Advances in Applied Mathematics **22,** 103–118 (1999)
Article ID aama.1998.0623, available online at http://www.idealibrary.com on IDEAL®

Identifying Linear Combinations of Ridge Functions

Martin D. Buhmann*

Mathematik, Lehrstuhl 8, Universität Dortmund, 44221 Dortmund, Germany

and

Allan Pinkus†

Department of Mathematics, Technion-Israel Institute of Technology, Haifa, 32000, Israel

Received July 1, 1997; accepted September 19, 1997

This paper is about an inverse problem. We assume we are given a function $f(\mathbf{x})$ which is some sum of ridge functions of the form $\sum_{i=1}^{m} g_i(\mathbf{a}^i \cdot \mathbf{x})$ and we just know an upper bound on m. We seek to identify the functions g_i and also to identify the directions $\mathbf{a}^i$ from such limited information. Several ways to solve this nonlinear problem are discussed in this work. © 1999 Academic Press

1. INTRODUCTION

A ridge function is a multivariate function

$$h: \mathbb{R}^n \to \mathbb{R},$$

of the simple form

$$h(x_1, \ldots, x_n) = g(a_1 x_1 + \cdots + a_n x_n) = g(\mathbf{a} \cdot \mathbf{x}),$$

where $g: \mathbb{R} \to \mathbb{R}$ and $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n \setminus \{\mathbf{0}\}$. In other words, it is a multivariate function constant on the parallel hyperplanes $\mathbf{a} \cdot \mathbf{x} = c$, $c \in \mathbb{R}$. The vector $\mathbf{a} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ is generally called the direction.

Ridge functions appear in various areas and under various guises. We find them in the area of partial differential equations (where they have

* E-mail: mdb@math.uni-dortmund.de.
† E-mail: pinkus@math.technion.ac.il.

0196-8858/99 $30.00

been known for many, many years under the name of *plane waves* [7]). We also find them used in computerized tomography [10], in statistics (where they appear in projection pursuit algorithms [5]), in neural networks, and of course in approximation theory. More about ridge functions may be found in Pinkus [11], and references therein.

When dealing with ridge functions, one is generally interested in one of three possible sets of functions.

The first is given by

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

That is, we fix a finite number of directions and we consider linear combinations of ridge functions with these directions. The functions g_i are the ''variables.'' This is a linear space.

The second set is

$$\mathcal{R}_m = \left\{ \sum_{i=1}^m g_i(\mathbf{a}^i \cdot \mathbf{x}) : \mathbf{a}^i \in \mathbb{R}^n \setminus \{\mathbf{0}\},\ g_i \in C(\mathbb{R}),\ i = 1, \ldots, m \right\}.$$

Here, we fix m and we choose both the functions g_i and the directions $\mathbf{a}^i$. This is not a linear space.

The third set is motivated by a model in neural networks. It is a subset of the second. We fix $\sigma \in C(\mathbb{R})$, called the *transfer function* in neural network literature, and we let

$$\mathcal{N}_m = \left\{ \sum_{i=1}^m c_i \sigma\left(\mathbf{a}^i \cdot \mathbf{x} - b_i\right) : \mathbf{a}^i \in \mathbb{R}^n \setminus \{\mathbf{0}\},\ c_i, b_i \in \mathbb{R},\ i = 1, \ldots, m \right\}.$$

Here we also fix m and choose both the directions $\mathbf{a}^i$ (called the *weights*), and the shifts b_i (called the *thresholds*). This is not a linear space.

One problem met with when dealing with function sets such as the preceding, is in knowing if and when a given function is in the set. That is, do we have any way of knowing whether any prescribed f is in $\mathcal{R}(\mathbf{a}^1, \ldots, \mathbf{a}^m)$ for some given directions $\mathbf{a}^1, \ldots, \mathbf{a}^m$, or in $\mathcal{R}_m$, or in $\mathcal{N}_m$? While the first set is linear, the latter two are not, and this problem is therefore far from trivial. A second much related problem is the following. Assume you know (or suppose) that f is in $\mathcal{R}(\mathbf{a}^1, \ldots, \mathbf{a}^m)$ for some given directions $\mathbf{a}^1, \ldots, \mathbf{a}^m$, or in $\mathcal{R}_m$, or in $\mathcal{N}_m$. How can we determine the unknowns, be they the functions, the directions, or the shifts? (If we know a method for determining these unknowns, we essentially have a method of finding whether we are in the appropriate set: we assume that we are in the appropriate set,

we identify the unknowns, and then we check whether the resulting function is in fact our original function.)

In this paper we address these questions and we give a rather generic method of answering the latter question. That is, we show that the problem can be solved. A major drawback is that this method is rather more theoretical than practical, although in principle our proofs are constructive.

In a previous paper [3], we considered a similar recovery problem. There we assumed that we were given a function G: $\mathbb{R}^n \to \mathbb{R}$ and a function f: $\mathbb{R}^n \to \mathbb{R}$, the latter we knew to be of the form

$$f(\mathbf{x}) = \sum_{j=1}^{m} c_j G(\mathbf{x} - \mathbf{t}_j), \qquad \mathbf{x} \in \mathbb{R}^n,$$

for some unknown coefficients $\{c_j\}_{j=1}^m \subset \mathbb{R}$ and shifts $\{\mathbf{t}_j\}_{j=1}^m \subset \mathbb{R}^n$, where m (or an upper bound on m) is known. The problem was to identify the coefficients and shifts. The techniques developed in that paper are used here. (In that paper we also considered $\mathcal{N}_m$. The result obtained (in Theorem 6 of [3]), while not technically in error, was meaningless. Thus this paper also allows us to redress this wrong.)

2. UNIQUENESS AND SMOOTHNESS

When representing a function as a sum of ridge functions, or when seeking to identify its various components, it is of fundamental importance to try to understand the extent to which any representation is unique. We therefore ask, if

$$f(\mathbf{x}) = \sum_{i=1}^{k} g_i(\mathbf{a}^i \cdot \mathbf{x}) = \sum_{i=1}^{l} h_i(\mathbf{b}^i \cdot \mathbf{x}),$$

what can be said about the $\{g_i\}_{i=1}^k$ and $\{\mathbf{a}^i\}_{i=1}^k$ relative to the $\{h_i\}_{i=1}^l$ and $\{\mathbf{b}^i\}_{i=1}^l$?

From linearity properties, this problem reduces to the following formulation: Assuming

$$\sum_{i=1}^{m} g_i(\mathbf{a}^i \cdot \mathbf{x}) = 0,$$

m (or at least an upper bound on it) being known, what can we say about the g_i? The following result is valid.

PROPOSITION 1. *If m is finite, the $\mathbf{a}^i$ are pairwise linearly independent, the $g_i \in C(\mathbb{R})$, $i = 1, \ldots, m$, and*

$$\sum_{i=1}^{m} g_i(\mathbf{a}^i \cdot \mathbf{x}) = 0,$$

for all $\mathbf{x} \in \mathbb{R}^n$, then each g_i is a polynomial of degree at most $m - 2$.

Remark. In fact the bound on the degree of the polynomial can be further reduced.

Proof. The proof of the proposition is elementary if each of the g_i lies in $C^{m-1}(\mathbb{R})$ (see Lemma 1 in Diaconis and Shahshahani [4]). It goes as follows. Fix $r \in \{1, \ldots, m\}$. For each $j \in \{1, \ldots, m\}$, $j \neq r$, let $\mathbf{c}^j \in \mathbb{R}^n$ satisfy

$$\mathbf{c}^j \cdot \mathbf{a}^j = 0 \quad \text{and} \quad \mathbf{c}^j \cdot \mathbf{a}^r \neq 0.$$

This is possible because the $\mathbf{a}^i$ are pairwise linearly independent. For general $\mathbf{c} = (c_1, \ldots, c_n) \in \mathbb{R}^n$, let

$$D_{\mathbf{c}} = \sum_{s=1}^{n} c_s \frac{\partial}{\partial x_s}.$$

Now

$$D_{\mathbf{c}} g(\mathbf{a} \cdot \mathbf{x}) = (\mathbf{c} \cdot \mathbf{a}) g'(\mathbf{a} \cdot \mathbf{x}).$$

Thus, because each g_i is sufficiently smooth,

$$\begin{aligned} 0 &= \prod_{\substack{j=1 \\ j \neq r}}^{m} D_{\mathbf{c}^j} \sum_{i=1}^{m} g_i(\mathbf{a}^i \cdot \mathbf{x}) \\ &= \sum_{i=1}^{m} \left(\prod_{\substack{j=1 \\ j \neq r}}^{m} (\mathbf{c}^j \cdot \mathbf{a}^i) \right) g_i^{(m-1)}(\mathbf{a}^i \cdot \mathbf{x}) \\ &= \prod_{\substack{j=1 \\ j \neq r}}^{m} (\mathbf{c}^j \cdot \mathbf{a}^r) \, g_r^{(m-1)}(\mathbf{a}^r \cdot \mathbf{x}). \end{aligned}$$

From our choice of the $\mathbf{c}^j$, it follows that

$$g_r^{(m-1)}(\mathbf{a}^r \cdot \mathbf{x}) = 0,$$

for all $\mathbf{x} \in \mathbb{R}^n$. This implies that

$$g_r^{(m-1)}(t) = 0,$$

for all $t \in \mathbb{R}$, and g_r is therefore a polynomial of degree at most $m - 2$.

If the g_i are only in $C(\mathbb{R})$, then the result remains valid. In fact we may even suppose g_i is just locally integrable, i.e., $g_i \in L^1_{\text{loc}}(\mathbb{R})$. To prove this, we use some very basic ideas from distribution theory. Choose $\mathbf{u} \in \mathbb{R}^n$ such that $\mathbf{a}^i \cdot \mathbf{u} = b_i \neq 0$, $i = 1, \ldots, m$. Then

$$0 = \sum_{i=1}^m g_i\big(\mathbf{a}^i \cdot (\mathbf{x} + t\mathbf{u})\big) = \sum_{i=1}^m g_i\big(\mathbf{a}^i \cdot \mathbf{x} + tb_i\big),$$

for every $t \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^n$. Let $\psi \in \mathscr{D} := C_o^\infty(\mathbb{R})$ (the infinitely smooth functions with compact support which we shall call *test functions*). Thus,

$$0 = \int_{-\infty}^{\infty} \sum_{i=1}^m g_i\big(\mathbf{a}^i \cdot \mathbf{x} + tb_i\big)\psi(t)\, dt = \sum_{i=1}^m \frac{1}{b_i} \int_{-\infty}^{\infty} g_i(s)\, \psi\left(\frac{s}{b_i} - \frac{\mathbf{a}^i \cdot \mathbf{x}}{b_i}\right) ds.$$

Fix $r \in \{1, \ldots, m\}$ and let the $\{\mathbf{c}^j\}_{j=1,\, j\neq r}^m$ be as in the previous text. Then

$$0 = \prod_{\substack{j=1 \\ j\neq r}}^m D_{\mathbf{c}^j} \left(\sum_{i=1}^m \frac{1}{b_i} \int_{-\infty}^{\infty} g_i(s)\, \psi\left(\frac{s}{b_i} - \frac{\mathbf{a}^i \cdot \mathbf{x}}{b_i}\right) ds \right),$$

which is the same as

$$0 = \int_{-\infty}^{\infty} g_r(s)\, \psi^{(m-1)}\left(\frac{s}{b_r} - \frac{\mathbf{a}^r \cdot \mathbf{x}}{b_r}\right) ds \times \frac{(-1)^{m-1}}{b_r^m} \prod_{\substack{j=1 \\ j\neq r}}^m (\mathbf{c}^j \cdot \mathbf{a}^r). \quad (2.1)$$

Because ψ is an arbitrary function in $\mathscr{D}$, it follows that

$$\int_{-\infty}^{\infty} g_r(s)\, \phi^{(m-1)}(s)\, ds = 0,$$

for all $\phi \in \mathscr{D}$. It is well known that this implies that g_r is a polynomial of degree at most $m - 2$. For completeness, here is a short proof.

Assume first for simplicity that $g_r = g \in C(\mathbb{R})$ and that

$$\int_{-\infty}^{\infty} g(s)\, \phi^{(m-1)}(s)\, ds = 0, \quad (2.2)$$

for all $\phi \in \mathscr{D}$. For each natural number k, let $\phi_k \in \mathscr{D}$ have support in $[-1/k, 1/k]$, $\phi_k \geq 0$, and $\int_{-\infty}^{\infty} \phi_k(s)\, ds = 1$. Thus $\{\phi_k\}_{k=1}^\infty$ is a sequence of

approximate identities, the elements of such a sequence being characterized by compact support that shrinks to {0} with growing index k, nonnegativity, and unit integral.

Let

$$g_{[k]}(t) = \int_{-\infty}^{\infty} g(s)\,\phi_k(t-s)\,ds.$$

Then $g_{[k]} \in C^\infty(\mathbb{R})$, and

$$g_{[k]}^{(m-1)}(t) = \int_{-\infty}^{\infty} g(s)\,\phi_k^{(m-1)}(t-s)\,ds = 0,$$

for each t. Therefore $g_{[k]}$ is a polynomial of degree at most $m-2$. In addition, $g_{[k]}$ converges uniformly to g as $k \to \infty$ on every finite interval. Therefore g is also a polynomial of degree at most $m-2$.

If g is not continuous but only in $L^1_{\text{loc}}(\mathbb{R})$ and satisfies (2.2) for any test function $\phi \in \mathscr{D}$, we can take Fourier transforms and we can get by Plancherel's identity,

$$\int_{-\infty}^{\infty} \hat{g}(x)\,x^{m-1}\hat{\phi}(x)\,dx = 0,$$

where $\phi \in \mathscr{D}$ and thus $\hat{\phi}$ are still arbitrary. Therefore, in the sense of distributions, $x^{m-1}\hat{g}(x) = 0$ which means $\hat{g}(x) = 0$ everywhere except at zero. Thus Theorem 3.20 in [8] proves the result, namely, that $\hat{g}$ is a finite linear combination of the delta function and its derivatives centered at the origin, their degrees being limited by $m-2$. (We are using here, of course, that the inverse Fourier transform of the kth derivative of the delta function centered at zero is an algebraic polynomial of degree k.)

Therefore we have proved that g must be a polynomial of degree less than $m-1$, almost everywhere. Returning to our g_r, it thus follows that each of the g_r is almost everywhere a polynomial. The points where any of the g_r might *not* be a polynomial of degree less than $m-1$ extend, through the inner product inside $g_r(\mathbf{a}^r \cdot \mathbf{x})$, to a hyperplane orthogonal to $\mathbf{a}^r$. Because the sum over i of these expressions is identically zero and because the directions $\mathbf{a}^i$ are mutually linearly independent, it is a straightforward consequence that each g_r must in fact be a polynomial everywhere. ∎

Remark. This question of uniqueness has been considered by Albertini, Sontag, Maillot [2], Sussman [12], and Fefferman [6] for $\mathscr{N}_m$, especially in the case where $\sigma(x) = \tanh(x)$. If the directions $\mathbf{a}^i$ are pairwise linearly independent, then we can apply Proposition 1. The condition of pairwise linear independence is not, however, natural in this setting. Proposition 1

does permit us to reduce the problem to studying when

$$\sum_{i=1}^{r} c_i \sigma\left(\alpha_i(\mathbf{a}\cdot\mathbf{x}) - b_i\right)$$

is a polynomial of degree at most $m-2$. Here a nonzero $\mathbf{a}$ is fixed; we have grouped all terms with linearly dependent directions. We are interested in conditions on σ, α_i, and b_i which then imply that the c_i are all zero.

Is Proposition 1 valid without any restriction on the g_i? Is it true that if $g_i\colon \mathbb{R}\to\mathbb{R}$, $i=1,\dots,m$, and the sum over the $g_i(\mathbf{a}^i\cdot\mathbf{x})$ vanishes identically for pairwise linearly independent $\mathbf{a}^i$, then each g_i is a polynomial of degree at most $m-2$? The answer, unfortunately, is no. It is well known, see Aczél [1, p. 35] that there exist highly noncontinuous functions $h\colon \mathbb{R}\to\mathbb{R}$ satisfying

$$h(x+y) = h(x) + h(y),$$

for all x and y in $\mathbb{R}$. These h are constructed using "Hamel bases." Thus, for example, the equation,

$$0 = g_1(x_1) + g_2(x_2) + g_3(x_1+x_2)$$

has highly nonpolynomial, noncontinuous solutions $g_1 = g_2 = -g_3 = h$.

The preceding text together with Proposition 1, begs the following interesting question. Namely, if

$$f(\mathbf{x}) = \sum_{i=1}^{m} g_i(\mathbf{a}^i\cdot\mathbf{x}), \tag{2.3}$$

and $f\in C^k(\mathbb{R}^n)$, do there then exist $\tilde{g}_i\in C^l(\mathbb{R})$ for which

$$f(\mathbf{x}) = \sum_{i=1}^{m} \tilde{g}_i(\mathbf{a}^i\cdot\mathbf{x})?$$

In addition, what is the relationship between k and l? For instance, we would wish the previous assertion to hold with $l=k$. The case $k=0$ is of special interest. However, here we will consider only $k\ge m-1$ and we will prove under mild assumptions that indeed $l=k$ in this case.

PROPOSITION 2. *Assume that f has the form* (2.3) *and that $f\in C^k(\mathbb{R}^n)$ for some $k\ge m-1$. If, in addition, $g_i\in L^1_{\mathrm{loc}}(\mathbb{R})$ for each i, then $g_i\in C^k(\mathbb{R})$.*

Proof. We use the $\mathbf{c}^i$ as in the proof of Proposition 1. Let ψ be an arbitrary test function. Then the expression

$$\int_{-\infty}^{\infty} \left(\prod_{\substack{j=1 \\ j \neq r}}^{m} D_{\mathbf{c}^j} \sum_{i=1}^{m} g_i(\mathbf{a}^i \cdot (\mathbf{x} + t\mathbf{u})) \right) \psi(t)\, dt$$

$$= \int_{-\infty}^{\infty} \left(\prod_{\substack{j=1 \\ j \neq r}}^{m} D_{\mathbf{c}^j} (f(\mathbf{x} + t\mathbf{u})) \right) \psi(t)\, dt \tag{2.4}$$

is identical to the right-hand side of (2.1). In other words, the distributional derivatives of f (f is a distribution as well as a function!) of total order $m - 1$ along the directions given by the $\mathbf{c}^j$ are identical to some distributional derivative of order $m - 1$ of g_r (times certain constants). It does not matter here that (2.4) is a univariate integral although a weak formulation of a derivative of the multivariate function f would require an integral against a multivariate test function. This is because f has continuous derivatives anyway. Now, the distributional derivative is equal to the continuous classical derivative of f according to Theorem 7.11 in [8, p. 195] because $f \in C^{m-1}(\mathbb{R}^n)$ and therefore the same must be true for the right-hand side. Thus $g_r \in C^{m-1}(\mathbb{R})$ and

$$\prod_{\substack{j=1 \\ j \neq r}}^{m} D_{\mathbf{c}^j} f(\mathbf{x}) = g_r^{(m-1)}(\mathbf{a}^r \cdot \mathbf{x}) \prod_{\substack{j=1 \\ j \neq r}}^{m} (\mathbf{c}^j \cdot \mathbf{a}^r),$$

where all $\mathbf{a}^r$ and $\mathbf{c}^j$ are known. As such $g_r^{(m-1)} \in C^{k-(m-1)}(\mathbb{R})$, $k \geq m - 1$. The result now follows. ∎

3. $\mathcal{R}(\mathbf{a}^1, \ldots, \mathbf{a}^m)$

A function $f(x, y)$ is of the form,

$$f(x, y) = \sum_{i=1}^{m} g_i(a_i x + b_i y), \qquad (x, y) \in \mathbb{R}^2,$$

for given (a_i, b_i), but unknown continuous g_i, $i = 1, \ldots, m$, if and only if

$$\prod_{i=1}^{m} \left(b_i \frac{\partial}{\partial x} - a_i \frac{\partial}{\partial y} \right) f(x, y) = 0, \tag{3.1}$$

in a distributional sense. This rather simple result is based on the same result in the case $m = 1$ which is straightforward. Note that this provides a

method of identifying $\mathcal{R}(\mathbf{a}^1, \ldots, \mathbf{a}^m)$ if $n = 2$. Unfortunately a simple characterization such as (3.1) does not hold in the case of three or more variables.

How can we determine if a function f (defined on $\mathbb{R}^n$) is of the form (2.3) for some given $\mathbf{a}^1, \ldots, \mathbf{a}^m \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, but unknown continuous $g_1, \ldots, g_m$? That is, how do we characterize $\mathcal{R}(\mathbf{a}^1, \ldots, \mathbf{a}^m)$? One answer may be found in Lin and Pinkus [9].

Let $\mathcal{P}(\mathbf{a}^1, \ldots, \mathbf{a}^m)$ denote the set of polynomials which vanish on all the lines $\{\lambda \mathbf{a}^i : \lambda \in \mathbb{R}\}$, $i = 1, \ldots, m$. This is an ideal.

THEOREM 3 (Lin and Pinkus [9]). *The continuous function* $f \in \mathcal{R}(\mathbf{a}^1, \ldots, \mathbf{a}^m)$ *if and only if*

$$f \in \overline{\operatorname{span}}\{q : q \text{ polynomial}, p(D)q = 0 \text{ for every } p \in \mathcal{P}(\mathbf{a}^1, \ldots, \mathbf{a}^m)\}.$$

This theorem, in and of itself, does not provide a simple method of checking whether a particular function is in $\mathcal{R}(\mathbf{a}^1, \ldots, \mathbf{a}^m)$. However, it follows from the theory of polynomial ideals that one need not check every $p \in \mathcal{P}(\mathbf{a}^1, \ldots, \mathbf{a}^m)$. It is a consequence of Hilbert's basis theorem, c.f. [13, p. 18], that it suffices to consider any set of $p \in \mathcal{P}(\mathbf{a}^1, \ldots, \mathbf{a}^m)$ which are generators for the polynomial ideal $\mathcal{P}(\mathbf{a}^1, \ldots, \mathbf{a}^m)$. Sets of generators are highly nonunique. But it is not difficult to show that there always exists a fairly simple set of generators of cardinality n. We shall, however, not further pursue these ideas here.

In Diaconis and Shahshahani [4] are to be found two additional theorems which characterize $\mathcal{R}(\mathbf{a}^1, \ldots, \mathbf{a}^m)$. The first is stated only for the case $n = 3$. But it is in fact valid for every n.

THEOREM 4 (Diaconis and Shahshahani [4]). *Let* $\mathbf{a}^1, \ldots, \mathbf{a}^m$ *be pairwise linearly independent vectors in* $\mathbb{R}^n$. *Let* Π^r *denote the hyperplane* $\{\mathbf{c} \in \mathbb{R}^n : \mathbf{c} \cdot \mathbf{a}^r = 0\}$, $r = 1, \ldots, m$. *A function* $f \in C^m(\mathbb{R}^n)$ *has the form,*

$$f(\mathbf{x}) = \sum_{i=1}^{m} g_i(\mathbf{a}^i \cdot \mathbf{x}) + P(\mathbf{x}), \tag{3.2}$$

for some polynomial P of degree less than m, if and only if

$$\prod_{i=1}^{m} D_{\mathbf{c}^i} f = 0,$$

for all choices of $\mathbf{c}^i \in \Pi^i$, $i = 1, \ldots, m$.

The free polynomial term in (3.2) is a consequence of Proposition 1. Another much more complicated characterization of $\mathcal{R}(\mathbf{a}^1, \ldots, \mathbf{a}^m)$ for certain particular choices of m (due to Royden) may be found at the end of the paper by Diaconis and Shahshahani [4].

Our goal here is more modest. Assume we are given $f \in \mathcal{R}(\mathbf{a}^1, \ldots, \mathbf{a}^m)$ of the form (2.3). We wish to identify the g_i.

If f is of the form (2.3) with pairwise linearly independent $\mathbf{a}^i$, then, from Proposition 1, these g_i are unique up to polynomials of degree at most $m-2$. How do we determine these g_i? Here is a recipe based on the ideas we have already discussed.

Assume first that $f \in C^{m-1}(\mathbb{R}^n)$ is of the form (2.3). Fix $r \in \{1, \ldots, m\}$ and let the $\{\mathbf{c}^j\}_{j=1}^m$, $j \neq r$, be as in the proof of Proposition 1. In particular, $\mathbf{c}^j \cdot \mathbf{a}^j = 0$ and $\mathbf{c}^j \cdot \mathbf{a}^r \neq 0$. Then we have

$$\begin{aligned}\prod_{\substack{j=1\\ j\neq r}}^m D_{\mathbf{c}^j} f(\mathbf{x}) &= \prod_{\substack{j=1\\ j\neq r}}^m D_{\mathbf{c}^j} \sum_{i=1}^m g_i(\mathbf{a}^i \cdot \mathbf{x}) \\ &= \sum_{i=1}^m \left(\prod_{\substack{j=1\\ j\neq r}}^m (\mathbf{c}^j \cdot \mathbf{a}^i) \right) g_i^{(m-1)}(\mathbf{a}^i \cdot \mathbf{x}) \\ &= \left(\prod_{\substack{j=1\\ j\neq r}}^m (\mathbf{c}^j \cdot \mathbf{a}^r) \right) g_r^{(m-1)}(\mathbf{a}^r \cdot \mathbf{x}).\end{aligned}$$

By assumption we know f and thus the left-hand side of the previous equation. We also know the constant,

$$\prod_{\substack{j=1\\ j\neq r}}^m (\mathbf{c}^j \cdot \mathbf{a}^r),$$

and the vector $\mathbf{a}^r$ itself. Thus this simple method gives us $g_r^{(m-1)}$. In other words we can determine g_r up to a polynomial of degree $m-2$, in agreement with Proposition 1.

This argument also works if f and the g_i are just in $L^1_{\mathrm{loc}}(\mathbb{R}^n)$ and $L^1_{\mathrm{loc}}(\mathbb{R})$, respectively. For this, we argue as in the proof of Proposition 1; i.e., we find a $\mathbf{u}$ as in that proof and we integrate against a test function $\psi \in \mathcal{D}$. Let this test function ψ scaled by b_r^{-1}, i.e., $\psi(b_r^{-1}\cdot)$, be an element from a sequence of approximate identities. In particular, $\int_{-\infty}^{\infty} \psi = b_r^{-1}$. Now, we can identify from the display (2.1) the $(m-1)$st derivative of g_r integrated against that ψ. Denote the result by $g_{r,\psi}^{(m-1)}$. So we have

$$\begin{aligned}&\frac{(-1)^{m-1}}{b_r^m} \prod_{\substack{j=1\\ j\neq r}}^m (\mathbf{c}^j \cdot \mathbf{a}^r) \int_{-\infty}^{\infty} g_r(t)\, \psi^{(m-1)}\left(\frac{t}{b_r} - \frac{\mathbf{a}^r \cdot \mathbf{x}}{b_r} \right) dt \\ &\qquad = \frac{1}{b_r} g_{r,\psi}^{(m-1)}(\mathbf{a}^r \cdot \mathbf{x}) \left(\prod_{\substack{j=1\\ j\neq r}}^m (\mathbf{c}^j \cdot \mathbf{a}^r) \right),\end{aligned}$$

which defines the $g_{r,\psi}^{(m-1)}$. In other words, integrating $(m-1)$ times gives $g_{r,\psi}$, namely, the g_r smoothed by the test function in the aforementioned way, i.e., also scaling ψ by b_r^{-1},

$$\int_{-\infty}^{\infty} g_r(t)\,\psi\left(\frac{t}{b_r} - \frac{\mathbf{a}^r \cdot \mathbf{x}}{b_r}\right) dt.$$

Letting the support of the approximate identity tend to $\{0\}$ while maintaining unit integral gives $g_r(\mathbf{a}^r \cdot \mathbf{x})$ modulo a polynomial of degree $m-2$. The latter is a consequence of our $(m-1)$-fold integration of the derivative.

Because we can do this for each $r \in \{1,\ldots,m\}$, we know that f is contained in the set of functions,

$$\sum_{i=1}^{m} \left(\tilde{g}_i(\mathbf{a}^i \cdot \mathbf{x}) + p_i(\mathbf{a}^i \cdot \mathbf{x})\right),$$

where each p_i is an arbitrary univariate polynomial of degree at most $m-2$ and $\tilde{g}_i$ is such that $\tilde{g}_i^{(m-1)} = g_i^{(m-1)}$, $i = 1,\ldots,m$. That is, for some choice of polynomials $p_i = \tilde{p}_i$ we have $g_i = \tilde{g}_i + \tilde{p}_i$. Alternatively, we may state that

$$f(\mathbf{x}) - \sum_{i=1}^{m} \tilde{g}_i(\mathbf{a}^i \cdot \mathbf{x}) = p(\mathbf{x}) \tag{3.3}$$

is a multivariate polynomial of total degree at most $m-2$ which may be determined and can be written in the form,

$$p(\mathbf{x}) = \sum_{i=1}^{m} \tilde{p}_i(\mathbf{a}^i \cdot \mathbf{x}),$$

for some choice of $\tilde{p}_i$ of degree at most $m-2$. These $\tilde{p}_i$ (and therefore the g_i) are not unique.

There seem to be various methods of determining appropriate $\tilde{p}_i$. Here is one such method. We know that

$$p(\mathbf{x}) = \sum_{i=1}^{m} \tilde{p}_i(\mathbf{a}^i \cdot \mathbf{x}) = \sum_{i=1}^{m} \sum_{j=0}^{m-2} \alpha_{ij}(\mathbf{a}^i \cdot \mathbf{x})^j,$$

for some choice of coefficients α_{ij}. If we can find appropriate α_{ij}, we have constructed suitable $\tilde{p}_i$, $i = 1,\ldots,m$. This can be done in the following fashion: Among the $\{(\mathbf{a}^i \cdot \mathbf{x})^j\}_{i=1,\,j=0}^{m,\,m-2}$, choose a basis, and then write p in terms of this basis. The coefficients α_{ij} for the polynomial in terms of this basis can be found explicitly because we know p. As such we have elucidated one recipe for determining appropriate functions g_i.

4. $\mathscr{R}_m$

Assume $f \in \mathscr{R}_m$, i.e.,

$$f(\mathbf{x}) = \sum_{i=1}^{m} g_i(\mathbf{a}^i \cdot \mathbf{x}), \tag{4.1}$$

for some set of continuous, but unknown, nonzero functions g_i, and unknown directions $\mathbf{a}^i \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, $i = 1, \ldots, m$. Can we determine the g_i and $\mathbf{a}^i$? To be more precise (because there is a problem of uniqueness) we wish to find some g_i and $\mathbf{a}^i$ such that (4.1) holds. For smooth g_i, we can do this, under certain further mild assumptions. In this section we explain how to find directions $\{\mathbf{a}^i\}_{i=1}^m$. We then refer the reader to the previous section for a recipe for determining the $\{g_i\}_{i=1}^m$ based on knowledge of the $\{\mathbf{a}^i\}_{i=1}^m$.

The case $m = 1$ is relatively simple. Let us see how it may be done because it is instructive for the more complicated issues to follow. Assume that

$$f(\mathbf{x}) = g(\mathbf{a} \cdot \mathbf{x}),$$

for some unknown $\mathbf{a} = (a_i)_{i=1}^n$ and g. Assume also that f (and therefore g) is continuously differentiable. Then

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = a_i g'(\mathbf{a} \cdot \mathbf{x}), \qquad i = 1, \ldots, n.$$

Taking ratios we have

$$\frac{\dfrac{\partial f}{\partial x_i}(\mathbf{x})}{\dfrac{\partial f}{\partial x_j}(\mathbf{x})} = \frac{a_i}{a_j}, \qquad i, j = 1, \ldots, n,$$

so long as a_j and $g'(\mathbf{a} \cdot \mathbf{x})$ do not vanish. The right-hand side is independent of $\mathbf{x}$ for every choice of $i, j \in \{1, \ldots, n\}$. Note that $\mathbf{a}$ and g are not uniquely determined. Specifically, we can always replace $\mathbf{a}$ by $c\mathbf{a}$ for any constant $c \neq 0$, and appropriately alter g. Thus, knowing all the ratios a_i/a_j effectively determines $\mathbf{a}$. Once given $\mathbf{a}$, we obtain g from Section 3.

The generalization to $m > 1$ is more complicated. We will use the following result, the proof of which may be found in Section 2 of Buhmann and Pinkus [3], see Theorem 1 and Corollary 2 therein.

THEOREM 5. *Assume that we are given numbers* $\{b_k\}_{k=0}^{2m-1}$ *which satisfy*

$$\sum_{i=1}^{m} c_i (d_i)^k = b_k, \qquad k = 0, \ldots, 2m-1, \tag{4.2}$$

for unknown nonzero $\{c_i\}_{i=1}^m$ *and unknown distinct* $\{d_i\}_{i=1}^m$. *The* d_i *are uniquely determined as follows. The function*

$$B(x) = \det \begin{pmatrix} b_0 & \cdots & b_{m-1} & b_m \\ \vdots & \ddots & \vdots & \vdots \\ b_{m-1} & \cdots & b_{2m-2} & b_{2m-1} \\ 1 & \cdots & x^{m-1} & x^m \end{pmatrix} \tag{4.3}$$

is a polynomial of exact degree m. *The* d_i, $i = 1, \ldots, m$, *are its* m *distinct zeros. The* c_i, $i = 1, \ldots, m$, *are easily calculated from the linear equations* (4.2) *once we know the* d_i, $i = 1, \ldots, m$.

Remark. Note that we are not saying that for every choice of $\{b_k\}_{k=0}^{2m-1}$ the system (4.2) has a solution. A more complete and detailed examination of this problem may be found in [3].

We now give a recipe for determining the $\{\mathbf{a}^i\}_{i=1}^m$ in (4.1), based on various (not unreasonable) assumptions. The basic assumptions which are used throughout are that the $\{\mathbf{a}^i\}_{i=1}^m$ are pairwise distinct, that each g_i is $C^{2m-1}(\mathbb{R})$ in some neighbourhood of 0, and that $g_i^{(2m-1)}(0) \neq 0$, $i = 1, \ldots, m$.

Let $\mathbf{c} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$. Assume that $\mathbf{c}$ has been chosen in such a way that the values

$$\left\{ (\mathbf{c} \cdot \mathbf{a}^i)^{2m-1} g_i^{(2m-1)}(0) \right\}_{i=1}^m \tag{4.4}$$

are nonzero and distinct. This is always possible with a suitable $\mathbf{c}$, because of the linear independence of the $\mathbf{a}^i$.

For each $\mathbf{d} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ and $k = 0, 1, \ldots, 2m-1$,

$$\begin{aligned} \left(D_{\mathbf{c}}^{2m-1-k} D_{\mathbf{d}}^k f\right)(\mathbf{0}) &= \sum_{i=1}^{m} (\mathbf{c} \cdot \mathbf{a}^i)^{2m-1-k} (\mathbf{d} \cdot \mathbf{a}^i)^k g_i^{(2m-1)}(0) \\ &= \sum_{i=1}^{m} \left[(\mathbf{c} \cdot \mathbf{a}^i)^{2m-1} g_i^{(2m-1)}(0) \right] \left[\frac{(\mathbf{d} \cdot \mathbf{a}^i)}{(\mathbf{c} \cdot \mathbf{a}^i)} \right]^k. \end{aligned}$$

If the

$$\left\{ \frac{(\mathbf{d} \cdot \mathbf{a}^i)}{(\mathbf{c} \cdot \mathbf{a}^i)} \right\}_{i=1}^{m} \tag{4.5}$$

are distinct, then it follows from Theorem 5 that they may be uniquely determined. Taking n linearly independent $\mathbf{d} = \mathbf{d}^j$, $j = 1, \ldots, n$, which satisfy the preceding, we obtain for each $i \in \{1, \ldots, m\}$ the n values,

$$\left\{ \frac{(\mathbf{d}^j \cdot \mathbf{a}^i)}{(\mathbf{c} \cdot \mathbf{a}^i)} \right\}, \qquad j = 1, \ldots, n. \tag{4.6}$$

This determines the

$$\frac{\mathbf{a}^i}{(\mathbf{c} \cdot \mathbf{a}^i)}, \qquad i = 1, \ldots, n;$$

i.e., it determines $\mathbf{a}^i$ up to multiplication by a nonzero, finite constant (in this case $(\mathbf{c} \cdot \mathbf{a}^i)^{-1}$). Thus the $\mathbf{a}^i$ are totally determined (see the remark in the case $m = 1$). This is our general ''recipe.''

Let us now consider certain of our requirements in further detail. (See also the discussion in the proof of Theorem 3 of Buhmann and Pinkus [3].) We need to be able to tell whether, for a given $\mathbf{d} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, the values (4.5) are distinct. This is straightforward, because under our other assumptions these values are distinct if and only if the associated polynomial $B(x)$ (in (4.3)) is of exact degree m and has this many distinct zeros. Why do we need the assumption that the coefficients in (4.4) be distinct, as well as being nonzero? The distinctness is not used in Theorem 5. It is, however, used for labeling. If two coefficients have the same values then we do not know how to assign the associated expressions (4.6) and thus we determine the $\mathbf{a}^i$. If the values in (4.4) are not distinct, we can alter the $\mathbf{c}$. From Theorem 5 we get the values of these coefficients, and so it is easily determined as to whether these coefficients are distinct. (The other option is to try all the different possibilities of assigning the terms (4.6) to the appropriate coefficients, and then matching the results obtained with the original f.) Further, we may weaken the demand that $g_i^{(2m-1)}(0) \neq 0$, $i = 1, \ldots, m$, by admitting any shifts away from the origin. Finally, the smoothness condition on g_i may be weakened by convolving g_i with a test function ψ and by considering $g_{i,\psi}$ instead of g_i, as in the previous section. Thus the condition $g_i^{(2m-1)}(0) \neq 0$ is replaced by demanding that there *exists* a test function ψ such that $g_{i,\psi}^{(2m-1)}$ does not vanish at the origin for all indices i.

5. $\mathcal{N}_m$

We are given $\sigma \in C(\mathbb{R})$ and the set $\mathcal{N}_m$ as defined in the Introduction. Because we do not know the $\mathbf{a}^i$, this is a specific subset of $\mathcal{R}_m$. We can and will assume, by using the methods of the previous section, that we are able to identify the $\mathbf{a}^i$ up to multiplication by constants. That is, $\mathbf{a}^i = d_i\tilde{\mathbf{a}}^i$ for some determined $\tilde{\mathbf{a}}^i$ but undetermined d_i. (Note that we must here again assume that the $\mathbf{a}^i$ are pairwise linearly independent.)

Furthermore we also assume that, using the methods of Section 3, we can in fact identify

$$g_r^{(m-1)}(\mathbf{a}^r \cdot \mathbf{x}) = c_r\sigma^{(m-1)}(\mathbf{a}^r \cdot \mathbf{x} - b_r) = c_r\sigma^{(m-1)}(d_r(\tilde{\mathbf{a}}^r \cdot \mathbf{x}) - b_r),$$

for each r. Thus, knowing already $\tilde{\mathbf{a}}^r$, we have reduced our problem to the following. Given σ and m and

$$c\sigma^{(m-1)}(dt - b), \tag{5.1}$$

find the constants c, b, and d.

Unfortunately, we know of no general methods for solving this problem. Numerous ad hoc methods present themselves depending on the particular σ.

For instance, assume σ is bounded and $\sigma(t) \to \alpha_{\pm}$ when $t \to \pm\infty$, where $\alpha_+ \neq \alpha_-$. (This is a typical case in neural network applications where usually $\alpha_- = 0$, $\alpha_+ = 1$.) Because σ is bounded, we can find

$$\tilde{c}\sigma(dt - b) + a, \tag{5.2}$$

from (5.1) by integration, where the uncertainty regarding the polynomial of degree at most $m - 2$ that comes from the integration is reduced to a constant a (because other nonconstant polynomials are ruled out by boundedness). Now, letting $t \to \pm\infty$ we obtain the two values f_+ and f_-, one of which corresponds to $\tilde{c}\alpha_+ + a$ and the other which corresponds to $\tilde{c}\alpha_- + a$, depending on the sign of d. Thus we obtain two options for $(\tilde{c}, a)$.

If, in addition, σ is strictly monotone, we can find b and d from (5.2) by evaluating at suitable t_1 and t_2. This will give us one or two possible choices for the constants a, b, c, and d. However, it should be noted that it is very possible that these constants are not uniquely determined.

ACKNOWLEDGMENT

We thank Joachim Stöckler and Burkhard Lenze for pointing out the oversight in [3].

REFERENCES

1. J. Aczél, ''Functional Equations and Their Applications,'' Academic Press, New York, 1966.
2. F. Albertini, E. D. Sontag, and V. Maillot, Uniqueness of weights for neural networks, *in* ''Artificial Neural Networks for Speech and Vision,'' (R. J. Mammone, Ed.), pp. 113–125, Chapman and Hall, London/New York, 1993.
3. M. D. Buhmann and A. Pinkus, On a recovery problem, *Ann. Num. Math.* **4** (1997), 129–142.
4. P. Diaconis and M. Shahshahani, On nonlinear functions of linear combinations, *SIAM J. Sci. Statist. Comput.* **5** (1984), 175–191.
5. D. L. Donoho and I. M. Johnstone, Projection-based approximation and a duality method with kernel methods, *Ann. Statist.* **17** (1989), 58–106.
6. C. Fefferman, Reconstructing a neural net from its output, *Rev. Mat. Iberoamericana* **10** (1994), 507–555.
7. F. John, ''Plane Waves and Spherical Means Applied to Partial Differential Equations,'' Interscience, New York, 1955.
8. D. S. Jones, ''The Theory of Generalised Functions,'' Cambridge Univ. Press, Cambridge, U.K., 1982.
9. V. Ya. Lin and A. Pinkus, Fundamentality of ridge functions, *J. Approx. Theory* **75** (1993), 295–311.
10. B. F. Logan and L. A. Shepp, Optimal reconstruction of a function from its projections, *Duke Math. J.* **42** (1975), 645–659.
11. A. Pinkus, Approximation by ridge functions, *in* ''Surface Fitting and Multiresolution Methods'' (A. Le Méhauté, C. Rabut, and L. L. Schumaker, Eds.), pp. 279–292, Vanderbilt Univ. Press, Nashville, 1997.
12. H. J. Sussman, Uniqueness of the weights for minimal feedforward nets with a given input–output map, *Neural Networks* **5** (1992), 589–593.
13. B. L. van der Waerden, ''Moderne Algebra Band II,'' Springer-Verlag, Berlin, 1931.