ and put

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

$$f, g \in S_p^m(\text{loc}).$$

Two functions in $S_p^m(\text{loc})$ are close in the metric $\rho_{p,\text{loc}}^m$ if their derivatives of orders $0 \leq |\alpha| \leq m$ are close in L_p metric on open bounded subsets of $\mathbb{R}^r$. We give conditions on G ensuring that single hidden layer feedforward netrorks can be used to approximate any unknown mapping and its derivatives to any desired degree of accuracy in the metric $\rho_{p,\text{loc}}^m$.

The spaces $S_p^m(U, \mu)$ are limited by the fact that they do not include functions that have derivatives everywhere except on sets of measure zero (e.g., piecewise differentiable functions). Interestingly, it turns out to be possible to approximate such functions arbitrarily well using multilayer feedforward networks. However, in order to discuss this possibility precisely, it is necessary to work with a genralized notion of the derivative.

In order to provide the proper generalization, we introduce the concepts of *distributions* and *distributional derivatives* due to Schwartz (1950). For all functions f in a broad class ($L_{1,\text{loc}}(U)$ specified below), we can associate a specific distribution, differentiable of all orders. When the function f is differentiable, the distributional derivatives correspond to the classical derivatives. However, even when the function f is not differentiable in the classical sense there is often a function in the original space corresponding to the distributional derivative. This function is called a "weak" or "generalized" derivative, and provides the generalization of the classical derivative needed for our discussion of the approximation capabilities of multilayer feedforward networks.

The formal definitions of a distribution and its derivatives make use of functions belonging to $C_0^\infty(U) \equiv C^\infty(U) \cap C_0(U)$, where $C_0(U)$ is the space of all functions in $C(U)$ with compact support. (The support of $f \in C(U)$ is defined as $\text{supp } f = \text{cl}\{x \in U : f(x) \neq 0\}$, where cl denotes the closure of the indicated set.) Functions in $C_0^\infty(U)$ have continuous derivatives of all orders and compact support.

A *distribution* on U over $\mathbb{R}$ is defined as a linear mapping $T : C_0^\infty(U) \to \mathbb{R}$ (i.e., $T(a\phi_1 + b\phi_2) = aT(\phi_1) + bT(\phi_2)$, $a, b \in \mathbb{R}$, $\phi_1, \phi_2 \in C_0^\infty(U)$). We construct the distributions used here in a straightforward manner. Let K be a compact subset of U. Let $L_1(K, \lambda)$ be the set of all measurable functions $f : U \to \mathbb{R}$ such that $\int_K |f| \, d\lambda < \infty$. The space of locally integrable functions on U is $L_{1,\text{loc}}(U) \equiv \cap \{L_1(K, \lambda) | K \subset U, K \text{ compact}\}$. For every f in $L_{1,\text{loc}}(U)$ we define the distribution T_f such that

$$T_f(\phi) \equiv \int_U f\phi \, d\lambda, \qquad \phi \in C_0^\infty(U).$$

This is readily verified to be a linear mapping from $C_0^\infty(U)$ to $\mathbb{R}$.

Further, for any distribution T we may define the distributional derivative $\partial^\alpha T$ such that

$$\partial^\alpha T(\phi) \equiv (-1)^{|\alpha|} T(D^\alpha \phi), \qquad \phi \in C_0^\infty(U).$$

Consequently, $\partial^\alpha T$ is also a linear mapping from $C_0^\infty(U)$ to $\mathbb{R}$. This definition is constructed so that when f belongs to $C^m(U)$, then $\partial^\alpha T_f = T_{D^\alpha f}$ for $|\alpha| \leq m$. In this case the distributional derivative corresponds precisely to the classical derivative. To see this, note that

$$\begin{aligned} \partial^\alpha T_f(\phi) &\equiv (-1)^{|\alpha|} T_f(D^\alpha \phi) \\ &\equiv (-1)^{|\alpha|} \int_U f(D^\alpha \phi) \, d\lambda \\ &= \int_U (D^\alpha f)\phi \, d\lambda \\ &\equiv T_{D^\alpha f}(\phi), \qquad \phi \in C_0^\infty(U). \end{aligned}$$

The key step is the equality, which follows from integration by parts and the fact that ϕ vanishes at the boundary of U because it has compact support.

Even when the classical derivative does not exist, there may exist an element h of $L_{1,\text{loc}}(U)$ such that $\partial^\alpha T_f = T_h$. In such cases, we write $h = \partial^\alpha f$ and call $\partial^\alpha f$ the weak or generalized derivative of f. (When $f \in C^m(U)$, $\partial^\alpha f = D^\alpha f$.) Showalter (1977, pp. 30–31) gives numerous examples of functions in $L_{1,\text{loc}}(U)$

having weak derivatives, but not classical derivatives. However, not all functions in $L_{1,\text{loc}}(U)$ have weak derivatives. Such functions will play no role in what follows.

We now have sufficient background to define the Sobolev spaces

$$W_p^m(U) \equiv \{f \in L_{1,\text{loc}}(U) | \partial^\alpha f \in L_p(U, \lambda), 0 \le |\alpha| \le m\}.$$

This is the collection of all functions having generalized derivatives belonging to $L_p(U, \lambda)$ of order up to m. Consequently, $W_p^m(U)$ includes $S_p^m(U, \lambda)$, as well as functions that do not have derivatives in the classical sense, such as piecewise differentiable functions.

Although it would be possible to define "weighted" Sobolev spaces $W_p^m(U, \mu)$ containing $S_p^m(U, \mu)$ for $\mu \neq \lambda$ in an obvious way, we leave formal consideration of these spaces aside in order to avoid certain unpleasant technicalities. (See Kufner, 1980, and Kufner and Sandig, 1987.)

The norm on $W_p^m(U)$ generalizes that on $S_p^m(U, \lambda)$; we write it as

$$\|f\|_{m,p,U} \equiv \left(\sum_{|\alpha| \le m} \|\partial^\alpha f\|_{p,U,\lambda}^p \right)^{1/p} \qquad f \in W_p^m(U).$$

For the metric on $W_p^m(U)$ we suppress the dependence on U and write

$$\rho_p^m(f, g) \equiv \|f - g\|_{m,p,U} \qquad f, g \in W_p^m(U).$$

Two functions are close in the Sobolev space $W_p^m(U)$ if all generalized derivatives are close in $L_p(U, \lambda)$ distance. In the next section, we give conditions on G and U ensuring that single hidden layer feedforward networks are indeed ρ_p^m-dense in $W_p^m(U)$. Consequently, single hidden layer feedforward networks are capable of approximating an unknown mapping and its generalized derivatives to any degree of accuracy under general conditions, provided that sufficiently many hidden units are available.

The conditions to be placed on U are that U is an open bounded subset of $\mathbb{R}^r$ and that the set of restrictions to U of functions in $C_0^\infty(\mathbb{R}^r)$ is ρ_p^m-dense in $W_p^m(U)$. This places further restrictions on U that have practical consequences for the construction of feedforward networks approximating an unknown mapping and its derivatives. The reason for this is that when U is such that $C_0^\infty(\mathbb{R}^r)$ is not ρ_p^m-dense in $W_p^m(U)$, then it is easy to construct examples of functions belonging to $W_p^m(U)$ which are impossible to approximate arbitrarily well with any feedforward network (or indeed with any smooth function).

For example, take $r = 1$ and let $U = (a, b) \cup (b, c)$, $a < b < c$, $a, b, c \in \mathbb{R}$. Let $f(x) = 0$, $x \in (a, b)$ and let $f(x) = 1$, $x \in (b, c)$. Now f belongs to $C^\infty(U)$, but no function g in $C^\infty(\mathbb{R})$ (recall that elements of $\Sigma(G)$ are defined on $\mathbb{R}$) can approximate f in $S_1^1(U, \lambda)$. Because U lies locally on both sides of the boundary point b we can have a jump in f with no corresponding jump in Df. No function g in $C^\infty(\mathbb{R})$ can exhibit this behavior, even approximately. To obtain an artibrarily accurate approximate using only one feedforward network is thus impossible. However, two networks, one for the region $U_1 = (a, b)$ and the other for the region $U_2 = (b, c)$ can deliver the desired approximations. This strategy of partitioning the domain U and applying a different feedforward net separately to each subdomain satisfying our regularity conditions is often feasible. It is important in practice to examine the input domain to see if this strategy is necessary.

Necessary and sufficient conditions ensuring that U is sufficiently regular that $C_0^\infty(\mathbb{R}^r)$ is ρ_p^m-dense in $W_p^m(U)$ are not presently known. However, there are a number of useful sufficient conditions available, all ruling out the possibility that U lies locally on both sides of its boundary. We give two examples: that U possess the "segment property," or that U is "starshaped with respect to a point."

Let U^c denote the complement of U in $\mathbb{R}^r$ and let the boundary of U be defined as $\partial U \equiv \text{cl } U \cap \text{cl } U^c$. The open set U has the *segment property* if for every x in ∂U there exist a neighborhood of x, denoted N_x, and a nonzero vector y_x in $\mathbb{R}^r$ such that if z belongs to cl $U \cap N_x$, then the segment $z + ty_x$, $0 < t < 1$ belongs to U. A domain possessing the segment property must have an $(r - 1)$-dimensional boundary and cannot lie locally on both sides of any part of its boundary.

THEOREM 2.1 (Adams, 1975, Theorem 3.18) *If U has the segment property, then $C_0^\infty(\mathbb{R}^r)$ is ρ_p^m-dense in $W_p^m(U)$ for $1 \le p < \infty$, $m = 0, 1, 2, \ldots$.* □

The domain U *is starshaped with respect to a point* when there exists x in U such that any ray with origin x has a unique intersection with the boundary ∂U.

THEOREM 2.2 (Maz'ja, 1985, Theorem 1.1.6.1). *If U is a bounded domain starshaped with respect to a point, then $C_0^\infty(\mathbb{R}^r)$ is ρ_p^m-dense in $W_p^m(U)$ for $1 \le p < \infty$, $m = 0, 1, 2, \ldots$.* □

Our results make fundamental use of one last function space, the space $C_\downarrow^\infty(\mathbb{R}^r)$ of rapidly decreasing functions in $C^\infty(\mathbb{R}^r)$. $C_\downarrow^\infty(\mathbb{R}^r)$ is defined as the set of all functions in $C^\infty(\mathbb{R}^r)$ such that for multi-indices α and β, $x^\beta D^\alpha f(x) \to 0$ as $|x| \to \infty$, where $x^\beta \equiv x_1^{\beta_1} x_2^{\beta_2} \ldots x_r^{\beta_r}$ and $|x| \equiv \max_{1 \le i \le r} |x_i|$. Note that $C_0^\infty(\mathbb{R}^r) \subset C_\downarrow^\infty(\mathbb{R}^r)$.

To summarize, the spaces of functions within which we study the approximation capabilities of multilayer feedforward networks are: (1) $C_\downarrow^\infty(\mathbb{R}^r)$; (2) $S_p^m(U, \mu)$ (functions in $C^m(U)$ having derivatives up to order m being $L_p(U, \mu)$-integrable) for particular choices of U and μ; (3) $S_p^m(\text{loc})$ (functions in $C^m(\mathbb{R}^r)$

with derivatives up to order m being $L_p(U, \lambda)$-integrable for all bounded subsets U of $\mathbb{R}^r$); and (4) $W_p^m(U)$ (functions having generalized derivatives up to order m being $L_p(U, \lambda)$ integrable). Associated with each of these spaces is an appropriate metric measuring distance between functions in a way that takes into account the closeness of derivatives of up to a specified order. Thus we consider the metric spaces $(C_\downarrow^\infty(\mathbb{R}^r), \rho_K^m)$, $(S_p^m(U, \mu), \rho_{p,\mu}^m)$, $(S_p^m(\text{loc}), \rho_{p,\text{loc}}^m)$, and $(W_p^m(U), \rho_p^m)$. We seek conditions on G and U ensuring that multilayer feedforward networks have approximation capabilities (i.e., a denseness property) in these spaces. For the case of $W_p^m(U)$, the restrictions on U have practical consequences for the possibility of approximation by multilayer feedforward networks.

3. MAIN RESULTS

All our results flow straightforwardly from our first result. We make use of a Fourier integral representation for single hidden layer feedforward networks with a continuum of hidden units proposed by Irie and Miyake (1988).

THEOREM 3.1. *Let $G \neq 0$ belong to $S_\downarrow^m(\mathbb{R}, \lambda)$ for some integer $m \geq 0$. Then $\Sigma(G)$ is m-uniformly dense on compacta in $C_\downarrow^\infty(\mathbb{R}^r)$.* □

Thus, as long as the hidden layer activation function G belongs to $S_\downarrow^m(\mathbb{R}, \lambda)$ and does not vanish everywhere, then $\Sigma(G)$ can approximate any function belonging to $C_\downarrow^\infty(\mathbb{R}^r)$ and its derivatives up to order m arbitrarily well on compact sets.

The conclusion of this result is in fact strong enough to deliver all desired corollaries regarding approximation in the spaces $S_p^m(U, \mu)$, $S_p^m(\text{loc})$, and $W_p^m(U)$. This follows, roughly speaking, from the denseness of $C_\downarrow^\infty(\mathbb{R}^r)$ in these spaces. However, the condition that G belong to $S_\downarrow^m(\mathbb{R}, \lambda)$ is uncomfortably strong. In particular, this condition rules out the familiar logistic or hyperbolic tangent squashing functions because these are not even members of $S_\downarrow^0(\mathbb{R}, \lambda)$. Indeed, no sigmoid choice for G is allowed by the present condition. Fortunately, the conditions on G can be considerably weakened. We use the following definition.

DEFINITION 3.2. *Let $l \in \{0\} \cup \mathbb{N}$ be given. G is l-finite if $G \in C^l(\mathbb{R})$ and $0 < \int |D^l G| \, d\lambda < \infty$.* □

The practical significance of G being l-finite is established by the following lemma.

LEMMA 3.3. *If G is l-finite then for all $0 \leq m \leq l$ there exists $H \in S_\downarrow^m(\mathbb{R}, \lambda)$, $H \neq 0$, such that $\Sigma(H) \subset \Sigma(G)$.* □

Consequently, it will suffice in Theorem 3.1 that G be l-finite. It follows that $\Sigma(G)$ contains a subset, namely $\Sigma(H)$, m-uniformly dense on compacta in $C_\downarrow^\infty(\mathbb{R}^r)$ for $0 \leq m \leq l$.

From this all our desired corollaries follow. Before stating them, however, it is useful to examine the content of the condition that G be l-finite. First note that the logistic and hyperbolic tangent squashers are l-finite for any $l \in \mathbb{N}$, so that these familiar hidden layer activation functions are covered by our theorems. Next, note that if we have already that $G \in S_\downarrow^m(\mathbb{R}, \lambda)$ then for $1 \leq k \leq m$ it follows that $\int D^k G \, d\lambda = 0$ (a consequence of the fundamental theorem of calculus). More generally, if $G \in C^{l-1}(\mathbb{R})$ and $\int |D^l G| \, d\lambda < \infty$ then $\int D^{l-1} G \, d\lambda = 0$, while if $\int D^{l-1} G \, d\lambda$ exists and is not equal to zero then $\int |D^l G| \, d\lambda = \infty$. To summarize, l-finite activation functions G with $\int D^l G \, d\lambda \neq 0$ have $\int |D^m G| \, d\lambda = \infty$ for all $m < l$, and for $m \geq l$ all l-finite activation functions G have $\int D^m G \, d\lambda = 0$ (provided $D^m G$ exists).

It is informative to examine cases not satisfying the conditions of the theorems. For example, if $G = \sin$ then $G \in C^\infty(\mathbb{R})$, but for all l, $\int |D^l G| \, d\lambda = \infty$. If G is a polynomial of degree m then again $G \in C^\infty(\mathbb{R})$, but for $l \leq m$ we have $\int |D^l G| \, d\lambda = \infty$, although $\int |D^l G| \, d\lambda = 0$ for $l > m$. Consequently, neither trigonometric functions nor polynomials are l-finite; the approximation results obtained here for l-finite activation functions thus have a character distinct from Fourier analysis (e.g., Edmunds and Moscatelli, 1977) and Nachbin's extension of the Stone–Weierstrass theorem (Llavona, 1979).

From Theorem 3.1 and Lemma 3.3 we obtain the following corollaries.

COROLLARY 3.4. *If G is l-finite, then for all $0 \leq m \leq l$, $\Sigma(G)$ is m-uniformly dense on compacta in $C_\downarrow^\infty(\mathbb{R}^r)$.* □

Let U be an open subset of $\mathbb{R}^r$ and let $C^* \subset C^\infty(U)$. By Corollary 3.4 we know that if G is l-finite then $\Sigma(G)$ is ρ_K^m-dense in $C_\downarrow^\infty(\mathbb{R}^r)$ for any compact set K. From this it follows that if the set of restrictions of elements of $C_\downarrow^\infty(\mathbb{R}^r)$ to U is ρ_K^m-dense in C^* then $\Sigma(G)$ is ρ_K^m-dense in C^*. The next corollary is an application of this technique. There are many others.

COROLLARY 3.5. *If G is l-finite, $0 \leq m \leq l$, and U is an open subset of $\mathbb{R}^r$ then $\Sigma(G)$ is m-uniformly dense on compacta in $S_p^m(U, \lambda)$ for $1 \leq p < \infty$.* □

COROLLARY 3.6. *If G is l-finite and μ is compactly supported, then for all $0 \leq m \leq l$, $\Sigma(G) \subset S_p^m(\mathbb{R}^r, \mu)$ and $\Sigma(G)$ is $\rho_{p,\mu}^m$-dense in $S_p^m(\mathbb{R}^r, \mu)$.* □

COROLLARY 3.7. *If G is l-finite, then for all $0 \leq m \leq l$, $\Sigma(G)$ is $\rho_{p,\text{loc}}^m$-dense in $S_p^m(\text{loc})$.* □

COROLLARY 3.8. *If G is l-finite, $0 \leq m \leq l$, U is an open bounded subset of $\mathbb{R}^r$ and $C_0^\infty(\mathbb{R}^r)$ is ρ_p^m-dense in $W_p^m(U)$ then $\Sigma(G)$ is also ρ_p^m-dense in $W_p^m(U)$.*

Theorems 2.1 and 2.2 can be applied to provide conditions on U ensuring that $C_0^\infty(\mathbb{R}^r)$ is ρ_p^m-dense as required.

These results rigorously establish that sufficiently complex multilayer feedforward networks with as few as a single hidden layer are capable of arbitrarily accurate approximation to an unknown mapping and its (generalized) derivatives in a variety of precise senses. The conditions imposed on G are relatively mild; the conditions required of U have practical implications.

The fact that $\Sigma(G)$ is m-uniformly dense on compacta in $C_1^\infty(\mathbb{R}^r)$ (hence $C_0^\infty(\mathbb{R}^r)$) has further consequences that we now note, but do not elaborate on. Specifically, it follows from Theorem 7.40 of Adams (1975) that $\Sigma(G)$ contains a subset dense in the fractional Sobolev space $W_p^s(U)$ for $s = m + \sigma$, $m \in \mathbb{N}$, $0 < \sigma < 1$, provided there exists a "strong $(m + 1)$-extension operator E" for U. Further, Theorem 8.28 of Adams (1975) applies to imply that $\Sigma(G)$ contains a subset dense in Orlicz–Sobolev spaces. The reader is referred to Chapters 7 and 8 of Adams (1975) for background and details.

In concluding this section we note that it follows trivially that all of the foregoing results hold for multi-output networks defined by letting β_j be a vector rather then a scalar. Also, identical conclusions hold for feedforward networks with more than one hidden layer under the same conditions on G, by arguments analogous to those of HSW (Corollary 2.7).

4. NETWORK IMPLEMENTATION AND SOME REMARKS ON LEARNING

Figure 1 provides a schematic representation of a single hidden layer feedforward network with two inputs, two hidden units and one output. We consider this architecture for the sake of simplicity and because it suffices to illustrate the relevant concepts. Figure 2 presents an augmentation of this network that possesses additional output nodes on which register the values of the first partial derivatives (with respect to inputs x_1 and x_2), denoted g_1 and g_2, of the network output function g. The connections of the original feedforward network have been drawn in dashed lines in Figure 2 to emphasize the additional connections required by this augmentation. Two features are noteworthy: (1) the addition of the derivative activation elements (to compute DG) at the hidden layer; and (2) the direct "connections" of the input to hidden weights γ to the multiplication elements above the hidden layer.

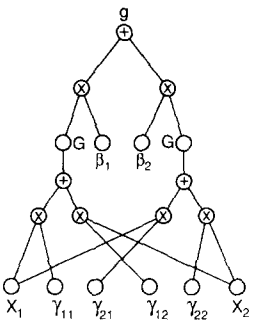

FIGURE 1. Feedforward Network. ○ input unit, ⊗ multiplication unit, G○ activation unit, ⊕ addition unit. Note: biases not shown.

The treatment of network connection strengths as "inputs" in these figures is motivated in part by a desire to make clear the nature of the relation between the original network and its augmentation. However, it turns out that practical implementations of the augmented network may benefit from precisely this sort of architecture. The reason is that weights obtained from any suitable learning procedure can be loaded directly into this network for use in applications.

Learning procedures delivering connection strengths implementing the approximations shown here to be possible are obtainable from results of Gallant (1987); see Gallant and White (1989) for details.

5. SUMMARY AND CONCLUDING REMARKS

Multilayer feedforward networks with as few as a single hidden layer and an appropriately smooth hidden layer activation function are capable of arbitrarily accurate approximation to an arbitrary function and its derivatives. In fact, these networks can

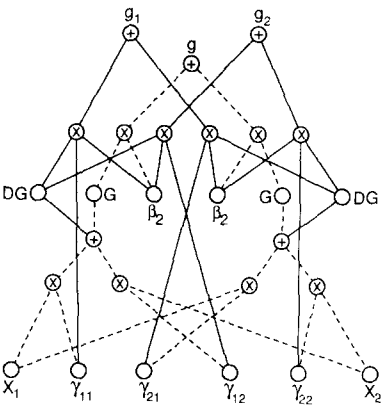

FIGURE 2. Derivative network. ○ input unit, ⊗ multiplication unit, G○ activation unit, ⊕ addition unit, DG○ activation derivative unit. Note: Biases not shown.

approximate functions that are not differentiable in the classical sense, but possess only a generalized derivative, as is the case for certain piecewise differentiable functions. These approximation results provide a previously missing theoretical justification for the use of multilayer feedforward networks in applications requiring approximation of a function and its derivatives.

REFERENCES

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

Billingsley, P. (1979). *Probability and measure*. New York: Wiley.

Cybenko G. (1989). Approximation by superpositions of a sigmodial function. In *Mathematics of control, signals and systems*, **2**, 303–314.

Dugundji, J. (1966). *Topology*. Boston: Allyn and Bacon.

Dym, H., & McKean, H.P. (1972). *Fourier series and integrals*. New York: Academic Press.

Edmunds D.E., & Moscatelli, V.B. (1977). Fourier approximations and embeddings in Sobolev space. *Dissertationes Mathematicae*, **145**, 1–46.

Elbadawi, I., Gallant, A.R., & Souza, G. (1983). An elasticity can be estimated consistently without a priori knowledge of functional form. *Econometrica*, **51**, 1731–1752.

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

Gallant, A.R. (1987). Identification and consistency in semi-nonparametric regression. In T. Bewley (Ed.), *Advances in econometrics, Fifth World Congress* (Vol. 1 pp. 145–170). New York: Cambridge University Press.

Gallant A.R., & White, H. (1989). On learning the derivatives of an unknown mapping with multilayer feedforward networks. UCSD Department of Economics Discussion Paper 89–53.

Gilstrap, L., & Dominy, R. (1989). A general explanation and interrogation system for neural networks. Poster Presentation, *International Joint Conference on Neural Networks*, Washington, D.C.

Halmos, P. (1974). *Measure theory*. New York: Springer-Verlag.

Hecht-Nielsen, R. (1989). Theory of the back-propagation neural network. In *Proceedings of the 1989 International Joint Conference on Neural Networks* (pp. I:593–606). New York: IEEE Press.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**, 359–368.

Irie, B., & Miyake, S. (1988). Capabilities of three-layered perceptrons. In *Proceedings of the 1988 IEEE International Conference on Neural Networks* (pp. I:641–648). New York: IEEE Press.

Jordan, M. (1989). Generic constraints on underspecified target trajectories. In *Proceedings of the 1989 International Joint Conference on Neural Networks* (pp. I:217–225). New York: IEEE Press.

Kufner, A. (1980). *Weighted Sobolev spaces*. Leipzig: B.G. Teubner.

Kufner, A., & Sandig, A.M. (1987). *Some applications of weighted Sobolev spaces*. Leipzig: B.G. Teubner.

Lapedes, A., & Farber, R. (1987). Nonlinear signal processing using neural networks: Prediction and system modeling. Technical Report LA-UR-87-2662, Los Alamos National Laboratory, Los Alamos, NM.

Llavona, J. (1979). Approximations of differentiable functions. In G.C. Rota (Ed.), *Studies in Analysis* (pp. 197–221). New York: Academic Press.

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

Schwartz, L. (1950). *Théorie des Distributions*. Paris: Hermann.

Showalter, R.E. (1977). *Hilbert space methods for partial differential equations*. London: Pitman.

Stinchcombe, M., & White, H. (1989). Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions. In *Proceedings of the 1989 International Joint Conference on Neural Networks*, Washington, D.C. (pp. I:613–617). New York: IEEE Press.

Ullah, A. (1988). Non-parametric estimation of econometric functionals. *Canadian Journal of Economics*, **21**, 625–658.

Varian, H. (1978). *Microeconomic analysis*. New York: Norton.

Vinod, H.D., & Ullah, A. (1985). Flexible production function estimation by non-parametric kernel estimators. Research Report, University of Western Ontario, Department of Economics.

MATHEMATICAL APPENDIX

In the proof of Theorem 3.1 we shall make use of Fourier transforms and some of their properties. An excellent exposition of these techniques is given by Dym and McKean (1972). Most of their theorems deal only with the univariate case explicitly; however, extensions to the multivariate case are straightforward.

Let f belong to $C_c^\infty(\mathbb{R}^r)$. The Fourier transform $f \to \hat{f}$ with

$$\hat{f}(a) = \int_{\mathbb{R}^r} e^{-2\pi i a^T x} f(x)\, dx, \qquad a \in \mathbb{R}^r,$$

maps $C_c^\infty(\mathbb{R}^r)$ onto $C_c^\infty(\mathbb{R}^r)$. (See Chapter 2.2 in Dym and McKean, 1972.) In particular, for all multi-indices α, both $D^\alpha f$ and $\hat{D^\alpha f}$ (the Fourier transform of $D^\alpha f$) are in $L_1(\mathbb{R}^r)$. Integration by parts gives

$$\hat{D^\alpha f}(a) = (2\pi i a)^\alpha \hat{f}(a)$$

and, by the Fourier inversion theorem,

$$D^\alpha f(x) = \int_{\mathbb{R}^r} e^{2\pi i a^T x} (2\pi i a)^\alpha \hat{f}(a)\, da.$$

Similarly, as $D^\alpha G \in L_1(\mathbb{R})$ for $q = m_2$ we may take Fourier transforms

$$\hat{D^\alpha G}(b) = \int_{\mathbb{R}} e^{2\pi i b p} D^\alpha G(p)\, dp$$

and again,

$$\hat{D^\alpha G}(b) = (2\pi i b)^\alpha \hat{G}(b).$$

In particular, if we had $\hat{G}(b) = 0$ for all b, then $G(b) \equiv 0$ by the uniqueness theorem. This is ruled out by assumption in Theorem 3.1. Thus, by continuity of $\hat{G}$, we can always find $b \neq 0$ such that $\hat{G}(b) \neq 0$.

Proof of Theorem 3.1. The proof is accomplished in four steps.

Step 1: Let $f \in C_c^\infty(\mathbb{R}^r)$ and fix $b \neq 0$ such that $\hat{G}(b) \neq 0$. Then

$$D^\alpha f(x) = \int_{S^{r-1}} \int_{\mathbb{R}} a^\alpha D^m G(a^T x - \theta) K_b(a, \theta)\, da\, d\theta,$$

where

$$K_b(a, \theta) = \mathrm{Re}\left[\frac{|b|^r \hat{f}(ba) e^{2\pi i b\theta}}{\hat{G}(b)}\right].$$

Proof. We have, using Fubini's Theorem,

$$\begin{aligned}
&\int_{a=-M}^{M}\int_{\theta=-\gamma M}^{\gamma M} a^\alpha D^\alpha G(a^t x-\theta)\frac{|b|^r\hat f(ba)e^{2\pi ib\theta}}{\hat G(b)}\,d\theta\,da\\
&=\int_{a=-M}^{M}\frac{a^\alpha|b|^r\hat f(ba)}{\hat G(b)}\left[\int_{\theta=-\gamma M}^{\gamma M} D^\alpha G(a^t x-\theta)e^{2\pi ib\theta}\,d\theta\right]da\\
&=\int_{a=-M}^{M}\frac{a^\alpha|b|^r\hat f(ba)}{\hat G(b)}\left[\int_{p=-\infty}^{\infty} D^\alpha G(p)e^{-2\pi ibp}e^{2\pi iba^t x}\,dp\right]da\\
&=\int_{a=-M}^{M}\frac{a^\alpha|b|^r\hat f(ba)}{\hat G(b)}\hat D^\alpha G(b)e^{2\pi iba^t x}\,da\\
&=\int_{a=-M}^{M}a^\alpha|b|^r\hat f(ba)(2\pi ib)^\alpha e^{2\pi iba^t x}\,da\\
&=\int_{a=-M}^{M}(2\pi iba)^\alpha\hat f(ba)e^{2\pi iba^t x}|b|^r\,da\\
&=\int_{a=-M}^{M}\hat D^\alpha f(ba)e^{2\pi iba^t x}\,d(ba)\\
&=D^\alpha f(x),
\end{aligned}$$

whence Step 1 by taking real parts.

Step 2: Let K be a compact subset of $\mathbb{R}^r$, $\kappa=\kappa(K)\equiv\sup\{|x|,\ x\in K\}<\infty$ and let $\gamma\equiv\kappa r+1$. Put

$$f_M(x)=\int_{a=-M}^{M}\int_{\theta=-\gamma M}^{\gamma M}G(a^t x-\theta)K_b(a,\theta)\,da\,d\theta,\qquad M\in\mathbb{N}.$$

Then for all α such that $|\alpha|\le m$,

$$D^\alpha f_M(x)=\int_{a=-M}^{M}\int_{\theta=-\gamma M}^{\gamma M}a^\alpha D^\alpha G(a^t x-\theta)K_b(a,\theta)\,da\,d\theta,$$

and, as $M\to\infty$, $D^\alpha f_M\to D^\alpha f$ uniformly on K.

Proof. The formula for $D^\alpha f_M$ is obtained by differentiating the representation of f_M with respect to x under the integral sign. Now

$$\begin{aligned}
D^\alpha f(x)-D^\alpha f_M(x)&=\int_{a=-M}^{M}\int_{\mathbb{R}}a^\alpha D^\alpha G(a^t x-\theta)K_b(a,\theta)\,da\,d\theta\\
&\quad+\int_{a=-M}^{M}\int_{\theta=-\gamma M}^{\gamma M}a^\alpha D^\alpha G(a^t x-\theta)K_b(a,\theta)\,da\,d\theta.
\end{aligned}$$

The absolute value of the first integral is less than or equal to

$$\begin{aligned}
&\int_{a=-M}^{M}\left[\int_{\mathbb{R}}|D^\alpha G(a^t x-\theta)|d\theta\right]\frac{|b|^r|a^\alpha\hat f(ba)|}{|\hat G(b)|}\,da\\
&\le\|D^\alpha G\|_{1,\mathbb{R}^r}\int_{ba=-bM}^{bM}\left[\frac{|(ba)^\alpha\hat f(ba)|}{|\hat G(b)|\,|b|^\alpha}\right]d(ba)\\
&=\frac{\|D^\alpha G\|_{1,\mathbb{R}^r}}{|\hat G(b)|\,|b|^\alpha}\int_{a=-bM}^{bM}|a^\alpha\hat f(a)|\,da.
\end{aligned}$$

To obtain an upper bound for the second integral, notice that if $|a|\le M$, $x\in K$, and $|\theta|>\gamma M$, we have $|a^t x|\le\kappa r M$ and thus

$$|a^t x-\theta|\ge\|\theta|-|a^t x\|>\gamma M-\kappa r M=M.$$

It follows that, for $x\in K$, the absolute value of the second integral is less than

$$\begin{aligned}
&\int_{a=-M}^{M}\left[\int_{\theta=-\gamma M}^{\gamma M}|D^\alpha G(a^t x-\theta)|\,d\theta\right]\frac{|b|^r|a^\alpha\hat f(ba)|}{|\hat G(b)|}\,da\\
&\le\frac{\|a^\alpha\hat f\|_{1,\mathbb{R}^r}}{|\hat G(b)|\,|b|^\alpha}\int_{a=-M}^{M}|D^\alpha G(u)|\,du.
\end{aligned}$$

Combining both inequalities, we obtain

$$\begin{aligned}
\sup_{x\in K}|D^\alpha f(x)-D^\alpha f_M(x)|&\le\frac{1}{|\hat G(b)|\,|b|^\alpha}\\
&\times\left(\|D^\alpha G\|_{1,\mathbb{R}^r}\int_{a=-bM}^{bM}|a^\alpha\hat f(a)|\,da+\|a^\alpha\hat f\|_{1,\mathbb{R}^r}\int_{a=-M}^{M}|D^\alpha G(u)|\,du\right)
\end{aligned}$$

which tends to 0 as $M\to\infty$ by integrability of $a^\alpha\hat f$ and $D^\alpha G$.

Step 3: For fixed M, consider the following Riemann sum approximations to f_M. Let

$$T_N=\{v=(v_0,\ldots,v_r):\ v_i \text{ is integer and } -N\le v_i<N,\ i=0,\ldots,r\}$$

and

$$S_{M,N}f(x)=\sum_{v\in T_N}\beta_{M,N,v}\,G(a_{M,N,v}^t x-\theta_{M,N,v})\in\Sigma^r(G),$$

with $a_{M,N,v}=(v_1,\ldots,v_r)^t M/N$, $\theta_{M,N,v}=v_0\gamma M/N$, $\beta_{M,N,v}=\gamma(M/N)^{r+1}K_b(a_{M,N,v},\theta_{M,N,v})$. Then for all α such that $|\alpha|\le m$, as $N\to\infty$, $D^\alpha S_{M,N}f\to D^\alpha f_M$ uniformly on K.

Proof. Introduce the notations

$$H_\alpha(x,a,\theta)=a^\alpha D^\alpha G(a^t x-\theta)K_b(a,\theta),$$

$$\begin{aligned}
B_{M,N,v}=\{(a,\theta):\ &v_0\gamma M/N\le\theta\le(v_0+1)\gamma M/N;\\
&vM/N\le a_i\le(v_i+1)M/N,\ i=1,\ldots,r\}.
\end{aligned}$$

Observe that $|T_N|=\#v$ in $T_N=(2N)^{r+1}$ and that

$$\bigcup_{v\in T_N}B_{M,N,v}=\{(a,\theta):\ a\in[-M,M]^r,\ \theta\in[-\gamma M,\gamma M]\},$$

$$\int_{B_{M,N,v}}da\,d\theta=\gamma(M/N)^{r+1}.$$

It follows that

$$\begin{aligned}
D^\alpha S_{M,N}f(x)&=\sum_{v\in T_N}\gamma(M/N)^{r+1}a_{M,N,v}^\alpha\\
&\quad\times D^\alpha G(a_{M,N,v}^t x-\theta_{M,N,v})K_b(a_{M,N,v},\theta_{M,N,v})\\
&=\sum_{v\in T_N}\int_{B_{M,N,v}}H_\alpha(x,a_{M,N,v},\theta_{M,N,v})\,da\,d\theta
\end{aligned}$$

and that

$$D^\alpha f_M(x)=\sum_{v\in T_N}\int_{B_{M,N,v}}H_\alpha(x,a,\theta)\,da\,d\theta.$$

For all $|\alpha|\le m$, the map $(a,x,\theta)\to H_\alpha(x,a,\theta)$ is continuous and therefore uniformly continuous for $(x,a,\theta)\in K\times[-M,M]^r\times[-\gamma M,\gamma M]$. In particular,

$$\begin{aligned}
\omega_{M,K}(\delta)\equiv\sup_{\alpha:|\alpha|\le m}\{&|H_\alpha(x,a,\theta)-H_\alpha(x,\tilde a,\tilde\theta)|:\\
&a,\tilde a\in[-M,M]^r,\ |a-\tilde a|\le\delta;\ \theta,\tilde\theta\\
&\in[-\gamma M,\gamma M],\ |\theta-\tilde\theta|\le\gamma\delta\}
\end{aligned}$$

tends to 0 as $\delta\to 0$. Hence, for all $x\in K$,

$$\begin{aligned}
&|D^\alpha f_M(x)-D^\alpha S_{M,N}f(x)|\\
&\le\sum_{v\in T_N}\int_{B_{M,N,v}}|H_\alpha(x,a,\theta)-H_\alpha(x,a_{M,N,v},\theta_{M,N,v})|\,da\,d\theta\\
&\le\sum_{v\in T_N}\int_{B_{M,N,v}}\omega_{M,K}(M/N)\,da\,d\theta\\
&=\omega_{M,K}(M/N)(2N)^{r+1}\gamma(M/N)^{r+1}\\
&=\gamma(2M)^{r+1}\omega_{M,K}(M/N)
\end{aligned}$$

and therefore

$$\sup_{x \in K} |D^\alpha f_M(x) - D^\alpha S_{M,N} f(x)| \leq$$

$$\gamma(2M)^{r+1} \omega_{M,K}(M/N) \to 0 \quad \text{as } N \to \infty.$$

Step 4: The result follows from Steps 2 and 3, first taking $N \to \infty$ and then $M \to \infty$. □

Proof of Lemma 3.3. For any function f mapping R to R and any $n \geq 0$, define a new function from R to R, the n-shifted difference of f, $\Delta^n(f)$, recursively as follows. For all $x \in \mathrm{R}$ $\Delta^0(f)(x) = f(x)$, and for all $n \geq 1$ and all $x \in \mathrm{R}$ $\Delta^n(f)(x) = \Delta^{n-1}(f)(x + a) - \Delta^{n-1}(f)(x)$ where $a > 0$ is some fixed real number. The proof will consist of showing that if G is l-finite then $\Delta^l(G) \neq 0$ and belongs to $S_1^m(\mathrm{R}, \lambda)$ for all $0 \leq m \leq l$.

We first record some useful properties of $\Delta^n(f)$: (1) For all $n \geq 0$, $\Delta^n(f) \in \Sigma(f)$; (2) If $f \in C^k(\mathrm{R})$ then for all $n > 0$, $\Delta^n(f) \in C^k(\mathrm{R})$; (3) If $f \in C^l(\mathrm{R})$ and $0 \leq m \leq l$, then for all $n \geq 0$, $\Delta^n(D^m f) = D^m \Delta^n(f)$.

We now proceed to the proof. When $l = 0$ the result is trivial. For $l \geq 1$ we show that if G is l-finite then $\Delta^1(G)$ is $(l - 1)$-finite (i.e., $0 < \int |\Delta^1(D^{l-1}G)|\, d\lambda < \infty$).

By assumption $0 < \int |D^l G|\, d\lambda < \infty$. For $x \in \mathrm{R}$ let $D^l G^+(x) = \max(D^l G(x), 0)$ and $D^l G^-(x) = -\min(D^l G(x), 0)$ so that $D^l G = D^l G^+ - D^l G^-$. Further, let $M^+(x) = \int_{-\infty}^{x} D^l G^+(t)\, dt$ and $M^-(x) = \int_{-\infty}^{x} D^l G^-(t)\, dt$. M^+ and M^- are continuous, nondecreasing functions bounded above by $\int D^l G^+\, d\lambda$ and $\int D^l G^-\, d\lambda$. By the fundamental theorem of calculus, $D^{l-1}G(x) = M^+(x) - M^-(x) + k$ for some constant k. Thus, $\Delta^1(D^{l-1}G) = M^+(t + a) - M^+(t) - (M^-(t + a) - M^-(t))$, so $|\Delta^1(D^{l-1}G)| \leq (M^+(t + a) - M^+(t)) + (M^-(t + a) - M^-(t))$. Integrating, we obtain

$$\int |\Delta^1(D^{l-1}G)|\, d\lambda \leq a \left[\int D^l G^+\, d\lambda + \int D^l G^-\, d\lambda \right] = a \int |D^l G|,$$

where the inequality follows from Billingsley (1979, Ex. 18.10, p. 205). Thus, $\int |\Delta^1(D^{l-1}G)|\, d\lambda < \infty$. All that is left is to show that $0 < \int |\Delta^1(D^{l-1}G)|\, d\lambda$. If $\int |\Delta^1(D^{l-1}G)|\, d\lambda = 0$ then $\Delta^1(D^{l-1}G)(x) = 0$ for all $x \in \mathrm{R}$. But this implies that $D^{l-1}G$ is periodic with period a, which in turn implies that $D^{l-1}G = 0$, contradicting the assumption that G is l-finite.

Inductive application of this argument shows that $0 < \int |\Delta^l(G)|\, d\lambda < \infty$. This implies that $\int D^m \Delta^l(G)\, d\lambda = 0$ for $1 \leq m \leq l$, proving that $\Delta^l(G) \in S_1^m(\mathrm{R}, \lambda)$. □

Proof of Corollary 3.4. By Lemma 3.3 there is an H in $\Sigma(G)$ satisfying the assumptions of Theorem 3.1. □

The proof of Corollary 3.5 uses the following lemma, which closely resembles the Arzela–Ascoli theorem (e.g., Dugundji, 1966, Theorem XII. 6.4). The Arzela–Ascoli theorem would allow us to prove that pointwise convergence implies uniform convergence on compacta. Our Lemma establishes that almost everywhere-λ convergence, generally not a topological concept, implies uniform convergence on compacta.

LEMMA A.1. *Let U be a nonempty open subset of* R^r. *If* $\{f_n: U \to \mathrm{R}\}$ *is equicontinuous on every compact subset of U and $f_n \to f$ a.e.-λ, $f \in C(U)$, then $f_n \to f$ uniformly on compact subsets of U.* □

Proof. Pick arbitrary compact $K \subset U$ and $\varepsilon > 0$. We show that there exists $N \in \mathrm{N}$ such that for all $n \geq N$ $\max_{x \in K} |f_n(x) - f(x)| < \varepsilon$.

Let $A = \{x \in U | f_n(x) \to f(x)\}$. Because $f_n \to f$ a.e.-λ, A is dense in U. For any $\eta > 0$, let $K_\eta = \{x \in \mathrm{R}^r | |x - y| < \eta \text{ for some } y \in K\}$. Because K is compact and U is open there exists $\eta > 0$ such that the compact set cl $K_{2\eta}$ is a subset of U. Pick such an η. Because $\{f_n\}$ is equicontinuous on cl $K_{2\eta}$, hence $K_{2\eta}$, and f is continuous, there exists $\delta > 0$ such that

$$\sup_{g \in \{f_n, f\}} \sup_{x, y \in K_{2\eta}, |x-y| < \delta} |g(x) - g(y)| < \varepsilon/3.$$

Because K_η is open and A is dense in U, the collection of sets $\{B(x, \delta) \cap K_{2\eta} | x \in K_\eta \cap A\}$ is an open cover of cl K_η, where $B(x, \delta) \equiv \{y \in \mathrm{R}^r | |y - x| < \delta\}$. Let $\{B(x_i, \delta) \cap K_{2\eta}\}_{i \in I}$ be a finite subcover. Because $x_i \in A$, we can pick N sufficiently large that for all $n \geq N$ $|f_n(x_i) - f(x_i)| < \varepsilon/3$. For every $i \in I$ we have

$$\sup_{x \in B(x_i, \delta) \cap K_{2\eta}} |f_n(x) - f(x)| \leq \sup_{x \in B(x_i, \delta) \cap K_{2\eta}} |f_n(x) - f_n(x_i)|$$

$$+ |f_n(x_i) - f(x_i)| + |f(x_i) - f(x)| < 3(\varepsilon/3) = \varepsilon$$

for all $n \geq N$. Because $K \subset \cup_{i \in I} (B(x_i, \delta) \cap K_{2\eta})$, the result follows. □

Proof of Corollary 3.5. It suffices to show that the set of restrictions of $C_1^\infty(\mathrm{R}^r)$ to U is ρ_K^m-dense in $S_p^m(U, \lambda)$. Let K be a compact subset of U and g an arbitrary element of $S_p^m(U, \lambda)$. Again, put $K_\eta = \{x \in \mathrm{R}^r | |x - y| < \eta \text{ for some } y \in K\}$. Because K is compact and U is open, $K \subset K_\eta \subset K_{2\eta} \subset U$ for all $\eta > 0$ sufficiently small. Pick such an η and let $\phi \in C_0^\infty(\mathrm{R}^r)$ satisfy $0 \leq \phi \leq 1$, $\phi(x) = 1$ if $x \in K_\eta$ and $\phi(x) = 0$ if $x \notin K_{2\eta}$. Then $h \equiv \phi \cdot g$ belongs to $C_1^m(\mathrm{R}^r)$ and $h(x) = g(x)$ for $x \in K_\eta$.

For $\varepsilon > 0$, set $\Psi^\varepsilon(x) = \exp(|x|^2 - \varepsilon^2)^{-1}$ if $|x| < \varepsilon$ and $\Psi^\varepsilon(x) = 0$ if $|x| \geq \varepsilon$. Define $\psi^\varepsilon = (\int \Psi^\varepsilon\, d\lambda)^{-1} \Psi^\varepsilon$ and set $h^\varepsilon(x) = \int h(x - y)\psi^\varepsilon(y)\, dy$. By the boundedness of $K_{2\eta}$ and Maz'ja (1985, 1.1.5, 1)–3), pp. 11–12) $h^\varepsilon \in C_0^\infty(\mathrm{R}^r)$ and $\rho_p^m(h^\varepsilon, h) \to 0$ as $\varepsilon \to 0$. Taking subsequences if necessary, we have for all $|\alpha| \leq m$ that $D^\alpha h^\varepsilon \to D^\alpha h$ a.e.-λ on K_η.

It is easy to check that for all $\varepsilon < \eta$ we have that for all $|\alpha| \leq m$ and all $\delta > 0$

$$\sup_{x, y \in K_\eta, |x-y| < \delta} |D^\alpha h^\varepsilon(x) - D^\alpha h^\varepsilon(y)|$$

$$\leq \sup_{x, y \in K_{2\eta}, |x-y| < \delta} |D^\alpha h(x) - D^\alpha h(y)|.$$

Because h vanishes outside of $K_{2\eta}$, this implies that $\{D^\alpha h^\varepsilon\}$ is an equicontinuous family of functions on K_η. Lemma A.1 thus applies, and the result follows. □

Proof of Corollary 3.6. Because μ has compact support K, $S_p^m(\mathrm{R}^r, \mu) = C^m(\mathrm{R}^r)$. Thus $\Sigma(G) \subset C^l(\mathrm{R}^r) \subset C^m(\mathrm{R}^r) = S_p^m(\mathrm{R}^r, \mu)$.

Let U be a bounded open set containing K. Let $C^m(\overline{U})$ denote the set of restrictions of functions in $C^m(\mathrm{R}^r)$ to U. Now $C^m(\overline{U}) \subset S_p^m(U, \lambda)$, so Corollary 3.5 implies that $\Sigma(G)$ is m-uniformly dense on compact subsets of U in $C^m(\overline{U})$. In particular, every element of $S_p^m(\mathrm{R}^r, \mu)$ can be m-uniformly approximated on $K = \text{supp}\, \mu$. □

Proof of Corollary 3.7. This follows from the definition of $\rho_{p,\mathrm{loc}}^m$ and Corollary 3.6. □

Proof of Corollary 3.8. Because $C_0^\infty(\mathrm{R}^r) \subset C_1^\infty(\mathrm{R}^r)$ and $C_0^\infty(\mathrm{R}^r)$ is ρ_p^m-dense in $W_p^m(U)$, it is sufficient to show that $\Sigma(G)$ is ρ_p^m-dense in $C_1^\infty(\mathrm{R}^r)$. Because U is bounded, cl(U) is compact. By Corollary 3.4, $\Sigma(G)$ is m-uniformly dense on compacta in $C_1^\infty(\mathrm{R}^r)$. In particular, for every $f \in C_1^\infty(\mathrm{R}^r)$ and for every $\varepsilon > 0$, there is a $g \in \Sigma(G)$ such that $\max_{|\alpha| \leq m} \sup_{x \in U} |D^\alpha f(x) - D^\alpha g(x)| < \varepsilon$. Thus $\rho_p^m(f, g_{|U}) \leq \sum_{|\alpha| \leq m} (|\varepsilon|^p \cdot \lambda(U))^{1/p} < \lambda(U)^{1/p}(m + 1)^r \varepsilon$. Because ε is arbitrary, the proof is complete. □