On Ridge Functions

Allan Pinkus

Technion

September 23, 2013

Foreword

In this lecture we will survey a few problems and properties associated with Ridge Functions. I hope to convince you that this is a subject worthy of further consideration, especially as regards to *Multivariate Approximation and Interpolation with Applications*

What is a Ridge Function?

- 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.

Where do we find Ridge Functions?

We see specific Ridge Functions in numerous multivariate settings without considering them as of interest in and of themselves.

• In multivariate Fourier series where the basic functions are of the form $e^{i(\mathbf{n}\cdot\mathbf{x})}$, for $\mathbf{n} \in \mathbb{Z}^n$, in the Fourier transform $e^{i(\mathbf{w}\cdot\mathbf{x})}$, and in the Radon transform.

• In PDE 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 of the form $f(\mathbf{x}) = e^{\mathbf{a}\cdot\mathbf{x}}$ if and only if $P(\mathbf{a}) = 0$.

• The polynomials $(\mathbf{a} \cdot \mathbf{x})^k$ are used in many settings.

Where do we use Ridge Functions?

• Approximation Theory – Ridge Functions should be of interest to researchers and students of approximation theory. The basic concept 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 ....

Where do we use Ridge Functions?

• Partial Differential Equations – Ridge Functions used to be called Plane Waves. For example, we see them in the book *Plane Waves and Spherical Means applied to Partial Differential Equations* by Fritz John. In general, linear combinations of ridge functions also appear in the study of hyperbolic constant coefficient pde's. As an example, assume the (a_i, b_i) are pairwise linearly independent vectors in $\mathbb{R}^2$. Then the general "solution" of the pde

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

are all functions of the form

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

for arbitrary f_i.

Where do we use Ridge Functions?

- Projection Pursuit – This is a topic in Statistics. Projection pursuit algorithms approximate a functions of n variables by functions of the form

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

where both the functions g_i and directions $\mathbf{a}^i$ are variables. The idea here is to "reduce dimension" and thus bypass the "curse of dimensionality".

Where do we use Ridge Functions?

- Neural Networks – One of the popular neuron models is that of a *multilayer feedforward neural net* with input, hidden and output layers. In its simplest case, and without the terminology used, one is interested in functions of the form

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

where $\sigma : \mathbb{R} \to \mathbb{R}$ is some given fixed univariate function. In this model, which is just one of many, we vary the w_{ij}, θ_i and α_i. For each θ and $\mathbf{w} \in \mathbb{R}^n$ we are considering linear combinations of

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

Thus, a lower bound on the degree of approximation by such functions is given by the degree of approximation by ridge functions.

Where do we use Ridge Functions?

- Computerized Tomography – The term Ridge Function was coined in a 1975 paper by Logan and Shepp, that was a seminal paper in computerized tomography. They 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 $L^2(K)$.

Problem: 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?

Answer: 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

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

if such 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 $i = 1, \ldots, r$, and among all such functions with the same data as G is the one of minimal $L^2(K)$ norm.

Properties of Ridge Functions

In the remaining part of this lecture I want to consider various properties of linear combinations of Ridge Functions. Namely,

- Density
- Representation
- Smoothness
- Uniqueness
- Interpolation

Density - Fixed Directions

• Ridge functions are dense in $C(K)$ for every compact $K \subset \mathbb{R}^n$.
E.g., span $\{e^{\mathbf{n}\cdot\mathbf{x}} : \mathbf{n} \in \mathbb{Z}_+^n\}$ is dense (Stone-Weierstrass).

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

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

Theorem (Vostrecov, Kreines)

$\mathcal{M}(\Omega)$ 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.

Density - Variable Directions

• 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

$$\|G - F\|_{L^\infty(K)} < \varepsilon.$$

(If Ω_j are the totality of all sets of ridge functions with k directions, then this is the problem of approximating with k arbitrary directions.)

• To each Ω_j, let r_j be the minimal degree of a non-trivial homogeneous polynomial vanishes on Ω_j. Then (Kroó)

$$\bigcup_{j \in J} \mathcal{M}(\Omega_j)$$

is dense in $C(\mathbb{R}^n)$, as explained above, if and only if

$$\sup_{j \in J} r(\Omega_j) = \infty.$$

Representation

- As previously, let Ω be any set of vectors in $\mathbb{R}^n$, and

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

The question we now ask is: What is $\overline{\mathcal{M}(\Omega)}$ when it is not all of $C(\mathbb{R}^n)$?

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

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

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

Representation

Theorem

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

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

- Thus, for example, $g(\mathbf{b} \cdot \mathbf{x}) \in \overline{\mathcal{M}(\Omega)}$ for some $\mathbf{b}$ and all continuous g if and only if all homogeneous polynomials vanishing on Ω also vanish on $\mathbf{b}$.
- For $n = 2$, $\Omega = \{(a_i, b_i)\}_{i=1}^r$ this gives us

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

for arbitrary smooth f_i if and only if

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

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?

Smoothness — $r = 1$, $r = 2$

• Assume $G \in C^k(\mathbb{R}^n)$. If $r = 1$ there is nothing to prove. That is, assume

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

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

• Let $r = 2$. 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

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

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.

Smoothness — $r \geq 3$

Recall that the Cauchy Functional Equation

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

has, as proved by Hamel (1905), 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

$$0 = f_1(x) + f_2(y) + f_3(x+y)$$

for all $(x, y) \in \mathbb{R}^2$. This Cauchy Functional Equation is critical in the analysis of our problem for all $r \geq 3$.

Smoothness

• 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.

Smoothness — Theorem

Theorem

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 $i = 1, \ldots, r$.

Uniqueness

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 $\mathbb{R}^n$?

Uniqueness

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?

Uniqueness — Theorem

Theorem

Assume

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

holds 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 $f_i \in \Pi^1_{r-2}$, $i = 1, \ldots, r$, where Π^1_{r-2} denotes the set of polynomials of degree at most $r - 2$.

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

Interpolation

Assume $\mathbf{a}^1, \ldots, \mathbf{a}^m \in \mathbb{R}^n$ are m fixed pairwise linearly independent directions, and

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

Interpolation at a finite number of points, for any given data, by functions from $\mathcal{M}(\mathbf{a}^1, \ldots, \mathbf{a}^m)$ was studied is a few papers in the mid 1990's.

The real question is: for which points can we not interpolate? The only cases well-understood are $m = 2$ for all n, and $m = 3$ if $n = 2$.

Interpolation

Recently the following problem was considered. Given arbitrary data on k straight lines in $\mathbb{R}^n$, can we interpolate from $\mathcal{M}(\mathbf{a}^1, \dots, \mathbf{a}^m)$ to arbitrary data on these straight lines?

• If $m = 1$, $k = 1$ and the line $\ell_1 = \{t\mathbf{b}^1 + \mathbf{c}^1 : t \in \mathbb{R}\}$, then one can interpolate iff $\mathbf{a}^1 \cdot \mathbf{b}^1 \neq 0$.

• If $m = 1$ or $m = 2$, then for $k > m$, one **cannot** interpolate $\mathcal{M}(\mathbf{a}^1, \dots, \mathbf{a}^m)$ to arbitrary data on these k straight lines.

• If $m = 2$ and $k = 2$, one can generally interpolate arbitrary data except when certain known (too detailed to list here) conditions hold.

• If $m = k = n = 2$, and the two lines are $\ell_j = \{t\mathbf{b}^j + \mathbf{c}^j : t \in \mathbb{R}\}$, $j = 1, 2$, then these conditions reduce to

$$(\mathbf{a}^1 \cdot \mathbf{b}^1)(\mathbf{a}^2 \cdot \mathbf{b}^2) + (\mathbf{a}^1 \cdot \mathbf{b}^2)(\mathbf{a}^2 \cdot \mathbf{b}^1) \neq 0,$$

and if the lines ℓ_1 and ℓ_2 intersect, then the data is consistent at the intersection point.

What we did not talk about!!

• In this short talk we touched upon only a few properties of Ridge Functions.

• Other important properties that have been studied and are being studied include degree of approximation, the inverse problem (identifying ridge functions and their directions), closure properties of $\mathcal{M}(\Omega)$, ridgelets and algorithms for approximation.

Thank you for your attention!!