Azerbaijan Journal of Mathematics
V. 3, No 2, 2013, July
ISSN 2218-6816

On Ridge Functions

A. Pinkus

Abstract. In this paper we survey some of the basic properties of linear combinations of ridge functions.

Key Words and Phrases: ridge functions, density, smoothness, representation

2010 Mathematics Subject Classifications: 41A02, 41A30, 41A63

1. Introduction

In this paper we will review a few of the basic properties associated with linear combinations of *Ridge Functions*. We hope the reader will find this subject worthy of further consideration.

A *Ridge Function*, in its simplest form, is any multivariate function

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

of the form

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

where $f : \mathbb{R} \to \mathbb{R}$, $\mathbf{x} = (x_1, \ldots, x_n)$, and $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n \backslash \{\mathbf{0}\}$. The vector $\mathbf{a} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ is generally called the *direction*. It is a multivariate function, constant on the hyperplanes $\mathbf{a} \cdot \mathbf{x} = c$, $c \in \mathbb{R}$. It is one of the simpler multivariate functions. Namely, a superposition of a univariate function with one of the simplest multivariate functions, the inner product.

More generally, we can also consider, for given d, $1 \leq d \leq n-1$, functions F of the form

$$F(\mathbf{x}) = f(A\mathbf{x}),$$

where A is a fixed $d \times n$ non-zero real matrix, and $f : \mathbb{R}^d \to \mathbb{R}$. For $d = 1$, this reduces to a ridge function. Many of the results reported on in this paper for $d = 1$ have their counterparts when $d > 1$.

We see ridge functions in numerous multivariate settings without considering them as of interest in and of themselves. For example, in multivariate Fourier series where the basic functions are of the form $e^{i(\mathbf{n}\cdot\mathbf{x})}$, for $\mathbf{n} \in \mathbb{Z}^n$, as the kernel of the Fourier transform

$e^{i(\mathbf{w}\cdot\mathbf{x})}$, and in the Radon transform. We see them in PDEs where, for example, if P is a constant coefficient polynomial in n variable, then

$$P\left(\frac{\partial}{\partial x_1},\ldots,\frac{\partial}{\partial x_n}\right) f = 0$$

has a solution $f(\mathbf{x}) = e^{\mathbf{a}\cdot\mathbf{x}}$ if and only if $P(\mathbf{a}) = 0$. And, of course, polynomials in the form $(\mathbf{a}\cdot\mathbf{x})^k$ are used in many settings.

2. Motivation

Where do we find ridge functions used in a more central way? Here are a few examples.

We see them in *Approximation Theory*. Ridge functions should be of interest to researchers and students of approximation theory. The basic concept in approximation theory is straightforward and simple. Approximate complicated functions by simpler functions. Among the class of multivariate functions linear combinations of ridge functions are a class of simpler functions. The questions one asks are the basic questions of approximation theory. Can one approximate arbitrarily well (density)? How well can one approximate (degree of approximation)? How does one approximate (algorithms)? Etc ....

Ridge functions are found in PDE theory where they were called *Plane Waves*. For example, we see them in the book by F. John [6]. In that book one finds representations of multivariate functions using integrals whose kernels are specific "plane waves" and applications thereof to partial differential equations. Plane waves are also discussed in Courant and Hilbert [1]. In general, linear combinations of ridge functions with fixed directions occur in the study of hyperbolic constant coefficient partial differential equations. As an example, assume the (a_i, b_i) are pairwise linearly independent vectors in $\mathbb{R}^2$. Then the general "solution" to the homogeneous partial differential equation

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

where the derivatives are understood in the sense of distributions, are all functions of the form

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

for arbitrary univariate functions f_i.

Ridge functions and ridge function approximations are studied in statistics in the analysis of large multivariate data sets. There they often go under the name of *projection pursuit*, (see e. g. Friedman and Stuetzle [2] and Huber [4]). Projection pursuit algorithms approximate a function of n variables by functions of the form

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

where both the directions $\mathbf{a}^i$ and the univariate functions f_i are variables. 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". The method of approximation, introduced by Friedman and Stuetzle [2] and called projection pursuit regression (PPR), is essentially a stepwise greedy algorithm that, at its kth stage, looks for a best (or good) approximation of the form $f_k(\mathbf{a}^k \cdot \mathbf{x})$, as we vary over both the univariate function f_k and the direction $\mathbf{a}^k$.

Ridge functions appear in many neural network models. One of the popular models in the theory of neural nets is that of a *multilayer feedforward perceptron* (MLP) neural net with input, hidden, and output layers. The simplest case (which is that of one hidden layer, r processing units and one output) considers, in mathematical terms, functions of the form

$$\sum_{i=1}^{r} \alpha_i \sigma(\mathbf{w}^i \cdot \mathbf{x} + \theta_i),$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is some given fixed univariate function, $\theta_i \in \mathbb{R}$, $\mathbf{w}^i \in \mathbb{R}^n \backslash \{\mathbf{0}\}$. In this model, which is just one of many, we are in general permitted to vary over the $\mathbf{w}^i$ and θ_i, in order to approximate an unknown function. Note that for each θ and $\mathbf{w}$ the function

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

is also a ridge function. Thus, a lower bound on the degree of approximation by such functions is given by the degree of approximation by ridge functions. See e. g. Pinkus [10] and references therein for more on this problem.

The term *ridge function* was coined in a 1975 paper by Logan and Shepp [9]. This was a seminal paper in computerized tomography. In tomography, or at least in tomography as the theory was initially constructed in the early 1980's, ridge functions were basic. The idea there was to try to reconstruct a given, but unknown, function $G(\mathbf{x})$ from the values of its integrals along certain planes or lines. Logan and Shepp considered ridge functions in the unit disk in $\mathbb{R}^2$ with equally spaced directions. We will consider some nice domain K in $\mathbb{R}^n$, and a function G belonging to $L^2(K)$. Assume that for some fixed directions $\{\mathbf{a}^i\}_{i=1}^r$ we are given

$$\int_{K \cap \{\mathbf{a}^i \cdot \mathbf{x} = \lambda\}} G(\mathbf{x})\, d\mathbf{x}$$

for each λ and $i = 1, \ldots, r$. That is, we see the *projections* of G along the hyperplanes $K \cap \{\mathbf{a}^i \cdot \mathbf{x} = \lambda\}$, λ a.e., $i = 1, \ldots, r$. What is a good method of reconstructing G based only on this information? It easily transpires, from basic orthogonality considerations, that the unique best $L^2(K)$ approximation

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

to G from the space

$$\mathcal{M}(\mathbf{a}^1, \ldots, \mathbf{a}^r) = \left\{ \sum_{i=1}^{r} f_i(\mathbf{a}^i \cdot \mathbf{x}) : \ f_i \text{ vary} \right\},$$

if such a best approximation exists, necessarily satisfies

$$\int_{K\cap\{\mathbf{a}^i\cdot\mathbf{x}=\lambda\}} G(\mathbf{x})\,d\mathbf{x} = \int_{K\cap\{\mathbf{a}^i\cdot\mathbf{x}=\lambda\}} f^*(\mathbf{x})\,d\mathbf{x}$$

for each λ and $i = 1,\dots,r$. That is, it has the same projections as G. Furthermore, since it is a best approximation in a Hilbert space, its norm is less than the norm of G. Thus, among all functions with the same data (projections) as G, this specific linear combination of ridge functions is the one of minimal $L^2(K)$ norm. In the unit disk in $\mathbb{R}^2$ with equally spaced directions, Logan and Shepp also give a closed-form expression for f^*.

3. Density of Ridge Functions

Ridge functions are dense in $C(K)$ for every compact $K \in \mathbb{R}^n$. For example, consider

$$span\{e^{\mathbf{n}\cdot\mathbf{x}} : \mathbf{n} \in \mathbb{Z}_+^n\}.$$

It easily follows from the Stone-Weierstrass Theorem that this class is dense.

Let Ω be any fixed set of vectors in $\mathbb{R}^n$, and

$$\mathcal{M}(\Omega) = \text{span}\{f(\mathbf{a}\cdot\mathbf{x}) : \mathbf{a} \in \Omega, \text{all } f\}.$$

A more interesting question is to try to determine necessary and sufficient conditions on the set of directions Ω for when we have density of $\mathcal{M}(\Omega)$ in $C(\mathbb{R}^n)$, in the topology of uniform convergence on compact subsets of $\mathbb{R}^n$. Why this norm and topology? Note that no ridge function (other than the identically zero function) is in any of the classical spaces $L^p(\mathbb{R}^n)$, for any $p \in [1,\infty)$. As such, the set of functions that we will approximate are functions of the class $C(\mathbb{R}^n)$, in the topology of uniform convergence on compact subsets. That is, we would like to have the property that for any given $G \in C(\mathbb{R}^n)$, any compact set $K \subset \mathbb{R}^n$, and any $\varepsilon > 0$, there exists an $F \in \mathcal{M}(\Omega)$ such that

$$\|G - F\|_K = \max_{\mathbf{x}\in K} |G(\mathbf{x}) - F(\mathbf{x})| < \varepsilon.$$

If we can prove density for this class of functions, then we also obtain density in many other spaces, such as $L^p(K)$, where K is any compact subset of $\mathbb{R}^n$, and every $p \in [1,\infty)$.

The following result may be found in Vostrecov and Kreines [12], see also Lin and Pinkus [8]. This result was, for many years, overlooked.

Theorem 3.1. *$\mathcal{M}(\Omega)$ is dense in $C(\mathbb{R}^n)$, in the topology of uniform convergence on compact subsets, if and only if no non-trivial homogeneous polynomial vanishes on Ω.*

Some simple consequences of this result are the following.

Corollary 3.2. *Assume $\Omega = \Omega_1 \cup \Omega_2$. Then $\mathcal{M}(\Omega)$ is dense in $C(\mathbb{R}^n)$, in the topology of uniform convergence on compact subsets, if and only if $\mathcal{M}(\Omega_j)$ is dense in $C(\mathbb{R}^n)$, in the topology of uniform convergence on compact subsets, for $j = 1$ and/or $j = 2$.*

Corollary 3.3. *If Ω contains only a finite number of distinct elements, then $\overline{\mathcal{M}(\Omega)} \neq C(\mathbb{R}^n)$.*

Corollary 3.4. *In $\mathbb{R}^2$, we have that $\mathcal{M}(\Omega)$ is dense in $C(\mathbb{R}^2)$, in the topology of uniform convergence on compact subsets, if and only if Ω contains an infinite number of pairwise linearly independent vectors.*

What about if the directions are also permitted to vary? Let Ω_j, $j \in J$, be sets of vectors in $\mathbb{R}^n$, and $\mathcal{M}(\Omega_j)$ be as above. We ask when, for each given $G \in C(\mathbb{R}^n)$, compact $K \subset \mathbb{R}^n$ and $\varepsilon > 0$, there exists an $F \in \mathcal{M}(\Omega_j)$, for some $j \in J$, such that

$$\|G - F\|_K < \varepsilon?$$

If $\Omega_j = \Omega$ for all $j \in J$, then this is exactly the problem considered previously. If the $\{\Omega_j\}_{j\in J}$ are the totality of all sets with at most k elements, then this is the problem of approximating with k arbitrary directions

To state the result we introduce the following quantity. To each Ω_j, let r_j be the minimal degree of the non-trivial homogeneous polynomials that vanish on Ω_j. If no non-trivial homogeneous polynomial vanishes on Ω_j, we set $r_j = \infty$.

The following result is from Kroó [7].

Theorem 3.5. *The set $\bigcup_{j\in J} \mathcal{M}(\Omega_j)$ is dense in $C(\mathbb{R}^n)$, in the above sense, if and only if*

$$\sup_{j\in J} r_j = \infty.$$

4. Representation

As in the previous section, let Ω be any set of fixed vectors in $\mathbb{R}^n$, and

$$\mathcal{M}(\Omega) = \operatorname{span}\{f(\mathbf{a}\cdot\mathbf{x}) : \mathbf{a} \in \Omega, \text{all } f\}.$$

The question we ask here is the following. *What is $\overline{\mathcal{M}(\Omega)}$ when it is not all of $C(\mathbb{R}^n)$?*

For any polynomial p we define

$$p(D) := p\left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}\right).$$

Let $\mathcal{P}(\Omega)$ be the set of all homogeneous polynomials that vanish on Ω, and let $\mathcal{C}(\Omega)$ be the set of all polynomials q such that

$$p(D)q = 0, \quad \text{all } p \in \mathcal{P}(\Omega).$$

Then we have the following.

Theorem 4.1. *On $C(\mathbb{R}^n)$, in the topology of uniform convergence on compact subsets, we have*

$$\overline{\mathcal{M}(\Omega)} = \overline{\mathcal{C}(\Omega)}.$$

As a consequence we have, for example, that $g(\mathbf{b}\cdot\mathbf{x}) \in \overline{\mathcal{M}(\Omega)}$ for some $\mathbf{b}$ and all continuous g if and only if all homogeneous polynomials vanishing on Ω also vanish on $\mathbf{b}$.

For $n = 2$ and $\Omega = \{(a_i, b_i)\}_{i=1}^r$, we have from Theorem 4.1 that

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

for arbitrary smooth f_i 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 fact, the set $\mathcal{M}(\Omega)$ is, to a great extent, determined by the polynomials it contains. To understand this relationship we have the following result.

Let Π_m^n denote the set of polynomials of *total degree* at most m in n variables, and let H_m^n denote the set of homogeneous polynomials of degree m in n variables. Then

Proposition 4.2. *We have the equality*

$$\operatorname{span}\{(\mathbf{a}\cdot\mathbf{x})^r : \mathbf{a} \in \Omega,\, r = 0, \ldots, m\} = \Pi_m^n$$

if and only if no non-trivial $p \in H_m^n$ *vanishes on* Ω.

5. Smoothness

Assume

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

where r is finite, and the $\mathbf{a}^i$ are pairwise linearly independent fixed vectors in $\mathbb{R}^n$. If G is of a certain smoothness class, what can we say about the smoothness of the f_i?

Let us first consider the simpler cases. Assume $G \in C^k(\mathbb{R}^n)$. If $r = 1$ then since

$$G(\mathbf{x}) = f_1(\mathbf{a}^1 \cdot \mathbf{x})$$

is in $C^k(\mathbb{R}^n)$ for some $\mathbf{a}^1 \neq \mathbf{0}$, it easily follows that $f_1 \in C^k(\mathbb{R})$.

Now assume $r = 2$. As the $\mathbf{a}^1$ and $\mathbf{a}^2$ are linearly independent, there exists a vector $\mathbf{c} \in \mathbb{R}^n$ satisfying $\mathbf{a}^1 \cdot \mathbf{c} = 0$ and $\mathbf{a}^2 \cdot \mathbf{c} = 1$. Thus for all $t \in \mathbb{R}$

$$G(t\mathbf{c}) = f_1(\mathbf{a}^1 \cdot t\mathbf{c}) + f_2(\mathbf{a}^2 \cdot t\mathbf{c}) = f_1(0) + f_2(t).$$

As $G(t\mathbf{c})$ is in $C^k(\mathbb{R})$, as a function of t, so is f_2. The same result holds for f_1.

For $r \geq 3$, the situation is very much different. Recall that the Cauchy Functional Equation

$$g(x+y) = g(x) + g(y)$$

has, as proved by Hamel [3], very badly behaved solutions. As such, setting $f_1 = f_2 = -f_3 = g$, we have very badly behaved (and certainly not in $C^k(\mathbb{R})$) f_i, $i = 1, 2, 3$, that satisfy

$$0 = f_1(x_1) + f_2(x_2) + f_3(x_1 + x_2)$$

for all $(x_1, x_2) \in \mathbb{R}^2$. This Cauchy Functional Equation turns out to be critical in the analysis of this problem for all $r \geq 3$.

Denote by $\mathcal{B}$ any class of real-valued functions f defined on $\mathbb{R}$ such that if there is a function $r \in C(\mathbb{R})$ such that $f - r$ satisfies the Cauchy Functional Equation, then $f - r$ is necessarily linear, i.e. $(f - r)(x) = Ax$ for some constant A, and all $x \in \mathbb{R}$. $\mathcal{B}$ includes, for example, the set of all functions that are continuous at a point, or monotonic on an interval, or bounded on one side on a set of positive measure, or Lebesgue measurable. The following is from Pinkus [11].

Theorem 5.1. *Assume $G \in C^k(\mathbb{R}^n)$ is of the form*

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

where r is finite, and the $\mathbf{a}^i$ are pairwise linearly independent vectors in $\mathbb{R}^n$. Assume, in addition, that each $f_i \in \mathcal{B}$. Then, necessarily, $f_i \in C^k(\mathbb{R})$ for $i = 1, \ldots, r$.

6. Uniqueness of the Representations

What can we say about the uniqueness of the representation? That is, when and for which functions $\{g_i\}_{i=1}^k$ and $\{h_i\}_{i=1}^\ell$ can we have distinct representations

$$G(\mathbf{x}) = \sum_{i=1}^{k} g_i(\mathbf{b}^i \cdot \mathbf{x}) = \sum_{j=1}^{\ell} h_i(\mathbf{c}^i \cdot \mathbf{x})$$

for all $\mathbf{x} \in \mathbb{R}^n$, where k and ℓ are finite, and the $\mathbf{b}^1, \ldots, \mathbf{b}^k, \mathbf{c}^1, \ldots, \mathbf{c}^\ell$ are $k + \ell$ pairwise linearly independent vectors in $\mathbb{R}^n$?

From linearity this is, of course, equivalent to the following. Assume

$$\sum_{i=1}^{r} f_i(\mathbf{a}^i \cdot \mathbf{x}) = 0$$

for all $\mathbf{x} \in \mathbb{R}^n$, where r is finite, and the $\mathbf{a}^i$ are pairwise linearly independent vectors in $\mathbb{R}^n$. What does this imply regarding the f_i?

Theorem 6.1. *Assume*

$$\sum_{i=1}^{r} f_i(\mathbf{a}^i \cdot \mathbf{x}) = 0$$

where r is finite, and the $\mathbf{a}^i$ are pairwise linearly independent vectors in $\mathbb{R}^n$. Assume, in addition, that $f_i \in \mathcal{B}$, for $i = 1, \ldots, r$. Then $f_i \in \Pi_{r-2}^1$, $i = 1, \ldots, r$.

That is, with minor smoothness assumptions we have uniqueness of representations up to polynomials of degree $r - 2$.

7. Conclusion

In these few pages we have touched upon certain basic properties of linear combinations of ridge functions. There are many, many other aspects to the study of ridge functions that we have not considered. For example, we have not discussed the question of characterizing best approximations from spaces of finite linear combination of ridge functions with fixed or variable directions in different normed linear spaces, nor the algorithms for finding such best approximations. We have not looked at the closure of the space $\mathcal{M}(\Omega)$ in, say, $L^p(K)$ for bounded K (it may or may not be closed). We did not review known results concerning interpolation at points or on lines by ridge functions. Some of these and related questions are addressed in the recent review article by Ismailov [5].

Acknowledgements

This paper is based on a lecture given at the *International Conference on Actual Problems of Mathematics and Informatics* held in Baku, Azerbaijan, on May 29–31, 2013. I wish to thank the organizers of that conference for their invitation and for their warm hospitality.

References

[1] R. Courant and D. Hilbert. Methods of Mathematical Physics, Vol. II. Interscience Publishers, Inc., New York, 1962.

[2] J. H. Friedman and W. Stuetzle. Projection pursuit regression. J. Amer. Statist. Assoc., 76: 817–823, 1981.

[3] G. Hamel. Eine Basis aller Zahlen und die unstetigen Lösungen der Funktionalgleichung $f(x+y) = f(x) + f(y)$. Math. Ann., 60: 459–462, 1905.

[4] P. J. Huber. Projection pursuit. Ann. Statist., 13: 435–475, 1985.

[5] V. E. Ismailov. A review of some results on ridge function approximation. Azerbaijan Journal of Mathematics, 3(1): 3–51, 2013.

[6] F. John. Plane Waves and Spherical Means applied to Partial Differential Equations. Interscience Publishers, Inc., New York, 1955.

[7] A. Kroo. On approximation by ridge functions. Constr. Approx., 13: 447–460, 1997.

[8] V. Ya. Lin and A. Pinkus. Fundamentality of ridge functions. J. Approx. Theory, 75: 295-311, 1993.

[9] B. F. Logan and L. A. Shepp. Optimal reconstruction of a function from its projections. Duke Math. J., 42: 645–659, 1975.

[10] A. Pinkus. Approximation theory of the MLP model in neural networks. Acta Numerica, 8: 143–195, 1999.

[11] A. Pinkus. Smoothness and Uniqueness in Ridge Function Representation. to appear in Indagationes Mathematicae.

[12] B. A. Vostrecov and M. A. Kreines. Approximation of continuous functions by superpositions of plane waves. Dokl. Akad. Nauk SSSR, 140: 1237–1240, 1961 = Soviet Math. Dokl., 2: 1326–1329, 1961.

Allan Pinkus
Department of Mathematics, Technion, Haifa, Israel
E-mail: pinkus@technion.ac.il