*The Annals of Statistics*
2003, Vol. 31, No. 5, 1561–1599

# RIDGELETS: ESTIMATING WITH RIDGE FUNCTIONS[1]

BY EMMANUEL J. CANDÈS

*Stanford University*

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—Candès [*Appl. Comput. Harmon. Anal.* **6** (1999) 197–218] 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 (damping); 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.

**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

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

where $\varepsilon$ 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

Received October 1999; revised November 2002.

[1] Supported by NSF Grant DMS-95-05151 and by AFOSR MURI 95-P49620-96-1-0028.

*AMS 2000 subject classifications.* Primary 62G07, 62C20; secondary 41A30.

*Key words and phrases.* Nonparametric regression, ridgelets, ridge functions, projection pursuit regression, minimax decision theory, radon transform, spatial inhomogeneities, edges, thresholding of ridgelet coefficients.

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, that is, $\|u_j\| = 1$. 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 residual 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: $|u^T x - t| \le h$, thus avoiding the problems relative to the inherent sparsity of the sample.

1.2. *Feedforward neural networks.* There are many different kinds of neural networks and one of the most commonly discussed classes of neural nets is the class of so-called feedforward neural networks. These neural nets are indeed very much in use in statistics for regression, classification, discrimination, and so on [see the survey of Cheng and Titterington (1994) and its accompanying 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 $\alpha_j$ and $b_j$ are scalars; and the $k_j$ are $d$-dimensional vectors. In that field, $\rho$ 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 $(\alpha^*, k^*, b_*)$ are solutions of the optimization problem

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

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 Euclidean norm.

Of course, PPR and feedforward 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 nonsmooth 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. 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). Important references—with evocative titles—would be "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).

   Even if one is willing to ignore the difficulty of implementing a stepwise addition of elements, one may wonder about the efficiency of such a procedure. It is well known that greedy procedures may have weak estimation properties as the inability to look ahead may cause initial errors that the fitting 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 parameters $\alpha_j, k_j, b_j$ 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 Candès (1999a). The ridgelet transform represents quite general functions as a superposition 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 expanding the data into a ridgelet series and applying 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 4). 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 5). 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 feedforward neural networks (we would like to emphasize that our analysis only applies to these types of neural nets). As a surprising example, the estimation of radial functions with projection based approaches will be discussed (Section 6). Here, we use the word "surprising" because our results suggest a different behavior than that which is expected from the literature.

Finally, some numerical experiments will illustrate the power of these new ideas (Section 7). The discussion Section 8 will tie the methodology presented here with the classical neural network approach, survey some extensions of the present work and identify areas for future research.

**2. Ridgelets.** This section introduces the ridgelet transforms and surveys some of their main properties. All of the forthcoming claims and results are proved in Candès (1999a). For now, $\hat{g}$ will denote the Fourier transform of $g$:

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

2.1. *The continuous ridgelet transform.* In $d$ dimensions, the ridgelet construction starts with a univariate function $\psi$ satisfying an oscillatory condition, namely,

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

a ridgelet is a function of the form

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

where $a$ and $b$ are scalar parameters and $u$ is a vector of unit length. Of course, a ridgelet is a ridge function and resembles a neuron but for the oscillatory behavior of the profile (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 $\{x,\ |u^T x - b| \le a\}$. Ridgelets are pictured in Figure 1 for various values of these parameters.

Define a ridgelet coefficient as

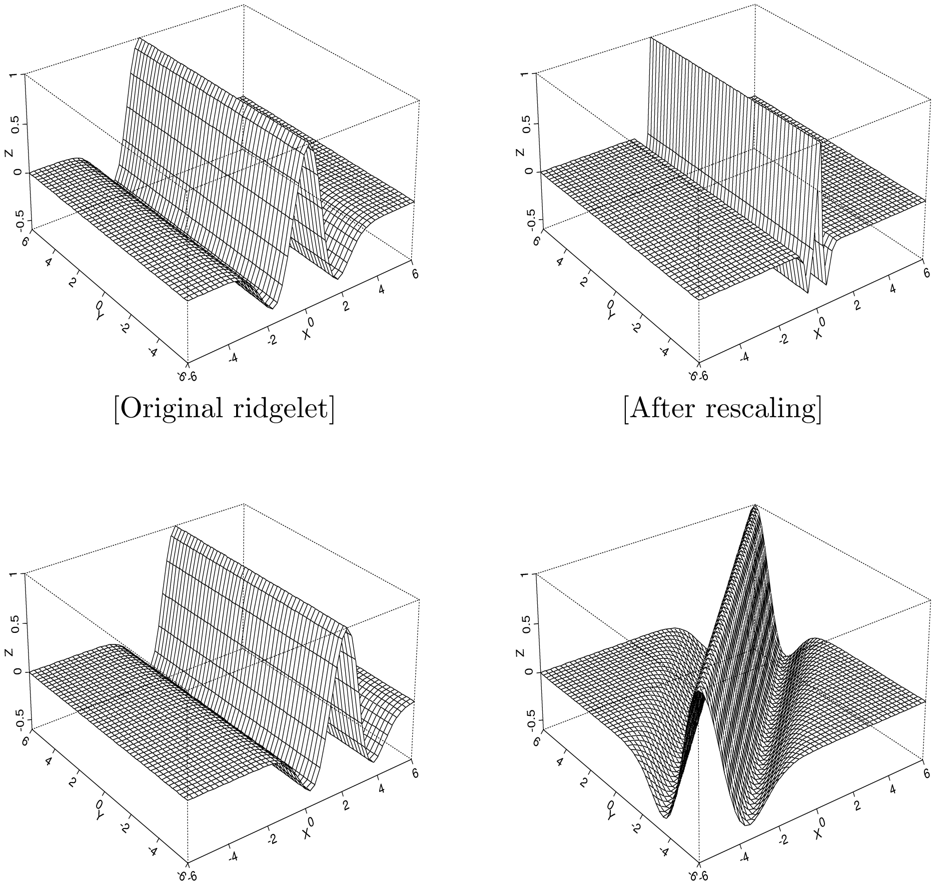
$$\mathcal{R}_f(a,u,b) = \int f(x)\, a^{-1/2}\psi\left(\frac{u^T x - b}{a}\right) dx; \tag{2.4}$$

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

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

where $d\mu(a,u,b) = da/a^{d+1}\, du\, db$ ($du$ being the uniform measure on the sphere) which holds true if $\psi$ is properly normalized, that is, $K_\psi = 1/(2\pi)^{d-1}$ in (2.2). Equation (2.5) expresses the idea that one can represent any function as a superposition of these ridgelets. Furthermore, this formula is stable as one has a Parseval relation

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



[Original ridgelet] [After rescaling]

[After shifting] [After rotation]

FIG. 1. *Ridgelets.*

2.2. *The discrete ridgelet transform.* Similar to the continuous transform, there is a discrete transform. Let $i$ be the triple $(j, \ell, k)$ where the indices run as follows:

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

and define the collection of discrete ridgelets

$$\psi_i(x) = 2^{j/2}\psi(2^j u_\ell^T x - k), \qquad i \in \mathcal{I}. \tag{2.7}$$

Note that the range of the parameter $\ell$ is scale dependent as it depends on $j$. Ridgelets are directional and, here, the interesting aspect is the discretization of the directional variable $u$; this variable is sampled at increasing resolution so that at scale $j$, the discretized set is a net of nearly equispaced points at a distance of order $2^{-j}$; a detailed exposition of the ridgelet construction is given in Candès (1999a).

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

$$A\,\|f\|^2\le\sum_{i\in\mathcal{I}}|\langle f,\psi_i\rangle|^2\le B\,\|f\|^2. \tag{2.8}$$

The previous equation says that the datum of the ridgelet transform at the points $(a=2^j,u=u_\ell,b=k2^{-j})_{(j,k,\ell)\in\mathcal{I}}$ 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 $(\tilde{\psi}_i)_{i\in\mathcal{I}}$ with the property

$$f=\sum_{i\in\mathcal{I}}\langle f,\tilde{\psi}_i\rangle\psi_i=\sum_{i\in\mathcal{I}}\langle f,\psi_i\rangle\tilde{\psi}_i, \tag{2.9}$$

which gives perfect and stable reconstruction.

2.3. *Why a discrete transform*? Various completeness theorems are known for the set of neurons $\mathcal{D}_{NN}=\{\rho(k^Tx-b),\ k\in\mathbb{R}^d,b\in\mathbb{R}\}$; 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 $f$, that is, for any $\varepsilon>0$, one can find parameter values $(k_j,b_j)_{1\le j\le J}$ such that

$$\left\|f-\sum_{j=1}^{J}\alpha_j\rho(k_j^Tx-b_j)\right\|_2<\varepsilon.$$

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 (1.4) and in any real implementation, one would probably need to restrict the search for a minimum over a grid. What are the properties of a restricted search? Is there a grid preserving the completeness property? If so, what is the proper spacing of this grid? In other words, what is the real complexity of the search (1.4)? In our opinion, the discretization (2.7) gives a precise answer to these questions.

## 3. Thresholding with noisy data.

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

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

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 $\mathcal{R}_\varepsilon(\mathcal{F})$ be the minimax mean-squared error in the white noise model

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

where of course the estimates $\hat{f}$ are restricted to be obtained through measurable procedures, that is, $\hat{f} = \widehat{F}(Y_\varepsilon)$, 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.

3.2. *The sequence space view.* 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 and Johnstone (1998) where the wavelet coefficients are to be estimated.

Thus, we define the empirical ridgelet coefficients

$$y_i = \langle Y_\varepsilon, \psi_i \rangle, \qquad i \in \mathcal{I},$$

which obey the Gaussian model

$$y_i = \theta_i + \varepsilon z_i, \qquad \theta_i := \langle f, \psi_i \rangle, \qquad i \in \mathcal{I}, \tag{3.4}$$

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 $V_{i,j} = \langle \psi_i, \psi_j \rangle$, for example, $y_i \sim N(\theta_i, \varepsilon^2 \|\psi_i\|^2)$. An estimate $(\hat{\theta})$ of the coefficient sequence automatically defines a function estimate by the reconstruction rule $\hat{f} = \sum_{i \in \mathcal{I}} \widehat{\theta_i}\, \tilde{\psi}_i$. A classical result in analysis gives

$$\|\widehat{f} - f\|_2^2 \le A^{-1} \|\widehat{\theta} - \theta\|^2_{\ell_2(\mathcal{I})}, \tag{3.5}$$

where $A$ is the constant appearing on the left-hand side of (2.8). 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.

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

$$\hat{\theta}_i = \eta_i(y_i) = \eta_i(\langle \psi_i, Y_\varepsilon \rangle), \tag{3.6}$$

yielding simple estimates of the form $\widehat{f} = \sum_{i \in \mathcal{I}} \eta_i(\langle \psi_i, Y_\varepsilon \rangle) \tilde{\psi}_i$; the scalar nonlinearities $\eta_i$ will be made explicit below.

3.4. *Thresholding in the white noise model.* Although the sequence model (3.4) does not assume independently distributed errors, existing work suggests the construction of level-dependent thresholding rules; see Johnstone (1999) for an excellent account. In particular, it is now well established that the quality of the estimation is linked to the sparsity of the vector $\theta$. In addition, Johnstone and Silverman (1997) show that scalar thresholding rules come close to the minimax risk provided suitable conditions about the correlation matrix.

Our exposition now closely follows the concept of oracle inequalities developed by Donoho and Johnstone (1994). We introduce some notation. Let $\eta_{\mathrm{S}}$ denote the soft threshold nonlinearity

$$\eta_{\mathrm{S}}(y, \lambda) = \mathrm{sgn}(y)\,(|y| - \lambda)_+ \tag{3.7}$$

and $r_{\mathrm{S}}(\varepsilon; \lambda, \mu)$ the risk of the latter rule, that is,

$$r_{\mathrm{S}}(\varepsilon; \lambda, \mu) = E[\eta_{\mathrm{S}}(Y, \lambda) - \mu]^2, \qquad Y \sim N(\mu, \varepsilon^2).$$

[In the case $\varepsilon = 1$, we will simply write $r_{\mathrm{S}}(\lambda, \mu)$.] We borrow the following lemma from Johnstone (1999).

LEMMA 3.1. *Let* $\bar{r}(\lambda, \mu) = \min\{r_{\mathrm{S}}(\lambda, 0) + \mu^2, 1 + \lambda^2\}$. *Then for any choice of threshold* $\lambda$ *and* $\mu \in \mathbb{R}$,

$$\tfrac{1}{2}\bar{r}(\lambda, \mu) \le r_{\mathrm{S}}(\lambda, \mu) \le \bar{r}(\lambda, \mu). \tag{3.8}$$

For any choice of threshold, we always have $r_{\mathrm{S}}(\lambda, \mu) \ge 1/2 \min(\mu^2, 1)$. Similar inequalities exist for hard thresholding rules. For instance, letting $\eta_{\mathrm{H}}(y, \lambda) = y\,\mathbb{1}_{\{|y| > \lambda\}}$, we have $r_{\mathrm{H}}(\lambda, \mu) \ge \xi(\lambda) \min(\mu^2, 1)$, where $\xi$ is some function bounded away from zero, $0 < \xi < 1$, which tends to 1 when its argument tends to $\infty$ [Donoho (1993)].

Let $(y_i)$ be as above ($y_i \sim N(\theta_i, \sigma_i^2)$) and suppose now that $\mathcal{I}'$ is a finite subset of $\mathcal{I}$. Put

$$\widehat{\theta}_i = \begin{cases} \eta_{\mathrm{S}}(y_i, \lambda \cdot \sigma_i), & i \in \mathcal{I}', \\ 0, & i \in \mathcal{I} \setminus \mathcal{I}'. \end{cases} \tag{3.9}$$

Then, for any $\lambda$ a simple rescaling argument shows that

$$E\|\hat{\theta}-\theta\|_{\ell_2}^2 \geq \sum_{i\in\mathcal{I}'} \tfrac{1}{2}\min(\varepsilon^2\sigma_i^2,\theta_i^2) + \sum_{i\in\mathcal{I}\setminus\mathcal{I}'} \theta_i^2. \tag{3.10}$$

Hence, the sparsity of the coefficient sequences (ridgelets, wavelets, etc.) automatically gives lower estimation bounds of thresholding rules. On the other hand, the choice $\lambda = \varepsilon\sqrt{2\log(\#\mathcal{I}')}$ gives the upper bound

$$E\|\hat{\theta}-\theta\|_{\ell_2}^2 \leq (1+2\log(\#\mathcal{I}'))\left(\varepsilon^2\bar{\sigma}^2 + \sum_{i\in cI'} \min(\theta_i^2,\varepsilon^2\sigma_i^2)\right) + \sum_{i\in\mathcal{I}\setminus\mathcal{I}'} \theta_i^2, \tag{3.11}$$

with $\bar{\sigma}^2$ a shorthand for $\{\#\mathcal{I}'\}^{-1}\sum_{i\in\mathcal{I}'}\sigma_i^2$. This is often referred to as the oracle inequality [Donoho and Johnstone (1994)].

## 4. Ridgelets and linear singularities.

4.1. *Linear singularities.* Consider the mutilated Gaussian defined as follows:

$$f(x) = \mathbb{1}_{\{u^T x \geq b\}} e^{-|x|^2/2}. \tag{4.1}$$

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. In this nontechnical section, we shall discuss the recovery of the mutilated Gaussian from noisy data (3.1) and will use the integrated mean-squared error (3.2) to measure performance. [We carefully acknowledge that the mutilated Gaussian is not supported in the unit cube and, therefore, does not fit into the statistical paradigm we set up. We chose the mutilated Gaussian merely for its evocative power and the conscious reader may substitute the Gaussian $e^{-|x|^2/2}$ with a nice $C^\infty$ function $g$ supported in the unit cube in definition (4.1).]

We are going to compare the performance of our ridgelet shrinkage estimator (3.6)–(3.9) to that of kernel smoothers [Stone (1977)] or wavelet-based estimators as proposed by Donoho and Johnstone. Write $\widehat{f}_{\mathrm{KS}}$ to denote an estimator obtained by kernel smoothing and similarly, $\widehat{f}_{\mathrm{WT}}$ for a wavelet shrinkage estimator.

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

$$MSE(\widehat{f}_{\mathrm{KS}}, f) \geq C\,(\varepsilon^2)^{1/(d+1)}. \tag{4.2}$$

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 (4.2). 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 smoothes out the edge or undersmoothes 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 ill-suited for linear procedures.

What about nonlinear procedures? Following Donoho and Johnstone, we now examine the situation for a wavelet thresholding estimator and argue that the performance of such an estimator obeys

$$MSE(\widehat{f}_{\mathrm{WT}}, f) \geq C\,(\varepsilon^2)^{1/d}, \tag{4.3}$$

regardless of the selection of a wavelet basis, thresholding rule (hard, soft, etc.) and thresholding parameter. This result is a direct consequence of (3.10) together with the fact that the wavelet coefficient sequence of $f$ is nonsparse. A simple calculation that we will omit shows that the number of coefficients whose squared modulus exceeds $\varepsilon^2$ is bounded below by $c \cdot \varepsilon^{-2+2/d}$ and therefore the proxy $\sum \min(\theta_i^2, \varepsilon^2)$ for the risk is above a constant times $\varepsilon^{2/d}$, which establishes (4.3).

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)] (one would certainly select a smaller bandwidth in a neighborhood of the discontinuity). The result is striking: such a nonlinear 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, and so on. This nice feature is certainly one of the reasons why they generated and continue to generate so much enthusiasm and now play a salient role in the literature of Statistics. In higher dimensions, however, there are various kinds of spatial inhomogeneities and our example surely illustrates an important one. It shows that, in some sense, wavelets present a distinguished feature which operates in dimension one but 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 two, 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.

We now turn our attention to a simple ridgelet thresholding estimate $\widehat{f}_{\mathrm{RT}}$ as in (3.6)–(3.9). Then

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

Unlike wavelets, ridgelets adapt very well to linear inhomogeneities. The reason is that the singularity causes highly concentrated or localized effect to the ridgelet representation, giving only a few significant coefficients to estimate. This phenomenon justifies a slogan which says that "ridgelets are provably 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 data over strips. For objects like (4.1), this seems to be a clear advantage, especially if the strip may be positioned along the edge.

4.2. *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 $\|g\|^2_{W_2^s} = \|g\|_2^2 + \|D^s g\|_2^2$. [When $s$ is not an integer, the norm is given via the Fourier transform $\hat{g}$ (2.1), $\|g\|^2_{W_2^s} = \int_{\mathbb{R}^d} (1 + |\xi|^{2s}) |\hat{g}(\xi)|^2 \, d\xi$.]

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

$$\begin{aligned} \mathcal{F}(C) = \big\{ f,\ &\|f\|_{W_2^s(\mathbb{R}^d \setminus L)} \le C, \\ &\text{for some hyperplane } L, \text{ and } \operatorname{supp} f \subset [0,1]^d \big\}. \end{aligned} \tag{4.4}$$

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 4.1. *Let $\mathcal{R}_{\mathrm{L}}(\varepsilon, \mathcal{F})$ be the minimax rate where the infimum* (3.3) *is restricted over linear procedures. Then, there exists a constant $C$ such that*

$$\mathcal{R}_{\mathrm{L}}(\varepsilon, \mathcal{F}) \ge C(\varepsilon^2)^{1/(d+1)}. \tag{4.5}$$

This fully justifies our claim (4.2).

REMARK. Linear estimation of discontinuous functions has been studied by Korostelev and Tsybakov [(1993), page 178] although their estimation problem is different from (4.4). They wish to recover elements of the form

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

where $\varphi$ 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 $\varphi$ to be linear. However, translated to our framework, their lower bound is of order $(\varepsilon^2)^{1/2}$ when, say, the singularity $\varphi$ is $C^\infty$ and the $f_i$'s are smooth enough, which is not the correct order (not sharp), as suggested by Theorem 4.1. Our method is different from theirs as ridgelets play a central role in the determination of (4.5).

PROOF OF THEOREM 4.1. 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 4.2. *We have*

$$\mathcal{R}_{\mathrm{L}}(\varepsilon, \mathcal{F}) = \mathcal{R}_{\mathrm{L}}\bigl(\varepsilon, \mathrm{Hull}(\mathcal{F})\bigr). \tag{4.6}$$

This is a classical result and we only sketch the argument—mainly to introduce some notation. For a linear estimator of the form $\hat{f} = TY$ the classical bias-variance decomposition gives

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

where $\|T\|_{\mathrm{HS}}$ is the Hilbert–Schmidt norm of the operator $T$ ($\|T\|_{\mathrm{HS}}^2 = \sum_n |Te_n|^2$ with $(e_n)$ any orthobasis of $L_2[0,1]^d$). The variance is independent of the estimand and for $f$ in the convex hull of $\mathcal{F}$ ($\sum_i a_i f_i$, $f_i \in \mathcal{F}$, $\sum_i |a_i| \le 1$), the squared bias obeys

$$\|(I - T)f\|_2^2 \le \sup_i \|(I - T)f_i\|_2^2 \le \sup_{g \in \mathcal{F}} \|(I - T)g\|_2^2,$$

which proves the 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 $\overline{\mathrm{Hull}(\mathcal{F})}$. 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 (4.5). A rigorous argument involves a delicate construction whose proof may be found in the Appendix.

LEMMA 4.3. *For any $\delta > 0$, there exist $m(\delta)$ orthogonal elements $\{g_\ell\} \in \overline{\mathrm{Hull}(\mathcal{F})}$ satisfying the following properties*:

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

We use this lemma to complete 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

$$\mathcal{R}_{\mathrm{L}}\big(\varepsilon, \mathrm{Hull}(\mathcal{F})\big) = \mathcal{R}_{\mathrm{L}}\big(\varepsilon, \overline{\mathrm{Hull}(\mathcal{F})}\big) \ge \mathcal{R}_{\mathrm{L}}(\varepsilon, \mathcal{V}_\delta).$$

Now, the linear minimax rate is given by

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

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, $\varepsilon^2\|T\|_{\mathrm{HS}}^2$. In the former case, we will use the upper bound on the bias to get a lower bound on the variance term, that is, $\|T\|_{\mathrm{HS}}^2$. Indeed, it is not hard to show that

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

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

$$\mathcal{R}_{\mathrm{L}}(\varepsilon, \mathcal{V}_\delta) \ge \tfrac{1}{2}\min\big(\delta^2, \varepsilon^2 m(\delta)\big).$$

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

$$\mathcal{R}_{\mathrm{L}}(\varepsilon, \mathcal{V}_{\delta_\varepsilon}) \ge C(\varepsilon^2)^{1/(d+1)}.$$

We trivially conclude that

$$\mathcal{R}_{\mathrm{L}}\big(\varepsilon, \mathrm{Hull}(\mathcal{F})\big) \ge \mathcal{R}_{\mathrm{L}}(\varepsilon, \mathcal{V}_{\delta_\varepsilon}) \ge C(\varepsilon^2)^{1/(d+1)}.$$

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 (3.6)–(3.9). We will work with a nice ridgelet frame (2.7) $(\psi_i)_{i \in \mathcal{I}}$ such that $\psi$ has enough vanishing moments and regularity. To simplify the analysis we take $\varphi$ and $\psi$ to be compactly supported. Hence, at a given scale $j$, the number of ridgelets that are nonzero on $[0, 1]^d$ is bounded by

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

for some fixed constant $C$.

To estimate the true ridgelet coefficients $\theta$ from our noisy data $y$ (3.4), we consider the diagonal projection as defined in Section 3. Set a thresholding zone

$$\mathcal{I}' = \{i,\ j(i) \le J_\varepsilon\}$$

and define

$$\widehat{\theta}_i = \begin{cases} \eta_S(y_i, \lambda\sigma_i), & i \in \mathcal{I}', \\ 0, & i \in \mathcal{I} \setminus \mathcal{I}', \end{cases} \tag{4.7}$$

with $\lambda = \varepsilon\sqrt{2\log(\#\mathcal{I}')}$ and where we recall that the $\sigma_i$'s are the $L_2$-norms of the ridgelets $\psi_i$. Thus, the estimator (4.7) sets to zero all the coefficients outside of a thresholding zone, namely, exceeding a given scale and applies a thresholding to the others.

THEOREM 4.4. *Consider the ridgelet thresholding estimate* $\hat{f}$ (3.6)–(3.9) [*with* (4.7) *as the choice of scalar nonlinearities*]. *Then*,

$$\sup_{\mathcal{F}} MSE(\hat{f}, f) \le C\big(1 + 2\log(\varepsilon^{-1})\big)(\varepsilon^2)^{2s/(2s+d)}.$$

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

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

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

$$\inf_{\hat{f}} \sup_{f \in \mathcal{W}(s,C)} MSE(f, \hat{f}) \ge c\,(\varepsilon^2)^{2s/(2s+d)}.$$

It is interesting to note that our estimator achieves an error of estimation which, ignoring log-like factors, is 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.

PROFF OF THEOREM 4.4. Following the argument developed in the previous section, we simply need to study the sparsity of the ridgelet coefficient sequence.

We invoke the oracle inequality (3.11) and the upper bound will result from the following two facts that are proven in Candès (2001): first,

$$\sum_i \min(\theta_i^2, \varepsilon^2) \le C(\varepsilon^2)^{2s/(2s+d)}, \tag{4.8}$$

and second, the organization of the Fourier transform gives

$$\sum_{j(i) > J_\varepsilon} \theta_i^2 \le C \max(2^{-2J_\varepsilon s}, 2^{-J_\varepsilon}), \tag{4.9}$$

which is truly a bound on the high-frequency energy of $\hat{f}$, that is, above the frequency cut-off $2^{J_\varepsilon}$. Since ridgelets are uniformly bounded in $L_2([0,1]^d)$, we may as well take the upper bound to be 1 so that $\sigma_i \le 1$ for any $i \in \mathcal{I}$. Finally, inequality (3.11) together with (4.9) gives

$$E\|\widehat{\theta} - \theta\|^2_{\ell_2(\mathcal{I})} \le C\big[1 + 2\log(2^{J_\varepsilon d})\big]\big[\varepsilon^2 + (\varepsilon^2)^{2s/(2s+d)}\big] + C2^{-2J_\varepsilon \min(1/2,s)}.$$

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

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

Finally, inequality (3.5) linking $E\|\widehat{\theta} - \theta\|^2_{\ell_2(\mathcal{I})}$ and $E\|\widehat{f} - f\|_2^2$ completes the proof of Theorem 4.4. □

4.3. *Why does this work?* It is beyond the scope of this paper to argue about the claim that ridgelets provide optimally sparse representations of linear singularities which is the content of (4.8). As a compromise, we now give an idea of the reason why the ridgelet coefficient sequence of the two-dimensional mutilated Gaussian (4.1) decays nearly exponentially. For simplicity, consider the centered and vertically mutilated Gaussian

$$f(x_1, x_2) = \mathbb{1}_{\{x_1>0\}} e^{-|x|^2/2}.$$

In two dimensions ridgelets take the form

$$\psi_{j,l,k}(x_1,x_2) = 2^{j/2}\psi\big(2^j(\cos\theta_{j,\ell}x_1 + \sin\theta_{j,\ell}x_2) - k\big),$$

where at scale $j$, the angular discretization step is of the order of $2^{-j}$, say $\theta_{j,\ell} = \alpha \cdot \ell \cdot 2^{-j}$, $\ell = 0, 1, 2, \dots, L$.

1. *Angular localization.* All the coefficients corresponding to ridgelets whose orientation differ from the singular orientation by a multiple of $2^{-j}$ are negligible.
2. *Spatial localization.* For each singular ridgelet orientation, that is, such that $d(\theta_{j,\ell}, \{0,\pi\})$ is less than a multiple of $2^{-j}$, the number of nonnegligible coefficients is of the order of $O(1)$.

In short, there are only $O(1)$ orientations and $O(1)$ locations per orientation that can possibly contribute nonnegligible coefficients. Altogether, there are only $O(1)$ nonnegligible coefficients per ridgelet scale.

The angular localization is perhaps best understood in Fourier space. Let $\hat{f}$ (resp. $\hat{\psi}$) be the Fourier transform of $f$ (resp. of the profile $\psi$ of a ridgelet). We have

$$\int f(x_1,x_2)\psi_{j,l,k}(x_1,x_2)\,dx_1\,dx_2$$
$$= (1/2\pi)^2 \int \hat{f}(\lambda\cos\theta_{j,\ell}, \lambda\sin\theta_{j,\ell})\overline{\hat{\psi}_{j,k}(\lambda)}\,d\lambda,$$

where $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$ so that $\psi_{j,l,k}(x) = \psi_{j,k}(\theta_{j,\ell}^T x)$. The point is that in Fourier space, $\hat{f}$ has very little energy along lines $(\lambda\cos\theta, \lambda\sin\theta)$ when $\theta$ is not pointing in the direction orthogonal to that of the singularity, namely, $\theta = 0$. Indeed, the Fourier transform obeys

$$|\hat{f}(\lambda\cos\theta, \lambda\sin\theta)| \le C\frac{e^{-\lambda^2\sin^2\theta}}{1+|\lambda|}$$

for some universal constant $C$. In frequency, $\hat{\psi}_{j,k}(\lambda)$ is localized near the dyadic subband $2^j \le \lambda \le 2^{j+1}$ and in that frequency range $\hat{f}$ obeys

$$|\hat{f}(\lambda\cos\theta_{j,\ell}, \lambda\sin\theta_{j,\ell})| \le C \cdot 2^{-j} \cdot e^{-2^{2j}\sin^2\theta_{j,\ell}}.$$

Since $\theta_{j,\ell} = \alpha \cdot \ell \cdot 2^{-j}$ this gives

$$|\hat{f}(\lambda\cos\theta_{j,\ell}, \lambda\sin\theta_{j,\ell})| \le C \cdot 2^{-j} \cdot e^{-\gamma\ell^2}$$

for some $\gamma > 0$. Hence, the Fourier transform decays exponentially yielding exponentially decaying coefficients.

We now turn our attention to the spatial localization. Consider a ridgelet orientation parallel to the singularity. Transverse to the ridge, a ridgelet is a wavelet with fine localization properties. For instance, suppose that the profile $\psi$ is of compact support. Then there is only a finite number of ridgelets whose support overlaps with the singularity. That is, all but a finite number of ridgelets do not feel the singularity and yield negligible coefficients as they basically analyze an object which is infinitely many times differentiable.

Now, ridgelets coefficients at scale $j$ obey

$$|\langle f, \psi_{j,l,k}\rangle| \le C \cdot 2^{-j/2}, \tag{4.10}$$

uniformly over all possible orientations and locations. This follows from

$$|\langle f, \psi_{j,l,k}\rangle| \le \int e^{-|x|^2/2}|\psi_{j,k}(\theta_{j,\ell}^T x)|\,dx = \int e^{-|x|^2/2}|\psi_{j,k}(x_1)|\,dx,$$

which holds because of spherical symmetry. The inequality

$$\int e^{-x_1^2/2}|\psi_{j,k}(x_1)|\,dx_1 \le \int e^{-x_1^2/2}|\psi_{j,k}(x_1)|\,dx_1 \le C \cdot 2^{-j/2},$$

where the constant only depends upon $\psi$, establishes (4.10). Hence, at scale $j$, there are only $O(1)$ coefficients of order $2^{-j/2}$.

This is of course a very loose description and a complete argument must accurately quantify the size of those so-called negligible coefficients and is very involved. In effect, the coefficient sequence does not exactly decay with an exponential rate but rather faster than any polynomial rate. At the heart of the analysis lies a geometrical argument describing a subtle balancing between the following two different phenomena: on the one hand, the Fourier transform decays away from the singular direction which tends to yield smaller coefficients; on the other, the coefficients become less sparse as increasingly many ridgelets "feel" the singularity as their orientation moves away from that of the singularity.

4.4. *Extensions*: *several singularities.* There are obvious extensions to Theorem 4.4. For instance, one could take finite superpositions of elements from our class of templates $\mathcal{F}(C)$ (4.4). 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 are unchanged, namely,

$$E\|\hat{f} - f\|_2^2 \leq C \log(\varepsilon^{-1})\, (\varepsilon^2)^{2s/(2s+d)}.$$

Again, we do not need to know how many of these hyperplanes there are or where they are.

Going toward 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 $f(x) = \mathbb{1}_{\{u^T x - b \geq 0\}}\, g(x)$ with $g \in W_2^s$.

DEFINITION 4.5. Let $\mathcal{S}_{\mathrm{H}}$ be the class of functions defined by

$$\mathcal{S}_{\mathrm{H}} = \Big\{ f = \sum a_i f_i \sum |a_i| \leq 1,\ \|f_i\|_{W_2^{(d+1)/2}(\mathbb{R}^2 \backslash L_i)} \leq C \Big\}. \tag{4.11}$$

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 4.6. *The ridgelet thresholding estimate* $\hat{f}$ (3.6)–(3.9) *is asymptotically nearly minimax over our model* $\mathcal{S}_{\mathrm{H}}$. *We have*

$$\sup_{\mathcal{S}_{\mathrm{H}}} MSE(\hat{f}, f) \leq C\big(1 + 2\log(\varepsilon^{-1})\big)(\varepsilon^2)^{(d+1)/(2d+1)}. \tag{4.12}$$

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 it were $(d+1)/2$ times differentiable (in an $L_2$ sense).

PROOF OF THEOREM 4.6. We first show that the sum of the absolute values of the ridgelet coefficients $\theta_i$ of any $f \in \mathcal{S}_{\mathrm{H}}$ is bounded as follows:

$$\sup_j 2^{j/2} \sum_{i: j(i) = j} |\theta_i| \leq C. \tag{4.13}$$

By convexity, it suffices to show (4.13) for $f$ of the form $f = f_0 + \mathbb{1}_{\{u^T x - b \geq 0\}} f_1$, a fact established in Candès (2001). In turn, this property implies that the ridgelet sequence is in $w\ell_p$ for $1/p = 1 + 1/(2d)$, or equivalently that $\sum_i \min(\varepsilon^2, \theta_i^2) \le C(\varepsilon^2)^{(d+1)/(2d+1)}$. The rest of the argument is now similar to that of Theorem 4.4.

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

**5. 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, Theorems 4.4 and 4.6. 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.

5.1. *New notion of smoothness.* Candès (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

$$\|f\|_{\dot{R}^s_{p,q}} \equiv \left(\int \left[\int |\mathcal{R}_f(a,u,b)|^p \, db\, du\right]^{q/p} \frac{da}{a^{q(s+d/2)+1}}\right)^{1/q} < \infty, \tag{5.1}$$

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]

$$Rf(u,t) = \int_{u^T x = t} f(x)\, dx. \tag{5.2}$$

The quantity (5.1) has a natural interpretation in terms of the smoothness of the Radon transform. Indeed, for $p = q$, we have the following equivalence:

$$\|f\|^p_{\dot{R}^s_{p,p}} \asymp \operatorname*{Ave}_u \|Rf(u,\cdot)\|^p_{\dot{B}^{s+(d-1)/2}_{p,p}}, \tag{5.3}$$

where $\dot{B}^{s+(d-1)/2}_{p,p}$ stands for the usual one-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

$$f(x) = \mathbb{1}_{\{x_1>0\}}(2\pi)^{-d/2}e^{-|x|^2/2}. \tag{5.4}$$

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 $\Phi$ 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 $(t, \theta)$ 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 $s < (d + 1)/2$ [Candès (1998)].

Indeed, typical elements of $R^s_{p,q}$ (at least when $p < 2$) look like our mutilated Gaussian (5.4), in that they exhibit the same kind of spatial inhomogeneities. For instance, the class $\mathcal{S}_{\mathrm{H}}$ of mutilated functions that we defined in Section 4 almost corresponds to one of these spaces. Indeed, we have

$$R^{(d+1)/2}_{1,1}(C_1) \subset \mathcal{S}_{\mathrm{H}} \subset R^{(d+1)/2}_{1,\infty}(C_2), \tag{5.5}$$

which means that membership in $\mathcal{S}_{\mathrm{H}}$ is roughly equivalent to membership in $R^{(d+1)/2}_{1,q}$ $(1 \le q \le \infty)$. 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 (5.1) or (5.3). The remainder of this section is devoted to a precise formulation of these heuristics.

5.2. *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 (5.1) is bounded by a fixed constant $C$. We have the following result:

THEOREM 5.1. *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,*

$$\mathcal{R}_\varepsilon(R^s_{p,q}(C)) \ge K(\varepsilon^2)^{2s/(2s+d)}, \tag{5.6}$$

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

(ii) *A simple thresholding estimator* (3.6)–(3.9) *achieves the optimal rate within a log-like factor, that is*,

$$\sup_{f \in R^s_{p,q}(C)} E\|\widehat{f} - f\|_2^2 \le K' \log(\varepsilon^{-1})(\varepsilon^2)^{2s/(2s+d)}, \tag{5.7}$$

*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.

5.3. *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 $(g_\ell)_{1\le\ell\le m}$ with $\|g_\ell\|_{L_2} = \delta$ such that

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

that is, by taking a functional analysis viewpoint, one can find a cube of sidelength $\delta$ 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 $\mathcal{H}$). We then consider the minimax risk $\mathcal{R}_\varepsilon(\mathcal{H})$ of this specific subproblem which turns out to be easily calculated as it is a direct consequence of Assouad's lemma.

LEMMA 5.2. *Let $\mathcal{H}(\varepsilon, \{g_\ell\})$ be the orthogonal hypercube of dimension $m$ and sidelength $\varepsilon$ defined as above ($\delta = \varepsilon$). Then*

$$\mathcal{R}_\varepsilon(\mathcal{H}) \ge \tfrac{1}{4}\Phi(-1/2)/4m\varepsilon^2, \tag{5.8}$$

*where $\Phi$ 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 5.2. 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 $g_\ell$'s; this fact follows from the simple following observation: by letting $P$ be the orthogonal projector onto that span, for any estimator we have

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

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

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

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

$$\tilde{y}_\ell = y_\ell / \varepsilon^2 = \xi_\ell + z_\ell, \tag{5.9}$$

and where, of course, $z_\ell \overset{\text{i.i.d.}}{\sim} N(0,1)$. Observe now that for an estimator of the form $\widehat{f} = \sum_\ell \widehat{\xi}_\ell g_\ell$, we have $\|\widehat{f} - f\|_2^2 = \varepsilon^2 \sum_\ell (\widehat{\xi}_\ell - \xi_\ell)^2$. Then, a rescaling argument gives that the minimax mean-squared error $\mathcal{R}_\varepsilon(\mathcal{H})$ equals $\varepsilon^2$ times the minimax mean-squared error of the problem (5.9), that is,

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

The latter problem (5.9) is now classical and a lower bound for its minimax mean-squared error is $\Phi(-1/2)\,m$. It is interesting to note that (5.9) has a strong flavor of a 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 $\mathcal{H}(\varepsilon, m(\varepsilon))$ yielding a sharp asymptotic lower bound. The lower bound (5.6) follows from the technical lemma:

LEMMA 5.3. *For any $\delta > 0$, there exists a hypercube $\mathcal{H}(\delta, \{g_\ell\}) \subset R^s_{p,q}(C)$ of sidelength $\delta$ and dimension $m(\delta) \ge K\delta^{-1/(s/d+1/2)}$.*

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 complete the proof of the first part of Theorem 5.1:

COROLLARY 5.4. *We have a lower bound on the minimax risk,*

$$\mathcal{R}_\varepsilon\big(R^s_{p,q}(C)\big) \ge c(\varepsilon^2)^{2s/(2s+d)}. \tag{5.10}$$

PROOF. We clearly have

$$\mathcal{R}_\varepsilon\big(R^s_{p,q}(C)\big) \ge \mathcal{R}_\varepsilon(\mathcal{H}) \ge \tfrac{1}{4}\Phi(-1/2)/4m(\varepsilon)\varepsilon^2,$$

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

5.4. *Upper bounds.* The proof of the upper bound closely follows the concepts presented in Section 3. For convenience, let us take exactly the same estimator as the one introduced at the beginning of Section 4, that is,

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

(see Section 4.2 for the value of the parameter $\lambda$).

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,

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

The upper bound will result from the following fact that is proven in Candès (1998): for any function $f \in \mathbb{R}^s_{p,q}(C)$, letting $\sigma = d(1/p - 1/2)$ we have

$$\|\theta\|_{\mathbf{r}^s_{p,q}} := \sum_i \left(\sum_{j\ge 0}\left(2^{j\sigma} \sum_{j(i)=j} |\theta_i(f)|^p\right)^{q/p}\right)^{1/q} \le C\|f\|_{R^s_{p,q}}. \tag{5.11}$$

Formally, $\|\theta\|_{\mathbf{r}^s_{p,q}}$ has the same structure as a discrete Besov norm, except that the sequence $\theta$ 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 $\varepsilon > 0$, we have

$$\sum_i \min(\varepsilon^2, \theta_i^2) \le \varepsilon^{2s/(2s+d)}; \tag{5.12}$$

and second,

$$\sum_{j(i)>J_\varepsilon} \theta_i^2 \le C 2^{-2J_\varepsilon s'/d}, \tag{5.13}$$

with

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

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

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

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

Finally, an application of the oracle inequality (3.11) together with (5.13) gives

$$E\|\widehat{\theta}-\theta\|^2_{\ell_2(\mathcal{I})} \le C[1+2\log(2^{J_\varepsilon d})][\varepsilon^2+(\varepsilon^2)^{2s/(2s+d)}]+C2^{-2J_\varepsilon s'}.$$

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

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

and, finally, inequality (3.5) 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 5.1 is complete.

We would like to close this section 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 $\eta_{\mathrm{ST}}$ with

$$\eta_{\mathrm{HT}}(y,\lambda) = y\mathbb{1}_{\{|y|\ge\lambda\}}, \tag{5.14}$$

would give exactly the same asymptotic performance.

5.5. *Adapting to the unknown degree of smoothness.* A special 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

$$\sup_{f\in\mathcal{F}_\nu(C)} E\|\hat{f}-f\|_2^2 \le K(\nu)C\log(\varepsilon^{-1})\,(\varepsilon^2)^{2s/(2s+d)}.$$

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

$$\sup_{\mathcal{F}_\nu(C)} E\|\hat{f}-f\|_2^2 \le K_0 C\log(\varepsilon)(\varepsilon^2)^{2s/(2s+d)} \qquad \forall\, \nu\in\mathcal{V}_0, \tag{5.15}$$

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

**6. Curved singularities.** As in Chapter 6 of Candès (1998), 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 4 by studying a simple example exhibiting a general phenomenon.

Let $f$ be the radial function defined, as follows:

$$f(x) = \mathbb{1}_{\{|x| \le 1/2\}} e^{-|x|^2/2}, \tag{6.1}$$

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 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, that is,

$$MSE(\widehat{f}_{\mathrm{KS}}, f) \ge C(\varepsilon^2)^{1/(d+1)}$$

for a linear smoother with any bandwidth, and

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

for any reasonable wavelet thresholding estimate.

The lower bound (3.10) introduced in Section 3 gives, in turn, a lower bound on the risk of a ridgelet thresholding estimator,

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

The reason for this slow convergence is, of course, that the ridgelet transform of (6.1) is not sparse. Candès [(1998), Chapter 6] proves that

$$\sum_i \min(\theta_i^2, \varepsilon^2) \ge C(\varepsilon^2)^{1/d}, \tag{6.2}$$

which supports the claim as discussed in an earlier section. We find this result somewhat unexpected.

First, it is uncommon 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 that existing between projection pursuit regression and nonlinear kernel regression with an adaptive choice of bandwidth: the nonlinear 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 (6.1), meaning that one would need only a small number of ridge functions to approximate the true regression surface. The problem is that objects with curved structure like (6.1) are not well approximated by ridge functions. Preliminary results about this heuristic may be found in Candès (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 $(\varepsilon^2)^{2s/(2s+1)}$.

## 7. Numerical experiments.

7.1. *A digital ridgelet transform.* Recent work developed an approximate digital implementation of the two-dimensional ridgelet transform. At the present stage, the algorithm takes data on a two-dimensional Cartesian grid and computes approximate ridgelet coefficients. Although the details of the algorithm have not been yet published, we will now give an outline of the algorithm we used for our numerical experiments and refer the reader to Donoho (1998) for an accurate description of this digital transform.

The starting point is the observation that the ridgelet transform is precisely the application of a one-dimensional wavelet transform to the slices of the Radon transform (5.2) where the variable $u$ is held constant and $t$ is varying. Mathematicaly speaking, the ridgelet coefficient (2.4) can be expressed as

$$\mathcal{R}_f(a,u,b) = \int Rf(u,t)\, a^{-1/2}\psi\left(\frac{t-b}{a}\right) dt. \tag{7.1}$$

A natural strategy for a digital ridgelet transform then consists of:

(i) developing a numerical Radon transform, and
(ii) applying a one-dimensional numerical wavelet transform.

The first step is delicate whereas the second is by now absolutely standard.

Our numerical evaluation of the Radon transform relies on the celebrated projection-slice theorem [Deans (1983)] which states that

$$\hat{f}(\lambda u_1, \dots, \lambda u_d) = \int Rf(t,u) e^{-i\lambda t}\, dt, \qquad u = (u_1, \dots, u_d).$$

Therefore, the Radon transform may be obtained by applying the one-dimensional inverse Fourier transform to the two-dimensional Fourier transform restricted to radial lines going through the origin. The idea behind the digital Radon transform is then to compute sampled values of the Fourier transform on a polar lattice, that is, on a lattice where the points lie on radial lines. This suggests deploying the following three-step procedure for calculating a two-dimensional discrete Radon transform, say, from gridded data $(f(i_1, i_2))$, $0 \le i_1, i_2 < n$:

1. *2D-FFT*. Compute the two-dimensional FFT of $f$ giving the array $(\hat{f}(k_1, k_2))$, $-n/2 \le k_1, k_2 \le n/2 - 1$.
2. *Cartesian to polar conversion*. Using an interpolation scheme, substitute the sampled values of the Fourier transform obtained on the square lattice with sampled values of $\hat{f}$ on a polar lattice.
3. *1D-IFFT*. Compute the one-dimensional IFFT on each line, that is, for each value of the angular parameter.

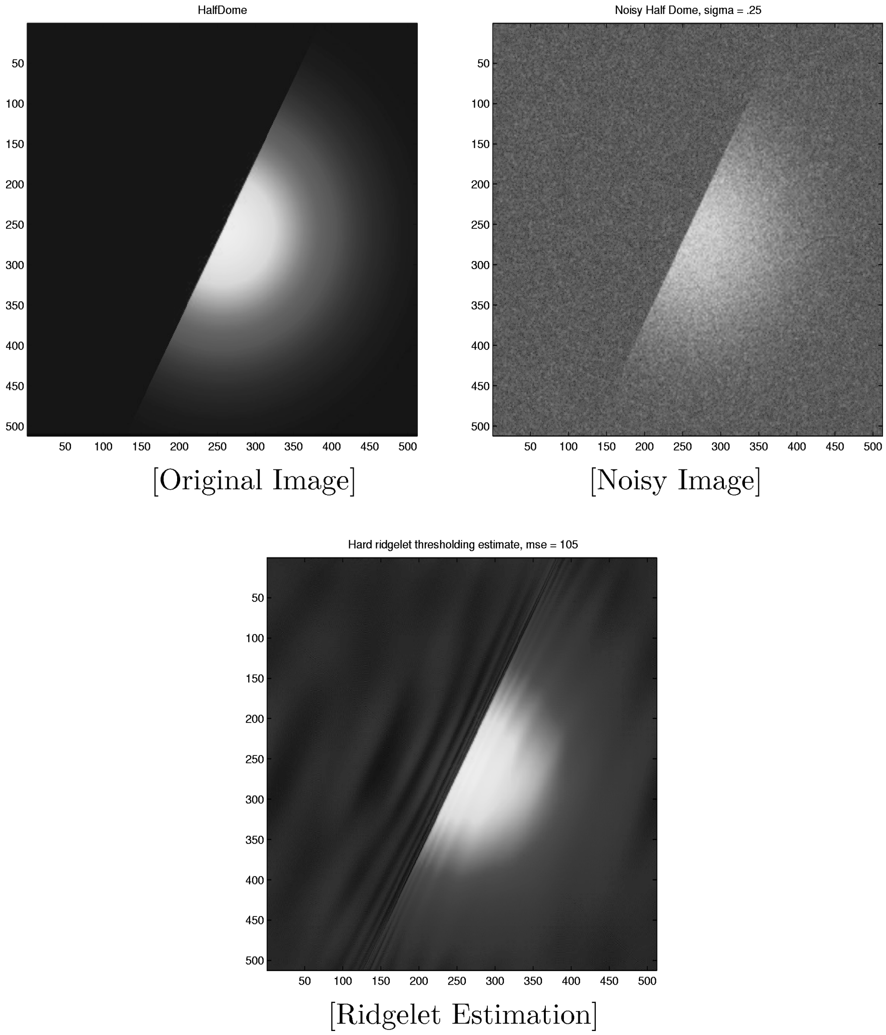
The $d$-dimensional version of the algorithm would be exactly similar. The *Cartesian to polar conversion* is at the heart of the matter and beyond the scope of this paper. Donoho (1998) and Starck, Candès and Donoho (2002) explore possible strategies. In these experiments, we used an interpolation scheme which is exact whenever the data $(f(i_1, i_2))_{0 \le i_1, i_2 < n}$ are sampled from a two-dimensional trigonometric polynomial of degree $n$.

We would like to close this section by listing a few properties of our digital ridgelet transform.

- The transform is not orthonormal but numerically tight; it expands an $n \times n$ digital array into an $n \times 2n$ array of coefficients.
- The transform has low complexity and runs with a number of operations of the order of $N \log N$ for an image of size $N = n^2$.
- There is an associated approximate inverse transform that reconstructs an image from the data of its discrete digital ridgelet coefficients. The inverse and the forward transform have the same order of complexity.

Last but not least, Donoho is to be credited for the major part of the work described in this section.

**8. 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 coefficients. This approach is very concrete and amenable to rigorous theory and bears great potential for applications; for example, ridgelets are making their way into image processing. 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



[Original Image]

[Noisy Image]

[Ridgelet Estimation]

FIG. 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.*

coefficients. We have identified many situations where the ridgelet shrinkage is optimal and, in addition, we have also been able to study its limitations.

8.1. *Connection with sigmoidal neural networks.* At this point, the connection between ridgelet thresholding and classical sigmoidal neural networks may seem loose although the philosophy is the same, namely, that of approximation by superposition of ridge functions. Quantitatively speaking, it may be perfectly reasonable to think that traditional neural networks enjoy superior approximation performance over the naive ridgelet thresholding approach. This paper, however, is part of a large body of work and other results suggesting that this does not happen. Ignoring boundary issues, another paper [Candès (2002)] claims that *there is no function which is approximated at a faster rate, in an asymptotic sense, with sigmoidal feedforward neural networks than with naive ridgelet thresholding.* First, this result suggests that ridgelets are a viable substitute for sigmoidal feedforward neural networks, at least theoretically. Second, it says that ridgelet analysis is probably the right tool for studying ridge function approximation and thereby legitimates the definition and claims about ridge spaces introduced in Section 5.

Further, it is now well understood that improved approximation properties usually translate into improved estimation bounds; see Donoho and Johnstone (1989) and Barron (1991), for instance. Therefore, it is very unlikely that methods based on possibly nonconstructive neural network complexity penalized fitting procedures with better estimation rates (in an asymptotic sense) than those obtained by naive ridgelet thresholding would exist.

8.2. *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 with discrete data; and, second, the implicit assumption is that the setting is in some sense uniform as the performance is evaluated with respect to 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

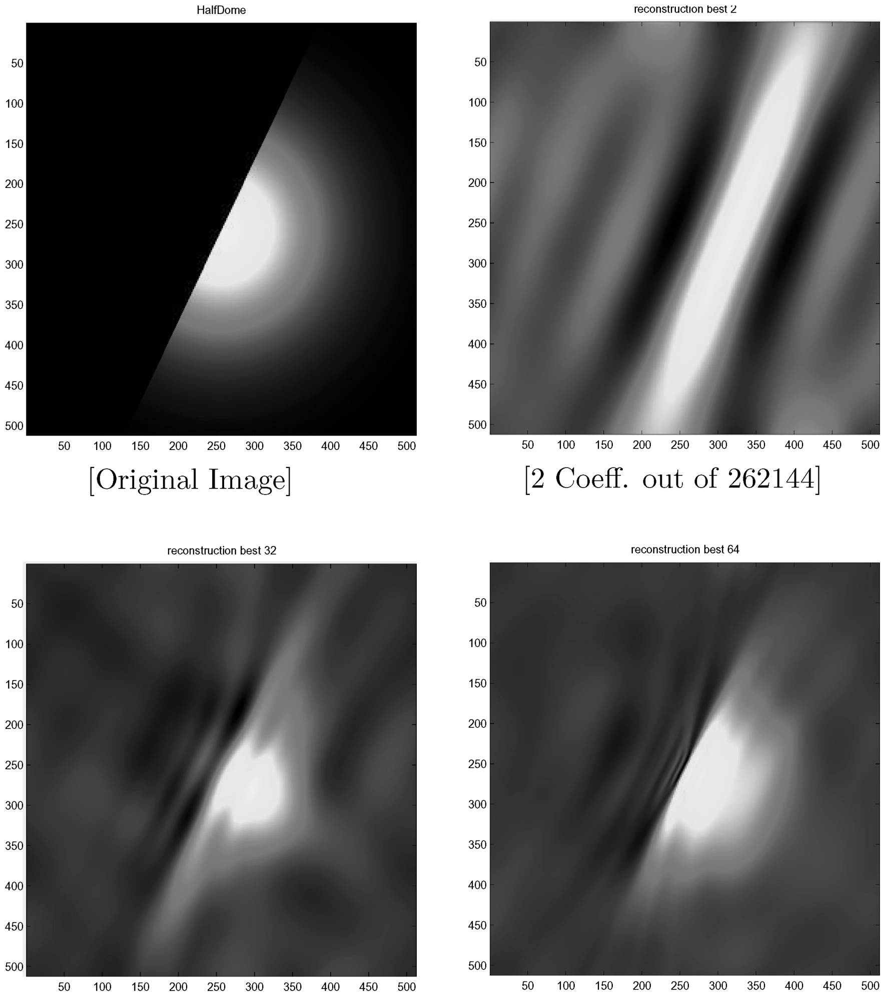
$$y_{i,j} = \tilde{f}(i,j) + \sigma z_{i,j},$$

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

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

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), e.g.]. However, the author is confident that a careful analysis will give discrete versions of Theorems 4.4 and 5.1. 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.

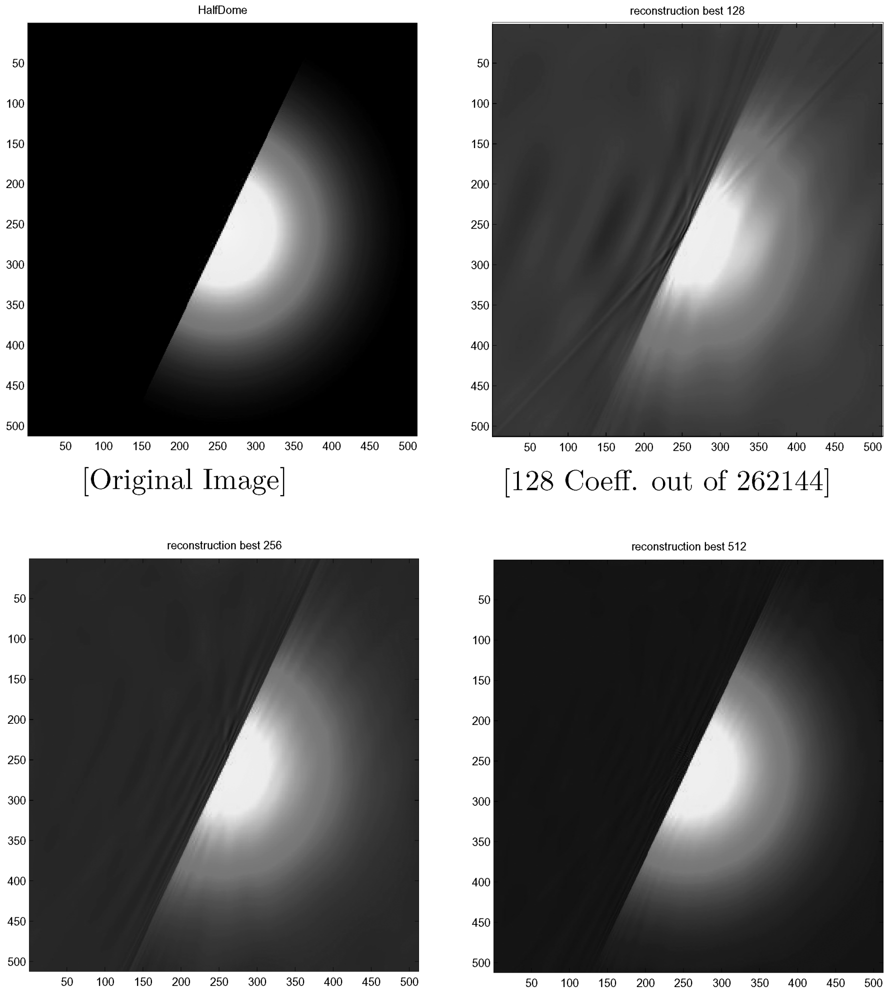
*Regular setting.* Even though one may expect to see ridgelet algorithms enjoy nice estimation bounds with data given on a regular grid, there does not seem to be a quick answer to the problem of dealing with irregularly spaced (heterogeneous) data points. This is indeed a fairly classical problem that 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)].



[Original Image] [2 Coeff. out of 262144]

[32 Coeff. out of 262144] [64 Coeff. out of 262144]

FIG. 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]

FIG. 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.*

As we can see, the issues that we raised are shared by many popular methods in current use and are far from being a distinguished feature 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.

8.3. *Curved edges.* Finally, ridgelets are optimal for estimating objects with singularities across hyperplanes (Section 4), but they fail to estimate efficiently objects with singularities across curved hypersurfaces (Section 6).

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 the 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 Candès (1999b) and is shown to provide better estimation bounds than (6.2).

A more promising approach is based on a new transform, namely, the curvelet transform pioneered by Candès and Donoho (2000). 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 refinements are grounded on the work presented in this paper.

8.4. *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.

## APPENDIX

In this appendix we will give rigorous proofs of some hypercube embedding results (Lemmas 4.3 and 5.3) needed to support the claims about lower rates of convergence (Sections 4 and 5). Lemma 5.3 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 $\varepsilon > 0$, how can we distribute points on the sphere such that balls of radius $\varepsilon$ and centered at these points do not 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 normalizing 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,

$$f(v) = c_d'(1 + v^2)^{-d/2}. \tag{A.1}$$

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 $\varepsilon_j = 2^{-(j-j_0)}$ and let $S_j$ be a set of points on the sphere $(u_\ell)$ satisfying the following properties:

(i) $\forall u_\ell, u_{\ell'} \in S_j$, $\|u_\ell \pm u_{\ell'}\| \geq \varepsilon_j$,

(ii) $B_1 \varepsilon_j^{-(d-1)} \leq |S_j| \leq B_2 \varepsilon_j^{-(d-1)}$,

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

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

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

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 $u_\ell$ and $u_{\ell'}$. We remark that the first property implies that

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

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

To further simplify the analysis, suppose $\psi \in \mathcal{S}(\mathbb{R})$ is compactly supported, $\operatorname{supp} \psi \subset [-1/2, 1/2]$, and has a sufficiently large number of vanishing moments. We normalize $\psi$ 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 $j$:

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

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 $f_{\ell,k} \subset \{x,\ |u_\ell \cdot x| \leq \sqrt{2}/2\}$ 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 A.1. *The cardinality of $A_j$ is bounded below by*

$$\# A_j \geq C 2^{jd}.$$

*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.*

$$\forall f \in A_j, \qquad \|f\|_2 \geq c_d; \tag{A.3}$$

(ii) *and if $j_0$ is chosen large enough*,

$$\forall f \in A_j, \qquad \sum_{g \in A_j, g \neq f} |\langle f, g\rangle| \leq \frac{c_d}{2}. \tag{A.4}$$

PROOF. 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

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

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

$$\begin{aligned}
&\sum_{k \in \mathbb{Z}} (1 + |x - \varepsilon k|)^{-1-\delta} (1 + |y - \eta k|)^{-1-\delta} \\
&\qquad \leq C_\delta \varepsilon^{-1} (1 + |y - x\eta\varepsilon^{-1}|)^{-1-\delta}.
\end{aligned} \tag{A.5}$$

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 Candès (1998) when $u_\ell \neq u_{\ell'}$ shows that one can find a constant $C_1(d)$ depending on $d$, $\psi$ and $w$ such that

$$\begin{aligned}&|\langle f_{\ell,k}, f_{\ell',k'}\rangle| \\ &\qquad \le C_1(d) 2^{-j(2d+1)}(1+v_{\ell,\ell'}^2)^{(2d+1)/2}\bigl(1+2^{-j}|v_{\ell,\ell'}k-(1+v_{\ell,\ell'}^2)^{1/2}k'|\bigr)^{-2}.\end{aligned}$$

Now, it follows from (A.5) that

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

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

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

where again $C_4(d)$ is a new constant $C(d,\psi,w)$. (Notice that we have sacrificed exactness for synthetic notation: in the second line of the array, read $|\{u_{\ell'}, 0 \le |v_{\ell,\ell'}| \le 1\}|$ instead of $|\{u_{\ell'}, 2^{\ell-1} \le |v_{\ell,\ell'}| \le 2^\ell\}|$ when the index $\ell$ 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 Candès (1998).

LEMMA A.2. *First, the elements $f_{\ell,k}$ satisfy*

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

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

$$\mathcal{C} = \Big\{ f,\ f = \sum_{\ell,k} \xi_{\ell,k} f_{\ell,k},\ |\xi_{\ell,k}| \le 1 \Big\}. \tag{A.6}$$

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

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

*where the constant $C$ depends at most on $s, p, q, \psi, w$ 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 5.3 one needs to construct a cube. The next lemma shows how to orthogonalize our parallelepiped.

LEMMA A.3. *Suppose we have $n$ vectors $\{f_i\}_{1 \le i \le n}$ in a Hilbert space such that for all $1 \le i \le n$*

(i) $\|f_i\| = 1$;
(ii) $\sum_{j \ne i} |\langle f_i, f_j \rangle| \le 1 - \delta < 1$.

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

PROOF. 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 $\delta$. 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 $\{e_i\}_{1 \le i \le n}$ defined by $e_i = H f_i$ is indeed an orthogonal basis of $\operatorname{span}(\{f_i\}_{1 \le i \le n})$ [see Meyer (1992) for a proof]. Furthermore, a trivial fact states that

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

Thus the embedding problem becomes: show that $\|x\|_\infty \le \delta \Rightarrow \|Hx\|_\infty \le 1$. This requires nothing but to prove that the norm of $H$, as an operator from $\ell_\infty \to \ell_\infty$, is bounded by $\delta^{-1}$. Recall

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

We now derive an upper bound 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

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

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

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

which gives

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

Finally,

$$\begin{aligned}
\|H\|_{(\ell_\infty,\ell_\infty)} &\le \frac{1}{\pi}\int_0^\infty \bigl(1-\|F\|_{(\ell_\infty,\ell_\infty)}\bigr)^{-1}(1+\lambda)^{-1}\lambda^{-1/2}\,d\lambda\\
&= \bigl(1-\|F\|_{(\ell_\infty,\ell_\infty)}\bigr)^{-1}.
\end{aligned}$$

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

Lemma 5.3 is now a mere consequence of the three preceding preparatory lemmas.

As far as the linear estimation is concerned, Lemma 4.3 essentially follows from Lemma A.2 and (5.5). Indeed, chasing definitions, the closed convex hull $\overline{\mathrm{Hull}(\mathcal{F})}$ contains $\mathcal{S}_{\mathcal{H}}$ which in turn contains a ball of $R^{(d+1)/2}_{1,1}$. Hence, it is sufficient to prove the appropriate embedding in a ball of $R^{(d+1)/2}_{1,1}$. By Lemma A.2 we have

$$\mathcal{C} = \biggl\{f,\ f = \sum_{\ell,k}\xi_{\ell,k}f_{\ell,k},\ \sum|\xi_{\ell,k}| \le 1\biggr\} \subset \Bigl\{f,\ \|f\|_{R^{(d+1)/2}_{1,1}} \le C2^{j/2}\Bigr\}.$$

We use the same orthogonalization procedure as in Lemma A.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 A.3) such that

$$\mathcal{C}' = \left\{f,\ f = \sum_{\ell,k} \xi_{\ell,k} g_{\ell,k},\ \sum |\xi_{\ell,k}| \le 1\right\} \subset \mathcal{C} \subset \left\{f,\ \|f\|_{R_{1,1}^{(d+1)/2}} \le C2^{j/2}\right\}.$$

The proof of this fact is identical to that of Lemma A.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 $\delta^{-1}$. 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 4.3 (the quantity $2^{-j/2}$ playing the role of $\delta$ in the statement of this lemma).

**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. I would like to thank the anonymous referees for their role in improving the original manuscript.

## REFERENCES

BARRON, A. R. (1991). Complexity regularization with application to artificial neural networks. In *Nonparametric Functional Estimation and Related Topics* (G. Roussas, ed.) 561–576. Kluwer, Dordrecht.

CANDÈS, E. J. (1998). Ridgelets: Theory and applications. Ph.D. dissertation, Dept. Statistics, Stanford Univ.

CANDÈS, E. J. (1999a). Harmonic analysis of neural netwoks. *Appl. Comput. Harmon. Anal.* **6** 197–218.

CANDÈS, E. J. (1999b). Monoscale ridgelets for the representation of images with edges. Technical report, Dept. Statistics, Stanford Univ.

CANDÈS, E. J. (2001). Ridgelets and the representation of mutilated Sobolev functions. *SIAM J. Math. Anal.* **33** 347–368.

CANDÈS, E. J. (2002). New ties between computational harmonic analysis and approximation theory. In *Approximation Theory X* (C. K. Chui, L. L. Schumaker and J. Stöckler, eds.) 87–153. Vanderbilt Univ. Press, Nashville, TN.

CANDÈS, E. J. and DONOHO, D. L. (2000). Curvelets—A surprisingly effective nonadaptive representation of objects with edges. In *Curve and Surface Fitting* (A. Cohen, C. Rabut and L. L. Schumaker, eds.) 105–120. Vanderbilt Univ. Press, Nashville, TN.

CHENG, B. and TITTERINGTON, D. M. (1994). Neural networks: A review from a statistical perspective (with discussion). *Statist. Sci.* **9** 2–54.

CONWAY, J. H. and SLOANE, N. J. A. (1988). *Sphere Packings, Lattices and Groups*. Springer, New York.

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*. Wiley, New York.

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

DONOHO, D. L. (1998). Digital ridgelet transform via rectopolar coordinate transform. Technical report, Dept. Statistics, Stanford Univ.

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. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81** 425–455.

DONOHO, D. L. and JOHNSTONE, I. M. (1995). Empirical atomic decomposition. Unpublished manuscript.

DONOHO, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879–921.

EFROIMOVICH, S. and PINSKER, M. (1982). Estimation of square-integrable density on the basis of a sequence of observations. *Problems Inform. Transmission* **17** 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 Univ. Press.

HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. and TSYBAKOV, A. (1998). *Wavelets, Approximation, and Statistical Applications. Lecture Notes in Statist.* **129**. Springer, New York.

IBRAGIMOV, I. A. and HASMINSKII, R. Z. (1981). *Statistical Estimation. Asymptotic Theory*. Springer, New York.

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

JOHNSTONE, I. M. and SILVERMAN, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *J. Roy. Statist. Soc. Ser. B* **59** 319–351.

JONES, L. K. (1997). The computational intractability of training sigmoidal neural networks. *IEEE Trans. Inform. Theory* **43** 167–173.

KOROSTELEV, A. P. and TSYBAKOV, A. B. (1993). *Minimax Theory of Image Reconstruction. Lecture Notes in Statist*. **82**. Springer, New York.

MEYER, Y. (1992). *Wavelets and Operators*. Cambridge Univ. Press.

PINSKER, M. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problems Inform. Transm.* **16** 120–133.

SILVERMAN, B. (1999). Wavelets in statistics: Beyond the standard assumptions. *R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci.* **357** 2459–2473.

STARCK, J., CANDÈS, E. and DONOHO, D. (2002). The curvelet transform for image denoising. *IEEE Trans. Image Process.* **11** 670–684.

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 Trans. Inform. Theory* **44** 2892–2900.

DEPARTMENT OF APPLIED
AND COMPUTATIONAL MATHEMATICS
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA 91125
E-MAIL: emmanuel@acm.caltech.edu