Ridgelets: Estimating with Ridge Functions

Emmanuel J. Candès

Department of Statistics
Stanford University
Stanford, California 94305–4065
emmanuel@stat.stanford.edu

Feedforward neural networks, projection pursuit regression, and more generally, estimation via ridge functions 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. To address problems inherent to these methods – ranging from the construction of neural networks to their efficiency and capability – Candes (1999d) developed a new system that allows the representation of arbitrary functions as superpositions of specific ridge functions, the *ridgelets*.

In a nonparametric regression setting, this article suggests expanding noisy data into a ridgelet series and applying a scalar nonlinearity to the coefficients (dumping); this is unlike existing approaches based on stepwise additions of elements. The procedure is simple, constructive, stable and spatially adaptive – and fast algorithms have been developed to implement it.

The ridgelet estimator is nearly optimal for estimating functions with certain kinds of spatial inhomogeneities. In addition, ridgelets help to identify new classes of estimands – corresponding to a new notion of smoothness – that are well suited for ridge functions estimation. While the results are stated in a decision theoretic framework, numerical experiments are also presented to illustrate the practical performance of the methodology.

Key Words and Phrases. Nonparametric regression, ridgelets, ridge functions, Projection Pursuit Regression, minimax decision theory, Radon transform, spatial inhomogeneities, edges, thresholding of ridgelet coefficients.

Acknowledgments. I am especially grateful to David Donoho for serving as my adviser and for many fruitful discussions. It is a pleasure to acknowledge conversations with Iain Johnstone. This research was supported by National Science Foundation grants DMS 95–05151 and DMS 98–72890 (KDI) and by AFOSR MURI 95–P49620–96–1–0028.

1 Introduction

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

$$Y_i = f(X_i) + \epsilon_i, \tag{1.1}$$

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 unit cube $[0, 1]^d$ is mostly empty, as discussed in the excellent paper of Friedman and Stuetzle (1981). In terms of estimation bounds, roughly speaking, the curse says, for example, that unless you have an enormous sample size N, you will get a poor mean-squared error.

1.1 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,

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

where the u_j's are vectors of unit length, i.e. $\|u_j\| = 1$. In its abstract version, the approximation process operates in a stepwise and greedy fashion. At stage m, it augments the fit f_{m-1} by adding a ridge function $g_m(u_m^T x)$ where u_m and g_m are chosen so that $g_m(u_m^T x)$ best approximates the residuals $f(x) - f_{m-1}(x)$.

For the sampling case and in a regression setup, there is a statistical analogy of the aforementioned greedy procedure. At stage m, the fit f_{m-1} is augmented by adding a ridge function $g_j(u_j^T x)$ obtained as follows: calculate the residuals of the $(m-1)$th fit $r_i = Y_i - \sum_{j=1}^{m-1} g_j(u_j^T X_i)$; and for a fixed direction u, plot the residuals r_i against $u^T x_i$; fit a smooth curve g and choose the best direction u, so as to minimize the residuals sum of squares $\sum_i (r_i - g(u^T X_i))^2$. 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: $|u^T x - t| \leq h$, thus avoiding the problems relative to the inherent sparsity of the sample.

1.2 Neural Networks

Neural networks 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). The idea is to approximate the regression surface by a superposition of ridge functions of the form

$$f = \sum_{j=1}^{m} \alpha_j \rho(k_j^T x - b_j), \tag{1.2}$$

where the m terms in the sum are called neurons; the α_j and b_j are scalars; and the k_j are d-dimensional vectors. In that field, ρ is usually sigmoidal, that is, bounded and monotone. A prevailing choice is the logistic function $\rho(t) = 1/(1 + e^{-t})$.

As far as constructing the approximation, the relaxed greedy algorithm is a popular approach: starting from $f_0(x) = 0$, it operates in a stepwise fashion running through steps $i = 1, \dots m$; we inductively define

$$f_i = \alpha^* f_{i-1} + (1 - \alpha^*)\rho(k^{*T} x - b^*), \tag{1.3}$$

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

$$\arg \min_{0 \le \alpha \le 1} \arg \min_{(k,b) \in \mathbb{R}^n \times \mathbb{R}} \|f - \alpha f_{i-1} + (1 - \alpha)\rho(k^T x - b)\|_2. \tag{1.4}$$

Thus, at the i-th stage, the algorithm substitutes to f_{i-1} a convex combination involving f_{i-1} and a new term, a neuron $\rho(k^T x - b)$, that results in the largest decrease in approximation error (1.4). In the sampling case, the L_2 norm $\|\cdot\|$ is replaced by the discrete euclidian norm.

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 $\rho(k^T x - b)$ resembles a step function when the norm $\|k\|$ of the weights is large. On the other hand, PPR can make better use of projections because of the freedom to choose a different profile g at each step.

1.3 Problems

This approach (approximating the regression surface by a sum of ridge functions) poses new and challenging questions both at a practical and theoretical level, ranging from the construction of neural networks to their efficiency and capability. We now detail some of these questions.

1. *How to construct neural networks?* In practice, minimizing (1.3) is rather problematic as the $(d + 2)$-dimensional error surface, as a function of the parameters, may exhibit several local minima. This fact is more and more acknowledged in the literature; we quote Barron

(Cheng and Titterington, 1994) "there is no known algorithm for network estimation that is proven to produce accurate estimates in a feasible amount of computation time." This statement is not pure rhethoric. Actually, there is an emergence of negative results about the computational feasibility of fitting neural nets. In a nutshell, the aim of this pioneering work is to show that it is impossible to design algorithms running in polynomial time that would produce 'accurate estimates' (the exact formulation is that this problem is NP-hard and it is a conjecture that NP-hard problems cannot be solved in polynomial time). The papers "The computational intractability of training sigmoidal neural networks" by Jones (1997) and "On the infeasibility of training neural networks with small mean-squared error" by Vu (1998) are important references in this area.

Even if one is willing to ignore the difficulty of implementing a stepwise addition of elements, we may wonder about the efficiency of such a procedure. It is well known that such procedures may have weak estimation properties because of their greedy nature; for instance, the inability to look ahead may cause initial errors that the algorithm keeps on trying to correct.

2. *Neural nets for which regression surface?* It would be of interest to be able to identify classes of functions for which neural networks are more efficient than other methods of estimation or, more ambitiously, a class $\mathcal{F}$ for which it could be proved that linear combinations of carefully selected ridge functions are minimax or nearly minimax over $\mathcal{F}$. In less technical terms, we would like to know for which estimands ridge function approximation and/or estimation makes much sense.

3. *Which rates should we expect?* There are very few results about quantitative rates of estimation. For instance, what is the performance of estimators of the form:

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

(where the paramaters α_j, k_j, b_j are estimated from data) in terms of the mean squared-error

$$MSE(f, \hat{f}) = E\|f - \hat{f}\|_2^2?$$

1.4 Overview

This paper is about these important questions. While existing approaches are based on stepwise construction of approximations, we develop a new approach based on a new transform, namely, the ridgelet transform introduced by Candes (1999d). The ridgelet transform represents quite general functions as superpositions of ridge functions in a stable and concrete way (section 2) and the point of this paper is to show how one can use this representation to construct estimators and derive precise estimation bounds.

When presented with noisy data, we suggest to expand the data into a ridgelet series and apply a scalar nonlinearity (soft or hard thresholding) to the coefficients (section 3). We want to investigate the performance of this simple, stable and constructive procedure.

Roughly speaking, our estimator is optimal for estimating multivariate regression surfaces that exhibit specific sorts of high-dimensional spatial inhomogeneities (section 5). Following this observation, we will introduce a new notion of smoothness that models these spatial inhomogeneities; it will be shown that thresholding the ridgelet series is nearly minimax for these new smoothness classes (section 6). In other words, projection based approaches make a lot of sense for estimating objects from these classes.

In addition, we will try to argue that the ridgelet transform gives decisive insights about the limitations of neural networks: as a surprising and remarkable example, the estimation of radial functions with projection based approaches will be discussed (section 7).

Finally, some numerical experiments will illustrate the power of these new ideas (section 8). The discussion section 9 will survey some extensions of the present work and identify areas for future research.

2 Ridgelets

In this section, $\hat{g}$ will denote the Fourier transform of g

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

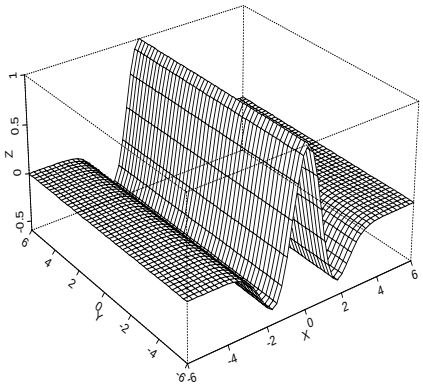
In d dimensions, the ridgelet construction starts with a univariate function ψ satisfying an oscillatory condition, namely,

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

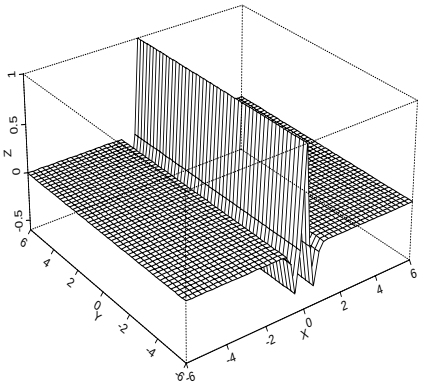
a ridgelet is a function of the form:

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

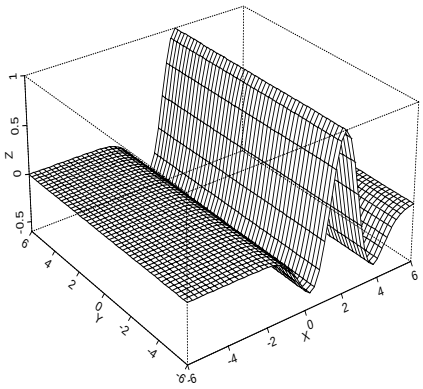
where a and b are scalar parameters and u is a vector of unit length. Of course, a ridgelet is a ridge function and resembles a neuron but for the oscillatory behavior of the profile (the profile of a neuron is sigmoidal, i.e. monotone increasing). A ridgelet has a scale a, an orientation u, and a location parameter b. Ridgelets are concentrated around hyperplanes: roughly speaking the ridgelet (2.3) is supported near the strip $\{x,\, |u^Tx - b| \le a\}$. Ridgelets are pictured on Figure 2 for various values of these parameters.



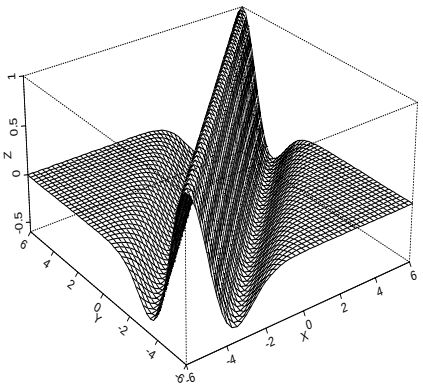
[Original ridgelet]



[After rescaling]



[After shifting]



[After rotation]

Figure 1: Ridgelets.

The nice thing is that one can represent any function as a superposition of these ridgelets: we define the ridgelet coefficient as

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

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

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

where $d\mu(a,u,b) = da/a^{d+1}\, du\, db$ (du being the uniform measure on the sphere); furthermore, this formula is stable as one has a Parseval relation

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

Similar to the continuous transform, there is a discrete transform. One can find a discrete set of parameters $(a_i, b_i, u_i)_{i\in\mathcal{I}}$ such that the collection $(\psi_{a_i,b_i,u_i})_{i\in\mathcal{I}}$ satisfies the following property: there exist two constants A and B such that for any f supported in the unit cube $[0,1]^d$ with finite L^2 norm, we have

$$A\,\|f\|^2 \le \sum_{i\in\mathcal{I}} |\langle f, \psi_{a_i,b_i,u_i}\rangle|^2 \le B\,\|f\|^2. \tag{2.7}$$

The previous equation says that the datum of the ridgelet transform at the points $(a_i, b_i, u_i)_{i\in\mathcal{I}}$ suffices to reconstruct the function perfectly. In this sense, this is analogous to the Shannon sampling theorem for the reconstruction of bandlimited functions. Indeed, standard arguments show that there exists a dual collection $(\widetilde{\psi_{a_i,b_i,u_i}})_{i\in\mathcal{I}}$ with the property

$$f = \sum_{i\in\mathcal{I}} \langle f, \widetilde{\psi_{a_i,b_i,u_i}}\rangle \psi_{a_i,b_i,u_i} = \sum_{i\in\mathcal{I}} \langle f, \psi_{a_i,b_i,u_i}\rangle \widetilde{\psi_{a_i,b_i,u_i}}, \tag{2.8}$$

where the notation $\langle\cdot,\cdot\rangle$ stands here and throughout the remainder of this paper for the usual inner product of L_2: $\langle f, g\rangle = \int f(x)g(x)dx$.

The discretization is as follows:

$$\{\psi_i(x) = 2^{j/2}\psi(2^j u_{j,\ell}^T x - k),\ j \ge j_0, u_{j,\ell} \in \Sigma_j, k \in \mathbb{Z}\}. \tag{2.9}$$

Ridgelets are directional and, here, the interesting aspect is the discretization of the directional variable u; this variable is sampled at increasing resolution so that at scale j, the discretized set is a net of nearly equispaced points at a distance of order 2^{-j}; a detailed exposition on the ridgelet construction is given in Candes (1999d). As in (2.9), we will often use the compact notation ψ_i ($i \in \mathcal{I}$) and, therefore, we will keep in mind that the index runs through an enumeration of the triples (j,ℓ,k). It will then be handy to use the notation $j(i)$ to refer to the scale of ψ_i.

It will be more convenient to work with coarse scale element and we will choose to work with a frame of the form:

$$\{\varphi(u_\ell^T x - k),\ 2^{j/2}\psi(2^j u_{j,\ell}^T x - k),\ j \geq 0, u_{j,\ell} \in \Sigma_j, k \in \mathbb{Z}\}, \tag{2.10}$$

2.1 Why a discrete transform?

Various completeness theorems are known for the set of neurons $\mathcal{D}_{NN} = \{\rho(k^T x - b),\ k \in \mathbb{R}^d, b \in \mathbb{R}\}$, see Cybenko (1989), for example. This says that for a given a square integrable function f supported in the unit cube, there exist finite linear combinations of neurons that are arbitrarily close to f: i.e., for any $\epsilon > 0$, one can find parameter values $(k_j, b_j)_{1 \leq j \leq J}$ such that

$$\|f - \sum_{j=1}^{J} \alpha_j \rho(k_j^T x - b_j)\|_2 < \epsilon.$$

In the introduction we have described a popular approach – the greedy algorithm – to compute these approximations. At each step, one would need to solve an optimization problem of the form

$$\min_{0 \leq \alpha \leq 1} \min_{(k,b) \in \mathbb{R}^n \times \mathbb{R}} \|f - \alpha f_{i-1} + (1-\alpha)\rho(k^T x - b)\|; \tag{2.11}$$

and in any real implementation, one would probably need to restrict the search for a minimum over a grid. What are the properties of a restricted search? Is there a grid preserving the completeness property? If so, what is the proper spacing of this grid? In other words, what is the real complexity of the search (2.11)? In our opinion, the discretization (2.9) gives a precise answer to these questions.

3 The paradigm

The aim of this paper is to discuss the use of the ridgelet representation for statistical estimation; the object of this section is to spell out the statistical model that will furnish the framework in which the results of the following sections are derived.

As in Ibragimov and Hasminskii (1981) or Efroimovich and Pinsker (1982), consider the following white noise model:

$$Y_\epsilon(dx) = f(x)dx + \epsilon W(dx), \quad x \in [0,1]^d. \tag{3.1}$$

Here, f is the object to be recovered and $W(dx)$ is the standard d-dimensional white noise. We will measure the performance of an estimator $\hat{f}$ by the classical integrated mean-squared error

$$MSE(f, \hat{f}) = E\|\widehat{f} - f\|^2_{L_2[0,1]^d}. \tag{3.2}$$

For a class $\mathcal{F}$ of objects, let $\mathcal{R}_\epsilon(\mathcal{F})$ be the minimax mean-squared error in the white noise model

$$\mathcal{R}_\epsilon(\mathcal{F}) = \inf_{\hat{f}} \sup_{\mathcal{F}} E\|\widehat{f} - f\|^2_{L_2[0,1]^d}, \tag{3.3}$$

where of course the estimates $\hat{f}$ are restricted to be obtained through measurable procedures, i.e. $\hat{f} = \widehat{F}(Y_\epsilon)$, with $\widehat{F}$ measurable.

The white noise model (3.1) is standard in the literature of mathematical statistics. The justification of this continuous setup is that it may be viewed as the limit of a number of nonparametric discrete models, see Johnstone (1999) for details. In the discussion section, we will comment, however, on the limits of this model.

3.1 The Sequence Model

A now classical approach to the study of nonparametric problems of the form (3.1)–(3.3) is to, first, transform the data and, second, analyze and/or solve the problem obtained after transformation, the latter problem being hopefully much easier than the original one. This approach has already proven to be very successful; see Pinsker (1980), for example, where the estimation problem is solved by looking at the estimation of the Fourier coefficients of the function f to be recovered and Donoho et al. where the wavelet coefficients are to be estimated. Our approach will be similar as we study the estimation of the ridgelet coefficients.

Projecting the white noise model (3.1) onto the ridgelet frame $(\psi_i)_{i\in\mathcal{I}}$ gives rise to a sequence space model obtained as follows: for any element of the frame ψ_i, we calculate the noisy coefficient $y_i = \langle Y_\epsilon, \psi_i\rangle$; it is clear that y_i is a Gaussian random variable with mean $\theta_i = \langle f, \psi_i\rangle$ and variance $\epsilon^2\|\psi_i\|_2^2 = \epsilon^2\sigma_i^2$. In other words, we have the following model:

$$y_i = \theta_i + \epsilon z_i, \qquad i \in \mathcal{I}, \tag{3.4}$$

where for a fixed and finite subset $I \subset \mathcal{I}$, $\{z_i\}_{i\in I}$ is a Gaussian vector with mean 0 and covariance matrix V, the Gramm matrix of the ridgelets $V_{i,j} = \langle \psi_i, \psi_j\rangle$. Now suppose that we are estimating the ridgelet coefficient θ from the data; i.e., we are considering an estimator of the form

$$\widehat{f} = \sum_{i\in\mathcal{I}} \widehat{\theta}_i\, \tilde{\psi}_i. \tag{3.5}$$

Then the lemma stated below implies that

$$\|\widehat{f} - f\|_2^2 \le A^{-1}\, \|\widehat{\theta} - \theta\|^2_{\ell_2(\mathcal{I})}, \tag{3.6}$$

where A is the constant appearing on the left-hand side of (2.7). Therefore, control of the risk $E\|\widehat{\theta} - \theta\|^2_{\ell_2(I)}$ at the coefficient level gives control of the integrated mean-squared error $E\|\widehat{f} - f\|_2^2$. As we will see, this observation is a key fact in establishing upper estimation bounds.

Lemma 3.1 *Let $(a_i)_{i\in\mathcal{I}}$ be a sequence in ℓ_2 and let*

$$\tilde{f} = \sum_{i\in\mathcal{I}} a_i \tilde{\psi}_i,$$

then we have

$$\|\tilde{f}\|_2^2 \leq A^{-1}\|a\|_{\ell_2}^2.$$

Proof of Lemma. We let $\tilde{F}$ be the synthesis operator defined by $\tilde{F}a = \sum a_i\tilde{\psi}_i$ and F be the analysis operator $Ff = (\langle f, \psi_i\rangle)_{i\in\mathcal{I}}$. The property (2.7) gives

$$\|\tilde{f}\|^2 = \|\tilde{F}a\|^2 \leq A^{-1}\|F\,\tilde{F}a\|_{\ell_2}^2.$$

Now, it is not hard to see that $F\,\tilde{F}$ is the orthogonal projector onto the range of F and has, therefore, a norm (as an operator from ℓ_2 onto itself) bounded by 1. Consequently, we have

$$\|\tilde{f}\|^2 \leq A^{-1}\|F\,\tilde{F}a\|_{\ell_2}^2 \leq A^{-1}\|a\|_{\ell_2}^2,$$

which is what needed to be shown. ∎

3.2 Ridgelet shrinkage

In the following sections, we will mostly consider shrinkage estimators, i.e. where the $\hat{\theta}_i$'s are obtained by applying some scalar nonlinearities (hard/soft-thresholding, etc.) to the noisy coefficients $y_i = \langle\psi_i, Y_\epsilon\rangle$: that is,

$$\hat{\theta}_i = \eta_i(y_i) = \eta_i(\langle\psi_i, Y_\epsilon\rangle),$$

yielding simple estimates of the form

$$\widehat{f} = \sum_{i\in\mathcal{I}} \eta_i(\langle\psi_i, Y_\epsilon\rangle)\tilde{\psi}_i, \tag{3.7}$$

. We stress that the shrinkage procedure yielding $\widehat{f}$ (3.7) is well defined, constructive and stable. The object of the next sections is to study the performance of this estimator.

4 Abstract statistical estimation

The goal of this section is to prove a lemma that will help to establish the main forthcoming results of the paper. The material presented here closely follows the concept of oracle inequalities developed by Donoho and Johnstone (1994).

Suppose that we have the following problem:

$$y_i = \theta_i + \epsilon z_i, \qquad i \in \mathcal{I}, \tag{4.1}$$

where for any finite subset $I \subset \mathcal{I}$, $\{z_i\}_{i\in I}$ is a Gaussian vector with mean 0 and covariance matrix $V_{i,j}$. In this section, $\mathcal{I}$ might be finite or countable. This model is fairly standard as we wish to estimate the mean θ of a Gaussian vector. Several authors have studied this problem and an excellent account may be found in Johnstone (1999). In particular, it is now well established that the quality of the estimation is linked to the sparsity of the vector θ.

We introduce some notation. Let η_S denote the soft threshold nonlinearity

$$\eta_{ST}(y, \lambda) = \mathrm{sgn}(y)\,(|y| - \lambda)_+ \tag{4.2}$$

and $r_S(\lambda, \mu)$ the risk of the latter rule: that is

$$r_S(\epsilon; \lambda, \mu) = E\left[\eta_S(Y, \lambda) - \mu\right]^2, \quad Y \sim N(\mu, \epsilon^2).$$

(In the case $\epsilon = 1$, we will simply write $r_S(\lambda, \mu)$.)

We borrow the following lemma from Johnstone (1999).

Lemma 4.1 *Let $\bar{r}(\lambda, \mu) = \min\{r_S(\lambda, 0) + \mu^2, 1 + \lambda^2\}$. Then for any choice of threshold λ and $\mu \in \mathbb{R}$,*

$$\frac{1}{2}\bar{r}(\lambda, \mu) \le r_S(\lambda, \mu) \le \bar{r}(\lambda, \mu). \tag{4.3}$$

There is a useful corollary to this lemma:

Corollary 4.2 *Let $\lambda_\delta = \sqrt{2\log\delta^{-1}}$. Then*

$$r_S(\lambda_\delta, \mu) \le (1 + 2\log\delta^{-1})\,(\delta + \mu^2 \wedge 1).$$

Now let $Y \sim N(\mu, \epsilon^2)$ and $\eta_S(Y, \lambda)$ be a soft thresholding estimator with parameter λ: a simple rescaling argument gives

$$r_S(\epsilon; \lambda, \mu) = \epsilon^2 r_S(\lambda/\epsilon, \mu/\epsilon).$$

Therefore, we have

$$r_S(\epsilon; \lambda, \mu) \ge \frac{1}{2}\min(\mu^2, \epsilon^2); \tag{4.4}$$

in addition, the choice of $\lambda_{\epsilon,\delta} = \epsilon\sqrt{2\log\delta^{-1}}$ as the value of the threshold yields the upper bound

$$r_S(\epsilon; \lambda_{\epsilon,\delta}, \mu) \le (1 + 2\log\delta^{-1})\,(\epsilon^2\delta + \mu^2 \wedge \epsilon^2).$$

Remark. A similar inequality exists for the hard thresholding rule as in this case one has

$$r_H(\lambda, \mu) \geq \xi(\lambda/\epsilon)\, \min(\mu^2, \epsilon^2),$$

where ξ is some function bounded away from zero, $0 < \xi < 1$, which tends to 1 when its argument tends to ∞ (Donoho, 1993).

Suppose now that $\mathcal{I}'$ is a finite subset of $\mathcal{I}$ and let $\hat{\theta}_{\mathcal{I}'}$ be the estimator defined by

$$\widehat{\theta}_i \;=\; \begin{cases} \eta_S(y_i, \lambda\sigma_i) & i \in \mathcal{I}' \\ 0 & i \in \mathcal{I} \setminus \mathcal{I}' \end{cases}, \tag{4.5}$$

where $\lambda = \epsilon\sqrt{2\log(\#\mathcal{I}')}$. Then, of course, for $i \in \mathcal{I}'$, we have

$$\frac{1}{2}\min(\epsilon^2\sigma_i^2, \theta_i^2) \leq E(\widehat{\theta}_i - \theta_i)^2 \leq (1 + 2\log(\#\mathcal{I}'))(\{\#\mathcal{I}'\}^{-1}\epsilon^2\sigma_i^2 + \theta_i^2 \wedge \epsilon^2\sigma_i^2).$$

Hence, we have established the following result:

Lemma 4.3 *Let $\widehat{\theta}$ be the threshold estimator (4.5). Then,*

$$\frac{1}{2}\sum_{i\in\mathcal{I}'}\min(\epsilon^2\sigma_i^2, \theta_i^2) + \sum_{i\in\mathcal{I}\setminus\mathcal{I}'}\theta_i^2 \leq E\|\hat{\theta} - \theta\|_{\ell_2}^2$$
$$\leq (1 + 2\log(\#\mathcal{I}'))(\epsilon^2\bar{\sigma}^2 + \sum_{i\in cI'}\theta_i^2 \wedge \epsilon^2\sigma_i^2) + \sum_{i\in\mathcal{I}\setminus\mathcal{I}'}\theta_i^2, \tag{4.6}$$

where $\bar{\sigma}^2$ is simply a shorthand for $\{\#\mathcal{I}'\}^{-1}\sum_{i\in\mathcal{I}'}\sigma_i^2$.

The right-hand side of the above inequality is often referred to as the oracle inequality (Donoho and Johnstone, 1994). Similar bounds exist for hard thresholding rules as well.

It easily follows from the previous lemma that no thresholding rule exists with better lower bounds than

$$E\|\hat{\theta} - \theta\|_{\ell_2}^2 \geq \frac{1}{2}\sum_i \epsilon^2\sigma_i^2 \wedge \theta_i^2. \tag{4.7}$$

Hence, in the context of the sequence model, the sparsity of the coefficient sequences (3.4) (ridgelets, wavelets, etc.) gives lower estimation bounds of thresholding rules.

5 Linear singularities: The ridgelet miracle

5.1 Linear Singularities

Consider the mutilated Gaussian defined, as follows:

$$f(x) = 1_{\{u^T x \geq b\}} e^{-|x|^2/2}. \tag{5.1}$$

This function is discontinuous along the hyperplane $u^T x = b$ and smooth away from this hyperplane. In some sense, this is a very simple object. We wish to recover this object from noisy data (3.1) and will use the integrated mean-squared error as a measure of performance (3.2). (We acknowledge that the mutilated gaussian is not supported in the unit cube and, therefore, does not fit into the statistical paradigm that we set up. We chose the mutilated gaussian for its evocative power, rather than anything else. The conscious reader may substitute the gaussian $e^{-|x|^2/2}$ with a nice C^∞ function g supported in the unit cube in definition (5.1).)

We are going to compare the perfomance of our ridgelet shrinkage estimator (3.7) to that of kernel smoothers (Stone, 1977) or wavelet-based estimators as proposed by Donoho and Johnstone. Introducing a bit of terminology, $\widehat{f}_{KS}$ will denote an estimator obtained by kernel smoothing; similarly, $\widehat{f}_{WT}$ will denote a wavelet shrinkage estimator.

So, suppose that one uses a kernel smoother to recover f, then it can be shown that its integrated mean-squared error is bounded below by

$$MSE(\widehat{f}_{KS}, f) \geq C\,(\epsilon^2)^{1/(d+1)}. \tag{5.2}$$

It is interesting to note that the above inequality holds for any choice of bandwidth: that is, even if one had available an oracle that would specify the optimal bandwidth, one would not be able to obtain better bounds than (5.2). The optimal choice of the bandwidth comes from the classical bias/variance trade-off: the smaller the bandwidth, the smaller the bias around the edge but the greater the variance of the smoother; vice versa, the greater the bandwidth, the greater the bias (around the edge). The kernel smoother either smooths out the edge or undersmooths the flat part of the estimand. This undesirable feature is shared by all linear estimators as in fact, the optimized kernel smoother is as good as a linear estimator can be (we will make this claim more precise in the next section). The poor performance has a simple interpretation: we quote from Donoho and Johnstone (1998) "linear estimators are based in some sense on the idea of spatial homogeneity of the estimand." Here, our example is not spatially homogeneous – having a sharp discontinuity – and therefore not suited for linear procedures.

What about non-linear procedures? Following Donoho and Johnstone, let us look at a wavelet thresholding estimator (soft or hard does not matter). We argue that the performance of such an estimator (for any choice of wavelet basis) satisfies

$$MSE(\widehat{f}_{WT}, f) \geq C\,(\epsilon^2)^{1/d}. \tag{5.3}$$

One may think about the wavelet-thresholding estimator as a local smoother where one would be able to pick the size of the bandwidth adaptively, depending on the spatial inhomogeneity of the data (Donoho and Johnstone, 1994). (Here, the bandwidth one would certainly select a smaller bandwidth in a neighborhood of the discontinuity). The result is striking: such a non-linear procedure offers *very little improvement* over linear ones.

In dimension one, wavelets deal remarkably well with spatial inhomogeneities: that is, estimands that might be discontinuous, spiky, etc. This nice feature is certainly one of the reasons why they generated and continue to generate so much enthusiasm. In higher dimensions, however, there are various kinds of spatial inhomogeneities and our example is certainly an important one. It shows that, in some sense, the 'wavelet miracle' that operates in dimension one does not extend to higher dimensions. Wavelets cannot deal efficiently with objects that exhibit the kind of inhomogeneity we have just described. Already in dimension 2, this simple example enlightens the difficulties of wavelet methods in dealing with edges in images. We are allowed to talk about the 'poor performance' of linear or wavelet procedures on this type of object because of the existence of others with much better estimation properties, as we are about to see.

Now let us consider a simple ridgelet thresholding estimate $\widehat{f}_{RT}$ as in (3.7) of the same object. Then,

$$MSE(\widehat{f}_{RT}, f) = O((\epsilon^2)^s), \qquad \forall\, s < 1.$$

Unlike wavelets, ridgelets adapt very well to linear inhomogeneities. The reason for this remarkable fact is that the singularity causes highly concentrated or localized effect to the ridgelet representation, giving only a few a significant coefficients to estimate. Ridgelets are optimal to recover structures organized along hyperplanes.

Rather than averaging data over isotropic neighborhoods like balls (kernel, wavelet methods), ridgelet estimates are constructed by averaging the data over strips. For objects like (5.1), it seems to be a clear advantage if the strip may be positioned along the edge.

5.2 Adaptivity

Let $L := \{x,\, u^T x - b = 0\}$ be an arbitrary hyperplane and consider a function f such that

$$\|f\|_{W_2^s(\mathbb{R}^d \setminus L)} \le C :$$

that is, f has some kind of regularity away from L but may be discontinuous at L. We recall that W_2^s is the Sobolev space of square integrable functions whose s-th derivative is also square integrable. The norm is given by $\|g\|_{W_2^s}^2 = \|g\|_2^2 + \|D^s g\|_2^2$. (When s is not an integer, the norm is given via the Fourier transform $\hat{g}$, $\|g\|_{W_2^s}^2 = \int_{\mathbb{R}^d} (1 + |\xi|^{2s}) |\hat{g}(\xi)|^2 \, d\xi$.)

We can then consider the collection of such templates: i.e., let $\mathcal{F}(C)$ be the class defined by

$$\mathcal{F}(C) = \{f,\, \|f\|_{W_2^s(\mathbb{R}^d \setminus L)} \le C, \text{ for some hyperplane } L, \text{ and } \mathrm{supp} f \subset [0,1]^d\}. \tag{5.4}$$

It is important to emphasize that the singular hyperplane is not fixed; two elements from $\mathcal{F}(C)$ may be singular along two different hyperplanes.

We now give a lower bound on the estimation error of linear procedures.

Theorem 5.1 *Let $\mathcal{R}_L(\epsilon, \mathcal{F})$ be the minimax rate where the infimum (3.3) is restricted over linear procedures. Then, there exists a constant C such that*

$$\mathcal{R}_L(\epsilon, \mathcal{F}) \geq C\,(\epsilon^2)^{1/(d+1)}. \tag{5.5}$$

This fully justifies our claim (5.2).

Remark. Linear estimation of discontinuous functions has been studied by Korostelev and Tsybakov (1993)[Page 178] although their estimation problem is different than (5.4). They wish to recover elements of the form

$$f(x_1, \cdots, x_d) = f_0(x_1, \cdots, x_d) + f_1(x_1, \cdots, x_d)\, 1_{\{x_d \geq \varphi(x_1, \cdots, x_{d-1})\}},$$

where φ is a smooth function and where we may assume – as we do – that the pieces f_i's $i \in \{0, 1\}$ belong to some Sobolev ball. This problem is more general than ours since our assumption requires φ to be linear. However, translated to our framework, their lower bound is of order $(\epsilon^2)^{1/2}$ when, say, the singularity φ is C^∞ and the f_i's are smooth enough, which is not the correct order (not sharp), as suggested by Theorem 5.1. Our method is different than theirs as ridgelets play a central role in the determination of (5.5).

Proof of Theorem. The proof is in two steps. We first argue that the minimax linear rate over the class $\mathcal{F}$ is the same as the minimax linear rate over the convex hull of $\mathcal{F}$; then, we give a lower bound on the linear minimax rate of the latter convex hull.

Lemma 5.2 *We have*

$$\mathcal{R}_L(\epsilon, \mathcal{F}) = \mathcal{R}_L(\epsilon, Hull(\mathcal{F})). \tag{5.6}$$

Proof of Lemma. Let T be a linear procedure yielding estimators of the form $\hat{f} = TY$. The classical bias-variance decomposition states that

$$MSE(f, \hat{f}) = \|f - E\hat{f}\|_2^2 + \mathrm{Var}\hat{f},$$

and since the estimator is linear, the risk can be rewritten as

$$MSE(f, \hat{f}) = \|(I - T)f\|_2^2 + \epsilon^2 \|T\|_{HS}^2,$$

where $\|T\|_{HS}$ is the Hilbert Schmidt norm of the operator T ($\|T\|_{HS}^2 = \sum_n |Te_n|^2$ with (e_n) any orthobasis of $L_2[0,1]^d$). Now, let g be in the convex hull of $\mathcal{F}$, i.e. g is a finite sum of the form:

$$g = \sum_i a_i f_i, \quad f_i \in \mathcal{F}, \quad \sum_i |a_i| \leq 1,$$

The bias of our linear estimate may be bounded as follows:

$$\|(I-T)f\|_2 = \|(I-T)(\sum_i a_i f_i)\|_2 \le \sum_i |a_i| \|(I-T)f_i\|_2.$$

The variance term is unchanged and

$$\begin{aligned} MSE(g,\hat{g}) &= \|(I-T)g\|_2^2 + \epsilon^2 \|T\|_{HS}^2 \\ &\le (\sum_i |a_i|)^2 \sup_i \|(I-T)f_i\|_2^2 + \epsilon^2 \|T\|_{HS}^2 \\ &\le \sup_{f \in \mathcal{F}} MSE(f,\hat{f}). \end{aligned}$$

This last inequality implies the desired result. ∎

We now give a lower bound on the linear minimax rate over the convex hull, which, of course, is the same as the one over the L_2-closure of the convex hull $\overline{Hull(\mathcal{F})}$. The basic idea is to observe that rescaled ridgelets of the form $\psi(2^j(u_{j,\ell}^T x - k))$ are in the closure of the convex hull of $\mathcal{F}$. Hence, for each scale $j \ge 0$, we have of the order of 2^{jd} nearly orthogonal elements with L_2 norms roughly equal to $2^{-j/2}$. There is a natural lower bound on the linear estimation of orthogonal functions; when j is chosen appropriately, this lower bound gives (5.5). A rigorous argument involves a delicate construction whose proof may be found in the appendix.

Lemma 5.3 *For any $\delta > 0$, there exist $m(\delta)$ orthogonal elements $\{g_\ell\} \in \overline{Hull(\mathcal{F})}$ satisfying the following properties:*

1. *For any $1 \le \ell \le m(\delta)$, $\|g_\ell\|_2 = \delta$, and*
2. *$m(\delta) \ge \delta^{-2d}$.*

We use this lemma to finish the proof of the theorem. To ease notation, we will set $\mathcal{V}_\delta = (g_\ell)_{1 \le \ell \le m(\delta)}$. We then have

$$\mathcal{R}_L(\epsilon, Hull(\mathcal{F})) = \mathcal{R}_L(\epsilon, \overline{Hull(\mathcal{F})}) \ge \mathcal{R}_L(\epsilon, \mathcal{V}_\delta)$$

Now, the linear minimax rate is given by

$$\mathcal{R}_L(\epsilon, \mathcal{V}_\delta) = \inf_T \sup_\ell \|(I-T)g_\ell\|_2^2 + \epsilon^2 \|T\|_{HS}^2.$$

There are two cases: either $\|I-T\|_2^2 \ge 1/2$ or $\|(I-T)\|_2^2 < 1/2$. In the first case, we bound the risk of the linear estimator T by the bias term, namely, $\delta^2/2$; in the second, we bound the risk by the variance, $\epsilon^2 \|T\|_{HS}^2$. In the former case, we will use the upper bound on the bias to get a lower bound on the variance term, i.e. $\|T\|_{HS}^2$. Indeed, it is not hard to show that

$$\|I-T\|_2^2 < 1/2 \Rightarrow \|T\|_{HS}^2 \ge m(\delta)/2,$$

where $m(\delta)$ is the cardinality of $\mathcal{V}_\delta$. In any event, we have that for any δ,

$$\mathcal{R}_L(\epsilon, \mathcal{V}_\delta) \geq \frac{1}{2} \min(\delta^2, \epsilon^2 m(\delta)).$$

We finish the proof by letting $\delta_\epsilon = \epsilon^{1/(d+1)}$. Using the fact that $m(\delta)$ is bounded below by δ^{-2d} gives

$$\mathcal{R}_L(\epsilon, \mathcal{V}_{\delta_\epsilon}) \geq C\,(\epsilon^2)^{1/(d+1)}.$$

We trivially conclude that

$$\mathcal{R}_L(\epsilon, Hull(\mathcal{F})) \geq \mathcal{R}_L(\epsilon, \mathcal{V}_{\delta_\epsilon}) \geq C\,(\epsilon^2)^{1/(d+1)}.$$

The proof of the theorem is complete. ∎

In stark contrast with linear procedures, shrinkage ridgelet estimates attain estimation bounds as if there were no discontinuity.

In order to give a precise statement, we need to polish the form of our ridgelet shrinkage estimator (3.7). We will work with a nice ridgelet frame (2.10) $(\psi_i)_{i\in\mathcal{I}}$ such that ψ has enough vanishing moments and regularity. To simplify the analysis we take φ and ψ to be compactly supported. Hence, at a given scale j, the number of ridgelets that are nonzero on $[0,1]^d$ is bounded by

$$\#\{\psi_i,\, j(i) = j\} \leq C\, 2^{jd},$$

for some fixed constant C.

To estimate the true ridgelet coefficients θ from our noisy data y (3.4), we consider the diagonal projection as defined in section 4. Set

$$\mathcal{I}' = \{i,\, j(i) \leq J_\epsilon\}$$

and define

$$\widehat{\theta}_i \;=\; \begin{cases} \eta_S(y_i, \lambda\sigma_i) & i \in \mathcal{I}' \\ 0 & i \in \mathcal{I} \setminus \mathcal{I}' \end{cases} \tag{5.7}$$

with $\lambda = \epsilon\sqrt{2\log(\#\mathcal{I}')}$ and where we recall that the σ_i's are the L_2 norms of the ridgelets ψ_i. Thus, the estimator (5.7) sets to zero all the coefficients exceeding a given scale and thresholds the others.

Theorem 5.4 *Consider the ridgelet thresholding estimate $\hat{f}$ (3.7) (with (5.7) as the choice of scalar nonlinearities). Then,*

$$\sup_{\mathcal{F}} MSE(\hat{f}, f) \leq (1 + 2\log(\epsilon^{-1}))(\epsilon^2)^{\frac{2s}{2s+d}}.$$

Our estimator gives the optimal rate – up to a logarithmic factor – since there is a lower bound on the estimation of compactly supported functions with square integrable s-th derivatives. Indeed, if we let

$$\mathcal{W}(s,C) = \{f,\, \|f\|_{W_2^s} \leq C\,, \mathrm{supp} f \subset [0,1]^d\}$$

be this class, its minimax rate is bounded below as follows:

$$\inf_{\hat{f}} \sup_{f \in \mathcal{W}(s,C)} MSE(f,\hat{f}) \geq c\,(\epsilon^2)^{\frac{2s}{2s+d}}.$$

It is quite remarkable that our estimator achieves an error of estimation that is almost (up to the logarithmic factor) as good as the one that one could obtain if an oracle told us the exact location of the discontinuity.

The ridgelet shrinkage procedure is entirely data driven: we do not need to know whether or not there is a singularity or if there is one, where it is. In addition, we do not need to know the degree of smoothness s of the regression surface away from the singularity. In this sense, the ridgelet estimator is spatially adaptive and, moreover, adapts to the unknown degree of smoothness.

Proof. Following the argument developed in the previous section, we simply need to study the sparsity of the ridgelet coefficient sequence.

We apply Lemma 4.3 and the upper bound will result from the following two facts that are proven in Candes (1999b): first,

$$\sum_i \min(\theta_i^2, \epsilon^2) \leq C\,(\epsilon^2)^{\frac{2s}{2s+d}}; \tag{5.8}$$

and, second,

$$\sum_{j(i)>J_\epsilon} \theta_i^2 \leq C\, \max(2^{-2J_\epsilon s}, 2^{-J_\epsilon}). \tag{5.9}$$

Since the ridgelets are uniformly bounded in $L_2([0,1]^d)$, we may as well take the upper bound to be 1 so that $\sigma_i \leq 1$ for any $i \in \mathcal{I}$. Finally, an application of Lemma 4.3, together with (5.9), gives

$$E\|\widehat{\theta} - \theta\|^2_{\ell_2(\mathcal{I})} \leq C\, \left[1 + 2\log(2^{J_\epsilon d})\right] \left[\epsilon^2 + (\epsilon^2)^{2s/(2s+d)}\right] + C\, 2^{-2J_\epsilon \min(1/2,s)}.$$

Suppose that $J_\epsilon = \lfloor 2\log(\epsilon^{-1})\rfloor$. Then, the approximation term $2^{-2J_\epsilon \min(1/2,s)} \leq (\epsilon^2)^{\min(2s,1)}$ is negligible when compared to the leading term $(\epsilon^2)^{2s/(2s+d)}$ of the mean-squared error. In short, we have

$$E\|\widehat{\theta} - \theta\|^2_{\ell_2(\mathcal{I})} \leq C\, \log(\epsilon^{-1})(\epsilon^2)^{2s/(2s+d)}.$$

Finally, inequality (3.6) linking $E\|\widehat{\theta} - \theta\|^2_{\ell_2(\mathcal{I})}$ and $E\|\widehat{f} - f\|_2^2$ finishes the proof of Theorem 5.4. ∎

There are obvious extensions of this result. For instance, one could take finite superposition of elements from our class of templates $\mathcal{F}(C)$ (5.4). Let the regression surface f be of the form

$$f = \sum_{i=1}^{m} a_i f_i,$$

where m is arbitrary, equal to 10 or 20, say, meaning that our regression surface is smooth away from 10 or 20 hyperplanes. Now, suppose that we observe f in the presence of noise and apply the ridgelet shrinkage estimator: the asymptotics is unchanged; namely,

$$E\|\hat{f} - f\|_2^2 \le C \, \log(\epsilon^{-1})) \, (\epsilon^2)^{\frac{2s}{2s+d}}.$$

Again, we do not need to know how many of these hyperplanes there are nor where they are.

Going towards more generality, there is an infinite dimensional version of these types of results. We can construct a class of functions whose typical elements are of the form $f(x) = 1_{\{u^T x - b \ge 0\}} \, g(x)$ with $g \in W_2^s$.

Definition 5.5 *Let $\mathcal{S}_H$ be the class of functions defined by*

$$\mathcal{S}_H = \{f = \sum a_i f_i \;\; \sum |a_i| \le 1, \, \|f_i\|_{W_2^{\frac{d+1}{2}}(\mathbb{R}^2 \setminus L_i)} \le C\}. \tag{5.10}$$

The model is meant to represent objects composed of singularities across hyperplanes: typical elements of our model are smooth away and discontinuous across these same hyperplanes. There may be an arbitrary number of singularities which may be located in all orientations and positions.

Theorem 5.6 *The ridgelet thresholding estimate $\hat{f}$ (3.7) is asymptotically nearly minimax over our model $\mathcal{S}_H$. We have*

$$\sup_{\mathcal{S}_H} MSE(\hat{f}, f) \le (1 + 2\log(\epsilon^{-1}))(\epsilon^2)^{\frac{d+1}{2d+1}}. \tag{5.11}$$

Our model is made up of functions that may be discontinuous along an arbitrary and possibly infinite number of hyperplanes, but the rate estimate still behaves as if they were $(d+1)/2$ times differentiable (in an L_2 sense).

Proof of the theorem. We first show that the sum of the absolute values of the ridgelet coefficients θ_i of any $f \in \mathcal{S}_H$ is bounded as follows:

$$\sup_j 2^{j/2} \sum_{i:j(i)=j} |\theta_i| \le C. \tag{5.12}$$

By convexity, it suffices to show (5.12) for f of the form $f = f_0 + 1_{\{u^T x - b \ge 0\}} f_1$, a fact estabished in Candes (1999b). In turn, this property implies that the ridgelet sequence is in $w\ell_p$ for $1/p =$

$1+1/(2d)$, or equivalently that $\sum_i \min(\epsilon^2, \theta_i^2) \le C\,(\epsilon^2)^{\frac{d+1}{2d+1}}$. The rest of the argument is now similar to that of Theorem 5.4.

The near-minimaxity follows from the mere observation that the class $\mathcal{S}_H$ contains $W_2^{(d+1)/2}$ whose minimax estimation rate is bounded below by $c\,\epsilon^{\frac{d+1}{2d+1}}$. This establishes the theorem. ∎

6 A Minimax Theorem

In the previous section, we argued that ridgelets – and, in a broader sense, ridge functions – were optimal for estimating functions with some special kinds of inhomogeneities, Theorem 5.4 and Theorem 5.6. This section shows that these results are part of a broader picture. The section is organized as follows: we first introduce new functional classes based on a new notion of smoothness; we then show that a simple ridgelet thresholding estimator is asymptotically nearly minimax for estimating objects from these classes.

6.1 New notion of smoothness

Candes (1998) introduces a family of spaces defined via the properties of the continuous ridgelet transform: we will say that a function f belongs to the homogeneous ridge space $\dot{R}^s_{p,q}$ for $p, q \ge 1$ if f is integrable and

$$\|f\|_{\dot{R}^s_{p,q}} \equiv \left(\int \left[\int |\mathcal{R}_f(a,u,b)|^p \, dbdu \right]^{q/p} \frac{da}{a^{q(s+d/2)+1}} \right)^{1/q} < \infty, \tag{6.1}$$

where $\mathcal{R}_f(a,u,b)$ is the ridgelet coefficient of f (2.4) (we recall that du is the uniform measure on the sphere).

In nonparametric estimation, there has recently been a great deal of interest in studying estimation procedures over Besov balls, see Härdle, Kerkyacharian, Picard, and Tsybakov (1998) and references therein. Besov norms measure the smoothness of an estimand f. Roughly, if s is an integer, $\|f\|_{B^s_{p,q}} \le C$ means that f is in some sense s times differentiable. (When s is not an integer, it says that the $[s]^{th}$ derivative of f has some kind of continuity properties.)

We recall the definition of the Radon transform Rf of an integrable function f (see Deans, 1983 for details)

$$Rf(u,t) = \int_{u^T x = t} f(x)\, dx.$$

The quantity (6.1) has a natural interpretation in terms of the smoothness of the Radon transform. Indeed, for $p = q$, we have the following equivalence:

$$\|f\|^p_{\dot{R}^s_{p,p}} \asymp \operatorname*{Ave}_u \|Rf(u,\cdot)\|^p_{\dot{B}^{s+(d-1)/2}_{p,p}}, \tag{6.2}$$

where $\dot{B}^{s+(d-1)/2}_{p,p}$ stands for the usual 1-dimensional homogeneous Besov norm. Instead of – classically – requiring smoothness on the estimand, we require smoothness on the Radon transform. Roughly speaking, s is associated with the number of derivatives of the Radon transform and, hence, is interpreted as a degree of smoothness and p, q are parameters that serve to measure smoothness. We would like to emphasize that this is very different from the classical pointwise notion of smoothness as we are about to see.

For instance, suppose one is given the function

$$f(x) = H(x_1)(2\pi)^{-d/2} e^{-|x|^2/2}. \tag{6.3}$$

From a classical viewpoint, this is not a smooth object: the first derivative is a singular measure. Let $\cos\theta$ be the first component of the unit vector u in the canonical basis, then the Radon transform of f is given by

$$Rf(t,u) = e^{-t^2/2}\Phi(t\cos\theta/|\sin\theta|),$$

where Φ is the cumulative distribution function of a standard normal variable $\Phi(t) = \int_{-\infty}^{t}(2\pi)^{-1/2}e^{-y^2/2}\,dy$. Except for values of (t,θ) in the neighborhood of the singular point $(0,0)$, the Radon transform of f is extremely smooth. In fact, according to our definition it has about $(d+1)/2$ derivatives as one can show that $f \in R^s_{1,1}$ for every choice of $s < (d+1)/2$ (Candes, 1998).

In fact, typical elements of $R^s_{p,q}$ (at least when $p < 2$) look like our mutilated gaussian (6.3), in that they exhibit the same kind of spatial inhomogeneities. For instance, the class S_H of mutilated functions that we defined in Section 5 almost corresponds to one of these spaces. Indeed, we have

$$R^{(d+1)/2}_{1,1}(C_1) \subset \mathcal{S}_{\mathcal{H}} \subset R^{(d+1)/2}_{1,\infty}(C_2), \tag{6.4}$$

which means that membership to S_H is roughly equivalent to membership to $R^{(d+1)/2}_{1,q}$ $(1 \le q \le \infty)$. Therefore, we should really think about these spaces as describing the kind of spatial inhomogeneities we introduced in the previous section.

Kernel smoothing techniques are well adapted to some functional classes and wavelet methods to others; likewise, we believe that ridge function estimation (and approximation) is especially well suited for objects having the smoothness displayed by (6.1) or (6.2). The remainder of this section is devoted to a precise formulation of this heuristics.

6.2 A minimax theorem

Let $R^s_{p,q}(C)$ be the ball of radius C: that is, the collection of elements supported in the unit cube $[0,1]^d$ whose norm (6.1) is bounded by a fixed constant C. We have the following result:

Theorem 6.1 *Consider the class $R^s_{p,q}(C)$ and assume $s > d(1/p - 1/2)_+$, a condition that guarantees that the class can be consistently estimated with an L_2 loss.*

- *(i) There is a lower bound on the minimax rate,*

$$\mathcal{R}_\epsilon(R^s_{p,q}(C)) \geq K(\epsilon^2)^{\frac{2s}{2s+d}}, \tag{6.5}$$

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

- *(ii) A simple thresholding estimator (3.7) achieves the optimal rate within a log-like factor; i.e.,*

$$\sup_{f \in R^s_{p,q}(C)} E\|\widehat{f} - f\|_2^2 \leq K' \log(\epsilon^{-1})(\epsilon^2)^{\frac{2s}{2s+d}}, \tag{6.6}$$

  where again K' might depend on s, p, q.

It is possible to get sharper lower bounds and show that a logarithmic factor is necessary for a certain range of the indices. However, we do not attempt to go that far in this paper.

6.3 Lower Bounds

The proof of the lower bound is classical and relies on a well-known result, namely, Assouad's lemma (Korostelev and Tsybakov, 1993)[Page 69]: that is, we specify a subproblem and use Assouad's lemma to calculate its difficulty. The idea is as follows: suppose that one can find m orthogonal functions $(g_\ell)_{1 \leq \ell \leq m}$ with $\|g_\ell\|_{L_2} = \delta$ such that

$$\mathcal{H}(\delta, \{g_\ell\}) \equiv \{f = \sum_{\ell=1}^m \xi_\ell g_l,\, \xi_\ell \in \{-1, 1\}\} \subset R^s_{p,q}(C);$$

that is, by taking a functional analysis viewpoint, one can find a cube of sidelength δ and dimension m (2^m vertices) embedded in the functional ball $R^s_{p,q}(C)$. Our subproblem is the same estimation problem but restricted to the cube $\mathcal{H}$ (the functions to be recovered are the vertices of $\mathcal{H}$). Let us consider the minimax risk $\mathcal{R}_\epsilon(\mathcal{H})$ of this specific subproblem which urns out to be easily calculated as it is a direct consequence of Assouad's lemma.

Lemma 6.2 *Let $\mathcal{H}(\epsilon, \{g_\ell\})$ be the orthogonal hypercube of dimension m and sidelength ϵ defined as above ($\delta = \epsilon$). Then*

$$\mathcal{R}_\epsilon(\mathcal{H}) \geq \Phi(-1/2)/4\, m\, \epsilon^2, \tag{6.7}$$

where Φ is the cumulative distribution function of the standard normal distribution.

As emphasized, the lemma is a variation on Assouad's lemma; moreover we would like to point out that our formulation is not new as it may be found in Donoho and Johnstone (1995).

Proof of Lemma. To find the minimax risk of (3.1) when f is assumed to be of the form $f = \sum_{\ell=1}^{m} \xi_\ell g_l$, with $\xi_\ell \in \{-1, 1\}$, we first note that we may only consider estimators that lie in the span of the g_ℓ's; this fact follows from the simple following observation: by letting P be the orthogonal projector onto that span, for any estimator we have

$$\|P\widehat{f} - f\|_2^2 \leq \|\widehat{f} - f\|_2^2.$$

Thus, the problem reduces to estimating the ξ_ℓ's from the noisy observations $y_\ell = \langle Y, g_\ell \rangle$, where

$$y_\ell = \epsilon^2 \xi_\ell + \epsilon^2 z_\ell,$$

or, equivalently, from the rescaled noisy observations $\tilde{y}$,

$$\tilde{y_\ell} = y_\ell/\epsilon^2 = \xi_\ell + z_\ell, \tag{6.8}$$

and where, of course, $z_\ell \overset{i.i.d.}{\sim} N(0,1)$. Observe now that for an estimator of the form $\widehat{f} = \sum_\ell \widehat{\xi}_\ell g_\ell$, we have $\|\widehat{f} - f\|_2^2 = \epsilon^2 \sum_\ell (\widehat{\xi}_\ell - \xi_\ell)^2$. Then, a rescaling argument gives that the minimax mean-squared-error $\mathcal{R}_\epsilon(\mathcal{H})$ equals ϵ^2 times the minimax mean-squared error of the problem (6.8); that is

$$\mathcal{R}_\epsilon(\mathcal{H}) = \inf_{\widehat{f}} \sup_{\mathcal{H}} E\|\widehat{f} - f\|_{L_2[0,1]^d}^2 = \epsilon^2 \inf_{\xi(\tilde{y})} \sup_{\xi \in \{-1,1\}^m} E \sum_\ell (\widehat{\xi}_\ell - \xi_\ell)^2.$$

The latter problem (6.8) is now classical and a lower bound for its minimax mean-squared-error is $\Phi(-1/2)\, m$. It is interesting to note that (6.8) has a strong flavor of an hypothesis testing problem as one tries to distinguish which of the 2^m hypotheses $\xi \in \{-1, 1\}^m$ is the correct one. ∎

The previous lemma will give the lower bound of estimation if one can find a sequence of 'fat' hypercubes $\mathcal{H}(\epsilon, m(\epsilon))$ yielding a sharp asymptotic lower bound. The lower bound (6.5) follows from the technical lemma:

Lemma 6.3 *For any $\delta > 0$, there exists a hypercube $\mathcal{H}(\delta, \{g_\ell\}) \subset R_{p,q}^s(C)$ of sidelength δ and dimension $m(\delta) \geq K\delta^{-1/(s/d+1/2)}$.*

The proof of this technical lemma is given in the appendix. Again, a slight perturbation of properly rescaled ridgelets builds up the vertices of this hypercube.

We now finish the proof of the first part of Theorem 6.1:

Corollary 6.4 *We have a lower bound on the minimax risk*

$$\mathcal{R}_\epsilon(R_{p,q}^s(C)) \geq c\, (\epsilon^2)^{\frac{2s}{2s+d}}. \tag{6.9}$$

Proof of Corollary. We clearly have

$$\mathcal{R}_\epsilon(R^s_{p,q}(C)) \geq \mathcal{R}_\epsilon(\mathcal{H}) \geq \Phi(-1/2)/4\, m(\epsilon)\, \epsilon^2,$$

and the lower bound follows since $m(\epsilon)$ might be chosen to be greater than $K\epsilon^{-1/(s/d+1/2)}$. ∎

6.4 Upper Bounds

The proof of the upper bound closely follows the concepts presented in section 4. For convenience, let us take exactly the same estimator as the one introduced at the beginning of section 5.2; i.e.,

$$\widehat{\theta}_i \;=\; \begin{cases} \eta_S(y_i, \lambda\sigma_i) & j(i) \leq J_\epsilon \\ 0 & j(i) > J_\epsilon \end{cases},$$

(see section 5.2 for the value of the parameter λ).

We suppose that the parameters s, p, q are fixed with $s > d(1/p - 1/2)_+$ and we consider the image of $R^s_{p,q}(C)$ through the analysis operator $f \mapsto (\theta_i(f))_{i\in\mathcal{I}}$, $\theta_i(f) = \langle f, \psi_i\rangle$: that is,

$$\Theta = \{\theta = (\theta_i(f))_{i\in\mathcal{I}},\ \|f\|_{R^s_{p,q}} \leq C\}.$$

The upper bound will result from the following fact that is proven in Candes (1999c): for any function $f \in \mathbb{R}^s_{p,q}(C)$, we have

$$\|\theta\|_{\mathbf{r}^s_{p,q}} := \sum_i \left(\sum_{j\geq 0} (2^{j\sigma} \sum_{j(i)=j} |\alpha_i|^p)^{q/p} \right)^{1/q} \leq C\, \|f\|_{R^s_{p,q}}. \tag{6.10}$$

Formally, $\|\theta\|_{\mathbf{r}^s_{p,q}}$ has the same structure than a discrete Besov norm, except that the sequence θ measures a radically different behavior.

Among other things, the finiteness of $\|\theta\|_{\mathbf{r}^s_{p,q}}$ for $\theta \in \Theta$ has two consequences: first for any $\epsilon > 0$, we have

$$\sum_i \min(\epsilon^2, \theta_i^2) \leq \epsilon^{\frac{2s}{2s+d}}; \tag{6.11}$$

and second,

$$\sum_{j(i)>J_\epsilon} \theta_i^2 \leq C\, 2^{-2J_\epsilon s'/d}, \tag{6.12}$$

with

$$s' = \begin{cases} s & p \geq 2 \\ s - d(1/p - 1/2) & p < 2 \end{cases}.$$

Since the ridgelets are uniformly bounded in $L_2([0,1]^d)$, the sparsity of the coefficient sequence gives

$$\sum_{j(i)\le J_\epsilon} \theta_i^2 \wedge \epsilon^2\sigma_i^2 \le C \sum_{j(i)\le J_\epsilon} \theta_i^2 \wedge \epsilon^2 \le C\,(\epsilon^2)^{2s/(2s+d)}.$$

(Compare with Lemma 2 in Donoho (1993).)

Finally, an application of Lemma 4.3 together with (6.12) gives

$$E\|\widehat{\theta}-\theta\|^2_{\ell_2(\mathcal{I})} \le C\,\left[1+2\log(2^{J_\epsilon d})\right]\left[\epsilon^2+(\epsilon^2)^{2s/(2s+d)}\right] + C\,2^{-2J_\epsilon s'}.$$

Suppose that $J_\epsilon = \lfloor 2\alpha\log(\epsilon^{-1})\rfloor$ with α chosen large enough so that $2\alpha s' > 2s/(2s+d)$. Then, the approximation term $2^{-2J_\epsilon s'} \le (\epsilon^2)^{2\alpha s'}$ is negligible when compared to the leading term $(\epsilon^2)^{2s/(2s+d)}$ of the mean-squared error. To summarize, we have

$$E\|\widehat{\theta}-\theta\|^2_{\ell_2(\mathcal{I})} \le C\,\log(\epsilon^{-1})(\epsilon^2)^{2s/(2s+d)},$$

and, finally, inequality (3.6) allows us to conclude that the worst case error of our simple thresholding estimator comes within a possible logarithmic factor of the minimax risk. The proof of Theorem 6.1 is complete. ∎

We would like to close this subsection by pointing out that a hard thresholding rule, similar in every aspect to the soft thresholding rule presented above but for the substitution of the nonlinearity η_{ST} with

$$\eta_{HT}(y,\lambda) = y\,1_{\{|y|\ge\lambda\}}, \tag{6.13}$$

would give exactly the same asymptotic performance.

6.5 Adapting to the unknown degree of smoothness

A remarkable feature of the ridgelet shrinkage estimator is its spatial adaptivity: the same estimator is simultaneously asymptotically nearly minimax over a wide range of smoothness classes $R^s_{p,q}$. In other words, no prior information on the parameters s, p, q is needed to obtain near-minimaxity; the estimator adapts to the unknown smoothness of the estimand.

A simple mathematical statement may clarify this point. Let $\nu = (s,p,q)$ denote the parameters describing the smoothness scale $R^s_{p,q}$, and $\mathcal{F}_\nu(C)$, the corresponding ball of radius C. We have just shown that there is an estimator such that

$$\sup_{f\in\mathcal{F}_\mathcal{V}(C)} E\|\hat{f}-f\|_2^2 \le K(\nu)\,C\,\log(\epsilon)\,(\epsilon^2)^{2s/(2s+d)}.$$

Suppose now that we are given a subset $\mathcal{V}_0$ of the parameter space satisfying $s - d(1/p - 1/2)_+ \geq s_0$ for any $\nu \in \mathcal{V}_0$ and some $s_0 > 0$. Then, there is a ridgelet thresholding estimator $\hat{f}$ with the property

$$\sup_{\mathcal{F}_{\mathcal{V}}(C)} E\|\hat{f} - f\|_2^2 \leq K_0 \, C \, \log(\epsilon) \, (\epsilon^2)^{2s/(2s+d)}, \quad \forall \nu \in \mathcal{V}_0, \tag{6.14}$$

for some constant K_0 depending only on $\mathcal{V}_0$.

7 Curved Singularities

As in Chapter 6, one may ask whether one can curve the singularity, still preserving the nice theoretical estimation bounds of ridgelet thresholding estimators. In statistics, projection pursuit regression and kernel regression are frequently used for estimating smooth multivariate functions from noisy observations. It is true that in some cases, projection-based approaches might be more accurate, as exemplified in the previous sections. In particular, there has been a large debate in the literature of statistics about their relative performances when the underlying estimand is radial, see Donoho and Johnstone (1989), for example. We will follow an approach similar to the one developed in section 5 by studying a simple example exhibiting a general phenomenon.

Let f be the radial function defined, as follows:

$$f(x) = 1_{\{|x| \leq 1/2\}} e^{-|x|^2/2}, \tag{7.1}$$

that is, a 'radially mutilated dome.' This surface is smooth away from the sphere $|x| = 1/2$, but singular across the latter sphere.

For kernel smoothing and wavelet wavelet thresholding procedures, the story is similar to the one presented in the previous section. That is, the risks scale in the same way as before; i.e.,

$$MSE(\widehat{f}_{KS}, f) \geq C \, (\epsilon^2)^{1/(d+1)}$$

for a linear smoother with any bandwidth, and

$$MSE(\widehat{f}_{WT}, f) \sim (\epsilon^2)^{1/d}$$

for any reasonable wavelet thresholding estimate.

Now, let us consider the risk of a ridgelet thresholding estimator. Using the results presented section 4, we argue that

$$MSE(\widehat{f}_{RT}, f) \sim (\epsilon^2)^{1/d}.$$

The reason is that the ridgelet transform of (7.1) is not sparse. Candes (1998)[Chapter 6] proves that

$$\sum_i \min(\theta_i^2, \epsilon^2) \geq C\,(\epsilon^2)^{1/d}, \tag{7.2}$$

which supports the claim as discussed in section 4.

This result is doubly surprising: first, it is spectacular that two distinct methods corresponding to radically different procedures give the same asymptotic estimation bounds. Of course, the duality existing between ridgelet and wavelet estimation is essentially the same as the one existing between projection pursuit regression and nonlinear kernel regression with an adaptive choice of bandwitdth: the non-linear ridgelet procedure estimates the regression surface by a superposition of ridge functions (chosen after averaging the noisy data over strips) while the wavelet estimator is based on a superposition of bumps (obtained after averaging the data over balls). And yet, both estimate the singular regression surface with the same degree of accuracy!

One might argue that the limit of performance is due not so much to the ridge function approach but to the specificity of the ridgelet shrinkage method. After all, other estimators with better estimation bounds may exist, even though this is unlikely. Indeed, to obtain good estimation bounds, finite linear combinations of ridge functions should provide a good model for objects like (7.1), meaning that one would need only a few number of ridge functions to approximate the true regression surface. The problem is that objects with curved structure like (7.1) are not well approximated by ridge functions. Preliminary results about this heuristics may be found in Candes (1998)[Chapter 7].

Second, this negative result clearly shows the limits of projection-based approaches. Superficially, it may be seen as a curse for it disproves a widespread and recurrent claim in the literature arguing that neural networks and related prediction methods are free from the curse of dimensionality. In a nutshell, the result says that unless the regression surface is $s \times d$ times differentiable, you cannot, in general, hope for a mean-squared error of order $(\epsilon^2)^{2s/(2s+1)}$.

8 Numerical experiments

A fast algorithm has been developed to code up the ridgelet transform: the details of the algorithm are not yet published. At the present stage, the algorithm works in the case of the dimension $d = 2$. The algorithm takes data on a cartesian grid and computes "pseudo-ridgelet" coefficients. Here, the discrete transform is not orthonormal but provides a frame of redundancy factor 2 and is numerically tight. Finally, this discrete transform has low complexity since it runs in $O(n^2 \log(n))$ flops for an $n \times n$ image. Of course, there is an associated inverse transform that reconstructs

an image from the data of its "pseudo-ridgelet" coefficients; its order of complexity is of the same order as the one of the forward transform. The major part of the work described in this paragraph has been done by David Donoho.

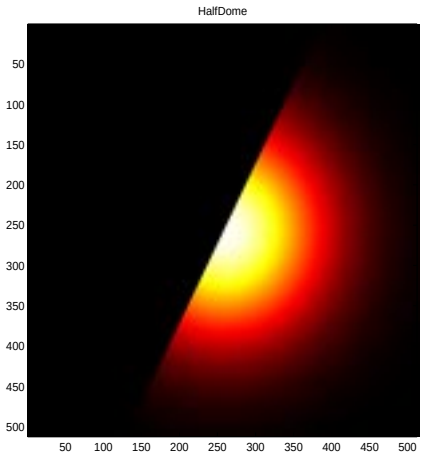
It follows that the local ridgelet transform can obviously be computed in $O(n^2 \log(n))$ flops and, hence, the method we have described in this paper appears to be very attractive from a practical point of view.

Let f be the half dome introduced earlier, namely,

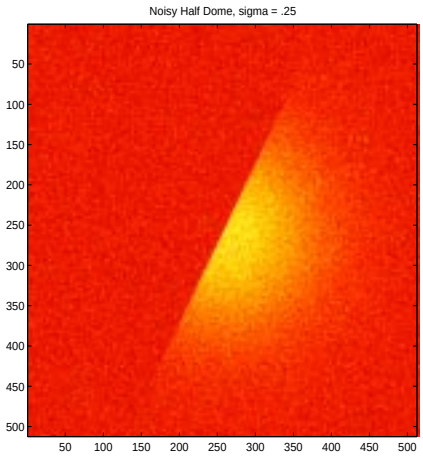
$$f(x, y) = 1_{\{x+2y>0\}} e^{-4(x^2+y^2)}.$$

The image of this function is stored as an array of 512 by 512 pixel values and is represented on Figure 2.

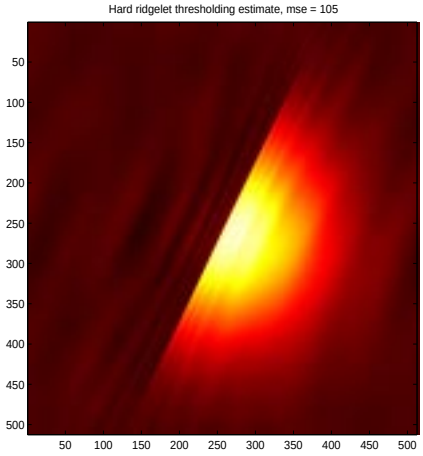
Figure 2 presents an application of ridgelets to the problem of recovering images from noisy data. The edge and the flat part of the image are well recovered from the data. There are not any visible artifacts near the edge.



[Original Image]



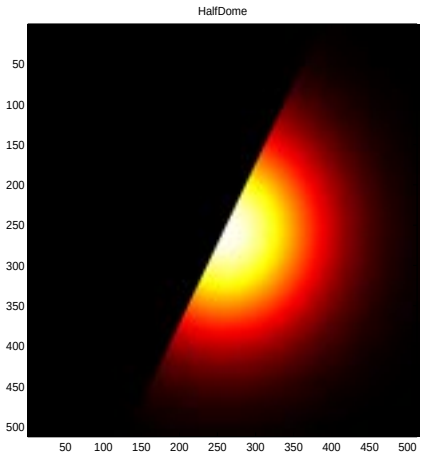
[Noisy Image]



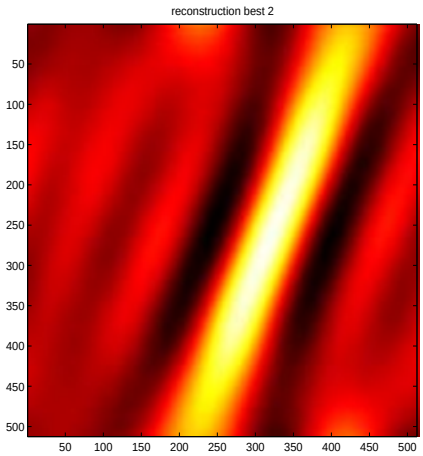
[Ridgelet Estimation]

Figure 2: The original image is presented together with its noisy version. The last figure represents our estimate obtained after thresholding the noisy ridgelet coefficients. Both the edge and the flat part of the image are well recovered.

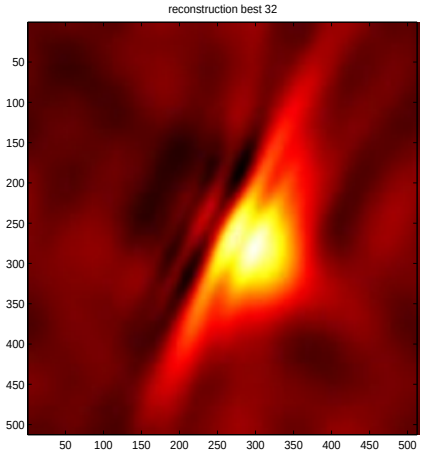
The next figures (Figures 3 and 4) display the successive approximations of the noiseless object f as they provide some insights about the nature of the method. The first ridgelet to enter into the approximation is the one with the largest coefficient; the second is the one with the second largest, and so on. The first selected ridgelets are nicely aligned with the edge.



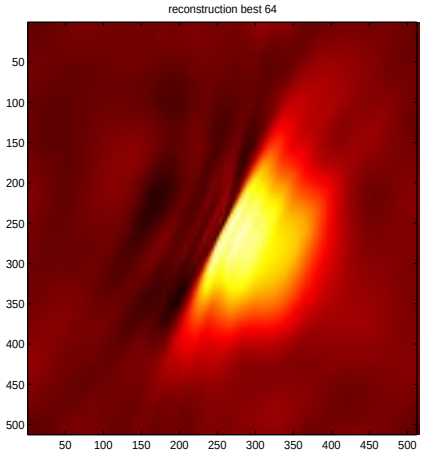
[Original Image]



[2 Coeff. out of 262144]

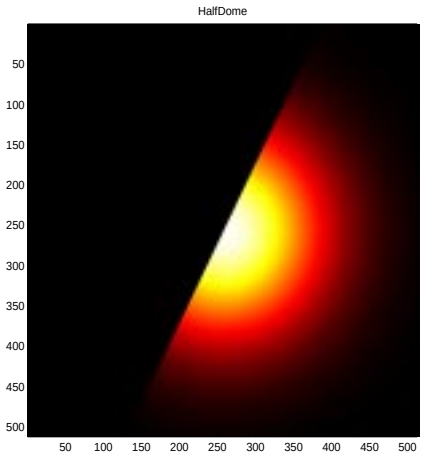


[32 Coeff. out of 262144]

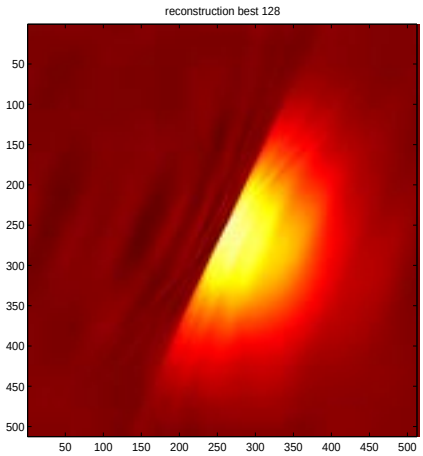


[64 Coeff. out of 262144]

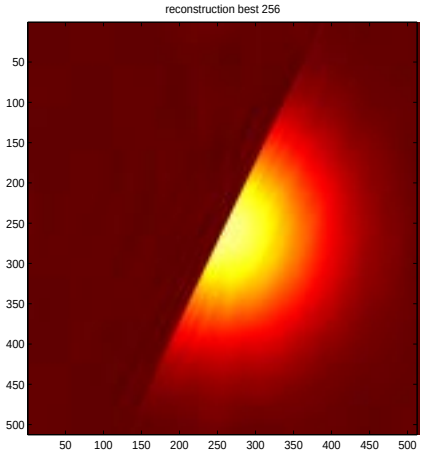
Figure 3: The original image is presented together with its approximations using successively 2, 32 and 64 coefficients. It is interesting to observe that the first ridgelets that are selected are aligned with the edge: they 'pick up' the edge.



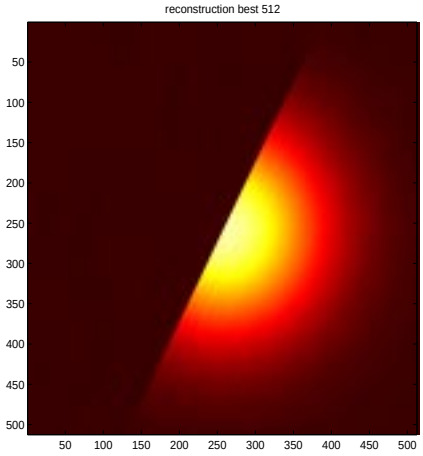
[Original Image]



[128 Coeff. out of 262144]



[256 Coeff. out of 262144]



[512 Coeff. out of 262144]

Figure 4: The original image is presented together with its approximations using successively 128, 256 and 512 coefficients. With only 128 coefficients (compression ratio of order 1/2000), the reconstruction of the edge is near-perfect.

9 Discussion

The point of this paper has been the quantitative study of the properties of estimation by finite linear combinations of ridgelets. In contrast to existing approaches based on stepwise addition of elements, we suggest a new approach based on a new tool, the ridgelet transform: expanding the noisy data into a ridgelet series and simply thresholding the noisy coefficient. We have shown that this is a powerful method for statistical estimation. Roughly speaking, one can read the estimation bounds from the sparsity of the ridgelet coefficients. We have identified many situations where the

ridgelet shrinkage is optimal and, in addition, we have also been able to study its limitations.

9.1 Choice of model

We would like to stress that the framework of all of our quantitative estimation results is that of the continuous white-noise model (3.1) of Ibragimov and Hasminskii (1981). Although this model is of common use in the literature, one may object that this model serves the author's purpose. There could be two main objections: first, it is not discrete while in practice one is presented discrete data; and, second, the implicit assumption is that the setting is in some sense uniform as the performance is evaluated with respect to the Lebesgue measure. Both of these objections are well founded and we shall attempt to address them both.

Discrete data. We present the situation in dimension two: suppose we observe noisy measurements

$$y_{i,j} = \tilde{f}(i,j) + \sigma z_{i,j},$$

where $z_{i,j} \overset{i.i.d.}{\sim} N(0,1)$ is a Gaussian noise term. In a lot of physical devices, the $\tilde{f}(i,j)$'s are gridded data of level-pixel averages

$$\tilde{f}(i,j) = \mathrm{Ave}\{f \mid [i/n, i+1/n) \times [j/n, j+1/n)\} \quad 0 \le i,j < n.$$

We wish to recover f with small per-pixel mean-squared error $MSE(\hat{f}, f) = En^{-2} \sum_{i,j} (\hat{f}(i,j) - \tilde{f}(i,j))^2$. The problem of recovering objects with edges from gridded data is not trivial (see Korostelev and Tsybakov, 1993, for example). However, the author is confident that a careful analysis will give discrete versions of Theorems 5.4 and 6.1. We hope to report on this in later papers. (Precise bounds will probably depend on the implementation that is chosen.) We would like to point out that although the benefit of wavelet methods was pointed out quite a while ago, it is only fairly recently that results have been transported from the continuous white-noise model to equispaced regular designs.

Regular setting. Even though one may expect to see ridgelet algorithms enjoy nice estimation bounds with data given on a regular grid, there does not seem to be a quick answer to the problem of dealing with irregularly spaced (heterogeneous) data points. This is indeed a fairly classical problem that a lot of theoretically motivated methods have to deal with. For instance, it is not always clear how to use the Fast Fourier transform or Fast wavelet transforms to handle nonequispaced data points on the real line. Although these issues have been around for a long time, their careful study is fairly recent Silverman, 1999.

As we can see, the issues that we raised are shared by many popular methods in current use and are far from being the sole appanage of ridgelet procedures. Practical work on those issues will undoubtedly be of great importance. The author hopes to report on some empirical work in a later paper.

9.2 Curved edges

Finally, ridgelets are optimal for estimating objects with singularities across hyperplanes (Section 5), but they fail efficiently to estimate objects with singularities across curved hypersurfaces (Section 7).

One can adapt to this situation by localizing the ridgelets. We divide the domain in question into squares and smoothly localize the function into smooth pieces supported on or near those squares either by partition of unity or by smooth orthonormal windowing. We then apply ridgelet methods to each piece. The idea is that, at sufficiently fine scale, a curving singularity looks straight, and so ridgelet analysis – appropriately localized – works well in such cases. This strategy has been fully developed in Candes (1999a) and is shown to provide better estimation bounds than (7.2).

A more promising approach is based on a new transform, namely, the curvelet transform pioneered by Candes and Donoho (1999). In two dimensions, the curvelet transform combines ideas from ridgelet and wavelet analysis to provide optimal representations of smooth functions with twice differentiable singularities. All of these refinments are grounded on the work presented in this paper.

9.3 A last word

In this paper, we presented the mathematical foundations and some early numerical experiments of a new approach. However, the previous comments made clear that this work opens up many challenging questions and, therefore, it should only be interpreted as a starting point for further investigation.

10 Appendix

In this appendix, we will give rigorous proofs of some hypercube embedding results (Lemma 5.3 and 6.3) needed to support the claims about lower rates of convergence (section 5 and 6). Theorem 6.3 is proved in the author's unpublished thesis and is reproduced here, with the argument of an intermediate technical result removed, however.

It is important to note that the proof of the existence of lower bounds of estimation does not need to be constructive. This observation greatly simplifies our argument. Interestingly, 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.

Let u be uniformly distributed on the unit sphere. Then, for any other unit vector u', the density of $u_1 = u \cdot u'$ is given by

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

where c_d is a renormalizing constant. A simple change of variables formula then gives the density of the tangent $v = u \cdot u'/\sqrt{1 - (u \cdot u')^2}$ between the vectors u and u', namely,

$$f(v) = c'_d(1 + v^2)^{-d/2}. \tag{10.1}$$

We now introduce discrete packing sets on the sphere with properties mimicking the continuous ones listed above. In all that follows, j_0 will denote a nonnegative integer whose value will be decided later. For a fixed $j \geq j_0$, let $\epsilon_j = 2^{-(j-j_0)}$ and let S_j be a set of points on the sphere (u_ℓ) satisfying the following properties:

1. $\forall u_\ell, u_{\ell'} \in S_j, \quad \|u_\ell \pm u_{\ell'}\| \geq \epsilon_j,$
2. $B_1\ \epsilon_j^{-(d-1)} \leq |S_j| \leq B_2\ \epsilon_j^{-(d-1)}$, and
3. for any $u \in \mathcal{S}^{d-1}$, and all $0 \leq m \leq j - j_0$,

$$|\{u_\ell,\ \ 2^{m-1} \leq \frac{|u \cdot u_\ell|}{(1 - (u \cdot u_\ell)^2)^{1/2}} \leq 2^m\}| \leq B_2\ \epsilon_j^{-(d-1)} \int_{2^{m-1} \leq |v| \leq 2^m} \frac{dv}{(1+v^2)^{d/2}}$$

In the above expressions, the constants B_1 and B_2 can be chosen to be independent of ϵ_j.

Let $v_{\ell,\ell'} = u_\ell \cdot u_{\ell'}(1 - (u_\ell \cdot u_{\ell'})^2)^{-1/2}$ be the absolute value of the tangent between u_ℓ and $u_{\ell'}$. We remark that the first property implies that

$$\{u_{\ell'},\ \ \frac{|u_\ell \cdot u_{\ell'}|}{(1 - (u_\ell \cdot u_{\ell'})^2)^{1/2}} \geq \epsilon_j^{-1}\} = \{u_\ell\}.$$

This fact is a mere consequence of

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

Indeed, suppose for instance that $v_{\ell,\ell'} \geq \epsilon_j^{-1}$. Then,

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

Therefore, $v_{\ell,\ell'} \geq \epsilon_j^{-1}$ implies $\|u_{\ell'} - u_\ell\| < \epsilon_j$. It then follows from the first property that this is equivalent to $\ell = \ell'$. The argument is identical in the case $v_{\ell,\ell'} \leq -\epsilon_j^{-1}$.

To further simplify the analysis, suppose $\psi \in \mathcal{S}(\mathbb{R})$ 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 $0 \le w \le 1$ and $w(x) = 1$ for any x with $\|x\| \le \sqrt{3}/2$. We now consider the set A_j of windowed ridgelets at scale j

$$A_j = \{f_{\ell,k}(x) = 2^{j/2}\psi(2^j u_\ell \cdot x - k)w(x), \;\; u_\ell \in S_j, \; k \in \mathbb{Z} \;\text{ and }\; |k|2^{-j} \le 1/2\}. \tag{10.2}$$

Finally, we will assume $j \ge 2$ so that $1/2 + 2^{-j}/2 \le \sqrt{2}/2$; from our assumptions it follows that supp $f_{\ell,k} \subset \{x, \;\; |u_\ell \cdot x| \le \sqrt{2}/2\}$ for any $f_{\ell,k}$ 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 10.1 *The cardinality of A_j is bounded below by*

$$\#A_j \ge C\, 2^{jd}.$$

Next, the elements of A_j satisfy the following two properties:

(i) there is a constant c_d (only depending upon the dimension d) s.t.

$$\forall \;\; f \in A_j, \qquad \|f\|_2 \ge c_d, \tag{10.3}$$

(ii) and if j_0 is chosen large enough,

$$\forall \;\; f \in A_j, \qquad \sum_{g \in A_j, g \ne f} |\langle f, g\rangle| \le \frac{c_d}{2}. \tag{10.4}$$

Proof of Lemma. The norm of $f_{\ell,k}$ being clearly invariant by rotation (w radial), one can assume that $u_\ell = e_1$, with e_1 being the first vector of the canonical basis of $\mathbb{R}^d$. We have

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

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 $0 < \eta \le \epsilon \le 1$, $x \in \mathbb{R}$, $y \in \mathbb{R}$, and $\delta > 0$ we have

$$\sum_{k \in \mathbb{Z}} (1 + |x - \epsilon k|)^{-1-\delta} (1 + |y - \eta k|)^{-1-\delta} \le C_\delta \epsilon^{-1} (1 + |y - x\eta\epsilon^{-1}|)^{-1-\delta}. \tag{10.5}$$

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

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

Next, an application of Lemma 10 from Candes (1998) when $u_\ell \neq u_{\ell'}$ shows that one can find a constant $C_1(d)$ depending on d, ψ and w such that,

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

Now, it follows from (10.5) that

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

for some new constant $C_2(d)$, depending only on d, ψ and w. Summing over $u_{\ell'}$ ($u_{\ell'} \neq u_\ell$) and making use of the third assumption on the u_ℓ's gives (recall $\epsilon_j = 2^{-(j-j_0)}$)

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

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 $|\{u_{\ell'}, 0 \leq |v_{\ell,\ell'}| \leq 1\}|$ instead of $|\{u_{\ell'}, 2^{\ell-1} \leq |v_{\ell,\ell'}| \leq 2^\ell\}|$ 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)*. ∎

The next lemma is proved in Candes (1998).

Lemma 10.2 *First, the elements $f_{\ell,k}$ satisfy*

$$\|f_{\ell,k}\|_{R^s_{p,q}} \leq C\ 2^{js} 2^{jd(1/2-1/p)}.$$

Second, let $\mathcal{C}$ the parallelepiped be defined by

$$\mathcal{C} = \{f, \;\; f = \sum_{\ell,k} \xi_{\ell,k} f_{\ell,k}, \;\; |\xi_{\ell,k}| \leq 1\}. \tag{10.6}$$

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

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

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

Note that the previous lemma shows how to construct a full parallelepiped embedded in $R^s_{p,q}$. However, in view of Lemma 6.3 one needs to construct a cube. The next lemma shows how to orthogonalize our parallelepiped.

Lemma 10.3 *Suppose we have n vectors $\{f_i\}_{1 \leq i \leq n}$ in a Hilbert space such that for all $1 \leq i \leq n$*

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

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

We consider the set $\mathcal{C} = \{\sum_{i=1}^n y_i f_i, \; \|y\|_\infty \leq 1\}$. 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 $\{e_i\}_{1 \leq i \leq n}$ defined by $e_i = H f_i$ is indeed an orthogonal basis of span$(\{f_i\}_{1 \leq i \leq n})$ (see 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 requires nothing 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 (1992) 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. ∎

Lemma 6.3 is now a mere consequence of the three preceding preparatory lemmas.

As far as the linear estimation is concerned, Lemma 5.3 essentially follows from Lemma 10.2 and (6.4). Indeed, chasing definitions, the closed convex hull $\overline{Hull(\mathcal{F})}$ contains $\mathcal{S}_\mathcal{H}$ which in turn contains a ball of $R_{1,1}^{(d+1)/2}$. Hence, it is sufficient to prove the appropriate embedding in a ball of $R_{1,1}^{(d+1)/2}$. By Lemma 10.2, we have

$$\mathcal{C} = \{f,\ \ f = \sum_{\ell,k}\xi_{\ell,k}f_{\ell,k},\ \ \sum|\xi_{\ell,k}| \leq 1\} \subset \{f,\ \|f\|_{R_{1,1}^{(d+1)/2}} \leq C\,2^{j/2}\}.$$

We use the same orthogonalization procedure as in Lemma 10.3 and conclude that one can construct a set of orthogonal functions $g_{\ell,k}$ (constructed in the same way as in the proof of Lemma 10.3) such that

$$\mathcal{C}' = \{f,\ \ f = \sum_{\ell,k}\xi_{\ell,k}g_{\ell,k},\ \ \sum|\xi_{\ell,k}| \leq 1\} \subset \mathcal{C} \subset \{f,\ \|f\|_{R_{1,1}^{(d+1)/2}} \leq C\,2^{j/2}\}.$$

The proof of this fact is identical to that of Lemma 10.3; keeping the notation of this lemma, one needs to check that the norm of H, as an operator from $\ell_1 \to \ell_1$ now, is bounded by δ^{-1}. We recall that

$$\|H\|_{(\ell_1,\ell_1)} = \sup_j \sum_i |H_{i,j}|,$$

and the desired bound on the norm is proved in the same way as before.

A simple rescaling finally gives Lemma 5.3 (the quantity $2^{-j/2}$ playing the role of δ in the statement of this lemma).

References

Candes, E. J. (1998). *Ridgelets: theory and applications.* Unpublished doctoral dissertation, Department of Statistics, Stanford University.

Candes, E. J. (1999a). *Monoscale ridgelets for the representation of images with edges* (Tech. Rep.). Department of Statistics, Stanford University.

Candes, E. J. (1999b). *On the representation of mutilated sobolev functions* (Tech. Rep.). Department of Statistics, Stanford University.

Candes, E. J. (1999c). *Ridgelet representations of new smoothness classes* (Tech. Rep.). Department of Statistics, Stanford University.

Candes, E. J. (1999d). Harmonic analysis of neural netwoks. *Applied and Computational Harmonic Analysis*, *6*, 197–218.

Candes, E. J., and Donoho, D. L. (1999). *Curvelets* (Tech. Rep.). Department of Statistics, Stanford University.

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.

Deans, S. R. (1983). *The Radon transform and some of its applications.* John Wiley & Sons.

Donoho, D., and Johnstone, I. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, *81*, 425–455.

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., 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., and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.*

Efroimovich, S., and Pinsker, M. (1982). Estimation of square-integrable [spectral] density based on a sequence of observations. *Problems of Information Transmission*, 182–196.

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.

Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (1998). *Wavelets, approximation, and statistical applications* (Vol. 129). New York: Springer-Verlag.

Ibragimov, I. A., and Hasminskii, R. Z. (1981). *Statistical estimation. Asymptotic theory.* New York-Berlin: Springer-Verlag.

Johnstone, I. M. (1999). *Wavelets and the theory of nonparametric function estimation.* (Available at: `http://www-stat.stanford.edu/~imj`)

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

Korostelev, A. P., and Tsybakov, A. B. (1993). *Minimax theory of image reconstruction* (Vol. 82). New York: Springer-Verlag.

Meyer, Y. (1992). *Wavelets and operators.* Cambridge University Press.

Pinsker, M. (1980). Optimal filtering of square-integrable signals in gaussian white noise. *Problems of Information Transmission*, 120–133.

Silverman, B. (1999). Wavelets in statistics: beyond the standard assumptions. *to appear Phil. Trans. R. Soc. Lond. A.*

Stone, C. J. (1977). Consistent nonparametric regression, with discussion. *Ann. Statist.*, *5*, 595–645.

Vu, V. H. (1998). On the infeasibility of training neural networks with small mean-squared error. *IEEE Transactions on Information Theory*, *44*, 2892–2900.