New Ties between Computational Harmonic Analysis and Approximation Theory

Emmanuel J. Candès

Abstract. Is the connection between approximation theory and harmonic analysis genuine? This question may seem a little provocative, especially in light of the recent literature about the significant interactions between wavelet analysis and nonlinear approximation. However, this connection might just as well be an accident as it merely seems to be limited to wavelets. For instance, do we know of any other constructions giving important insights into central problems in approximation theory? The work surveyed in this paper suggests that the connection between applied harmonic analysis and approximation theory is much larger than that wavelets or Fourier analysis imply. We introduce some fundamentally new systems for representing functions, and will show how well those new constructions connect with central problems in approximation theory. Especially, we will explore the subject of ridge function approximation and the problem of representing or approximating objects with spatial inhomogeneities.

§1. Introduction

This paper is the companion to the lecture delivered at the 10th Conference on Approximation Theory on March 28th, 2001. Rather than being about a specific problem, this is merely an account of some ideas that have been part of my intellectual life for about 5 years. Many of these ideas formed as I was a Ph. D. student in the Department of Statistics at Stanford University. David L. Donoho served as my advisor and this paper owes a lot to his intellectual generosity.

This article is loosely organized around a central theme, namely, the connection between computational harmonic analysis (CHA) and approximation theory (AT). The rapid development of wavelets in the eighties together with the key observation that wavelets can be construed as

Approximation Theory X: Wavelets, Splines, and Applications
Charles K. Chui, Larry L. Schumaker, and Joachim Stöckler (eds.), pp. 87–153.
Copyright © 2002 by Vanderbilt University Press, Nashville, TN.
ISBN 0-8265-1416-2.
All rights of reproduction in any form reserved.

the building blocks of many functional classes and approximation spaces [20,48] fueled an explosive literature on this subject, see [29] for a wonderful survey. The importance of this literature was amplified by the fact that most of the ideas discussed bear great practical significance for signal and image processing. By now, we are all familiar with wavelet success stories such as their inclusion in JPEG 2000, the new still-picture compression standard, or the celebrated wavelet shrinkage for removing noise from signals or images.

1.1 An accidental connection?

By the late nineties, the connection between wavelet and approximation theory had been explored rather extensively. A concern –at least to the author– was whether the connection between CHA and approximation theory was genuine. Indeed, there was no evidence at the time of any other CHA construction giving important insights into central problems in approximation theory. In other words, was this tie between wavelets and approximation theory merely an accident?

¿From the CHA viewpoint, the domain of expertise of wavelets was clearly understood. At the same time, the concern that wavelets were perhaps not the answer to every problem in approximation theory developed. For instance, wavelets offer poor representations of two-dimensional objects that are singular along curves. An example, which to my knowledge should be attributed to Yves Meyer, shows that in two dimensions, nonlinear wavelet approximations of the indicator function of the unit disk converge 'slowly.' We used the word slowly because one can exhibit other methods of approximation such as using superpositions of indicators of triangles with all possible shapes which converge at a much faster rate. In the introduction to this lecture, Ron DeVore addressed this issue and perhaps best summarized that feeling when he said that "wavelets had hit a wall." Implications are clear. The challenge is to find new tools in CHA for problems that wavelets are not able to address efficiently. We should acknowledge, however, that this is an act of faith in which we express the belief that there are many possible CHA constructions beyond wavelets. Against this optimism, we often heard in conferences that not only wavelets but CHA had run its course and that we need ideas outside of CHA. The recent work of Mallat and his collaborators on bandelets [40] indeed explores other directions.

Conversely, from the approximation theory viewpoint, wavelets brought a remarkable sense of closure to an important line of research about spline and rational approximation as well as related topics in mathematical analysis such as applied functional analysis and interpolation theory. For instance, wavelet, free-knot spline or rational function approximation are, in some sense, optimal over the scale of Besov spaces and give the same speed of convergence.

The scope of approximation theory, however, is of course much larger than wavelet or spline approximation. In particular, nonlinear approximation received an increasing degree of attention, perhaps motivated by the close relationship existing between nonlinear approximation and the subject of data compression. We shall now describe a typical set of problems in this area.

1.2 Nonlinear approximation

Given a dictionary of generally overcomplete elements $\mathcal{D} = \{g_\lambda, \lambda \in \Lambda\}$, we are interested in the best m-term approximation

$$\inf_{f_m \in \Sigma_m} \|f - f_m\|_{L_2}, \tag{1.1}$$

where Σ_m is the set of objects that are linear combinations of at most m elements of $\mathcal{D}$

$$\Sigma_m = \{g,\ g = \sum_{i=1}^{m} \alpha_i g_{\lambda_i}\}. \tag{1.2}$$

Note that Σ_m is not a linear subspace. If g and h are both in Σ_m, then $f + g$ is generally not a member of Σ_m but rather, of the larger set Σ_{2m}. Let a target function f be given. How do we construct the best or near-best m-term approximation to f? What do we know about the degree of approximation

$$d_m(f, \mathcal{D}) \equiv \inf_{f_m \in \Sigma_m} \|f - f_m\|_{L_2} \quad \text{as} \quad m \to \infty? \tag{1.3}$$

Next, instead of being interested in the performance for a specific target f, one might study, instead, the quality of the approximation of a functional class $\mathcal{F}$ by the dictionary $\mathcal{D}$. For example, what do we know about

$$d_m(\mathcal{F}, \mathcal{D}) \equiv \sup_{f \in \mathcal{F}} d_m(f, \mathcal{D})? \tag{1.4}$$

There are also interesting characterization issues. For instance, given a dictionary $\mathcal{D}$ can we characterize those functions that can be approximated at a given rate by members of Σ_m? And conversely, given a functional class $\mathcal{F}$, how do we design a dictionary which, in some sense, is best for approximating its elements? Problems of this nature, whether they are about the characterization of approximation spaces or about the existence of constructive procedures for obtaining near-best m term approximations, are usually hopelessly difficult. There is a lack of universal principles which the author finds a little discouraging. Indeed, each new problem calls for a lot of hard work and progress are generally at the expense of a lot of hard and cumbersome analysis.

1.3 Our claim

I claim that there is much more to CHA than wavelets and related systems. Indeed, the work surveyed in this paper will introduce a wealth of new systems, and we will show how well those new constructions connect with central problems in approximation theory. Further, we suggest that the connection between CHA and approximation theory is much larger than what wavelets imply. We begin our exploration of this renewed connection with the problem of approximating with superpositions of ridge functions.

1.4 Ridge functions

In the seventies, Logan and Shepp [43] coined the terminology ridge function to designate functions of the form $g(k\cdot x) = g(k_1x_1+k_2x_2+\ldots+k_dx_d)$. In other words, a ridge function is a multivariate function constant on the hyperplanes $k\cdot x = t$ where t is a real valued constant. Ridge functions are also known under the name of planar waves. Over the last twenty years or so, ridge functions have appeared rather frequently in the scientific literature.

Computerized tomography. First, ridge functions play an important role in the literature of computerized tomography. Logan and Shepp [43] considered the problem of reconstructing a two-dimensional function $f(x)$ from its projections $(Rf(\cdot,\theta_1)\ ,\ Rf(\cdot,\theta_2),\ \ldots\ ,\ Rf(\cdot,\theta_k))$ for fixed and distinct directions $\theta_1,\ldots,\theta_k$ which belong to $\mathbf{S}^1$, the unit sphere of $\mathbb{R}^2$. Here Rf is the Radon transform

$$Rf(t,\theta) = \int_{\theta\cdot x=t} f(x)\,dx, \quad t\in\mathbb{R},\ \theta\in\mathbf{S}^1. \tag{1.5}$$

This problem is, indeed, a mathematical idealization of several imaging problems [18] as in tomography. Define the set of 'candidates' as all functions whose Radon transform match the given projections for every θ_j, $j=1,\ldots,k$. We use the method of regularization and seek that object f^* with minimum L_2 norm among all possible candidates. The result proved by Logan and Shepp is that f^* is a superposition of at most k ridge functions.

Statistics. In statistics, ridge functions were introduced to overcome the adverse effect of the curse of dimensionality. A central problem there is that of estimating an unknown regression surface given data $(x_i,y_i)_{i=1}^N$, where x_i is a d-dimensional input variable and y_i is a real-valued output response and the model

$$y_i = f(x_i) + \epsilon_i; \tag{1.6}$$

here, ϵ is a stochastic and noisy contribution which is assumed to have zero-mean so that the problem is to estimate the conditional mean of y given x

as $f(x) = E(y|x)$. Friedman and Stuetzle [31] suggest approximating the unknown regression function f by a sum of ridge functions

$$f(x) \sim \sum_{j=1}^{m} g_j(k_j \cdot x),$$

where the k_j's are vectors of unit length, i.e. $\|k_j\| = 1$. In its abstract version, the approximation process operates in a stepwise and greedy fashion. At stage m, it augments the fit f_{m-1} by adding a ridge function $g_m(k_m \cdot x)$, where k_m and g_m are chosen so that $g_m(k_m \cdot x)$ best approximates the residuals $f(x) - f_{m-1}(x)$.

For completeness, we should stress that the original paper presented a principle for fitting finite linear combinations of ridge functions in the sampling case. Given data $(x_i, y_i)_{i=1}^N$, the procedure is analogous to that described above. At stage m, the fit f_{m-1} is augmented by adding a ridge function $g_j(k_j \cdot x)$ obtained as follows: calculate the residuals of the $(m-1)$th fit $r_i = x_i - \sum_{j=1}^{m-1} g_j(k_j \cdot x_i)$; and for a fixed direction a, plot the residuals r_i against $k \cdot x_i$; fit a smooth curve g and choose the best direction a, so as to minimize the residuals sum of squares $\sum_i (r_i - g(k \cdot x_i))^2$. The algorithm stops when the improvement is small.

Partial differential equations. Planar waves appear frequently in the study of partial differential equations and especially in certain hyperbolic problems [35]. To give a flavor of this connection, we follow [33] and consider the differential operator with constant coefficients

$$L = \sum_{\alpha} a_\alpha D^\alpha,$$

where D is either one of the partial derivatives $\partial_1, \partial_2, \ldots, \partial_d$. Suppose we want to solve

$$Lu = f, \tag{1.7}$$

where f is a nice object, namely, f belongs to the Schwartz class $\mathcal{S}(\mathbb{R}^d)$. Start with the equation

$$Lu = f_\theta, \tag{1.8}$$

where $\theta \in \mathbf{S}^{d-1}$ and $f_\theta(x) = f_\theta(x \cdot \theta)$ is a ridge function. Then, looking for solutions of the form $u_\theta(x) = u_\theta(x \cdot \theta)$ amounts to solving an *ordinary* differential equation. If we can synthesize the right hand-side of (1.7) as a superposition of ridge functions, then under some conditions, we would be able to write down a solution to (1.7) as a superposition of ridge functions as well. For completeness, such synthesizing principles are possible by means of the Radon transform.

As an example, consider the Cauchy problem for the wave equation with constant coefficients

$$\Delta u = \frac{\partial^2 u}{\partial t^2}, \quad u(x,0) = f_0(x),\ u_t(x,0) = f_1(x),$$

where f_0 and f_1 are given smooth functions. Then one has available an explicit formula which expresses the Cauchy problem as a continuous superposition of ridge functions [33]. Note also that for $h \in C^2$ and $\theta \in \mathbf{S}^{d-1}$, the ridge function

$$v^{\pm}(x,t) = h(x \cdot \theta \pm t)$$

satisfies $\Delta v = (\partial^2/\partial t^2)v$.

Neural networks. There is no real definition of a neural network. However, the most commonly studied neural network is the one hidden-layer feedforward neural network which is a name given to a function constructed by the rule

$$f_m = \sum_{j=1}^{m} \alpha_j \sigma(k_j \cdot x - b_j); \tag{1.9}$$

the α_j and b_j are scalars and the k_j are d-dimensional vectors. A one hidden-layer feedforward neural network is then a superposition of m ridge functions which are often called neurons. The activation function σ is usually sigmoidal, a terminology which means that σ is bounded and monotone with e.g. $\lim_{t\to-\infty} \sigma(t) = 0$ and $\lim_{t\to\infty} \sigma(t) = 1$. Popular choices for σ are the Heavyside $\sigma(t) = 1_{\{t>0\}}$ and the logistic function $\sigma(t) = 1/(1+e^{-t})$.

Approximation theory. Many of the problems in the neural nets methodology are of an approximation theoretic nature, hence the significant place of ridge functions in the literature of approximation theory. Throughout the remainder of this paper, $\mathcal{D}_{NN}$ will denote the collection of neurons (neural network dictionary)

$$\mathcal{D}_{NN} = \{\sigma(a \cdot x - b),\ k \in \mathbb{R}^d, b \in \mathbb{R}\}. \tag{1.10}$$

The most fundamental questions in this field are a recast of those formulated in a preceding paragraph. For instance, one can ask about the capabilities of this dictionary. As an example, we may be curious about the speed of convergence of finite linear combination of neurons to a given target f as the number of neurons n is increasing and tending to infinity. Or about how well does $\mathcal{D}_{NN}$ approximate a given functional class $\mathcal{F}$. In addition, a central question is about the existence of very concrete procedures with good approximation properties. Later sections will review

some results in this area but in truth, very little is known about neural network as a form of approximation. We quote from Petrushev [53] "It is surprising, therefore, that the most fundamental questions concerning the efficiency of approximation by ridge functions are unanswered." There seems to be a general consensus about this, see Pinkus [54] who writes "there is very, very little known about the degree of approximation by ridge functions."

We should emphasize that in order to talk about the speed of convergence, one should, of course, better verify that $\mathcal{D}_{NN}$ is, in some sense, complete. Various completeness results are known for neural networks [13,17,32,34]. These results are fairly recent and we introduce two of them. First, Cybenko [42] shows that the span of $\mathcal{D}_{NN}$ is dense in $C(K)$ for any compact set $K \subset \mathbb{R}^d$, for any continuous σ obeying $\lim_{t\to-\infty} \sigma(t) = 0$ and $\lim_{t\to\infty} \sigma(t) = 1$. This means that for every function $f \in C(K)$ and every $\epsilon > 0$, one can find g in the span of $\mathcal{D}_{NN}$ with the property

$$\sup_K |f(x) - g(x)| < \epsilon.$$

Note that this result gives completeness in other spaces such as in $L_2(K)$ since $C(K)$ is dense in $L_2(K)$. Second, Leshno, Lin, Pinkus and Schocken [42] obtained a rather definitive result concerning the density property of of $\mathcal{D}_{NN}$. They proved that continuous activation σ yields the density in $C(K)$ if, and only if, σ is not a polynomial.

Many mathematical problems in the field of neural networks are of an approximation-theoretic nature. We shall first discuss the problem of constructing neural net approximations, and then review what is known about their degree of approximation.

1.5 Construction of neural networks

Fundamental questions remain open about the computational efficiency of neural networks. First and foremost, it is unclear how to construct neural networks. A previous article [4] pointed out this major conceptual weakness, and our exposition will closely follow the argument presented in that paper. The problem here is that we do not really know how to represent a multivariate function as a superpositions of neurons $\sigma(k\cdot x-b)$. This is in stark contrast with some other areas of approximation theory such as polynomial, Fourier or wavelet approximations.

Because of the lack of synthesizing principle, any approximation procedure amounts to minimizing highly multimodal error surfaces. For instance, to find the best m-term approximation

$$f_m(x) = \sum_{j=1}^{m} \alpha_j \sigma(k_j \cdot x - b_j),$$

one has to solve the following optimization problem

$$\inf \ \|f - f_m\|_{L_2}$$

where the infimum ranges over the coefficients $(\alpha_j)_{j=1}^m$ and the network parameters $(k_j)_{j=1}^m$ and $(b_j)_{j=1}^m$. Solving such problems belong to the realm of dreams. We quote Barron [16]: "There is no known algorithm... Gradient search and its variants produce a local optimum of dubious scientific merit." There is more. In a practical setting where one is given sampled data $(x_i, y_i)_{i=1}^N$, Vu [62] showed that finding the minimum of

$$\sum_i (y_i - f(x_i))^2$$

or even an *approximate* minimum is NP-hard, as soon as $m \geq 2$. Vu improved upon the pioneering work of Jones [37]; the title of Jones' article, "The computational intractability of training sigmoidal neural networks" surely drives our point home. The aim of this line of research is to show that it is impossible to design algorithms running in polynomial time that would produce 'accurate estimates' (the exact formulation is that this problem is NP-hard and it is a conjecture that NP-hard problems cannot be solved in polynomial time).

A possibly more reasonable method of approximation is the greedy algorithm as discussed above. This algorithm and its variants synthesize the approximation f_m through a greedy stepwise addition of terms; begin with $f_0(x) = 0$, the relaxed greedy algorithm inductively defines for each $i = 1, \ldots, m$

$$f_i = \alpha^* f_{i-1} + (1 - \alpha^*)\sigma(k^* \cdot x - b^*), \tag{1.11}$$

where (α^*, k^*, b^*) are solutions of the optimization problem

$$\min_{0 \leq \alpha \leq 1} \min_{(k,b) \in \mathbb{R}^n \times \mathbb{R}} \|f - \alpha f_{i-1} + (1 - \alpha)\sigma(k \cdot x - b)\|_2. \tag{1.12}$$

At stage i, the algorithm updates the current approximation f_{i-1} with a convex combination involving f_{i-1} and a new term, a neuron $\sigma(k \cdot x - b)$, that results in the largest decrease in approximation error (1.12). As we pointed out, this strategy sounds more concrete, but is still subject to some rather serious objections.

- *Algorithm?* At each stage, there are many feasible choices (α_i, k_i, b_i), and the minimization (1.12) involves a nonlinear search over those parameters. This is a nonconvex problem, and to the author's knowledge there is no obvious practical algorithm for solving (1.12). (In a discrete setting, there is work showing that the number of local minima may be bounded below by $C \cdot N^d$, where N is the sample size and

d the dimension of the space [1]). In a realistic implementation, one would need to discretize the set of parameters to perform a search. How fine does the discretization of the network parameters need to be?

- *Stability.* The greedy approach does not yield stable decompositions. A small perturbation of the input function f will typically produce radically different parameter values. In other words, these parameters do not have any reliable scientific meaning.

- *Efficiency.* The work of DeVore and Temlyakov proves that the greedy algorithm obeys very weak approximation bounds even when good approximations exist. To be more concrete, it is possible to synthesize a target function as a superposition of only two elements of the neural net dictionary and prove that the greedy algorithm, producing a sequence of m-term approximations, will converge at the rate $O(m^{-1/2})$ and not faster. And the target is only a superposition of two terms!

  These weak properties are well-known to engineers and statisticians. In statistics, this says that stepwise regression may be severely inefficient for model selection. In signal processing, where the greedy algorithm is also known under the name of "matching pursuit," it is known that the inability to look ahead may cause initial errors the algorithm will keep on trying to correct. Chen et al. [14,15] give a sequence of rather spectacular computational examples in this direction.

We say "that there is no obvious practical algorithm for solving (1.12)" in light of the scientific standards in common use in numerical optimization. There, the word algorithm has a very precise meaning, namely, that of a procedure which, in a given number of steps, gives the minimum guaranteed or an approximate minimum with a ticket which says how far we are from the minimum, see the literature on Linear Programming (LP) or Semi-Definite Programming (SDP), for example.

It is not the author's intention to sound sweepingly negative. As a matter of fact, the neural network methodology is used quite successfully in many applications. The point here is that, by and large, this methodology seems removed from mathematical and scientific standards. Success is often measured in terms of how well does a particular methodology perform on a specific example rather than on the scientific underpinning it provides. This is not a criticism, merely a fact. The neural network methodology is a guiding principle and researchers in this field are happy trying everything that works well in practice. However, true scientific progress has to do with a better understanding of these mathematical models.

1.6 Approximation by ridge functions

In this section, we will consider the degree of approximation with n-term approximations from $\mathcal{D}_{NN}$. For each n, put

$$\Sigma_n = \Big\{ \sum_{j=1}^{n} \alpha_j \sigma(k_j \cdot x - b_j),\ \alpha_j \in \mathbb{R}, k_j \in \mathbb{R}^d, b_j \in \mathbb{R} \Big\}.$$

There is a body of work which tells us that ridge function or neural network approximation is as efficient as other means of approximation such as splines or polynomials for approximating classical smoothness classes which we now define.

We first introduce some notations. Let $\alpha = (\alpha_1, \ldots, \alpha_d)$ be a d-uple of nonnegative integers and D^α be the partial derivative $D^\alpha = \partial_1^{\alpha_1} \cdots + \partial_d^{\alpha_d}$. Put $|\alpha| = \alpha_1 + \ldots \alpha_d$. Further, we set m to be a nonnegative integer and Ω to be an open set of $\mathbb{R}^d$. The Sobolev space $W_p^m(\Omega)$ is the completion of $C^m(\Omega)$ with respect to the norm

$$\|f\|_{W_p^m} = \|f\|_{L_p(\Omega)} + \sum_{\alpha:|\alpha|=m} \|D^\alpha f\|_{L_p(\Omega)}, \quad p \in [1, \infty]. \tag{1.13}$$

An object f belongs to the space W_p^m if it has finite Sobolev norm or, in other words, if f and all of its partial derivatives up to order m are in L_p. Interpolation theory allows the extension of Sobolev norms to the half-line $m \geq 0$ and we will omit this definition. Finally, define Sobolev balls $W_p^m(C)$ by

$$W_p^m(C) = \{f \in W_p^m,\ \|f\|_{W_p^m} \leq C\}$$

meaning that one get control of the size of f and its derivatives up to order m. Unless specified otherwise, we will take Ω to be the unit ball D of $\mathbb{R}^d$, $D = \{x,\ x_1^2 + \cdots + x_d^2 \leq 1\}$.

Mhaskar [49] proves the following result. Assume that the activation function σ is C^∞ and that σ is not a polynomial. Then for each $p \in [1, \infty]$ and $m \geq 1$

$$d_n(W_p^m(C), \mathcal{D}_{NN}; L_p) \equiv \sup_{f \in W_p^m(C)} \inf_{g \in \Sigma_n} \|f - g\|_{L_p} \leq K_{s,p} \cdot C \cdot n^{-m/d}. \tag{1.14}$$

There are converse results as well. For instance, letting σ be the logistic sigmoid, $\sigma(t) = 1/(1 + e^{-t})$, Maiorov and Meir [45] prove that

$$d_n(W_p^m(C), \mathcal{D}_{NN}; L_p) \geq K'_{s,p} \cdot C \cdot (n \log n)^{-m/d}. \tag{1.15}$$

This shows that the degree of approximation $n^{-m/d}$ is, in some sense, optimal. In fact, [19] showed that the Sobolev balls cannot be approximated

at a rate faster than $n^{-m/d}$ by any reasonable means of approximation. By 'reasonable,' we mean a method of approximation which depends continuously on f, see the above reference for details. Also, the exponent m/d is that appearing in the Kolmogorov ϵ-entropy, and in the minimax risk of estimation over Sobolev balls.

We would like to mention another interesting result due to Petrushev [53] which is an extension of an earlier work from DeVore, Oskolkov and Petrushev [21]. Set $I = [-1, 1]$. We let X_n be a linear space of univariate functions in $L_2(I)$ of dimension n, and let Θ_n be a finite subset of the unit sphere of $\mathbb{R}^d$. The collection of functions of the form

$$g(x) = \sum_{\theta \in \Theta_n} \alpha_\theta \rho_\theta(\theta \cdot x), \quad \rho_\theta \in X_n,\ \theta \in \Theta_n\}$$

is a linear space Y_n of dimension $\leq n \times \#\Theta_n$. We suppose that X_n obeys

$$\inf_{h \in X_n} \|g - h\|_{L_2(I)} \leq C_s \cdot n^{-s} \cdot \|g\|_{W_2^s(I)}. \tag{1.16}$$

Then Petrushev proves that for appropriately chosen sets Θ_n of cardinality $O(n^{d-1})$, we have

$$\inf_{g \in Y_n} \|f - g\|_{L_2(D)} \leq C_r \cdot n^{-r} \cdot \|g\|_{W_2^r(D)}, \quad r = s + (d-1)/2. \tag{1.17}$$

Note that the cardinality of Y_n obeys $\#Y_n = O(n^d)$. For instance we may construct X_n as the univariate space spanned by $\sigma(nt - k)$, $0 \leq k \leq n$ and assume that the activation function σ is chosen such that (1.16) holds. Then define Y_n as described above, i.e.

$$Y_n \ = \ \text{span}\{\sigma(n\, x \cdot \theta_n - k),\, \theta_n \in \Theta_n, 0 \leq k < n\}.$$

Then Y_n obeys (1.17). Petrushev remarks that there is "an unexpected gain of $(d-1)/2$ in the approximation order." To describe this phenomenon, consider the case where σ is the Heavyside $\sigma(t) = 1_{\{t>0\}}$, and observe that for this choice of sigmoid, X_n obeys (1.16) for $s = 1$. Then Y_n approximates elements with $1 + (d-1)/2$ derivatives at the optimal rate, even though σ, and by the same token the elements of Σ_n, are discontinuous. We shall return to this point later. The point of Petrushev's findings is that the condition (1.16) is about the approximation of univariate functions which automatically translates into corresponding approximation properties for multivariate smoothness classes by superpositions of ridge functions.

We remark that (1.17) is a result about linear approximation which, in some sense, is an extension of Mhaskar's theorem (1.14) since it allows

for very general X_n. On the other hand, however, it is only established for $p = 2$.

We would like to point the reader to other references in this direction of research such as [44] and [52] which give other results with means of approximation which are not continuous. See also, Pinkus [55] for a rapid tour of those references.

This line of research shows, essentially, that neural networks enjoy the same degree of approximation as polynomials over classical smoothness classes. In fact, polynomials play a central role in the proofs of the above results. For instance, the key observation behind Mhaskar's theorem (1.14) is that for every integer $n \geq m$ there exists a polynomial $\pi_n(f)$ of coordinatewise degree not exceeding n such that for every $f \in W_p^m$, we have

$$\|f - \pi_n(f)\|_{L_p} \leq C \cdot n^{-m} \cdot \|f\|_{W_p^m}.$$

The idea is that one can then approximate each monomial of $\pi_n(f)$ with finite differences of neurons, see [49]. There is a similar idea underlying the result (1.17) of Petrushev. The starting point is again to use polynomials to approximate elements of W_2^s, and to decompose those into 'ridge' polynomials. The profile of those 'ridge' polynomials are of course polynomial and one can use X_n to approximate those univariate polynomials.

The point of this research, however, is not very clear to the author. We have available a very concrete, stable and extraordinarily well-understood method of approximation by polynomials. Why should we prefer obscure methods of approximations by neural networks which are often unstable [49] and algorithmically nonconstructive? Isn't it possible to introduce function classes over which neural networks are arguably better than other means of approximation? In some sense, this is the spirit of the work of Barron we introduce next.

1.7 Barron's results

Barron [2] published a result about the degree of approximation of neural networks which had an enormous impact upon the community. Letting $\hat{f}$ be the Fourier transform of f, Barron introduces the class of functions

$$\mathcal{B}(C) = \Big\{ f \in L_2([0,1]^d) : \int_{\mathbf{R}^d} |\xi| |\hat{f}(\xi)| \, d\xi \leq C \Big\}. \tag{1.18}$$

(Here and throughout the remainder of the paper, the notation $|\cdot|$ stands for the Euclidean norm.) Barron shows that for this class, the relaxed greedy algorithm (1.11)–(1.12) produces a sequence of approximations with the property

$$\|f - f_n\|_2 \leq 2 \cdot C \cdot n^{-1/2}, \tag{1.19}$$

where f_n is the output of the algorithm at stage n. One notices that the exponent of convergence does not depend upon the dimension d, which launched a curious discussion about the fact that neural networks defeated the curse of dimensionality.

The argument, here, is that Barron's class is included in the convex hull of the neural net dictionary. In other words, every $f \in \mathcal{B}(C)$ can be written as

$$f = \sum_j \alpha_j \sigma(k_j \cdot x - b_j), \quad \text{with} \quad \sum_j |a_j| \leq C. \tag{1.20}$$

A stochastic argument which goes back to Maurey [56] then shows that if f belongs to the convex hull of $\{\sigma(k \cdot x - b)\}$, there exists a sequence of n-term approximations which converge at the rate $n^{-1/2}$. Further, in a remarkable paper, Jones [36] showed that the greedy algorithm converges at this same rate.

On the one hand, we already highlighted the lack of constructive character of the greedy algorithm which does not turn (1.19) into a very constructive result. On the other hand, the dictionary of orthogonal sinusoids is optimal for approximating elements of $\mathcal{B}(C)$, which is hardly surprising since the class $\mathcal{B}(C)$ is defined by means of the Fourier transform. We let $\mathcal{D}_F$ be the classical orthobasis $\{e^{i2\pi k\cdot x},\ k \in \mathbb{Z}^d\}$, and let f_n^F be the approximation obtained by keeping only the n largest terms of the trigonometric series –ironically, trigonometric exponentials are ridge functions. We have

$$\sup_{f\in\mathcal{B}(C)} \|f - f_n^F\|_{L_2} \leq C \cdot n^{-1/2-1/d}. \tag{1.21}$$

Roughly, this follows from the equivalence (f is compactly supported)

$$\int_{\mathbb{R}^d} |\xi||\hat{f}(\xi)|\, d\xi \asymp \sum_{k\in\mathbb{Z}^d} |k||\hat{f}(2\pi k)|,$$

which is a simple consequence of a famous theorem about the sampling of bandlimited functions due to Polỳa and Plancherel [57]. Therefore, $f \in \mathcal{B}(C)$ implies that the Fourier coefficients $c_k(f)$ of f obey

$$\sum_k |k|\, |c_k(f)| \leq C. \tag{1.22}$$

The inequality (1.22) gives a bound on the decay coefficient sequence of f. Skipping the details and letting $|c(f)|_{(n)}$ be the nth largest entry in the sequence, (1.22) gives

$$\sum_{m>n} |c(f)|_{(m)}^2 \leq C \cdot n^{-(1+2/d)},$$

and, therefore, (1.21).

Actually, Makovoz [46] later improved the bound (1.19), and showed

$$d_n(\mathcal{B}(C), \mathcal{D}_{NN}) \leq C \cdot n^{-1/2-1/2d}. \tag{1.23}$$

This upperbound is still not as good as the rate provably obtained by trigonometric approximations (1.21).

In short, these results are not very satisfying as one may exhibit other well-established method of approximation, namely, the thresholding of Fourier series, with at least as good approximation properties.

1.8 Key issues

Neural networks are used everyday for approximation, prediction, pattern recognition, etc. in the applied sciences and engineering. This is a gigantic field; there are several journals in the field with several thousands of papers published on neural networks, many annual conferences and several dozens of textbooks on this subject. And yet, there seems to be no real theoretical basis, and also practical issues such as the construction of neural nets need to be addressed. From the viewpoint of approximation, there is a need to understand the properties of neural net expansions, to understand what they can and what they cannot do, and where they do well and where they do not. How and which type of functions should we approximate with ridge functions? How do we develop approximation bounds?

1.9 A different approach

The work surveyed in this paper suggests that there is a very different way to go about this problem. We develop transforms which allow the representation of arbitrary objects by superpositions of ridge functions. Those transforms can be used to construct stable approximations. An analogy may be helpful to illustrate this shift in emphasis [27].

In one dimension, consider the problem of approximating a function f by dilations and translations $\sigma(at-b)$, $a > 0, b \in \mathbb{R}$ of a single template σ. If σ is sigmoidal, this is far from being obvious. Now, substitute the sigmoid σ with an oscillatory profile ψ. Harmonic analysis tells us that this becomes a much better posed problem. Indeed, the wavelet transform allows the representation of rather arbitrary signals as superpositions of elements of the form $\psi(at-b)$. Moreover, one has available wavelet orthobases and fast algorithms which yield very concrete procedures for obtaining stable n-term approximations. This is especially relevant for practical applications. The philosophy is the same, namely, that of approximating objects with dilations and translations of a single (or a few) templates. Only the shape of this template has changed.

Just as the wavelet transform allows the representations of arbitrary functions with superpositions of dilations and translations of a single function, the ridgelet transform will allow the representations of multivariate objects with dilated, rotated and translated versions $\psi(au \cdot x - b)$ of a single ridge function, say, $\psi(x_1)$. Like in wavelet theory, there is both a discrete and a continuous transform which we now introduce.

§2. Ridgelets

This section introduces the ridgelet transforms and surveys some of their main properties. All of the forthcoming claims and results are proved in [4]. For now, $\hat{g}$ will denote the Fourier transform of g,

$$\hat{g}(\xi) = \int_{\mathbb{R}^d} f(x) e^{-ix\cdot\xi}\, dx. \tag{2.1}$$

2.1 The continuous ridgelet transform

In d dimensions, the ridgelet construction starts with a univariate function ψ satisfying an admissibility condition, namely,

$$K_\psi = \int |\hat{\psi}(\xi)|^2/|\xi|^d \, d\xi < \infty; \tag{2.2}$$

this condition says that ψ is oscillatory and has vanishing moments up to about $d/2$. Here, the number of vanishing moments grows linearly with the dimension of the space. Sigmoidal activation functions in use in the theory of neural networks are not admissible. A ridgelet is a function of the form

$$\frac{1}{a^{1/2}} \psi\left(\frac{u\cdot x - b}{a}\right), \tag{2.3}$$

where a and b are scalar parameters and u is a vector of unit length. Of course, a ridgelet is a ridge function and resembles a neuron but for the oscillatory behavior of the profile. A ridgelet has a scale a, an orientation u, and a location parameter b. Ridgelets are concentrated around hyperplanes: roughly speaking the ridgelet (2.3) is supported near the strip $\{x, |u \cdot x - b| \le a\}$. Ridgelets are pictured in Fig. 1 for various values of these parameters.

Define a ridgelet coefficient as

$$\mathcal{R}_f(a,u,b) = \int f(x)\, a^{-1/2} \psi(\frac{u\cdot x - b}{a})\, dx; \tag{2.4}$$

then for any $f \in L_1 \cap L_2(\mathbb{R}^d)$, we have

$$f(x) = \int \mathcal{R}_f(a,u,b) a^{-1/2} \psi(\frac{u\cdot x - b}{a})\, d\mu(a,u,b), \tag{2.5}$$

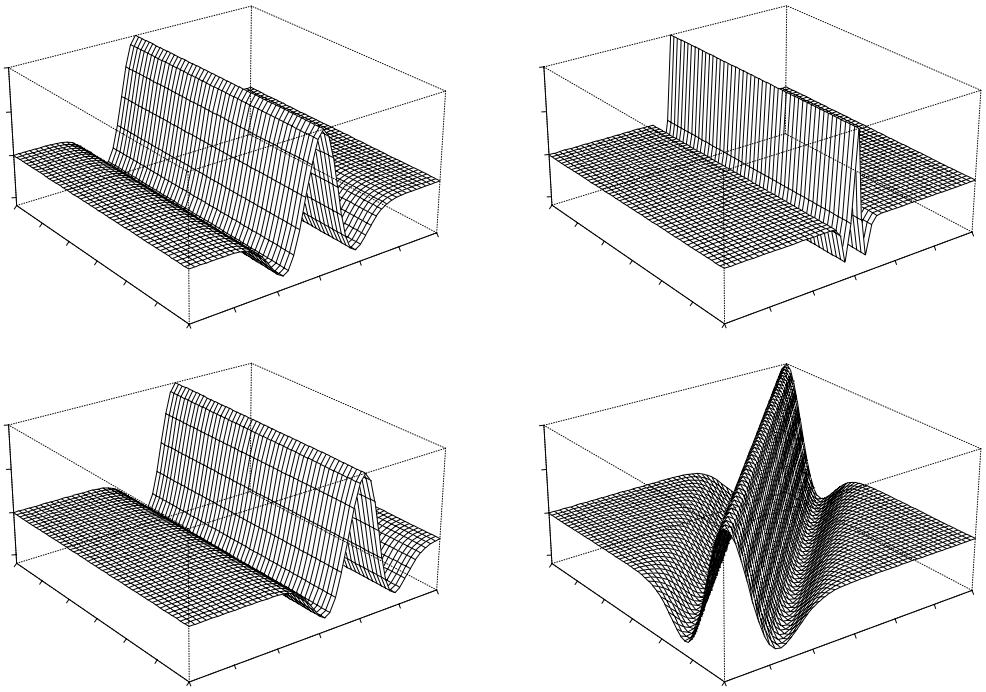

Fig. 1. Ridgelets.

where $d\mu(a,u,b) = da/a^{d+1}\, du\, db$ (du being the uniform measure on the sphere) which holds true if ψ is properly normalized, i.e. $K_\psi = 1/(2\pi)^{d-1}$ in (2.2). The formula (2.5) expresses the idea that one can represent any function as a superposition of these ridgelets. In fact, this formula has been independently discovered by two other researchers, namely, Murata [50] and Rubin [58]. The former is working on neural networks while the latter is working in the field of mathematical analysis. The interpretation of the formula is classical. For $f \in L_1 \cap L_2(\mathbb{R}^d)$, put

$$f_{\epsilon,\delta}(x) = \int_{\epsilon<a<\delta} \int_{\mathbf{S^{d-1}}} \int_{\mathbb{R}} \mathcal{R}_f(a,u,b) a^{-1/2} \psi(\frac{u\cdot x - b}{a})\, d\mu(a,u,b);$$

$f_{\epsilon,\delta}$ is well-defined and obeys

$$\|f - f_{\epsilon,\delta}\|_{L_2} \to 0 \text{ as } \epsilon \to 0,\ \delta \to \infty.$$

Furthermore, the reproducing formula is stable as one has a Parseval relation

$$\|f\|_2^2 = \int |\mathcal{R}_f(a,u,b)|^2 d\mu(a,u,b). \tag{2.6}$$

Like in Fourier or wavelet analysis, this says that a perturbation of the function (resp. the coefficient sequence) has a well-controlled effect on the coefficient sequence (resp. the reconstructed object).

As in Littlewood-Paley or wavelet theory, one may want to discretize the scale on a a dyadic lattice. Let us choose a profile ψ obeying

$$\sum_{j\in\mathbb{Z}} \frac{|\hat{\psi}(2^{-j}\xi)|^2}{|2^{-j}\xi|^{d-1}} = K'_\psi, \tag{2.7}$$

a condition which greatly resembles the admissibility condition (2.2) introduced earlier. Note that if one is given a function Ψ obeying

$$\sum_{j\in\mathbb{Z}} |\hat{\Psi}(2^{-j}\xi)|^2 = c$$

as in wavelet theory, ψ defined by $\hat{\psi}(\xi) = |\xi|^{(d-1)/2}\hat{\Psi}(\xi)$ will verify (2.7). Assume the special normalization $2K'_\psi = (2\pi)^{-(d-1)}$ in (2.7). Then

$$f(x) = \sum_{j\in\mathbb{Z}} 2^{jd} \int \mathcal{R}_f(2^{-j}, u, b) 2^{j/2} \psi(2^j(u\cdot x - b))\, du\, db, \tag{2.8}$$

where again the inequality holds in an L_2 sense; for $f \in L_1 \cap L_2(\mathbb{R}^d)$, the partial sums of the right-hand side are square integrable and converge to f in L_2.

Finally, as in wavelet theory, we may introduce some special coarse scale ridgelets. We choose a profile φ such that $\xi \in \mathbb{R}$

$$\frac{|\hat{\varphi}(\xi)|^2}{|\xi|^{d-1}} + \sum_{j\geq 0} \frac{|\hat{\psi}(2^{-j}\xi)|^2}{|2^{-j}\xi|^{d-1}} = K'_\psi. \tag{2.9}$$

Note that the above equality implies $|\hat{\varphi}(\xi)|^2 \leq |\xi|^{d-1}$, which is very much *unlike* Littlewood-Paley or wavelet theory: our coarse scale ridgelets are also oscillating since $\hat{\varphi}$ must have some decay near the origin, that is, φ itself must have some vanishing moments.

Let (φ, ψ) be a pair obeying (2.9) with $2K'_\psi = (2\pi)^{-(d-1)}$. Then

$$\begin{aligned} f = &\int \mathcal{R}^0_f(u,b)\varphi(u\cdot x - b)\, du\, db \\ &+ \sum_{j\geq 0} 2^{jd} \int \mathcal{R}_f(2^{-j}, u, b) 2^{j/2} \psi(2^j(u\cdot x - b))\, du\, db, \end{aligned} \tag{2.10}$$

with

$$\mathcal{R}^0_f(u,b) = \int f(x)\varphi(u\cdot x - b)\, dx,$$

and $\mathcal{R}_f$ as before.

2.2 The discrete ridgelet transform

Similar to the continuous transform, there is a discrete transform. Let R be the triple (j, ℓ, k) where the indices run as follows

$$R \in \mathcal{R} := \{(j, \ell, k),\ j, k \in \mathbb{Z}, j \geq j_0, \ell \in \Lambda_j\},$$

and define the collection of discrete ridgelets

$$\psi_R(x) = 2^{j/2}\psi(2^j u_{j,\ell} \cdot x - k), \quad R \in \mathcal{R}. \tag{2.11}$$

Note that the range of the parameter ℓ is scale dependent as it depends on j. Ridgelets are directional and, here, the interesting aspect is the discretization of the directional variable u; this variable is sampled at increasing resolution so that at scale j, the discretized set is a net of nearly equispaced points at a distance of order 2^{-j}; a detailed exposition on the ridgelet construction is given in [4].

The key result is that the discrete collection of ridgelets $(\psi_R)_{R\in\mathcal{R}}$ is complete in $L_2[0,1]^d$, and any function f can be reconstructed from the knowledge of its coefficients $(\langle f, \psi_R\rangle)_{R\in\mathcal{R}}$. (The notation $\langle\cdot,\cdot\rangle$ stands here and throughout this paper for the usual inner product of L_2: $\langle f, g\rangle = \int f(x)g(x)dx$.) There exist two constants A and B such that for any $f \in L_2[0,1]^d$, we have

$$A\,\|f\|^2 \leq \sum_{R\in\mathcal{R}} |\langle f, \psi_R\rangle|^2 \leq B\,\|f\|^2. \tag{2.12}$$

The previous equation says that the datum of the ridgelet transform at the points $(a = 2^j, u = u_{j,\ell}, b = k2^{-j})_{(j,k,\ell)\in\mathcal{R}}$ suffices to reconstruct the function perfectly. In this sense, this is analogous to the Shannon sampling theorem for the reconstruction of bandlimited functions. Indeed, standard arguments show that there exists a dual collection $(\tilde{\psi}_R)_{R\in\mathcal{R}}$ with the property

$$f = \sum_{R\in\mathcal{R}} \langle f, \tilde{\psi}_R\rangle \psi_R = \sum_{R\in\mathcal{R}} \langle f, \psi_R\rangle \tilde{\psi}_R, \tag{2.13}$$

which gives perfect and stable reconstruction.

2.3 Frequency-side picture

In d-dimensions, ridgelets are localized around planes of codimension 1. To make things concrete, consider the situation in two dimensions where the discretization of the angular variable is as follows: we let $u_{j,\ell} = (\cos\theta_{j,\ell}, \sin\theta_{j,\ell})$ and for a fixed discretization step $\alpha > 0$, set

$$\theta_{j,\ell} = \alpha\,\pi\,2^{-j}\,\ell, \quad \ell = 0, 1, \ldots, \lfloor 2^j/\alpha \rfloor.$$

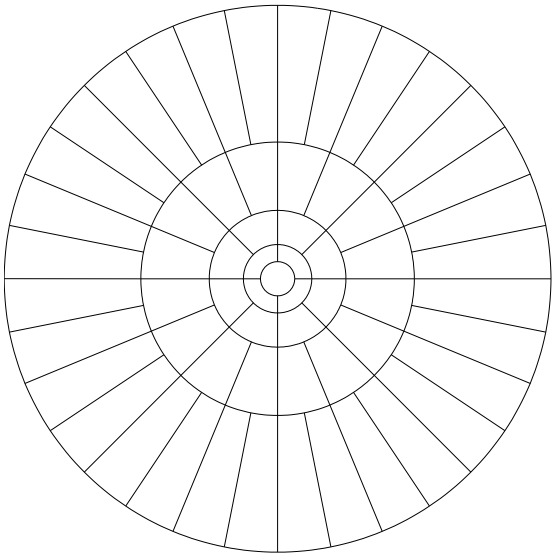

Fig. 2. Illustration of ridgelet sampling scheme in the frequency plane.

Ridgelets are oriented in the codirection $\theta_{j,\ell} = \ell/2^j$ and located near the line $x_1 \cos\theta_{j,\ell} + x_2 \sin\theta_{j,\ell} - k2^{-j} = 0$. Their width is about 2^{-j} so that we can think of ridgelets as a collection of fat lines at different scales.

The frequency-side picture also gives some nice insights about the organization of the ridgelet transform. In the Fourier domain, a ridge function $\rho(k \cdot x)$ is, in fact, supported on the radial line (λk), $\lambda \in \mathbb{R}$. Then, the Fourier transform of a ridgelet ψ_R (polar coordinates) is given by

$$(\hat{\psi}_R)(\lambda\cos\theta, \lambda\sin\theta) = \hat{\psi}_I(2^{-j}\lambda)\delta(\theta - \theta_{j,\ell}), \quad \lambda \in \mathbb{R},\ \theta \in [0, 2\pi), \quad (2.14)$$

where ψ_I is the dyadic wavelet $\psi_I(t) = 2^{j/2}\psi(2^j t - k)$, $I = (j,k)$. Recall the Fourier transform of ψ_I

$$\hat{\psi}_I(\lambda) = 2^{-j/2}\,\psi(2^{-j}\lambda)\,e^{\mathbf{i}k2^{-j}}.$$

Therefore, in the frequency plane, ridgelets are supported near the dyadic segments represented on Fig. 2. The angular sampling principle is explicit; the number of ridgelet orientations is doubled from one dyadic corona to the next. We close this section by summarizing the main points of the discrete ridgelet transform.

- Like in Littlewood-Paley theory, divide the frequency domain into dyadic annuli $|x| \in [2^j, 2^{j+1})$.
- Sample each annulus with radial lines $\theta = \alpha\pi 2^{-j}\ell$.
- The angular resolution increases with the spatial resolution as illustrated in Fig. 2.

2.4 Connection with neural nets

The introduction described a popular approach – the greedy algorithm – to compute neural net approximations. At each step, one would need to solve an optimization problem of the form (1.12), and in any real implementation, one would probably need to restrict the search for a minimum over a grid. What are the properties of a restricted search? Is there a grid preserving the completeness property? If so, what is the proper spacing of this grid? In other words, what is the real complexity of the search (1.12)? In some sense, the discretization (2.11) gives a precise answer to these questions.

§3. Orthonormal Ridgelets

In dimension 2, Donoho [26] introduced a new orthonormal basis whose elements he called 'orthonormal ridgelets.' We quote from [10]: "Such a system can be defined as follows: let $(\psi_{j,k}(t) : j \in \mathcal{Z}, k \in \mathcal{Z})$ be an orthonormal basis of Meyer wavelets for $L^2(\mathbb{R})$ [41], and let $(w^0_{i_0,\ell}(\theta),\ \ell = 0, \ldots, 2^{i_0} - 1;\ w^1_{i,\ell}(\theta),\ i \geq i_0,\ \ell = 0, \ldots, 2^i - 1)$ be an orthonormal basis for $L^2[0, 2\pi)$ made of periodized Lemarié scaling functions $w^0_{i_0,\ell}$ at level i_0 and periodized Meyer wavelets $w^1_{i,\ell}$ at levels $i \geq i_0$. (We suppose a particular normalization of these functions). Let $\hat{\psi}_{j,k}(\omega)$ denote the Fourier transform of $\psi_{j,k}(t)$, and define ridgelets $\rho_\lambda(x)$, $\lambda = (j, k; i, \ell, \varepsilon)$ as functions of $x \in \mathbb{R}^2$ using the frequency-domain definition

$$\hat{\rho}_\lambda(\xi) = |\xi|^{-\frac{1}{2}} (\hat{\psi}_{j,k}(|\xi|) w^\varepsilon_{i,\ell}(\theta) + \hat{\psi}_{j,k}(-|\xi|) w^\varepsilon_{i,\ell}(\theta + \pi))/2 \ . \tag{3.1}$$

Here the indices run as follows: $j, k \in \mathbb{Z}$, $\ell = 0, \ldots, 2^{i-1} - 1$; $i \geq i_0$, $i \geq j$. Notice the restrictions on the range of ℓ and on i. Let λ denote the set of all such indices λ. It turns out that $(\rho_\lambda)_{\lambda \in \Lambda}$ is a complete orthonormal system for $L^2(\mathbb{R}^2)$."

Orthonormal ridgelets are *not* ridge functions, although there is a close connection between 'pure' and orthonormal ridgelets. Using a formulation emphasizing the resemblance with (3.1), the frequency representation of a pure ridgelet is given by

$$\hat{\psi}_{j,\ell,k}(\xi) = (\hat{\psi}_{j,k}(|\xi|)\delta(\theta - \alpha\pi 2^{-j}\ell) + \hat{\psi}_{j,k}(-|\xi|)\delta(\theta + \pi - \alpha\pi 2^{-j}\ell))/2. \tag{3.2}$$

In the ridgelet construction, the angular variable θ is uniformly sampled at each scale; the sampling step being proportional to the scale 2^{-j}. In contrast, the sampling idea is replaced by the wavelet transform for orthonormal ridgelets. The orthonormal ridgelet basis is a tensor basis in the polar coordinate system. It is properly renormalized by the Jacobian

underlying the Cartesian to polar change of coordinates so that it yields an orthobasis of $L^2(\mathbb{R}^2)$. It is interesting to note that the restriction on the range, namely, $i \geq j$ in the definition (3.1), gives the same angular scaling as in the original ridgelet construction.

§4. Ridgelet Thresholding

Whereas the construction of neural networks involved a delicate stepwise construction of approximations, ridgelet analysis gives an *explicit and stable* formula for representing a function as a superposition of ridge functions. Further, the expansion (2.13) allows the construction of simple finite approximations using naive ideas like thresholding. For instance, we may truncate the exact series

$$f = \sum_R \theta_R \tilde{\psi}_R, \quad \theta_R = \langle f, \psi_R \rangle,$$

by extracting the terms corresponding to the n largest coefficients; that is,

$$f_n^R = \sum_\lambda \theta_R 1_{\{|\theta_R| \geq \delta\}} \tilde{\psi}_R, \tag{4.1}$$

where δ is chosen so that $\#\{R, |\theta_R| \geq \delta\} = n$, i.e. δ is about the nth largest entry of the sequence $|\theta|$ (in case of ties, i.e. $|\theta_R| = \delta$ for several R's, any subset will do).

We note that (4.1) is a simple, constructive and stable method of approximation, and we propose this strategy as an alternative to the ill-posed construction of neural networks. The philosophy, however, is the same; that of approximating multivariate objects by finite linear combinations of dilations, translations and rotations of a single ridge function.

§5. Ridgelets and Neural Networks

Ridgelet thresholding may seem like a very naive strategy. As a first impulse, one would imagine that it would not be able to match the approximation performance of possibly very abstract neural network approximation strategies. This section shows that, in some sense, this intuition is wrong. Ignoring boundary issues, we claim that there is no such function which is approximated at a faster rate, in an asymptotic sense, with sigmoidal feedforward neural networks than with ridgelet thresholding. We now detail this claim and refer the reader to [6] for further reference.

Assume σ is the logistic function $\sigma(t) = (1 + e^{-t})$, and consider a sequence $(f_n)_{n \geq 1}$ of feedforward neural networks

$$f_n^{NN} = \sum_{i=1}^{n} \alpha_i^n \sigma(k_i^n \cdot x - b_i^n) \tag{5.1}$$

with coefficients having polynomial growth, that is

$$\sup_i |\alpha^n| = O(n^\beta), \quad \text{for some} \quad \beta \geq 0. \tag{5.2}$$

We emphasize that the parameters of the network, namely the directions and locations, are allowed to change as the number of terms or neurons in the approximation increases. The restriction concerning the growth of the coefficients α^n (5.2) is very mild. It says that the coefficients have at most polynomial growth in the number of terms which prevents the consideration of very wild approximants that would practically be irrelevant, see the discussion in [6].

Theorem 5.1. *Let $f \in L_2(\mathbb{R}^d)$ be supported in the unit ball D of $\mathbb{R}^d$, and suppose that there is a sequence (f_n) of feedforward neural networks (5.1) obeying*

$$\|f - f_n^{NN}\|_{L_2(2D)} = O(n^{-r}).$$

Then the n-term approximations f_n^R (4.1) obtained by simply thresholding the exact ridgelet series of f obeys

$$\|f - f_n^R\|_{L_2(2D)} = O(n^{-r+\delta}), \quad \textit{for any} \quad \delta > 0.$$

In short, ridgelets thresholding guarantees at least the same rate as the best neural net approximation.

We already argued that the best approximation f_n^{NN} is merely existential; we may think about it as an ideal approximation able to pick the best possible n directions $(k_1, k_2, \ldots, k_n)$. A remarkable feature of the ridgelet thresholding procedure is that it does not know anything about those directions and yet performs nearly equally as well. That is, ridgelet thresholding gives rates of approximation which rival those attainable by very abstract and complicated procedures.

Important Remark. We would like to point out that *this result is not limited to the use of the logistic function.* Actually, Theorem 5.1 still holds whether one uses a sigmoidal and monotone C^∞ activation function. It is also true if σ is C^∞ and vanishes at infinity, e.g. σ belongs to the Schwartz class. Moreover, σ may not be continuous. The theorem is still valid in the case where σ is the Heavyside, say.

We note that the formulation of Theorem 5.1 avoids discussing problems related to the geometry of the domain D, say. We mention, in passing, that other formulations of Theorem 5.1 are possible with other boundary conditions. Boundaries introduce, of course, artificial discontinuities which require a special treatment and, at the moment, we are not interested in studying those boundary effects. (Nearly everyone is aware of boundary issues. If by any chance, the reader is not familiar with these, consider the

following situation. In two dimensions, we let f be a superb C^∞ function supported on the unit disk D. Then, wavelet orthobases of $L_2(\mathbb{R}^2)$ yield poor approximations of f although it is C^∞ on the disk. The problem is that the extension of f to the plane introduces a discontinuity along ∂D.)

Theorem 5.1 is proved in [6], and we will briefly outline the argument in a later section. A preliminary version of Theorem 5.1 was first introduced in [3].

§6. Ridgelets and Ridge Functions

At the heart of Theorem 5.1 lies a key property about the ridgelet representation of ridge functions which we now present.

6.1 Ridgelets and sigmoidal ridge functions

Let σ be a C^∞ sigmoidal function, and construct the ridge function

$$f(x) = \sigma(k \cdot x - b); \tag{6.1}$$

f is not in any L_p and, therefore, the inner product between a ridge function and a ridgelet does not make much sense. We localize this ridge function near the unit disk with a multiplication by a smooth window w in C_0^∞. The key-property is that the sorted ridgelet coefficients $|\alpha(fw)|_{(n)}$ of $f\,w$ decay faster than any negative power of n. We establish that for any $p > 0$,

$$\sum_R |\alpha_R(fw)|^p \le C_p, \tag{6.2}$$

where the constant C_p may be chosen independently of the parameters k and b. It is well-known that the sparsity of the decomposition controls the quality of partial reconstructions, see [22] for example. Suppose that w is identically equal to one over the unit D and vanishes out of $2D$, and let f_n^R be the truncated ridgelet series of $f\,w$, keeping the n-largest terms in the ridgelet series. Then for any $s > 0$, we have

$$\|f - f_n^R\|_{L_2(D)} \le C_s \cdot n^{-s}, \tag{6.3}$$

where again the constant C_s is may be chosen independently of the parameters of the ridge function (6.1).

As we pointed out, the estimates (6.2)–(6.3) hold uniformly over k and b. As k varies, the ridge functions (6.1) exhibit very different behaviors. When $|k|$ is small, say $|k| \le 1$, f is a very gentle function; its derivatives up to any order are well behaved. As $|k|$ increases, however, f becomes less smooth and in the limit is discontinuous, e.g.

$$\sigma(a \cdot (\theta \cdot x - b)) \to H(\theta \cdot x - b), \text{ as } a \to \infty, \quad \theta \in \mathbf{S}^{d-1}$$

where H is the Heavyside $H(t) = 1_{\{t>0\}}$. It is then interesting to note that the ridgelet sequence is equally sparse in either situation. First, ridgelets provide sparse representations of smooth functions, which explains (6.2) whenever the parameter $|k|$ is not too large. Then as $|k|$ increases, a singularity develops. In the ridgelet domain, however, very few coefficients actually feel this singularity just as in one dimension, only a few wavelets would feel a point singularity. This localization phenomenon is responsible for the sparsity of the sequence, see Section 8 and [7] for further reference.

6.2 Ridgelets and 'Ridge-Wavelets'

Theorem 5.1 is not limited to smooth sigmoidal activation functions. Likewise, the properties (6.1)–(6.3) hold for other profiles. For instance, let ψ^M be the father Meyer wavelet [48], and consider the Meyer ridge function

$$f(x) = a^{1/2}\psi^M(a\theta \cdot x - b), \quad a > 0, \theta \in \mathbf{S}^{d-1}, \quad b \in \mathbb{R}. \tag{6.4}$$

Recall that the Meyer ψ^M belongs to the Schwartz class and is bandlimited, i.e. its Fourier transform is compactly supported. Then, the ridgelet coefficient sequence of the ridge Meyer wavelet (6.4) also obeys the estimate (6.2).

The ℓ_p-summability of ridge-wavelets (6.4) is especially interesting because it says that ridgelets are nearly orthogonal. Indeed, (6.2) applied to a smooth and oscillatory profile guarantees that the Gramm matrix

$$T(R, R') := \langle \psi_R, \psi_{R'} \rangle_{L_2(w)} := \int \psi_R(x)\psi_{R'}(x)\, w(x)\, dx \tag{6.5}$$

is nearly diagonal; Specifically, suppose ψ belongs to the Schwartz class and has vanishing moments up to any order. Then

$$\sup_R \sum_{R'} |T(R, R')|^p \le A_p. \tag{6.6}$$

This key property is proved in the Appendix. In other words, (6.6) shows that the rows are very sparse as they are uniformly bounded in ℓ_p for any $p > 0$.

The results collected so far bring up an interesting question which Theorem 5.1 partially addresses. The grail would be to show that if a function can be well-approximated by ridge-wavelets, or ridge-free-knots-splines, it can be well approximated by ridgelet thresholding. We now discuss some results in this direction.

Inspired by Donoho [27], we consider the ridge function

$$f_\theta = f(\theta \cdot x), \quad \theta \in \mathbf{S}^{d-1}, \tag{6.7}$$

and assume that the profile f is supported over the interval $[-1,1]$, say. We might develop an ideal n-term approximation of f as follows: we use an orthobasis of wavelets (ψ_I) for approximating the profile and let f_n be an n-term wavelet approximation of the ridge profile; we construct $f_{n,\theta}$ by the rule

$$f_{n,\theta} = f_n(\theta \cdot x), \tag{6.8}$$

Assume now that nonlinear wavelet partial reconstructions of the profile obey

$$\|f - f_n\|_{L_2} \le C \cdot n^{-s}, \tag{6.9}$$

for a constant C not depending on n. The point is that (6.9) is about one-dimensional functions and automatically gives the corresponding multivariate degree of approximation

$$\|f - f_{n,\theta}\|_{L_2(D)} \le C \cdot n^{-s}. \tag{6.10}$$

Again, let w be a fixed window in $\mathcal{S}(\mathbb{R}^d)$ with the property that w is identically equal to one over the unit ball D and vanishes out of $2D$. Consider now the ridgelet approximation f_n^R obtained by thresholding the ridgelet expansion of $f\,w$, keeping the n-largest terms. Then f_n^R obeys

$$\|f - f_n^R\|_{L_2(D)} \le C \cdot n^{-s}. \tag{6.11}$$

This is a remarkable property. Although thresholding does not know about the special direction θ, thresholding gives the rate (6.10) attainable by ideal procedures which know about θ.

6.3 Orthonormal ridgelets and ridge functions

Donoho [27] explored a similar situation in a slightly different setting. In two dimensions, consider the orthonormal ridgelet expansion

$$f_\theta = \sum_\lambda \alpha_\lambda \rho_\lambda$$

of the ridge function

$$f_\theta(x) = f(x_1 \cos\theta + x_2 \sin\theta), \quad \theta \in \mathbf{S}^1.$$

Let $n(\delta)$ be

$$n(\delta) = \sum_\Lambda 1_{\{|\alpha_\lambda| \|\rho_\lambda\|_{L_\infty(D)} > \delta\}},$$

and set

$$f_\delta = \sum_\Lambda \alpha_\lambda \, 1_{\{|\alpha_\lambda| \|\rho_\lambda\|_{L_\infty(D)} > \delta\}} \, \rho_\lambda.$$

To construct f_δ, we reorder the coefficients such that $|\alpha_\lambda|\|\rho_\lambda\|_{L_\infty(D)}$ is nonincreasing and keep the $n(\delta)$ largest. The reason for this special ordering is that we will be interested in L_∞ rather than L_2-errors of approximation. We thereby define a sequence of approximants $(\overline{f}_n)$ by letting $\overline{f}_{n(\delta)}$ be the $n(\delta)$ term approximation from f_θ. Suppose that f belongs to the homogeneous Besov space $\dot{B}^s_{p,p}$, with $s = 1/p$ and $0 < p < 1$, and vanishes at $\pm\infty$. Donoho showed

$$\|f_\theta - \overline{f}_n\|_{L_\infty(D)} \le C\,\|f\|_{\dot{B}^s_{p,p}}\, n^{-(s-1)}, \quad n = 1, 2, \ldots \tag{6.12}$$

He compares this with an ideal ridge approximation knowing θ constructed as in the previous section. That is, one would use wavelets to construct an n-term approximation f_n of the profile and synthesize an approximation of f_θ with

$$f_{\theta,n} = f_n(x_1 \cos\theta + x_2 \sin\theta).$$

The error of the ideal 1-dimensional approximation obeys

$$\|f - f_n\|_{L_\infty([-1,1])} \le C\,\|f\|_{\dot{B}^s_{p,p}}\, n^{-(s-1)}, \quad s = 1/p.$$

We now have an isometry

$$\|f - f_n\|_{L_\infty([-1,1])} = \|f_\theta - f_{n,\theta}\|_{L_\infty(D)}$$

and, therefore, the ideal approximation $f_{n,\theta}$ obeys

$$\|f_\theta - f_{n,\theta}\|_{L_\infty(D)} \le C\,\|f\|_{\dot{B}^s_{p,p}}\, n^{-(s-1)}, \quad s = 1/p. \tag{6.13}$$

Again, thresholding does not know the direction θ, and yet, does just as well as a method with full knoweledge of the structure of f_θ. The assumption about the profile f, namely $f \in \dot{B}^s_{p,p}$, says that, in some sense, the univariate rate $n^{-(s-1)}$ is optimal. Therefore, one cannot fundamentally improve the bound (6.13).

These results are of the same flavor as those presented in the last section. The difference here is that one uses an L_∞-norm to measure the degree of approximation as opposed to the L_2-norm. This is hardly a major distinction, however, as results similar to those presented in this section are likely to hold in L_2. The important point here is that the guiding principle is the same; that is, ridgelet series or orthoridgelet series of ridge-wavelets are very sparse.

6.4 Why does Theorem 5.1 hold?

In some sense, the results developed in this section are very special cases of Theorem 5.1. The theorem considers the case of ideal approximations which can carefully select the best n directions to approximate f. (For instance, if f were truly a superposition of finitely many ridge functions, the approximation might select those directions.) In Theorem 5.1, those directions are arbitrary whereas Sections 6.1 and 6.2 involved only a single direction.

Theorem 5.1 is a statement about the sparsity of α, the ridgelet coefficient sequence of f. Define the ℓ_p norm of an arbitrary sequence (a_n) by

$$\|a\|_{\ell_p}^p = \sum_n |a_n|^p. \tag{6.14}$$

Recall that if $|a|_{(m)}$ denotes the m largest entry in the sequence $(|a_n|)$, we have

$$\sum_{n>m} |a|_{(n)}^2 \le K_p \cdot m^{-2r} \cdot \|a\|_{\ell_{p^*}}^2, \tag{6.15}$$

where r and p^* are related to each other via $1/p^* = r + 1/2$, e.g. compare with Lemma 1 in [22].

To prove the theorem it is then sufficient to establish that for any p such that $1/p < r + 1/2$, (r is the exponent appearing in the statement of Theorem 5.1), the ridgelet coefficient sequence α obeys

$$\|\alpha\|_{\ell_p} \le C_p. \tag{6.16}$$

To see why this implies Theorem 5.1, observe that the frame property gives the following inequality which is classical and proved in [7], say: letting (a_λ) be a sequence in ℓ_2 we have

$$\|f\|_2^2 \le A^{-1}\|a\|_{\ell_2}^2, \quad f = \sum_\lambda a_\lambda \tilde{\psi}_\lambda,$$

where A is the constant appearing on the left-hand side of (2.12). We apply this inequality to the rest $(f - f_n^R)$ and obtain

$$\|f - f_n^R\|_{L_2}^2 \le A^{-1} \sum_{m>n} |\alpha(f)|_{(m)}^2,$$

which gives $\|f - f_n^R\|_{L_2}^2 = O(n^{-r'})$ for any $r' > r$ thanks to (6.16) and (6.15).

The strategy for proving our theorem will be to establish a key property about the sparsity of ridgelet coefficients of sigmoidal functions. For each $p > 0$, [6] proves (6.2) and, here, we simply sketch the idea behind

this estimate. First, we show that the one-dimensional expansion of a neuron in a nice wavelet basis is very sparse. Specifically, we prove that

$$\sigma(at-b) = \sum_k b_k \varphi_k(t) + \sum_{j\geq 0}\sum_k a_{jk}\psi_{j,k}(t), \tag{6.17}$$

with for each $p > 0$,

$$\|b\|_{\ell_\infty} \leq C_\infty, \quad \|a\|_{\ell_p} \leq C_p.$$

Then, for $\theta \in \mathbf{S}^{d-1}$, we then decompose a neuron as a superposition of ridge-wavelets by the rule

$$\sigma(a\,\theta\cdot x - b) = \sum_k b_k \varphi_k(\theta\cdot x) + \sum_{j\geq 0}\sum_k a_{jk}\psi_{j,k}(\theta\cdot x).$$

The point is that ridgelet coefficients of 'ridge-wavelets' are very sparse. This can be quantified, and roughly speaking, this is the content of (6.6); indeed, the infinite matrices mapping the a_{jk}'s and b_k's into ridgelet coefficients α_R are nearly diagonal and preserve sparsity. A careful argument then gives (6.2).

To prove the theorem, we then start by writing f as the telescoping sum

$$f = \sum_{n\geq 0} g_n, \quad g_n = f_{2^n} - f_{2^{n-1}}. \tag{6.18}$$

where f_{2^n} is the best approximation using 2^n neurons as in (5.1). Each term g_n is a finite linear combination of at most $2^n + 2^{n-1}$ neurons whose coefficients obey (5.2) and by assumption, the L_2-norm of g_n converges to zero at the rate 2^{-nr}. An argument which uses these two special facts then shows

$$\|\alpha(g_n)\|_{\ell_p} \leq B^p\,\epsilon_n^p, \tag{6.19}$$

with

$$\sum_n \epsilon_n^p \leq 1, \quad 1/p < r + 1/2.$$

For $p \leq 1$, say, the p-triangle inequality gives

$$\|\alpha(f)\|_{\ell_p}^p \leq \sum_n \|\alpha(g_n)\|_{\ell_p}^p \leq B^p\,\epsilon_n^p \leq B^p, \tag{6.20}$$

which is what we seeked to establish.

§7. Ridgelets Analysis

7.1 CHA and Approximation Theory

Harmonic analysis is concerned with developing new ways for representing functions which may have great potential in approximation theory. Typical approximation theoretic questions are about how well one can approximate an object by finite linear combinations of given templates. Harmonic analysis, however, brings emphasis on a whole different set of issues. To name just a few, some of the key concepts are

- analysis,
- synthesis,
- stability, and
- discretization.

To be complete, one should add that much of the literature in CHA is also concerned with the development of rapid algorithms for computing these new representations. Those issues are perhaps not central in approximation theory. Nevertheless, this shift in emphasis helps in reformulating important approximation theoretic questions, as we have seen. It also suggests new ways of approaching these problems.

In addition, CHA provides some new tools which may very helpful in gaining a renewed mathematical understanding of approximation theoretic problems. As an example, the ridgelet transform proves to be very powerful for studying the capabilities and the limitations of ridge function approximations. In addition, this transform is also very helpful for identifying functional classes that are well approximated by ridge functions.

7.2 Examples

To illustrate our purpose, consider the following problem. In d-dimensions, let f be the indicator function of the unit ball

$$f(x) = 1_{\{|x| \leq 1\}}, \tag{7.1}$$

and let us ask about the rate of convergence of neural networks

$$d_n(f, D^{NN}) \text{ as } n \to \infty.$$

This may seem like a trivial question. After all, it is hard to think of a simpler object than f. In truth, quantifying –giving a sharp upper bound, say– the rate of decay of

$$d_n(f, \mathcal{D}_{NN}), \quad n \to \infty$$

is certainly not an easy task.

In some sense, ridgelet analysis solves this type of problem rather effortlessly. We simply need to calculate and quantify the size of the ridgelet coefficients of f. The degree of approximation is immediately read off the decay of the coefficient sequence. We will carry out this program on the example (7.1).

First, recall the definition of the Radon transform Rf of an integrable function f (see [18] for details)

$$Rf(u,t) = \int_{u\cdot x=t} f(x)\,dx, \quad u \in \mathbf{S}^{d-1},\ t \in \mathbb{R}. \tag{7.2}$$

One way to calculate ridgelet coefficients is to observe that the ridgelet transform is precisely the application of a 1-dimensional wavelet transform to the slices of the Radon transform where the angular variable u is held constant and t is varying [4]. Mathematically speaking, the ridgelet coefficient (2.4) can be expressed as

$$\mathcal{R}_f(a,u,b) = \int Rf(u,t)\, a^{-1/2}\psi\left(\frac{t-b}{a}\right) dt. \tag{7.3}$$

Loosely speaking, ridgelet analysis is some kind of wavelet analysis in the Radon domain.

A simple calculation then shows for f, the indicator of the unit ball, that

$$R_u f_\alpha(t) = c_d(1-t^2)^{(d-1)/2}. \tag{7.4}$$

We now study the sparsity of the vector of ridgelet coefficients. Without loss of generality, we will take ridgelets with a profile ψ compactly supported in the spatial domain. We will assume further that ψ is R times differentiable and has vanishing moments through order D. Finally, let p^* be defined by $p^* = 2 - 2/d$. We show that the coefficient sequence (α_R) of f is in ℓ_p for any $p > p^*$ provided that $\min(R, D)$ is sufficiently large.

Define g by

$$g(t) = (1-t^2)^{(d-1)/2},$$

and as usual let $\psi_{j,k}(t)$ denote the one dimensional wavelet $2^{j/2}\psi(2^j t - k)$. We have

$$|\langle g, \psi_{j,k}\rangle| \le C\, 2^{-jd/2}\, \left(1 + (|k| - 2^j)\right)^{-2}. \tag{7.5}$$

The proof is a simple integration by parts and we omit it. Since $\alpha_R = c_d\langle g, \psi_{j,k}\rangle$, we have

$$|\alpha_R| \le C\, 2^{-jd/2}\, \left(1 + (|k| - 2^j)\right)^{-2},$$

and, therefore, for $p > p^*$,

$$\sum_k |\alpha_R|^p \le C\, 2^{-jdp/2} \sum_k \left(1 + (|k| - 2^j)\right)^{-2p} \le C\, 2^{-jdp/2}.$$

We sum this inequality over the angular variable (the number of orientations is of the order $2^{j(d-1)}$), and obtain

$$\sum_{\ell} \sum_{k} |\alpha_R|^p \le C\, 2^{-jdp/2}\, 2^{j(d-1)}.$$

In short, $\sum_R |\alpha_R|^p$ is finite provided $dp/2 > d-1$, or equivalently $p > p^*$.

In fact, careful bookkeeping also gives

$$\|\alpha\|_{w\ell_{p^*}} \le A_p, \tag{7.6}$$

where the weak-ℓ_p or Marcinkiewicz quasi-norm is defined as follows: let $|\theta|_{(n)}$ be the nth largest entry in the sequence $(|\theta_n|)$; we set

$$|\theta|_{w\ell_p} = \sup_{n>0} n^{1/p} |\theta|_{(n)}. \tag{7.7}$$

It is well known that inequality (6.15) holds with the weak-ℓ_p in place of the ℓ_{p^*} norm, and therefore, thresholding the ridgelet series gives

$$\|f - f_n^R\|_{L_2} \le C\, n^{-\frac{1}{2(d-1)}}. \tag{7.8}$$

The ridgelet sequence does not belong to any weak-ℓ_p space for $p < p^*$ and thus, the rate $n^{-\frac{1}{2(d-1)}}$ is the best one can hope for. In light of Theorem 5.1, this also gives a lower bound on the degree of approximation with neural networks, provided that σ is a smooth sigmoidal function.

We may generalize this example and consider different types of singularities. For instance, let

$$f_\alpha = (1 - |x|)_+^\alpha, \quad \alpha > -1/2;$$

$\alpha = 0$ is our previous example, $\alpha = 1$ says the first derivative is discontinuous at $|x| = 1$, etc. The condition $\alpha > -1/2$ ensures that f_α is square-integrable. Then, the ridgelet coefficient sequence of f_α is in $\mathrm{w}\ell_{p^*(\alpha)}$ for $1/p^*(\alpha) - 1/2 = (\alpha + 1/2)/(d-1)$ and thresholding the ridgelet series gives

$$\|f - f_n^R\|_{L_2} \le C\, n^{-(\alpha+1/2)/(d-1)}. \tag{7.9}$$

We take another example. In two dimensions, let f now be the indicator function of the unit square

$$f(x) = 1_{[0,1]^2}(x).$$

How well can we approximate f by superpositions of ridge functions? Again, this is far from trivial. Simple calculations show that the ridgelet sequence obeys $\|\alpha\|_{w\ell_{2/3}} \le C$ and, therefore,

$$\|f - f_n^R\|_{L_2} \le C\, n^{-1}. \tag{7.10}$$

We might multiply examples of this kind at will but hope that the power of the ridgelet transform is now quite clear. In another direction, the ridgelet transforms helps identifying those functions which are well approximated by ridge functions, as we are about to see.

§8. Ridgelets and Linear Singularities

Traditional methods based on wavelets, Fourier series, local cosine transforms, or splines fail at efficiently representing objects which are discontinuous along lines in dimension 2, planes in dimension 3, and so on. To detail this claim, let us examine a very simple example. On $[0,1]^d$, suppose that we want to represent the simple object

$$f(x) = 1_{\{u\cdot x > t_0\}}\, g(x) \quad g \in \overline{W}_2^s([0,1]^d), \tag{8.1}$$

where $\overline{W}_2^s([0,1]^d)$ is the closure of $C_0^\infty([0,1]^d)$ with respect to the W_2^s-Sobolev norm. The object f is singular on the hyperplane $u \cdot x = t_0$ (u is a unit vector) but may be very smooth elsewhere. Suppose for instance that one wishes to represent this object in a wavelet basis. Then, the vector of wavelet coefficients is not sparse. In fact, the number of wavelet coefficients exceeding $1/n$ is greater than $c\, n^{2(1-1/d)}$. This immediately translates into lower bounds for nonlinear approximations. Letting f_n^W be the best n-term partial reconstruction of f, the L_2-squared error of such an approximation obeys

$$\|f - f_n^W\|_{L_2}^2 \geq c\, n^{-\frac{1}{2(d-1)}}. \tag{8.2}$$

This lower bound holds even when g is as nice as we want, i.e., $g \in C^\infty$. Of course, one could develop an ideal approximation which would track the singularity $\{u \cdot x = t_0\}$ and partition the space into two halves along this hyperplane. One would then use specially adapted polynomials, splines, or wavelets to those half-spaces, and obtain a degree of approximation as if there were no singularity, i.e. of the order of $n^{-s/d}$. Hence, wavelet thresholding performs very badly vis a vis these ideal strategies.

Whereas the presence of the singularity had a dramatic effect on the sparsity of wavelet coefficients, *it does not ruin the sparsity of the ridgelet series.* Indeed, let us consider let α ($\alpha_R = \langle f, \psi_R\rangle$) denote the ridgelet coefficient sequence of f. Then, [7] shows the sequence α is sparse as if f were not singular in the sense that

$$\#\{i,\, s.t.\, |\alpha_i| \geq 1/n\} \leq C\, n^p\, \|g\|_{H^s}, \quad with\ \ 1/p = s/d + 1/2, \tag{8.3}$$

where the constant C does not depend on f. There is a direct consequence of this result. Consider the n-term obtained by naive thresholding. Then,

$$\|f - f_n^R\| \leq C\, n^{-s/d}\, \|g\|_{H^s}, \tag{8.4}$$

where, again, the constant C is independent of f. Just as wavelets are optimal for representing objects with point-like discontinuities, ridgelets provide optimally sparse representations of objects with discontinuities

along hyperplanes. To my knowledge, there is not any other system with similar features.

The bound (8.4) also highlights remarkable spatial adaptivity. Ridgelet thresholding does not know whether or not there is a singularity and if there is one, where it is. In addition, ridgelet thresholding does not need to know the degree of smoothness s of smooth part of the object. And yet, ridgelet thresholding does as just as well as an ideal approximation which would know about the location of a possible singularity, and about the smoothness away from the singularity.

We would like to point out that in two dimensions, both results (8.3) and (8.4) continue to hold with orthonormal ridgelets in place of 'pure' ridgelets (2.11), see [7].

§9. Ridge Spaces

This section introduces a new scale of functional spaces which we believe are not studied in classical analysis. The aim here is to show that this new scale is closely related to ridge function approximation.

9.1 Definition

We start with a few classical definitions/conditions, and will assume that they hold throughout the remainder of this section:

(i) $\psi \in \mathcal{S}(\mathbb{R})$.

(ii) supp $\hat{\psi} \subset \{1/2 \le |\xi| \le 2\}$.

(iii) $|\hat{\psi}(\xi)| \ge c$ if $3/5 \le |\xi| \le 5/3$. These conditions are standard in Littlewood-Paley theory, see [30].

As before, we let ψ_j denote the univariate function $2^{j/2}\psi(2^j \cdot)$ and set $R_j(u,b)$ as

$$R_j(u,b) = \int f(x)\, \psi_j(u \cdot x - b)\, dx \quad u \in \mathbf{S}^{d-1}, b \in \mathbb{R}.$$

For $s \in \mathbb{R}$, $p, q > 0$, define

$$\|f\|_{\dot{R}^s_{p,q}} = \left(\sum_{j \in \mathbb{Z}} \left(2^{js} 2^{jd/2} \|R_j\|_{L_p} \right)^q \right)^{1/q}, \tag{9.1}$$

which is well defined for integrable functions, say. Here $\|R_j\|_{L_p}$ is of course the L_p-norm defined by $\|R_j\|_{L_p} = \left(\int_{\mathbf{S}^{d-1} \times \mathbb{R}} |R_j(u,b)|^p \, dudb \right)^{1/p}$. There is an obvious modification in the case where $q = \infty$. It is not difficult to see that these expressions are norms for $1 \le p \le \infty$, $1 \le q \le \infty$ and quasi

norms in general. Note that it follows from the reproducing formula (2.8) that for $f \in L_1 \cap L_2$, $\|f\|_{\dot{R}^s_{p,q}} = 0$ if only if $f = 0$. Finally, one can show that replacing ψ with $\psi^\#$ obeying the properties of ψ listed introduced above yields equivalent norms or quasi-norms.

Sobolev spaces are defined by means of the Fourier transform. Similarly, Besov spaces may be defined by means of the wavelet transform. Likewise, what we have done here is merely to define a norm based on the ridgelet transform. The definition of $\|\cdot\|_{\dot{R}^s_{p,q}}$ bears much resemblance with Besov norms. Actually if $d = 1$, $\|\cdot\|_{\dot{R}^s_{p,q}}$ is the homogeneous Besov norm with the same indices. In higher dimensions, the quantities $\|\cdot\|_{\dot{R}^s_{p,q}}$ measure a very different behavior than that measured by Besov norms. We will not explore these differences here and simply point to [3] for further reference.

At first, the definition may seem rather internal. It is possible, however, to give an external characterization of these spaces, at least in the case where $p = q$. Let $Rf(u,t)$ be the Radon transform (7.2) of f. Then, Section 7 showed that

$$R_j(u,b) = (Rf(u,\cdot) * \tilde{\psi}_j)(b),$$

where $\tilde{\psi}_j(t) = \psi_j(-t)$ allowing an automatic substitution in (9.1). Now for $p = q$, this gives

$$\begin{aligned} \|f\|^p_{\dot{R}^s_{p,p}} &= \int \left(\sum_j 2^{jsp} 2^{jdp/2} \|Rf(u,\cdot) * \tilde{\psi}_j\|^p_{L_p(\mathbb{R})} \right) du \\ &\sim \int \|Rf(u,\cdot)\|^p_{\dot{B}^{s+(d-1)/2}_{p,p}} \, du, \end{aligned} \tag{9.2}$$

where $\dot{B}^{s+(d-1)/2}_{p,p}$ stands for the usual one-dimensional homogeneous Besov norm. This interpretation makes clear that s is a smoothness parameter and that both parameters p, q serve to measure smoothness. Here, smoothness has to be understood in a non-classical way; we are not talking about the local behavior of a function but rather about its behavior near lines (or if one is in dimension $d > 2$, near hyperplanes). We observe that this characterization highlights an interesting aspect: the condition does not require any particular smoothness of the Radon transform along the directional variable u. As an aside comment, (9.2) also supports the claim that one obtains equivalent quantities for two ψ and $\psi^\#$ verifying the properties listed at the beginning of the section.

It is interesting to notice that for $p = q = 2$, $\|\cdot\|_{R^s_{2,2}}$ is an equivalent norm for the homogeneous Sobolev norm $\dot{W}^s_2$,

$$\|f\|^2_{\dot{W}^s_2} = \int_{\mathbb{R}^d} |\hat{f}(\xi)|^2 \, |\xi|^{2s} \, d\xi.$$

This follows from (9.2) and the fact that

$$\int \|Rf(u,\cdot)\|^2_{\dot{B}^{s+(d-1)/2}_{2,2}} \sim \|f\|^2_{\dot{W}^s_2},$$

see [51].

Definition 9.1. *Let D be the unit ball of $\mathbb{R}^d$, say. The space $\overline{R}^s_{p,q}(D)$ will denote the closure of $C^\infty_0(D)$ of f with respect to the homogeneous norm $\|\cdot\|_{\dot{R}^s_{p,q}}$.*

Other definitions, with other 'boundary conditions' are of course possible, see [3].

It is beyond the scope of this paper to explore the properties of the new spaces $\overline{R}^s_{p,q}$, such as investigating for instance embedding relationships or interpolation properties, etc. Instead, we will show that those new functional classes are optimally approximated by ridge functions and especially, by ridgelet thresholding.

We mention, however, that these functional classes may model objects with a special kind of inhomogeneity. For instance, consider a linearly mutilated object as in Section 8 $f_i(x) = 1_{\{u_i\cdot x - b_i \geq 0\}}\, g_i(x)$ with $g_i \in \overline{W}^s_2$. and linear combinations of these. Take the class $\mathcal{F}$ of those f which may be decomposed as a convex combination of our templates

$$f = \sum_i a_i f_i, \quad \sum |a_i| \leq 1, \text{ with } \|g_i\|_{W^2_s} \leq C.$$

It turns out that this intuitive class of objects nearly corresponds to a ball in one of our functional classes. In fact, [3] proves that there exist two constants such that

$$R^{(d+1)/2}_{1,1}(C_1) \subset \mathcal{F} \subset R^{(d+1)/2}_{1,\infty}(C_2). \tag{9.3}$$

Now, the bracketing classes are nearly the same which means that membership in $\mathcal{F}$ would be roughly equivalent to membership in a ball in $R^{(d+1)/2}_{1,q}$ $(1 \leq q \leq \infty)$. Therefore, we should really think about these spaces as describing the kind of spatial inhomogeneities we introduced in Section 8.

9.2 Approximation

Let $\overline{R}^s_{p,q}(C)$ be the ball

$$\overline{R}^s_{p,q}(C) = \{f \in \overline{R}^s_{p,q},\ \|f\|_{\dot{R}^s_{p,q}} \leq C\}. \tag{9.4}$$

We are going to characterize the degree of approximation of these functional classes in the L_2 metric. We give both an upper and a lower bound of approximation which are of the same order. We will consider a range of indices where $s > d(1/p - 1/2)$, as this guarantees that elements of $\overline{R}^s_{p,q}$ are square integrable, see [3].

Theorem 9.2. *Consider the class $\overline{R}^s_{p,q}(C)$, and assume $s > d(1/p - 1/2)_+$. Then*

- *For any "reasonable" dictionary $\mathcal{D}$*

$$d_n(\overline{R}^s_{p,q}(C), \mathcal{D}) \geq K_1\, n^{-s/d}, \tag{9.5}$$

  where the constant K_1 depends at most on s, p, q.
- *Ridgelet thresholding achieves the optimal rate i.e.*

$$\sup_{f \in \overline{R}^s_{p,q}(C)} \|f - f_n^R\|_{L_2(D)} \leq K_2\, n^{-s/d}, \tag{9.6}$$

  where again K_2 might depend on s, p, q.

Essentially, what the theorem says is that no other "reasonable" dictionary exists with better approximation estimates for the classes $\overline{R}^s_{p,q}(C)$ than what can be obtained via ridgelet thresholding. We must clarify, however, the meaning of the word "reasonable": when considering lower approximation bounds by finite linear combinations from dictionaries, we remark that we must only consider certain kinds of dictionaries. We quote from [28]. "If one allows infinite dictionaries (even discrete countable ones), we would then be considering dictionaries $\mathcal{D} = \{g_\lambda,\ \lambda \in \Lambda\}$ enumerating a dense subset of all common functional classes (including $\overline{R}^s_{p,q}(C)$.), and which can perfectly reproduce any f with a singleton: $d_1(f, \mathcal{D}) = 0$". Thus when we say "reasonable dictionary," we have in mind that one considers only sequences of dictionaries whose size grow polynomially in the number of terms to be kept in the approximation (1.3).

There are other ways of looking at the lower bound. Let $\mathcal{F}$ be a compact set of functions in $L^2(D)$. We recall that the Kolmogorov ϵ-entropy $N(\epsilon, \mathcal{F})$ of the class $\mathcal{F}$ is the minimum number of bits that is required to specify any element f from $\mathcal{F}$ within an accuracy of ϵ. It may be defined as follows. Let $N(\epsilon, \mathcal{F})$ be the minimum number of L_2-balls of radius ϵ one needs to cover $\mathcal{F}$. Formally,

$$N(\epsilon, \mathcal{F}) = \min\{n :\ \exists (f_i)_{i=1}^n \text{ such that } \forall f \in \mathcal{F},\ \inf_i \|f - f_i\| \leq \epsilon\};$$

the Kolmogorov entropy is

$$L(\epsilon, \mathcal{F}) = \lceil \log_2 N(\epsilon, \mathcal{F}) \rceil. \tag{9.7}$$

Then inequality (9.5) says that the Kolmogorov entropy of $\overline{R}^s_{p,q}(C)$ is bounded below by

$$L(\epsilon, \overline{R}^s_{p,q}(C)) \geq c \cdot \epsilon^{-1/(s/d)}.$$

It also gives correponding lower bounds on rates of estimation, see [5]. Finally, following [19], it is most likely that there is no method of approximation which depends continuously on f with better properties than (9.5).

The upper bound says that we have an asymptotically near-optimal procedure for binary encoding elements of $\overline{R}^s_{p,q}(C)$: let $L(\epsilon, \overline{R}^s_{p,q}(C))$ be the minimum number of bits necessary to store in a lossy encoding/decoding system in order to be sure that the decoded reconstruction of every $f \in \overline{R}^s_{p,q}(C)$ will be accurate to within ϵ (in an L_2 sense). Then, a coder-decoder based on simple uniform quantization (depending on ϵ) of the coefficients α_i followed by simple run length coding achieves both a distortion smaller than ϵ and a codelength that is optimal up to multiplicative factors like $\log(\epsilon^{-1})$.

9.3 About the lower bounds

The proof of the lower bound uses an argument which is rather classical in statistics and perhaps in rate distortion theory. The goal is to construct a "fat" hypercube which is embedded in the functional class you wish to approximate. More specifically, a result from [28] shows how the hypercube embedding limits the approximation error.

Theorem 9.3. *Suppose that the class $\mathcal{F}$ contains embedded hypercubes of dimension $n(\delta)$ and side δ, and that*

$$n(\delta) \geq K\,\delta^{-2/(2r+1)}, \qquad 0 < \delta < \delta_0.$$

Let $\mathcal{D}_k$ be a family of finite dictionaries indexed by $k = k_0, k_0 + 1, \ldots$ obeying the size estimate $\#\mathcal{D}_k \leq Bk^\beta$. Let $\pi(t)$ be a polynomial. Then

$$d_n(\mathcal{F}, \mathcal{D}_{\pi(n)}) \geq K'\,n^{-r}.$$

In our situation, the construction of embedded hypercubes involves properties of the ridgelet frame and is proved in [5]. Letting $\nu = (s, p, q)$. We construct a sequence of cubes $\mathcal{H}^\nu_j$ indexed by $j \geq j_0$ and dimension of the order of 2^{jd} such that

$$\mathcal{H}^\nu_j \subset \overline{R}^s_{p,q}(C), \quad \nu = (s, p, q),$$

and of sidelength $\delta_{j\nu}$ obeying the hypothesis of Theorem 9.3. If ridgelets were orthogonal, the vertices of the cube $\mathcal{H}^\nu_j$ would be properly renormalized ridgelets at scale j. The actual construction $\mathcal{H}^\nu_j$ involves the 'orthogonalization' of the ridgelet frame (so that the vertices are 'perturbed ridgelets') and is delicate. The argument is not reproduced here.

9.4 Proof of the upper bound

Let (ψ_R) be a nice ridgelet frame such that ψ is at least R times differentiable and has vanishing moments up to order D. We define a discrete norm on the coefficient sequence $\alpha_R = \langle f, \psi_R \rangle$ as follows

$$\|\alpha\|_{\dot{\mathbf{r}}^s_{p,q}} \equiv \left(\sum_j \left(2^{j\sigma} \Big(\sum_{|R|=j} |\alpha_R|^p \Big)^{1/p} \right)^q \right)^{1/q}, \quad \sigma = s + d(1/2 - 1/p). \tag{9.8}$$

Here, the notation $\sum_{|R|=j}$ means that the summation extends to all those coefficients at a fixed scale j. The norm (9.8) is, of course, the discrete analog of (9.1). Recall that we may interpret the frame coefficients as being the samples from the continuum of ridgelet coefficients, namely

$$\alpha_R = \mathcal{R}_f(2^{-j}, u_{j\ell}, k2^{-j}).$$

To have the discrete analogy, one would like to have a sampling theorem which says that the discrete Riemann sum is equivalent to the corresponding integral, i.e.

$$\|\alpha\|_{\dot{\mathbf{r}}^s_{p,q}} \sim \|f\|_{\dot{R}^s_{p,q}}$$

valid for elements of $\overline{R}^s_{p,q}$. Notice that we have already established this equivalence in the case where $s = 0$ and $p = q = 2$ since the frame property gives

$$\|\alpha\|_{\ell_2} \sim \|f\|^2_{L_2} \sim \|f\|^2_{\dot{R}^0_{2,2}}.$$

In fact, a slight modification of the argument underlying the proof of the frame property (2.12) [4] would give the same equivalence for arbitrary s, $s > 0$.

Lemma 9.4. *Let $f \in \overline{R}^s_{p,q}$. There is $r^*(s)$ so that, if we use a ridgelet frame with $min(R, D) \geq r^*(s)$, then there is a constant C possibly depending on s, p, q such that*

$$\|\alpha\|_{\dot{\mathbf{r}}^s_{p,q}} \leq C \|f\|_{\dot{R}^s_{p,q}}. \tag{9.9}$$

Proof: It is simpler to prove the lemma for a ridgelet frame $(\psi_R)_{R\in\mathcal{R}}$ with a compactly supported profile ψ. In addition, to be compactly supported, we will assume that ψ has as many derivatives and vanishing moments as we want.

Further, we will take $\varphi^\#$ and $\psi^\#$ compactly supported and obeying (2.9) with $2K'_\psi = (2\pi)^{-(d-1)}$ so that the following identity holds:

$$f(x) = \int R^0(u,b)\varphi^\#(u \cdot x - b)\, du\, db + \sum_{j\geq 0} 2^{jd} \int R^1_j(u,b) 2^{j/2} \psi^\#(2^j(u \cdot x - b))\, du\, db,$$

with

$$R^0(u,b) = \int f(x)\varphi^{\#}(u\cdot x - b)\,dx,$$
$$R^1_j(u,b) = \int f(x)2^{j/2}\psi^{\#}(2^j(u\cdot x - b))\,dx.$$

We recall that the reproducing formula is valid for square integrable and compactly supported functions, say.

Inspired by the discrete ridgelet transform, we further decompose our reproducing formula. We keep the same notations as for the ridgelet frames, and recall that R indexes the triples (j,ℓ,k) and $u_{j,\ell}$ is the collection of sampled orientations at scale j. At each scale j, we introduce a Voronoi partition of the sphere, and let $S_{j,\ell}$ be those directions u on $\mathbf{S}^d$ which are closer to $u_{j,\ell}$ (Euclidean distance) than any other sampled orientations $u_{j,\ell'}$. We then define the 'cell' Q_R by

$$Q_R = S_{j,\ell}\otimes[k2^{-j},\,(k+1)2^{-j}].$$

With these notations, we set

$$f = \sum_{R\in\mathcal{R}^0} m^0_R + \sum_{R\in\mathcal{R}} m^1_R,$$

where $\mathcal{R}^0 = \{R\in\mathcal{R},\ |R| = 0\}$ and

$$m^0_R = \int_{Q_R} R^0(u,b)\varphi^{\#}(u\cdot x - b)\,du\,db$$
$$m^1_R = 2^{jd}\int_{Q_R} R_j(u,b)2^{j/2}\psi^{\#}(2^j(u\cdot x - b))\,du\,db.$$

The reader will object that we have not defined cells allowing a proper definition of m^0_R; we take the same cells as those in m^1_R for $j = 0$.

Now, let w be a C^∞ window identically equal to one over D and vanishing outside of $2D$. We recall that $f = fw$ since f is compactly supported in D. For notational convenience, we let $\mathcal{R}^1$ be a copy of $\mathcal{R}$ and write

$$\alpha_{R'} = \sum_{\varepsilon\in\{0,1\}}\sum_{R\in\mathcal{R}^\varepsilon}\langle\psi_{R'}, m^\varepsilon_R w\rangle.$$

We introduce some notations. Define

$$T_1(R',R) = \sup_{Q_R}\left|\int_{\mathbb{R}^d} 2^{j/2}\psi^{\#}(2^j(u\cdot x - b))\,\psi_{R'}(x)w(x)\,dx\right|. \tag{9.10}$$

and

$$\beta^1_R, = 2^{jd}\int_{Q_R}|R^1_j(u,b)|\,du\,db, \tag{9.11}$$

and similarly for T_0 and β_R^0. Note that per scale, there is only a finite number of β_R^1's which are possibly nonzero because we have chosen a pair $(\varphi^\#, \psi^\#)$ where both are compactly supported; at scale j, β_R^1 vanishes whenever $k > c2^{-j}$ where the constant c is a function of the support of $\psi^\#$. A similar conclusion applies to β_R^0. As a consequence, the number of nonzero β_R^0's is at most $O(1)$ and that of nonzero β_R^1's is at most 2^{jd} per scale.

Observe now that

$$|\langle \psi_{R'}, m_R^\varepsilon w\rangle| \le T_\varepsilon(R', R)\, \beta_R^\varepsilon.$$

This follows from

$$\begin{aligned}
|\langle \psi_{R'}, m_R^1 w\rangle| &= \\
&= \left| 2^{jd} \int_{Q_R} \int_{\mathbb{R}} R_j^1(u,b) 2^{j/2} \psi^\#(2^j(u\cdot x - b))\, \psi_{R'}(x) w(x)\, dx\, du\, db \right| \\
&\le 2^{jd} \int_{Q_R} |R_j^1(u,b)| \left| \int_{\mathbb{R}} 2^{j/2} \psi^\#(2^j(u\cdot x - b)) \psi_{R'}(x) w(x)\, dx \right| du\, db,
\end{aligned}$$

and similarly for $\varepsilon = 0$. In short,

$$|\alpha_R| \le \sum_\varepsilon \sum_{R' \in \mathcal{R}^\varepsilon} T_\varepsilon(R, R') \beta_{R'}^\varepsilon.$$

Lemma 9.5. *Let $1 \le p, q \le \infty$ and $s > d(1/2 - 1/p)$. The operators T_0 and T_1 obey*

$$\|T_0 \beta^0\|_{\dot{\mathbf{r}}^s_{p,q}} \le A_0 \, \|\beta^0\|_{\ell_p}, \quad \|T_1 \beta^1\|_{\dot{\mathbf{r}}^s_{p,q}} \le A_1 \, \|\beta^1\|_{\dot{\mathbf{r}}^s_{p,q}}.$$

This lemma is at the heart of the proof of (9.9) and we postpone its demonstration. It implies

$$\|\alpha\|_{\dot{\mathbf{r}}^s_{p,q}} \le A_0 \|\beta^0\|_{\ell_p} + A_1 \|\beta^1\|_{\dot{\mathbf{r}}^s_{p,q}},$$

and (9.9) follows from the claim that for $f \in \overline{R}^s_{p,q}$

$$\|\beta^0\|_{\ell_p} + \|\beta^1\|_{\dot{\mathbf{r}}^s_{p,q}} \le A \, \|f\|_{\dot{R}^s_{p,q}}. \tag{9.12}$$

We shall now argue about (9.12).

Recall the general probability inequality

$$\int_\Omega |g(x)|\, \mu(dx) \le \mu(\Omega)^{1-1/p} \left(\int_\Omega |g(x)|^p \, \mu(dx) \right)^{1/p}.$$

We apply this inequality and obtain

$$|\beta_R^1| \leq C \left(2^{jd} \int_{Q_R} |R_j^1(u,b)|^p \, du \, db \right)^{1/p}.$$

This follows from the ridgelet discretization which gives

$$\sup_R |Q_R| \leq C \, 2^{-jd}.$$

Therefore,

$$\sum_{|R|=j} |\beta_R^1|^p \leq C \, 2^{jd} \int_{\mathbf{S}^{d-1}} \int_{\mathbb{R}} |R_j^1(u,b)|^p \; du \, db.$$

Similarly

$$\sum_{R \in \mathcal{R}^0} |\beta_R^0|^p \leq C \int_{\mathbf{S}^{d-1}} \int_{\mathbb{R}} |R_j^0(u,b)|^p \; du \, db.$$

Hence, we have proved that

$$\|\beta^0\|_{\ell_p} + \|\beta^1\|_{\dot{\mathbf{r}}^s_{p,q}} \leq C \left(\|R^0\|_{L_p} + \left(\sum_{j \geq 0} (2^{js} 2^{jd/2} \|R_j\|_{L_p})^q \right)^{1/q} \right). \tag{9.13}$$

The right-hand side of the above inequality is the definition (9.1) but for the coarse scale variation where the sum over the negative indices j is replaced by the single term $\|R^0\|_{L_p}$. In fact, the right-hand side of (9.13) may be taken as the definition of the inhomogeneous $R^s_{p,q}$-norm. For $f \in \overline{R}^s_{p,q}$, we have

$$\|R^0\|_{L_p} + \left(\sum_{j \geq 0} \left(2^{js} 2^{jd/2} \|R_j\|_{L_p} \right)^q \right)^{1/q} \sim \|f\|_{\dot{R}^s_{p,q}} \tag{9.14}$$

where the $\sim$ symbol has the meaning of norm equivalence $\dot{R}^s_{p,q}$-norm. This norm equivalence is proved in [3], and holds for we defined $\overline{R}^s_{p,q}$ as the closure of $C_0^\infty(D)$ with respect to the . The proof is straightforward and is totally analogous to the following situation: for $f \in \overline{W}^s_2$ (see Section 8 for the definition of $\overline{W}^s_2$), we have

$$\|f\|_{W^s_2} = \|f\|_{L_2} + \|f\|_{\dot{W}^s_2} \sim \|f\|_{\dot{W}^s_2}.$$

(The reader familiar with Besov norms will also know that if one defines Besov spaces over the unit ball $\overline{B}^s_{p,q}$ by the closure of $C_0^\infty(D)$ with respect

to the homogeneous Besov norm, say, then over that space, the homogeneous and inhomogeneous Besov norms are equivalent.) Of course, (9.14) finishes the proof of Lemma 9.4. □

Proof of Lemma 9.5. We have

$$T_1(R', R) = \sup_{(u,b)\in Q_R} K(j, u, b; j', u_{R'}, b_{R'}),$$

with

$$K(j, u,b; j', u', b') = \\ = \left| \int_{\mathbb{R}^d} 2^{j/2}\psi^{\#}(2^j(u\cdot x - b))2^{j'/2}\psi(2^{j'}(u'\cdot x - b')\, w(x)\, dx \right|.$$

Recall that $Q_R = S_{u_R} \otimes [k2^{-j}, (k+1)2^{-j}]$ where the diameter of S_{u_R} does not exceed $C\, 2^{-j}$. Now, Lemma 14.1 below gives an upper bound on the quantity $K(j, u, b; j', u', b')$. We then apply this lemma and take the supremum of the right hand-side of (14.8) over the cell Q_R. It is easy to verify that this gives an upper estimate as in (14.13), i.e. with the notations of the Appendix,

$$T_1(R', R) \le C 2^{-(j'+j)(n+1/2)}\, \delta_j^{2n+1}(\ell, \ell')\, \left(1 + |k'2^{-j'}|\right)^{-m} \\ \left(1 + \delta_j(\ell, \ell')|k'2^{-j'}\cos\theta - k2^{-j}|\right)^{-m}.$$

We first show the lemma at 'the corners', namely for $p, q \in \{1, \infty\}$. We will then establish the general case by interpolation. We let $\alpha^\varepsilon = T_\varepsilon \beta^\varepsilon$.

Suppose first that $p = 1$. Then

$$\sum_{|R|=j} |\alpha_R^1| \le \sum_{R'} \sum_{R} T_1(R, R')|\beta_{R'}^1|.$$

The Appendix establishes (14.18)

$$\sum_{|R|=j} |T(R; R')| \le C\, 2^{-|j-j'|((n+1/2)-d)},$$

where $n \le \min(R, D)$. Therefore,

$$\sum_{|R|=j} |\alpha_R^1| \le C \sum_{R'} 2^{-|j-j'|((n+1/2)-d)}|\beta_{R'}^1| \\ = C \sum_{j'} 2^{-|j-j'|((n+1/2)-d)} \sum_{|R'|=j'} |\beta_{R'}^1|.$$

Letting

$$a_j = 2^{js}2^{-jd/2}\sum_{|R|=j}|\alpha_R^1|, \quad b_j = 2^{js}2^{-jd/2}\sum_{|R|=j}|\beta_R^1|,$$

we have

$$a_j \le C\sum_{j'}\epsilon_{j,j'}\,b_{j'}, \quad \epsilon_{j,j'} = 2^{-|j-j'|((n+1/2)-d)}\,2^{(j-j')s}.$$

For $q=\infty$ we have

$$a_j \le C\sup_{j'\ge 0} b_{j'}\sum_{j'\ge 0}\epsilon_{j,j'} \le B\,\|b\|_{\ell_\infty},$$

provided that n is large enough so that $(n+1/2)>d$. For $q\le 1$ we have (again $(n+1/2)>d$)

$$\sum_j \epsilon_{j,j'}^q \le B_q^q,$$

and, therefore, the q-triangle inequality gives

$$\sum_{j\ge 0} a_j^q \le \sum_{j\ge 0}\sum_{j'\in\mathbf{Z}}\epsilon_{j,j'}^q b_{j'}^q \le B_q^q\,\|b\|_{\ell_q}^q.$$

The case for $p=\infty$ is nearly identical. We have

$$\sum_{|R'|=j'} T(R,R') \le C\,2^{-|j-j'|((n+1/2)-d)},$$

and, therefore,

$$|\alpha_R^1| \le \sum_{j'} 2^{-|j-j'|((n+1/2)-d)}\sup_{|R'|=j'}|\beta_{R'}^1|.$$

Letting

$$a_j = 2^{js}2^{jd/2}\sup_{|R|=j}|\alpha_R^1|, \quad b_j = 2^{js}2^{jd/2}\sup_{|R|=j}|\beta_R^1|,$$

gives

$$a_j \le C\sum_{j'}\epsilon_{j,j'}b_{j'}, \quad \epsilon_{j,j'} = 2^{-|j-j'|(n-(s+d/2))}.$$

We use the exact same argument as before. The ℓ_q summability for any $q>0$ of $\epsilon_{j,j'}$ with respect to either j or j' (provided $n>s+d/2$) gives

$$\|a\|_{\ell_q} \le C\,\|b\|_{\ell_q}.$$

We have now proved the property for $p, q \in \{1, \infty\}$.

We finish the proof by interpolation. The sequence norm $\dot{\mathbf{r}}^s_{p,q}$ is a weighted norm of the type $\ell_q(\ell_p)$. We write

$$\alpha^1 = T_1 \beta^1,$$

and as we observed earlier, at scale j the coefficients α^1_R and β^1_R vanish for $k \geq c \cdot 2^j$. Therefore, at each scale, there is a set of indices of cardinality at most $O(2^{jd})$ which can contribute possibly nonzero coefficients. This shows that there is, of course, a full analogy with the Besov scale which is also a weighted space of the type $\ell_q(\ell_p)$ with 2^{jd} coefficients per scale. Therefore, the sequence spaces $\dot{\mathbf{r}}^s_{p,q}$ are clearly interpolation spaces with the same interpolation properties as for the d-dimensional Besov scale. The boundedness of T_1 in the $\dot{\mathbf{r}}^s_{p,q}$ norm for arbitrary $1 \leq p, q \leq \infty$ follows.

The story is nearly identical for T_0. The operator T_0 obeys the estimates developed in the Appendix, namely (14.22). Further, the Appendix (14.23) shows that

$$\sum_{|R'|=j'} \sum_{|R|=0} |T_0(R', R)|^p \leq C\, 2^{-j'((n+1/2)p-d)},$$

where n is as large as we want provided we have enough regularity and vanishing moments. Now, using the same arguments as before, the previous estimates on T_0 show that

$$\|T_0 \beta^0\|_{\dot{\mathbf{r}}^s_{p,q}} \leq C\, \|\beta^0\|_{\ell_p}, \quad p, q \in \{1, \infty\}.$$

Interpolation allows the extension of the above inequality to arbitrary parameters $1 \leq p, q \leq \infty$. □

Corollary 9.6. *Assume $s > d(1/p - 1/2)_+$ and let p^* be defined by $1/p^* = s/d + 1/2$. Then for f in $\overline{R}^s_{p,q}(D)$ we have*

$$\|\alpha\|_{w\ell_{p^*}} \leq C\, \|f\|_{\dot{R}^s_{p,q}(D)}, \tag{9.16}$$

where $w\ell_{p^*}$ is the weak-ℓ_{p^*} quasi-norm (7.7).

Proof: We just showed that $\|\alpha\|_{\dot{\mathbf{r}}^s_{p,q}} \leq C\, \|f\|_{\dot{R}^s_{p,q}(D)}$, and we have

$$\|\alpha\|_{w\ell_{p^*}} \leq C\, \|\alpha\|_{\dot{\mathbf{r}}^s_{p,q}}$$

for $s > d(1/p - 1/2)_+$. This follows from a similar statement about the weak-ℓ_p boundedness of wavelet coefficient sequences taken from d-dimensional Besov spaces since $r^s_{p,q}$ and $b^s_{p,q}$-Besov balls have exactly the same structure. The proof then consists of a minor adaptation of a result in [22]. We do not reproduce it here. □

Of course, (9.16) gives that ridgelet thresholding satisfies

$$\|f - f^R_n\| \leq C\, n^{-s/d}\, \|f\|_{\dot{R}^s_{p,q}(D)},$$

which proves (9.6), the second part of Theorem 9.2.

§10. Beyond Ridgelets

10.1 Limitations of ridge function approximation

Ridgelet analysis may give very precise information about the degree of approximation by finite linear combination of ridge functions. As we have seen, we may use the ridgelet transform as a fundamental tool to identify those target functions that are well-suited or ill-suited for ridge function approximation. Objects with curved singularities are in the latter class.

We recall that Section 7 developed the following sharp lower bound: letting f be the indicator function of the unit ball, we have

$$\|f - f_n^R\| \sim n^{-1/(2(d-1))}.$$

This is a general phenomenon. If f is singular along a smooth manifold of codimension 1, we cannot, in general, hope for better rates than $n^{-1/(2(d-1))}$. (By the way, it is remarkable that d-dimensional wavelets give the same degree of approximation, i.e. $n^{-1/(2(d-1)}$.) In two dimensions, this says that wavelets and ridgelets are inefficient for representing edges as the rate is only of the order $n^{-1/2}$. We are entitled to say that they are inefficient because of the existence of other approximation strategies with provably better performances, as we are about to see.

Finding nonadaptive and optimally sparse representations of objects with curved singularities has been a special concern of mine; this is our next stop. The beautiful thing is that although ridgelets fail at representing efficiently smooth singularities, they will be the cornerstone of refined transforms which will not. In short, they pave the way to more sophisticated constructions.

10.2 CHA and curved singularities

It is well-known that, for objects with discontinuities, wavelets offer an improvement on traditional representations like sinusoids, but wavelets are far from optimal.

To make this concrete, consider approximating such an f from the best n-terms in a Fourier expansion adapted to $[-\pi, \pi]^2$ (say). The squared error of such an n-term expansion $\tilde{f}_n^F$ would obey

$$\|f - \tilde{f}_n^F\|_2^2 \asymp n^{-1/2}, \quad n \to \infty. \tag{10.1}$$

For comparison, consider an approximation $\tilde{f}_n^W$ from the best n-terms in a wavelet expansion; then

$$\|f - \tilde{f}_n^W\|_2^2 \asymp n^{-1}, \quad n \to \infty, \tag{10.2}$$

which is considerably better. However, from [28,25,33] we know that there exist dictionaries of (nonorthogonal) elements, and procedures for selecting from those dictionaries that will yield m-term approximations obeying

$$\|f - \tilde{f}_n^D\|_2^2 \asymp n^{-2}, \quad n \to \infty. \tag{10.3}$$

The next section constructs tight frames of curvelets. These frames are based on the ridgelet transform and provide optimal representations of C^2 objects smooth away from C^2 curves. Indeed, thresholding the curvelet expansion gives (10.3) up to a logarithmic factor.

§11. Curvelet Construction

We now briefly discuss the curvelet frame; for more details, see [8]. The construction combines several ingredients, which we briefly review:

- *Ridgelets*, a method of analysis very suitable for objects which are discontinuous across straight lines.
- *Multiscale Ridgelets*, a pyramid of analyzing elements which consists of ridgelets renormalized and transported to a wide range of scales and locations.
- *Bandpass Filtering*, a method of separating an object into a series of disjoint scales.

Sections 2 and 3 introduced ridgelets, and we first briefly discuss multiscale ridgelets and bandpass filtering. We then describe the combination of these three components. There is a difference between this construction and the one given in [8] at large scales.

11.1 Multiscale ridgelets

Think of ortho ridgelets as objects which have a "length" of about 1 and a "width" which can be arbitrarily fine. The multiscale ridgelet system renormalizes and transports such objects, so that one has a system of elements at all lengths and all finer widths.

The construction begins with a smooth partition of energy function w with $w(x_1, x_2) \geq 0$, $w \in C_0^\infty([-1,1]^2)$ obeying $\sum_{k_1,k_2} w^2(x_1 - k_1, x_2 - k_2) \equiv 1$. Define a transport operator associated with index Q indicating a dyadic square $Q = (s, k_1, k_2)$ of the form $[k_1/2^s, (k_1+1)/2^s) \times [k_2/2^s, (k_2+1)/2^s)$, by $(T_Q f)(x_1, x_2) = f(2^s x_1 - k_1, 2^s x_2 - k_2)$. The Multiscale Ridgelet with index $\mu = (Q, \lambda)$ is then

$$\psi_\mu = 2^s \cdot T_Q(w \cdot \rho_\lambda)$$

In short, one transports the normalized, windowed orthoridgelet.

Letting $\mathcal{Q}_\int$ denote the dyadic squares of side 2^{-s}, we can define the subcollection of Monoscale Ridgelets at scale s:

$$\mathcal{M}_s = \{(Q, \lambda) : Q \in \mathcal{Q}_s, \lambda \in \Lambda\}.$$

It is immediate from the orthonormality of the ridgelets that each system of monoscale ridgelets makes a tight frame, in particular obeying the Parseval relation

$$\sum_{\mu \in \mathcal{M}_s} \langle \psi_\mu, f \rangle^2 = \|f\|_{L^2}^2.$$

It follows that the dictionary of multiscale ridgelets at all scales, indexed by

$$\mathcal{M} = \cup_{s \geq 1} \mathcal{M}_s,$$

is not frameable, as we have energy blow-up:

$$\sum_{\mu \in \mathcal{M}} \langle \psi_\mu, f \rangle^2 = \infty. \tag{11.1}$$

The Multiscale Ridgelets dictionary is simply too massive to form a good analyzing set. It lacks inter-scale orthogonality – $\psi_{(Q,\lambda)}$ is not typically orthogonal to $\psi_{(Q',\lambda')}$ if Q and Q' are squares at different scales and overlapping locations. In analyzing a function using this dictionary, the repeated interactions with all different scales causes energy blow-up (11.1).

The construction of curvelets solves this problem by in effect disallowing the full richness of the Multiscale Ridgelets dictionary. Instead of allowing all different combinations of 'lengths' and 'widths', we allow only those where $width \approx length^2$.

11.2 Subband filtering

Our remedy to the 'energy blow-up' (11.1) is to decompose f into subbands using standard filterbank ideas. Then we assign one specific monoscale dictionary $\mathcal{M}_\int$ to analyze one specific (and specially chosen) subband.

We define coronae of frequencies $|\xi| \in [2^{2s}, 2^{2s+2}]$, and subband filters D_s extracting components of f in the indicated subbands; a filter P_0 deals with frequencies $|\xi| \leq 1$. The filters decompose the energy exactly into subbands:

$$\|f\|_2^2 = \|P_0 f\|_2^2 + \sum_s \|D_s f\|_2^2.$$

The construction of such operators is standard [61]; the coronization oriented around powers 2^{2s} is nonstandard – and essential for us. Explicitly, we build a sequence of filters Φ_0 and $\Psi_{2s} = 2^{4s}\Psi(2^{2s}\cdot)$, $s = 0, 1, 2, \ldots$ with

the following properties: Φ_0 is a lowpass filter concentrated near frequencies $|\xi| \le 1$; Ψ_{2s} is bandpass, concentrated near $|\xi| \in [2^{2s}, 2^{2s+2}]$; and we have

$$|\hat{\Phi}_0(\xi)|^2 + \sum_{s \ge 0} |\hat{\Psi}(2^{-2s}\xi)|^2 = 1, \quad \forall \xi.$$

Hence, D_s is simply the convolution operator $D_s f = \Psi_{2s} * f$.

11.3 Definition of curvelet transform

Assembling the above ingredients, we are able to sketch the definition of the Curvelet transform. We let M' consist of M merged with the collection of integral triples (s, k_1, k_2, e) where $s \le 0$, $e \in \{0, 1\}$, indexing all dyadic squares in the plane of side $2^s > 1$.

The curvelet transform is a map $L^2(\mathbb{R}^2) \mapsto \ell^2(M')$, yielding Curvelet coefficients $(\alpha_\mu : \mu \in M')$. These come in two types. At coarse scales we have wavelet coefficients:

$$\alpha_\mu = \langle W_{s,k_1,k_2,e}, P_0 f \rangle, \qquad \mu = (s, k_1, k_2) \in M' \backslash M,$$

where each $W_{s,k_1,k_2,e}$ is a Meyer wavelet, while at fine scale we have Multiscale Ridgelet coefficients of the bandpass filtered object:

$$\alpha_\mu = \langle D_s f, \psi_\mu \rangle, \qquad \mu \in M_s, s = 1, 2, \ldots.$$

Note well that for $s > 0$, each coefficient associated to scale 2^{-s} derives from the subband filtered version of f – $D_s f$ – and not from f. Several properties are immediate;

- Tight Frame:
$$\|f\|_2^2 = \sum_{\mu \in M'} |\alpha_\mu|^2.$$
- Existence of Coefficient Representers (Frame Elements): There are $\gamma_\mu \in L^2(\mathbb{R}^2)$ so that
$$\alpha_\mu \equiv \langle f, \gamma_\mu \rangle.$$
- L^2 Reconstruction Formula:
$$f = \sum_{\mu \in M'} \langle f, \gamma_\mu \rangle \gamma_\mu.$$
- Formula for Frame Elements: for $s \le 0$, $\gamma_\mu = P_0 \phi_{s,k_1,k_2}$, while for $s > 0$,
$$\gamma_\mu = D_s \psi_\mu, \qquad \mu \in \mathcal{Q}_s. \tag{11.2}$$
In short, fine-scale curvelets are obtained by bandpass filtering of Multiscale Ridgelets coefficients where the passband is rigidly linked to the scale of spatial localization.

- Anisotropy Scaling Law: By linking the filter passband $|\xi| \approx 2^{2s}$ to the scale of spatial localization 2^{-s} imposes that: (1) most curvelets are negligible in norm (most multiscale ridgelets do not survive the bandpass filtering D_s); (2) the nonnegligible curvelets obey $length \approx 2^{-s}$ while $width \approx 2^{-2s}$. In short, the system obeys approximately the scaling relationship

$$width \approx length^2.$$

  Note: it is at this last step that our 2^{2s} coronization scheme comes fully into play.
- Oscillatory Nature. Both for $s > 0$ and $s \leq 0$, each frame element has a Fourier transform supported in an annulus away from 0.

§12. Curvelets and Curved Singularities

12.1 Functions which are C^2 away from C^2 edges

We now formally specify a class of objects with discontinuities along edges; our notation and exposition are taken from [25,28,23]; related models were introduced some time ago in the mathematical statistics literature by [38,39]. It is clear that nothing in the arguments below would depend on the specific assumptions we make here, but the precision allows us to make our arguments uniform over classes of such objects.

A star-shaped set $B \subset [0,1]^2$ has an origin $b_0 \in [0,1]^2$ from which every point of B is 'visible'; i.e. such that the line segment $\{(1-t)b_0 + tb : t \in [0,1]\} \subset B$ whenever $b \in B$. This geometrical regularity is useful; it forces very simple interactions of the boundary with dyadic squares at sufficiently fine scales. We use this to guarantee that 'sufficiently fine' has a uniform meaning for every B of interest.

We define $\mathsf{Star}^2(A)$, a class of star-shaped sets with 2-smooth boundaries, by imposing regularity on the boundaries using a kind of polar coordinate system. Let $\rho(\theta) : [0, 2\pi) \to [0,1]$ be a radius function and $b_0 = (x_{1,0}, x_{2,0})$ be an origin with respect to which the set of interest in star-shaped. Define $\Delta_1(x) = x_1 - x_{1,0}$ and $\Delta_2(x) = x_2 - x_{2,0}$; then define functions $\theta(x_1, x_2)$ and $r(x_1, x_2)$ by

$$\theta = \tan^{-1}(-\Delta_2/\Delta_1); \qquad r = ((\Delta_1)^2 + (\Delta_2)^2)^{1/2}.$$

For a starshaped set, we have $(x_1, x_2) \in B$ iff $0 \leq r \leq \rho(\theta)$. In particular, the boundary ∂B is given by the curve

$$\beta(\theta) = (\rho(\theta)\cos(\theta) + x_{1,0}, \rho(\theta)\sin(\theta) + x_{2,0}). \tag{12.1}$$

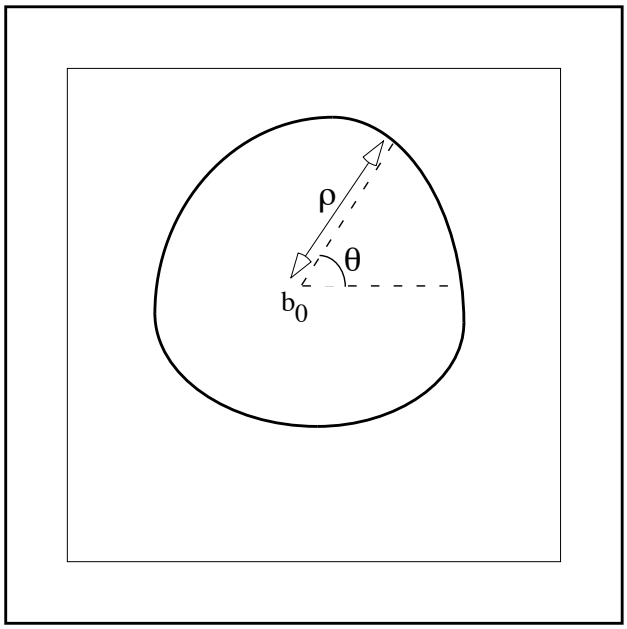

Fig. 3. Typical star-shaped set, and associated notation.

Fig. 3 gives a graphical indication of some of the objects just described. The class $\mathsf{Star}^2(A)$ of interest to us can now be defined by

$$\mathsf{Star}^2(A) = \{B : B \subset [\frac{1}{10}, \frac{9}{10}]^2, \quad \frac{1}{10} \leq \rho(\theta) \leq \frac{1}{2}, \quad \theta \in [0, 2\pi), \\ \rho \in \mathsf{Hölder}^2(A)\}.$$

Here the condition $\rho \in \mathsf{Hölder}^2(A)$ means that ρ is continuously differentiable and

$$|\rho'(\theta) - \rho'(\theta')| \leq A \cdot |\theta - \theta'|, \quad \theta, \theta' \in [0, 2\pi).$$

The actual objects of interest to us are functions which are twice continuously differentiable except for discontinuities along edges ∂B of star-shaped sets. We define $C_0^2(A)$ to be the collection of twice continuously differentiable functions supported strictly inside $[0, 1]^2$.

Definition 12.1. *Let $\mathcal{E}^2(A)$ denote the collection of functions f on $\mathbb{R}^2$ which are supported in the square $[0, 1]^2$ and obey*

$$f \;=\; f_1 + f_2 \cdot 1_B \tag{12.2}$$

where $B \in \mathsf{Star}^2(A)$, and each $f_i \in C_0^2(A)$. We speak of $\mathcal{E}^2(A)$ as consisting of functions which are C^2 away from a C^2 edge.

12.2 Near-optimality of curvelet expansions

The point is that the curvelet coefficient sequence $(c_\mu)_{\mu\in M}$ of an object f that is C^2 away from a C^2 edge, is in some sense, as sparse as if f were not singular.

Theorem 12.2. *Let $\mathcal{E}^2(A)$ be the collection (12.2) of objects which are C^2 away from a C^2 curve. There exists a constant C such that for every $f \in \mathcal{E}^2(A)$, the curvelet coefficient sequence $(c_\mu)_{\mu\in M}$ of f obeys*

$$\#\{\mu, \, , |c_\mu| \geq \epsilon\} \leq C \, \log(\epsilon^{-1}) \, \epsilon^{-2/3}. \tag{12.3}$$

There is a natural companion to this theorem. Let f_m be the m-term approximation of f obtained by extracting from the curvelet series

$$f = \sum_\mu \langle f, \gamma_\mu \rangle \gamma_\mu,$$

the terms corresponding to the m largest coefficients. Then,

Theorem 12.3. *Under the assumptions of Theorem 12.2, we have*

$$\|f - f_m\|_{L_2}^2 \leq C \, m^{-2} \, (\log m)^3. \tag{12.4}$$

The proof of Theorem 12.3 is very technical and over 35 pages long. We cannot possibly give an idea of the argument here.

Comparing with the results (10.1)–(10.3) we see that m-term approximations in the curvelet frame are almost rate optimal, and in fact perform far better than m-term sinusoid or wavelet approximations, in an asymptotic sense.

12.3 Significance

I believe that Theorem 12.3 is very significant. In fact, it comes as a little surprise. Finding optimal representations for smooth objects singular along smooth edges has been a long standing problem in CHA. Some expressed the belief that the way to go about singularities was to use adaptive bases as in [25], i.e. bases that would depend upon the object to be approximated, see the discussion in [9]. In other words, the existence of orthobases, frames, or tight frames yielding approximation rates similar to (12.4) by naive thresholding was thought to be highly doubtful. In some sense, Theorem 12.3 disproves this conjecture.

Beyond Theorem 12.3, the curvelet transform introduces a new data structure or, mathematically speaking, fundamentally new linear functionals. It has great potential in various fields outside of approximation theory, such as statistical estimation, mathematical analysis, partial differential equations, scientific computing, and data processing [59,11,60,12].

§13. Discussion

13.1 Renewed understanding

In this paper, we presented a whole set of new ideas and showed how one can deploy these ideas to tackle central problems in approximation theory.

First, we replaced a delicate and rather ill-posed problem of constructing ridge function approximations by a well-posed and constructive procedure, namely, the thresholding of ridgelet expansions. We proved that ridgelet thresholding is a viable substitute for traditional neural network approximations, as not only, it is more constructive, but it also rivals —at least asymptotically— rates attainable by very abstract approximation neural net procedures.

Second, we were able to identify those objects that are well approximated by ridgelet thresholding. We proved that ridgelet provide optimally sparse representations of smooth objects with discontinuities along hyperplanes of codimension 1. We introduced a new scale of functional spaces which have an intimate relationship with ridge-wavelet, ridge-free-knot splines or ridgelet approximations. We explained why, in some sense, this relationship is similar to the role that the Besov scale plays vis a vis free-knot splines or wavelet approximations.

13.2 Multiscale geometric analysis (MGA)

At the same time, ridgelet analysis gives decisive insights about the limitations and capabilities of ridge function approximation. For instance, we proved that ridgelets are suboptimal for approximating objects with curved inhomogeneities. In this manuscipt, we used those limitations as a motivation for further CHA constructions such as the curvelet transform.

We would like to emphasize that there is a wealth of new multiscale constructions which use the ridgelet transform as a core component. In addition to the curvelet transform, we would like to mention the possibility of constructing ridgelet packets which would involve the joint recursive dyadic splitting of both time and frequency. It is not the scope of this paper to discuss such constructions. However, the reader familiar with ideas such as Recursive Dyadic Partitioning on the one hand, and the curvelet pyramid on the other, will see that one can obtain a pyramid of windowed and filtered ridgelets at all lengths, and widths and that fast algorithms for searching sparse decompositions are very likely to exist.

In another direction, [24] explores a natural extension of the ideas developed here, namely, the construction of k-plane ridgelets in spaces of arbitrary dimension d.

There are undoubtedly many other possible multiscale constructions and, in some sense, it is only the beginning of a new subject we thought

–together with David Donoho– about calling MGA as in *Multiscale Geometric Analysis*. This is a rather unexplored territory with many research opportunities both at the theoretical and practical level.

13.3 Other connections between approximation theory and CHA

There are other opportunities for building bridges between approximation theory and CHA. We close this paper by describing one potential connection which is especially dear to my heart.

It seems of interest to bring together concepts of approximation and statistical theory with those of time-frequency analysis. There are interesting issues such as: how well can we approximate or estimate certain classes of rapidly oscillating functions where the oscillation rate may be changing over time? Can we construct approximations or estimators that will match the best theoretical performances? Time-frequency analysis provides a natural framework for studying these problems, but the tools to address them remain, I believe, to be constructed.

In time-frequency analysis, signals of primary interest oscillate rapidly and their frequency of oscillations is also rapidly changing over time. Those signals commonly referred to as chirps take the general form

$$f(t) = A(t)\cos(N\phi(t)); \tag{13.2}$$

here N is a (large) base frequency, $\phi(t)$ is time-varying and the amplitude $A(t)$ is slowly varying.

The method of choice for representing chirps is to use Gabor frames of the form

$$g_{m,n}(t) = w((t - mL)/L)\, e^{i2\pi n\, L\, t}, \quad n, m \in \mathbb{Z};$$

that is smoothly windowed sinusoids. This representation breaks the time-frequency plane into congruent rectangles and assigns one basis function to each rectangle. Schematically, Gabor approximations correspond to piecewise approximation of the instantaneous frequency. In a sense, Gabor frames are as good for approximating smooth chirps as wavelets are for approximating objects with edges. This brings up a central question: is there a nonadaptive representation which provides optimally sparse decompositions of chirps just as curvelets do for objects with smooth edges? To put it bluntly: is there something beyond Gabor? A mathematical result in this direction – either positive or negative – would certainly be very significant.

13.4 Closing

During my Ph.D. at Stanford, I recall that on several occasions I asked David Donoho the reason why he was not writing a book about the connections between harmonic analysis, statistical estimation, approximation theory and data compression. At the time (1998), he was drafting a manuscript entitled "Harmonic analysis and data compression" [29] for a special issue of *IEEE Trans. Inform. Theory* in the honor of the 50th birthday of Shannon's seminal paper on communication theory. He answered rather vaguely.

Two years later, he obviously had a very different attitude about this as he gave ten lectures about this same topic at the CBMS-NSF Regional Conference 2000 in Applied Mathematics. Much of those lectures were about ridgelets and curvelets. As if the picture suddenly became more complete, and much richer...

§14. Appendix

14.1 Fundamental estimates

The purpose of this technical section is to show that the kernel

$$T(R, R') = \langle \psi_R, \psi_{R'} \rangle_{L_2(w)}, \quad R, R' \in \mathcal{R} \tag{14.1}$$

is "almost diagonal." The fact that $\langle \psi_R, \psi_{R'} \rangle_{L_2(w)}$ decays rapidly as the "distance" between the indices (R, R') increases is a crucial fact of our analysis.

Let f_u and $g_{u'}$ be two ridge functions given by $f_u(x) = f(u \cdot x)$ and $g_{u'}(x) = g(u' \cdot x)$, respectively. Of course, $\langle f_u, g_{u'} \rangle$ makes no sense, as the ridge functions are not square-integrable. We then take a fixed window w in $\mathcal{S}(\mathbb{R}^d)$ and look at the inner product with respect to the signed measure w. We set

$$\langle f_u, g_{u'} \rangle_{L_2(w)} = \int f(u \cdot x) g(u' \cdot x) w(x)\, dx, \tag{14.2}$$

and derive a simpler expression for this quantity. Let Q be an orthogonal change of coordinates ($x = Qx'$) such that

$$u' \cdot x = x'_1 \qquad \text{and} \qquad u \cdot x = (u \cdot u') x'_1 - \sqrt{1 - (u \cdot u')^2}\, x'_2,$$

and put

$$v = u \cdot u' / \sqrt{1 - (u \cdot u')^2}; \tag{14.3}$$

letting θ be the angle between the unit vectors u and u' gives

$$\cos\theta = u \cdot u', \quad v = \cos\theta / |\sin\theta|. \tag{14.4}$$

With these notations, the inner product (14.2) can be rewritten as

$$\langle f_u, g_{u'} \rangle_{L_2(w)} = \int f(x'_1 \cos\theta - x'_2 |\sin\theta|)\; g(x'_1)\; w(Qx')\, dx'_1 dx'_2 \ldots dx'_d. \tag{14.5}$$

Define $w_Q(x_1, x_2)$ by $\int w(Qx)\, dx_3 \ldots dx_d$. Then,

$$\langle f_u, g_{u'} \rangle_{L_2(w)} = \int f(x_1 \cos\theta - x_2 |\sin\theta|)\; g(x_1)\; w_Q(x_1, x_2)\, dx_1 dx_2. \tag{14.6}$$

Examine now the partial derivatives of w_Q. Let D_1 be the partial derivative with respect to the first variable, i.e. x_1, and similarly for D_2. It is trivial to see that w_Q belongs to $\mathcal{S}(\mathbb{R}^2)$. Moreover, for any n_1, n_2, and each $m > 0$, there is a constant C depending upon w, n_1, n_2, and m such that

$$\begin{aligned} \sup_Q |D_1^{n_1} D_2^{n_2} w_Q(x_1, x_2)| &\le C\; (1 + |x_1| + |x_2|)^{-2m} \\ &\le C\; (1 + |x_1|)^{-m} (1 + |x_2|)^{-m}. \end{aligned} \tag{14.7}$$

In this section we will assume that ψ satisfies a few standard conditions: namely, ψ is R times differentiable and for every nonnegative integer m there is a constant C (depending on m) so that

$$|(D^n \psi)(t)| \le C(1 + |t|)^{-m}$$

for some constant C (depending only upon m and n). Further, we will suppose that ψ has vanishing moments through order D. The next lemma gives an upper bound on the inner product

$$K(j, u, b; j', u', b') := \int \psi_j(u \cdot x - b) \psi_{j'}(u' \cdot x - b') w(x)\, dx,$$

where $\psi_j(t) = 2^{j/2} \psi(t)$.

Lemma 14.1. *Assume $j' \ge j$ and suppose $n < \min(R, D)$. Then, for each $m \ge 0$, there is a constant C (depending on n and m) so that*

$$|K(j, u, b; j', u', b')| \le C\, 2^{-(j'+j)(n+1/2)}\, \delta_j^{2n+1}(\theta)\, T_j(u, b; u', b'); \tag{14.8}$$

here, θ is the angle between u and u' and

$$\delta_j(\theta) = \min(2^j, |\sin\theta|^{-1})$$

and

$$T_j(u, b; u', b') = (1 + |b'|)^{-m}\; (1 + \delta_j(\theta) |b' \cos\theta - b|)^{-m}\,.$$

Proof: Equation (14.6) gives

$$K(j,u,b;j',u',b') = \int_{\mathbb{R}^2} \psi_{j'}(x_1-b')\psi_j(x_1\cos\theta - x_2|\sin\theta| - b) w_Q(x_1,x_2)\,dx_1dx_2.$$

Suppose first that $|\sin\theta| \geq 2^{-j}$ so that $\delta_j(\theta) = |\sin\theta|^{-1}$. The change of variables $x_1' = x_1, x_2' = x_1\cos\theta - x_2|\sin\theta|$ allows rewriting the kernel as

$$K(j,u,b;j',u',b') = \int_{\mathbb{R}^2} \psi_{j'}(x_1'-b')\,\psi_j(x_2'-b)\,w_Q(x_1', |\sin\theta|^{-1}(x_1'\cos\theta - x_2')\,dx_1'dx_2'.$$

Now let $\tilde{w}_Q$ be the function defined by $\tilde{w}_Q(x_1,x_2) = w_Q(x_1, |\sin\theta|^{-1}(x_1\cos\theta - x_2))$. For each pair of integers n_1, n_2. We check that $\tilde{w}_Q$ obeys

$$|D_1^{n_1}D_2^{n_2}\tilde{w}_Q)(x_1,x_2)| = C\,|\sin\theta|^{-(n_1+n_2)} (1+|x_1|+|\sin\theta|^{-1}|x_1\cos\theta - x_2|)^{-2m}, \tag{14.9}$$

where again, the constant C does not depend on Q. This property is directly inherited from (14.7) and we omit the proof. Let n be an integer. By assumption, the wavelet ψ is of class $\mathcal{C}^R$ for $R > n$ and has at least n vanishing moments. It follows from standard wavelet estimates that one can find a constant C such that

$$|K(j,u,b;j',u',b')| \leq C\,2^{-(j'+j)(n+1/2)}|\sin\theta|^{-(2n+1)} (1+|b'|+|\sin\theta|^{-1}|b'\cos\theta - b|)^{-2m}. \tag{14.10}$$

The proof is essentially an integration by parts. We have

$$\begin{aligned} K(j,u,b;j',u',b') &= \langle \psi_{j'}(\cdot - b')\otimes\psi_j(\cdot - b), \tilde{w}_Q\rangle \\ &= \langle D_1^{-n}\psi_{j'}(\cdot - b')\otimes D_2^{-n}\psi_j(\cdot - b), D_1^nD_2^n\tilde{w}_Q\rangle. \end{aligned}$$

Now the inequality follows from the estimate about the size of the partial derivatives of $\tilde{w}_Q$ and the localization properties of $D^{-n}\psi$ as for each n and m, there is a constant C such that

$$|D^{-n}\psi_j(t)| \leq C\,2^{-j(n-1/2)}\,(1+2^j|t|)^{-m}.$$

The point is that the bulk of the mass of $|D_1^{-n}\psi_{j'}(\cdot - b')\otimes D_2^{-n}\psi_j(\cdot - b)|$ is concentrated near the rectangle $[b'\pm c2^{-j'}]\times[b\pm c2^{-j}]$. The upper bound

on the function $D_1^n D_2^n \tilde{w}_Q$ is slowly varying over this effective support since $|\sin\theta|^{-1} \le 2^j$. The bound (14.10) follows from the fact that

$$\|D_1^{-n}\psi_{j'}(\cdot - b') \otimes D_2^{-n}\psi_j(\cdot - b)\|_{L_1(\mathbb{R}^2)} \le 2^{-(j+j')/2}\|\psi\|_{L_1(\mathbb{R}^2)}$$

and the fact that over the effective support, the upper bound is about $C\,(1 + |b'| + |\sin\theta|^{-1}|b'\cos\theta - b|)^{-2m}$. This reasoning may be turned into a rigorous argument and the calculations giving (14.10) are absolutely standard, see [47] for instance.

The inequality (14.10) together with

$$(1+|b'|+|\sin\theta|^{-1}|b'\cos\theta-b|)^{-2m} \le (1+|b'|)^{-m}(1+|\sin\theta|^{-1}|b'\cos\theta-b|)^{-m}$$

give the result for $\delta_j(\theta) = |\sin\theta|^{-1}$.

The argument is a little different when the ridgelets are nearly parallel; that is, when $|\sin\theta| \le 2^{-j}$ or equivalently $\delta_j(\theta) = 2^j$. We apply Fubini's theorem and express the inner product as

$$K(j,u,b;j',u',b') = \int \psi_{j'}(t-b')\,g(t)\,dt,$$

where

$$g(t) = \int_{\mathbb{R}} \psi_j(t\cos\theta - t'\sin\theta - b) w_Q(t,t')\,dt'.$$

The function g is smooth and well localized. On the one hand, for any n there is a constant C such that for any $\ell \le n$

$$\left|\frac{d^\ell}{dt^\ell}\psi_j(t\cos\theta - t'\sin\theta - b)\right| \le C\ 2^{j(\ell+1/2)}(1 + 2^j|t\cos\theta - t'\sin\theta - b|)^{-m}.$$

On the other,

$$\left|\frac{\partial^{n-\ell}}{\partial t^{n-\ell}} w(t,t')\right| \le C\ (1 + |t| + |t'|)^{-2m} \le C\ (1+|t|)^{-m}(1+|t'|)^{-m}.$$

Now, an integration by parts gives

$$\left|\frac{d^n}{dt^n} g(t)\right| \le C\ 2^{j(n+1/2)}(1+|t|)^{-m}(1 + 2^j|t\cos\theta - b|)^{-m}. \qquad (14.12)$$

Again, we apply classical wavelet analysis techniques for bounding the coefficients of a smooth and well-localized function. The argument is essentially an integration by parts as explained above. We write

$$K(j,u,b;j',u',b') = \langle D^{-n}\psi_{j'}(\cdot - b'), D^n g\rangle,$$

and use (14.12) and (14.11) to obtain

$$|K(j,u,b;j',u',b')| = \left| \int \psi_{j'}(t-b')g(t)\,dt \right| \\ \leq C2^{-(j'-j)(n+\frac{1}{2})}(1+|b'|)^{-m}(1+2^j|b'\cos\theta - b|)^{-m}.$$

This last inequality is the content of (14.8) when $|\sin\theta| \leq 2^{-j}$. The lemma is proved. □

For $j' \geq j$, Lemma 14.1 gives the upper bound on the entries of the Gramm matrix (14.1)

$$|T(R,R')| \leq C2^{-(j'+j)(n+1/2)}\,\delta_j^{2n+1}(\ell,\ell')\,L(R,R'), \tag{14.13}$$

with

$$L(R,R') = \left(1+|k'2^{-j'}|\right)^{-m}\ \left(1+\delta_j(\ell,\ell')|k'2^{-j'}\cos\theta - k2^{-j}|\right)^{-m}. \tag{14.14}$$

Here, δ_j is as before, namely,

$$\delta_j(\ell,\ell') = \min(2^j, 1/|\sin\theta|)$$

where θ is the angle between $u_{j,\ell}$ and $u_{j',\ell'}$. The estimate for $j' < j$ is obtained by symmetry.

Lemma 14.2. *Suppose ψ is C^∞ and has vanishing moments up to any order. Then for any $p > 0$, we have*

$$\sup_{R'} \sum_{R} |T(R,R')|^p \leq A_p. \tag{14.15}$$

(If $r = \min(R,D)$ is finite, then (14.15) holds for $p > p^*$ with $(r+1/2)p^* = d$.)

Proof: First, for $j' \geq j$, (14.14) gives

$$\sum_k |L(R,R')|^p \leq \sum_k \left(1+\delta_j(\ell,\ell')|k'2^{-j'}\cos\theta - k2^{-j}|\right)^{-mp} \\ \leq C\,2^j\,|\delta_j(\ell,\ell')|^{-1}, \tag{14.16}$$

provided that $mp > 1$. Second, for $j > j'$ and m large enough so that $mp > 1$, we have

$$\sum_k |L(R,R')|^p \\ = \sum_k \left(1+|k2^{-j}|\right)^{-mp}\left(1+\delta_{j'}(\ell,\ell')|k2^{-j}\cos\theta - k'2^{-j'}|\right)^{-mp} \\ \leq C\,2^j\ \min(1,(|\delta_{j'}(\ell,\ell')\cos\theta|^{-1}).$$

In particular, this last upper bound gives

$$\sum_k |L(R,R')|^p \le \begin{cases} C\,2^j, & |\cos\theta| < 1/\sqrt{2} \\ C\,2^j\,|\delta_{j'}(\ell,\ell')|, & |\cos\theta| \ge 1/\sqrt{2}. \end{cases} \tag{14.17}$$

We now sum the inequalities (14.16)–(14.17) over the angular variable. To develop upper bound, we shall use a little lemma whose proof is postponed.

Lemma 14.3. *Let $\beta > d-1$. Then,*

$$\sum_\ell |\delta_j(\ell,\ell')|^\beta \le C\,2^{j\beta}.$$

and, for $j' \le j$

$$\sum_\ell |\delta_{j'}(\ell,\ell')|^\beta \le C\,2^{j'\beta}\,2^{(j-j')(d-1)}.$$

Suppose $j' \ge j$. The bound (14.16) yields

$$\sum_k |T(R,R')|^p \le C\,2^{-(j+j')(n+1/2)p}\,2^j\,|\delta_j(\ell,\ell')|^{(2n+1)p-1}.$$

For $(2n+1)p > d$, Lemma 14.3 gives

$$\sum_\ell |\delta_j(\ell,\ell')|^{(2n+1)p-1} \le C\,2^{j((2n+1)p-1)},$$

and, therefore,

$$\sum_\ell \sum_k |T(R,R')|^p \le C\,2^{-(j+j')(n+1/2)p}\,2^{j(2n+1)p} = C\,2^{-(j'-j)(n+1/2)p}.$$

Let us turn our attention to the case $j' \le j$. For $|\cos\theta| > 1/\sqrt{2}$, the bound (14.17) yields

$$\sum_k |T(R,R')|^p \le C\,2^{-(j+j')(n+1/2)p}\,2^j\,|\delta_{j'}(\ell,\ell')|^{(2n+1)p-1}.$$

For $(2n+1)p > d$, Lemma 14.3 now gives

$$\sum_\ell |\delta_{j'}(\ell,\ell')|^{(2n+1)p-1} \le C\;2^{j'((2n+1)p-1)}\;2^{(j-j')(d-1)},$$

and, therefore,

$$\sum_{\ell:|\cos\theta|>1/\sqrt{2}} \sum_k |T(R,R')|^p \le C\, 2^{-(j+j')(n+1/2)p}\, 2^{j'((2n+1)p}\, 2^{(j-j')d}$$
$$= C\, 2^{-(j-j')((n+1/2)p-d)}.$$

Moreover, For $|\cos\theta| \ge 1/\sqrt{2}$, the bound (14.17) yields

$$\sum_k |T(R,R')|^p \le C\, 2^{-(j+j')(n+1/2)p}\, 2^j\, |\delta_{j'}(\ell,\ell')|^{(2n+1)p}.$$

Observe that in the range $|\cos\theta| \ge 1/\sqrt{2}$, we have $|\delta_{j'}(\ell,\ell')| \le \sqrt{2}$ since $\delta_{j'}(\ell,\ell') \le 1/|\sin\theta|$. The last inequality then becomes

$$\sum_k |T(R,R')|^p \le C\, 2^{-(j+j')(n+1/2)p}\, 2^j,$$

which is independent of ℓ. At scale j, the total number of orientations is $O(2^{j(d-1)})$ and, hence,

$$\sum_{\ell:|\cos\theta|\le 1/\sqrt{2}} \sum_k |T(R,R')|^p \le C\, 2^{-(j+j')(n+1/2)p}\, 2^{jd}.$$

Take $2(n+1/2)p > d$, then for $j' \ge 0$, we have $2^{j'(2n+1)p-d)} \ge 1$ which gives

$$\sum_{\ell:|\cos\theta|\le 1/\sqrt{2}} \sum_k |T(R,R')|^p \le C\, 2^{-(j+j')(n+1/2)p}\, 2^{jd}\, 2^{j'(2n+1)p-d)}$$
$$= C\, 2^{-(j-j')((n+1/2)p-d)}.$$

Collecting our results, we showed that for $j' \le j$

$$\sum_{\ell:|\cos\theta|>1/\sqrt{2}} \sum_k |T(R,R')|^p \le C\, 2^{-(j-j')((n+1/2)p-d)}.$$

To summarize, we developed the following upper bound:

$$\sum_{|R|=j} |T(R,R')|^p \le \begin{cases} C2^{-|j-j'|((n+1/2)p-d)}, & j' \le j \\ C2^{-|j-j'|(n+1/2)p}, & j' \ge j. \end{cases} \tag{14.18}$$

It is now clear that choosing n such that $(n+1/2)p > d$ gives

$$\sum_R |T(R,R')|^p \le C,$$

where C is independent of R'. This finishes the proof of the lemma. □

Proof of Lemma 14.3: Let u_0 be a fixed direction and let θ be the angle between u_0 and a running point u on the sphere. Define

$$\begin{aligned} \Lambda_0 &= \{u, |\sin\theta| \geq 1/2\}, \\ \Lambda_m &= \{u, 2^{-(m+1)} \leq |\sin\theta| < 2^{-m}\},\ m = 1, 2, \ldots \end{aligned} \tag{14.19}$$

and let $\Lambda_{j,m}$ denote the set of indices obeying

$$\Lambda_{j,m} := \{\ell,\, u_{j,\ell} \in \Lambda_m\} \tag{14.20}$$

with the convention that for $m = j$, we will say that $\ell \in \Lambda_{j,m}$ if $u_{j,\ell} \in \Lambda_m$, for some $m \geq j$.

We recall that at scale j, the set of discrete angular variables $\{u_{j,\ell}, \ell \in \Lambda_j\}$ consists of points approximately uniformly distributed on the sphere. In particular, for $m = 0, 1, \ldots, j$, we have

$$\#\,\Lambda_{j,m} \leq C\,\left(1 + 2^{j(d-1)}\mu(\Lambda_m)\right), \tag{14.21}$$

where μ is the uniform probability measure on the unit sphere of $\mathbb{R}^d$. In other words, the number of points falling in the set Λ_m, is essentially bounded — up to a multiplicative constant — by the total number of sampled points on the sphere times the area of the set Λ_m. As one can see, this says that the points $u_{j,\ell}$ are approximately equi-distributed on the sphere.

To calculate $\mu(\Lambda_m)$, we introduce the spherical coordinates defined by $x_1 = \cos\theta_1$, $x_2 = \sin\theta_1\cos\theta_2$, …, $x_d = \sin\theta_1\sin\theta_2\ldots\sin\theta_{d-1}$, $0 \leq \theta_1, \ldots, \theta_{d-2} \leq \pi$, $0 \leq \theta_{d-1} < 2\pi$. Let θ_1 be the angle between the fixed orientation u_0 and the running point u. With these notations, the distribution of θ_1 is proportional to $(\sin\theta_1)^{d-2}\,d\theta_1$, where the normalizing constant guarantees that the density integrates up to one. Then,

$$\mu(\Lambda_m) = c_d \int_{2^{-(m+1)} \leq |\sin\theta| \leq 2^{-m}} |\sin\theta|^{d-2}\,d\theta.$$

This last equation gives the upper bound

$$\mu(\Lambda_m) \leq C\,2^{-m(d-1)},$$

which also holds for the special case $m = 0$. Therefore, from (14.21) we deduce a bound on the cardinality of $\Lambda_{j,m}$:

$$\#\Lambda_{j,m} \leq C(1 + 2^{(j-m)(d-1)}), \quad m = 0, 1, 2, \ldots, j.$$

Observe that the above inequality is also valid in the special case where $m \in \{0, j\}$.

After these preliminaries, we are in a position to prove the lemma. Let $\beta > d-1$. Over the subset of indices such that $\{\ell : |\sin\theta| \geq 2^{-j}\}$, we have

$$\sum_{\ell:|\sin\theta|\geq 2^{-j}} |\delta_j(\ell,\ell')|^\beta \leq \sum_{m=0}^{j-1} 2^{(m+1)\beta}\, \#\Lambda_{j,m}$$
$$\leq C \sum_{m=0}^{j-1} 2^{(m+1)\beta}(1 + 2^{(j-m)(d-1)}) \leq C\, 2^{j\beta}.$$

And for those ℓ's such that $|\sin\theta| \leq 2^{-j}$, we have

$$\sum_{\ell:|\sin\theta|\leq 2^{-j}} |\delta_j(\ell,\ell')|^\beta \leq 2^{j\beta}\, \#\Lambda_{j,j} \leq C\, 2^{j\beta}.$$

The last two inequalities yield the first part of the lemma.

Next, assume $j' \leq j$. Over the subset of indices such that $\{\ell : |\sin\theta| \geq 2^{-j'}\}$, we have

$$\sum_{\ell:|\sin\theta|\geq 2^{-j'}} |\delta_{j'}(\ell,\ell')|^\beta \leq \sum_{m=0}^{j'-1} 2^{(m+1)\beta}\, \#\Lambda_{j,m} \leq 2^{j'\beta}\, 2^{(j-j')(d-1)}.$$

And for those ℓ's such that $|\sin\theta| \leq 2^{-j'}$, we have

$$\sum_{\ell:|\sin\theta|\leq 2^{-j'}} |\delta_{j'}(\ell,\ell')|^\beta \leq 2^{j'\beta} \sum_{m=j'}^{j} \#\Lambda_{j,m} \leq C\, 2^{j'\beta}\, 2^{(j-j')(d-1)}.$$

The last two inequalities yield the second part of the lemma. □

We close this section by proving similar estimates for the inner product

$$K_0(u,b;j',u',b') := \int \phi(u\cdot x - b)\psi_{j'}(u'\cdot x - b')w(x)\,dx,$$

where φ is a coarse scale function as in Section 9. Let φ be R times differentiable such that, for each $m \geq 0$, its derivatives up to order R obey

$$|(D^n\varphi)(t)| \leq C(1+|t|)^{-m}.$$

The same arguments as in the proof of Lemma 14.8 show that

$$|K_0(u,b;j',u',b')| \leq C\, 2^{-j'(n+1/2)} \left(1+|b'|\right)^{-m} \left(1+|b'\cos\theta - b|\right)^{-m}. \tag{14.22}$$

In particular, this last inequality gives ($j' \geq 0$)

$$|T_0(R', R)| \leq C\, 2^{-j'(n+1/2)} \left(1 + |k'2^{-j'}|\right)^{-m} \left(1 + |k'2^{-j'} \cos\theta - k|\right)^{-m},$$

with T_0 as in Section 9. Then, for $m > 1/p$, we have

$$\sum_k |T_0(R', R)|^p \leq C\, 2^{-j'(n+1/2)p} \left(1 + |k'2^{-j'}|\right)^{-mp},$$

and further

$$\sum_k \sum_{k'} |T_0(R', R)|^p \leq C\, 2^{-j(n+1/2)p}\, 2^{j'}.$$

We sum the previous inequalities over the angular variables. At scale j', the total number of orientations is $O(2^{j'(d-1)})$ and, hence,

$$\sum_{|R'|=2^{j'}} \sum_{|R|=0} |T_0(R', R)|^p \leq C\, 2^{-j'(n+1/2)p}\, 2^{j'd}. \tag{14.23}$$

It is clear that for $(n + 1/2)p > d$, T_0 is maps bounded ℓ_∞ sequences into ℓ_p bounded sequences, say.

Acknowledgments. This work was supported by an Alfred P. Sloan Fellowship. I would like to dedicate this paper to all the people who have helped me in my young career, and especially to David L. Donoho.

References

1. Auer, P., M. Hebster, and M. K. Warmuth, Exponentially many local minima for single neurons, in *Advances in Neural Information Processing Systems*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (eds.), volume 8, The MIT Press, 1996, 316–322.
2. Barron, A. R., Universal approximation bounds for superpositions of a sigmoidal function, IEEE Trans. Inform. Theory **39** (1993), 930–945.
3. Candès, E. J., Ridgelets: theory and applications, PhD thesis, Department of Statistics, Stanford University, 1998.
4. Candès, E. J., Harmonic analysis of neural networks, Appl. Comput. Harmonic Anal. **6** (1999), 197–218.
5. Candès, E. J., Ridgelets: Estimating with ridge functions, Technical report, Department of Statistics, Stanford University, 1999. Submitted for publication.

6. Candès, E. J., Ridgelets and sigmoidal neural networks, Submitted for publication, see `http://www.acm.caltech.edu/~emmanuel/publications.html`, 2001.
7. Candès, E. J., Ridgelets and the representation of mutilated Sobolev functions, SIAM J. Math. Anal. **33** (2001), 347–368.
8. Candès, E. J., and D. L. Donoho. Curvelets, manuscript, see `http://www-stat.stanford.edu/donoho/Reports/1998/curvelets.zip`, 1999.
9. Candès, E. J., and D. L. Donoho, Curvelets—a surprisingly effective nonadaptive representation for objects with edges, in *Curve and Surface Fitting: Saint-Malo 1999*, Albert Cohen, Christophe Rabut, and Larry L. Schumaker (eds.), Vanderbilt University Press, Nashville, 2000, 105–120.
10. Candès, E. J., and D. L. Donoho, Ridgelets: the key to higher dimensional intermittency?, Phil. Trans. R. Soc. Lond. A. **357** (1999), 2495–2509.
11. Candès, E. J., and D. L. Donoho, Recovering edges in ill-posed inverse problems: Optimality of curvelet frames, Technical report, Department of Statistics, Stanford University, 2000, to appear Ann. Statist.
12. Candès, E. J., and D. L. Donoho. Curvelets and curvilinear integrals, J. Approx. Theory **113** (2001), 59–90.
13. Carroll, S. M., and B. W. Dickinson, Construction of neural nets using the Radon tranform, in *Proceedings of the IEEE 1989 International Joint Conference on Neural Networks,* IEEE, New York, 1989, 661–667.
14. Chen, S. S., Basis Pursuit, PhD thesis, Department of Statistics, Stanford University, 1995.
15. Chen, S. S., D. L. Donoho, and M. A. Saunders, Atomic decomposition by basis pursuit, SIAM J. Sci. Comput. **20** (1999), 33–61.
16. Cheng, B., and D. M. Titterington, Neural networks: a review from a statistical perspective. With comments and a rejoinder by the authors, Stat. Sci. **9** (1994), 2–54.
17. Cybenko, G., Approximation by superpositions of a sigmoidal function, Math. Control Signals Systems **2** (1989), 303–314.
18. Deans, S. R., *The Radon Transform and Some of Its Applications,* John Wiley & Sons, 1983.
19. DeVore, R. A., R. Howard, and C. A. Micchelli, Optimal nonlinear approximation, Manuscripta Mathematica **63** (1989), 469–478.
20. DeVore, R. A., B. Jawerth, and V. Popov, Compression of wavelet decompositions, Amer. J. Math. **114** (1992), 737–785.

21. DeVore, R. A., K. I. Oskolkov, and P. P. Petrushev, Approximation by ridge functions and neural networks, Ann. Numer. Math. **4** (1997), 261–287. The heritage of P. L. Chebyshev: a Festschrift in honor of the 70th birthday of T. J. Rivlin.

22. Donoho, D. L., Unconditional bases are optimal bases for data compression and for statistical estimation, Appl. Comput. Harmonic Anal. **1** (1993), 100–115.

23. Donoho, D. L., Sparse components analysis and optimal atomic decomposition, Technical report, Department of Statistics, Stanford University, 1998.

24. Donoho, D. L., Tight frames of k-plane ridgelets and the problem of representing objects that are smooth away from d-dimensional singularities in $\mathbb{R}^n$, Proc. Natl. Acad. Sci. USA **96** (1999), 1828–1833.

25. Donoho, D. L., Wedgelets: nearly minimax estimation of edges, Ann. Statist, **27** (1999), 859–897.

26. Donoho, D. L., Orthonormal ridgelets and linear singularities, SIAM J. Math. Anal. **31** (2000), 1062–1099.

27. Donoho, D. L., Ridge functions and orthonormal ridgelets, J. Approx. Theory **111** (2001), 143–179.

28. Donoho, D. L., and I. M. Johnstone, Empirical atomic decomposition, manuscript, 1995.

29. Donoho, D. L., M. Vetterli, R. A. DeVore, and I. Daubechies, Data compression and harmonic analysis, IEEE Trans. Inform. Theory **44** (1998), 2435–2476.

30. M. Frazier, B. Jawerth, and G. Weiss, *Littlewood-Paley Theory and the Study of Function Spaces,* volume 79 of NSF-CBMS Regional Conf. Ser. in Mathematics, American Math. Soc., Providence, RI, 1991.

31. Friedman, J. H., and W. Stuetzle, Projection pursuit regression. J. Amer. Statist. Assoc. **76** (1981), 817–823.

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

33. Helgason, S., *The Radon Transform*, Birkhäuser, Boston, second edition, 1999.

34. Hornik, K., M. Stinchcombe, and H. White, Multilayer feedforward networks are universal approximators, Neural Networks **2** (1989), 359–366.

35. John, F., *Plane Waves and Spherical Means Applied to Partial Differential Equations*, Interscience Publishers, Inc., New York, 1955.

36. 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.

37. Jones, L. K., The computational intractability of training sigmoidal neural networks, IEEE Transactions on Information Theory **43** (1997), 167–173.

38. Khas'minskii, R. Z., and V. S. Lebedev, On the properties of parametric estimators for areas of a discontinuous image, Problems Control Inform. Theory, 1990.

39. Korostelev, A. P., and A. B. Tsybakov, *Minimax Theory of Image Reconstruction,* volume 82 of Lecture Notes in Statistics, Springer-Verlag, New York, 1993.

40. Le Pennec, E., and S. Mallat, Image compression with geometrical wavelets, in *Proceedings of International Conference on Image Processing* (ICIP) 2000.

41. Lemarié, P. G., and Y. Meyer, Ondelettes et bases Hilbertiennes, Rev. Mat. Iberoamericana **2** (1986), 1–18.

42. Leshno, M., V. Y. Lin, A. Pinkus, and S. Schoken, Multilayer feedforward networks with a nonpolynomial activation function can approximate any function, Neural Networks **6** (1993), 861–867.

43. Logan, B. F., and L. A. Shepp, Optimal reconstruction of a function from its projections, Duke Math. J. **42** (1975), 645–659.

44. Maiorov, V., and A. Pinkus, Lower bounds for approximation by MLP neural networks, Neurocomputing **25** (1999), 81–91.

45. Maiorov, V. E., and R. Meir, On the near optimality of the stochastic approximation of smooth functions by neural networks, Adv. Comput. Math. **13** (2000), 79–103.

46. Makovoz, Y., Random approximants and neural networks, J. Approx. Theory **85** (1996), 98–109.

47. Meyer, Y., *Ondelettes et Opérateurs: II. Opérateurs de Calderón Zygmund*, Hermann, 1990.

48. Meyer, Y., *Wavelets and Operators*, Cambridge University Press, 1992.

49. Mhaskar, H. N., Neural networks for optimal approximation of smooth and analytic functions, Neural Computation **8** (1996), 164–177.

50. Murata, N., An integral representation of functions using three-layered networks and their approximation bounds, Neural Networks **9** (1996), 947–956.

51. Natterer, F., *The Mathematics of Computerized Tomography*, B. G. Teubner, John Wiley & Sons, 1986.

52. Oskolkov, K., Chebyshev—Fourier analysis and optimal quadrature formulas, Proc. Steklov Math. Inst. **219** (1997), 269–285. (in Russian).
53. Petrushev, P. P., Approximation by ridge functions and neural networks, SIAM J. Math. Anal. **30** (1999), 155–189.
54. Pinkus, A., Approximating by ridge functions, in *Surface Fitting and Multiresolution Methods*, A. Le Méhauté, C. Rabut, and L. L. Schumaker (eds.), Vanderbilt University Press, Nashville, 1997, 279–292.
55. Pinkus, A., Approximation theory of the MLP model in neural networks, Acta Numerica **8** (1999), 143–196.
56. Pisier, G., Remarques sur un résultat non publié de B. Maurey, in *Séminaire d'Analyse Fonctionelle, 1980–1981*, Centre des Mathématiques, Ecole Polytechnique, Palaiseau, France.
57. Plancherel, M., and G. Pólya, Fonctions entières et intégrales de Fourier multiples, Commentarii Math. Helv. **10** (1938), 110–163.
58. Rubin, B., The Calderón reproducing formula, windowed X-ray transforms and Radon transforms in L^p-spaces, J. Fourier Anal. Appl. **4** (1998), 175–197.
59. Starck, J. L., E. J. Candès, and D. L. Donoho, The curvelet transform for image denoising, IEEE Transactions on Image Processing, 2000, to appear.
60. Starck, J. L., E. J. Candès, and D. L. Donoho, Very high quality image restoration, in *Wavelet Applications in Signal and Image Processing IX*, A. Aldroubi, A. F. Laine, and M. A. Unser (eds.), Proc. SPIE 4478, 2001.
61. Vetterli, M., and J. Kovacevic, *Wavelets and Subband Coding*, Prentice Hall, Englewood Cliffs, NJ, 1995.
62. Vu, V. H., On the infeasibility of training neural networks with small mean-squared error, IEEE Transactions on Information Theory **44** (1998), 2892–2900.

Emmanuel J. Candès
Applied and Computational Mathematics
California Institute of Technology
Pasadena, California 91125
emmanuel@acm.caltech.edu
http://www.acm.caltech.edu/∼emmanuel