Approximating By Ridge Functions

Allan Pinkus

Abstract. This paper surveys certain aspects of the study of ridge functions. We hope it will also encourage some readers to consider researching problems in this area. After first explaining what ridge functions are and giving various motivations for their study, we turn to the problem of presenting algorithms for approximating by ridge functions. We then touch upon the topic of determining the degree of approximation by ridge functions, and that of recognizing functions which are linear combinations of ridge functions.

§1. Introduction

This short paper is an introduction to certain aspects of the study of ridge functions. After first explaining what ridge functions are and giving various motivations for their study, we turn to the problem of presenting algorithms for approximating by ridge functions. We then touch upon the topic of determining the degree of approximation by ridge functions, and that of recognizing when we have a function which is a linear combination of ridge functions.

This paper is short for the very simple reason that at present rather little is known. Our goal here is to present a partial review, to try to convince you, the reader, of the significance of the subject, and to encourage some of you to consider researching problems in this area.

A `Ridge Function`, in its simplest form, is a multivariate function

$$f : \mathbb{R}^n \to \mathbb{R}$$

of the form

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

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`. We should also

Surface Fitting and Multiresolution Methods
A. Le Méhauté, C. Rabut, and L. L. Schumaker (eds.), pp. 1–14.

ISBN 0-8265-1294-1.

regard a ridge function as a composition of one of the simplest multivariate functions, namely the inner product $\mathbf{a}\cdot\mathbf{x}$, with an arbitrary univariate function g.

Ridge functions appear in various areas and under various guises. We will review some of these shortly. Note that we often use ridge functions without calling them by name. The functions $e^{\mathbf{a}\cdot\mathbf{x}}$, $e^{i\mathbf{a}\cdot\mathbf{x}}$, and $(\mathbf{a}\cdot\mathbf{x})^k$ are well-known to us.

What are the approximation sets we wish to consider? One can of course choose a particular univariate function g, and then vary over a number of directions. That is, consider for a fixed g the space

$$\mathcal{G}_r(g) = \left\{ \sum_{i=1}^r \alpha_i g(\mathbf{a}^i \cdot \mathbf{x}) : \alpha_i \in \mathbb{R},\, \mathbf{a}^i \in \mathbb{R}^n \setminus \{\mathbf{0}\},\, i = 1, \ldots, r \right\}.$$

(Note that this is not a linear space.) However, of the various possibilities, this is the one which we will not consider.

Rather we will look at two sets of ridge functions. The first is given by

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

That is, we fix a finite number of directions and 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

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

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

§2. Motivations

The name "ridge function" is rather recent. However these functions have been considered for some time now but under the name of `Plane Waves`. See, for example, the well-known book by Fritz John "Plane Waves and Spherical Means Applied to Practical Differential Equations" [7]. Plane waves are also discussed in the classic Courant and Hilbert, "Methods of Mathematical Physics, Vol. II", [2]. In general, linear combinations of ridge functions with fixed directions occur in the study of hyperbolic constant coefficient partial differential equations. For example, assume that the (a_i, b_i) are pairwise linearly independent vectors in $\mathbb{R}^2, i = 1, \ldots, r$. Then the general "solution" of the homogeneous partial differential equation

$$\prod_{i=1}^r \left(a_i \frac{\partial}{\partial x} + b_i \frac{\partial}{\partial y} \right) f = 0 \tag{2.1}$$

are all functions of the form

$$f(x,y) = \sum_{i=1}^{r} g_i(b_i x - a_i y)\,, \tag{2.2}$$

for arbitrary g_i. (Here is one way of determining whether an arbitrary function f is of the form (2.2) for given (a_i, b_i), $i = 1, \ldots, r$, and arbitrary g_i. Check whether it satisfies the homogeneous equation (2.1).)

The term "ridge function" was coined in a 1975 paper by Logan and Shepp [13]. This was a seminal paper in computerized tomography. In tomography, or at least in tomography as the theory was initially constructed in the early 80's, ridge functions were basic. However, they were used in a somewhat special form. The general idea therein was to take a given multivariate function $H(\mathbf{x})$ and to try to reconstruct it from the values of its integrals along certain planes or lines. If we are given planes or lines which are all parallels of a given plane or line, then these integrals can be considered as a ridge function (or some fairly simple generalizations thereof) based on a single direction. The idea is to reconstruct $H(\mathbf{x})$ from a set of such ridge functions. Thus the ridge functions so obtained are *not* arbitrary. These functions all come from one function H, and thus must satisfy certain moment type conditions. Any reconstruction algorithm must take into account the special nature of the data. Ridge functions also enter tomography from a slightly different direction. Logan and Shepp showed how ridge functions solve a "minimum norm approximation problem" in $L^2(D)$ (D the disk in $\mathbb{R}^2$) connected with best reconstruction of a function based on projection type data.

Ridge functions and ridge function approximation are studied in Statistics. There they often go under the name of `Projection Pursuit`. The interested reader may consult Friedman and Stuetzle [5], Huber [6], and Donoho and Johnstone [4]. Projection pursuit algorithms approximate a function of n variables by functions of the form

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

where the $\mathbf{a}^i$ and g_i are the variables. In other words, projection pursuit algorithms are interested in approximation from $\mathcal{R}_r$. The idea here is to "reduce dimension" and thus bypass the "curse of dimensionality". The $\mathbf{a}^i \cdot \mathbf{x}$ is considered as a projection of $\mathbf{x}$. The directions $\mathbf{a}^i$ are chosen to "pick out the salient features". We will later consider some of the ideas and algorithms developed in the theory of projection pursuit.

The past five years have seen an explosion of interest in the subject of `(Artificial) Neural Networks`. This is a highly interdisciplinary area of research with a selection of problems and models which touch upon various mathematical questions. One of the popular models is that of a *multilayer feedforward neural net* with input, hidden, and output layers. Stripping away the terminology of neural networks, the simplest case (which is that of one

hidden layer, r processing units and one output) considers, in mathematical terms, functions of the form

$$\sum_{i=1}^{r} \alpha_i \sigma(\sum_{j=1}^{n} w_{ij} x_j + \theta_i),$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is some given fixed univariate function (called the *activation function*). In this model, which is just one of many, we are in general permitted to vary over the w_{ij} and θ_i. Note that for each $\theta \in \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ the function

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

is a ridge function. Thus a lower bound on the degree of approximation by such functions is given by a lower bound on the degree of approximation by ridge functions, i.e., from $\mathcal{R}_r$. And, in fact, there exist σ for which the "degrees of approximation" are essentially the same.

Finally, ridge functions are and should be of interest to the approximation theorist. The basic idea and concept is a simple one. We wish to approximate complicated functions by simple functions. Multivariate functions are complicated things. We look for various classes of simpler functions, and ridge functions are one such class. The questions we ask when approximating by ridge functions, or by other "candidates" for approximation, are straightforward questions. Can we approximate (density)? How well can it be done (degree of approximation)? How do we go about approximating (algorithms for approximation)? Can it be done quickly and efficiently?

We now have a fairly complete understanding of the first question, i.e., that of density. The interested reader should look at Vostrocev and Kreines [18], Lin and Pinkus [12], and Kroo [10] for various results in this direction. However, the remaining questions are as yet unanswered.

§3. Approximation Algorithms

A. Fixed Directions

We are interested in methods of approximating from

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

in some norm and over some domain in $\mathbb{R}^n$.

There seems to be one essential method of approximation. This method goes under different names in different settings. Particular variants have been called, among other things, the Von Neumann Alternating Algorithm, the Cyclic Coordinate Algorithm, the Schwarz Domain Decomposition method, and the Diliberto-Straus Algorithm. Other variants may be found in the tomography literature. The idea in our case is the following:

Set

$$\mathcal{M}_i = \{g(\mathbf{a}^i \cdot \mathbf{x}) : g \in X_i\}$$

(for the appropriate space of functions X_i). Let P_i be a best approximation operator to $\mathcal{M}_i$, i.e., to each f the element $P_i f$ is a best approximation to f from $\mathcal{M}_i$. (This demand on P_i may at times be weakened.) The major assumption underlying this method is that each P_i is fairly easily computable. Now set

$$E = (I - P_r) \cdots (I - P_2)(I - P_1).$$

That is, we first find the error in approximating from $\mathcal{M}_1$, then the error in approximating this new function from $\mathcal{M}_2$, etc.... E represents a one time cycle of the process through the $\mathcal{M}_i$. In other words

$$Ef = f - g_1 - g_2 - \cdots - g_r,$$

where g_j is a best approximation to $E - g_1 - g_2 - \cdots - g_{j-1}$ from $\mathcal{M}_j$, $j = 1, \ldots, r$.

The algorithm then iterates E. That is, we consider

$$\lim_{m \to \infty} E^m f \, .$$

The hope and expectation is that this algorithm will converge, and converge to

$$f - g^*,$$

where g^* is a best approximation to f from

$$\mathcal{M}_1 + \cdots + \mathcal{M}_r = \mathcal{R}(\mathbf{a}^1, \ldots, \mathbf{a}^r) \, .$$

(One variant of this algorithm is to consider at each step all the $\{I - P_i\}_{i=1}^r$ applied to the existing error, and then choose the "best" one.)

Consider the uniform norm on a rectangular domain in $\mathbb{R}^2$ with sides parallel to the axes. If $r = 2$ and the $\mathbf{a}^i$ are the unit directions, then this algorithm is essentially a specific case of the Diliberto-Straus Algorithm, and it converges to the desired quantity. It will, with the proper assumptions, converge to the desired quantity for any two arbitrary directions. However, in general it need not converge to the correct quantity if the number of directions r is at least three. It seems that if $r \geq 3$ then this algorithm, in the uniform norm, may prematurely stop. That is, it thinks that it is at a best global approximation from $\mathcal{M}_1 + \cdots + \mathcal{M}_r$, and it is not. It is at a best approximation from each $\mathcal{M}_i$, $i = 1, \ldots, r$, separately. In this case this does not imply that it is at a best approximation from $\mathcal{M}_1 + \cdots + \mathcal{M}_r$. It would be interesting to determine conditions under which this algorithm is valid in this setting. Moreover to date there are no known algorithms for finding best approximations from $\mathcal{R}(\mathbf{a}^1, \ldots, \mathbf{a}^r)$ in the uniform norm (or in the L^1-norm).

Having dwelt on the defects of this "algorithm", let us now consider its advantages. The following two results are valid. Both may be proven using the methods of proof found in Chapter 3 of Light, Cheney [11].

Theorem 1. *Assume each $\mathcal{M}_i$, $i = 1, \ldots, r$, is a closed linear subspace in a uniformly convex and smooth Banach space X. Let P_i denote the best approximation operator from $\mathcal{M}_i$, $i = 1, \ldots, r$. Assume, in addition, that*

$$\mathcal{M}_1 + \cdots + \mathcal{M}_r$$

is closed. Then the algorithm described above converges as desired.

A few comments with regards to the assumptions. The closure of each $\mathcal{M}_i$ is necessary in order for a best approximation to exist. The uniform convexity implies that each $\mathcal{M}_i$ is an existence and unicity space (proximinal and Chebyshev). As such the operator P_i is well-defined. To the best of my knowledge, it is not known whether the assumption concerning the closure of $\mathcal{M}_1 + \cdots + \mathcal{M}_r$ is in fact necessary. Claims have been made but not properly substantiated. However this property is used in the proof of this theorem. The closure question is far from trivial. See, for example, Petersen, Smith, Solmon [16], Boman [1], and references therein.

The second result is a strengthening of Theorem 1 in the Hilbert space setting.

Theorem 2. *Assume each $\mathcal{M}_i$, $i = 1, \ldots, r$, is a closed linear subspace of a Hilbert space H. Let P_i denote the best approximation operator from $\mathcal{M}_i$, $i = 1, \ldots, r$. Then the algorithm, as described above, converges to the best approximation from*

$$\overline{\mathcal{M}_1 + \cdots + \mathcal{M}_r}.$$

Furthermore, if $\mathcal{M}_1 + \cdots + \mathcal{M}_r$ is closed, then the rate of convergence is geometric.

If $f \in H$, g^* is the best approximation to f from $\mathcal{M}_1 + \cdots + \mathcal{M}_r$, and $E^m f = f - g_m$, then we say that the rate of convergence is `geometric` if there exist constants C and θ, $0 \le \theta < 1$, such that

$$\|g_m - g^*\| \le C\theta^m.$$

The algorithm we have been considering is useful only if a best approximation from each $\mathcal{M}_i$ is (easily) calculable. We claim that this is not an unreasonable assumption in our setting. Let

$$\mathcal{M} = \{g(\mathbf{a} \cdot \mathbf{x}) : g \in X_{\mathcal{M}}\}$$

be a closed linear subspace of X. Given $f \in X$ we are essentially looking for g^* such that on the hyperplane $\mathbf{a} \cdot \mathbf{x} = c$ the constant $g^*(c)$ is a best approximation to f (for c as applicable). Finding a best approximation by a constant is not an impossible task. However, what may not be *a priori* guaranteed is that the resulting g^* be necessarily in the appropriate $X_{\mathcal{M}}$.

As an example of how we can calculate the best approximation operator P_i, let $1 < p < \infty$, $d\mu$ be a finite positive measure on $\mathbb{R}^n$, and consider

$X = L^p(\mathbb{R}^n, d\mu)$. Given $f \in L^p(\mathbb{R}, d\mu)$ and $\mathbf{a} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, how does one find a best ridge function $g^*(\mathbf{a} \cdot \mathbf{x})$ with which to approximate f in $L^p(\mathbb{R}^n, d\mu)$?

Since g^* is constant on the hyperplane $\mathbf{a} \cdot \mathbf{x} = c$, one takes its value thereon to be the constant of best approximation to f on this hyperplane. In the $L^p(\mathbb{R}^n, d\mu)$ norm, this is the unique constant K for which

$$\int_{\mathbf{a}\cdot\mathbf{x}=c} |f(\mathbf{x}) - K|^{p-1} sgn(f(\mathbf{x}) - K) d\tilde{\mu}(\mathbf{x}) = 0,$$

where $d\tilde{\mu}$ is the appropriate "restriction" of the measure $d\mu$ to this hyperplane.

Note that for $p = 2$ this reduces to

$$\int_{\mathbf{a}\cdot\mathbf{x}=c} (f(\mathbf{x}) - K) d\tilde{\mu}(\mathbf{x}) = 0\,.$$

Thus

$$g^*(c) = \frac{\int_{\mathbf{a}\cdot\mathbf{x}=c} f(\mathbf{x}) d\tilde{\mu}(\mathbf{x})}{\int_{\mathbf{a}\cdot\mathbf{x}=c} d\tilde{\mu}(x)}\,,$$

i.e., the "average" value on the hyperplane. A simple calculation also shows, via the orthogonality, that

$$\|f - g^*(\mathbf{a}\cdot)\|^2_{L^2(\mathbb{R}^n, d\mu)} = \|f\|^2_{L^2(\mathbb{R}^n, d\mu)} - \|g^*(\mathbf{a}\cdot)\|^2_{L^2(\mathbb{R}^n, d\mu)}\,.$$

This formula will prove useful in what follows.

B. Variable Directions

When considering $\mathcal{R}(\mathbf{a}^1, \ldots, \mathbf{a}^r)$, we might ask ourselves if our true aim is to find a best approximation from this fixed subspace $\mathcal{R}(\mathbf{a}^1, \ldots, \mathbf{a}^r)$? An odd question to say the least. But the answer is sometimes yes and sometimes no. Sometimes we are interested in the fixed subspace $\mathcal{R}(\mathbf{a}^1, \ldots, \mathbf{a}^r)$, in and of itself. We want a best (or good) approximation and nothing more. Often we use exactly the same terminology when we are really interested in something a bit different, namely in a process wherein r is also tending to infinity. (If this sounds strange, think of the problem of best approximation from the space of polynomials of degree n. The same dichotomy appears here, especially from an algorithmic point of view. Are we interested in p_n, the best polynomial approximant of degree n to a fixed function, or are we interested in a sequence $\{p_n\}_{n=0}^{\infty}$ of good approximations (each p_n of degree n) which approximate our function as $n \to \infty$, and good methods of constructing the sequence?)

Assume that we are given a sequence of vectors $\{\mathbf{a}^i\}_{1=1}^{\infty}$, $\mathbf{a}^i \in \mathbb{R}^n \setminus \{\mathbf{0}\}$. Let

$$\mathcal{A} = \text{span}\left\{g_i(\mathbf{a}^i \cdot \mathbf{x}) : g_i \in X_i,\ i = 1, 2, \ldots\right\}$$

(for the appropriate space of functions X_i).

We are interested in finding better and better approximations to a given function f from $\mathcal{A}$. Here the directions are fixed and not variable. But this

is, nevertheless, not the same problem we discussed previously. We could consider the problem

$$\lim_{r\to\infty} \inf_{g\in\mathcal{R}(\mathbf{a}^1,\ldots,\mathbf{a}^r)} \|f-g\|\,.$$

However, the order of the $\mathbf{a}^i$ really plays no particular role in ridge function approximation. As such we should rather consider

$$\lim_{r\to\infty} \inf_{1\le j_1<\cdots<j_r} \inf_{g\in\mathcal{R}(\mathbf{a}^{j_1},\ldots,\mathbf{a}^{j_r})} \|f-g\|\,,$$

or some variant thereof. In this sense the above problem of approximation with fixed directions is very much similar to the problem of approximation from the set of ridge functions with variable directions, i.e., approximation from

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

The difference is that in the first case the set of permissible directions is limited. However, many of the ideas, with regard to constructing good sequences of approximants, are the same.

The set $\mathcal{R}_r$ is non-linear because of the variable directions $\mathbf{a}^i$. This non-linearity causes very serious problems in any attempt to construct good approximations. The only work, of which I am aware, in this area is due to statisticians. (That is to say, it is a projection pursuit algorithm.) It was suggested by Friedman, Stuetzle [5] in 1981. It does not attempt to construct a best approximation from $\mathcal{R}_r$, but gives a method of obtaining a sequence of $g^r\in\mathcal{R}_r$ which converge to a given f.

This framework is $L^2(\mathbb{R}^n,d\mu)$ where, as previously, $d\mu$ is a finite positive measure on $\mathbb{R}^n$. Assume we are given $f\in L^2(\mathbb{R}^n,d\mu)$, and have already determined $g_1^*,\ldots,g_{r-1}^*$ and directions $\mathbf{b}^1,\ldots,\mathbf{b}^{r-1}$. That is, we are given

$$f(\mathbf{x})-\sum_{i=1}^{r-1} g_i^*(\mathbf{b}^i\cdot\mathbf{x})\,.$$

How can we choose a function g_r^* and direction $\mathbf{b}^r$ so as to minimize

$$\left\| \left((f-\sum_{i=1}^{r-1} g_i^*(\mathbf{b}^i\cdot\,) \right) - g(\mathbf{a}\cdot\,) \right\|_{L^2(\mathbb{R}^n,d\mu)}$$

as we vary over g and $\mathbf{a}$?

From the result of the previous subsection, we know that for a given fixed direction $\mathbf{a}$ (which we can always assume to be a unit vector), the optimal $g_{\mathbf{a}}$ is given by

$$g_{\mathbf{a}}(c)=\frac{\int_{\mathbf{a}\cdot\mathbf{x}=c} f_{r-1}(x)d\tilde{\mu}(\mathbf{x})}{\int_{\mathbf{a}\cdot\mathbf{x}=c} d\tilde{\mu}(\mathbf{x})}\,,$$

where $d\tilde{\mu}$ is the appropriate restriction of the measure $d\mu$ to the hyperplane $\mathbf{a} \cdot \mathbf{x} = c$, and

$$f_{r-1}(\mathbf{x}) = f(\mathbf{x}) - \sum_{i=1}^{r-1} g_i^*(\mathbf{b}^i \cdot \mathbf{x}) .$$

Furthermore, as noted,

$$\|f_{r-1} - g_{\mathbf{a}}(\mathbf{a}\cdot)\|^2_{L^2(\mathbb{R}^n, d\mu)} = \|f_{r-1}\|^2_{L^2(\mathbb{R}^n, d\mu)} - \|g_{\mathbf{a}}(\mathbf{a}\cdot)\|^2_{L^2(\mathbb{R}^n, d\mu)} .$$

Thus minimizing the resulting error is equivalent to maximizing

$$\|g_{\mathbf{a}}(\mathbf{a}\cdot)\|_{L^2(\mathbb{R}^n, d\mu)}$$

over the (compact) set of unit vectors. As such, theoretically it is doable. From a computational point of view it is difficult.

This is a "greedy" algorithm. At the rth step it looks at the best of the possible approximants to the upgraded f_{r-1}. It does not look for the best approximation from $\mathcal{R}_r$.

The algorithm converges. That is,

$$\lim_{r\to\infty} \|f_r\|_{L^2(\mathbb{R}^n, d\mu)} = 0 .$$

However, because of the non-linearity it is worth trying to mollify the demands of the algorithm while maintaining its convergence. In 1987, L. K. Jones [8] proved the following result.

Let $0 < \rho < 1$ be fixed. Assume that at the rth step we have obtained

$$g_r^*(\mathbf{b}^r \cdot \mathbf{x}),$$

where g_r^* is optimal for $\mathbf{b}^r$ (i.e., $g_r^* = g_{\mathbf{b}^r}$) but $\mathbf{b}^r$ is not quite an optimal direction. Assume, however, that we do know that

$$\|g_r^*(\mathbf{b}^r\cdot)\|_{L^2(\mathbb{R}^n, d\mu)} \geq \rho \sup_{\|\mathbf{a}\|=1} \|g_{\mathbf{a}}(\mathbf{a}\cdot)\|_{L^2(\mathbb{R}^n, d\mu)} .$$

Then

$$\lim_{r\to\infty} \|f - \sum_{i=1}^{r} g_i^*(\mathbf{b}^i\cdot)\|_{L^2(\mathbb{R}^n, d\mu)} = 0 .$$

§4. Degree of Approximation

There is very, very little known about degree of approximation by ridge functions. The problems are both interesting and important. For example, recall that in the model for multilayer feedforward neural nets with one hidden layer a lower bound on the degree of approximation is given by the lower bound on the degree of approximation by ridge functions (and this lower bound may be attained).

One of the major issues (and this is certainly the same for any reasonable approximating set in many variables) is in deciding what criteria to use in measuring "goodness of approximation". This is intertwined with the problem of identifying classes of functions which are better and more appropriately approximated by ridge functions. The usual classes considered are the Sobolev spaces, determined by smoothness conditions. To understate the case somewhat, it is far from clear that this is a reasonable class in this setting. It seems, using the criteria of Sobolev spaces, that $\mathcal{R}(\mathbf{a}^1,\ldots,\mathbf{a}^r)$ (and $\mathcal{R}_r$) attain approximation orders (in the worst case) essentially the same as those of the polynomials they contain, in the following sense.

In $\mathbb{R}^n$ the dimension of the space π_k^n of polynomials of degree at most k is $\binom{n+k}{k}$. The space H_k^n of homogeneous polynomials of degree k has dimension $\binom{n+k-1}{k}$. Set $r = \dim H_k^n = \binom{n+k-1}{k} \asymp k^{n-1}$, and let the $\mathbf{a}^1,\ldots,\mathbf{a}^r$ be chosen so that

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

(The set of $\{\mathbf{a}^1,\ldots,\mathbf{a}^r\}$ for which this assumption does not hold is rather sparse.) Then

$$\pi_k^n \subset \mathcal{R}(\mathbf{a}^1,\ldots,\mathbf{a}^r)\,,$$

and

$$\pi_{k+1}^n \not\subset \mathcal{R}(\mathbf{a}^1,\ldots,\mathbf{a}^r)\,.$$

(Note that $\mathcal{R}(\mathbf{a}^1,\ldots,\mathbf{a}^r)$ contains π_k^n and not only H_k^n. This is an important difference.)

Let $B_p^{s,n}$ denote the usual subset of the Sobolev space consisting of all functions defined on $K = [0,1]^n$ which have a.e. in their domain of definition all partial derivatives up to order s and such that

$$\|f\|_{L^p(K)} + \sum_{|\mathbf{k}|\le s} \|D^{\mathbf{k}} f\|_{L^p(K)} \le 1\,,$$

where $|\mathbf{k}| = k_1 + \cdots + k_n$ and

$$D^{\mathbf{k}} f = \frac{\partial^{|\mathbf{k}|} f}{\partial x_1^{k_1} \cdots \partial x_n^{k_n}}\,.$$

Now it is well-known that

$$\sup_{f \in B_p^{s,n}} \inf_{p \in \pi_k^n} \|f - p\|_{L^p(K)} \le C k^{-s}$$

for some constant C. Since $r \asymp k^{n-1}$ and $\pi_k^n \subset \mathcal{R}(\mathbf{a}^1,\ldots,\mathbf{a}^r)$, it thus follows that for any $\{\mathbf{a}^1,\ldots,\mathbf{a}^r\}$ as above,

$$\sup_{f \in B_p^{s,n}} \inf_{g \in \mathcal{R}(\mathbf{a}^1,\ldots,\mathbf{a}^r)} \|f - g\|_{L^p(K)} \le C r^{-s/(n-1)}\,.$$

This is an elementary calculation. The more interesting question is whether this in fact is the correct order. That is, is the lower bound on the degree of approximation of the same order? It seems that it is essentially of that order, see Maiorov [14]. (Other results in this direction may be found in DeVore, Oskolkov Petrushev [3].)

Another method of obtaining error estimates developed as a consequence of this next result. This is due to Maurey, is contained in a paper of Pisier [17], was later independently proven by L. K. Jones [9] in a different form, and we quote it here in a slightly refined manner as it appears in Makovoz [15].

Theorem 3. *Let* $\Phi = \{\phi_1, \phi_2, \ldots\}$ *be an arbitrary bounded sequence of elements in a Hilbert space* H. *Let*

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

For every r *and* $f \in H$ *of the form*

$$f = \sum_i c_i \phi_i, \qquad \sum_i |c_i| < \infty,$$

there is a $g = \sum_{j=1}^r a_j \phi_{i_j}$ *with* $\sum_{j=1}^r |a_j| \leq \sum_i |c_i|$ *such that*

$$\|f - g\|_H \leq \frac{2\varepsilon_r(\Phi)}{\sqrt{r}} (\sum |c_i|).$$

We can translate this result into a ridge function "meta"theorem as follows.

Assume f lies in the closure of the convex hull of a bounded set of ridge functions $g(\mathbf{a} \cdot \mathbf{x})$, i.e., satisfying $\|g(\mathbf{a}\cdot)\|_{L^2(\mathbb{R}^n, d\mu)} \leq c$ for some fixed c.

Then

$$\inf_{g \in \mathcal{R}_r} \|f - g\|_{L^2(\mathbb{R}^n, d\mu)} \leq \frac{Ac}{\sqrt{r}}$$

for some absolute constant A.

The important and surprising fact worth noting here is that the bound is independent of n. Have we found a method of defeating the "curse of dimensionality"? Undoubtedly not. The "dimensionality" factor has been transferred to the class of functions being approximated. It seems that the function sets

$$\text{co}\{g(\mathbf{a}\cdot) : \|g(\mathbf{a}\cdot)\|_{L^2(\mathbb{R}^n, d\mu)} \leq c\}$$

are, in some sense, more and more constrained as the dimension increases. The problem remains to understand the nature of these sets. In [9], L. K. Jones gives an algorithm, close in character to the Friedman, Stuetzle algorithm, for which the bound is attained.

We recall that in the Friedman, Stuetzle algorithm we chose, at the rth step, a function g_r^* and a direction $\mathbf{b}^r$ so as to minimize

$$\| \left(f - \sum_{i=1}^{r-1} g_i^*(\mathbf{b}^i \cdot) \right) - g(\mathbf{a} \cdot) \|_{L^2(\mathbf{R}^n, d\mu)}$$

as we vary over g and $\mathbf{a}$. The Jones variant which gives the bound $Ac/\sqrt{r}$ for the requisite f has us, at step r, minimizing

$$\| f - (1-\alpha) \left(\sum_{i=1}^{r-1} g_i^*(\mathbf{b}^i \cdot) \right) - \alpha g(\mathbf{a} \cdot) \|_{L^2(\mathbf{R}^n, d\mu)}$$

over a function g, direction $\mathbf{a}$, and $\alpha \in [0,1]$.

§5. Recognizing Linear Combinations of Ridge Functions

As mentioned in Section 2, a function $f(x,y)$ is of the form

$$f(x,y) = \sum_{i=1}^{r} g_i(a_i x + b_i y)$$

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

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

in a "generalized" sense. Unfortunately such a simple characterization does not carry over to the case of three or more variables.

How can we determine if a function f (defined on $\mathbb{R}^n$) is of the form

$$f(\mathbf{x}) = \sum_{i=1}^{r} g_i(\mathbf{a}_i \cdot \mathbf{x})$$

for some given $\mathbf{a}^1, \ldots, \mathbf{a}^r \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, but unknown $g^1, \ldots, g^r$? That is, how do we identify $\mathcal{R}(\mathbf{a}^1, \ldots, \mathbf{a}^r)$? The answer is known and may be found in Lin, Pinkus [12]. Let $\mathcal{P}(\mathbf{a}^1, \ldots, \mathbf{a}^r)$ denote the set of polynomials which vanish on all the rays $\{\lambda \mathbf{a}^i : \lambda \in \mathbb{R}\}$, $i = 1, \ldots, r$.

Theorem 4. *The continuous function $f \in \mathcal{R}(\mathbf{a}^1, \ldots, \mathbf{a}^r)$ if and only if*

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

It follows from the theory of polynomial ideals that one need not check every $p \in \mathcal{P}(\mathbf{a}^1, \ldots, \mathbf{a}^r)$.

Can we determine when $f \in \mathcal{R}_r$ for a fixed r? (Recall that as the directions are also variables, $\mathcal{R}_r$ is a nonlinear set and more difficult to classify.) The case $r = 1$ is relatively simple. Assume

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

for some unknown $\mathbf{a}$ and g. Assume f (i.e., 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

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

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

References

1. Boman, J., On the closure of spaces of sums of ridge functions and the range of the X-ray transform, Ann. Inst. Fourier, Grenoble **34** (1984), 207–239.
2. Courant, R., and D. Hilbert, *Methods of Mathematical Physics, Vol. II*, Interscience Publishers, Inc., New York, 1962.
3. DeVore, R. A., Oskolkov, K. I., and P. P. Petrushev, Approximation by feed-forward neural networks, preprint.
4. Donoho, D. L., and I. M. Johnstone, Projection-based approximation and a duality method with kernel methods, Ann. Statist. **17** (1989), 58–106.
5. Friedman, J. H., and W. Stuetzle, Projection pursuit regression, J. Amer. Statist. Assoc. **76** (1981), 817–823
6. Huber, P. J., Projection pursuit, Ann. Statist. **13** (1985), 435–475.
7. John, F., *Plane Waves and Spherical Means Applied to Partial Differential Equations*, Interscience Publishers, Inc., New York, 1955.
8. Jones, L. K., On a conjecture of Huber concerning the convergence of projection pursuit regression, Ann. Statist. **15** (1987), 880–882.
9. Jones, L. K., A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training, Ann. Statist. **20** (1992), 608–613.
10. Kroo, A., On approximation by ridge functions, to appear in Const. Approx.

11. Light, W. A., and E. W. Cheney, *Approximation Theory in Tensor Product Spaces*, Lecture Notes in Mathematics, 1169, Springer-Verlag, Berlin, 1985.
12. Lin, V. Ya. and A. Pinkus, Fundamentality of ridge functions, J. Approx. Theory **75** (1993), 295–311.
13. Logan, B. F., and L. A. Shepp, Optimal reconstruction of a function from its projections, Duke Math. J. **42** (1975), 645–659.
14. Maiorov, V. E., On best approximation by ridge functions, preprint.
15. Makovoz, Y., Random approximants and neural networks, J. Approx. Theory **85** (1996), 98–109.
16. Petersen, B. E., K. T. Smith, and D. C. Solmon, Sums of plane waves, and the range of the Radon transform, Math. Ann. **243** (1979), 153–161.
17. Pisier, G. Remarques sur un resultat non publié de B. Maurey, Seminaire D'Analyse Fonctionnelle, 1980-1981, École Polytechnique, Centre de Mathématiques, Palaiseau, France.
18. Vostrecov, B. A. and M. A. Kreines, Approximation of continuous functions by superpositions of plane waves, Dokl. Akad. Nauk SSSR **140** (1961), 1237–1240 = Soviet Math. Dokl. **2** (1961), 1326–1329.

Allan Pinkus
Department of Mathematics
Technion–Israel Institute of Technology
Haifa, 32000
Israel
pinkus@math.technion.ac.il