RIDGELETS: THEORY AND APPLICATIONS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Emmanuel Jean Candes

August 1998

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

David L. Donoho
(Principal Adviser)

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Iain M. Johnstone

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

George C. Papanicolaou

Approved for the University Committee on Graduate Studies:

Abstract

Single hidden-layer feedforward neural networks have been proposed as an approach to bypass the curse of dimensionality and are now becoming widely applied to approximation or prediction in applied sciences. In that approach, one approximates a multivariate target function by a sum of ridge functions; this is similar to projection pursuit in the literature of statistics. This approach poses new and challenging questions both at a practical and theoretical level, ranging from the construction of neural networks to their efficiency and capability. The topic of this thesis is to show that *ridgelets*, a new set of functions, provide an elegant tool to answer some of these fundamental questions.

In the first part of the thesis, we introduce a special admissibility condition for neural activation functions. Using an admissible neuron, we develop two linear transforms, namely the continuous and discrete ridgelet transforms. Both transforms represent quite general functions f as a superposition of ridge functions in a stable and concrete way. A frame of "nearly orthogonal" ridgelets underlies the discrete transform.

In the second part, we show how to use the ridgelet transform to derive new approximation bounds. That is, we introduce a new family of smoothness classes and show how they model "real-life" signals by exhibiting some specific sorts of high-dimensional spatial inhomogeneities. Roughly speaking, finite linear combinations of ridgelets are optimal for approximating functions from these new classes. In addition, we use the ridgelet transform to study the limitations of neural networks. As a surprising and remarkable example, we discuss the case of approximating radial functions.

Finally, it is explained in the conclusion why these new ridgelet expansions offer decisive improvements over traditional neural networks.

Acknowledgements

First, I would like to thank my advisor David Donoho whose most creative and original thinking have been for me a great source of inspiration. I admire his deep and penetrating views on so many areas of the mathematical sciences and feel particularly indebted to him for sharing his thoughts with me. Beyond the unique scientist, there is the friend whose kindness and generosity throughout my stay at Stanford have been invaluable. I also extend my gratitude to his wife, Miki.

I feel privileged to have had so many fantastic teachers and professors who nurtured my love and interest for science. I owe special thanks to Patrick David and to Professor Yves Meyer who shared their enthusiasm with me – a quality that I hope will be a lifetime companion.

I would also like to thank Professors Jerome Friedman, Iain Johnstone and George Papanicolaou for serving on my orals committee and for having, together with Professor Darrell Duffie, written letters of recommendation on my behalf.

I wish to thank all the people of the Department of Statistics for creating such a world-class scientific environment in which it is so easy to blossom; especially, the faculty which greatly enriched my scientific experience by exposing me to new areas of research.

A short acknowledgement seems to be very little to thank my parents for their constant love and support, and for the never-failing confidence they had in me.

My days at Stanford would not have been the same without Helen, for the countless little things she did so that I would feel "at home." I praise the courage she found to read and suggest improvements to this manuscript.

Finally, my deepest gratitude goes to my wife, Chiara, whose encouragement, humor and love have made these last four years a pure enjoyment.

Contents

List of Figures

Chapter 1

Introduction

Let $f(x) : \mathbf{R}^d \to \mathbf{R}$ be a function of d variables. In this thesis, we are interested in constructing convenient approximations to f using a system called *neural networks*. This problem is of wide interest throughout the mathematical sciences and many fundamental questions remain open. Because of the extensive use of neural networks, we will address questions from various perspectives and use these as guidelines for the present work.

1.1 Neural Networks

A single hidden-layer feedforward neural network is the name given a function of d-variables constructed by the rule

$$f_m(x) = \sum_{i=1}^{m} \alpha_i \rho(k_i \cdot x - b_i), \tag{1.1}$$

where the m terms in the sum are called neurons; the α_i and b_i are scalars; and the k_i are d-dimensional vectors. Each neuron maps a multivariate input $x \in \mathbf{R}^d$ into a real valued output by composing a simple linear projection $x \to k_i \cdot x - b_i$ with a scalar nonlinearity ρ, called the activation function. Traditionally, ρ has been given a sigmoid shape, $\rho(t) = e^t/(1+e^t)$, modeled after the activation mechanism of biological neurons. The vectors k_i specify the 'connection strengths' of the d inputs to the i-th neuron; the b_i specify activation thresholds. The use of this model for approximating functions in applied sciences, engineering, and finance is large and growing; for examples, see journals such as *IEEE Trans. Neural Networks*.

From a mathematical point of view, such approximations amount to taking finite linear combinations of atoms from the dictionary $\mathcal{D}_{Ridge} = \{\rho(k \cdot x - b); k \in \mathbf{R}^d, b \in \mathbf{R}\}$ of elementary *ridge functions*. As is known, any function of d variables can be approximated arbitrarily well by such combinations (Cybenko, 1989; Leshno, Lin, Pinkus, and Schocken, 1993). As far as constructing these combinations, a frequently discussed approach is the greedy algorithm that, starting from $f_0(x) = 0$, operates in a stepwise fashion running through steps $i = 1, \ldots m$; we inductively define

$$f_i = \alpha^* f_{i-1} + (1 - \alpha^*)\rho(k^* \cdot x - b^*), \tag{1.2}$$

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

$$\arg \min_{0 \le \alpha \le 1} \arg \min_{(k,b) \in \mathbf{R}^n \times \mathbf{R}} \|f - \alpha f_{i-1} + (1 - \alpha)\rho(k \cdot x - b)\|_2. \tag{1.3}$$

Thus, at the i-th stage, the algorithm substitutes to f_{i-1} a convex combination involving f_{i-1} and a term from the dictionary $\mathcal{D}_{Ridge}$ that results in the largest decrease in approximation error (1.3). It is known that when $f \in L^2(D)$ with D a compact set, the greedy algorithm converges (Jones, 1992b); it is also known that for a relaxed variant of the greedy algorithm, the convergence rate can be controlled under certain assumptions (Jones, 1992a; Barron, 1993). There are unfortunately two problems with the conceptual basis of such results.

First, they lack the constructive character which one ordinarily associates with the word "algorithm." In any assumed implementation of minimizing (1.3) one would need to search for a minimum within a discrete collection of k and b. What are the properties of procedures restricted to such collections? Or, more directly, how finely discretized must the collection be so that a search over that collection gives results similar to a minimization over the continuum? In some sense, applying the word "algorithm" for abstract minimization procedures in the absence of an understanding of this issue is a misnomer.

Second, even if one is willing to forgive the lack of constructivity in such results, one must still face the lack of stability of the resulting decomposition. An approximant $f_N(x) = \sum_{i=1}^N \alpha_i \rho(k_i \cdot x - b_i)$ has coefficients which in no way are continuous functionals of f and do not necessarily reflect the size and organization of f (Meyer, 1992).

1.2 Approximation Theory

Let alone the most delicate problem of their construction, one can look at neural networks from the viewpoint of approximation: that is, to investigate the efficiency of approximation of a function f by finite linear combinations of neurons taken from the dictionary $\mathcal{D}_{Ridge}$. Although this issue has received overwhelming attention (Barron, 1993; Cybenko, 1989; DeVore, Oskolkov, and Petrushev, 1997; Mhaskar, 1996; Mhaskar and Micchelli, 1995), there are surprisingly very few decisive results about the quantitative rates of these approximations.

First, there is a series of results which essentially amount to saying that neural networks are at least as efficient as polynomials for approximating functions (Mhaskar, 1996; Mhaskar and Micchelli, 1995); the argument being simply that since one can find good approximations of polynomials using neural networks, whenever there is a good polynomial approximation of a target function f, there is in principle a corresponding neural net approximation. Second, in a celebrated result, Barron (1993) and Jones (1992b) have been able to bound the convergence rate of the greedy algorithm (1.2)-(1.3), when f is restricted to satisfy some smoothness condition, namely f is a square integrable function over the unit ball Ω_d of $\mathbf{R}^d$ such that $\int_{\mathbf{R}^n} |\xi||\widehat{f}(\xi)|d\xi \leq C$ (here, $\widehat{f}$ denotes the Fourier transform of f). For this class, they show

$$\|f - f_N\|_2 \leq 2CN^{-1/2}, \tag{1.4}$$

where f_N is the output of the algorithm at stage N. Their result, however, also raises a set of challenging questions which we will now discuss.

The greedy algorithm. The work of DeVore and Temlyakov (1996) shows that the greedy algorithm has unfortunately very weak approximation properties. Even when good approximations exist, the greedy algorithm cannot be guaranteed to find them, even in the extreme case where f is just a superposition of a few, say ten, elements of our dictionary $\mathcal{D}_{Ridge}$.

Neural nets for which functions? It can be shown that for the class Barron considers, a simple N-term trigonometric approximation would give better rates of convergence; namely, $O(N^{-(1/2+1/d)})$ (and, of course, there is a real and fast algorithm). So, it would be of interest to be able to identify functional classes for which neural networks are more efficient than other methods of approximation or more ambitiously a class $\mathcal{F}$ for which it could be proved that linear combinations of elements of $\mathcal{D}_{Ridge}$ give the best rate of approximation over $\mathcal{F}$.

In Chapter 5, we will see how one can formalize this statement.

Better rates? Are there classes of functions (other than trivial ones) that can be approximated in $O(N^{-r})$ for $r > 1/2$? In other words, if one is willing to restrict further the set of functions to be approximated, can we guarantee better rates of convergence?

Therefore, 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. This is one of the main goals of the present thesis.

1.3 Statistical Estimation

In a nonparametric regression problem, one is given a pair of random variables (X, Y) where, say, X is a d-dimensional vector and Y is real valued. Given data $(X_i, Y_i)_{i=1}^N$, and the model

$$Y_i = f(X_i) + \epsilon_i, \tag{1.5}$$

where ϵ is the noisy contribution, one wishes to estimate the unknown smooth function f.

It is observed that well-known regression methods such as kernel smoothing, nearest-neighbor, spline smoothing (see Härdle, 1990 for details) may perform very badly in high dimensions because of the so-called curse of dimensionality. The curse comes from the fact that when dealing with a finite amount of data, the high-dimensional ball Ω_d is mostly empty, as discussed in the excellent paper of Friedman and Stuetzle (1981). In terms of estimation bounds, roughly speaking, the curse says that unless you have an enormous sample size N, you will get a poor mean-squared error, say.

1.3.1 Projection Pursuit Regression (PPR)

In an attempt to avoid the adverse effects of the curse of dimensionality, Friedman and Stuetzle (1981) suggest approximating the unknown regression function f by a sum of ridge functions,

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

where the u_j's are vectors of unit length, i.e. $\|u_j\| = 1$. The algorithm, the statistical analogy of (1.2)-(1.3), also operates in a stepwise fashion. At stage m, it augments the fit f_{m-1} by adding a ridge function $g_j(u_j \cdot x)$ obtained as follows: calculate the residuals of the $m-1$th fit $r_i = Y_i - \sum_{j=1}^{m-1} g_j(u_j \cdot X_i)$; and for a fixed direction u plot the residuals r_i against $u \cdot x_i$; fit a smooth curve g and choose the best direction u, so as to minimize the residuals sum of squares $\sum_i (r_i - g(u \cdot X_i))^2$. The algorithm stops when the improvement is small.

The approach was revolutionary because instead of averaging the data over balls, PPR performs a local averaging over narrow strips: $|u \cdot x - t| \leq h$, thus avoiding the problems relative to the sparsity of the high-dimensional unit ball.

1.3.2 Neural Nets Again

Neural nets are also very much in use in statistics for regression, classification, discrimination, etc. (see the survey of Cheng and Titterington, 1994 and its joined discussion). In regression, where the training data is again of the form (X_i, Y_i), neural nets fit the data with a sum of the form

$$\hat{y}(x) = \sum_{j=1}^{m} \alpha_j \rho(k_j \cdot x - b_j),$$

where $k_j \in \mathbf{R}^d$ and $b_j \in \mathbf{R}$, so that the fit is exactly like (1.1). Again, the sigmoid is most commonly used for ρ.

Of course, PPR and neural nets regression are of the same flavor as both attempt to approximate the regression surface by a superposition of ridge functions. One of the main differences is perhaps that neural networks allow for a non-smooth fit since $\rho(k \cdot x - b)$ resembles a step function when the norm $\|k\|$ of the weights is large. On the other hand, PPR can make better use of projections since it bears the freedom to choose a different profile g at each step.

1.3.3 Statistical Methodology

In approximation theory, given a dictionary $\mathcal{D} = \{g_\lambda,\ \lambda \in \Lambda\}$ (where Λ denotes some index set), one tries to build up an approximation by taking out finite linear combinations

$$f_N(x) = \sum_{i=1}^{N} \alpha_i g_{\lambda_i}(x).$$

Likewise, in statistics, almost all current nonparametric regression methods use selection of elements from $\mathcal{D}$ to construct an estimate

$$\hat{f}(x) = \sum_{i=1}^{N} \alpha_i g_{\lambda_i}(x)$$

of the unknown f (1.5). Following Breiman's discussion (Cheng and Titterington, 1994), examples include cases where $\mathcal{D}$ is a set of indicator functions $\mathcal{D} = \{1_{\{x \in R_\lambda\}}\}$ where the R_λ's are rectangles (CART); the case where elements of $\mathcal{D}$ are products of univariate splines $\mathcal{D} = \{\prod_{j=1}^{d} \pm(x_j - \lambda_{i,j})_+\}$ (MARS); and many others including the neural nets dictionary $\mathcal{D}_{Ridge}$. One of the most remarkable and beautiful examples concerns the case where $\mathcal{D}$ is a wavelet basis, as in this case; both fast algorithms and near-optimal theoretical results are available, see Donoho, Johnstone, Kerkyacharian, and Picard (1995).

PPR and neural nets are used every day in data analysis, but not much is known about their capability. We feel that there is a need to get an intellectual understanding of these projection-based methods. What can neural networks achieve? For which kinds of regression surface f will they give good estimates? How can a good subset of neurons $\rho(k \cdot x - b)$ be selected? It is common sense that PPR or neural nets will have a small prediction error if, and only if, superpositions of ridge functions like (1.1) approximate the regression surface rather well. In fact, the connection between approximation theory and statistical estimation is very deep (see, for instance, Hasminskii and Ibragimov, 1990; Donoho and Johnstone, 1989; Donoho, 1993; Donoho and Johnstone, 1995) to the point that in some cases, the two problems become hardly distinguishable, as shown in Donoho (1993), for example. Therefore, a lot of questions are common with the ones spelled out in the previous section.

1.4 Harmonic Analysis

It is well known that trigonometric series provide poor reconstructions of singular signals. For instance, let $H(x)$ be the step function $1_{\{x>0\}}$ on the interval $[-\pi, \pi]$. The best L_2 N-term approximation of H by trigonometric series gives only a L_2 error of order $O(N^{-1/2})$. One of the many reasons that make wavelets so attractive is that they are the best bases for representing objects composed with singularities (see the discussion of Mallat's heuristics in Donoho, 1993). In a nice wavelet basis, the L_2 approximation error is $O(N^{-s})$ for every possible choice of s. However, under a certain viewpoint, the picture changes dramatically when the dimension is greater than one. In the unit Q of $\mathbf{R}^d$, say that we want to represent again the step function $H(u \cdot x - t)$; then, $O(\epsilon^{-2(d-1)})$ wavelets are needed to give a reconstruction error of order ϵ (i.e. convergence in $O(N^{-\frac{1}{2(d-1)}})$ of N-term expansions). Translated into the framework of image compression, it says that both wavelets bases and Fourier bases are severely inefficient at representing edges in images.

In harmonic analysis, there has recently been much interest in finding new dictionaries and ways of representing functions by linear combinations of elements of those. Examples include wavelets, wavelet-packets, Gabor functions, brushlets, etc. However, there aren't any representations that represent objects like $H(u \cdot x - t)$ efficiently. From this point of view, it would be interesting to develop one which would represent step functions as well as wavelets do in one dimension.

1.5 Achievements

The thesis is about the important issues that have just been addressed. Our goal here is to apply the concepts and methods of modern harmonic analysis to tackle these problems, starting with the primary one: the problem of constructing neural networks.

Using techniques developed in group representations theory and wavelet analysis, we develop two concrete and stable representations of functions f as superpositions of ridge functions. We then use these new expansions to study finite approximations.

1.5.1 A Continuous Representation

In Chapter 2, we develop the concept of *admissible neural activation function* $\psi : \mathbf{R} \to \mathbf{R}$. Unlike traditional sigmoidal neural activation functions which are positive and monotone

increasing, such an admissible activation function is oscillating, taking both positive and negative values. In fact, our condition requires for ψ a number of vanishing moments which are proportional to the dimension d, so that an admissible ψ has zero integral, zero 'average slope,' zero 'average curvature,' etc. in high dimensions.

We show that if one is willing to abandon the traditional sigmoidal neural activation function ρ, which typically has no vanishing moments and is not in L^2, and replace it by an admissible neural activation function ψ, then any reasonable function f may be represented exactly as a *continuous* superposition from the dictionary $\mathcal{D}_{Ridgelet} = \{\psi_\gamma : \gamma \in \Gamma\}$ of *ridgelets* $\psi_\gamma(x) = a^{-1/2}\psi(\frac{u\cdot x - b}{a})$ where the ridgelet parameter $\gamma = (a, u, b)$ runs through the set $\Gamma \equiv \{(a, u, b); \quad a, b \in \mathbf{R}, a > 0, u \in \mathbf{S}^{d-1}\}$ with $\mathbf{S}^{d-1}$ denoting the unit sphere of $\mathbf{R}^d$. In short, we establish a continuous reproducing formula

$$f = c_\psi \int \langle f, \psi_\gamma \rangle \psi_\gamma \mu(d\gamma), \tag{1.6}$$

for $f \in L^1 \cap L^2(\mathbf{R}^d)$, where c_ψ is a constant which depends only on ψ and $\mu(d\gamma) \propto da/a^{n+1}dudb$ is a kind of uniform measure on Γ; for details, see below. We also establish a Parseval relation

$$\|f\|^2 = c_\psi \int |\langle f, \psi_\gamma \rangle|^2 \mu(d\gamma). \tag{1.7}$$

These two formulas mean that we have a well-defined *continuous Ridgelet transform* $\mathcal{R}(f)(\gamma) = \langle f, \psi_\gamma \rangle$ taking functions on $\mathbf{R}^d$ isometrically into functions of the ridgelet parameter $\gamma = (a, u, b)$.

1.5.2 Discrete Representation

We next develop somewhat stronger admissibility conditions on ψ (which we call *frameability* conditions) and replace this continuous transform by a discrete transform (Chapter 3). Let D be a fixed compact set in $\mathbf{R}^d$. We construct a special countable set $\Gamma_d \subset \Gamma$ such that every $f \in L^2(D)$ has a representation

$$f = \sum_{\gamma \in \Gamma_d} \alpha_\gamma \psi_\gamma, \tag{1.8}$$

with equality in the $L^2(D)$ sense. This representation is stable in the sense that the coefficients change continuously under perturbations of f which are small in $L^2(D)$ norm. Underlying the construction of such a discrete transform is, of course, a quasi-Parseval relation, which in this case takes the form

$$A\|f\|^2_{L^2(D)} \leq \sum_{\gamma\in\Gamma_d} |\langle f, \psi_\gamma\rangle_{L^2(D)}|^2 \leq B\|f\|^2_{L^2(D)}; \tag{1.9}$$

Equation (1.8) follows by use of the standard machinery of frames (Duffin and Schaeffer, 1952; Daubechies, 1992). Frame machinery also shows that the coefficients α_γ are realizable as bounded linear functionals $\alpha_\gamma(f)$ having Riesz representers $\tilde{\psi}_\gamma(x) \in L^2(D)$. These representers are not ridge functions themselves; but by the convergence of Neumann series underlying the frame operator, we are entitled to think of them as *molecules* made up of linear combinations of ridge atoms, where the linear combinations concentrate on atoms with parameters γ' "near" γ.

1.5.3 Applications

As a result of Chapters 2 and 3, we are, roughly speaking, in a position to efficiently construct finite approximations by ridgelets which give good approximations to a given function $f \in L^2(D)$. One can see where the tools we have constructed are heading: from the exact series representation (1.8), one aims to extract a finite linear combination which is a good approximation to the infinite series; once such a representation is available, one has a stable, mathematically tractable method of constructing approximate representations of functions f based on systems of neuron-like elements.

New functional classes. Rephrasing a comment made in section 1.2, it is natural to ask for which functional classes do ridgelets make sense. That is, what are the classes they approximate best? To explain further what we mean, suppose we are given a dictionary $\mathcal{D} = \{g_\lambda,\ \lambda \in \Lambda\}$. For a function f, we define its approximation error by N-elements of the dictionary $\mathcal{D}$ by

$$\inf_{(\alpha_i)_{i=1}^N} \inf_{(\lambda_i)_{i=1}^N} \|f - \sum_{i=1}^N \alpha_i g_{\lambda_i}\|_H \equiv d_N(f, \mathcal{D}). \tag{1.10}$$

Suppose now that we are interested in the approximation of classes of functions; characterize the rate of approximation of the class $\mathcal{F}$ by N elements from $\mathcal{D}$ by

$$d_N(\mathcal{F}, \mathcal{D}) = \sup_{f \in \mathcal{F}} d_N(f, \mathcal{D}). \tag{1.11}$$

In Chapter 4, we introduce a new scale of functional classes, not currently studied in harmonic analysis, which are "quasi-approximation spaces" for ridgelets. That is, we show that (Chapter 5):

(i) *Optimality.* There is a dictionary of ridgelet-like elements, namely the dual-ridgelet dictionary $\mathcal{D}_{Dual-Ridge} = \{\tilde{\psi}_\gamma\}_{\gamma \in \Gamma_d}$, that is optimal for approximating functions from these classes. In other words, there isn't any other dictionary with better approximation properties in the sense of (1.11).

(ii) *Constructive approximation.* There is an approximation scheme that is optimal for approximating functions from these classes. From the exact series representation

$$f = \sum_{\gamma \in \Gamma_d} \langle f, \psi_\gamma \rangle \tilde{\psi}_\gamma,$$

extract the N-term approximation $\tilde{f}_N$ where one only keeps the dual-ridgelet terms corresponding to the N largest ridgelet coefficients $\langle f, \psi_\gamma \rangle$; then, the approximant $\tilde{f}_N$ achieves the optimal rate of approximation over our new classes.

In Chapter 4, we give a description of these new spaces in terms of the smoothness of the Radon-transform of f. Furthermore, we explain how these spaces model functions that are singular across hyperplanes when there may be an arbitrary number of hyperplanes which may be located in any spatial positions and may have any orientations.

Specific examples. We study degrees of approximations over some specific examples. For example, we will show in Chapter 5 that the goals set in section 1.4 are fulfilled. Although ridgelets are optimal for representing objects with singularities across hyperplanes, they fail to represent efficiently singular radial objects (Chapter 6); i.e., when singularities are associated with spheres and more generally with curved hypersurfaces. In some sense, we cannot curve the singular sets.

Superiority over traditional neural nets. In Neural Networks, one considers approximations by finite linear combinations taken from the dictionary $\mathcal{D}_{NN} = \{\rho(k \cdot x - b),\ k \in$

$\mathbf{R}^n, b \in \mathbf{R}\}$, where ρ is the univariate sigmoid, see Barron (1993) for example. It is shown that for any function $f \in L_2(\Omega_d)$, there is a ridgelet approximation which is at least as good - and perhaps much better - as the best ideal approximation using Neural Networks.

1.5.4 Innovations

Underlying our methods is the inspiration of modern harmonic analysis – ideas like the Calderón reproducing formula and the Theory of Frames. We shall briefly describe what is new here – that which is not merely an 'automatic' consequence of existing ideas.

First, there is, of course, a general machinery for getting continuous reproducing formulas like (1.6), via the theory of square-integrable group representations (Duflo and Moore, 1976; Daubechies, Grossmann, and Meyer, 1986). Such a theory has been applied to develop wavelet-like representations over groups other than the usual $ax+b$ group on $\mathbf{R}^d$, see Bernier and Taylor (1996). However, the particular geometry of ridge functions does not allow the identification of the action of Γ on ψ with a linear group representation (notice that the argument of ψ is real, while the argument of ψ_γ is a vector in $\mathbf{R}^d$). As a consequence, the possibility of a straightforward application of well-known results is ruled out. As an example of the difference, our condition for admissibility of a neural activation function for the continuous ridgelet transform is much stronger – requiring about $d/2$ vanishing moments in dimension d – than the usual condition for admissibility of the mother wavelet for the continuous wavelet transform, which requires only one vanishing moment in any dimension.

Second, in constructing frames of ridgelets, we have been guided by the theory of wavelets, which holds that one can turn continuous transforms into discrete expansions by adopting a strategy of discretizing frequency space into dyadic coronae (Daubechies, 1992; Daubechies, Grossmann, and Meyer, 1986); this goes back to Littlewood-Paley (Frazier, Jawerth, and Weiss, 1991). Our approach indeed uses such a strategy for dealing with the location and scale variables in the Γ_d dictionary. However, in dealing with ridgelets there is also an issue of discretizing the directional variable u that seems to be a new element: u must be discretized more finely as the scale becomes finer. The existence of frame bounds under our discretization shows that we have achieved, in some sense, the 'right' discretization, and we believe this to be new and of independent interest.

Third, as emphasized in the previous two paragraphs, one has available a new tool to analyze and synthesize multivariate functions. While wavelets and related methods

work well in the analysis and synthesis of objects with local singularities, ridgelets are designed to work well with conormal objects: objects that are singular across some family of hypersurfaces, but smooth along them. This leads to a more general and superficial observation: the association between neural nets representations and certain types of spatial inhomogeneities seems, here, to be a new element.

Next, there is a serious attempt in this thesis to characterize and identify functional classes that can be approximated by neural nets at a certain rate. Unlike well grounded area of approximation theory, neural network theory does not solve the delicate characterization issue. In wavelet or spline theory, it is well known that the efficiency of the approximation is characterized by classical smoothness (Besov spaces). In contrast, it is necessary in addressing characterization issues of neural nets approximation to abandon the classical measure of smoothness. Instead, we propose a new one and define a new scale of spaces based on our new definition. In addition to providing a characterization framework, these spaces to our knowledge are not studied in classical analysis and their study may be of independent interest.

We conclude this introduction by underlining perhaps the most important aspect of the present thesis: ridgelet expansion and approximation are both constructive and effective procedures as opposed to existential approximations commonly discussed in the neural networks literature (see section 1.1).

Chapter 2

The Continuous Ridgelet Transform

In this chapter we present results regarding the existence and the properties of the continuous representation (1.6). Recall that we have introduced the parameter space

$$\Gamma = \{\gamma = (a, u, b); \quad a, b \in \mathbf{R}, a > 0, u \in \mathbf{S}^{d-1}\},$$

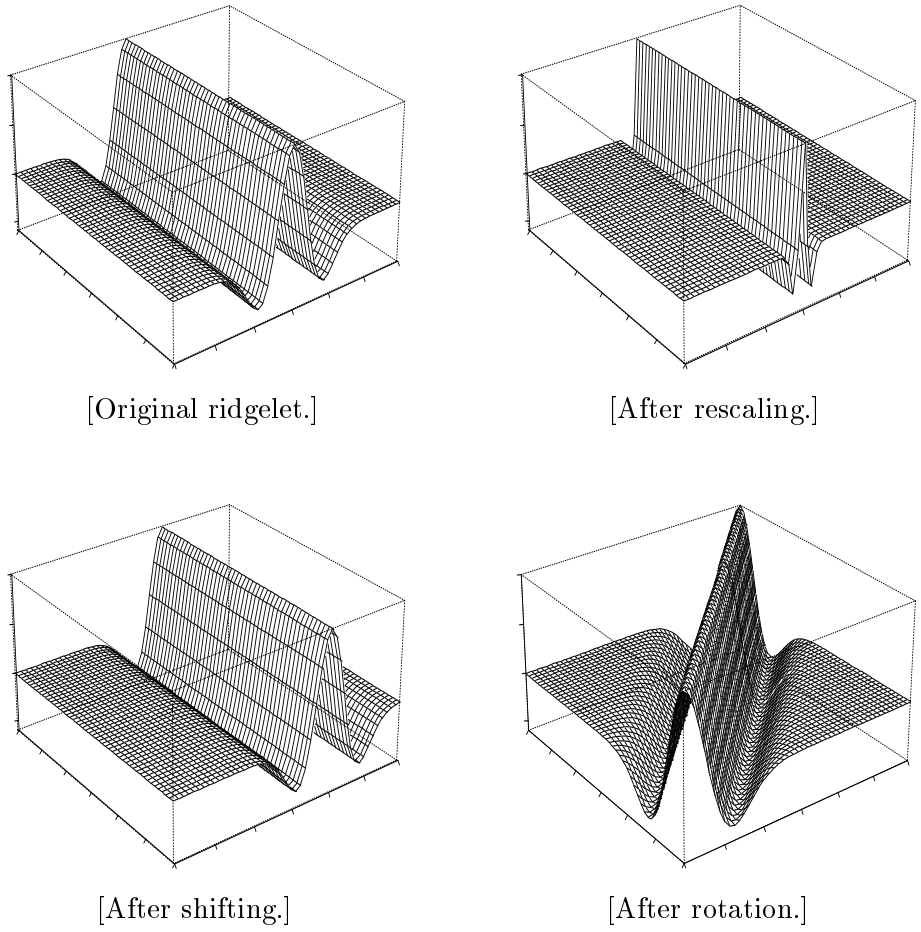
and the notation $\psi_\gamma(x) = a^{-1/2}\psi(\frac{u \cdot x - b}{a})$. Of course, the parameter $\gamma = (a, u, b)$ has a natural interpretation: a indexes the scale of the ridgelet; u, its orientation and b, its location. The measure $\mu(d\gamma)$ on neuron parameter space Γ is defined by $\mu(d\gamma) = \dfrac{da}{a^{d+1}}\, \sigma_d du\, db$, where σ_d is the surface area of the unit sphere $\mathbf{S}^{d-1}$ in dimension d and du the uniform probability measure on $\mathbf{S}^{d-1}$. As usual, $\widehat{f}(\xi) = \int e^{-ix\cdot\xi} f(x)dx$ denotes the Fourier transform of f and $\mathcal{F}(f)$ as well. To simplify notation, we will consider only the case of multivariate $x \in \mathbf{R}^d$ with $d \geq 2$. Finally, we will always assume that $\psi : \mathbf{R} \to \mathbf{R}$ belongs to the Schwartz space $\mathcal{S}(\mathbf{R})$. The results presented here hold under weaker conditions on ψ, but we avoid study of various technicalities in this chapter.

We now introduce the key definition of this chapter.

Definition 1 *Let $\psi : \mathbf{R} \to \mathbf{R}$ satisfy the condition*

$$K_\psi = \int \frac{|\widehat{\psi}(\xi)|^2}{|\xi|^d} d\xi < \infty. \tag{2.1}$$

Then ψ is called an **Admissible Neural Activation Function**.



[Original ridgelet.] [After rescaling.]

[After shifting.] [After rotation.]

Figure 2.1: Ridgelets.

We will call the ridge function ψ_γ generated by an admissible ψ a **ridgelet**.

2.1 A Reproducing Formula

We start by the fundamental reconstruction principle that will be extended to more general functions in the next section.

Theorem 1 (Reconstruction) *Suppose that f and $\widehat{f} \in L^1(\mathbf{R}^d)$. If ψ is admissible, then*

$$f = c_\psi \int \langle f, \psi_\gamma \rangle \psi_\gamma \mu(d\gamma), \tag{2.2}$$

where $c_\psi = \pi(2\pi)^{-d}K_\psi^{-1}$.

Remark 1. In fact, for $\psi \in \mathcal{S}(\mathbf{R})$, the admissibility condition (2.1) is essentially equivalent to the requirement of vanishing moments:

$$\int t^k \psi(t)dt = 0, \quad k \in \{0, 1, \dots, \left[\frac{d+1}{2}\right] - 1\}.$$

This clearly shows the similarity of (2.1) to the 1-dimensional wavelet admissibility condition (Daubechies, 1992, Page 24); however, unlike wavelet theory, the number of necessary vanishing moments grows linearly in the dimension d.

Remark 2. If $\rho(t)$ is the sigmoid function $e^t/(1+e^t)$, then ρ is *not admissible*. Actually no formula like (2.2) can hold if one uses neurons of the type commonly employed in the theory of Neural Networks. However, $\rho^{(m)}(t)$ is an admissible activation function for $m \geq [\frac{d}{2}] + 1$. Hence, sufficiently high derivatives of the functions used in Neural Networks theory do lead to good reconstruction formulas.

Proof of Theorem 1. The proof uses the Radon Transform R_u defined by: $R_u f(t) = \int f(tu + U^\perp s)ds$ with $s = (s_1, \dots, s_{d-1}) \in \mathbf{R}^{d-1}$ and $U^\perp$ an $d \times (d-1)$ matrix containing as columns an orthonormal basis for $u^\perp$.
With a slight abuse of notation, let $\psi_a(x) = a^{-\frac{1}{2}}\psi(\frac{x}{a})$ and $\widetilde{\psi}(x) = \psi(-x)$. Put $w_{a,u}(b) = \widetilde{\psi}_a * R_u f(b)$ and let $\mathrm{I} = \int \langle f, \psi_\gamma \rangle \psi_\gamma(x)\mu(d\gamma) = \int \psi_a(u \cdot x - b) w_{a,u}(b) \frac{da}{a^{d+1}}\, \sigma_d du\, db$. Recall $\widehat{R_u f} = \widehat{f}(\xi u)$ and, hence, if $\widehat{f} \in L^1(\mathbf{R}^d)$, $\widehat{R_u f} \in L^1(\mathbf{R})$. Then, $\mathrm{I} = \int \psi_a * (\widetilde{\psi}_a * R_u f)(u \cdot x) \frac{da}{a^{d+1}} \sigma_d du$. Noting that $\psi_a * (\widetilde{\psi}_a * R_u f) \in L^1(\mathbf{R})$ and that its 1-dimensional Fourier transform is given by $a|\widehat{\psi}(a\xi)|^2 \hat{f}(\xi u)$, we have

$$\mathrm{I} = \frac{1}{2\pi} \int \exp\{i\xi u \cdot x\} \widehat{f}(\xi u) a |\widehat{\psi}(a\xi)|^2 \frac{da}{a^{d+1}}\, \sigma_d du\, d\xi.$$

If ψ is real valued, $\overline{\widehat{\psi}(-\xi)} = \widehat{\psi}(\xi)$; hence,

$$\mathrm{I} = \frac{1}{\pi} \int \exp\{i\xi u \cdot x\} \widehat{f}(\xi u) a |\widehat{\psi}(a\xi)|^2 1_{\{\xi > 0\}} \frac{da}{a^{d+1}}\, \sigma_d du\, d\xi.$$

Then, by Fubini

$$\begin{aligned} \mathrm{I} &= \frac{1}{\pi}\int \exp\{i\xi u\cdot x\}\widehat{f}(\xi u)\left\{\int |\widehat{\psi}(a\xi)|^2\frac{da}{a^d}\right\}1_{\{\xi>0\}}d\xi\sigma_d du \\ &= \frac{1}{\pi}\int \exp\{i\xi u\cdot x\}\widehat{f}(\xi u)K_\psi|\xi|^{d-1}1_{\{\xi>0\}}d\xi\sigma_d du \\ &= \frac{1}{\pi}K_\psi\int_{\mathbf{R}^d}\exp\{ix\cdot k\}\widehat{f}(k)dk \\ &= \frac{1}{\pi}K_\psi(2\pi)^d f(x). \qquad \blacksquare \end{aligned}$$

Integral representations like (2.2) have been independently discovered in Murata (1996).

2.2 A Parseval Relation

Theorem 2 (Parseval relation) *Assume $f\in L^1\cap L^2(\mathbf{R}^d)$ and ψ admissible. Then*

$$\|f\|_2^2 = c_\psi\cdot\int|\langle f,\psi_\gamma\rangle|^2\mu(d\gamma).$$

Proof. With $w_{a,u}(b)$ defined as in the proof of Theorem 1, we then have

$$\int|\langle f,\psi_\gamma\rangle|^2\mu(d\gamma) = \int|w_{a,u}(b)|^2\frac{da}{a^{d+1}}\,\sigma_d du\,db = \mathrm{I},$$

say. Using Fubini's theorem for positive functions,

$$\int|w_{a,u}(b)|^2\frac{da}{a^{d+1}}\,\sigma_d du\,db = \int\|w_{a,u}\|_2^2\frac{da}{a^{d+1}}\,\sigma_d du. \tag{2.3}$$

$w_{a,u}$ is integrable, being the convolution between two integrable functions, and belongs to $L^2(\mathbf{R})$ since $\|w_{a,u}\|_2\le\|f\|_1\|\psi_a\|_2$; its Fourier transform is then well defined and $\widehat{w}_{a,u}(\xi) = \overline{\widehat{\psi}}_a(\xi)\widehat{f}(\xi u)$. By the usual Plancherel theorem, $\int|w_{a,u}(b)|^2db = \frac{1}{2\pi}\int|\widehat{w}_{a,u}(\xi)|^2d\xi$ and, hence,

$$\mathrm{I} = \frac{1}{2\pi}\int|\widehat{f}(\xi u)|^2|\widehat{\psi}_a(\xi)|^2\frac{da}{a^{d+1}}\,\sigma_d du\,d\xi = \frac{2}{2\pi}\int_{\{\xi>0\}}|\widehat{f}(\xi u)|^2|\widehat{\psi}(a\xi)|^2\frac{da}{a^d}\,\sigma_d du\,d\xi.$$

Since $\int |\widehat{\psi}(a\xi)|^2 \frac{da}{a^d} = K_\psi |\xi|^{d-1}$ (admissibility), we have

$$\mathrm{I} \;=\; \frac{2K_\psi}{2\pi} \int |\widehat{f}(\xi u)|^2 \xi^{d-1} d\xi du = \frac{1}{\pi} K_\psi (2\pi)^d \|f\|_2^2. \qquad \blacksquare$$

The assumptions on f in the above two theorems are somewhat restrictive, and the basic formulas can be extended to an even wider class of objects. It is classical to define the Fourier Transform first for $f \in L^1(\mathbf{R}^d)$ and only later to extend it to all of L^2 using the fact that $L^1 \cap L^2$ is dense in L^2. By a similar density argument, one obtains

Proposition 1 *There is a linear transform $\mathcal{R}$: $L^2(\mathbf{R}^d) \to L^2(\Gamma, \mu(d\gamma))$ which is an L^2-isometry and whose restriction to $L^1 \cap L^2$ satisfies*

$$\mathcal{R}(f)(\gamma) = \langle f, \psi_\gamma \rangle.$$

For this extension, a generalization of the Parseval relationship (1.7) holds.

Proposition 2 (Extended Parseval) *For all $f, g \in L^2(\mathbf{R}^d)$,*

$$\langle f, g \rangle = c_\psi \int \mathcal{R}(f)(\gamma) \mathcal{R}(g)(\gamma) \mu(d\gamma). \tag{2.4}$$

Proof of Proposition 2. Notice that one needs only to prove the property for a dense subspace of $L^2(\mathbf{R}^d)$; i.e., $L^1 \cap L^2(\mathbf{R}^d)$. So let $f, g \in L^1 \cap L^2$; we can write

$$\int \mathcal{R}(f)(\gamma) \mathcal{R}(g)(\gamma) \mu(d\gamma) = \int \langle \widetilde{\psi}_a * f, \widetilde{\psi}_a * g \rangle \frac{da}{a^{d+1}} \sigma_d du = \mathrm{I}.$$

Applying Plancherel

$$\begin{aligned} \mathrm{I} \;&=\; \frac{1}{2\pi} \int \langle \widehat{\widetilde{\psi}_a * f}, \widehat{\widetilde{\psi}_a * g} \rangle \frac{da}{a^{d+1}} \sigma_d du \\ &=\; \frac{1}{2\pi} \int \widehat{f}(\xi u) \widehat{g}(\xi u) a |\widehat{\psi}(a\xi)|^2 \frac{da}{a^{d+1}} \sigma_d du \, d\xi \end{aligned}$$

and, by Fubini, we get the desired result. $\blacksquare$

Relation (2.4) allows identification of the integral $c_\psi \int \langle f, \psi_\gamma \rangle \psi_\gamma \mu(d\gamma)$ with f by duality. In fact, taking the inner product of $c_\psi \int \langle f, \psi_\gamma \rangle \psi_\gamma \mu(d\gamma)$ with any $g \in L^2(\mathbf{R}^d)$ and exchanging

the order of inner product and integration over γ, one obtains

$$\langle c_\psi \left[\int \langle f, \psi_\gamma \rangle \psi_\gamma \mu(d\gamma) \right], g \rangle = c_\psi \int \langle f, \psi_\gamma \rangle \langle g, \psi_\gamma \rangle \mu(d\gamma) = \langle f, g \rangle$$

which by the Riesz theorem leads to $f \equiv c_\psi \int \langle f, \psi_\gamma \rangle \psi_\gamma \mu(d\gamma)$ in the prescribed weak sense.

The theory of wavelets and Fourier analysis contain results of a similar flavor: for example, the Fourier inversion theorem in $L^2(\mathbf{R}^d)$ can be proven by duality. However, there exists a more concrete proof of the Fourier inversion theorem. Recall, in fact, that if $f \in L^1 \cap L^2(\mathbf{R}^d)$ and if we consider the truncated Fourier expansion $\widehat{f}_K(\xi) = \widehat{f}(\xi) 1_{\{|\xi| \leq K\}}$, then $\widehat{f}_K \in L^1(\mathbf{R}^d)$ and $\|\overline{\mathcal{F}}(\widehat{f}_K) - (2\pi)^d f\|_{L^2} \to 0$ as $K \to \infty$. This argument provides an interpretation of the Fourier inversion formula that reassures about its practical relevance.

We now give a similar result for the convergence of truncated ridgelet expansions. For each $\varepsilon > 0$, define $\Gamma_\varepsilon := \{\gamma = (a, u, b) : \varepsilon \leq a \leq \varepsilon^{-1}, u \in \mathbf{S}^{d-1}, b \in \mathbf{R}\} \subset \Gamma$.

Proposition 3 *Let $f \in L^1(\mathbf{R}^d)$ and $\{\alpha_\gamma\} = \{\langle f, \psi_\gamma \rangle\}_{(\gamma \in \Gamma)}$; then for every $\varepsilon > 0$*

$$\alpha_\gamma 1_{\Gamma_\varepsilon}(\gamma) \in L^1(\Gamma, \mu(d\gamma)).$$

Proof. Notice that $\alpha_\gamma = (\widetilde{\psi}_a * R_u f)(b)$; then

$$\begin{aligned} \int_{\Gamma_\varepsilon} |\alpha_\gamma| \mu(d\gamma) &= \int |w_{a,u}(b)| \frac{da}{a^{d+1}} \sigma_d du\, db \\ &\leq \sigma_d \|f\|_1 \int_\varepsilon^{\varepsilon^{-1}} \|\psi\|_1 \frac{da}{a^{d+\frac{1}{2}}} < \infty, \end{aligned}$$

where we have used $\|w_{a,u}\|_1 \leq \|\widetilde{\psi}_\alpha\|_1 \|f\|_1 = a^{1/2} \|\psi\|_1 1 \|f\|_1$. ∎

The above proposition shows that for any $f \in L^1(\mathbf{R}^d)$, the expression

$$f_\varepsilon \equiv c_\psi \int_{\Gamma_\varepsilon} \langle f, \psi_\gamma \rangle \psi_\gamma \mu(d\gamma)$$

is meaningful, since $\{\psi_\gamma\}_{\gamma \in \Gamma}$ is uniformly L^∞ bounded over Γ_ε. The next theorem makes more precise the meaning of the reproducing formula.

Theorem 3 *Suppose $f \in L^1 \cap L^2(\mathbf{R}^d)$ and ψ admissible.*

(1) $f_\varepsilon \in L^2(\mathbf{R}^d)$, and

(2) $\|f - f_\varepsilon\|_2 \to 0$ *as* $\varepsilon \to 0$.

Proof of Theorem 3.

Step 1 Letting $\phi_\lambda(x) = (\frac{1}{2\pi\lambda})^{\frac{d}{2}} \exp\{-\frac{\|x\|^2}{2\lambda}\}$ and defining f_ε^λ as

$$f_\varepsilon^\lambda = c_\psi \int_{\Gamma_\varepsilon} \langle f * \phi_\lambda, \psi_\gamma \rangle \psi_\gamma \mu(d\gamma),$$

we start proving that $f_\varepsilon^\lambda \in L^2(\mathbf{R}^d)$. Notice that $R_u(f * \phi_\lambda) = R_u f * R_u \phi_\lambda$ and $R_u\phi_\lambda(t) = \frac{1}{(2\pi\lambda)^{1/2}} \exp\{-\frac{t^2}{2\lambda}\}$. Now $\mathcal{F}(R_u f * R_u\phi_\lambda)(\xi) = (\widehat{R_u f} \cdot \widehat{R_u\phi_\lambda})(\xi) = \widehat{f}(\xi u) \exp\{-\frac{\lambda}{2}\xi^2\}$. Repeating the argument in the proof of Theorem 1, we get

$$f_\varepsilon^\lambda = \frac{c}{\pi} \int_{\{\xi>0\},\mathbf{S}^{d-1}} \left\{ \int_{\varepsilon \le a \le \varepsilon^{-1}} \frac{da}{a^d} |\widehat{\psi}(a\xi)|^2 \right\} \exp\{i\xi u \cdot x - \frac{\lambda}{2}\xi^2\} \widehat{f}(\xi u) \sigma_d d\xi du.$$

Note that for $\xi \neq 0$, we have $\int_\varepsilon^{\varepsilon^{-1}} |\widehat{\psi}(a\xi)|^2 \frac{da}{a^d} = |\xi|^{d-1} \int_{\varepsilon|\xi|}^{\varepsilon^{-1}|\xi|} |\widehat{\psi}(t)|^2 \frac{dt}{t^d}$ (which we will abbreviate as $K_\psi |\xi|^{d-1} c_\varepsilon(|\xi|)$) and $c_\varepsilon(|\xi|) \uparrow 1$ as $\varepsilon \to 0$. After the change of variable $k = |\xi|u$, we obtain

$$f_\varepsilon^\lambda = \frac{c_\psi}{\pi} K_\psi \int \exp\{ik \cdot x - \frac{\lambda\|k\|^2}{2}\} c_\varepsilon(\|k\|) \widehat{f}(k) dk,$$

which allows the interpretation of f_ε^λ as the "conjugate" Fourier transform of an L^2 element and therefore the conclusion $f_\varepsilon^\lambda \in L^2(\mathbf{R}^d)$.

Step 2 We aim to prove that $f_\varepsilon^\lambda \to f_\varepsilon$ pointwise and in $L^2(\mathbf{R}^d)$. The dominated convergence theorem leads to

$$c_\varepsilon(\|k\|)\widehat{f}(k) \exp\{-\frac{\lambda}{2}\|k\|^2\} \longrightarrow c_\varepsilon(\|k\|)\widehat{f}(k) \quad \text{in} \quad L^2(\mathbf{R}^d) \quad \text{as} \quad \lambda \to 0.$$

Then by the Fourier Transform isometry, we have $f_\varepsilon^\lambda \to (2\pi)^{-d}\overline{FT}(c_\varepsilon \widehat{f})$ in $L^2(\mathbf{R}^d)$. It remains to be proved that this limit, which we will abbreviate with g_ε, is indeed f_ε:

$$\begin{aligned}
|f_\varepsilon^\lambda(x) - f_\varepsilon(x)| &= c_\psi \int_{\Gamma_\varepsilon} (\langle f * \phi_\lambda, \psi_\gamma\rangle - \langle f, \psi_\gamma\rangle)\psi_\gamma \mu(d\gamma) \\
&\leq c_\psi \sup_{\gamma\in\Gamma_\varepsilon} |\psi_\gamma(x)| \int_\varepsilon^{\varepsilon^{-1}} \int_{\mathbf{S}^{d-1}} \|\widetilde{\psi}_a * (R_u f * R_u\phi_\lambda - R_u f)\|_1 \frac{da}{a^{d+1}}\sigma_d du \\
&\leq c_\psi \varepsilon^{-\frac{1}{2}}\|\psi\|_\infty \int_\varepsilon^{\varepsilon^{-1}} \int_{\mathbf{S}^{d-1}} \|\widetilde{\psi}_a\|_1 \|R_u f * R_u\phi_\lambda - R_u f\|_1 \frac{da}{a^{d+1}}\sigma_d du \\
&= c_\psi \varepsilon^{-\frac{1}{2}}\|\psi\|_\infty \int_\varepsilon^{\varepsilon^{-1}} \frac{da}{a^{d+\frac{1}{2}}} \|\psi\|_1 \int_{\mathbf{S}^{d-1}} \|R_u f * R_u\phi_\lambda - R_u f\|_1 \sigma_d du.
\end{aligned}$$

Then for a fixed u, $\|R_u f * R_u\phi_\lambda - R_u f\|_1 \to 0$ as $\lambda \to 0$ and

$$\begin{aligned}
\|R_u f * R_u\phi_\lambda - R_u f\|_1 &\leq \|R_u f\|_1 + \|R_u f * R_u\phi_\lambda\|_1 \\
&\leq 2\|R_u f\|_1 \leq 2\|f\|_1.
\end{aligned}$$

Thus by the dominated convergence theorem, $\int_{\mathbf{S}^{d-1}} \|R_u f * R_u\phi_\lambda - R_u f\|_1 \sigma_d du \to 0$.
From $|f_\varepsilon^\lambda(x) - f_\varepsilon(x)| \leq \delta(\varepsilon)\|\psi\|_\infty \psi\|_1 \int_{\mathbf{S}^{d-1}} \|R_u f * R_u\phi_\lambda - R_u f\|_1 \sigma_d du$, we obtain $\|f_\varepsilon^\lambda - f_\varepsilon\|_\infty \to 0$ as $\lambda \to 0$. Note that the convergence is in $C(\mathbf{R}^d)$ as the functions are continuous. Finally, we get $f_\varepsilon = g_\varepsilon$ and, therefore, f_ε is in $L^2(\mathbf{R}^d)$ by completeness.
To show that $\|f_\varepsilon - f\|_2 \to 0$ as $\varepsilon \to 0$, it is necessary and sufficient to show that $\|\widehat{f_\varepsilon} - \widehat{f}\|_2 \to 0$,

$$\|\widehat{f_\varepsilon} - \widehat{f}\|_2^2 = \int |\widehat{f}(k)|^2 (1 - c_\varepsilon(\|k\|)^2 dk.$$

Recalling that $0 \leq c_\varepsilon \leq 1$ and that $c_\varepsilon \uparrow 1$ as $\varepsilon \to 0$, the convergence follows. ∎

2.3 A Semi-Continuous Reproducing Formula

We have seen that any function $f \in L_1 \cap L_2(\mathbf{R}^d)$ might be represented as a continuous superposition of ridge functions

$$f = c_\psi \int \langle f(x), \frac{1}{a}\psi(\frac{u\cdot x - b}{a})\rangle \frac{1}{a}\psi(\frac{u\cdot x - b}{a})\frac{da}{a^d} du db, \tag{2.5}$$

and the sense in which the above equation holds. Now, one can obtain a semi-continuous version of (2.5) by replacing the continuous scale by a dyadic lattice. The motivation for

doing so will appear in the later chapters. Let us choose ψ such that

$$\sum_{j\in\mathbf{Z}} \frac{|\hat{\psi}(2^{-j}\xi)|^2}{|2^{-j}\xi|^{d-1}} = 1. \tag{2.6}$$

Of course, this condition greatly resembles the admissibility condition (2.1) introduced earlier. If one is given a function Ψ such that

$$\sum_{j\in\mathbf{Z}} |\hat{\Psi}(2^{-j}\xi)|^2 = 1,$$

it is immediate to see that ψ defined by $\hat{\psi}(\xi) = |\xi|^{(d-1)/2}\hat{\Psi}(\xi)$ will verify (2.6). Now, using the same argument as for Theorems 1 and 3, the property (2.6) implies

$$f \propto \sum_{j\in\mathbf{Z}} 2^{j(d-1)} \int \langle f(x), 2^j\psi(2^j(u\cdot x - b))\rangle 2^j\psi(2^j(u\cdot x - b))dudb,$$

where again if $f \in \mathcal{S}(\mathbf{R}^d)$, the inequality holds in a pointwise way and more generally if $f \in L_1 \cap L_2(\mathbf{R}^d)$, the partial sums of the right-hand side are square integrable and converge to f in L_2. Finally, as in wavelet theory, it will be rather useful to introduce some special coarse scale ridgelets. We choose a profile φ so that

$$|\hat{\varphi}(\xi)|^2 = \sum_{j<0} 2^{j(d-1)}|\hat{\psi}(2^{-j}\xi)|^2.$$

As a consequence, we have that for any $\xi \in \mathbf{R}$

$$|\hat{\varphi}(\xi)|^2 + \sum_{j\geq 0} 2^{j(d-1)}|\hat{\psi}(2^{-j}\xi)|^2 = |\xi|^{d-1}. \tag{2.7}$$

Notice, 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. (In fact, φ is "almost" an Admissible Neural Activation Function: compare with (2.1)).

For a pair (φ, ψ) satisfying (2.7), we have the following semi-continuous reproducing

formula

$$f \propto \int \langle f(x), \varphi(u \cdot x - b)\rangle \varphi(u \cdot x - b) + \sum_{j \geq 0} 2^{j(d-1)} \int \langle f(x), \psi_j(u \cdot x - b)\rangle \psi_j(u \cdot x - b) du db, \tag{2.8}$$

where as in Littlewood Paley theory, ψ_j stands for $2^j \psi(2^j \cdot)$. At this point, the reader knows in which sense (2.8) must be interpreted.

Chapter 3

Discrete Ridgelet Transforms: Frames

The previous chapter described a class of neurons, the ridgelets $\{\psi_\gamma\}_{\gamma\in\Gamma}$, such that

(i) any function f can be reconstructed from the continuous collection of its coefficients $\langle f, \psi_\gamma\rangle$, and

(ii) any function can be decomposed in a continuous superposition of neurons ψ_γ.

The purpose of this chapter is to achieve similar properties using only a discrete set of neurons $\Gamma_d \subset \Gamma$.

3.1 Generalities about Frames

The theory of frames (Daubechies, 1992; Young, 1980) deals precisely with questions of this kind. In fact, if $\mathcal{H}$ is a Hilbert space and $\{\varphi_n\}_{n\in N}$ a frame, an element $f \in \mathcal{H}$ is completely characterized by its coefficients $\{\langle f, \varphi_n\rangle\}_{n\in N}$ and can be reconstructed from them via a simple and numerically stable algorithm. In addition, the theory provides an algorithm to express f as a linear combination of the frame elements φ_n.

Definition 2 *Let $\mathcal{H}$ be a Hilbert space and let $\{\varphi_n\}_{n\in N}$ be a sequence of elements of $\mathcal{H}$. Then $\{\varphi_n\}_{n\in N}$ is a* frame *if there exist $0 < A,\ B < \infty$ such that for any $f \in \mathcal{H}$*

$$A\|f\|_{\mathcal{H}}^2 \leq \sum_{n\in N} |\langle f, \varphi_n\rangle_{\mathcal{H}}|^2 \leq B\|f\|_{\mathcal{H}}^2 \tag{3.1}$$

in which case A and B are called frame bounds.

Let $\mathcal{H}$ be a Hilbert space and $\{\varphi_n\}_{n\in N}$ a frame with bounds A and B. Note that $A\|f\|^2_{\mathcal{H}} \leq \sum |\langle f, \varphi_n\rangle|^2$ implies that $\{\varphi_n\}_{n\in N}$ is a complete set in $\mathcal{H}$. A frame $\{\varphi_n\}_{n\in N}$ is said to be tight if we can take $A = B$ in Definition 1. Furthermore, if $\{\varphi_n\}_{n\in N}$ is a basis for $\mathcal{H}$, it is called a Riesz basis. Simple examples of Frames include Orthonormal Bases, Riesz Bases, finite concatenations of several Riesz Bases,etc.

The following results are stated without proofs and can be found in Daubechies (1992, Page 56) and Young (1980, Page 184). Define the coefficient operator F: $\mathcal{H} \to l^2(N)$ by $F(f) = (\langle f, \varphi_n\rangle)_{n\in N}$. Suppose that F is a bounded operator ($\|Ff\| \leq B\|f\|_{\mathcal{H}}$). Let F^* be the adjoint of F and let $G = F^*F$ be the *Frame Operator*; then $A\,\mathrm{Id} \leq G \leq B\,\mathrm{Id}$ in the sense of orders of positive definite operators. Hence, G is invertible and its inverse G^{-1} satisfies $B^{-1}\mathrm{Id} \leq G^{-1} \leq A^{-1}\mathrm{Id}$. Define $\widetilde{\varphi}_n = G^{-1}\varphi_n$; then $\{\widetilde{\varphi}_n\}_{n\in N}$ is also a frame (with frames bounds B^{-1} and A^{-1}) and the following holds:

$$f = \sum_{n\in N} \langle f, \widetilde{\varphi}_n\rangle_{\mathcal{H}} \varphi_n = \sum_{n\in N} \langle f, \varphi_n\rangle_{\mathcal{H}} \widetilde{\varphi}_n. \tag{3.2}$$

Moreover, if $f = \sum_{n\in N} a_n\varphi_n$ is an another decomposition of f, then $\sum_{n\in N} |\langle f, \widetilde{\varphi}_n\rangle|^2 \leq \sum_{n\in N} |a_n|^2$. To rephrase Daubechies, the frame coefficients are the most economical in an L^2 sense. Finally, $G = \frac{A+B}{2}(I - R)$ where $\|R\| < 1$, and so G^{-1} can be computed as $G^{-1} = \frac{2}{A+B}\sum_{k=0}^{\infty} R^k$.

3.2 Discretization of Γ

The special geometry of ridgelets imposes differences between the organization of ridgelet coefficients and the organization of traditional wavelet coefficients.

With a slight change of notation, we recall that $\psi_\gamma = a^{1/2}\psi(a(u\cdot x - b))$. We are looking for a countable set Γ_d and some conditions on ψ such that the quasi-Parseval relation (1.9) holds. Let $\mathcal{R}(f)(\gamma) = \langle f, \psi_\gamma\rangle$; then $\mathcal{R}(f)(\gamma) = \langle R_u f, \psi_{a,b}\rangle$ with $\psi_{a,b}(t) = a^{1/2}\psi(a(t-b))$. Thus, the information provided by a ridgelet coefficient $\mathcal{R}(f)(\gamma)$ is the one-dimensional wavelet coefficient of $R_u f$, the Radon transform of f. Applying Plancherel, $\mathcal{R}(f)(\gamma)$ may

be expressed as

$$\mathcal{R}(f)(\gamma) = \frac{1}{2\pi}\langle \widehat{R_u f}, \widehat{\psi}_{a,b}\rangle = \frac{a^{-1/2}}{2\pi}\int \widehat{f}(\xi u)\widehat{\psi}(\xi/a)\exp\{ib\xi\}d\xi, \tag{3.3}$$

which corresponds to a one-dimensional integral in the frequency domain (see Figure 1).

In fact, it is the line integral of $\widehat{f}\widehat{\psi}_{a,0}$, modulated by $\exp\{ib\xi\}$, along the line $\{tu : t \in \mathbf{R}\}$. If, as in the Littlewood-Paley theory (Frazier, Jawerth, and Weiss, 1991) $a = 2^j$ and $\mathrm{supp}(\psi) \subset [1/2, 2]$, it emphasizes a certain dyadic segment $\{t : 2^j \le t \le 2^{j+1}\}$. In contrast, in the multidimensional wavelets case where the wavelet $\psi_{a,b} = a^{-\frac{d}{2}}\psi(\frac{x-b}{a})$ with $a > 0$ and $b \in \mathbf{R}^d$, the analogous inner product $\langle f, \psi_{a,b}\rangle$ corresponds to the average of $\widehat{f}\widehat{\psi}_a$ over the whole frequency domain, emphasizing the dyadic corona $\{\xi : 2^j \le |\xi| \le 2^{j+1}\}$.

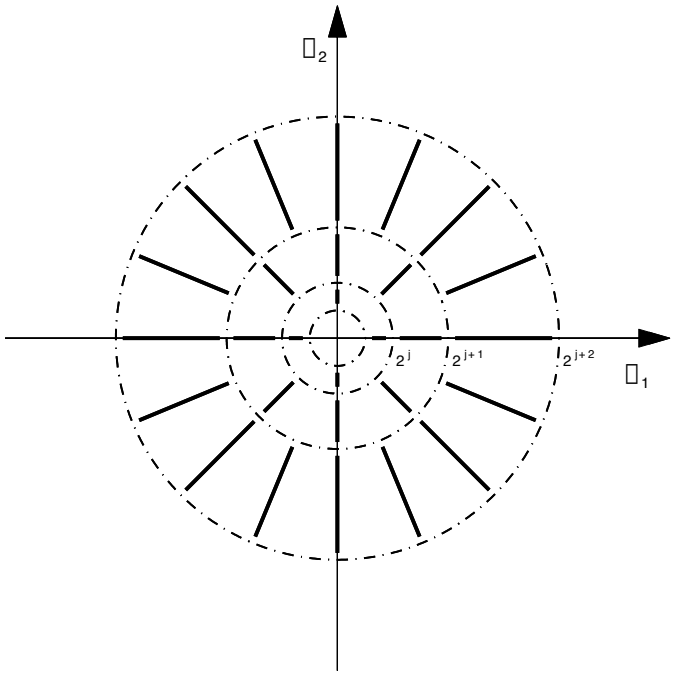


Figure 3.1: Diagram schematically illustrating the ridgelet discretization of the frequency plane (2-dimensional case). The circles represent the scales 2^j (we have chosen $a_0 = 2$) and the different segments essentially correspond to the support of different coefficient functionals. There are more segments at finer scales.

Now, the underlying object $\hat{f}$ must certainly satisfy specific smoothness conditions in order for its integrals on dyadic segments to make sense. Equivalently, in the original domain

f must decay sufficiently rapidly at ∞. In this chapter, we take for our decay condition that f be compactly supported, so that $\hat{f}$ is band limited. From now on, we will only consider functions supported on the unit cube $Q = \{x \in \mathbf{R}^d, \|x\|_\infty \leq 1\}$ with $\|x\|_\infty = \max_i |x_i|$; thus $\mathcal{H} = L^2(Q)$.

Guided by the Littlewood-Paley theory, we choose to discretize the scale parameter a as $\{a_0^j\}_{j \geq j_0}$ ($a_0 > 1$, j_0 being the coarsest scale) and the location parameter b as $\{kb_0 a_0^{-j}\}_{k,j \geq j_0}$. Our discretization of the sphere will also depend on the scale: the finer the scale, the finer the sampling over $\mathbf{S}^{d-1}$. At scale a_0^j, our discretization of the sphere, denoted Σ_j, is an ε_j-net of $\mathbf{S}^{d-1}$ with $\varepsilon_j = \epsilon_0 a_0^{-(j-j_0)}$ for some $\epsilon_0 > 0$. We assume that for any $j \geq j_0$, the sets Σ_j satisfy the following *Equidistribution Property*: two constants $k_d, K_d > 0$ must exist s.t. for any $u \in \mathbf{S}^{d-1}$ and r such that $\epsilon_j \leq r \leq 2$

$$k_d \left(\frac{r}{\varepsilon_j}\right)^{d-1} \leq \quad |\{B_u(r) \cap \Sigma_j\}| \quad \leq \quad K_d \left(\frac{r}{\varepsilon_j}\right)^{d-1}. \tag{3.4}$$

On the other hand, if $r \leq \epsilon_j$, then from $B_u(r) \subset B_u(\epsilon_j)$ and the above display, $|\{B_u(r) \cap \Sigma_j\}| \leq K_d$. Furthermore, the number of points N_j satisfies $k_d \left(\frac{2}{\varepsilon_j}\right)^{d-1} \leq N_j \leq K_d \left(\frac{2}{\varepsilon_j}\right)^{d-1}$. Essentially, our condition guarantees that Σ_j is a collection of N_j almost equispaced points on the sphere $\mathbf{S}^{d-1}$, N_j being of order $a_0^{(j-j_0)(d-1)}$. The discrete collection of ridgelets is then given by

$$\psi_\gamma(x) = a_0^{j/2} \psi(a_0^j u \cdot x - kb_0), \quad \gamma \in \Gamma_d = \{(a_0^j, u, kb_0 a_0^j),\ j \geq j_0, u \in \Sigma_j,\ k \in \mathbf{Z}\}. \tag{3.5}$$

In our construction, the coarsest scale is determined by the dimension of the space $\mathbf{R}^d$. Defining ℓ as $\sup\{\frac{\pi}{2k},\ k \in N \text{ and } \frac{\pi}{2k} < \frac{log2}{2d}\}$, we choose j_0 s.t. $a_0^{j_0+1} \leq \ell < a_0^{j_0+2}$. Finally, we will set $\epsilon_0 = 1/2$ so that $\epsilon_j = a_0^{-(j-j_0)}/2$.

Remark. Here, we want to be as general as possible and that is the reason why we do not restrict the choice of a_0. However, in Littlewood Paley or wavelet theory, a standard choice corresponds to $a_0 = 2$ (dyadic frames). Likewise, and although we will prove that there are frames for any choice of a_0, we will always take $a_0 = 2$ in the analysis we develop in the forthcoming chapters.

3.3 Main Result

We now introduce a condition that allows us to construct frames.

Definition 3 *The function ψ is called* **frameable** *if $\psi \in C^1(\mathbf{R})$ and*

- $\inf_{1\le|\xi|\le a_0} \sum_{j\ge 0} \left|\widehat{\psi}(a_0^{-j}\xi)\right|^2 \left|a_0^{-j}\xi\right|^{-(d-1)} > 0$
- $|\widehat{\psi}(\xi)| \le C|\xi|^{\alpha}(1+|\xi|)^{-\gamma}$ *where* $\alpha > \frac{d-1}{2}, \gamma > 2+\alpha$.

This type of condition bears a resemblance to conditions in the theory of wavelet frames (compare, for example, Daubechies, 1992, Page 55.) In addition, this condition looks like a discrete version of the admissible neural activation condition described in the previous section.

There are many frameable ψ. For example, sufficiently high derivatives (larger than $d/2+1$) of the sigmoid are frameable.

Theorem 4 (Existence of Frames) *Let ψ be frameable. Then there exists $b_0^* > 0$ so that for any $b_0 < b_0^*$, we can find two constants $A, B > 0$ (depending on ψ, a_0, b_0 and d) so that, for any $f \in L^2(Q)$ (where Q denotes the unit cube of $\mathbf{R}^d$),*

$$A\|f\|_2^2 \le \sum_{\gamma\in\Gamma_d} |\langle f, \psi_\gamma\rangle|^2 \le B\|f\|_2^2. \tag{3.6}$$

The theorem is proved in several steps. We first show:

Lemma 1

$$\Big|\sum_{\gamma\in\Gamma_d} |\langle f,\psi_\gamma\rangle|^2 - \frac{1}{2\pi b_0}\int_{\mathbf{R}} \sum_{j\ge j_0, u\in\Sigma_j} |\hat{f}(\xi u)|^2 |\hat{\psi}(a_0^{-j}\xi)|^2 d\xi\Big|$$
$$\le \frac{1}{2\pi}\sqrt{\int_{\mathbf{R}} \sum_{j\ge j_0,u\in\Sigma_j} |\hat{f}(\xi u)|^2|\hat{\psi}(a_0^{-j}\xi)|^2 d\xi|}\sqrt{\int_{\mathbf{R}} \sum_{j\ge j_0,u\in\Sigma_j} |\hat{f}(\xi u)|^2 |a_0^{-j}\xi|^2 |\hat{\psi}(a_0^{-j}\xi)|^2 d\xi} \tag{3.7}$$

The argument is a simple application of the analytic principle of the large sieve (Montgomery, 1978). Note that it presents an alternative to Daubechies' proof of one-dimensional dyadic affine frames (Daubechies, 1992). We first recall an elementary lemma that we state without proof.

Lemma 2 *Let g be a real valued function in $C^1[0,\delta]$ for some $\delta > 0$: then,*

$$|g(\delta/2) - \frac{1}{\delta}\int_0^\delta g(x)dx| \le \frac{1}{2}\int_0^\delta |g'(x)|dx.$$

Again, let $\psi_j(x)$ be $a_0^{j/2}\psi(a_0^j x)$. The ridgelet coefficient is then $\langle f, \psi_\gamma\rangle = (R_u f * \psi_j)(kb_0a_0^{-j})$. For simplicity we denote $F_j = |R_u f * \psi_j|^2$. Applying the lemma gives

$$\left|F_j(kb_0a_0^{-j}) - \frac{a_0^j}{b_0}\int_{(k-1/2)b_0a_0^{-j}}^{(k+1/2)b_0a_0^{-j}} F_j(b)db\right| \le \frac{1}{2}\int_{(k-1/2)b_0a_0^{-j}}^{(k+1/2)b_0a_0^{-j}} |F_j'(b)|db.$$

Now, we sum over k:

$$\left|\sum_k |(R_u f * \psi_j)(kb_0a_0^{-j})|^2 - \frac{a_0^j}{b_0}\int_{\mathbf{R}} |(R_u f * \psi_j)(b)db\right|$$
$$\le \int_{\mathbf{R}} |(R_u f * \psi_j)(b)|\ |(R_u f * (\psi_j)')(b)|db \le \|R_u f * \psi_j\|_2 \|(R_u f * (\psi_j)')\|_2.$$

Applying Plancherel, we have

$$\left|\sum_k |(R_u f * \psi_j)(kb_0a_0^{-j})|^2 - \frac{1}{2\pi b_0}\int_{\mathbf{R}} |\hat{f}(\xi u)|^2|\hat{\psi}(a_0^{-j}\xi)|^2 d\xi\right|$$
$$\le \frac{1}{2\pi}\sqrt{\int_{\mathbf{R}} |\hat{f}(\xi u)|^2|\hat{\psi}(a_0^{-j}\xi)|^2 d\xi}\sqrt{\int_{\mathbf{R}} |\hat{f}(\xi u)|^2|a_0^{-j}\xi|^2|\hat{\psi}(a_0^{-j}\xi)|^2 d\xi}.$$

Hence, if we sum the above expression over $u \in \Sigma_j$ and j and apply the Cauchy-Schwartz inequality to the right-hand side, we get the desired result. ∎

We then show that there exist $A',\ B' > 0$ s.t. for any $f \in L^2(Q)$; we have

$$A'\|\widehat{f}\|_2^2 \le \sum_{j\ge j_0, u\in\Sigma_j}\int_{-\infty}^{\infty}\left|\widehat{f}(\xi u)\right|^2\left|\widehat{\psi}(a_0^{-j}\xi)\right|^2 d\xi \quad\le\quad B'\|\widehat{f}\|_2^2\ ; \tag{3.8}$$

$$\sum_{j\ge j_0, u\in\Sigma_j}\int_{-\infty}^{\infty}\left|\widehat{f}(\xi u)\right|^2\left|a_0^{-j}\xi\right|^2\left|\widehat{\psi}(a_0^{-j}\xi)\right|^2 d\xi \quad\le\quad B'\|\widehat{f}\|_2^2\ . \tag{3.9}$$

Thus, if b_0 is chosen small enough, Theorem 4 holds.

3.4 Irregular Sampling Theorems

Relationship (3.8) is, in fact, a special case of a more abstract result which holds for general multivariate entire functions of exponential type. An excellent presentation of entire functions may be found in Boas (1952). In the present section, $B_1^2(\mathbf{R}^d)$ denotes the set of square integrable functions whose Fourier Transform is supported in $[-1,1]^d$ and $Q_a(d) = \{x, \|x-a\|_\infty \leq \ell\}$, the cube of center a and volume $(2\ell)^d$. Finally, let $\{z_m\}_{m\in\mathbf{Z}^d}$ be the grid on $\mathbf{R}^d$ defined by $z_m = 2\ell m$.

Theorem 5 *Suppose $F \in B_1^2(\mathbf{R}^d)$ and $\ell < \frac{log2}{d}$ with $\frac{\pi}{2\ell}$ an integer; then $\forall a \in \mathbf{R}^d$,*

$$\sum_{m\in\mathbf{Z}^d} \min_{Q_{a+z_m}(\ell)} |F(x)|^2 \geq c_\ell^2 \sum_{m\in\mathbf{Z}^d} \max_{Q_{a+z_m}(\ell)} |F(x)|^2, \tag{3.10}$$

where c_ℓ can be chosen equal to $2e^{-\ell d} - 1$.

In fact, a more general version of this result holds for any exponent $p > 0$. (In this case, the constants ℓ and c_ℓ will depend on p). The requirement that $\pi/2\ell$ must be an integer simplifies the proof but this assumption may be dropped.

Proof of Theorem 5. First, note that by making use of $F_a(x) = F(x-a)$, we just need to prove the result for $a = 0$. The proof is then based on the lemma stated below, which is an extension to the multivariate case of a theorem of Paley and Wiener on non-harmonic Fourier series (Young, 1980, Page 38). Then with $|F(\lambda_m^-)| = \min_{Q_{z_m}(\ell)} |F(x)|$ (resp. $|F(\lambda_m^+)| = \max_{Q_{z_m}(\ell)} |F(x)|$), we have (using Lemma 3)

$$\sum_{m\in\mathbf{Z}^d} |F(\lambda_m^-)|^2 \geq (1/2\ell)^d (1-\rho_\ell)^2 \|F\|_2^2 \geq \left(\frac{1-\rho_\ell}{1+\rho_\ell}\right)^2 \sum_{m\in\mathbf{Z}^d} |F(\lambda_m^+)|^2.$$

And $\frac{1-\rho_\ell}{1+\rho_\ell} = 2e^{-\ell d} - 1$.

Lemma 3 *Let $F \in B_1^2(\mathbf{R}^d)$ and $\{\lambda_m\}_{m\in\mathbf{Z}^d}$ be a sequence of $\mathbf{R}^d$ such that $\sup_{m\in\mathbf{Z}^d} \|\lambda_m - m\pi\|_\infty < \frac{log2}{d}$; then*

$$(1-\rho_\ell)^2 \pi^{-d} \|F\|_2^2 \leq \sum_{m\in\mathbf{Z}^d} |F(\lambda_m)|^2 \leq (1+\rho_\ell)^2 \pi^{-d} \|F\|_2^2, \tag{3.11}$$

for $\rho_\ell = e^{\ell d} - 1 < 1$.

Proof of Lemma 3. The Polya-Plancherel theorem (see Plancherel and Pólya, 1938, Page 116) gives

$$\sum_{m\in\mathbf{Z}^d} |F(m\pi)|^2 = \pi^{-d}\|F\|_2^2.$$

Let k denote the usual multi-index $(k_1,\ldots,k_d)$ and let $|k| = k_1+\cdots+k_d$, $k! = k_1!\ldots k_d!$ and $x^k = x_1^{k_1}\ldots x_d^{k_d}$. For any k, $\partial^k F$ is an entire function of type π. Moreover, Bernstein's inequality gives $\|\partial^k F\|_2 \le \|F\|_2$; see Boas (1952, Page 211) for a proof. Since F is an entire function of exponential type, F is equal to its absolutely convergent Taylor expansion. Letting s be a constant to be specified below, we have

$$\begin{aligned} F(\lambda_m) - F(m\pi) &= \sum_{|k|\ge 1} \frac{\partial^k F(m\pi)}{k!}(\lambda_m - m)^k \\ &= \sum_{|k|\ge 1} \frac{\partial^k F(m\pi)}{k!}(\lambda_m - m)^k \frac{s^{|k|}}{s^{|k|}}. \end{aligned}$$

Applying Cauchy-Schwarz and summing over m, we get

$$\begin{aligned} \sum_{m\in\mathbf{Z}^d} |F(\lambda_m) - F(m\pi)|^2 &\le \sum_{m\in\mathbf{Z}^d}\sum_{|k|\ge 1} \frac{|\partial^k F(m\pi)|^2}{k! s^{2|k|}} \sum_{|k|\ge 1} \frac{\|\lambda_m - m\|_\infty^{2|k|} s^{2|k|}}{k!} \\ &\le \sum_{|k|\ge 1} \frac{\pi^{-d}\|F\|_2^2}{k! s^{2|k|}} \sum_{|k|\ge 1} \frac{\ell^{2|k|} s^{2|k|}}{k!} \\ &= \pi^{-d}\|F\|_2^2 (e^{d\frac{1}{s^2}} - 1)(e^{d\ell^2 s^2} - 1). \end{aligned}$$

We choose $s^2 = \frac{1}{\ell}$. If $\rho_\ell = e^{\ell d} - 1 < 1$; then

$$\sum_{m\in\mathbf{Z}^d} |F(\lambda_m) - F(m\pi)|^2 \le \rho_\ell^2 \pi^{-d}\|F\|_2^2$$

and, by the triangle inequality, the expected result follows.

Let μ be a measure on $\mathbf{R}^d$; μ will be called *d-uniform* if there exist α, $\beta > 0$ such that $\alpha \le \mu(Q_{z_m}(\ell))/(2\ell)^d \le \beta$. The following result is completely equivalent to the previous theorem.

Corollary 1 *Fix $\ell < \frac{log 2}{d}$ with $\frac{\pi}{2\ell}$ an integer. Let $F \in B_1^2(\mathbf{R}^d)$ and μ be an d-uniform*

measure with bounds α, β. Then

$$\alpha c_\ell \|F\|_2^2 \leq \int |F|^2 d\mu \leq \frac{\beta}{c_\ell} \|F\|_2^2. \tag{3.12}$$

3.5 Proof of the Main Result

We notice that the frameability condition implies that

(i) $\sup_{1\leq|\xi|\leq a_0} \sum_{j\in\mathbf{Z}} \frac{\left|\widehat{\psi}(a_0^j\xi)\right|^2}{\left|a_0^j\xi\right|^{d-1}} < \infty$, and

(ii) $\sup_{1\leq|\xi|\leq a_0} \sum_{j\geq 0} \left|\widehat{\psi}(a_0^j\xi)\right|^2 < \infty$,

and respectively (i′) and (ii′) where $\widehat{\psi}(\xi)$ is replaced by $\xi\widehat{\psi}(\xi)$.
For any measurable set A, let μ_ψ be the measure defined as

$$\mu_\psi(A) = \sum_{j\geq j_0, u\in\Sigma_j} \int \left|\widehat{\psi}(a_0^{-j}\xi)\right|^2 1_A(\xi u) d\xi.$$

And similarly, we can define μ'_ψ by changing $\widehat{\psi}(\xi)$ into $\xi\widehat{\psi}(\xi)$. Then,

$$\sum_{j\geq j_0, u\in\Sigma_j} \int \left|\widehat{f}(\xi u)\right|^2 \left|\widehat{\psi}(a_0^{-j}\xi)\right|^2 d\xi = \int \left|\widehat{f}\right|^2 d\mu_\psi$$

and likewise for μ'_ψ.

Proposition 4 *If ψ is frameable, μ_ψ and μ'_ψ are d-uniform and therefore there exist $A', B' > 0$ s.t. (3.8)-(3.9) hold.*

We only give proof for the measure μ_ψ, the proof for μ'_ψ being entirely parallel. Let ρu be the standard polar form of x. In this section, we will denote by $\Delta_x(r,\delta)$ the sets defined by $\Delta_x(r,\delta) = \{y = \rho' u', 0 \leq \rho' - \rho \leq r, \|u' - u\| \leq \delta\}$. These sets are truncated cones. The proof uses the technical Lemma 4.

Lemma 4 *For ψ frameable,*

$$0 < \inf_{\|x\|\geq\ell} \mu_\psi\left(\Delta_x(\ell, \frac{\ell}{2\|x\|})\right) \leq \sup_{\|x\|\geq\ell} \mu_\psi\left(\Delta_x(\ell, \frac{\ell}{2\|x\|})\right) < \infty,$$

and respectively for μ'_ψ.

Proof. To simplify the notations, we will use ρ for $\|x\|$ and u for $x/\|x\|$. Let j_x be defined by $a_0^{-(j_x-j_0)} \leq \ell/\rho < a_0 a_0^{-(j_x-j_0)}$. Hence, if $j \geq j_x$, $\forall \epsilon \in \{-1,1\}$, the *Equidistribution Property* (3.4) implies that

$$k_d\Big(\frac{a_0^{(j-j_0)}\ell}{\rho}\Big)^{d-1} \leq |\{B_{\varepsilon u}(\ell/2\rho) \cap \Sigma_j\}| \leq K_d\Big(\frac{a_0^{(j-j_0)}\ell}{\rho}\Big)^{d-1}.$$

We have

$$\begin{aligned}
\mu_\psi(\Delta_x(\ell,\ell/2\rho)) &= \sum_{j\geq j_0 \leq, u\in\Sigma_j} \int \Big|\widehat{\psi}(a_0^{-j}\xi)\Big|^2 1_{\Delta_x(\ell,\ell/2\rho)}(\xi u)d\xi \\
&\geq \sum_{j\geq j_x} k_d\Big(\frac{a_0^{(j-j_0)}\ell}{\rho}\Big)^{d-1} \int_{\rho\leq|\xi|\leq\rho+\ell} \Big|\widehat{\psi}(a_0^{-j}\xi)\Big|^2 d\xi \\
&\geq k_d(a_0^{-j_0}\ \ell)^{d-1} \int_{\rho\leq|\xi|\leq\rho+\ell} \Big(\frac{|\xi|}{\rho}\Big)^{d-1} \sum_{j'\geq 0} \frac{|\widehat{\psi}(a_0^{-j'}a_0^{-j_x}\xi)|^2}{|a_0^{-j'}a_0^{-j_x}\xi|^{d-1}} d\xi.
\end{aligned}$$

Now, since by assumption, $\ell \leq \rho$, we have $\forall\ |\xi| \in [\rho,\rho+\ell], \ell a_0^{-(j_0+1)} \leq |a_0^{-j_x}\xi| \leq 2\ell a_0^{-j_0}$. We recall that $\ell a_0^{-(j_0+1)} \geq 1$. Therefore,

$$\begin{aligned}
\mu_\psi(\Delta_x(\ell,\ell/2\rho)) &\geq k_d(a_0^{-j_0}\ \ell)^{d-1}\ 2\ell \inf_{\ell a_0^{-(j_0+1)}\leq|\xi|\leq 2\ell a_0^{-j_0}} \sum_{j'\geq 0} \frac{|\widehat{\psi}(a_0^{-j'}\xi)|^2}{|a_0^{-j'}\xi|^{d-1}} \\
&\geq k_d(a_0^{-j_0}\ \ell)^{d-1}\ 2\ell \inf_{1\leq|\xi|\leq a_0} \sum_{j'\geq 0} \frac{|\widehat{\psi}(a_0^{-j'}\xi)|^2}{|a_0^{-j'}\xi|^{d-1}}.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\sum_{j\geq j_x, u\in\Sigma_j} \int \Big|\widehat{\psi}(a_0^{-j}\xi)\Big|^2 1_{\Delta_x(\ell,\ell/2\rho)}(\xi u)d\xi \\
&\quad\leq K_d(a_0^{-j_0}\ \ell)^{d-1} 2^{d-1} 2\ell \sup_{\ell a_0^{-(j_0+1)}\leq|\xi|\leq 2\ell a_0^{-j_0}} \sum_{j'\geq 0} \frac{|\widehat{\psi}(a_0^{-j'}\xi)|^2}{|a_0^{-j'}\xi|^{d-1}} \\
&\quad\leq K_d(a_0^{-j_0}\ \ell)^{d-1} 2^{d-1} 2\ell \sup_{1\leq|\xi|\leq a_0} \sum_{j'\in\mathbf{Z}} \frac{|\widehat{\psi}(a_0^{-j'}\xi)|^2}{|a_0^{-j'}\xi|^{d-1}}.
\end{aligned}$$

We finally consider the case of the j's s.t. $j_0 \leq j < j_x$. We recall that in this case, we have $|\{B_{\varepsilon u}(\ell/2\rho) \cap \Sigma_j\}| \leq K_d$, and thus

$$\begin{aligned}
\sum_{j_0 \leq j < j_x, u \in \Sigma_j} \int \left|\widehat{\psi}(a_0^{-j}\xi)\right|^2 1_{\Delta_x(\ell,\ell/2\rho)}(\xi u) d\xi &\leq K_d \int_{\rho \leq |\xi| \leq \rho+\ell} \sum_{j_0 \leq j < j_x} \left|\widehat{\psi}(a_0^{j_x-j} a_0^{-j_x}\xi)\right|^2 \\
&\leq K_d 2\ell \sup_{\ell a_0^{-(j_0+1)} \leq |\xi| \leq 2\ell a_0^{-j_0}} \sum_{j'>0} \left|\widehat{\psi}(a_0^{j'}\xi)\right|^2 \\
&\leq K_d 2\ell \sup_{1 \leq |\xi| \leq a_0} \sum_{j'>0} \left|\widehat{\psi}(a_0^{j'}\xi)\right|^2 .
\end{aligned}$$

The lemma follows. ∎

Proof of Proposition 4. Now, we recall that $\{z_m\}_{m \in \mathbf{Z}^d}$ is the grid on $\mathbf{R}^d$ defined by $z_m = 2\ell m$ and we show that $\sup_m \mu_\psi(Q_{z_m}(\ell)) < \infty$ and that $\inf_m \mu_\psi(Q_{z_m}(\ell)) > 0$.
Again, we shall use the polar coordinates i.e. $z_m = \rho_m u_m$. For $m \neq 0$, let z'_m be $\rho'_m u_m$ with $\rho'_m = \rho_m - \ell/2$. Then, we have that $\Delta_{z'_m}(\ell, \ell/2\rho'_m) = \{\rho' u' \text{ s.t } |\rho' - \rho_m| \leq \ell/2,\ \|u' - u_m\| \leq \ell/2\rho'_m\} \subset B_{z_m}(\ell) \subset Q_{z_m}(\ell)$. To see the first inclusion, we can check that $\|\rho' u' - \rho_m u_m\|^2 = (\rho' - \rho_m)^2 + \rho'\rho_m \|u' - u_m\|^2$. Then we use the fact that $\rho'/\rho'_m \leq 5/3$ and $\rho_m/\rho'_m \leq 4/3$ to prove the inclusion.

For $m \neq 0$, let $\{x_j^{(m)}\}_{1 \leq j \leq J_m}$ with $\|x_j^{(m)}\| \geq \ell$ s.t $Q_{z_m}(\ell) \subset \cup_{1 \leq j \leq J_m} \Delta_{x_j^{(m)}}(\ell, \ell/2\|x_j^{(m)}\|)$ and $T_{d,m}$ be the minimum number of j's such that the above inclusion is satisfied. By rescaling, we see that the numbers $T_{d,m}$ are independent of ℓ. Moreover, it is easy to check that if δ is chosen small enough, then any set $\Delta_x(\ell, \ell/2\|x\|)$ (where again $\|x\| \geq \ell$) contains a ball of radius δ. (Although we don't prove it here, δ maybe chosen equal to $\ell/2$.) Therefore, the numbers $T_{d,m}$ are bounded above and we let $T_d = \sup_{m \neq 0} T_{d,m}$. It follows that for all $m \neq 0$, $(m \in \mathbf{Z}^d)$ we have

$$\begin{aligned}
0 < \inf_{\|x\| \geq \ell} \mu_\psi \left(\Delta_x(d, \frac{\ell}{2\|x\|})\right) &\leq \mu_\psi \left(\Delta_{z'_m}(\ell, \ell/2\rho'_m)\right) \\
&\leq \mu_\psi \left(Q_{z_m}(\ell)\right) \leq T_n \sup_{\|x\| \geq \ell} \mu_\psi \left(\Delta_x(\ell, \frac{\ell}{2\|x\|})\right) < \infty.
\end{aligned}$$

Finally, we need to prove the result for the cube $Q_0(\ell)$. In order to do so, we need to establish two last estimates:

$$\begin{aligned}
\mu_\psi\left(B_0(\ell)\right) &= \sum_{j\geq j_0} |\Sigma_j| \int_{\{|\xi|\leq \ell\}} \left|\widehat{\psi}(a_0^{-j}\xi)\right|^2 d\xi \\
&\geq k_d a_0^{(j-j_0)(d-1)} \int_{\{|\xi|\leq \ell\}} \sum_{j\geq j_0} \left|\widehat{\psi}(a_0^{-j}\xi)\right|^2 d\xi \\
&= k_d \int_{\{|\xi|\leq \ell\}} |a_0^{-j_0}\xi|^{d-1} \sum_{j'\geq 0} \frac{|\widehat{\psi}(a_0^{-j'}a_0^{-j_0}\xi)|^2}{|a_0^{-j'}a_0^{-j_0}\xi|^{d-1}} d\xi \\
&\geq k_d \int_{\{\ell/a_0\leq|\xi|\leq \ell\}} |a_0^{-j_0}\xi|^{d-1} \sum_{j'\geq 0} \frac{|\widehat{\psi}(a_0^{-j'}a_0^{-j_0}\xi)|^2}{|a_0^{-j'}a_0^{-j_0}\xi|^{d-1}} d\xi \\
&\geq k_d\, 2\ell(1-1/a_0)\, (\ell a_0^{-(j_0+1)})^{d-1} \inf_{\ell a_0^{-(j_0+1)}\leq|\xi|\leq \ell a_0^{-j_0}} \sum_{j'\geq 0} \frac{|\widehat{\psi}(a_0^{-j'}a_0^{-j_0}\xi)|^2}{|a_0^{-j'}a_0^{-j_0}\xi|^{d-1}}.
\end{aligned}$$

Repeating the argument of Lemma 4 finally gives

$$\mu_\psi\left(B_0(\ell)\right) \geq k_d\, 2\ell(1-1/a_0)\, (\ell a_0^{-(j_0+1)})^{d-1} \inf_{1\leq|\xi|\leq a_0} \sum_{j'\geq 0} \frac{|\widehat{\psi}(a_0^{-j'}\xi)|^2}{|a_0^{-j'}\xi|^{d-1}}.$$

After similar calculations, we can prove that

$$\mu_\psi\left(B_0(\ell)\right) \leq K_d 2\ell(\ell a_0^{-j_0})^{d-1} \sup_{\ell a_0^{-(j_0+1)}\leq|\xi|\leq \ell a_0^{-j_0}} \sum_{j'\geq 0} \frac{|\widehat{\psi}(a_0^{-j'}\xi)|^2}{|a_0^{-j'}\xi|^{d-1}}.$$

Again, let $\{x_j\}_{1\leq j\leq J}$ with $\|x_j\|\geq \ell$ s.t $Q_0(\ell)\subset \cup_{1\leq j\leq J}\Delta_{x_j}(\ell,\ell/2\|x_j\|)\cup B_0(\ell)$ and T_d^0 be the minimum number of j's needed. We then have

$$0<\mu_\psi\left(B_0(\ell)\right)\leq \mu_\psi\left(Q_0(\ell)\right)\leq \mu_\psi\left(B_0(\ell)\right)+T_n^0 \sup_{\|x\|\geq\ell} \mu_\psi\left(\Delta_x(\ell,\frac{\ell}{2\|x\|})\right)<\infty.$$

This completes the proof of Proposition 4. ∎

Although we do not prove it here, we may replace the frameability condition by one slightly weaker. For any traditional one-dimensional wavelet φ which satisfies the sufficient conditions listed in Daubechies (1992, Pages 68-69), define ψ via $\widehat{\psi}(\xi)\equiv \mathrm{sgn}(\xi)|\xi|^{\frac{d-1}{2}}(1+\xi^2)^{-\frac{d-1}{4}}\widehat{\varphi}(\xi)$; then Theorem 4 holds for such a ψ.

3.6 Discussion

3.6.1 Coarse Scale Refinements

In Neural Networks, the goal is to synthesize or represent a function as a superposition of neurons from the dictionary $\mathcal{D}_{Ridge} = \{\rho(k \cdot x - b); k \in \mathbf{R}^d, b \in \mathbf{R}\}$, the activation function ρ being fixed. That is, all the elements of $\mathcal{D}_{Ridge}$ have the same profile ρ. Likewise, as we wanted to keep this property, there is a unique profile ψ for all the elements of our ridgelet frame. However, it will be rather useful to introduce a different profile φ for the coarse-scale elements. For instance, following section 2.3, let us consider a function φ satisfying the following assumptions:

- $\hat{\varphi}(\xi)/|\xi|^{(d-1)/2} = O(1)$ and $|\hat{\varphi}(\xi)|/|\xi|^{(d-1)/2} \geq c$ if $|\xi| \leq 1$,
- $\hat{\varphi}(\xi) = O((1+|\xi|)^{-2})$.

Clearly, for a frameable ψ , the collection

$$\{\varphi(u_i \cdot x - kb_0),\ 2^{j/2}\psi(2^j u_i^j \cdot x - kb_0),\ j \geq 0, u_i^j \in \Sigma_j, k \in \mathbf{Z}\}, \tag{3.13}$$

(where again Σ_j is a set of "quasi-equidistant" points on the sphere, the resolution being $\theta_0 2^{-j}$) is a frame for $L_2(Q)$. The advantage of this description over the other (3.5) is the fact that the coarse scale corresponds to $j = 0$ (and not upon some funny index j_0 which depends on the dimension). In our applications, we shall generally use (3.13) for its greater comfort. As we will see, in addition to the frameability condition, we often require φ and ψ to have some regularity and ψ to have a few vanishing moments.

We close this section by introducing a few notations that we will use throughout the rest of the text. Indeed, it will be helpful to use the notation ψ_γ for $\varphi(u_i \cdot x - kb_0)$. We will make this abuse possible in saying that $\varphi(u_i \cdot x - kb_0)$ corresponds to the scale $j = -1$. For $j \geq -1$ then, denote also by Γ_j the index set for the jth scale,

$$\Gamma_j = \{(j, u_i^j, k),\ \ u_i^j \in \Sigma_j, k \in \mathbf{Z}\}. \tag{3.14}$$

(Note, finally, that $\gamma \in \Gamma_{-1}$, $\psi_\gamma(x)$ is in fact $\varphi(u_i \cdot x - kb_0)$.)

3.6.2 Quantitative Improvements

Our goal in this chapter has been merely to provide a qualitative result concerning the existence of frames of ridgelets. However, quantitative refinements will undoubtedly be important for practical applications.

The frame bounds ratio. The coefficients a_γ in a frame expansion may be computed via a Neumann series expansion for the frame operator; see Daubechies (1992). For computational purposes, the closer the ratio of the upper and lower frame bounds to 1, the fewer terms will be needed in the Neumann series to compute a dual element within an accuracy of ϵ. Thus for computational purposes, it may be desirable to have good control of the frame bounds ratio. Of course, the proof presented in section 3.5 provides only crude estimates for the upper bound of the frame bounds ratio. The interest of this method is that it uses general ideas, stated in section 3.4, which may be applied in a variety of different settings. The author is confident that further detailed studies will allow proof of versions of Theorem 4 with tighter bounds. Although, such refinements are beyond the scope of the present study, we next present a result that supports our certitude.

Actually, a reasonably simple calculation shows that the frame bounds ratio can be made arbitrarily small in dimension 2. In this case, let us consider a 2-dimensional frame,

$$\{\varphi(x_1\cos\theta_i + x_2\sin\theta_i - kb_0),\ 2^{j/2}\psi(2^j(x_1\cos\theta_i^j + x_2\sin\theta_i^j - kb_0), \\ j\geq 0, \theta_i^j = 2\pi\theta_0 i 2^{-j}, k\in\mathbf{Z}\}, \quad (3.15)$$

where at scale j the θ_i^j's are equispaced points on the circle, with step $2\pi\theta_0 2^{-j}$ (take θ_0^{-1} to be an integer).

Proposition 5 *Let (ψ_γ) be the frame defined by (3.15) and suppose that the pair (φ,ψ) satisfies (2.7), say. Then,*

$$\Big|\sum_\gamma |\langle f,\psi_\gamma\rangle|^2 - \frac{1}{2\pi b_0\theta_0}\|\hat f\|_2^2\Big| \leq C\,(\theta_0^{-1}+b_0^{-1})\|\hat f\|_2^2, \quad (3.16)$$

where the constant C depends at most upon φ and ψ. It follows from (3.16) that the frame bounds ratio is bounded by $\frac{C}{2\pi}(\theta_0+b_0)$ for the same constant C.

The result links the decay of the frame bounds ratio to the oversampling factor $(\theta_0 b_0)^{-1}$. The proof of the proposition may be found in the Appendix.

The oversampling. The redundancy of the frame that one can construct by this strategy depends heavily on the quality of the underlying "quasi-uniform" sampling of the sphere at each scale j. The construction of quasi-uniform discrete point sets on spheres has received considerable attention in the literature; see Conway and Sloane (1988) and additional references given in the bibliography. Quantitative improvements of our results would follow from applying some of the known results obtained in that field.

Computations. Another area for investigation has to do with rapid calculation of groups of coefficients. Note that if the sets Σ_j for $j \geq j_0$ present some symmetries, it may not be necessary to compute $\widetilde{\psi}_\gamma$ for all $\gamma \in \Gamma_d$; many dual elements would simply be translations, rotations and rescalings of each other. This type of relationship would be important to pursue for practical applications.

3.6.3 Sobolev Frames

Actually, there is a trivial extension of Theorem 4 to the case of the Sobolev spaces H^s. So, let us consider the space $H_0^s(Q)$ for $s \geq 0$. Recall that

$$\|f\|_{H_0^s}^2 = \int_{\mathbf{R}^2} \|k\|_2^{2s} |\hat{f}(k)|^2 \, dk$$

is an equivalent norm on $H_0^s(Q)$.

To fix the ideas, we will consider a frame of the type (3.13) as discussed in section 3.6.1.

Theorem 6 *Suppose ψ is frameable and that both b_0 and θ_0 are small enough so that $\{\psi_\gamma\}$ is a frame for $L_2(Q)$. In addition, assume that $\hat{\psi}(\xi)/|\xi|^s$ satisfies the conditions of Definition 3. Then, we can find two constants $A_s, B_s > 0$ so that, for any $f \in H_0^s(Q)$*

$$A\|f\|_{H_0^s}^2 \leq \sum_{\gamma \in \Gamma_d} 2^{2js} \, |\langle f, \psi_\gamma \rangle|^2 \leq B\|f\|_{H_0^s}^2. \tag{3.17}$$

Proof of Theorem. In fact applying Lemma 1, one immediately sees that

$$\Big| \sum_{\gamma \in \Gamma_d} 2^{2js} |\langle f, \psi_\gamma \rangle|^2 - \frac{1}{2\pi b_0} \int_{\mathbf{R}} \sum_{j \geq 0, u \in \Sigma_j} |\xi|^{2s} |\hat{f}(\xi u)|^2 \frac{|\hat{\psi}(2^{-j}\xi)|^2}{|2^{-j}\xi|^{2s}} d\xi \Big|$$

$$\leq \frac{1}{2\pi} \sqrt{\int_{\mathbf{R}} \sum_{j \geq 0, u \in \Sigma_j} |\xi|^{2s} |\hat{f}(\xi u)|^2 \frac{|\hat{\psi}(2^{-j}\xi)|^2}{|2^{-j}\xi|^{2s}} d\xi |} \sqrt{\int_{\mathbf{R}} \sum_{j \geq 0, u \in \Sigma_j} |\xi|^{2s} |\hat{f}(\xi u)|^2 |2^{-j}\xi|^2 \frac{|\hat{\psi}(2^{-j}\xi)|^2}{|2^{-j}\xi|^{2s}} d\xi}.$$

Taking the coarse scale into account amounts to add (in the previous display) $|\hat{f}(\xi u)|^2|\hat{\varphi}(\xi)|^2$ to the quantity $\sum_{j\geq 0, u\in\Sigma_j} |\xi|^{2s}|\hat{f}(\xi u)|^2|2^{-j}\xi|^2 \frac{|\hat{\psi}(2^{-j}\xi)|^2}{|2^{-j}\xi|^{2s}}$ at each of its occurence. Now, the rest of the proof is absolutely identical to the one of Theorem 4. Our irregular sampling applies and gives the desired result provided that $\hat{\psi}(\xi)/|\xi|^s$ satisfies the conditions of Definition 3. ∎

Remark. Anticipating the next chapter, suppose for instance that φ and ψ satisfy the conditions listed at the very beginning of Chapter 4, section 4.1; then our constructed L_2 frame (Theorem 4) is also a frame for every H_0^s with $s \geq 0$.

3.6.4 Finite Approximations

The frame dictionary $\mathcal{D}_{\Gamma_d} = \{\psi_\gamma, \gamma \in \Gamma_d\}$ may be used for constructing approximations of certain kinds of multivariate functions. It would be interesting to know the "Approximation Space" associated to this frame; that is, the collection of multivariate functions f obeying

$$\|f - f_N\|_2 \leq CN^{-r}, \tag{3.18}$$

where f_N is an appropriately chosen superposition of dictionary elements

$$f_N = \sum_{i=1}^{N} \lambda_{i,N}\psi_{\gamma_{i,N}}. \tag{3.19}$$

Based on obvious analogies with the orthogonal basis case, one naturally expects that functions f of this type can be characterized by their frame coefficients, saying (3.18) is possible if, and only if, the frame coefficients $\{\alpha_\gamma\}_{\gamma\in\Gamma_d}$ belong to the Lorentz weak l^p space $l_{p,\infty}$, with $r = (1/p - 1/2)_+$. Work to establish those conditions under which the above would hold is in progress.

It would also be interesting to establish results saying that (3.18) is equivalent to a weak l^p condition on the frame coefficients even when the approximant (3.19) is not restricted to use only $\gamma \in \Gamma_d$. If one could establish that any continuous choices $\gamma_{i,N} \in \Gamma$ would still only lead to f with weak-l^p conditions on frame coefficients, then one would know that the frame system is really an effective way of obtaining high-quality nonlinear approximations.

Chapter 4

Ridgelet Spaces

Littlewood-Paley theory is at the heart of the study of a wide range of classical functional spaces (Stein, 1970; Frazier, Jawerth, and Weiss, 1991). In this chapter, we do not use the classical d-dimensional Littlewood-Paley analysis to study the regularity of an object f, but rather use its one-dimensional version to study the regularity of the Radon transform Rf. Doing so, we introduce a new scale of functional spaces that we believe are not studied in classical analysis. Most of the analysis in this chapter serves to provide an external characterization of these spaces together with an intuitive description.

As we will see in the next chapter, these new spaces are the ones that make sense for neural nets.

4.1 New Spaces

We start out this section by listing a few classical definitions/conditions and will assume that they hold throughout the remainder of the chapter.

(i) $\varphi, \psi \in \mathcal{S}(\mathbf{R})$.

(ii) $\text{supp}\ \hat{\varphi} \subset \{|\xi| \leq 1\}$.

(iii) $\text{supp}\ \hat{\psi} \subset \{1/2 \leq |\xi| \leq 2\}$.

(iv) $\hat{\varphi}(\xi)/|\xi|^{(d-1)/2} = O(1)$ and $|\hat{\varphi}(\xi)|/|\xi|^{(d-1)/2} \geq c$ if $|\xi| \leq 5/6$,
$|\hat{\psi}(\xi)| \geq c$ if $3/5 \leq |\xi| \leq 5/3$.

These conditions are standard in Littlewood-Paley theory except for the first part of (iv). Instead of requiring $|\hat{\varphi}|$ to be 1 in the neighborhood of the origin, we want $|\hat{\varphi}|/|\xi|^{(d-1)/2}$ to be 1 near the origin. As we now recall, the reason for this modification originates in section 1.3. At times, we will choose pairs (φ, ψ) satisfying the additional condition (2.7), namely

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

Recall that (v) allows to represent any function in $L_1 \cap L_2(\mathbf{R}^d)$ as a semi-continuous superposition of ridgelets (2.8), i.e.

$$f \propto \int \langle f(x), \phi(u\cdot x - b)\rangle \phi(u\cdot x - b) + \sum_{j\geq 0} 2^{j(d-1)} \int \langle f(x), \psi_j(u\cdot x - b)\rangle \psi_j(u\cdot x - b) du db, \tag{4.1}$$

where again the notation ψ_j stands for $2^j\psi(2^j\cdot)$.

Keeping in mind that $R_u f$ denotes the Radon transform of f defined by

$$R_u f(t) = \int_{u\cdot x = t} f(x) dx,$$

we now turn to the main definition of this section.

Definition 4 *For $s \geq 0$ and $p, q > 0$, we say that $f \in R^s_{p,q}$ if $f \in L_1$ and*

$$\operatorname*{Ave}_u \|R_u f * \varphi\|_p < \infty \qquad \text{and} \qquad \left\{ 2^{js} 2^{j(d-1)/2} \left(\operatorname*{Ave}_u \|R_u f * \psi_j\|_p^p \right)^{1/p} \right\} \in \ell_q(\mathbf{N}). \tag{4.2}$$

Note that we have defined our space as a subset of L_1. The major reason is that we truly understand the ridgelet transform for elements of L_1; it is beyond the scope of this thesis to extend the transform to distributions. In our definition, it would be possible to drop the membership to L_1 but the discussion that follows would have been more complicated and even more technical without further enlightening our business.

Next, we define

$$\|f\|_{R^s_{p,q}} = \operatorname*{Ave}_u \|R_u f * \varphi\|_p + \left\{ \sum_{j\geq 0} \left(2^{js} 2^{j(d-1)/2} \left(\operatorname*{Ave}_u \|R_u f * \psi_j\|_p^p \right)^{1/p} \right)^q \right\}^{1/q} \tag{4.3}$$

and we will prove later that this is a norm for a suitable range of parameters s, p and q.

There is an homogeneous version of our definition where the norm (4.2) is replaced by

$$\|f\|_{\dot{R}^s_{p,q}} = \left\{ \sum_{j \in \mathbf{Z}} \left(2^{js} 2^{j(d-1)/2} \left(\operatorname*{Ave}_u \|R_u f * \psi_j\|_p^p \right)^{1/p} \right)^q \right\}^{1/q}. \tag{4.4}$$

To be consistent with the literature of Functional Analysis, we will call these classes $\dot{R}^s_{p,q}$.

Remark 1. It is not hard to show that the definition is independent of the choice of the pair (φ, ψ) provided it satisfies (i)-(iv). Furthermore, it is not necessary to assume (ii)-(iii); i.e., to assume that $\hat{\varphi}$ and $\hat{\psi}$ are compactly supported. As we will show later, one can get equivalent definitions to (4.2)-(4.3) with the relaxed assumption that ψ has a sufficient number of vanishing moments.

Remark 2. Like the Sobolev scales (respectively Besov scales) where a norm is defined based on the Fourier transform (respectively wavelet transform), what we have done here is merely to define a norm based on the ridgelet transform. Let $R_j(u,b)$ the ridgelet coefficient defined by $w_j(u,b)(f) = \langle f(x), \psi_j(u \cdot x - b) \rangle = R_u f * \check{\psi}_j(b)$ for $j \geq 0$ ($v(u,b)(f) = \langle f(x), \varphi(u \cdot x - b) \rangle$). The quantity $\|f\|_{R^s_{p,q}}$ may simply be rewritten as

$$\|f\|_{R^s_{p,q}} = \left(\int |v(u,b)(f)|^p du db \right)^{1/p} + \left\{ \sum_{j \geq 0} \left(2^{js} 2^{j(d-1)/2} \left(\int |w_j(u,b)(f)|^p du db \right)^{1/p} \right)^q \right\}^{1/q},$$

which is a measure of the size of these coefficients.

Remark 3. For $p, q \geq 1$, there are of course obvious analogies with the Besov scale. In fact if $d = 1$, the definitions fully agree – at least within this range of parameters. In higher dimensions, as we will see, the spaces $R^s_{p,q}$ correspond to a different notion of singularity; Besov norms and analogs measure the smoothness of an object where the smoothness is understood as the local regularity of the function. Here, our norm also corresponds to a certain notion of smoothness, but in a different way. The scope of the remainder of this chapter is to attempt to give an accurate intuition of what this new notion is about.

Remark 4. At first, the definiton may seem rather internal. In fact, it is possible to give an external characterization of these spaces in the case where $p = q$. As usual, let $R_u f$ denote the Radon transform of f. Then, roughly speaking, we have for $q = p \geq 1$

$$\|f\|^p_{\mathbf{R}^s_{p,p}} \sim \operatorname*{Ave}_u \|R_u f\|^p_{\mathbf{B}^{s+(d-1)/2}_{p,p}}. \tag{4.5}$$

We explain further what we mean by (4.5). In fact, one can easily be convinced that the $R^s_{p,q}$ norm (4.3) dominates its homogeneous version (4.4). Moreover, it is clear that

$$\|f\|^p_{\dot{\mathbf{R}}^s_{p,p}} \asymp \operatorname*{Ave}_u \ \|R_u f\|^p_{\dot{\mathbf{B}}^{s+(d-1)/2}_{p,p}}$$

in the usual sense of norm equivalence ($\dot{\mathbf{B}}^{s+(d-1)/2}_{p,p}$ is the corresponding homogeneous Besov space). On the other hand, it is trivial to see that

$$\|f\|^p_{\mathbf{R}^s_{p,p}} \leq C \operatorname*{Ave}_u \ \|R_u f\|^p_{\mathbf{B}^{s+(d-1)/2}_{p,p}}.$$

That is the sense of our sloppy equivalence (4.5). In this thesis, we are mainly concerned by the representation, the approximation or the estimation of functions of d-variables that are compactly supported, say, in the unit ball Ω_d. We should note that in this case, the distinction we have just made is irrelevant as both norms (the homogeneous one and the non-homogeneous one) are equivalent and, therefore, (4.5) holds in the usual sense of norm equivalence.

Along the same lines of consideration, one could quickly remark that the space $R^s_{2,2}(\mathbf{R}^d)$ is, indeed, the classical Sobolev space $H^s(\mathbf{R}^d) = B^s_{2,2}(\mathbf{R}^d)$ since it easily follows from our definition that $\|f\|_{R^s_{2,2}(\mathbf{R}^d)}$ is an equivalent norm for H^s. One could go at length over the properties of the new spaces $R^s_{p,q}$, investigating for instance embedding relationships or interpolation properties, etc. However, the goal of this chapter is to show that the spaces $R^s_{p,q}$ characterize objects with a special kind of inhomogeneities rather than exploring these issues.

Our characterization highlights an interesting aspect: the condition does not require any particular smoothness of the Radon transform along the directional variable u.

Example. We now consider a simple example to illustrate the difference between traditional measures of smothness and our new criteria. Let $f(x) = 1_{\{x_1>0\}}(2\pi)^{-d/2}e^{-\|x\|^2/2}$. From a classical point of view, this function has barely one derivative (the first derivative is a singular measure). To quantify this idea, one can show that $f \in \mathbf{B}^s_{1,1}$ if, and only if, $s < 1$. On the other hand, this object is quite smooth in regard to (4.3). In fact, $f \in \mathbf{R}^s_{1,1}$ as long as $s < 1 + (d-1)/2$. We now give a detailed proof of this claim for we will make an extensive use of some details appearing in the argument.

Proposition 6

$$f(x) = 1_{\{x_1>0\}}(2\pi)^{-d/2}e^{-\|x\|^2/2} \in \mathbf{R}^s_{1,1} \qquad \textit{iff} \qquad s < 1 + (d-1)/2.$$

Proof of Proposition. Using the external characterization that we developed (Remark 4), it is sufficient to show that

$$\sigma < d \;\Rightarrow\; \operatorname*{Ave}_u \|R_u f\|^1_{\mathbf{B}^\sigma_{1,1}} < \infty \qquad \text{and} \qquad \sigma \geq d \;\Rightarrow\; \operatorname*{Ave}_u \|R_u f\|^1_{\mathbf{B}^\sigma_{1,1}} = \infty.$$

Let u be a unit vector. Next, let Q be an orthogonal matrix whose first two columns are u and u', where u' is a unit vector, orthogonal to u and belonging to the span of $\{e_1, u\} - e_1$ being the first element of the canonical basis of R^d. We will choose u' so that $u' \cdot e_1 \leq 0$. Now, let x' be the change of coordinates defined by $x' = Q^T x$. We then rewrite the Radon transform as

$$R_u f(t) = \int_{u\cdot x = t} f(x)dx = \int_{x'_1 = t} f(Qx')dx'.$$

So that in our case, the above expression becomes

$$\begin{aligned} R_u f(t) &= \int_{x'_1=t} 1_{\{u\cdot e_1\, x'_1 + u'\cdot e_1\, x'_2\}}(2\pi)^{-d/2}e^{-\|x'\|^2/2}dx'_2...dx'_d \\ &= (2\pi)^{-1/2}e^{-t^2/2}\int 1_{\{u\cdot e_1\, t + u'\cdot e_1\, x'_2\}}(2\pi)^{-1/2}e^{-x'^2_2/2}dx'_2 \end{aligned}$$

To simplify the notations, let ϕ denote the Gaussian density ($\phi(t) = (2\pi)^{-1/2}e^{-t^2/2}$) and Φ be the cumulative distribution function, i.e. $\Phi(t) = \int_{t'\leq t}\phi(t')dt'$. With these notations, we have

$$R_u f(t) = \phi(t)\Phi(\frac{u\cdot e_1\, t}{\sqrt{1-(u\cdot e_1)^2}}).$$

(In case $u = e_1$, the RHS is replaced by $\phi(t)1_{\{t>0\}}$.) We now use the following lemma.

Lemma 5 *Let u be uniformly distributed on the unit sphere. Then the density of $u_1 = u\cdot e_1$ is given by*

$$f(u_1) = c_d(1-u_1^2)^{(d-3)/2},$$

where c_d is a renormalizing constant.

Proof. We give a non-rigorous proof, similar in spirit to what can be found in physics textbooks. We want to calculate the measure of the infinitesimal surface $\delta\mathcal{S}_t = \mathcal{S}^{d-1} \cap \{t \leq u_1 \leq t+dt\}$. The density is clearly symmetric around 0 and, therefore, we may assume that $t \leq 0$ without loss of generality. Let T_u be the tangent hyperplane to the sphere at a point u. Let u' be the unit vector of T_u that belongs to the span of $\{u, e_1\}$. We then clearly have $|u' \cdot e_1|^2 = 1 - |u \cdot e_1|^2 = 1 - u_1^2$; that is, the cosine of the angle between any vector of the tangent hyperplane and e_1 is $1 - u_1^2$. Then, the infinitesimal measure of $\delta\mathcal{S}_t$ is, to the first order of approximation, the one of a cylinder of radius $(1-t^2)^{1/2}$ and sidelength $dt/(1-t^2)^{1/2}$; therefore, its measure is proportional to $(1-t^2)^{(d-2)/2} dt/(1-t^2)^{1/2} = (1-t^2)^{(d-3)/2}$. ∎

Corollary 2 *Again, let u be uniformly distributed on the unit sphere and set $v = u_1/\sqrt{1-u_1^2}$. The density of v is then given by*

$$f(v) = c'_d (1+v^2)^{-d/2}. \tag{4.6}$$

The proof is a simple consequence of the change of variables formula.

End of proof of Proposition 6. With the same notation as in Corollary 1, recall that

$$R_u f(t) = \phi(t)\Phi(vt).$$

Using the rescaling properties of Besov spaces, it is not hard to show that for any $\sigma > 1$ we have that there exist two constants $C_1, C_2 > 0$ (not depending on v) such that

$$C_1 v^{\sigma-1} \leq \|R_u f\|_{\dot{B}^\sigma_{1,1}} \qquad \text{and} \qquad \|R_u f\|_{B^\sigma_{1,1}} \leq C_2(1+v^{\sigma-1}).$$

Using Corollary 1, we have that for $\sigma < d$,

$$\operatorname*{Ave}_u \|R_u f\|_{B^\sigma_{1,1}} \leq C \int (1+v^{\sigma-1})(1+v^2)^{-d/2} < \infty.$$

And conversely, for $\sigma \geq d$,

$$\infty = C \int v^{\sigma-1}(1+v^2)^{-d/2} \leq \operatorname*{Ave}_u \|R_u f\|_{\dot{B}^\sigma_{1,1}},$$

which is what needed to be established. ∎

To close this section, we remark that

Proposition 7 *For $s \geq 0$ and $p, q \geq 1$, the space $R^s_{p,q}$ equipped with the norm $\|f\|_{\overline{R}^s_{p,q}} \equiv \|f\|_1 + \|f\|_{R^s_{p,q}}$ is a Banach space.*

The proof of the proposition is in the Appendix.

4.1.1 Spaces on Compact Domains

As our thesis deals essentially with compactly supported functions (see Chapters 2, 5 and 6), it is necessary to define properly the restriction of our new spaces to compact domains as we will make an extensive use of their properties in the sequel. The definition of this restriction is classical.

Definition 5 *Let Ω be a bounded set of $\mathbf{R}^d$. We define*

$$R^s_{p,q}(\Omega) = \{f \in L_1(\Omega),\ \exists g \in R^s_{p,q}(\mathbf{R}^d) \ \ with \ \ g_{|\Omega} = f\}. \tag{4.7}$$

Now, the following trivial proposition will serve as a preliminary for a subsequent result.

Proposition 8 *The space $R^s_{p,q}(\Omega)$ equipped with the norm*

$$\|f\|_{\overline{R}^s_{p,q}(\Omega)} = \inf \ \|g\|_{\overline{R}^s_{p,q}} = \inf \ \|g\|_{L_1(\mathbf{R}^d)} + \|g\|_{R^s_{p,q}(\mathbf{R}^d)} \tag{4.8}$$

– infimum taken over all $g \in R^s_{p,q}(\mathbf{R}^d)$ in the sense of (4.7) – is a Banach space.

Proof of Proposition 8. It is clear that $R^s_{p,q}(\Omega)$ is a linear space and that $\|f\|_{R^s_{p,q}(\Omega)}$ is a norm. We need to check for completeness. Let $\{f_n\}_{n\geq 1}$ be a Cauchy sequence and assume without loss of generality that

$$\|f_{n+1} - f_n\|_{\overline{R}^s_{p,q}(\Omega)} \leq 3^{-n}, \qquad n \geq 0,$$

(we take $f_0 = 0$). Let $\{g_n\}_{n\geq 1} \in R^s_{p,q}(\mathbf{R}^d)$ be a corresponding sequence in the sense of (4.7)-(4.8) such that

$$\|g_{n+1} - g_n\|_{\overline{R}^s_{p,q}(\mathbf{R}^d)} \leq 2^{-n}.$$

Then we have

$$g = \lim_{n\to\infty} g_n = \sum_{n=1}^{\infty}(g_{n+1} - g_n),$$

with

$$\|\sum_{n=1}^{\infty}(g_{n+1} - g_n)\|_{\overline{R}^s_{p,q}(\mathbf{R}^d)} \leq \sum_{n=1}^{\infty}\|g_{n+1} - g_n\|_{\overline{R}^s_{p,q}(\mathbf{R}^d)} \leq \sum_{n=1}^{\infty} 2^{-n} < \infty.$$

Let $f = g_{|\Omega}$. It follows easily from the last two equations that f is the limit element of f_n. ∎

Remark. Let Ω_1 be a bounded domain such that $\overline{\Omega} \subset \Omega_1$. Suppose moreover, that both Ω and Ω_1 are nice in a sense that, say, their boundary is C^∞. In our definition (4.7)-(4.8), one can restrict $g \in R^s_{p,q}(\mathbf{R}^d)$ to be such that

$$g_{|\Omega} = f \qquad \text{and} \qquad \text{supp } g \subset \Omega_1.$$

Doing so, we obtain the same space $R^s_{p,q}(\Omega)$ and an equivalent quasi norm. The reason is that the multiplication of an element $g \in R^s_{p,q}(\mathbf{R}^d)$ by a fixed window $w \in C^\infty_0$ is a bounded operation from $R^s_{p,q}(\mathbf{R}^d)$ to itself. We will prove this later. In the application we have in mind, we shall consider the unit ball Ω_d as our domain Ω and $2\Omega_d$ as Ω_1.

4.2 $R^s_{p,q}$, A Model For A Variety of Signals

In the previous section, we introduced a new scale of functional classes and presented some basic properties. In this section, we will give a more intuitive characterization of these spaces: we show how they model objects having a very special kind of singularities. To describe these singularities, we introduce a definition due to Donoho (1993).

Definition 6 *Let $\alpha > -1/2$. We say that $\sigma : \mathbf{R} \to \mathbf{R}$ is a normalized singularity of degree α if $\sigma(t)$ is C^R on $\mathbf{R} \setminus \{0\}$ where $R = \max(2, 2+\alpha)$ and*

- $|\sigma(t)| \leq |t|^\alpha \ \forall t$, *and*
- $\left|\frac{d^m}{dt^m}\,\sigma(t)\right| \leq (m + |\alpha|!)\,|t|^{\alpha-m}, \quad t \neq 0, \ m = 1, 2, \ldots, R.$

A normalized singularity is a smooth C^R function which may or may not be singular at the origin. Following Donoho (1993), we list a few examples of normalized singularities: $|t|^\alpha$, $|t|^\alpha 1\{t > 0\}$, $|t|^\alpha w(t)$ where $w(t)$ is a "nice" smooth function properly normalized, etc. Additional examples are given in the reference cited above.

We make use of the definition to construct a class of functions whose typical elements are of the form $\sigma(u \cdot x - t)$.

Definition 7

$$\mathcal{S}_H^\alpha = \{\sum a_i \sigma_i(u_i \cdot x - t_i), \quad \sum |a_i| \leq 1, \ \|\sigma_i\|_1 \leq 1\}, \tag{4.9}$$

where the σ_i's are normalized singularities of degree α.

Our class is a kind of projection model. It is synthesized out of ridge functions that may or may not be singular across hyperplanes (the parameter α measuring the type of discontinuities allowed). Therefore, the model is meant to represent objects composed of singularities across hyperplanes. There may be an arbitrary number of singularities which may be located in all orientations and positions. The conditions on the coefficients a_i's and on the L_1-norm of the σ_i's guarantee that the elements of $\mathcal{S}_H^\alpha$ have finite energy. Now, the main theorem of this section allows us to bracket this very intuitive class between two spaces of the type $R_{p,q}^s$.

Theorem 7 *There exist $C_1, C_2 > 0$ such that*

$$\mathbf{R}_{1,1}^{1+\alpha+(d-1)/2}(C_1) \subset \mathcal{S}_H^\alpha \subset \mathbf{R}_{1,\infty}^{1+\alpha+(d-1)/2}(C_2), \tag{4.10}$$

when these spaces are restricted to the unit ball Ω_d.

(The notation $\mathbf{R}_{1,1}^{1+\alpha+(d-1)/2}(C_1)$ stands for the set of elements of $\mathbf{R}_{1,1}^{1+\alpha+(d-1)/2}(\Omega_d)$ such that $\|f\|_{\mathbf{R}_{1,1}^{1+\alpha+(d-1)/2}} \leq C_1$ and similarly for $\mathbf{R}_{1,\infty}^{1+\alpha+(d-1)/2}(C_2)$.)

For any s the difference between $R_{1,1}^s$ and $R_{1,\infty}^s$ is very subtle. Of course, we have that $\|.\|_{R_{1,\infty}^s}$ is dominated by $\|.\|_{R_{1,1}^s}$; but actually there isn't much gap between these two spaces. (For instance, we shall see that they have the same approximation bounds.) So, the theorem really identifies a very natural class of function whose typical elements may have some discontinuities accross hyperplanes with classes of our new scale.

Remark 1. Note that the result provides some intuition about the smoothness parameter s. As the functions get smoother, i.e., α increases, the parameter s increases accordingly.

The proof of the theorem is fairly involved and requires several steps and composes the remainder of this section. At the heart of the argument, there is the atomic decomposition of the space $\mathbf{R}^s_{1,1}$.

4.2.1 An Embedding Result

In Chapter 2 (Theorems 1 and 3), we derived a reproducing formula and established the sense in which it hold. Our condition requires that the function f one wishes to represent as a continuous (or semi-continuous) superposition of ridge functions must belong to the space $L_1 \cap L_2(\mathbf{R}^d)$. To make use of our reproducing formula, it is most useful to know for which value of the smoothness parameter s is the space $R^s_{1,1}(\Omega_d)$ embedded in $L_2(\Omega_d)$. The following lemma answers this question.

Lemma 6 *Suppose that $f \in \mathbf{R}^s_{1,1}(\Omega_d)$ for $s \geq d/2$. Then $f \in L_2(\Omega_d)$ and*

$$\|f\|_2 \leq C \, \|f\|_{\mathbf{R}^s_{1,1}(\Omega_d)} \tag{4.11}$$

Moreover, the result is sharp in a sense that for $s < d/2$, there are elements in $R^s_{1,1}(\Omega_d)$ that are not square integrable.

Proof of Lemma. It is sufficient to prove the desired norm inequality (4.11) in $R^s_{1,1}(\mathbf{R}^d)$ and such that $g \subset 2\Omega_d$. Indeed, let f be in $R^s_{1,1}(\Omega_d)$; then for any $g \in R^s_{1,1}(\mathbf{R}^d)$ such that $g_{|\Omega_d} = f$ and supp $g \subset 2\Omega_d$, we have

$$\|f\|_2 \leq \|g\|_2 \leq C\|g\|_{R^s_{1,1}(\mathbf{R}^d)}.$$

And taking the infimum over the g's would give the desired result.

Step 1. We then prove the result in the case where $g \in L_2$. We know that

$$\|g\|_2^2 \propto \operatorname*{Ave}_u \|R_u g * \varphi\|_2^2 + \sum_{j \geq 0} 2^{j(d-1)} \operatorname*{Ave}_u \|R_u g * \psi_j\|_2^2.$$

Now it is clear that $R_u g * \varphi$ (resp. $R_u g * \psi_j$) is an analytic function of exponential type 1 (resp. 2^j) and, therefore,

$$\|R_u g * \varphi\|_2 \leq C\|R_u g * \varphi\|_1 \qquad \text{and} \qquad \|R_u g * \psi_j\|_2 \leq C 2^{j/2} \|R_u g * \psi_j\|_1.$$

Consequently,

$$\|g\|_2^2 \leq C \int \left\{ \|R_u g * \varphi\|_1 \|R_u g * \varphi\|_2 + \sum_{j\geq 0} 2^{j(d-1)} 2^{j/2} \|R_u g * \psi_j\|_1 \|R_u g * \psi_j\|_2 \right\} du.$$

But for any $u \in \mathcal{S}^{d-1}$, we have $\|R_u g * \varphi\|_2 \leq C\|g\|_2$ and $\|R_u g * \psi_j\|_2 \leq C\|g\|_2$ for some positive constant C, see for example Candes (1996). Thus the inequality becomes

$$\|g\|_2 \leq C \int \left\{ \|R_u g * \varphi\|_1 + \sum_{j\geq 0} 2^{j(d-1)} 2^{j/2} \|R_u g * \psi_j\|_1 \right\} du.$$

The last inequality holds for any g satisfying the conditions specified above and, therefore, we can conclude to the validity of (4.11).

Step 2. We finish the proof by density. Suppose g is in $R^s_{1,1}(\mathbf{R}^d)$ and that supp $g \subset 2\Omega_d$, we show that

$$\|g\|_2 \leq C\|g\|_{R^s_{1,1}(\mathbf{R}^d)}. \tag{4.12}$$

So, let Δ be a non-negative, radial, infinitely differentiable and compactly supported function in $\mathbf{R}^d$ such that $\int \Delta = 1$. Further, let δ be its Radon transform (because of spherical symmetry, the Radon transform does not depend on u). We define Δ_ϵ to be $\epsilon^{-d}\Delta(\cdot/\epsilon)$ and similarly $\delta_\epsilon = \epsilon^{-1}\delta(\cdot/\epsilon)$. We first show that

$$\lim_{\epsilon\to 0} \|g * \Delta_\epsilon - g\|_{R^s_{1,1}(\mathbf{R}^d)} = 0. \tag{4.13}$$

We have $R_u(g*\Delta_\epsilon)*\psi_j = R_u g*\psi_j*\delta_\epsilon$ (and similarly for the coarse scale, i.e. the convolution with φ). Now, it is immediate to check that for any j, u and b, $(R_u g * \psi_j * \delta_\epsilon)(b) \to (R_u g * \psi_j)(b)$ as $\epsilon \to 0$ (and similarly when ψ_j is replaced by φ). Further, we show that $\lim \|g * \Delta_\epsilon\|_{R^s_{1,1}(\mathbf{R}^d)} = \|g\|_{R^s_{1,1}(\mathbf{R}^d)}$; in other words

$$\lim_{\epsilon\to 0} \operatorname*{Ave}_u \left\{ \|R_u g * \varphi * \phi_\epsilon\|_1 + \sum_{j\geq 0} \|R_u g * \psi_j * \phi_\epsilon\|_1 \right\} = \operatorname*{Ave}_u \left\{ \|R_u g * \varphi\|_1 + \sum_{j\geq 0} \|R_u g * \psi_j\|_1 \right\}. \tag{4.14}$$

By Fatou's lemma, we get that $\liminf \|g * \Delta_\epsilon\|_{R^s_{1,1}(\mathbf{R}^d)} \geq \|g\|_{R^s_{1,1}(\mathbf{R}^d)}$. Conversely, since

$$\|R_u g * \psi_j * \phi_\epsilon\|_1 \leq \|R_u g * \psi_j\|_1 \|\phi_\epsilon\|_1 = \|R_u g * \psi_j\|_1,$$

we have that $\limsup \|g * \Delta_\epsilon\|_{R^s_{1,1}(\mathbf{R}^d)} \leq \|g\|_{R^s_{1,1}(\mathbf{R}^d)}$, which proves (4.14). Finally, Scheffé's lemma establishes the claim (4.13).

We recall that since g is assumed to be in L_1, then $g * \Delta_\epsilon$ is in L_2. Now, from Step 1, we have

$$\|g * \Delta_\epsilon\|_2 \leq C \|g * \Delta_\epsilon\|_{R^s_{1,1}},$$

and thus we deduce that $\{g * \Delta_\epsilon\}$ is a Cauchy sequence in L_2 and therefore converges in L_2. On the other hand it is trivial that $g * \Delta_\epsilon$ converges to g in L_1. Then, $g * \Delta_\epsilon$ converges to g in L_2 and, therefore, we have proved (4.12). The lemma follows. ∎

This lemma has a rather useful corollary.

Corollary 3 *The space* $\mathbf{R}^s_{1,1}(\Omega_d)$ *equipped with the norm*

$$\|f\|_{\mathbf{R}^s_{1,1}}(\Omega_d) = \inf\{\|g\|_{R^s_{1,1}(\mathbf{R}^d)} \; \exists g \in R^s_{p,q}(\mathbf{R}^d) \;\; \textit{with} \;\; g_{|\Omega} = f\} \tag{4.15}$$

is a Banach space.

Proof of Corollary. For compactly supported functions, the L_2 norm dominates the L_1 norm. Now from the previous lemma it is clear that the norm (4.15) and the one of Proposition 8 are equivalent. This proves the claim. ∎

Actually, it is not hard to see that one can extend Lemma 6 and its corollary. For $p \leq 2$, we have $R^s_{p,q}(\Omega_d) \subset L_2(\Omega_d)$ as long as $s > 1/p - 1/2$ (the argument being essentially the same as the one spelled out in the proof of Lemma 6) with continuous injection. Moreover, for $p > 2$, the same is true as long as, this time, $s > 0$. Although we do not prove these results here, we will use them in Chapter 5.

4.2.2 Atomic Decomposition of $R^s_{1,1}(\Omega_d)$

We start out with a lemma that helps to establish our decomposition result.

Lemma 7 *The set* $\{\varphi(u \cdot x - b),\ 2^{-js} 2^{j\frac{d-1}{2}} \psi_j(u \cdot x - b),\ \ j \geq 0, u \in \mathcal{S}^{d-1}, b \in \mathbf{R}\}$ *is bounded in* $R^s_{1,1}(\Omega_d)$.

Proof of Lemma. Let w be in $C_0^\infty(2\Omega_d)$ such that its restriction to the unit ball is 1 ($w_{|\Omega_d} = 1$). It is enough to show that, say, the set $\{w(x)\varphi(u\cdot x-b), w(x)2^{-js}2^{j\frac{d-1}{2}}\psi_j(u\cdot x-b)\}$ is bounded in $R^s_{1,1}(\mathbf{R}^d)$. The analysis is greatly simplified if one replaces the window w by the d dimensional Gaussian density Φ_d; in fact, it is sufficient to prove the lemma with the gaussian window. Again, the reason is that the multiplication of an element in $R^s_{1,1}(\mathbf{R}^d)$ with a fixed C_0^∞ function is a bounded operation.

Lemma 8 *Suppose that $f \in R^s_{1,1}$ and let χ be a bounded C^∞ function together with its derivatives. Then there exists a constant c_χ such that*

$$\|f\chi\|_{R^s_{1,1}} \leq c_\chi \|f\|_{R^s_{1,1}}.$$

We postpone the proof of this intuitive lemma. As we shall see in the next chapter, we can prove Lemma 7 directly without the help of the intermediate Lemma 8.

End of proof of Lemma 7. The norm $\|\Phi_d(x)\psi_j(u\cdot x-b)\|_{R^s_{1,1}}$ being clearly invariant, we will without loss of generality assume that $u = e_1$. Finally g_γ will denote $R_u\{\Phi_d * \psi_j(u\cdot x-b)\}$. Following the discussion presented in the last chapter, it is sufficient to show that there exists a positive constant C such that

$$\operatorname*{Ave}_u \|g_\gamma\|_{B^{s+(d-1)/2}_{1,1}} \leq C2^{js}2^{-j\frac{d-1}{2}}. \tag{4.16}$$

In the remainder proof, σ will stand for the quantity $s + (d-1)/2$. Reproducing the calculations from the last chapter, we rapidly get

$$g_\gamma(t) = \phi(t)\int \phi(y)2^j\psi(2^j(u_1 t - \sqrt{1-u_1^2}y - b))\,dy$$

We recall that v denotes the ratio $u_1/\sqrt{1-u_1^2}$, so that $\sqrt{1-u_1^2} = (1+v^2)^{-1/2}$.
case 1. $a = 2^j(1+v^2)^{-1/2} \geq 1$. Then g_γ can be written as

$$g_\gamma(t) = \phi(t)(1+v^2)\,(\phi * \psi_a)\,(vt - (1+v^2)^{1/2}b).$$

Next, we use the following result from Triebel (1983)[Page 127]:

$$\|f * g\|_{B^\sigma_{1,1}} \leq \|f\|_{B^\sigma_{1,1}}\|g\|_{B^0_{1,\infty}}.$$

Applying this inequality gives

$$\begin{aligned}\|\phi * \psi_a\|_{B^\sigma_{1,1}} &\leq \|\phi\|_{B^\sigma_{1,1}} \|a\psi(a\,\cdot)\|_{B^0_{1,\infty}} \\ &\leq C_\psi \|\phi\|_{B^\sigma_{1,1}} \leq C,\end{aligned}$$

where the constant C_ψ depends only on ψ. Therefore,

$$\|g_\gamma\|_{B^\sigma_{1,1}} \leq C(1+v^2)^{1/2}(1+v^{\sigma-1}).$$

case 2. $2^j(1+v^2)^{-1/2} \leq 1$. We rewrite things slightly differently.

$$g_\gamma(t) = \phi(t)\int \phi(y)2^j\psi(2^j(u_1 t - \sqrt{1-u_1^2}y))dy = \phi(t)\int \phi(y)2^j\psi(2^j(u_1 t - \epsilon y))dy = \phi(t)g_\epsilon(t),$$

where obviously, $\epsilon = 2^j\sqrt{1-u_1^2} = 2^j(1+v^2)^{-1/2} \leq 1$ and $g_\epsilon(t) = \int \phi(y)\psi(t-\epsilon y)dy$. For k a non-negative integer, we have

$$\begin{aligned}|g_\epsilon^{(k)}(t)| &= \left|\int \phi(y)\psi^{(k)}(t-\epsilon y)dy\right| \\ &\leq \int \phi(y)|\psi^{(k)}(t-\epsilon y)|dy \\ &\leq C\int \phi(y)(1+|t-\epsilon y|)^{-m}dy \text{ for some positive } m \\ &= C\left(\int_{|t-\epsilon y|>|t|/2} \phi(y)(1+|t-\epsilon y|)^{-m}dy + \int_{|t-\epsilon y|\leq|t|/2} \phi(y)(1+|t-\epsilon y|)^{-m}dy\right) \\ &\leq C\left(\int_{|t-\epsilon y|>|t|/2} \phi(y)(1+|t|/2)^{-m}dy + \int_{|t-\epsilon y|\leq|t|/2} |\phi(y)\|dy\right) \\ &\leq C\left((1+|t|/2)^{-m} + \int_{|\epsilon y|>|t|/2} |\phi(y)\|dy\right) \\ &\leq C\left((1+|t|)^{-m} + (1+|\epsilon^{-1}t|)^{-m}\right)\end{aligned}$$

In the case where $\epsilon \leq 1$, the last expression implies that $|g_\epsilon^{(k)}(t)| \leq C(1+|t|)^{-m}$. Therefore, $g_\epsilon \in \mathcal{S}(\mathbf{R})$ and is bounded in almost every known space and particularly $B^\sigma_{1,1}$.

Now, the rescaling properties of Besov spaces gives that for $\sigma > 0$, there exists some positive constant C such that for $|\lambda| \geq 1$,

$$\|f(\lambda\,\cdot)\|_{B^\sigma_{1,1}} \leq C|\lambda|^\sigma \|f\|_{B^\sigma_{1,1}}$$

for any $f \in B^\sigma_{1,1}$. It is then obvious that, from $g_\gamma(t) = 2^j \phi(t)\, g_\epsilon(2^j u_1 t)$, we can deduce

$$\|g_\gamma\|_{B^\sigma_{1,1}} \leq C 2^{j(\sigma+1)},$$

for some positive constant C.

Combining the two estimates from Case 1 and 2, we get that for $\sigma > 1$,

$$\begin{aligned}
\operatorname*{Ave}_u \|g_\gamma\|_{B^\sigma_{1,1}} &\propto \int \|g_\gamma\|_{B^\sigma_{1,1}} \frac{dv}{(1+v^2)^{d/2}} \\
&\leq C\left(\int_{|v|\leq 2^j} \|g_\gamma\|_{B^\sigma_{1,1}} \frac{dv}{(1+v^2)^{d/2}} + \int_{|v|>2^j} \|g_\gamma\|_{B^\sigma_{1,1}} \frac{dv}{(1+v^2)^{d/2}} \right) \\
&\leq C\left(\int_{|v|\leq 2^j} (1+|v|^{\sigma-1}) \frac{dv}{(1+v^2)^{(d-1)/2}} + \int_{|v|>2^j} 2^{j\sigma} \frac{dv}{(1+v^2)^{d/2}} \right) \\
&\leq C\left(2^{j\sigma} 2^{-j(d-1)} + 2^{j\sigma} 2^{-j(d-1)} \right).
\end{aligned}$$

Now if we recall that $\sigma = s + (d-1)/2$, we have proven (4.16). The story is the same for the coarse scale ridgelets $\varphi(u \cdot x - b)$ and thus the proof of the lemma is complete. ∎

Theorem 8 (Atomic Decomposition of $R^s_{1,1}(\Omega_d)$) *Suppose that $f \in R^s_{1,1}(\Omega_d)$ for $s \geq d/2$. Then, there exist numerical sequences such that*

$$f(x) = \sum_{k=1}^{\infty} \alpha_k 2^{-j_k s} 2^{j_k \frac{d-1}{2}} \psi_{j_k}(u_k \cdot x - b_k), \tag{4.17}$$

with the convergence in $R^s_{1,1}$. Moreover,

$$\sum_{k=1}^{\infty} |\alpha_k| \leq 2(2\pi)^{-d}\, \|f\|_{R^s_{1,1}}.$$

(We recall the convention we adopted in section 3.6.1: for $j_k = -1$, $\psi_{j_k}(u_k \cdot x - b_k)$ stands for $\varphi(u_k \cdot x - b_k)$.)

We follow Frazier, Jawerth, and Weiss (1991) since our proof is a minor adaptation of their argument. In order to prove our claim, we shall use a lemma from functional analysis that can be obtained from the Hahn Banach theorem:

Lemma 9 *Let K be a closed, convex and bounded subset of a Banach space B over the reals. If x does not belong to K, then there exists a continuous real valued linear functional such that* $\sup_{w\in K} l(w) < l(x)$.

Proof of the Theorem. So, let f be in $R^s_{1,1}(\Omega_d)$ for $s \geq d/2$. We then know that $f \in L_1 \cap L_2\ (\Omega_d)$ and that we have

$$f \propto \int \langle f(x), \varphi(u\cdot x - b)\rangle \varphi(u\cdot x - b) + \sum_{j\geq 0} 2^{j(d-1)} \int \langle f(x), \psi_j(u\cdot x - b)\rangle \psi_j(u\cdot x - b) dudb, \tag{4.18}$$

where the equality means has the following sense: $S_0(f)(x) = \int \langle f(x), \varphi(u\cdot x - b)\rangle \varphi(u\cdot x - b)dudb \quad \in L_2$, for any $j \geq 0$, $\Delta_j(f)(x) = 2^{j(d-1)} \int \langle f(x), \psi_j(u\cdot x - b)\rangle \psi_j(u\cdot x - b)dudb \quad \in L_2$, and

$$S_J(f) = S_0(f) + \sum_{0\leq j\leq J-1} \Delta_j(f) \to f \ \text{ in } L_2 \ \text{ as } \ J \to \infty$$

In addition, the convergence does also take place in $R^s_{1,1}(\Omega_d)$. To see that, suppose $g \in R^s_{1,1}(\mathbf{R}^d)$ such that $g_{|\Omega_d} = f$. With the same notation as above, we have

$$g = S_0(g) + \sum_{j\geq 0} \Delta_j(g).$$

The convexity of the norm implies

$$\begin{aligned} \|\Delta_j(g)\|_{R^s_{1,1}(\Omega_d)} &\leq 2^{j(d-1)} \int |w_j(g)(u,b)| dudb \ \sup_{u,b} \|\psi_j(u\cdot x - b)\|_{R^s_{1,1}(\Omega_d} \\ &\leq C 2^{j(d-1)} 2^{js} 2^{-j\frac{d-1}{2}} \int |w_j(g)(u,b)| dudb \\ &= C 2^{js} 2^{j\frac{d-1}{2}} \int |w_j(g)(u,b)| dudb \end{aligned}$$

where the last inequality is a consequence of Lemma 7. We then immediatly have that

$$\|g - S_J(g)\|_{R^s_{1,1}(\Omega_d)} \leq C \sum_{j>J} C 2^{js} 2^{j\frac{d-1}{2}} \int |w_j(g)(u,b)| du db \quad \to 0 \quad \text{as } J \to \infty$$

by definition of the norm $\|.\|_{R^s_{1,1}(\mathbf{R}^d)}$. Since g (resp. $S_J(g)$) coincides with f (resp. $S_J(f)$) on the unit ball, the convergence is established.

We are now in position to finish up the proof of the theorem. Let A be the set defined by

$$A = \{\pm\varphi(u \cdot x - b), u \in \mathbf{S}^{d-1},\ b \in \mathbf{R}\} \cup \{\pm 2^{-js} 2^{j\frac{d-1}{2}} \psi_j(u \cdot x - b), j \geq 0,\ u \in \mathbf{S}^{d-1},\ b \in \mathbf{R}\}.$$

Lemma 7 implies that A is a bounded subset of $R^s_{1,1}(\Omega_d)$. Let $\Gamma = \{\sum_{k=1}^m \rho_k f_k,\ \rho_k \geq 0,\ \sum_{k=1}^m \rho_k = 1,\ f_k \in A,\ m \in \mathbf{N}\}$ and $K = \overline{\Gamma}$; then K is closed, convex and bounded. Suppose $\|f\|_{R^s_{1,1}(\Omega_d)} \leq (2\pi)^d$. For every l in the dual of $R^s_{1,1}(\Omega_d)$, we have

$$\begin{aligned}
l(f) &= (2\pi)^{-d} \int v(g)(u,b) l(\varphi(u \cdot x - b)) \\
&\quad + \sum_{j \geq 0} 2^{j(d-1)} 2^{js} 2^{-j\frac{d-1}{2}} \int w_j(g)(u,b) l(2^{-js} 2^{\frac{d-1}{2}} \psi_j(u \cdot x - b)) du db \\
&\leq (2\pi)^{-d} \sup_{h \in A} l(h) \left(\int |v(g)(u,b)| + 2^{js} 2^{j\frac{d-1}{2}} |w_j(g)(u,b)| du db \right) \\
&\leq (2\pi)^{-d} \sup_{h \in A} l(h) \|g\|_{R^s_{1,1}},
\end{aligned}$$

where again g satisfies the condition listed above. Taking the infimum over g yields

$$l(f) \leq (2\pi)^{-d} \sup_{h \in A} l(h) \|f\|_{R^s_{1,1}(\Omega_d)} \leq \sup_{h \in A} l(h).$$

Hence, by Lemma 9, $f \in K$. We can then conclude that

$$f(x) = \sum_{k=1}^{m_1} \rho_k^1 2^{-j_k^1 s} 2^{j_k^1 \frac{d-1}{2}} \psi_{j_k^1}(u_k^1 \cdot x - b_k^1) + f_1(x),$$

where $f_1 \in R^s_{1,1}(\Omega_d)$, $\|f_1\|_{R^s_{1,1}(\Omega_d)} \leq \frac{1}{2}(2\pi)^d$ and $\sum_{k=1}^{m_1} \rho_k^1 = 1$. We then iterate this process with f replaced by $2f_1$, and so on, obtaining

$$f(x) = \sum_{n=1}^{\infty} 2^{-n+1} \sum_{k=1}^{m_n} \rho_k^n 2^{-j_k^n s} 2^{j_k^n \frac{d-1}{2}} \psi_{j_k^n}(u_k^n \cdot x - b_k^n) + f_1(x)$$

and

$$\sum_{n=1}^{\infty} 2^{-n+1} \sum_{k=1}^{m_n} \rho_k^n = \sum_{n=1}^{\infty} 2^{-n+1} = 2.$$

This can be rewritten in the form

$$f(x) = \sum_{k=1}^{\infty} \alpha_k 2^{-j_k s} 2^{j_k \frac{d-1}{2}} \psi_{j_k}(u_k \cdot x - b_k),$$

with $\sum_{k=1}^{\infty} |\alpha_k| = 2$. The general case is now immediate. ∎

To cut a long story short, if we are given a function f on, say, the unit ball Ω_d which can be extended to a function in $R^s_{1,1}$ in the whole space, then we have

$$f(x) = \sum_{k=1}^{\infty} \alpha_k 2^{-j_k s} 2^{j_k \frac{d-1}{2}} \psi_{j_k}(u_k \cdot x - b_k). \tag{4.19}$$

4.2.3 Proof of the Main Result

End of proof of Theorem. Now, $s = 1 + \alpha + (d-1)/2$ with $\alpha \geq -1/2$. We can therefore rewrite (4.19) as

$$f(x) = \sum_{k=1:j_k=-1}^{\infty} \alpha_k \varphi(u_k \cdot x - b_k) + \sum_{k=1:j_k \geq 0}^{\infty} \alpha_k 2^{-j_k \alpha} \psi(2^{j_k}(u_k \cdot x - b_k)).$$

Now that we have been able to establish the atomic decomposition formula, the rest of the proof is almost identical to the one of Donoho (1993). We reproduce it here for the sake of completeness. We first, suppose that $-1/2 \leq \alpha \leq 0$. Define

$$\zeta = \max \left(\sup_{t \neq 0} \sup_{0 \leq m \leq R} \left| \frac{d^m}{dt^m} \psi(t) \right| \leq (m + |\alpha|!) \, |t|^{\alpha - m}, \|\psi\|_1 \right). \tag{4.20}$$

As $\psi \in \mathcal{S}(\mathbf{R})$ is regular, $\zeta < \infty$. It follows that with $\sigma = \psi/\zeta$, we have the representation

$$a^{-\alpha}\psi(a(t-b)) = a^{-\alpha}\zeta\sigma(a(t-b)),$$

where σ is a normalized singularity of degree α and by construction $\|\sigma\|_1 \leq 1$. We note that if σ is a normalized singularity of degree α, then $a^{-\alpha}\sigma(at)$ is again a normalized singularity of degree α. Hence

$$a^{-\alpha}\psi(a(t-b)) = \zeta\tilde{\sigma}(t-b),$$

where $\tilde{\sigma}$ is a normalized singularity of degree α with $\|\tilde{\sigma}\|_1 \leq a^{-1-\alpha} \leq 1$ if say, $a \geq 1$. Similarly, defining ζ' as in (4.20) with φ in place of ψ gives

$$\varphi(t) = \zeta'\sigma_0(t)$$

where again, σ_0 is a normalized singularity of degree α so that $\|\sigma_0\| \leq 1$. From (4.19), one can now check that we may write

$$f = \sum_k a_k \sigma_k(u_k \cdot x - b_k)$$

where $a_k = \zeta\alpha_k$ if $j_k \geq 0$ and $a_k = \zeta'\alpha_k$ if $j_k = -1$. Now

$$\sum_k |a_k| = \zeta' \sum_{k:j_k=-1} |\alpha_k| + \zeta \sum_{k:j_k\geq 0} |\alpha_k| \leq \max(\zeta,\zeta')2(2\pi)^{-d}\|f\|_{R^{1+\alpha+(d-1)/2}_{1,1}}.$$

Hence setting $C_1 = \frac{1}{2}\max(\zeta,\zeta')^{-1}(2\pi)^d$ gives the desired result.

In the case $\alpha > 0$, we argue not that ψ is itself a normalized α-singularity, but instead that it is a limit of such singularities,

$$\psi = \sum_{l=1}^{\infty} w_l\sigma_l,$$

with $\|\sigma_l\|_1 \leq 1$ and $\zeta = \sum_l |w_l|$ – and a similar decomposition for φ exists with say, ζ'

instead of ζ. The representation (4.19) may be rewritten as

$$\begin{aligned} f(x) &= \sum_k \alpha_k 2^{-j_k\alpha} \sum_l w_l \sigma_l(2^{j_k}(u_k \cdot x - b_k)) \\ &= \sum_k \alpha_k \sum_l w_l \tilde{\sigma}_l(u_k \cdot x - b_k) \\ &= \sum_i a_i \tilde{\sigma}_i(u_i \cdot x - b_i) \end{aligned}$$

where the index i runs through an enumeration of doubles (k,l), $a_i = \alpha_k w_l$, $b_i = b_k$, $u_i = u_k$ and $\tilde{\sigma}_i(\cdot - b_i) = 2^{-j_k\alpha}\sigma_l(2^{j_k}(\cdot - b_k))$. We remark that $\|\tilde{\sigma}_i\|_1 \leq 2^{-j_k(1+\alpha)}$ so that $\|\sigma_l\|_1 \leq 1$. The reasoning continues as before giving the inclusion for $C_1 = \frac{1}{2}(2\pi)^d \max(\zeta, \zeta')^{-1}$.

We now turn to the proof of the right inclusion. So let f be the superposition of normalized singularities of degree α; i.e.,

$$f(x) = \sum_i a_i \sigma_i(u_i \cdot x - b_i),$$

with $\sum_i |a_i| = 1$. We are going to show that f is bounded in $R^{1+\alpha+(d-1)/2}_{1,\infty}(\Omega_d)$ by some constant C. Of course, it is enough to establish that $\sigma(u \cdot x - b)$ is uniformly bounded in $R^{1+\alpha+(d-1)/2}_{1,\infty}(\Omega_d)$ for any α-singularity σ. Again, it is therefore sufficient to show that $\sigma(u \cdot x - b)\Phi_d(x)$ (Φ_d standard gausssian kernel) is bounded in $R^{1+\alpha+(d-1)/2}_{1,\infty}(\mathbf{R}^d)$. The norm being clearly invariant by rotation, we will work with $\sigma(x_1 - b)$ that we will denote σ by a slight abuse of notations. We want to prove

$$\operatorname*{Ave}_u \|R_u(\Phi_d\sigma) * \varphi\|_1 \leq C \quad \text{and} \quad \sup_{j\geq 0} 2^{j(\alpha+d)} \operatorname*{Ave}_u \|R_u(\Phi_d\sigma) * \psi_j\|_1 \leq C.$$

The basic calculation that we have already used gives

$$\begin{aligned} R_u(\Phi_d\sigma)(t) &= \phi(t) \int \phi(y)\sigma(u_1x_1 - (1-u_1^2)^{1/2}y - b)\, dy \\ &= \phi(t)(1-u_1^2)^{\alpha/2} \int \phi(y)\tilde{\sigma}(u_1x_1/(1-u_1^2)^{1/2} - y - b/(1-u_1^2)^{1/2})\, dy \\ &= \phi(t)(1+v^2)^{-\alpha/2}(\phi * \tilde{\sigma})(vt - (1+v^2)^{1/2}b), \end{aligned}$$

where we recall that $v = u_1/(1-u_1^2)^{1/2}$. Here, $\tilde{\sigma}$ is of course a normalized singularity of

degree α. Suppose now that $|v| \geq 2^j$; we then have that

$$\|\phi * \tilde{\sigma}\|_{\dot{B}^{1+\alpha}_{1,\infty}} \leq C \|\phi\|_{\dot{B}^{0}_{1,\infty}} \|\tilde{\sigma}\|_{\dot{B}^{1+\alpha}_{1,\infty}},$$

and a rescaling argument gives

$$\|\phi * \tilde{\sigma}(v \cdot - (1+v^2)^{1/2} b\|_{\dot{B}^{1+\alpha}_{1,\infty}} \leq C |v|^{\alpha} \|\phi\|_{\dot{B}^{0}_{1,\infty}} \|\tilde{\sigma}\|_{\dot{B}^{1+\alpha}_{1,\infty}}.$$

Therefore, it can be shown that

$$\|R_u(\Phi_d \sigma)\|_{\dot{B}^{1+\alpha}_{1,\infty}} \leq C \|\phi\|_{\dot{B}^{0}_{1,\infty}} \|\tilde{\sigma}\|_{\dot{B}^{1+\alpha}_{1,\infty}}.$$

Now, a calculation in Donoho (1993) shows that the homogeneous Besov norm $\|\cdot\|_{\dot{B}^{1+\alpha}_{1,\infty}}$ is uniformely bounded over all the normalized singularities of degree α. Hence, for $|v| \geq 2^j$ we have

$$\|R_u(\Phi_d \sigma) * \psi_j\|_1 \leq C\, 2^{-j(1+\alpha)}.$$

For $|v| \leq 2^j$, the picture is slightly different. Let s be greater than $\alpha+d$, e.g take $s = 1+\alpha+d$;

$$\|\phi * \tilde{\sigma}\|_{\dot{B}^{s}_{1,\infty}} \leq C \|\phi\|_{\dot{B}^{d}_{1,\infty}} \|\tilde{\sigma}\|_{\dot{B}^{1+\alpha}_{1,\infty}}.$$

From this, we deduce that (rescaling property)

$$\|R_u(\Phi_d \sigma)\|_{\dot{B}^{s}_{1,\infty}} \leq C\, (1+v^2)^{d/2} \|\phi\|_{\dot{B}^{d}_{1,\infty}} \|\tilde{\sigma}\|_{\dot{B}^{1+\alpha}_{1,\infty}},$$

which implies

$$\|R_u(\Phi_d \sigma) * \psi_j\|_1 \leq C\, 2^{-j(\alpha+1+d)} (1+v^2)^{d/2},$$

because of a previous observation. Averaging our two estimates yields (recall that the

measure is $\propto dv/(1+v^2)^{d/2}$)

$$\begin{aligned}
\operatorname*{Ave}_u \|R_u(\Phi_d\sigma) * \psi_j\|_1 &\leq C\left(2^{-j(\alpha+1+d)}\int_{|v|\leq 2^j}\frac{(1+v^2)^{d/2}}{(1+v^2)^{d/2}}dv + 2^{-j(1+\alpha)}\int_{|v|>2^j}\frac{dv}{(1+v^2)^{d/2}}\right)\\
&\leq C\left(2^{-j(\alpha+1+d)}2^j + 2^{-j(1+\alpha)}2^{-j(d-1)}\right)\\
&\leq C2^{-j(\alpha+d)}
\end{aligned}$$

which is what needed to be shown. Of course, we also have for the coarse scale (ϕ denoting the 1-dimensional gaussian density)

$$\begin{aligned}
\operatorname*{Ave}_u \|R_u\Phi_d\sigma * \varphi\|_1 \leq \operatorname*{Ave}_u \|R_u\Phi_d\sigma\|_1\|\varphi\|_1 &\leq \operatorname*{Ave}_u \|\Phi_d\sigma\|_1\|\varphi\|_1\\
&= \operatorname*{Ave}_u \|\phi\sigma(\cdot - b)\|_1\|\varphi\|_1 \leq C\,\|\sigma(\cdot - b)\|_1 \leq C.
\end{aligned}$$

Theorem 7 is now completely proved. ∎

Chapter 5

Approximation

This chapter shows how we can apply the ridgelet transform to derive new approximation bounds: that is, we derive both a lower bound and an upper bound of approximation for our new functional classes. These two bounds are essentially the same and show that in a sense, the ridgelet-dictionary is optimal for approximating functions from these classes.

5.1 Approximation Theorem

In order to state our approximation result, we need a bit of terminology. Suppose we are given a dictionary $\mathcal{D} = \{g_\lambda,\ \lambda \in \Lambda\}$. For a function f, we define its approximation error by N-elements of the dictionary $\mathcal{D}$ by

$$\inf_{(\alpha_i)_{i=1}^N} \inf_{(\lambda_i)_{i=1}^N} \|f - \sum_{i=1}^{N} \alpha_i g_{\lambda_i}\|_H \equiv d_N(f, \mathcal{D}). \tag{5.1}$$

Suppose now that we are interested in the approximation of classes of functions. We characterize the rate of approximation of the class $\mathcal{F}$ by N elements from $\mathcal{D}$ by

$$d_N(\mathcal{F}, \mathcal{D}) = \sup_{f \in \mathcal{F}} d_N(f, \mathcal{D}), \tag{5.2}$$

that is, the worst case error over $\mathcal{F}$.

Next, suppose we are given a frame $(\psi_\gamma) \in \Gamma_d$. We know that any element of, say $L_2(\Omega_d)$

might be represented as

$$f = \sum_{\gamma \in \Gamma_d} \langle f, \psi_\gamma \rangle \tilde{\psi}_\gamma = \sum_{\gamma \in \Gamma_d} \langle f, \tilde{\psi}_\gamma \rangle \psi_\gamma. \tag{5.3}$$

To simplify, α_γ will denote the ridgelet coefficient $\langle f, \psi_\gamma \rangle$. Finally, $\tilde{f}_N$ will denote the N-term dual-ridgelet expansion where one only keeps the dual ridgelets corresponding to the N largest ridgelet coefficients. That is,

$$\tilde{f}_N = \sum_{\gamma \in \Gamma_d} \alpha_\gamma 1_{\{|\alpha_\gamma| \geq |\alpha|_{(N)}\}} \tilde{\psi}_\gamma. \tag{5.4}$$

In this framework, we are interested in the L_2 approximation of $\mathcal{F} = R^s_{p,q}(C) = \{f,\ \|f\|_{R^s_{p,q}} \leq C\}$ over, say, the unit ball Ω_d. That is to say, we measure the approximation error in $L_2(\Omega_d)$. We have the following result:

Theorem 9 *Consider the class $\mathbf{R}^s_{p,q}(C)$ and assume $s > d(1/p - 1/2)_+$.*

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

$$d_N(R^s_{p,q}(C), \mathcal{D}) \geq K_1 N^{-s/d}, \tag{5.5}$$

  where the constant K_1 depends at most upon s, p, q.

- *(ii) Simple thresholding achieves the optimal rate i.e.*

$$\|f - \tilde{f}_N\|_{L_2(\Omega_d)} \leq K_2 N^{-s/d}, \tag{5.6}$$

  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 $R^s_{p,q}(C)$, than what can be obtained using $\mathcal{D}_{Dual-Ridge}$. 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 Donoho and Johnstone (1995). "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 $R^s_{p,q}(C)$.), 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 (5.1)–(5.2).

Remark. In Chapter 4, we considered a very natural class $\mathcal{S}_H^\alpha$ of objects having a special kind of singularities (Definition 7) and showed that they almost correspond to balls in $R_{p,q}^s$ (Theorem 7); it follows from Theorems 7 and 9 that these classes have the same lower and upper bounds, as the L_2 approximation rate (5.5)-(5.6) does not depend on the parameters p and q. Further, our result implies that thresholding the dual-ridgelet expansion is optimal for approximating objects from these classes.

5.2 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. (In statistics for example, this technique is often referred to as the construction of Assouad's cube.) More specifically, a result from Donoho and Johnstone (1995) shows how the hypercube embedding limits the approximation error.

Theorem 10 *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

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

In our situation, the construction of embedded hypercubes involves properties of the frame. Roughly speaking, for a fixed scale j, a subsequence of the frame $(\psi_\gamma)_{\gamma \in \Gamma_j}$ properly rescaled is an hypercube of the desired dimension. In order to prove this fact, we need first to establish a number of key estimates about the decay of the kernel $\langle \psi_\gamma, \psi_{\gamma'} \rangle$. We will notice that some of these key estimates support a number of claims that have been made in Chapter 4.

5.2.1 Fundamental Estimates

The purpose of this very technical section is then to show that the kernel $\langle \psi_\gamma, \psi_{\gamma'} \rangle$ for $\gamma, \gamma' \in \Gamma$ is "almost diagonal." As we will see, the fact that $\langle \psi_\gamma, \psi_{\gamma'} \rangle$ decays rapidly as the "distance" between the indices (γ, γ') increases is a crucial fact of our analysis.

Of course, $\langle \psi_\gamma, \psi_{\gamma'} \rangle$ makes no sense, as the ridgelets ψ_γ are not square-integrable. So as usual, we take a fixed window w in $\mathcal{S}(\mathbf{R}^d)$ and look at the inner product with respect to the signed measure w. The goal of this section is to give upper bounds on the kernel:

$$K_w(\gamma, \gamma') = \int \psi_j(u \cdot x - b)\psi_{j'}(u' \cdot x - b')w(x)\, dx. \tag{5.7}$$

(Recall the convention about our notations introduced in section 3.6.1: $\psi_{-1}(u \cdot x - b)$ stands for $\varphi(u \cdot x - b)$.) Let Q be an orthogonal change of coordinates ($x = Qx'$) such that

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

Finally, we set v to be $u \cdot u'/\sqrt{1 - (u \cdot u')^2}$. (We recall that for a fixed u', and a uniform distribution of u on the sphere, the density of v is proportional to $(1 + v^2)^{-d/2}$.) With our notations, the kernel (5.7) can be rewritten as

$$\begin{aligned} K_w(\gamma, \gamma') &= \int \psi_j((1 + v^2)^{-1/2}(vx'_1 - x'_2) - b)\ \psi_{j'}(x'_1 - b')\ w(Qx')\, dx' \\ &= \int \psi_j((1 + v^2)^{-1/2}(vx'_1 - x'_2) - b)\ \psi_{j'}(x'_1 - b')\ w_Q(x'_1, x'_2)\, dx'_1 dx'_2, \end{aligned}$$

where $w_Q(x'_1, x'_2)$ is of course $\int w(Qx')\, dx'_3 \ldots dx'_d$. It is trivial to see that w_Q belongs to $\mathcal{S}(\mathbf{R}^2)$ and that for all Q, and any n_1, n_2, there is a constant C (depending on w, n_1, n_2, and m) with

$$\left| \frac{\partial^{n_1+n_2}}{\partial x_1^{n_1} \partial x_2^{n_2}} w_Q(x_1, x_2) \right| \leq C\ (1 + |x_1| + |x_2|)^{-2m} \leq C\ (1 + |x_1|)^{-m}(1 + |x_2|)^{-m}.$$

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

$$\left| \frac{d^n}{dt^n} \psi(t) \right| \leq C(1 + |t|)^{-m}$$

for some constant C (depending only upon m and n) – and similarly for φ. Further, we will suppose that ψ has vanishing moments through order D.

Lemma 10 *Assume $j' \geq j$ and suppose $n < \min(R, D)$. Then, for each $n, m \geq 0$, there is a constant C (depending on n and m) so that:*

(i) Suppose $2^j(1+v^2)^{-1/2} \geq 1$, then

$$|K_w(\gamma, \gamma')| \leq C\, 2^{-j'n} 2^{-jn} (1+v^2)^{n+1/2} (1+|b'|)^{-m} (1+|vb' - (1+v^2)^{1/2} b|)^{-m}.$$

(ii) Suppose $2^j(1+v^2)^{-1/2} \leq 1$, then

$$|K_w(\gamma, \gamma')| \leq C\, 2^{-j'n} 2^{j(n+1)} (1+|b'|)^{-m} \left(1 + 2^j \left| b' - \frac{(1+v^2)^{1/2}}{v} b \right| \right)^{-m}.$$

Proof of Lemma.

Case (i). The change of variables $x_1' = x_1, x_2' = u_1 x_1 - u_2 x_2$ allows rewriting the kernel as

$$K_w(\gamma, \gamma') = \int_{\mathbf{R}^2} \psi_j(x_2' - b) \psi_{j'}(x_1' - b') w_Q(x_1', v x_1' - (1+v^2)^{1/2} x_2')\, dx_1' dx_2'.$$

Now let $\tilde{w}_Q$ be the function defined by $\tilde{w}_Q(x_1', x_2') = w_Q(x_1', v x_1' - (1+v^2)^{1/2} x_2')\, dx_1' dx_2'$. It is fairly clear that

$$\left| \frac{\partial^{n_1+n_2}}{\partial x_1'^{n_1} \partial x_2'^{n_2}} \tilde{w}_Q(x_1', x_2') \right| = C\, (1+v^2)^{(n_1+n_2)/2} (1 + |x_1'| + |v x_1' - (1+v^2)^{1/2} x_2'|)^{-2m},$$

where again, the constant C does not depend on Q. 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. Then from standard wavelet estimates, it follows that one can find a constant C such that

$$|K_w(\gamma, \gamma')| \leq C\, 2^{-j'n} 2^{-jn} (1+v^2)^{n+1/2} (1 + |b'| + |vb' - (1+v^2)^{1/2} b|)^{-2m},$$

which implies the desired result.

Case (ii). Let us consider the function

$$g_\gamma(x_1) = \int_{\mathbf{R}} \psi_j(u_1 x_1 - u_2 x_2 - b) w_Q(x_1, x_2)\, dx_2,$$

so that $K_w(\gamma,\gamma') = \int \psi(2^{j'}(x_1 - b'))g_\gamma(x_1)dx_1$. For any n, there is a constant C such that for any $\ell \le n$

$$\left|\frac{d^\ell}{dx_1^\ell}\psi_j(u_1x_1 - u_2x_2 - b)\right| \le C\, 2^{j(\ell+1)}(1 + 2^j|u_1x_1 - u_2x_2 - b|)^{-m}$$

and

$$\left|\frac{\partial^{n-\ell}}{\partial x_1^{n-\ell}}w(x_1,x_2)\right| \le C\,(1 + |x_1| + |x_2|)^{-2m} \le C\,(1+|x_1|)^{-m}(1+|x_2|)^{-m}.$$

Now the result from Chapter 4 guarantees that

$$\left|\frac{d^n}{dx_1^n}g_\gamma(x_1)\right| \le C\, 2^{j(n+1)}(1+|x_1|)^{-m}(1+2^j|u_1x_1 - b|)^{-m}.$$

And again standard wavelet estimates give in this case:

$$\begin{aligned}
|K_w(\gamma,\gamma')| &= \left|\int \psi_{j'}(x_1 - b')g_\gamma(x_1)\,dx_1\right| \\
&\le C\, 2^{-j'n}2^{j(n+1)}(1+|b'|)^{-m}(1+2^j|u_1b' - b|)^{-m} \\
&\le C\, 2^{-j'n}2^{j(n+1)}(1+|b'|)^{-m}\left(1 + 2^j\left|b' - \frac{(1+v^2)^{1/2}}{v}b\right|\right)^{-m}.
\end{aligned}$$

The lemma is proved. ∎

Corollary 4 *Let $j' \ge j$ and assume $n < \min(R, D)$. Then, there is a constant C depending on n and p so that:*

(i) Suppose $2^j(1+v^2)^{-1/2} \ge 1$.

$$\begin{aligned}
\left(\int |K_w(\gamma,\gamma')|^p\,db'\right)^{1/p} &\le C\, 2^{-j'n}2^{-jn}(1+v^2)^{n+1/2-1/2p}, \\
\left(\int |K_w(\gamma,\gamma')|^p\,db\right)^{1/p} &\le C\, 2^{-j'n}2^{-jn}(1+v^2)^{n+1/2-1/2p}.
\end{aligned}$$

(ii) Case $2^j(1+v^2)^{-1/2} \leq 1$.

$$\begin{aligned}\left(\int |K_w(\gamma,\gamma')|^p\, db'\right)^{1/p} &\leq C\, 2^{-j'n} 2^{j(n+1)} 2^{-j/p},\\ \left(\int |K_w(\gamma,\gamma')|^p\, db\right)^{1/p} &\leq C\, 2^{-j'n} 2^{j(n+1)} 2^{-j/p}.\end{aligned}$$

Proof of Corollary. The proof is absolutely straightforward as one just needs to integrate the upper estimates obtained in Lemma 10.

Corollary 5 *Let* $j' \geq j$ *and suppose* $n < \min(R, D)$. *For* $2n > d/p - 1$, *there is a constant* C *(depending on* n *and* p*) so that*

$$\begin{aligned}\left(\int |K_\Phi(\gamma,\gamma')|^p\, db'\, du'\right)^{1/p} &\leq C\, 2^{-(j'-j)n} 2^j 2^{-jd/p},\\ \left(\int |K_\Phi(\gamma,\gamma')|^p\, db\, du\right)^{1/p} &\leq C\, 2^{-(j'-j)n} 2^j 2^{-jd/p}.\end{aligned}$$

Proof of Corollary. It is not hard to show that

$$\int_{(1+v^2)^{1/2}\leq 2^j} (1+v^2)^{(n+1/2)p-1/2} \frac{dv}{(1+v^2)^{d/2}} \leq C_{n,p} 2^{j[(2n+1)p-d]},$$

if n is large enough. On the other hand,

$$\int_{(1+v^2)^{1/2}\geq 2^j} \frac{dv}{(1+v^2)^{d/2}} \leq C\, 2^{-j(d-1)}.$$

Combining these two results we have

$$\left(\int |K_w(\gamma,\gamma')|^p\, db'\, du'\right)^{1/p} \leq C\, 2^{-(j'-j)n} 2^j 2^{-jd/p} + C\, 2^{-(j'-j)n} 2^j 2^{-jd/p},$$

which established the first result. The second result is similar in every point. ∎

From Corollary 5, one can deduce the result announced in Chapter 4 (Lemma 7).

Proposition 9 *Let* $s > 0$ *and* $0 < p, q \leq \infty$. *For any* $j \geq 0$, $u \in \mathbf{S}^{d-1}$, $b \in \mathbf{R}$, *if* $\min(R, D)$ *is large enough, then*

$$\|\psi_j(u \cdot x - b)\|_{R^s_{p,q}(\Omega_d)} \leq C\, 2^{js} 2^{j/2} 2^{jd(1/2-1/p)},$$

for some constant C not depending on j, u, b.

Proof of Proposition. Let w be a nonnegative C^∞ window supported on $2\Omega_d$ such that $w_{|\Omega_d} = 1$. It suffices to prove that $\|w(x)\psi_j(u \cdot x - b)\|_{R^s_{p,q}(\mathbf{R}^d)} = O(2^{js}2^{j/2}2^{jd(1/2-1/p)})$. But

$$\|w(x)\psi_j(u \cdot x - b)\|_{R^s_{p,q}} = \left(\int |K_w(\gamma, \gamma')|^p \, db' \, du' \right)^{1/p} + \left\{ 2^{j's} 2^{j'(d-1)/2} \left(\int |K_w(\gamma, \gamma')|^p \, db' \, du' \right)^{1/p} \right\}_{\ell_q}$$

and applying our estimates (Corollary 5) gives

$$\begin{aligned} &\|w(x)\psi_j(u \cdot x - b)\|_{R^s_{p,q}} \\ &\leq C \left(2^{-jnq} + \sum_{j' \geq 0} (2^{j's} 2^{j'(d-1)/2} 2^{-|j-j'|n} 2^{\min(j,j')(1-d/p)})^q \right) \\ &\leq C 2^{j(s+(d-1)/2+1-d/p)q} \left(\sum_{j' \leq j} 2^{-(j-j')(n+s+(d-1)/2+1-d/p)q} + \sum_{j'>j} 2^{-(j'-j)(n-s-(d-1)/2)q} \right) \\ &\leq \left[C_q 2^{js} 2^{j/2} 2^{jd(1/2-1/p)} \right]^q, \end{aligned}$$

where the last inequality holds as long as $n > s + (d-1)/2$. As a conclusion, we have that

$$\|w(x)\psi_j(u \cdot x - b)\| \leq C 2^{js} 2^{j/2} 2^{jd(1/2-1/p)},$$

as expected.

5.2.2 Embedded Hypercubes

For the frame construction, the discretization of the directional variable u involved the covering of the sphere by balls. Indeed, at a given scale j, the discrete set of directions Σ_j is a covering set of the sphere: a set such that the collection of balls centered at the set points and of radius proportional to 2^{-j} cover the whole sphere $\mathcal{S}^{d-1}$. We have discussed the importance of good covering sets and their relationship to the control of the frame bounds ratio.

Inversely, the lower bounds involve properties of packing sets of the sphere: for a fixed $\epsilon > 0$, how can we distribute points on the sphere such that balls of radius ϵ and centered

at these points don't overlap? The maximum number of points we can distribute is called the packing number. Again, there is a considerable literature (Conway and Sloane, 1988) on this matter that the reader can refer to. In the sequel, we shall only make use of trivial facts about this packing problem.

The purpose of this section is to establish a technical fact that would guarantee that a certain functional class cannot be approximated by finite linear combinations of a given dictionary $\mathcal{D}$ faster than a certain rate. It is important to note that the proof of this fact (or in other words of the existence of lower bounds of approximation) does not need to be constructive. This observation greatly simplifies our argument.

For a fixed j, let S_j be a set of points on the sphere $\{u_i\}$ satisfying the following properties.

(i) $\forall u_i, u_{i'} \in S_j, \quad \|u_i \pm u_{i'}\| \geq 2^{-(j-j_0)}$.

(ii) $|S_j| \geq B_1 \, 2^{(j-j_0)(d-1)}$.

(iii) For any $u \in \mathcal{S}^{d-1}$, and all $0 \leq l \leq j - j_0$,

$$\begin{aligned} |\{u_i, \;\; 0 \leq \frac{|u \cdot u_i|}{(1-(u \cdot u_i)^2)^{1/2}} \leq 1\}| &\leq B_2 \, 2^{(j-j_0)(d-1)} \int_{|v| \leq 1} \frac{dv}{(1+v^2)^{d/2}} \\ |\{u_i, \;\; 2^{l-1} \leq \frac{|u \cdot u_i|}{(1-(u \cdot u_i)^2)^{1/2}} \leq 2^l\}| &\leq B_2 \, 2^{(j-j_0)(d-1)} \int_{2^{l-1} \leq |v| \leq 2^l} \frac{dv}{(1+v^2)^{d/2}} \end{aligned}$$

In the above expression, the constants B_1 and B_2 can be chosen to not depend on j and j_0. We remark that the first property (i) implies that

$$\{u_{i'}, \;\; \frac{|u_i \cdot u_{i'}|}{(1-(u_i \cdot u_{i'})^2)^{1/2}} \geq 2^{j-j_0}\} = \{u_i\}.$$

This fact is a mere consequence of

$$\|u_{i'} \pm u_i\|^2 = 2(1 \pm u_i \cdot u_{i'}).$$

Indeed let, $v_{i,i'} = u_i \cdot u_{i'}(1-(u_i \cdot u_{i'})^2)^{-1/2}$. Suppose for instance that $v_{i,i'} \geq 2^{j-j_0}$. We have

$$\begin{aligned} \|u_{i'} - u_i\|^2 &= 2\left(1 - \frac{v_{i,i'}}{(1+v_{i,i'}^2)^{1/2}}\right) \\ &= 2\frac{1}{(1+v_{i,i'}^2)^{1/2}(v_{i,i'} + (1+v_{i,i'}^2)^{1/2})} \leq \frac{1}{(1+v_{i,i'}^2)}. \end{aligned}$$

Therefore, $v_{i,i'} \geq 2^{j-j_0}$ implies $\|u_{i'} - u_i\| < 2^{-(j-j_0)}$. From (i), it follows that it is equivalent to $i = i'$. The argument is identical in the case $v_{i,i'} \leq -2^{j-j_0}$.

To simplify the analysis, suppose $\psi \in \mathcal{S}(\mathbf{R})$ compactly supported supp $\psi \subset [-1/2, 1/2]$, and has a sufficiently large number of vanishing moments. We normalize ψ such that $\|\psi\|_2 = 1$. Further, let $w \in C_0^\infty(\Omega_d)$ be a radial window such that $0 \leq w \leq 1$ and $w(x) = 1$ for any x with $\|x\| \leq \sqrt{3}/2$. We now consider the set A_j of windowed ridgelets at scale j

$$A_j = \{f_{i,k}(x) = 2^{j/2}\psi(2^j u_i \cdot x - k)w(x), \;\; u_i \in S_j, \; k \in \mathbf{Z} \;\text{ and }\; |k|2^{-j} \leq 1/2\}. \tag{5.8}$$

Finally, we will assume $j \geq 2$ so that $1/2 + 2^{-j}/2 \leq \sqrt{2}/2$; from our assumptions it follows that supp $f_{i,k} \subset \{x, \;\; |u_i \cdot x| \leq \sqrt{2}/2\}$ for any $f_{i,k}$ in A_j.

We show that if j_0 is large enough then the elements of A_j are "almost" orthogonal. That is, we prove the following result.

Lemma 11 *(i) There is a constant c_d (only depending upon the dimension d) s.t.*

$$\forall \;\; f \in A_j, \qquad \|f\|_2 \geq c_d. \tag{5.9}$$

(ii) If j_0 is chosen large enough,

$$\forall \;\; f \in A_j, \qquad \sum_{g \in A_j, g \neq f} |\langle f, g \rangle| \leq \frac{c_d}{2}. \tag{5.10}$$

Proof of Lemma. The norm of $f_{i,k}$ being clearly invariant by rotation (w radial), one can

assume that $u_i = e_1$. We have

$$\begin{aligned}
&\int 2^j \psi^2(2^j(x_1 - k2^{-j}))\; w^2(x)\, dx \\
&\geq \int_{|x_1|\leq\sqrt{2}/2} \int_{x_2^2+\ldots x_d^2\leq(1/2)^2} 2^j \psi^2(2^j(x_1 - k2^{-j}))\; w^2(x)\, dx_1 dx_2 \ldots dx_d \\
&\geq \int_{|x_1|\leq\sqrt{2}/2} 2^j \psi^2(2^j(x_1 - k2^{-j}))dx_1 \int_{x_2^2+\ldots x_d^2\leq(1/2)^2} 1\, dx_1 dx_2 \ldots dx_d \\
&= \|\psi\|_2^2\; c_d \;=\; c_d,
\end{aligned}$$

where c_d might be chosen to be the volume of a $d-1$ dimensional ball of radius $1/2$. This proves (i).

Before proceeding further, observe that if $0 < \eta \leq \epsilon \leq 1$, $x \in \mathbf{R}$, $y \in \mathbf{R}$, and $\delta > 0$ we have

$$\sum_{k\in\mathbf{Z}} (1 + |x - \epsilon k|)^{-1-\delta}(1 + |y - \eta k|)^{-1-\delta} \leq C_\delta \epsilon^{-1}(1 + |y - x\eta\epsilon^{-1}|)^{-1-\delta}. \tag{5.11}$$

By construction, it is pretty clear that the supports of $\psi(2^j u_i \cdot x - k)$ and $\psi(2^j u_i \cdot x - k')$ do not overlap when $k \neq k'$. Therefore,

$$\sum_{k', k'\neq k} |\langle f_{i,k}, f_{i,k'}\rangle| = 0.$$

Next, a simple application of our fundamental estimates when $u_i \neq u_i'$ shows that one can find a constant $C_1(d)$ depending on d, ψ and w such that (note that in this case $(1 + v_{i,i'}^2)^{1/2} \leq 2^j$),

$$|\langle f_{i,k}, f_{i',k'}\rangle| \leq C_1(d)\; 2^{-j(2d+1)}(1 + v_{i,i'}^2)^{\frac{2d+1}{2}}(1 + 2^{-j}|v_{i,i'}k - (1 + v_{i,i'}^2)^{1/2}k'|)^{-2}$$

Now, it follows from (5.11) that

$$\begin{aligned}
\sum_{k'} |\langle f_{i,k}, f_{i',k'}\rangle| &\leq C_2(d)\; 2^{-j(2d+1)}(1 + v_{i,i'}^2)^{\frac{2d+1}{2}} 2^j (1 + v_{i,i'}^2)^{-1/2} \\
&= C_2(d)\; 2^{-2jd}(1 + v_{i,i'}^2)^d
\end{aligned}$$

for some new constant $C_2(d)$, depending only on d, ψ and w. Summing over $u_{i'}$ $(u_{i'} \neq u_i)$

and making use of our assumption (iii) gives

$$\begin{aligned}
\sum_{f_{i',k'}\in A_j, f_{i',k'}\neq f_{i,k}} |\langle f_{i,k}, f_{i',k'}\rangle| &= \sum_{u_{i'}, u_{i'}\neq u_i} \sum_{k'} |\langle f_{i,k}, f_{i',k'}\rangle| \\
&\leq C_2(d)\ 2^{-2jd} \sum_{\ell=0}^{j-j_0} (1+2^{2\ell})^d |\{u_{i'},\ \ 2^{\ell-1} \leq |v_{i,i'}| \leq 2^\ell\}| \\
&\leq C_2(d)\ 2^{-2jd} B_2 2^{(j-j_0)(d-1)} \sum_{\ell=0}^{j-j_0} (1+2^{2\ell})^d \int_{2^{\ell-1}\leq|v|\leq 2^\ell} \frac{dv}{(1+v^2)^{d/2}} \\
&\leq C_3(d)\ 2^{(j-j_0)(d-1)} 2^{-2jd} \sum_{\ell=0}^{j-j_0} 2^{\ell(2d+1-d)} \\
&\leq C_4(d)\ 2^{(j-j_0)(d-1)} 2^{-2jd} 2^{(j-j_0)(2d-(d-1))} \\
&= C_4(d)\ 2^{-j_0 2d},
\end{aligned}$$

where again $C_4(d)$ is a new constant $C(d, \psi, w)$. (Notice that we have sacrificed exactness for synthetic notations: in the second line of the array, read $|\{u_{i'}, 0 \leq |v_{i,i'}| \leq 1\}|$ instead of $|\{u_{i'}, 2^{\ell-1} \leq |v_{i,i'}| \leq 2^\ell\}|$ when the index ℓ equals 0.) Therefore by choosing j_0 large enough, one can make sure that the quantity $C_d\ 2^{-j_0 2d}$ is dominated by c_d, which proves (ii). ∎

Lemma 12 *Let $\mathcal{C}$ the parallelepiped be defined by*

$$\mathcal{C} = \{f,\ \ f = \sum_{i,k} \xi_{i,k} f_{i,k},\ \ |\xi_{i,k}| \leq 1\}. \tag{5.12}$$

Then, for any f in $\mathcal{C}$ and triplet s, p, q; $s > 0, 0 < p, q \leq \infty$, we have

$$\|f\|_{R^s_{p,q}(\Omega_d)} \leq C\ 2^{js} 2^{jd/2},$$

where the constant C depends at most on s, p, q, ψ, w and the dimension d.

Proof of Lemma. Let f be defined by $f = \sum_{i,k} \xi_{i,k} f_{i,k}$ and let us calculate the norm of f. First notice that since supp $f \subset \Omega_d$, then $\|f\|_{R^s_{p,q}(\Omega_d)} = \|f\|_{R^s_{p,q}(\mathbf{R}^d)}$. Further, since $\|f\|_{R^s_{p,q}} \leq \|f\|_{R^s_{p,q'}}$ whenever $q' \leq q$, we only need to prove the result for say $0 < q \leq 1$. Finally, as to simplify notations, we show the result for $p = 1$; as the argument for $p \neq 1$ is absolutely parallel, we will just indicate where to modify the proof to handle this case. We

have

$$\begin{aligned}
|w_{j'}(u',b')(f)| &= \Big| \int \sum_{i,k} \xi_{i,k} f_{i,k}(x) 2^{j'} \psi(2^{j'}(u' \cdot x - b')) dx \Big| \\
&\leq \sum_{i,k} |\xi_{i,k}| \, \Big| \int 2^{j/2} \psi(2^j (u_i \cdot x - k2^{-j}))\, 2^{j'} \psi(2^{j'}(u' \cdot x - b')) w(x) dx \Big| \\
&\leq \sum_{i,k} \Big| \int 2^{j/2} \psi(2^j (u_i \cdot x - k2^{-j}))\, 2^{j'} \psi(2^{j'}(u' \cdot x - b')) w(x) dx \Big|
\end{aligned}$$

To be consistent with the preceding section and to simplify the notations, let $K_w(\gamma_{i,k}, \gamma')$ be defined as

$$K_w(\gamma_{i,k}, \gamma') = \int 2^j \psi(2^j (u_i \cdot x - k2^{-j}))\, 2^{j'} \psi(2^{j'}(u' \cdot x - b')) w(x) dx.$$

(Because of an exigence of consistency, we want to bring to the reader's attention that the normalization has changed, i.e. $2^{j/2}$ has been replaced by 2^j.) Once again we apply our fundamental estimates. Let v_i be $u_i \cdot u'/(1 - (u_i \cdot u')^2)^{1/2}$. Suppose for now that $j' \geq j$, we have for $v_i \leq 2^j$

$$|K_w(\gamma_{i,k}, \gamma')| \leq C 2^{-j'n} 2^{-jn} (1 + v_i^2)^{n+1/2} (1 + |b'|)^{-2} (1 + |v_i b' - (1 + v_i^2)^{1/2} k 2^{-j}|)^{-2}.$$

And summing over k gives

$$\begin{aligned}
\sum_k |K_w(\gamma_{i,k}, \gamma')| &\leq 2^{-j'n} 2^{-jn} (1 + v_i^2)^{n+1/2} (1 + |b'|)^{-2} 2^j (1 + v_i^2)^{-1/2} \\
&= 2^{-j'n} 2^{-j(n-1)} (1 + v_i^2)^n (1 + |b'|)^{-2}.
\end{aligned}$$

Recall our assumption (ii) states that for $\ell > 0$,

$$|\{u_i, \;\; 2^{\ell-1} \leq |v_i| \leq 2^\ell\}| \leq B_2 \, 2^{(j-j_0)(d-1)} \int_{2^{\ell-1} \leq |v| \leq 2^\ell} \frac{dv}{(1 + v^2)^{d/2}},$$

and, similarly, for the interval $0 \leq |v| \leq 1$. Repeating the argument of the previous lemma

gives

$$\begin{aligned}\sum_{i,|v_i|\leq 2^{j-j_0}} \sum_k |K_w(\gamma_{i,k},\gamma')| &\leq C2^{-j'n}2^{-j(n-1)}B2^{(j-j_0)(d-1)} \sum_{\ell=0}^{j-j_0} 2^{\ell(2n-(d-1))}(1+|b'|)^{-2} \\ &\leq C2^{-j'n}2^{jn}2^{j}(1+|b'|)^{-2}.\end{aligned}$$

Now for $v_i \geq 2^{j-j_0}$, since our estimate for $v_i \geq 2^j$ dominates the one for $v_i \leq 2^j$, one can check that

$$|K_w(\gamma_{i,k},\gamma')| \leq C2^{-j'n}2^{j(n+1)}(1+|b'|)^{-2}\left(1+2^j\left|b'-\frac{(1+v_i^2)^{1/2}}{v_i}k2^{-j}\right|\right)^{-2}$$

and, again, summing over k yields

$$\sum_k |K_w(\gamma_{i,k},\gamma')| \leq C2^{-j'n}2^{j(n+1)}(1+|b'|)^{-2}.$$

Now, we know that $|v_i| \geq 2^{j-j_0}$ implies $\|u'-u_i\| \leq 2^{-(j-j_0)}$ and, similarly, when the index i is replaced by i'. Suppose $S = \{i, \quad |v_i| \geq 2^{j-j_0}\} \neq \emptyset$ and let i_0 be one of its elements. Then, by the triangular inequality $|v_i| \geq 2^{j-j_0} \implies \|u_{i_0}-u_i\| \leq 2\ 2^{j-j_0}$, it follows that the cardinality of S can be bounded by a constant depending at most on the dimension d. Therefore, we have

$$\sum_{i,|v_i|\geq 2^{j-j_0}} \sum_k |K_w(\gamma_{i,k},\gamma')| \leq C2^{-j'n}2^{j(n+1)}(1+|b'|)^{-2}.$$

Combining these two estimates yields

$$\sum_i \sum_k |K_w(\gamma_{i,k},\gamma')| \leq C2^{-(j'-j)n}2^{j}(1+|b'|)^{-2}.$$

Similarly, for $j' \leq j$, an identical reasoning allows getting similar upper bounds (suppose $n > d$). First, assume that $|v_i| \leq 2^{j'}$. From

$$|K_w(\gamma_{i,k},\gamma')| \leq C\ 2^{-jn}2^{-j'n}(1+v_i^2)^{n+1/2}(1+|k2^{-j}|)^{-2}(1+|v_ik2^{-j}-(1+v_i^2)^{1/2}b'|)^{-2}.$$

one can deduce (applying (5.11)) that

$$\sum_k |K_w(\gamma_{i,k}, \gamma')| \leq C\ 2^{-jn} 2^{-j'n} 2^j (1 + v_i^2)^n (1 + |b'|)^{-2}.$$

And therefore, already well-known arguments imply that

$$\sum_{i,\, |v_i| \leq 2^{j'}} \sum_k |K_w(\gamma_{i,k}, \gamma')| \leq C\ 2^{-(j-j')(n-(d-1))} 2^j (1 + |b'|)^{-2}.$$

Similarly, in the case where $v_i \geq 2^{j'}$, we have

$$|K_w(\gamma_{i,k}, \gamma')| \leq C 2^{-jn} 2^{j'(n+1)} (1 + |k2^{-j}|)^{-2} \left(1 + 2^j \left| k2^{-j} - \frac{(1 + v_i^2)^{1/2}}{v_i} b' \right| \right)^{-2},$$

giving

$$\sum_k |K_w(\gamma_{i,k}, \gamma')| \leq C\ 2^{-jn} 2^{j'(n+1)} (1 + |b'|)^{-2}.$$

Again, summing over the i's yields

$$\sum_{i; |v_i| \geq 2^{j'}} \sum_k |K_w(\gamma_{i,k}, \gamma')| 2^{-(j-j')(n-d)} 2^j (1 + |b'|)^{-2}.$$

Combining the results, we have

$$\sum_i \sum_k |K_w(\gamma_{i,k}, \gamma')| \leq C 2^{-(j-j')(n-d)} 2^j (1 + |b'|)^{-2}.$$

The following display summarizes the situation:

$$|w_{j'}(u', b')(f)| \leq \begin{cases} C 2^{-(j-j')(n-d)} 2^{j/2} (1 + |b'|)^{-2} & j' \leq j \\ C 2^{-(j'-j)n} 2^{j/2} (1 + |b'|)^{-2} & j' \geq j \end{cases}. \tag{5.13}$$

In the case $p < 1$, we just need to replace $(1 + |b'|)^{-2}$ with $(1 + |b'|)^{-2/p}$ at each of its occurence in (5.13) as it follows from the same argument using, however, Lemma 10 with sharper bounds. A direct consequence of the above inequalities together with the preceding

comment gives that for any $p > 0$

$$\left(\int |w_{j'}(u',b')(f)|^p \; db'du' \right)^{1/p} \leq \left\{ \begin{array}{ll} C2^{-(j-j')(n-d)}2^{j/2} & j' \leq j \\ C2^{-(j'-j)n}2^{j/2} & j' \geq j \end{array} \right. .$$

As long as n is chosen to be greater than $\max(s + (d-1)/2, d)$, it is immediate to establish from here that for any $q > 0$,

$$\left(\int |v(u',b')(f)|^p \; db'du' \right)^{1/p} + \left\{ 2^{j's} 2^{j' \frac{d-1}{2}} \left(\int |w_{j'}(u',b')(f)|^p \; db'du' \right)^{1/p} \right\}_{\ell_q} \leq C 2^{jd(\frac{s}{d}+\frac{1}{2})},$$

which is our claim. ∎

Note that the previous lemma shows how to construct a full parallelepiped embedded in $R^s_{p,q}(\Omega_d)$. However, in view of Theorem 10 one needs to construct a cube. The next lemma shows how to orthogonalize our parallelepiped.

Lemma 13 *Suppose we have d vectors $\{f_i\}_{1 \leq i \leq d}$ in a Hilbert space such that for all $1 \leq i \leq d$*

(i) $\|f_i\| = 1$,

(ii) $\sum_{j \neq i} |\langle f_i, f_j \rangle| \leq 1 - \delta < 1$.

We consider the set $\mathcal{C} = \{\sum_{i=1}^d y_i f_i, \;\; \|y\|_\infty \leq 1\}$. Then there exists a hypercube $\mathcal{H}$ of sidelength δ that is included in $\mathcal{C}$.

Proof of Lemma. Let us consider the symmetric matrix G defined by $G_{i,j} = \langle f_i, f_j \rangle$. Applying the Gershgorin Theorem, we deduce from the hypotheses (i) and (ii) that all the eigenvalues of G must be greater or equal to δ. Therefore G is a positive definite matrix and we can talk about $H = G^{-1/2}$. It is an easy exercise to see that the collection of vectors $\{e_i\}_{1 \leq i \leq d}$ defined by $e_i = Hf_i$ is indeed an orthogonal basis of $\mathrm{span}(\{f_i\}_{1 \leq i \leq d})$ (See Meyer Meyer, 1992, Page 25 for a proof.) Furthermore, a trivial fact states that

$$\sum_i x_i e_i = \sum_i x'_i f_i \quad \text{whenever} \quad x' = Hx.$$

Thus the embedding problem becomes: show that $\|x\|_\infty \leq \delta \implies \|Hx\|_\infty \leq 1$. This is nothing else but to prove that the norm of H, as an operator from $\ell_\infty \to \ell_\infty$, is bounded

by δ^{-1}. Recall,

$$\|H\|_{(\ell_\infty,\ell_\infty)} = \sup_i \sum_j |H_{i,j}|.$$

We now derive an upperbound of $\|H\|_{(\ell_\infty,\ell_\infty)}$. We have

$$H = \frac{1}{\pi}\int_0^\infty (G+\lambda I)^{-1}\lambda^{-1/2}d\lambda,$$

(see Meyer for a justification of this fact). The previous relationship implies that

$$\|H\|_{(\ell_\infty,\ell_\infty)} \leq \frac{1}{\pi}\int_0^\infty \|(G+\lambda I)^{-1}\|_{(\ell_\infty,\ell_\infty)}\lambda^{-1/2}d\lambda.$$

Now $G = I - F$, $G + \lambda I = (1+\lambda)I - F = (1+\lambda)(I - (1+\lambda)^{-1}F)$. The standard inversion formula for matrices (Neuman series) states

$$(G+\lambda I)^{-1} = (1+\lambda)^{-1}(I + \sum_{k\geq 1}(1+\lambda)^{-k}F^k),$$

which gives

$$\begin{aligned}
\|(G+\lambda I)^{-1}\|_{(\ell_\infty,\ell_\infty)} &\leq (1+\lambda)^{-1}(\|I\|_{(\ell_\infty,\ell_\infty)} + \sum_{k\geq 1}(1+\lambda)^{-k}\|F^k\|_{(\ell_\infty,\ell_\infty)}) \\
&\leq (1+\lambda)^{-1}(1 + \sum_{k\geq 1}(1+\lambda)^{-k}\|F\|^k_{(\ell_\infty,\ell_\infty)}) \\
&\leq (1+\lambda)^{-1}\frac{1}{1-\|F\|_{(\ell_\infty,\ell_\infty)}}.
\end{aligned}$$

Finally,

$$\|H\|_{(\ell_\infty,\ell_\infty)} \leq \frac{1}{\pi}\int_0^\infty (1-\|F\|_{(\ell_\infty,\ell_\infty)})^{-1}(1+\lambda)^{-1}\lambda^{-1/2}d\lambda = (1-\|F\|_{(\ell_\infty,\ell_\infty)})^{-1}.$$

By assumption we have $\|F\|_{(\ell_\infty,\ell_\infty)} \leq 1-\delta$ implying $\|H\|_{(\ell_\infty,\ell_\infty)} \leq \delta^{-1}$, which is precisely what needed to be proved. ∎

We are now in a position to state the main theorem of this section.

Theorem 11 *Suppose $1 \leq p, q \leq \infty$. Let $R^s_{p,q}(1)$ be the unit ball of $R^s_{p,q}(2\Omega_d)$. Then there exists an hypercube of sidelength δ and dimension $m(\delta)$ embedded in $R^s_{p,q}(1)$. That is we*

can find $m(\delta)$ *orthogonal functions* $g_{i,m,\delta}$, $i = 1, \ldots, m(\delta)$ *with* $\|g_{i,m,\delta}\|_2 = \delta$, *such that*

$$\mathcal{H}(\delta; \{g_i\}) = \{f, \quad f = \sum_i \xi_i g_{i,m,\delta}, \ |\xi| \leq 1\} \quad \subset R^s_{p,q}(1),$$

and

$$m(\delta) \geq K \delta^{-\frac{1}{s/d+1/2}}.$$

Proof of Theorem. The proof is a mere consequence of the three preceding preparatory lemmas. In view of Theorem 10, this is exactly what needed to be shown to prove the first part of Theorem 9.

5.3 Upper Bounds

In this thesis, we will prove the upper bound (5.6) for the classes $R^s_{p,q}$ in the following cases: $(s,p,q) \in \{(s,1,q), (s,2,2), (s,\infty,q), \ q > 0\}$, where this set is of course subject to the additional condition $s > d(1/p - 1/2)_+$ as it is a required hypothesis of our theorem.

5.3.1 A Norm Inequality

Let a frame be constructed as in Chapter 3, section 3.6.1. We define a discrete norm on the frame coefficients $\alpha_\gamma = \langle f, \psi_\gamma \rangle$ as follows

$$\|\alpha\|_{\mathbf{r}^s_{p,q}} \equiv \|(2^{js} 2^{jd(1/2-1/p)} (\sum_{\gamma \in \Gamma_j} |\alpha_\gamma|^p)^{1/p})_{j \geq -1}\|_{\ell_q}. \tag{5.14}$$

The norm (5.14) is the discretized version of the continuous norm (4.3).

For compactly supported functions, we have the following lemma.

Lemma 14 *There is a constant* C *possibly depending on* s, p, q *such that*

$$\|\alpha\|_{\mathbf{r}^s_{p,q}} \leq C \|f\|_{R^s_{p,q}}. \tag{5.15}$$

We prove the lemma for the subset of triplets (s,p,q) defined in the first paragraph of this section.

Proof of Lemma. Without loss of generality, we may assume that supp $f \subset \Omega_d$. For a

frame of $L_2(\Omega_d)$, let A be the Analysis operator defined by $Af = (\alpha_\gamma)_{\gamma_d \in \Gamma_d}$. To prove the lemma one has to prove that

$$\|Af\|_{\mathbf{r}^s_{p,q}} \leq C \|f\|_{R^s_{p,q}}.$$

Case $p = 1$. We first treat the case $p = 1$. Introducing a slight change of notations, let $R_j(u,b)$ be the ridgelet coefficient defined by $R_j(u,b)(f) = \langle f, 2^{j/2}\psi(2^j(u \cdot x - b))\rangle$ (with the convention that $R_{-1}(u,b)(f) = \langle f, \varphi(u \cdot x - b)\rangle$). Suppose that w is a nice C^∞ function, $0 \leq w \leq 1$, supp $w \subset 2\Omega_d$ and $w_{|\Omega_d} = 1$. The trivial equality $f = fw$ together with the semi-continuous reproducing formula (2.8) imply that

$$\langle f, \psi_\gamma \rangle \propto \int R_{-1}(u',b')(f)\langle \phi(u' \cdot x - b'), w\psi_\gamma\rangle du'db' + \sum_{j' \geq 0} 2^{j'd} \int R_{j'}(u',b')(f)\langle 2^{j'/2}\psi(2^{j'}(u' \cdot x - b')), w\psi_\gamma\rangle du'db'.$$

In view of the Parseval relationship (Proposition 2), the above relationship is fully justified as the membership to $R^s_{p,q}$ for $s > d(1/p - 1/2)_+$ implies (for compactly supported functions) membership to L_2 (see last paragraph of section 4.2.1). Then, observe that

$$\sum_{\gamma \in \Gamma_j} |\alpha_\gamma| \leq \int |R_{-1}(u',b')(f)| \sum_{\gamma \in \Gamma_j} |\langle \phi(u' \cdot x - b'), w\psi_\gamma\rangle| du'db' + \sum_{j' \geq 0} 2^{j'd} \int |R_{j'}(u',b')(f)| \sum_{\gamma \in \Gamma_j} |\langle 2^{j'/2}\psi(2^{j'}(u' \cdot x - b')), w\psi_\gamma\rangle| du'db'$$

Now apply a properly renormalized version of (5.13) to get

$$\sum_{\gamma \in \Gamma_j} |\alpha_\gamma| \leq \sum_{-1 \leq j' \leq j} 2^{j'd} 2^{-(j-j')(n-d-1/2)} \int |R_{j'}(u',b')(f)| du'db' + \sum_{j' > j} 2^{j'd} 2^{-(j-j')(n+1/2)} \int |R_{j'}(u',b')(f)| du'db'$$

Then,

$$2^{js}2^{-jd/2}\sum_{\gamma\in\Gamma_j}|\alpha_\gamma| \le \sum_{-1\le j'\le j} 2^{-(j-j')(n-d/2-s-1/2)}2^{j's}2^{j'd/2}\int |R_{j'}(u',b')(f)|du'db' + \sum_{j'>j} 2^{-(j-j')(n+d/2-s-1/2)}2^{j's}2^{j'd/2}\int |R_{j'}(u',b')(f)|du'db'$$

and from there, it is then easy to see that for any $q>0$

$$\|\alpha\|_{\mathbf{r}^s_{1,q}}\|(2^{js}2^{-jd/2}\sum_{\gamma\in\Gamma_j}|\alpha_\gamma|)_j\|_{\ell_q} \le \|(2^{j's}2^{j'd/2}\int |R_{j'}(u',b')(f)|du'db')_{j'}\|_{\ell_q} = \|f\|_{R^s_{1,q}}.$$

(The cases where $q\le 1$ and $q=\infty$ are clear, and interpolation does the rest.)

Case $p=\infty$. In this case, we have

$$\begin{aligned}
|\alpha_\gamma| &\le \int |R_{-1}(u',b')(f)||\langle\varphi(u'\cdot x-b'),w(x)\psi_\gamma(x)\rangle|du'db' \\
&\qquad + \sum_{j'\ge 0}2^{j'd}\int |R_j(u',b')(f)||\langle 2^{j'/2}\psi(2^{j'}(u'\cdot x-b')),w(x)\psi_\gamma(x)\rangle|du'db' \\
&\le \|R_{-1}(f)\|_\infty\int|\langle\varphi(u'\cdot x-b'),w(x)\psi_\gamma(x)\rangle|du'db' \\
&\quad + \sum_{j'\ge 0}2^{j'd}\|R_{j'}(f)\|_\infty\int|\langle 2^{j'/2}\psi(2^{j'}(u'\cdot x-b')),w(x)\psi_\gamma(x)\rangle|du'db'
\end{aligned}$$

Recall the estimates relative to the decay of the kernel (Corollary 5)

$$\begin{aligned}
\int|\langle 2^{j'/2}\psi(2^{j'}(u'\cdot x-b')),w(x)\psi_\gamma(x)\rangle|du'db' &\le 2^{-|j-j'|(n+1/2)}2^{-\min(j,j')d} \\
&= 2^{-jd/2}2^{-j'd/2}2^{-|j-j'|(n-(d-1)/2)}.
\end{aligned}$$

Thus, making use of this result we obtain

$$\begin{aligned}
2^{js}2^{jd/2}|\alpha_\gamma| &\le \sum_{j'\ge -1}2^{js}2^{j'd/2}2^{-|j-j'|(n-(d-1)/2)}\|R_{j'}(f)\|_\infty \\
&\le \sum_{j'\ge -1}2^{j's}2^{j'd/2}2^{-|j-j'|(n-s-(d-1)/2)}\|R_{j'}(f)\|_\infty.
\end{aligned}$$

Now again, it is clear that the above inequality gives for any $q > 0$

$$\|\alpha\|_{\mathbf{r}^s_{\infty,q}} = \|(2^{js}2^{jd/2} \sup_{\gamma\in\Gamma_j} |\alpha_\gamma|)_j\|_{\ell_q} \leq \|(2^{j's}2^{j'd/2} \int |R_{j'}(u',b')(f)|du'db')_{j'}\|_{\ell_q} = \|f\|_{R^s_{\infty,q}}. \tag{5.16}$$

Case $p = q = 2$. This case is a direct consequence of Theorem 6. ∎

Extension. It seems reasonable to suspect that one can extend Lemma 14 by interpolation. By definition, the space $R^s_{p,q}$ is a weighted space of the type $\ell_q(L_p)$: using the notations of the lemma, one can rewrite the $R^s_{p,q}$ norm of an object as

$$\|f\|_{R^s_{p,q}} = \|2^{js}2^{jd/2}\|R_j(f)\|_{L_p(\mathbf{S}^{d-1}\times\mathbf{R})}\|_{\ell_q} = \left(\sum_{j\geq -1}\left[\int_{\mathbf{S}^{d-1}\times\mathbf{R}} |R_j(u,b)|^p dudb\right]^{q/p}\right)^{1/q}$$

with obvious modification in the case $p = \infty$ or/and $q = \infty$. (Again, there is a full analogy with the Besov scale; replace $R_j(f)$ with $f * \varphi_j$ and the weights $2^{js}2^{jd/2}$ with 2^{js}.) Similarly, the sequence space $\mathbf{r}^s_{p,q}$ is a weighted space of the type $\ell_q(\ell_p)$. From Lemma 14, we know that for any $s, q > 0$, the operator A is bounded from $R^s_{1,q}$ (respectively, $R^s_{\infty,q}$) to $\mathbf{r}^s_{1,q}$ (respectively, $\mathbf{r}^s_{\infty,q}$). Now, the sequence spaces $\mathbf{r}^s_{p,q}$ are clearly interpolation spaces with well-known properties and one expects the same interpolation properties to be true for the spaces $R^s_{p,q}$. If that were true, one would be able to prove Lemma 14 for the full range of parameters $s > 0$, $1 \leq p \leq \infty$, $q > 0$ – at least when $R^s_{p,q} \subset L_2(\Omega_d)$ – since bounds at the "corners" are already established. Work in this direction is in progress and we hope to report on it shortly.

Further, we have learned that in the case where $p = q = 2$, the norms $\|\cdot\|_{\mathbf{r}^s_{2,2}}$ and $\|\cdot\|_{R^s_{2,2}}$ were actually equivalent; therefore, this favors the conjecture we have just brought up and leads to a more ambitious one: namely, the norm equivalence between $\|\cdot\|_{\mathbf{r}^s_{p,q}}$ and $\|\cdot\|_{R^s_{p,q}}$.

Lemma 14 has a very interesting corollary. Let us consider a frame like in section 3.6.1 such that both φ and ψ are compactly supported and that ψ has enough vanishing moments so that Lemma 14 holds. The assumption of compact support will simplify the proof of the corollary but is not a necessary hypothesis.

Corollary 6 *Assume* $s > d(1/p - 1/2)_+$ *and let* p^* *be defined by* $1/p^* = s/d + 1/2$, *then*

for f in $R^s_{p,q}(\Omega_d)$ we have

$$\|\alpha\|_{w\ell_{p^*}} \leq C\|f\|_{R^s_{p,q}(\Omega_d)}, \tag{5.17}$$

where $w\ell_{p^*}$ is the weak-ℓ_{p^*} quasi-norm (5.20).

In fact, one should be a bit more careful in the above statement: for a function f in $R^s_{p,q}(\Omega_d)$, extend it to a function in $R^s_{p,q}(\mathbf{R}^d)$ supported in $2\Omega_d$; then take the sequence α to be the frame coefficients of the extended f (where the frame is for $L_2(2\Omega_d)$).

Proof of Corollary. The proof is a minor adaptation of a result in (Donoho, 1993). We do not reproduce it here. ∎

Now, the proof of the upper bound is almost complete as it is a direct consequence of the result presented in the next section.

5.3.2 A Jackson Inequality

Besides the approximation of functional classes, one can be interested in knowing achievable rates of approximation for a given target f. The next simple lemma establishes that the sparsity of the ridgelet coefficients imply a minimum rate of convergence (i.e. a lower bound on the exponent of approximation (7.1)) of truncated N-term expansions.

Suppose we are given a frame $(\psi_\gamma) \in \Gamma_d$. We know that any element of, say $L_2(\Omega_d)$ might be represented as

$$f = \sum_{\gamma\in\Gamma_d} \langle f, \psi_\gamma\rangle\tilde{\psi}_\gamma = \sum_{\gamma\in\Gamma_d} \langle f, \tilde{\psi}_\gamma\rangle\psi_\gamma. \tag{5.18}$$

To simplify, α_γ will denote the ridgelet coefficient $\langle f, \psi_\gamma\rangle$. Finally, $\tilde{f}_N$ will denote the N-term ridgelet expansion where one only keeps the terms corresponding to the N largest coefficients. That is

$$\tilde{f}_N(x) = \sum_{\gamma\in\Gamma_d} \alpha_\gamma 1_{\{|\alpha_\gamma|\geq|\alpha|_{(N)}\}}\tilde{\psi}_\gamma. \tag{5.19}$$

We recall that for a sequence (θ_n), 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

$$|\theta|_{w\ell_p} = \sup_{n>0} n^{1/p} |\theta|_{(n)}. \tag{5.20}$$

Lemma 15 *Let $0 < p < 2$ and suppose that the sequence of coefficients $(\alpha_\gamma)_{\gamma \in \Gamma_d}$ has a bounded weak-ℓ_p norm. Then there is a constant C (depending at most on p) such that*

$$\|f - \tilde{f}_N\|_2 \leq C N^{-r} |\alpha|_{w\ell_p}$$

with $r = 1/p - 1/2$.

Proof of Lemma. Let A (respectively S) be the analysis operator (respectively synthesis operator) defined by

$$\begin{array}{rcl} A: \ L_2(\Omega_d) & \to & \ell_2(\Gamma_d) \\ f & \mapsto & \langle f, \psi_\gamma \rangle \end{array} \qquad \begin{array}{l} S: \ \ell_2(\Gamma_d) \to L_2(\Omega_d) \\ \theta_\gamma \mapsto \sum_\gamma \theta_\gamma \tilde{\psi}_\gamma. \end{array}$$

Then, the frame decomposition (6.7) tells us that, first, SA is the identity of $L_2(\Omega_d)$, and, second, $\tilde{K} = AS$ is the orthogonal projector onto the range of A. Again, let $\alpha_\gamma = \langle f, \psi_\gamma \rangle$ be the ridgelet coefficients of f and $\alpha^{(N)}$ be the truncated sequence of the Nth largest coefficients, i.e. $\alpha_\gamma^{(N)} = \alpha_\gamma 1_{\{|\alpha_\gamma| \geq |\alpha|_{(N)}\}}$. Then, of course, $\tilde{f}_N = S\alpha^{(N)}$. Now the frame property gives

$$\begin{aligned} \|f - \tilde{f}_N\|_{L_2(\Omega_d)} &\leq B \, \|A(f - \tilde{f}_N)\|_{\ell_2(\Gamma_d)} \\ &= B \, \|\alpha - \tilde{K}\alpha^{(N)}\|_{\ell_2(\Gamma_d)} \\ &= B \, \|\tilde{K}(\alpha - \alpha^{(N)})\|_{\ell_2(\Gamma_d)} \\ &\leq B \, \|\alpha - \alpha^{(N)}\|_{\ell_2(\Gamma_d)}, \end{aligned}$$

since the norm of $\tilde{K}$ is 1. Now, Lemma 1 in Donoho (1993) allows us to conclude that

$$\|\alpha - \alpha^{(N)}\|^2_{\ell_2(\Gamma_d)} = \sum_{n>N} |\alpha|^2_{(n)} \leq a(p) \, |\alpha|^2_{w\ell_p},$$

and therefore, we have reached the desired conclusion. ∎

Lemma 15 together with Corollary 6 finishes the proof of the second part of Theorem 9.

5.4 Applications and Examples

This section uses the preceding lemma to study the approximation of specific examples of singular objects. Let w be a window in $C_0^\infty(\Omega_d)$ and let us consider for a moment the function $f_\alpha(x) = (x_1)_+^\alpha w(x)$. Our example f_α is a C^∞ function away from the hyperplane $x_1 = 0$ but which is singular across this same hyperplane: the singularity being of degree α. As we will see, the representations of such objects in a ridgelet frame are extremely sparse. In some sense, ridgelets are optimal to represent objects of this kind.

In all that follows, we will use the notations from Chapter 3. We will assume that φ and ψ are R times differentiable and that ψ has vanishing moments through order D. Now repeating the argument of section 5.2.1, we have

$$\begin{aligned} R_\gamma(f) &= \int_{\mathbf{R}^d} f_\alpha(x) 2^j \psi(2^j(u \cdot x - b))\, dx \\ &= \int_{\mathbf{R}^2} 2^j \psi(2^j(x_1' - b))(1+v^2)^{-\alpha/2}(vx_1' - x_2')_+^\alpha w_Q(x_1', x_2'), \end{aligned}$$

where w_Q is exactly as in the beginning paragraphs of section 5.2.1. So, $R_\gamma(f)$ is the wavelet coefficient of h_v where h_v is defined by

$$h_v(t) \equiv (1+v^2)^{-\alpha/2} \int y_+^\alpha w_Q(t, y - vt).$$

From the last display one can see that

$$\frac{d^n}{dt^n} h_v(t) = (1+v^2)^{-\alpha/2} \int y_+^\alpha \sum_\ell \binom{n}{\ell} v^\ell (\partial_1^{n-\ell} \partial_2^\ell w_Q)(t, y - vt)\, dy.$$

Now, a single integration by parts argument shows that

$$\Big| \int_{\mathbf{R}} y_+^\alpha v^\ell (\partial_1^{n-\ell} \partial_2^\ell w_Q)(t, y - vt)\, dy \Big| \le C\, |v|^\ell (1 + |vt|)^{\alpha - \ell} 1_{\{|t| \le 1\}},$$

estimates that can be turned into

$$\Big| \int_{\mathbf{R}} y_+^\alpha v^\ell (\partial_1^{n-\ell} \partial_2^\ell w_Q)(t, y - vt)\, dy \Big| \le C(1 + |v|^2)^{n/2} (1 + (1+v^2)^{1/2}|t|)^{\alpha - n} 1_{\{|t| \le 1\}},$$

for any $\ell \leq n$ and some constant C depending on n. Then

$$\left|\frac{d^n}{dt^n} h_v(t)\right| \leq C\,(1+|v|^2)^{(n-\alpha)/2}(1+(1+v^2)^{1/2}|t|)^{\alpha-n} 1_{\{|t|\leq 1\}},$$

where again the constant C might be chosen to only depend on n. In the remainder of the proof, we will suppose $\alpha - n < -1$.

Now, let v be such that $(1+v^2)^{1/2} \leq 2^j$. Standard wavelet calculations show that

$$|R_\gamma(f)| \leq C\,2^{-j(n+1/2)}(1+v^2)^{(n-\alpha)/2}(1+|(1+v^2)^{1/2}k2^{-j}|)^{\alpha-n}.$$

For $(1+v^2)^{1/2} \geq 2^j$, a simple calculation, parallel to what we have done, argues

$$|R_\gamma(f)| \leq C\,2^{-j(\alpha+1/2)}(1+|k|)^{\alpha-n}.$$

We are going to show that, for a nice frame, the ridgelet coefficients of f are in ℓ_p for any $p > 0$. So, let $p > 0$ be fixed. Our estimates imply that for $(1+v^2)^{1/2} \leq 2^j$,

$$\sum_k |R_\gamma(f)|^p \leq C^p\,2^{-jp(n+1/2)}(1+v^2)^{p(n-\alpha)/2}2^j(1+v^2)^{-1/2},$$

if n is chosen large enough so that $(\alpha - n)p < -1$. Next we recall that for $s \geq 0$

$$\sum_{i\in\mathcal{I}_j:(1+v_i^2)^{1/2}\leq 2^j} (1+v_i^2)^{s/2} \leq C\,\max(2^{js}, 2^{j(d-1)}),$$

and hence, summing over i's gives

$$\begin{aligned}\sum_{i\in\mathcal{I}_j}\sum_k |R_\gamma(f)|^p &\leq C^p\,2^{-jp(n+1/2)}\,2^j 2^{jp(n-\alpha)}2^{-j} \\ &= C^p\,2^{-jp(1/2+\alpha)},\end{aligned}$$

if n satisfies the additional condition: $n - \alpha > d/p$.

Hence, we have proved that the ridgelet coefficients $R_\gamma(f)_{\gamma\in\Gamma_d}$ – for a frame as in Chapter 3, with φ and ψ being R times differentiable and ψ having vanishing moment trough order D – satisfy

$$\|(R_\gamma(f))_{\gamma\in\Gamma_d}\|_{\ell_p} \leq C,$$

provided $d/p + \alpha < \min(R, D)$ which is what we claimed.

Applying the Jackson inequality (Lemma 15) gives

Proposition 10 *For any $\alpha > -1/2$, for any $u \in \mathbf{S}$, $b \in \mathbf{R}$ let $f(x) = (u \cdot x - b)_+^\alpha$. As usual $\tilde{f}_N$ will denote the truncation of f to the dual-ridgelet terms corresponding to the N largest ridgelet coefficients. Then,*

$$\|f - \tilde{f}_N\|_{L_2(\Omega_d)} = O(N^{-r}) \qquad \text{for any } r > 0.$$

(Note that for $\alpha \leq -1/2$, f is not square integrable.) Various extensions of results of this kind are, of course, possible.

Chapter 6

The Case of Radial Functions

In Chapter 5 we have seen that ridgelets are optimal to represent functions that may be singular across hyperplanes when there may be an arbitrary number of hyperplanes with any orientations and locations. A natural question one may ask: can we curve these hyperplanes (singular sets)? In other words, how are good ridgelets to represent objects being singular across curved manifolds? To answer this question, the study of radial objects is particularly attractive for it is both enlightening and simple. We will give accurate degrees of approximation of radial functions by ridgelets. Finally, we will establish a remarkable result: the rates of approximation by ridgelets to radial functions which are smooth away from spheres are identical to the rates achieved by wavelets.

The analysis will illustrate why ridge functions (neural networks) are not free from the curse of dimensionality.

Again, we shall restrict our attention to the case of compactly supported objects.

6.1 The Radon Transform of Radial Functions

Let f be a radial function $f(x) = \varphi(\|x\|)$ where φ is an even univariate and real valued function. Throughout this chapter we will suppose that supp $f \subset \Omega_d$, the unit ball of $\mathbf{R}^d$.

Recall that $R_u f$ is the Radon transform of f. We have

$$\begin{aligned} R_u f(t) &= \int_{u\cdot x=t} \varphi(\|x\|)\, dx \\ &= \int_{\mathbf{R}^{d-1}} \varphi(\sqrt{t^2+\|y\|^2})\, dy \\ &= \int_{S^{d-2}} \int_0^\infty \varphi(\sqrt{t^2+\rho^2})\rho^{d-2}\, d\rho du \\ &= |S^{d-2}| \int_0^\infty \varphi(\sqrt{t^2+\rho^2})\rho^{d-2}\, d\rho \\ &= |S^{d-2}| \int_{|t|}^\infty (r^2-t^2)^{(d-3)/2}\varphi(r) r\, dr. \end{aligned} \tag{6.1}$$

To simplify the notation, we will let T_d be the operator defined as

$$(T_d\varphi)(t) = \int_{|t|}^\infty (r^2-t^2)^{(d-3)/2}\varphi(r) r\, dr, \tag{6.2}$$

so that $R_u f(t) = |S^{d-2}|(T_d\varphi)(t)$. The operator T_d is an integral operator whose kernel is $r(r^2-t^2)_+^{(d-3)/2}$; from (6.2) one sees easily that for $t \geq 0$, we have

$$(T_d\varphi)(\sqrt{t}) = \frac{1}{2}\int_0^\infty (r-t)_+^{(d-3)/2}\varphi(\sqrt{r})\, dr.$$

Therefore, if we let U_d be the convolution operator defined by

$$(U_d g)(t) = \int_0^\infty (r-t)_+^{(d-3)/2} g(r)\, dr \tag{6.3}$$

and C be the change of coordinates $g(t) \mapsto g(\sqrt{t})$ (C^{-1} corresponding of course to $g(t) \mapsto g(t^2)$), we have

$$T_d\varphi = \frac{1}{2} C^{-1}\, U_d\, C\varphi.$$

Now, suppose that we first look at functions vanishing in a neighborhood of the origin. For instance, let $\rho > 0$ such that $\varphi_{|[0,\rho]} = 0$. We have for $1 \leq p, q \leq \infty$

$$\|T_d\varphi\|_{B_{p,q}^{s+(d-1)/2}} \leq C_\rho \|\varphi\|_{B_{p,q}^s},$$

where, in addition to ρ, the constant C_ρ depends on s, p, q. To see this, observe that for

supp $g \subset [\rho, 1]$, the operator $C : g(t) \mapsto g(\sqrt{t})$ is bounded from $B^s_{p,q}(0,1)$ to itself for any choice of s, p, q. In addition, $C^{-1} : g(t) \mapsto g(t^2)$ is bounded from $B^s_{p,q}(0,1)$ to $B^s_{p,q}(-1,1)$, again for any possible triplet s, p, q. The only thing remaining to prove is that U_d is bounded from $B^s_{p,q}(0,1)$ to $B^{s+(d-1)/2}_{p,q}(0,1)$. This is not very difficult and is shown in the Appendix.

As a matter of fact, we probably do not need the additional assumption φ to vanish in a neighborhood of the origin.

Proposition 11 *Suppose φ is an even function such that $\varphi \in B^s_{p,q}(-1,1)$. Then $T_d\varphi \in B^{s+(d-1)/2}_{p,q}(-1,1)$ and*

$$\|T_d\varphi\|_{B^{s+(d-1)/2}_{p,q}(-1,1)} \leq C\|\varphi\|_{B^s_{p,q}(-1,1)}, \tag{6.4}$$

for some constant C depending at most on s, p, q and the dimension d. Moreover, suppose in addition that φ vanishes in $[-\rho, \rho]$; then we have the reverse inequality

$$\|T_d\varphi\|_{B^{s+(d-1)/2}_{p,q}(-1,1)} \geq C\|\varphi\|_{B^s_{p,q}(-1,1)}, \tag{6.5}$$

where, here, the constant C may also depend on ρ.

Proof of Proposition. We give a proof of the first part of the proposition in the case of odd dimension d. For the time being, we omit the proof when d is even. The reader who might doubt the validity of the result in the latter case is invited to add the additional condition φ to be supported away from 0 in the reminder of the chapter (since, we have proved the result in this case).

First, observe that the membership to $B^s_{p,q}$ is equivalent to both memberships to L_p and $\dot{B}^s_{p,q}$. Now the following semi-norms are equivalent:

$$\|f\|_{\dot{B}^s_{p,q}} \asymp \|f'\|_{\dot{B}^{s-1}_{p,q}}$$

and, therefore,

$$\|f\|_{B^s_{p,q}} \asymp \|f\|_p + \|f'\|_{\dot{B}^{s-1}_{p,q}},$$

in the sense of equivalent norms. Now suppose further that f is compactly supported, say,

in the interval $[-1,1]$. For $p \geq 1$, we have the Poincaré inequality which states that

$$\|f\|_p \leq C(p)\|f'\|_p,$$

and therefore, for compactly supported functions, $\|f'\|_p + \|f'\|_{\dot{B}^{s-1}_{p,q}}$ dominates $\|f\|_{B^s_{p,q}}$.

Now, we prove the proposition by induction. For $d = 3$, it is not hard to see that

$$\frac{d}{dt}(T_3\varphi)(t) = -t\phi(t) \qquad \text{for } \; t \neq 0.$$

Following our observation, if $\varphi \in B^s_{p,q}$ it is then obvious that $T_3\varphi \in B^{s+1}_{p,q}$ and that we have

$$\|T_3\varphi\|_{B^{s+1}_{p,q}} \leq C\|\varphi\|_{B^s_{p,q}},$$

for some constant possibly depending on s, p, q.

Next, suppose that $T_d\varphi \in B^{s+1}_{p,q}$ and satisfies (6.4). Again, one can see easily that

$$\frac{d}{dt}(T_{d+2}\varphi)(t) = -t(T_d\varphi)(t) \qquad \text{for } \; t \neq 0.$$

Therefore, the same argument as the one for $d = 3$ yields (6.4) for $d + 2$.

We turn to the proof of the lower bound (6.5). Let χ be a C^∞ function such that $0 \leq \chi \leq 1$, $\chi = 1$ whenever $|t| \geq 1$, and $\chi = 0$ for $|t| \leq 1/2$ (the notation χ_ϵ will stand for $\chi(\cdot/\epsilon)$). Now, there are well-known inversion formulas for the Radon transform. In particular, when the function f is radial ($f(x) = \varphi(\|x\|)$, it is easily established that (Natterer, 1986)

$$\varphi(t) \propto L^{d-1}(T_d \circ T_d\varphi)(t), \tag{6.6}$$

where L is the operator $\frac{1}{t}\frac{d}{dt}$. Again, we assume that $\varphi \in B^s_{p,q}$ and let us consider $g = \chi_{\rho/2}(T_d \circ T_d)\varphi$ (note that the first part of our proposition implies that $g \in B^{s+d-1}_{p,q}$). We claim that

$$\|Lg\|_{B^{s+d-2}_{p,q}} \leq C\|g\|_{B^{s+d-1}_{p,q}}.$$

To see this, recall that the differentiation operator $\frac{d}{dt}$ is a bounded map between $B^s_{p,q}$ and $B^{s-1}_{p,q}$ for any $-\infty < s < \infty$, as it can be obtained as a corollary of two theorems from

Triebel (1983, Pages 57–61). Further, the multiplication by $1/t$ is of course a bounded operation in $B^{s-1}_{p,q}$ restricted to its elements supported in, say, $\{\rho/4 \leq |t| \leq 1\} \supset$ supp g. By induction, we get that

$$\|L^{d-1}g\|_{B^{s-d}_{p,q}} \leq C\|g\|_{B^s_{p,q}}.$$

But of course, it follows from (6.6) that for $|t| \geq \rho/2$, we have $L^{d-1}g(t) = \varphi(t)$ implying $\varphi = \chi_\rho L^{d-1}g$. Therefore, to summarize, we have

$$\begin{aligned}\|\varphi\|_{B^s_{p,q}} = \|\chi_\rho L^{d-1}g\|_{B^s_{p,q}} &\leq C\,\|L^{d-1}g\|_{B^s_{p,q}} \leq C\,\|g\|_{B^{s+d}_{p,q}} \\ &= C\,\|\chi_{\rho/2}(T_d \circ T_d)\varphi\|_{B^{s+d}_{p,q}} \leq C\,\|(T_d \circ T_d)\varphi\|_{B^{s+d}_{p,q}} \leq C\,\|T_d\varphi\|_{B^{s+(d-1)/2}_{p,q}},\end{aligned}$$

where the last inequality comes from the first part of the proposition. This completes the proof of the proposition. ∎

6.2 The Approximation of Radial Functions

Proposition 11 has important consequences. For instance, suppose we have a frame of $L_2(\Omega_d)$ (see Chapter 3)

$$\{\varphi(u_i^j \cdot x - kb_0),\ 2^{j/2}\psi(2^j u_i^j \cdot x - kb_0),\ \ j \geq 0, u_i^j \in \Sigma_j, k \in \mathbf{Z}\},$$

where we recall that $|\Sigma_j| = O(2^{j(d-1)})$. We will take both φ and ψ to be compactly supported. Following Chapter 2, we will denote the element of the frame ψ_γ for $\gamma \in \Gamma_d$. We know that any element of, say, $L_2(\Omega_d)$ might be represented as

$$f = \sum_{\gamma \in \Gamma_d} \langle f, \psi_\gamma \rangle \tilde{\psi}_\gamma = \sum_{\gamma \in \Gamma_d} \langle f, \tilde{\psi}_\gamma \rangle \psi_\gamma. \tag{6.7}$$

To simplify, α_γ will denote the ridgelet coefficient $\langle f, \psi_\gamma \rangle$. Finally, let $\tilde{f}_N$ be the N-term ridgelet expansion where one only keeps the terms corresponding to the N largest coefficients (5.19). We have the following theorem:

Theorem 12 *Let f be a radial function supported in the unit ball $f(x) = \varphi(\|x\|)$ such that*

$\varphi \in B^s_{p,q}$. Suppose that $s > d(1/p - 1/2)_+$. Then we have the L_2 error of approximation

$$\|f - \tilde{f}_N\|_2 \leq C\ N^{-s/d} \|\varphi\|_{B^s_{p,q}}. \tag{6.8}$$

Remark 1. In fact, the result (6.8) is probably sharp: that is for any $s' > s$, there must be a function $\varphi \in B^s_{p,q}$ with, say, $\|\varphi\|_{B^s_{p,q}} = 1$ such that

$$\limsup \|f - \tilde{f}_N\|_2\ N^{s'/d} = \infty. \tag{6.9}$$

At the end of this section, we will argue that there is some strong evidence supporting this conjecture. This suggests that neural networks are also subject to the curse of dimensionality: to obtain approximation to f at rates N^{-r}, we must assume that f is $r \cdot d$ times differentiable.

Remark 2. We remark that the rate of convergence is identical to the one obtained using truncated wavelet expansions. This remarkable fact will be examined further in the discussion section.

Proof of Theorem. We start by proving the upper bound. Notice that by construction the collection

$$\{\varphi(t - kb_0),\ 2^{j/2}\psi(2^j t - kb_0),\ j \geq 0,\ k \in \mathbf{Z}\}$$

is a frame of $L_2[-1, 1]$. Now, for an univariate function $g : \mathbf{R} \to \mathbf{R}$, let $\alpha_{j,k} = \langle g, \psi_{j,k}\rangle$ and $\beta_k = \langle g, \phi_k\rangle$. Although we do not give a proof here, the quantity

$$\|f\|_{\mathbf{b}^s_{p,q}} \equiv \|(\beta_k)_k\|_{\ell_p} + \|(2^j 2^{j(1/2-1/p)}(\sum_k |\alpha_{j,k}|^p)^{1/p})_{j\geq 0}\|_{\ell_q} \tag{6.10}$$

is an equivalent norm to the norm of f in $B^s_{p,q}$, for any g supported in $[-1, 1]$.

As we suppose that $f(x) = g(\|x\|)$ is supported in the unit ball, we recall that

$$\alpha_\gamma = \langle f, \psi_\gamma\rangle = \langle R_{u_i^j} f, \psi_{j,k}\rangle = |S^{d-2}|\langle T_d g, \psi_{j,k}\rangle = |S^{d-2}|\alpha_{j,k}(T_d g).$$

For a g in $B^s_{p,q}$, we know that its image $T_d g$ is in $B^{s+(d-1)/2}_{p,q}$; in addition, its norm $\|T_d g\|_{\mathbf{b}^{s+(d-1)/2}_{p,q}}$ is bounded by $\|g\|_{B^s_{p,q}}$. Further, it is fairly clear that for any $q > 0$,

$$\|T_d g\|_{\mathbf{b}^{s+(d-1)/2}_{p,\infty}} \leq \|T_d g\|_{\mathbf{b}^{s+(d-1)/2}_{p,q}} \leq C\ \|g\|_{\mathbf{b}^s_{p,q}}.$$

Following Donoho (1993), one may bound the number of coefficients, at a given scale j, whose absolute values exceed a certain threshold $\epsilon > 0$ as displayed below:

$$\#\{k,\, |\alpha_{j,k}| > \epsilon\} \leq c_{b_0} \min(2^j, (2^{-j(s+(d-1)/2+1/2-1/p)} \|T_d g\|_{\mathbf{b}^{s+(d-1)/2}_{p,\infty}} \epsilon^{-1})^p).$$

Thus, the number of ridgelet coefficients α_γ at scale j that exceed ϵ in absolute value is larger than $\epsilon > 0$ is bounded as follows:

$$\begin{aligned} &\#\{\gamma,\, |\alpha_\gamma| > \epsilon \ \text{ and } \ |\gamma| = j\} \\ &\quad \leq \quad c_{b_0} C 2^{j(d-1)} \min(2^j, (2^{-j(s+(d-1)/2+1/2-1/p)} \|T_d g\|_{\mathbf{b}^{s+(d-1)/2}_{p,\infty}} \epsilon^{-1})^p) \\ &\quad = \quad c_{b_0} C \min(2^{jd}, (2^{-j(s+d(1/2-1/p)} \|T_d g\|_{\mathbf{b}^{s+(d-1)/2}_{p,\infty}} \epsilon^{-1})^p), \end{aligned} \tag{6.11}$$

since the number of directions u_i^j is bounded by $C2^{j(d-1)}$ for any scale j. Summing (6.11) over scales j, a bit of algebra shows that (provided $s > d(1/p - 1/2)_+$)

$$\#\{\gamma,\, |\alpha_\gamma| > \epsilon\} \leq C \, \|T_d g\|_{\mathbf{b}^{s+(d-1)/2}_{p,\infty}} \epsilon^{-\frac{1}{s/d+1/2}}. \tag{6.12}$$

where C possibly depends upon d, s, p and the sampling resolution b_0. Let p^* be defined by $1/p* = s/d + 1/2$. The inequality (6.12) merely states that the weak-ℓ_{p*} norm of α is bounded by $\|T_d g\|_{\mathbf{b}^{s+(d-1)/2}_{p,\infty}}$ up to a multiplicative constant. The supporting Lemma 15 leads to the desired conclusion. ∎

Of course, one would like to have a converse statement or, in other words, give lower bounds of the squared error of approximation $\|f - \tilde{f}_N\|_2$. In the remarks directly following the statement of our theorem, we have articulated a reasonable conjecture (6.9) expressing the sharpness of the bound (6.8). Unfortunately, we are not able to prove this conjecture yet. However, we have available a recipe for constructing radial functions whose ridgelet coefficients sequences are not sparse. For instance, let p' be a real such that $1/p' > 1/p^* = s/d + 1/2$. Making use of Proposition 11 and Lemma 2 in Donoho (1993), one can find a $\varphi \in B^s_{p,q}$ ($\|\varphi\|_{B^s_{p,q}} = 1$) such that the sequence of ridgelet coefficient α is not in $w\ell_{p'}$. Now, the problem is that we don't have a converse to Lemma 15: that is, if we know something about the rate of approximation by finite linear combinations of ridgelets (or their duals), does it imply that the ridgelet coefficients have a certain decay? Let us be even more precise.

Suppose f is such that there is a sequence $f_N = \sum_{i=1}^{N} \lambda_{i,N} \psi_{\gamma_i,N}$ with the property

$$\|f - \tilde{f}_N\|_2 \leq C\, N^{-r}.$$

Does it imply that $\{\langle f, \psi_\gamma \rangle\} \in w\ell_p(\Gamma_d)$ with $r = (1/p - 1/2)_+$? In Approximation Theory, this is often referred to as a Bernstein type of inequality whereas Lemma 15 is a kind of abstract Jackson inequality. Suppose the ridgelets were orthogonal; then the Bernstein inequality would be trivial. The delicate issue here is, of course, that the ridgelets are possibly linearly dependent. Work in this direction is in progress.

6.3 Examples

In this section, we consider a class of radial functions (f_α) defined by $f_\alpha(x) = (1 - \|x\|^2)_+^\alpha$. Away from the sphere of radius $r = 1$, these functions are smooth but are singular across this same sphere. Here, the index parameter α is simply the degree of the singularity (see Definition 6). In all that follows, we will suppose $\alpha > -1/2$, so that f_α is square integrable.

First, a simple calculation shows that

$$\begin{aligned} R_u f_\alpha(t) &= |S^{d-2}| \int_{|t|}^{1} (r^2 - t^2)^{(d-3)/2} (1 - r^2)^\alpha r\, dr \\ &= \frac{1}{2}(1 - t^2)^{\alpha + (d-1)/2} \int_0^1 (1 - v)^{(d-3)/2} v^\alpha dv \\ &= c_{\alpha,d} (1 - t^2)^{\alpha + (d-1)/2}. \end{aligned}$$

We now check the sparsity of the representation of f_α in a ridgelet frame. Again, suppose that we are using ridgelets such that their profiles φ and ψ have compact supports. We will assume that φ and ψ are R times differentiable and that ψ has vanishing moments through order D. Finally let p^* be defined by $1/p^* = 1/2 + (\alpha + 1/2)/(d-1)$. We show that the coefficients α_γ are in $w\ell_{p*}$. If $\min(R, D)$ is large enough ($\min(R, D) \geq \max(2/p^*, 2 + \alpha + (d-1)/2)$ suffices), we have

$$\begin{aligned} \alpha_\gamma = \langle f, \psi_\gamma \rangle &= c_{\alpha,d} \int (1 - t^2)^{\alpha + (d-1)/2} 2^{j/2} \psi(2^j t - k b_0) \\ &\leq C_{\alpha,d}\, 2^{-j(1/2 + \alpha + (d-1)/2)} \left(1 + \left| |k b_0| - 2^j \right| \right)^{-2/p*} \\ &= C_{\alpha,d} 2^{-j(d-1)/p^*} \left(1 + \left| |k b_0| - 2^j \right| \right)^{-2/p*}. \end{aligned} \tag{6.13}$$

The proof of the first inequality is an application of integration by parts and we omit it.

So, let $\epsilon > 0$ and let us bound the number of coefficients α_γ that are greater or equal to ϵ in absolute value. For a given scale j and orientation u_i^j, it follows from (6.13) that

$$\#\{k, |\alpha_\gamma| \geq \epsilon\} \leq C\, 2^{-j(d-1)/2} \epsilon^{-p*/2}, \tag{6.14}$$

where the constants may depend on the parameter b_0. Moreover, (6.13) implies that $\#\{k, |\alpha_\gamma| \geq \epsilon\} = 0$ whenever j verify

$$C_{\alpha,d}\, 2^{-j(d-1)/p^*} < \epsilon \Leftrightarrow C_{\alpha,d}^{p*/2}\, 2^{-j(d-1)/2} < \epsilon^{p*/2}.$$

Then, of course, if we fix the scale j, we trivially get that

$$\#\{\gamma \text{ with } |\gamma| = j, |\alpha_\gamma| \geq \epsilon\} \leq C\, 2^{j(d-1)/2} \epsilon^{-p*/2} 1_{\{2^{j(d-1)/2} \leq (C_{\alpha,d}\epsilon^{-1})^{p*/2}\}}.$$

And summing across scales gives

$$\#\{\gamma\ , |\alpha_\gamma| \geq \epsilon\} = O(\epsilon^{-p*}),$$

which estabishes the claim.

As a consequence of the last display, a direct application of Lemma 15 gives

Proposition 12 *Suppose $\alpha > -1/2$ and let f be the spherical singular function $f(x) = (1 - \|x\|^2)_+^\alpha$. As in the preceding section, let $\tilde{f_N}$ be the truncated ridgelet expansion of f. Then,*

$$\|f - \tilde{f_N}\|_2 = O(N^{-(\alpha+1/2)/(d-1)}). \tag{6.15}$$

Suppose one were to use wavelets to approximate the singular function f from the above Proposition. So let $(\varphi_k, \psi_{j,k}^\epsilon)$ ($\lambda = (j,k)$, $j \geq 0, k \in \mathbf{Z}^d$) be a nice isotropic wavelet basis in $\mathbf{R}^d$ with compact support, enough regularity and vanishing momemts. Standard techniques show that for any $\delta > 0$, we have

$$|\langle f, \psi_\lambda \rangle| \leq C(\delta, \alpha) 2^{-j} 2^{-j\alpha} \left(1 + \left|\|k\| - 2^j\right|\right)^{-\delta}.$$

As a straightforward consequence of the above estimate, one can see that the best wavelet

approximation converges at the rate $O(N^{-(\alpha+1/2)/(d-1)})$. In this case, it can be shown that this rate cannot be improved.

6.4 Discussion

In this chapter, we have demonstrated in several places that ridgelets and multi-dimensional wavelets were equally as good, at least asymptotically, to approximate radial functions. For instance, in the previous section, we have shown explicitly that the rate of approximation of, say, $f_\alpha(x) = (1 - \|x\|^2)_+^\alpha$, for $\alpha > -1/2$ using ridgelets coincides with the one obtained using multi-dimensional wavelets – regardless of the degree of the singularity α and/or of the dimension d. In fact, this conclusion is not limited to the case of radial functions. For instance, one could consider a simple extension of our radial example: let $\mathcal{F}$ be the set of all diffeomorphisms $h : \mathbf{R}^d \to \mathbf{R}^d$, with, say, $h \in C^r$ for some $r > 0$ and such that all the partial derivatives of h up to order r are bounded by some constant C. Now, for any $h \in \mathcal{F}$, consider $g_\alpha = f_\alpha \circ h$. So, g_α is a smooth function away from a hypersurface Σ with a minimum of curvature and singular across Σ (roughly speaking, one sees a singularity of degree α when one crosses Σ along a normal vector to the surface). It then turns out that if r is large enough, the conclusion of Proposition 12 remains unchanged for this larger class of functions. That is, the rate of approximation of g by ridgelets is $O(N^{-(\alpha+1/2)/(d-1)})$ and so is the rate of convergence of wavelet approximations.

In the author's opinion, this result is most remarkable and surprising. The two discussed approximation schemes are, of course, very much unlike: as one is building up an approximation using oscillatory bumps at various scales whereas the other is using ridge functions localized around strips of various scales. And yet, both seem to represent functions that are smooth away from curved hypersurfaces, but that may be singular across these hypersurfaces, with the same degree of accuracy.

We close this chapter by pointing out that our findings seem to contradict – at least at a superficial level – the results obtained by Donoho and Johnstone (1989). In fact, the model they consider is somewhat different as they study the squared error of approximation with respect to the gaussian measure whereas we are concerned with compactly supported functions. Hence, their results are not inconsistent with ours. First, the behavior at infinity of the functions they are treating may account for the divergence of our conclusions. Second, their study only concerns the case of $\mathbf{R}^2$ and, perhaps, the effects they describe are specific

to this case.

Chapter 7

Concluding Remarks

7.1 Ridgelets and Traditional Neural Networks

The purpose of this section is to show that ridgelet approximations offer decisive superiority over traditional neural networks. We recall that in the latter case one considers finite approximations from $\mathcal{D}_{NN} = \{\rho(k \cdot x - b),\ k \in \mathbf{R}^n, b \in \mathbf{R}\}$, where ρ is the univariate sigmoid. As we have already shown, neural networks cannot outperform ridgelets over the classes $R^s_{p,q}(C)$. However, outside of this paradigm, it is of interest to compare for any function f the rates of approximations using either $\mathcal{D}_{NN}$ or $\mathcal{D}_{Ridgelets}$.

In Chapter 5, we defined the approximation error of a a function f by N-elements of the dictionary $\mathcal{D}$:

$$\inf_{(\alpha_i)_{i=1}^N} \inf_{(\lambda_i)_{i=1}^N} \|f - \sum_{i=1}^{N} \alpha_i g_{\lambda_i}\|_H \equiv d_N(f, \mathcal{D}).$$

In many interesting cases, $d_N(f, \mathcal{D}) \asymp N^{-r}$ or $d_N(f, \mathcal{D}) \asymp N^{-r} \log^{\delta}(N)$. A crude measure of convergence is the *exponent of approximation.*

$$r^*(\mathcal{F}, \mathcal{D}) = \sup\{r,\ d_N(f, \mathcal{D}) = O(N^{-r})\} \tag{7.1}$$

This measure is, indeed, insensitive to log-like factors. Now, one would like to know if there exists a function f for which

$$r^*(f, \mathcal{D}_{NN}) > r^*(f, \mathcal{D}_{Ridgelets}).$$

The answer to this question is generally negative. In order to give a precise statement, we only consider finite linear approximations with bounded coefficients: i.e., we restrict the minimization (7.1) to $|\alpha_i| \leq C$ for some positive C. In practice, this restriction makes perfect sense since one is not going to store unbounded coefficients. Also, note that the relaxed greedy algorithm (1.2)-(1.3) produces coefficients bounded by 1. With this restriction, one can show

Theorem 13 *For any given f,*

$$r^*(f, \mathcal{D}_{NN}) \leq r^*(f, \mathcal{D}_{Ridgelets}). \tag{7.2}$$

Whenever traditional neural networks give a good approximation of f, there is at least an equally good - and perhaps much better - approximation using ridgelets. In short, there seems to be very little advantage in using traditional neural activation functions.

Proof of Theorem. The idea behind the theorem is very simple: any neuron $\rho(k \cdot x - b)$ can be approximated in $O(N^{-s})$ for any $s > 0$ using the dual ridgelet dictionary; therefore, one can get a very good approximation of any finite linear combination of such neurons.

Step 1. Let $\gamma_0 = (a_0, u_0, b_0)$ be in $\Gamma = \mathbf{R}^+ \times \mathbf{S}^{d-1} \times \mathbf{R}$ and let ρ_{γ_0} be the neuron defined by $\rho_{\gamma_0}(x) = \rho(a_0(u_0.x - b_0))$. Consider now a ridgelet frame for $L_2(2\Omega_d)$ as in Chapter 3, section 3.6.1, and a d dimensional window w, $0 \leq w \leq 1$, supp $w \subset 2\Omega_d$ and $w_{|\Omega_d} = 1$. The frame decomposition theorem allows us to write

$$\rho_{\gamma_0}(x)w(x) = \sum_{\gamma \in \Gamma_d} \alpha_{\gamma_0,\gamma} \tilde{\psi}_\gamma(x) \tag{7.3}$$

in the sense of $L_2(2\Omega_d)$. From the exact series displayed above, extract – as usual – the finite sum $\rho_{\gamma_0,N}$ corresponding to the N largest coefficients. The first part of the proof consists of proving that

$$\sup_{\gamma_0} \|\rho_{\gamma_0} - \rho_{\gamma_0,N}\|_{L_2(\Omega_d)} \leq CN^{-s} \tag{7.4}$$

for any $s > 0$. To prove the claim, we will try to use a maximum number of results proved in the previous chapters.

Let us consider for a moment the Meyer wavelet basis (Daubechies, 1992) $(\tilde{\varphi}_{0,k}, \tilde{\psi}_{j,k})$ on the real line. For scalars $a_0 > 0$ and b_0, ρ_{a_0,b_0} will denote the univariate function

$\rho(a_0(t-b_0))$. As one can guess, the wavelet coefficients sequence of $\rho_{a_0,b_0}w$ is extremely sparse. We claim that

$$|\langle \rho_{a_0,b_0}, \psi_{j,k}\rangle| \leq \begin{cases} C2^{-j/2}(1+|k-2^j b_0|)^{-\delta} & a_0 \geq 2^j \\ Ca_0 2^{-j}2^{-j/2}(1+|k-2^j b_0|)^{-\delta} & a_0 \leq 2^j \end{cases}.$$

First, notice that $\rho(t) = (1+e^{-t})^{-1} = 1/2 + \sigma(t)$ where σ, according to Definition 6, is an α-singularity of degree 1 (up to a renormalizing constant). To see this, remark that $|\sigma(t)|$ is obviously bounded by $|t|$ and an induction argument shows that for $n \geq 1$,

$$\Big|\frac{d^n}{dt^n}\sigma(t)\Big| \leq C(n)e^{-|t|},$$

which suffices to establish the singularity property. Then, for any positive a_0, $a_0^{-1}\sigma(a_0(t-b_0))$ is also a singularity of degree 1. For a singularity of degree 1, we know that

$$|\langle a_0^{-1}\sigma(a_0(t-b_0)), \psi_{j,k}\rangle| \leq C\, 2^{-3j/2}(1+|k-2^j b_0|)^{-\delta},$$

where the constant C may depend on δ. Then, of course, $\psi_{j,k}$ is orthogonal to the constant function; hence,

$$|\langle \rho_{a_0,b_0}, \psi_{j,k}\rangle| \leq C\, a_0 2^{-3j/2}(1+|k-2^j b_0|)^{-\delta},$$

which is the part of the result corresponding to the case $a_0 \leq 2^j$. As far as the case $a_0 \geq 2^j$ is concerned, we observe that

$$\frac{d}{dt}\rho_{a_0,b_0}(t) = a_0\rho'(a_0(t-b_0))$$

and, therefore, the above derivative is bounded in absolute value by $a_0 e^{-a_0|t-b_0|}$. Now since ψ has one vanishing moment, a simple integration by parts gives

$$|\langle \rho_{a_0,b_0}w, \psi_{j,k}\rangle| \leq C\, 2^{-j/2}(1+|k-2^j b_0|)^{-\delta},$$

which proves the second part of the claim. It follows from the previous inequalities that for every choice of p, $\|\alpha\|_{\ell_p}$ is finite and uniformly bounded (over all the possible choices of a_0 and b_0). Furthermore, it is trivial that $\beta_k = \langle \rho_{a_0,b_0}, \tilde{\varphi}_{0,k}\rangle$ satisfies $|\beta_k| \leq \|\varphi\|_1$.

Now, back in terms of our frame decomposition, the property we have just shown is very interesting since it implies that the sequence $(\alpha_{\gamma_0,\gamma})_{\gamma\in\Gamma_d}$ in (7.3) is also uniformly bounded (over all choices of γ_0) in ℓ_p. This merely follows from Chapter 5 (section 5.2.1) as it is a consequence of Lemma 10. Then, Lemma 15 allows us to conclude and establish the validity of (7.4).

Step 2. We are now in a position to finish up the proof of the theorem. We show that for any $r < r^*(f, \mathcal{D}_{NN})$ there is a sequence f_N of N-term ridgelet expansions with the property

$$\|f - f_N\| = O(N^{-r}).$$

It is not difficult to construct such a sequence since by assumption we know the existence of a sequence $g_m(x) = \sum_{i=1}^m \alpha_{i,m}\rho(k_{i,m}\cdot x - b_{i,m})$ (that we shall write $g_m = \sum_i \alpha_{i,m}\rho_{\gamma_{i,m}}$), such that

$$\|f - g_m\|_2 \leq Cm^{-r'}$$

with, say, $r' = (r + r^*(f, \mathcal{D}_{NN}))/2$. For a fixed N, let m be defined by $m = N^{1-\delta}$, where $(1-\delta)r' = r$. Next define f_N as follows:

$$f_N = \sum_{i=1}^m \alpha_{i,m} \sum_{\gamma} \rho_{\gamma_{i,m},N^\delta}.$$

Then, of course, f_N is a ridgelet expansion of order N terms and

$$\begin{aligned}
\|f - f_N\| &\leq \|f - g_m\| + \sum_{i=1}^m |\alpha_{i,m}| \|\rho_{\gamma_{i,m}} - \rho_{\gamma_{i,m},N^\delta}\| \\
&= O(m^{-r'}) + mO(m^{-s}) \\
&= O(N^{-(1-\delta)r'}) + O(N^{-(1-\delta)(s-1)}) = O(N^{-r})
\end{aligned}$$

since s may be chosen arbitrarily large (in particular $s - 1 > r'$.) This finishes the proof of the theorem. ∎

Remark. In fact, the proof of the theorem shows that the assumption requiring the coefficient α_i to be bounded may be dropped. The same result is true if we restrict the coefficients to grow polynomially: that is, in the best approximation (7.1) we only require $|\alpha_i| \leq BN^\beta$ for some constants B and β.

7.2 What About Barron's Class?

In Chapter 1, we did mention a well-known result from Barron (1993) – see section 1.2; we recall that in his paper, Barron studies rates of approximation of a special class of square-integrable functions supported in the unit cube Q of $\mathbf{R}^d$ by finite superpositions of elements taken from $\mathcal{D}_{NN}$: namely, he considers objects satisfying

$$\|f\|_B \equiv \int_{\mathbf{R}^d} \|k\| |\hat{f}(k)| \, dk \leq C. \tag{7.5}$$

In this section, we will refer to this class as the Barron class $B(C)$ and to $\|\cdot\|_B$ as the Barron norm. His work shows that

$$d_N(B(C), \mathcal{D}_{NN}) = O(N^{-1/2}). \tag{7.6}$$

In fact, the Fourier dictionary is optimal for approximating elements of $B(C)$. (This is not surprising since the class $B(C)$ is defined by means of the Fourier transform.) Indeed, one can show that

$$d_N(B(C), \mathcal{D}_{Fourier}) = O(N^{-1/2-1/d}), \tag{7.7}$$

where the Fourier dictionary is, of course, $\mathcal{D}_{Fourier} \equiv \{e^{i2\pi n \cdot x}, \ n \in \mathbf{Z}^d\}$. Moreover, thresholding the Fourier expansion (keeping the N terms corresponding to the N largest Fourier coefficients) gives the optimal rate of approximation. In the next paragraph, we will show how to use the tools developed in this thesis to prove that (7.7) is indeed a lower bound. As far as the upper bound is concerned, observe that the Barron norm is actually equivalent to

$$\|f\|_B \asymp \sum_{n \in \mathbf{Z}^d} \|n\| |\hat{f}(2\pi n)|,$$

as it is a simple consequence of a famous theorem due to Plancherel and Pólya (1938). Next, a bound on the right-hand side of the above display implies a bound on the weak-$\ell_{1+1/d}$ quasi-norm (5.20) of the sequence $(|\hat{f}(2\pi n)|)_{n \in \mathbf{Z}^d}$. Thus, Fourier series outperform neural networks over the $B(C)$ model of functions.

In his analysis, Barron shows that the rate (7.6) may be obtained with expansions

of the form $f_N(x) = \sum_{i=1}^N \alpha_i \rho(k_i \cdot x - b_i)$, where the α_i's may be restricted to satisfy $\sum_{i=1}^N |\alpha_i| \leq C'$ for some constant C'. Hence, a straightforward application of Theorem 13 gives at the minimum $d_N(B(C), \mathcal{D}_{Dual-Ridge}) = O(N^{-s})$ for any $s < 1/2$. Moreover, a bit of ridgelet analysis shows that the rate of approximation of the Barron class by any reasonable dictionary (see Chapter 5, section 5.1) is bounded below by $O(N^{-1/2-1/d})$. The reason is the following: let us consider a pair (φ, ψ) satisfying the conditions (i)-(iv) listed at the very beginning of Chapter 4 together with (2.7); namely,

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

Then we have,

$$\begin{aligned}
\|f\|_B &= \int_{\mathbf{R}^d} \|k\| |\hat{f}(k)|\, dk \\
&= \int_{\xi>0} \int_{\mathbf{S}^{d-1}} |\hat{f}(\xi u)| |\xi|^d \, d\xi du \\
&= \int_{\xi>0} \int_{\mathbf{S}^{d-1}} |\hat{f}(\xi u)| |\hat{\varphi}(\xi)|^2 |\xi| d\xi du + \sum_{j\geq 0} 2^{j(d-1)} \int_{\xi>0} \int_{\mathbf{S}^{d-1}} |\hat{f}(\xi u)| |\hat{\psi}(2^{-j}\xi)|^2 |\xi| \, d\xi du \\
&\leq \left(\int |\hat{\varphi}(\xi)|^2 |\xi|^2 \, d\xi du \int |\hat{f}(\xi u)|^2 |\hat{\varphi}(\xi)|^2 \, d\xi du \right)^{1/2} \\
&\qquad + \sum_{j\geq 0} 2^{j(d-1)} \left(\int |\hat{\psi}(2^{-j}\xi)|^2 |\xi|^2 \, d\xi du \int |\hat{f}(\xi u)|^2 |\hat{\psi}(2^{-j}\xi)|^2 \, d\xi du \right)^{1/2} \\
&\leq C \left(\left[\int |\hat{f}(\xi u)|^2 |\hat{\varphi}(\xi)|^2 \, d\xi du \right]^{1/2} + \sum_{j\geq 0} 2^{j(d+1/2)} \left[\int |\hat{f}(\xi u)|^2 |\hat{\psi}(2^{-j}\xi)|^2 \, d\xi du \right]^{1/2} \right) \\
&= C \|f\|_{R_{2,1}^{d/2+1}}. \qquad (7.8)
\end{aligned}$$

Then, it follows from (7.8) that the Barron class $B(C)$ contains $R_{2,1}^{d/2+1}(C_1)$ for some constant C_1, and since the rate of approximation of the latter class is bounded below by $O(N^{-1/2-1/d})$ (Theorem 9), so is $B(C)$. We hope that this example emphasizes further the use and the potential power of the ridgelet analysis that we have developed in this thesis.

7.3 Unsolved Problems

Along the way, we have encountered a couple of issues that were left unanswered or unproved. In this section, we address a specific one.

One might be actually interested in synthesizing functions out of the ridgelet dictionary rather than out of the dual-ridgelet dictionary for approximating a target f. For instance, instead of considering N-term expansions of the form

$$\tilde{f}_N = \sum_{\gamma \in \Gamma} \alpha_\gamma 1_{\{|\alpha_\gamma| \geq |\alpha|_N\}} \tilde{\psi}_\gamma,$$

one may want to consider expansions like

$$f_N = \sum_{\gamma \in \Gamma} \tilde{\alpha}_\gamma 1_{\{|\tilde{\alpha}_\gamma| \geq |\tilde{\alpha}|_N\}} \psi_\gamma,$$

where α_γ (respectively $\tilde{\alpha}_\gamma$) stand for the ridgelet coefficients $\langle f, \psi_\gamma \rangle$ (respectively the dual ridgelet coefficients$\langle f, \tilde{\psi}_\gamma \rangle$). A natural question is whether or not these two approaches generally give the same results. That is, for a given f are the approximation errors $\|f - \tilde{f}_N\|$ and $\|f - f_N\|$ of about the same order? We know that

$$\begin{aligned} \alpha \in w\ell_p \quad &\Rightarrow \quad \|f - \tilde{f}_N\| = O(N^{-r}) \qquad \text{for } 1/p = r + 1/2 \qquad \text{and} \\ \tilde{\alpha} \in w\ell_p \quad &\Rightarrow \quad \|f - f_N\| = O(N^{-r}) \qquad \text{for } 1/p = r + 1/2, \end{aligned}$$

and, thus, the problem unquestionably takes place at the coefficient level. In this thesis, we have mainly studied dual-ridgelet expansions because one has a direct access to the ridgelet coefficients α since the ridgelets are known explicitly. It is much more delicate to work out similar estimates for the dual coefficient sequence.

In this direction, it would be interesting to know if both sequences have the same structure; that is, let $\|\cdot\|$ be a norm on the sequence space – like (5.14) for example. Is the norm based on the ridgelet coefficient sequence equivalent to the one based on the dual sequence? Especially, in view of the previous display, one would like to know, for example, if for a fixed $p > 0$,

$$\|\alpha\|_{\ell_p} \asymp \|\tilde{\alpha}\|_{\ell_p}, \tag{7.9}$$

the aim being to transfer knowledge about the size and/or organization of α to $\tilde{\alpha}$. In fact, one direction of (7.9) turns out to be easy. To fix ideas, suppose we are given a frame of $L_2(2\Omega_d)$ and a function f in $L_2(\Omega_d)$, such that $\tilde{\alpha} \in \ell_p$. Let $0 \leq w \leq 1$ be a C^∞ window, supported in $2\Omega_d$, so that its restriction to Ω_d is 1. From the frame decomposition

$$f = fw = \sum_{\gamma \in \Gamma_d} \alpha_\gamma \tilde{\psi}_\gamma = \sum_{\gamma \in \Gamma_d} \tilde{\alpha}_\gamma \psi_\gamma,$$

it follows that

$$\alpha_\gamma = \sum_{\gamma' \in \Gamma_d} \langle w\psi_\gamma, \psi_{\gamma'} \rangle \tilde{\alpha}_{\gamma'} \quad \text{and} \quad \tilde{\alpha}_\gamma = \sum_{\gamma' \in \Gamma_d} \langle w\tilde{\psi}_\gamma, \tilde{\psi}_{\gamma'} \rangle \alpha_{\gamma'}.$$

Now, the estimates obtained in section 5.2.1 imply that for any $p > 0$,

$$\|\alpha\|_{\ell_p} \leq C_p \, \|\tilde{\alpha}\|_{\ell_p},$$

provided that ψ has a sufficiently large number of derivatives and vanishing moments. To establish the converse, it would be sufficient to show that, say,

$$w\tilde{\psi}_\gamma = \sum_{\gamma'} \lambda_{\gamma,\gamma'} \psi_{\gamma'}$$

with

$$\sup_\gamma \sum_{\gamma'} |\lambda_{\gamma,\gamma'}|^p \quad \vee \quad \sup_{\gamma'} \sum_{\gamma} |\lambda_{\gamma,\gamma'}|^p \leq C_p,$$

which expresses that the dual frame has a sparse representation in the original one. The author has proven some results in this direction but they are too fragmentary to be reported.

7.4 Future Work

7.4.1 Nonparametric Regression

We anticipate the translation of our approximation results to statistical estimation results. Such connections have been made clear in section 1.3. For example, as in Donoho and

Johnstone (1995), consider the following white noise model

$$Y_\epsilon(dx) = f(x)dx + \epsilon W(dx), \quad x \in [0,1]^d.$$

Here, f is the object to be recovered and $W(dx)$ is the standard d-dimensional white noise. For a class $\mathcal{F}$ of objects, let $\mathcal{R}_\epsilon(\mathcal{F})$ be the minimax risk

$$\mathcal{R}_\epsilon(\mathcal{F}) = \inf_{\widehat{f}} \sup_{\mathcal{F}} E\|\widehat{f} - f\|_2^2.$$

Following ideas developed in Donoho, Johnstone, Kerkyacharian, and Picard (1995) and Donoho and Johnstone (1995), we expect to derive the asymptotics of $\mathcal{R}_\epsilon$ (as $\epsilon \to 0$) for the classes $R^s_{p,q}$. More importantly, consider estimators of the form

$$\widehat{f} = \sum_{\gamma \in \Gamma_d} \eta_\gamma(\langle \psi_\gamma, Y_\epsilon \rangle)\tilde{\psi}_\gamma,$$

where $(\psi_\gamma)_{\gamma \in \Gamma_d}$ is a nice ridgelet frame and where the functions η are scalar nonlinearities (hard/soft-thresholding, etc.) depending upon ϵ and the parameters s, p, q. We then expect such estimators to be adaptively nearly minimax, possibly within log-like factors, for estimating objects from these classes. In short, thresholding the noisy ridgelet coefficients should be nearly optimal for the estimation of objects exhibiting the spatial inhomogeneities described in Chapter 4.

7.4.2 Curved Singularities

Finally, ridgelets are optimal for representing objects with singularities across hyperplanes (Chapter 5), but they fail to represent efficiently objects with singularities across curved hypersurfaces (Chapter 6). Work in progress investigates possible refinements of the ridgelet construction to handle curved singularities; we hope to report later on this program.

Appendix A

Proofs and Results

Chapter 3

Proof of Proposition 5.

Let α_γ be the ridgelet coefficient $\langle f, \psi_\gamma \rangle$. We recall that $\alpha_\gamma = (R_{\theta_i^j} f * \psi_j)(bk2^{-j})$ (and in case $j = -1$, the convolution is in fact with φ). We will use the polar notation in its standard form; i.e. $f(t,\theta) = f(t\cos\theta, t\sin\theta)$. From Lemma 1, we got

$$\Big| \sum_k |\alpha_{j,\theta_i^j,k}|^2 - \frac{1}{2\pi b_0} \int_{\mathbf{R}} |\hat{f}(\xi,\theta_i^j)|^2 |\hat{\psi}(2^{-j}\xi)|^2 d\xi \Big| \\ \leq \frac{1}{2\pi} \sqrt{\int_{\mathbf{R}} |\hat{f}(\xi,\theta_i^j)|^2 |\hat{\psi}(2^{-j}\xi)|^2 d\xi} \sqrt{\int_{\mathbf{R}} |\hat{f}(\xi,\theta_i^j)|^2 |2^{-j}\xi|^2 |\hat{\psi}(2^{-j}\xi)|^2 d\xi}.$$

Mimicking the argument of Lemma 1 (large sieve principle), we have

$$\begin{aligned} \Big| \sum_i |\hat{f}(\xi,\theta_i^j)|^2 - \frac{2^j}{2\pi\theta_0} \int_{[0,2\pi]} |\hat{f}(\xi,\theta)|^2 d\theta \Big| &\leq \int_{[0,2\pi]} |\hat{f}(\xi,\theta)| |\frac{d}{d\theta} \hat{f}(\xi,\theta)| d\theta \qquad (A.1) \\ &\leq \int_{[0,2\pi]} |\hat{f}(\xi,\theta)| |\xi| \|\nabla \hat{f}(\xi,\theta)\| d\theta \\ &\leq \sqrt{\int_{[0,2\pi]} |\hat{f}(\xi,\theta)|^2 |\xi| d\theta} \sqrt{\int_{[0,2\pi]} \|\nabla \hat{f}(\xi,\theta)\|^2 |\xi| d\theta}, \end{aligned}$$

where the $\nabla \hat{f}$ is the gradient of the 2-dimensional Fourier transform of f.

As a consequence of (A.1), we have a sampling result for the angular variable of the

same flavor.

$$
\begin{aligned}
&\Big| \int \sum_j \sum_i |\hat{f}(\xi,\theta_i^j)|^2 |\hat{\psi}(2^{-j}\xi)|^2 - \frac{2^j}{2\pi\theta_0} \int\int_{[0,2\pi]} \sum_j |\hat{f}(\xi,\theta)|^2 |\hat{\psi}(2^{-j}\xi)|^2 d\theta d\xi \Big| \qquad (A.2)\\
&\leq \int \sum_j \Big| \int \sum_i |\hat{f}(\xi,\theta_i^j)|^2 |\hat{\psi}(2^{-j}\xi)|^2 - \frac{2^j}{2\pi\theta_0} \int_{[0,2\pi]} |\hat{f}(\xi,\theta)|^2 |\hat{\psi}(2^{-j}\xi)|^2 d\theta \Big| d\xi\\
&\leq \int \sum_j \left(\sqrt{\int_{[0,2\pi]} |\xi| |\hat{f}(\xi,\theta)|^2 |\hat{\psi}(2^{-j}\xi)|^2 d\theta} \sqrt{\int_{[0,2\pi]} \|\nabla \hat{f}(\xi,\theta)\|^2 |\xi| |\hat{\psi}(2^{-j}\xi)|^2 d\theta} \right) d\xi\\
&\leq \int \left(\sqrt{\int_{[0,2\pi]} |\xi| |\hat{f}(\xi,\theta)|^2 \sum_j |\hat{\psi}(2^{-j}\xi)|^2 d\theta} \sqrt{\int_{[0,2\pi]} \|\nabla \hat{f}(\xi,\theta)\|^2 |\xi| \sum_j |\hat{\psi}(2^{-j}\xi)|^2 d\theta} \right) d\xi\\
&\leq \sqrt{\int\int_{[0,2\pi]} |\xi| |\hat{f}(\xi,\theta)|^2 \sum_j |\hat{\psi}(2^{-j}\xi)|^2 d\theta d\xi} \sqrt{\int\int_{[0,2\pi]} \|\nabla \hat{f}(\xi,\theta)\|^2 |\xi| \sum_j |\hat{\psi}(2^{-j}\xi)|^2 d\theta d\xi}.
\end{aligned}
$$

Adding the coarse scale term amounts to substituting $\sum_j |\hat{\psi}(2^{-j}\xi)|^2$ by $|\hat{\varphi}(\xi)|^2 + \sum_j |\hat{\psi}(2^{-j}\xi)|^2$ in the last inequality. Recall that by assumption we have

$$|\hat{\varphi}(\xi)|^2 + \sum_{j\geq 0} 2^j |\hat{\psi}(2^{-j}\xi)|^2 = |\xi|,$$

and suppose

$$\sup |\hat{\varphi}(\xi)|^2 + \sum_j |\hat{\psi}(2^{-j}\xi)|^2 \leq B.$$

The last inequality of (A.2) together with the conditions on φ and ψ imply

$$
\begin{aligned}
&\Big| \int \sum_j \sum_i |\hat{f}(\xi,\theta_i^j)|^2 |\hat{\psi}(2^{-j}\xi)|^2 d\xi - \frac{2^j}{2\pi\theta_0} \int_{[0,2\pi]} |\hat{f}(\xi,\theta)|^2 |\hat{\psi}(2^{-j}\xi)|^2 d\theta d\xi \Big|\\
&= \Big| \int \sum_j \sum_i |\hat{f}(\xi,\theta_i^j)|^2 |\hat{\psi}(2^{-j}\xi)|^2 d\xi - \frac{2}{2\pi\theta_0} \|\hat{f}\|^2 \Big|\\
&\leq B \sqrt{\int |\xi| |\hat{f}(\xi,\theta)|^2 d\theta d\xi} \sqrt{\int \|\nabla \hat{f}(\xi,\theta)\|^2 |\xi| d\theta d\xi}\\
&\leq B\sqrt{2} \|\hat{f}\|_2 2 \|\hat{f}\|_2 = 2\sqrt{2} B \|\hat{f}\|_2^2. \qquad (A.3)
\end{aligned}
$$

Again, using A.1, one can check that

$$\Big|\sum_\gamma |\alpha_\gamma|^2 - \frac{1}{2\pi b_0}\int_{\mathbf{R}} \sum_j \sum_i |\hat f(\xi,\theta_i^j)|^2 |\hat\psi(2^{-j}\xi)|^2 d\xi\Big|$$
$$\leq \sum_j \sum_i \frac{1}{2\pi}\sqrt{\int_{\mathbf{R}} |\hat f(\xi,\theta_i^j)|^2|\hat\psi(2^{-j}\xi)|^2 d\xi}\sqrt{\int_{\mathbf{R}} |\hat f(\xi,\theta_i^j))|^2|2^{-j}\xi|^2|\hat\psi(2^{-j}\xi)|^2 d\xi}$$
$$\leq \frac{1}{2\pi}\sqrt{\int_{\mathbf{R}} \sum_j \sum_i |\hat f(\xi,\theta_i^j)|^2|\hat\psi(2^{-j}\xi)|^2 d\xi}\sqrt{\int_{\mathbf{R}} \sum_j \sum_i |\hat f(\xi,\theta_i^j)|^2|2^{-j}\xi|^2|\hat\psi(2^{-j}\xi)|^2 d\xi}.$$

But we have just proved (A.3) that

$$\int_{\mathbf{R}} \sum_j \sum_i |\hat f(\xi,\theta_i^j)|^2|\hat\psi(2^{-j}\xi)|^2 d\xi \leq \frac{2}{2\pi\theta_0}\|\hat f\|_2^2 + 2\sqrt{2}B\|\hat f\|_2^2.$$

Similarly, one can show without effort that

$$\int_{\mathbf{R}} \sum_j \sum_i |\hat f(\xi,\theta_i^j)|^2|2^{-j}\xi|^2|\hat\psi(2^{-j}\xi)|^2 d\xi \leq \frac{2B'}{2\pi\theta_0}\|\hat f\|_2^2$$

for some constant B' only depending on ψ. Therefore, we can deduce from all this that

$$\Big|\sum_\gamma |\alpha_\gamma|^2 - \frac{1}{2\pi b_0}\int_{\mathbf{R}} \sum_j \sum_i |\hat f(\xi,\theta_i^j)|^2|\hat\psi(2^{-j}\xi)|^2 d\xi\Big| \leq C\,\frac{1}{\theta_0}\|\hat f\|_2^2 \tag{A.4}$$

for some C.

Finally, combining (A.3) and (A.4) gives

$$\Big|\sum_\gamma |\alpha_\gamma|^2 - \frac{2}{(2\pi)^2 b_0\theta_0}\|\hat f\|_2^2\Big| \leq \frac{C}{2\pi\theta_0}\|\hat f\|_2^2 + \frac{C}{b_0}\|\hat f\|_2^2,$$

which is the desired result. ∎

Chapter 4

Proof of Proposition 8. The fact that $R^s_{p,q}(\mathbf{R}^d)$ is a normed space is trivial. We now prove the completeness. Let $\{f_n\}_{n\geq 0}$ be a Cauchy sequence in $R^s_{p,q}(\mathbf{R}^d)$. Then, of course, $\{f_n\}_{n\geq 0}$ is a Cauchy sequence in $L_1(\mathbf{R}^d)$ and, therefore, converges to a limit f in L_1. Let $w_j(f)(u,b)$ (resp. $v(u,b)$) be $(R_u f * \psi_j)(b)$ (resp. $(R_u f * \varphi)(b)$). It follows that $w_j(f_n)(u,b)$ converges

to $w_j(f)(u,b)$ a.e. (and similarly for $v(f_n)(u,b)$). On the other hand, $\{w_j(f_n)(u,b)\}_{n\geq 0}$ is a Cauchy sequence in $L_p(\mathbf{S}^{d-1}\times\mathbf{R})$ and therefore the limiting function, which coincides with $w_j(f)(u,b)$, is in $L_p(\mathbf{S}^{d-1}\times\mathbf{R})$. Now, it follows by standard arguments that f belongs to $R^s_{p,q}(\mathbf{R}^d)$ and that f_n converges in $R^s_{p,q}(\mathbf{R}^d)$ to f. Hence, $R^s_{p,q}(\mathbf{R}^d)$ is complete.

Chapter 6

Lemma 16 *The operator U_d defined by*

$$(U_d g)(t) = \int_0^\infty (r-t)_+^{(d-3)/2} g(r)\, dr$$

is bounded from $B^s_{p,q}(0,1)$ to $B^{s+(d-1)/2}_{p,q}(0,1)$.

Proof of Lemma. Let $(\varphi_{j_0,k}), \psi_{j,k}$, $0\leq k<2^j$, $j\geq j_0$, be a wavelet basis on the interval $(0,1)$. We will assume that $(\varphi_{j_0,k}), \psi_{j,k}$, are R times differentiable and are of compact support (support width $\leq C2^{-j}$.) In addition, we will suppose that the $\psi_{j,k}$'s have vanishing moments through order D. We will denote by $\beta_{j_0,k}=\langle f,\varphi_{j_0,k}\rangle$ and $\alpha_{j,k}=\langle f,\psi_{j,k}\rangle$ the wavelet coefficients of f. We recall that each Besov space $B^s_{p,q}(0,1)$ with $(1/p-1)<s<\min(R,D)$ and $0<p,q\leq\infty$ is characterized by its coefficients in a sense that

$$\|f\|_{B^s_{p,q}(0,1)} \equiv \|(\beta_{j_0,k})_k\|_{\ell_p} + \|(2^j 2^{j(1/2-1/p)}(\sum_k |\alpha_{j,k}|^p)^{1/p})_{j\geq j_0}\|_{\ell_q} \tag{A.5}$$

is an equivalent norm to the norm of $B^s_{p,q}(0,1)$.

Further, suppose one wants to bound the norm of an operator $T: B^s_{p,q}(0,1)\to B^\sigma_{p,q}(0,1)$ for all possible choices of $1\leq p,q\leq\infty$. (The norm being defined as $\sup \|T_f\|_{B^\sigma_{p,q}}/\|f\|_{B^s_{p,q}}$.) The theory of interpolation tells that it is sufficient to check the boundedness for cases $(p,q)\in\{(1,1),(\infty,\infty),(1,\infty),(\infty,1)\}$. That is what we shall verify for U_d. Now let $g_{j_0}(t)=\int (y-t)_+^{(d-3)/2}\varphi_{j_0,0}(y)$ and $g_{j_0}(t)=\int (y-t)_+^{(d-3)/2}\psi_{j_0,0}(t)$. We have for $t\in\mathbf{R}$

$$g_{j_0}(t) = \int (y-t)_+^{(d-3)/2} 2^{j_0/2}\varphi(2^{j_0}y).$$

Moreover, for $\ell\leq\min(R,D)$ we have

$$g_{j_0}(t) = \int (y-t)_+^{(d-3)/2} 2^{j_0/2}\psi(2^{j_0}y).$$

Then of course,

$$|\tau(\lambda,\lambda')| \leq 2^{-\min(j,j')(d-1)/2} 2^{-|j-j'|(n+1/2)} \left(1 + 2^{\min(j,j')}|k2^{-j} - k'2^{-j'}|\right)^{-2}.$$

From this last inequality, it follows that

$$\sup_{k'\in\Lambda_{j'}} \sum_{k\in\Lambda_j} |\tau(\lambda,\lambda')| \leq \begin{cases} C2^{-j'(d-1)/2}2^{-(j-j')(n-1/2)} & j' \leq j \\ C2^{-j(d-1)/2}2^{-(j-j')(n+1/2)} & j' \geq j \end{cases} \tag{A.6}$$

and

$$\sup_{k\in\Lambda_j} \sum_{k'\in\Lambda_{j'}} |\tau(\lambda,\lambda')| \leq \begin{cases} C2^{-j'(d-1)/2}2^{-(j-j')(n+1/2)} & j' \leq j \\ C2^{-j(d-1)/2}2^{-(j'-j)(n-1/2)} & j' \geq j \end{cases}. \tag{A.7}$$

Case $(1,1)$. In this case we have that

$$\|U_d\|_{B^s_{1,1},B^{s+(d-1)/2}_{1,1}} \leq \sup_{\lambda'} \sum_{\lambda} 2^{j(s+(d-1)/2-1/2)}2^{-j'(s-1/2)}|\tau(\lambda,\lambda')|.$$

But (A.7) yields

$$\begin{aligned} &\sup_{\lambda'} \sum_{\lambda} 2^{j(s+(d-1)/2-1/2)}2^{-j'(s-1/2)}|\tau(\lambda,\lambda')| \\ &\quad\leq \sup_{\lambda'} \sum_{j_0\leq j\leq j'} 2^{-(j'-j)(s-1/2)}2^{-(j-j')(n+1/2)} + \sum_{j>j'} 2^{(j-j')(s+(d-1)/2-1/2)}2^{-(j-j')(n-1/2)} \\ &\quad\leq C, \end{aligned}$$

where the last inequality holds as long as $n > s+(d-1)/2$. This fact implies the boundedness for $(1,1)$.

Case (∞,∞). This time,

$$\|U_d\|_{B^s_{\infty,\infty},B^{s+(d-1)/2}_{\infty,\infty}} \leq \sup_{\lambda} \sum_{\lambda'} 2^{j(s+(d-1)/2+1/2)}2^{-j'(s+1/2)}|\tau(\lambda,\lambda')|.$$

Now, (A.6) gives

$$\begin{aligned}
&\sup_{\lambda}\sum_{\lambda'} 2^{j(s+(d-1)/2+1/2)}2^{-j'(s+1/2)}|\tau(\lambda,\lambda')| \\
&\quad\leq \sup_{\lambda}\sum_{j_0\leq j'\leq j} 2^{(j-j')(s+(d-1)/2+1/2)}2^{-(j-j')(n+1/2)} + \sum_{j'>j} 2^{-(j'-j)(s+1/2)}2^{-(j'-j)(n-1/2)} \\
&\quad\leq C,
\end{aligned}$$

where, again, the last inequality is verified if $n > s + (d-1)/2$. We obtain the boundedness for (∞,∞).

Case $(1,\infty)$. This time

$$\|U_d\|_{B^s_{1,\infty},B^{s+(d-1)/2}_{1,\infty}} \leq \sup_{j'\geq j_0}\sum_{j\geq j_0} 2^{j(s+(d-1)/2-1/2)}2^{-j'(s-1/2)} \sup_{k\in\Lambda_j}\sum_{k'\in\Lambda_{j'}}|\tau(\lambda,\lambda')|.$$

Again, similar calculations as in the above two cases give

$$\begin{aligned}
&\sup_{j'\geq j_0}\sum_{j\geq j_0} 2^{j(s+(d-1)/2-1/2)}2^{-j'(s-1/2)} \sup_{k\in\Lambda_j}\sum_{k'\in\Lambda_{j'}}|\tau(\lambda,\lambda')| \\
&\quad\leq \sum_{j_0\leq j\leq j'} 2^{(j-j')(s-1/2)}2^{-(j'-j)(n-1/2)} + \sum_{j>j'} 2^{(j-j')(s+(d-1)/2-1/2)}2^{-(j-j')(n+1/2)} \\
&\quad\leq C,
\end{aligned}$$

where again, the last inequality is verified if $n > s + (d-3)/2$.

The case $(\infty,1)$ is analogous. The lemma is proved. ∎

References

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory*, *39*, 930–945.

Benveniste, A., and Zhang, Q. (1992). Wavelet networks. *IEEE Transactions on Neural Networks*, *3*, 889–898.

Bernier, D., and Taylor, K. F. (1996). Wavelets from square-integrable representations. *SIAM J. Math. Anal*, *27*, 594–608.

Boas, R. P., Jr. (1952). *Entire functions.* New York: Academic Press.

Candes, E. J. (1996). Harmonic analysis of neural netwoks. To appear in *Applied and Computational Harmonic Analysis.*

Cheng, B., and Titterington, D. M. (1994). Neural networks: a review from a statistical perspective. With comments and a rejoinder by the authors. *Stat. Sci.*, *9*, 2–54.

Conway, J. H., and Sloane, N. J. A. (1988). *Sphere packings, lattices and groups.* New York: Springer-Verlag.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems*, *2*, 303–314.

Daubechies, I. (1992). *Ten lectures on wavelets.* Philadelphia, PA: Society for Industrial and Applied Mathematics.

Daubechies, I., Grossmann, A., and Meyer, Y. (1986). Painless nonorthogonal expansions. *J. Math. Phys.*, *27*, 1271–1283.

Deans, S. R. (1983). *The Radon transform and some of its applications.* John Wiley & Sons.

DeVore, R. A., Oskolkov, K. I., and Petrushev, P. P. (1997). Approximation by feed-forward neural networks. *Ann. Numer. Math.*, *4*, 261–287.

DeVore, R. A., and Temlyakov, V. N. (1996). Some remarks on greedy algorithms. *Adv. Comput. Math.*, *5*, 173–187.

Donoho, D. L. (1993). Unconditional bases are optimal bases for data compression and for statistical estimation. *Applied and Computational Harmonic Analysis*, *1*, 100–115.

Donoho, D. L. (1996). Unconditional bases and bit-level compression. *Applied and Computational Harmonic Analysis*, *3*, 388–392.

Donoho, D. L. (1997). *Fast ridgelet transform in two dimensions* (Tech. Rep.). Department of Statistics, Stanford CA 94305–4065: Stanford University.

Donoho, D. L., and Johnstone, I. M. (1989). Projection-based approximation and a duality with kernel methods. *Ann. Statist.*, *17*, 58–106.

Donoho, D. L., and Johnstone, I. M. (1995). *Empirical atomic decomposition.*

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B*, *57*, 301–369.

Duffin, R. J., and Schaeffer, A. C. (1952). A class of nonharmonic Fourier Series. *Trans. Amer. Math. Soc.*, *72*, 341-366.

Duflo, M., and Moore, C. C. (1976). On the regular representation of a nonunimodular locally compact group. *J. Functional Analysis*, *21*, 209–243.

Feichtinger, H. G., and Gröchenig, K. (1988). A unified approach to atomic decompositions via integrable group representations. In *Function spaces and applications (Lund, 1986).* Berlin-New York: Springer.

Frazier, M., Jawerth, B., and Weiss, G. (1991). *Littlewood-Paley theory and the study of function spaces* (Vol. 79). Providence, RI: American Math. Soc.

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

Härdle, W. (1990). *Applied nonparametric regression.* Cambridge, England: Cambridge University Press.

Hasminskii, R., and Ibragimov, I. (1990). On density estimation in the view of Kolmogorov's ideas in approximation theory. *J. Amer. Statist. Assoc.*, *18*, 999–1010.

Holschneider, M. (1991). Inverse Radon transforms through inverse wavelet transforms. *Inverse Problems*, *7*, 853–861.

Huber, P. J. (1985). Projection pursuit, with discussion. *Ann. Statist.*, *13*, 435–525.

Jones, L. K. (1992a). A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann. Statist.*, *20*, 608–613.

Jones, L. K. (1992b). On a conjecture of Huber concerning the convergence of projection pursuit regression. *Ann. Statist.*, *15*, 880–882.

Katznelson, Y. (1968). *An introduction to harmonic analysis.* New York: Wiley.

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

Mallat, S., and Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, *41*, 3397–3415.

Meyer, Y. (1992). *Wavelets and operators.* Cambridge University Press.

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

Mhaskar, H. N., and Micchelli, C. A. (1995). Degree of approximation by neural and translation networks with a single hidden layer. *Adv. in Appl. Math.*, *16*, 151–183.

Montgomery, H. L. (1978). The analytic principle of the large sieve. *Bull. Amer. Math. Soc.*, *84*, 547–567.

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

Natterer, F. (1986). *The mathematics of computerized tomography.* B. G. Teubner; John Wiley & Sons.

Pati, Y. C., and Krishnaprasad, P. S. (1993). Analysis and synthesis of feedforward neural networks using discrete affine wavelet transformations. *IEEE Transactions on Neural Networks*, *4*, 73–85.

Peyrin, F., Zaim, M., and Goutte, R. (1993). Construction of wavelet decompositions for tomographic images. *J. Math. Imaging Vision*, *3*, 105–122.

Plancherel, M., and Pólya, G. (1938). Fonctions entières et intégrales de Fourier multiples. *Commentarii Math. Helv.*, *10*, 110–163.

Stein, E. M. (1970). *Singular integrals and differentiability properties of functions* (Vol. 30). Princeton, N.J.: Princeton University Press.

Triebel, H. (1983). *Theory of function spaces.* Basel: Birkhäuser Verlag.

Wagner, G. (1993). On a new method for constructing good point sets on spheres. *Discrete Comput. Geom.*, *9*, 111–129.

Young, R. M. (1980). *An introduction to nonharmonic Fourier series.* New York: Academic Press.