Submitted to the Annals of Statistics

SPATIALLY-ADAPTIVE SENSING IN NONPARAMETRIC REGRESSION

By Adam D. Bull

Statistical Laboratory, University of Cambridge

While adaptive sensing has provided improved rates of convergence in sparse regression and classification, results in nonparametric regression have so far been restricted to quite specific classes of functions. In this paper, we describe an adaptive-sensing algorithm which is applicable to general nonparametric-regression problems. The algorithm is spatially-adaptive, and achieves improved rates of convergence over spatially-inhomogeneous functions. Over standard function classes, it likewise retains the spatial adaptivity properties of a uniform design.

1. Introduction. In many statistical problems, such as classification and regression, we observe data $Y_1, Y_2, \ldots$, where the distribution of each Y_n depends on a choice of design point x_n. Typically, we assume the x_n are fixed in advance. In practice, however, it is often possible to choose the design points sequentially, letting each x_n be a function of the previous observations $Y_1, \ldots, Y_{n-1}$.

We will describe such procedures as *adaptive sensing*, but they are also known by many other names, including sequential design, adaptive sampling, active learning, and combinations thereof. The field of adaptive sensing has seen much recent interest in the literature: compared with a fixed design, adaptive sensing algorithms have been shown to provide improvements in sparse regression (Iwen, 2009; Haupt, Castro and Nowak, 2011; Malloy and Nowak, 2011b; Boufounos et al., 2012; Davenport and Arias-Castro, 2012) and classification (Cohn, Atlas and Ladner, 1994; Castro and Nowak, 2008; Beygelzimer, Dasgupta and Langford, 2009; Koltchinskii, 2010; Hanneke, 2011). Recent results have also focused on the limits of adaptive sensing (Arias-Castro, Candes and Davenport, 2011; Malloy and Nowak, 2011a; Castro, 2012).

In this paper, we will consider the problem of nonparametric regression, where we aim to estimate an unknown function $f : [0, 1] \to \mathbb{R}$ from observa-

AMS 2000 subject classifications: Primary 62G08; secondary 62L05, 62G20

Keywords and phrases: Nonparametric regression, adaptive sensing, sequential design, active learning, spatial adaptation, spatially-inhomogeneous functions

tions

$$Y_n \coloneqq f(x_n) + \varepsilon_n, \quad \varepsilon_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2).$$

While previous authors have also considered this model under adaptive sensing, their results have either been restricted to quite specific classes of functions f, or have not provided improved rates of convergence (Faraway, 1990; Cohn, Ghahramani and Jordan, 1996; Hall and Molchanov, 2003; Castro, Willett and Nowak, 2006; Efromovich, 2008).

In the following, we will describe a new algorithm for adaptive sensing in nonparametric regression. Our algorithm will be based on standard wavelet techniques, but with an adaptive choice of design points: we will aim to codify, in a meaningful way, the intuition that we should place more design points in regions where f is hard to estimate.

While many such heuristics are possible, we would like to construct an algorithm with good theoretical justification; in particular, we will be interested in attaining improved rates of convergence. In general, however, it is known that in nonparametric regression, adaptive sensing cannot provide improved rates over standard classes of functions. Castro, Willett and Nowak (2006) prove such a result for L^2 loss; we will show the same is true locally uniformly.

In the following, we will argue that the fault here lies not with adaptive sensing, but rather with the functions considered. In the field of *spatial adaptation*, unknown functions are often assumed to be *spatially inhomogeneous*: they may be rougher, and thus harder to estimate, in some regions of space than in others. The seminal paper of Donoho and Johnstone (1994) provides examples, which we have reproduced in Figure 1; these mimic the kinds functions observed in imaging, spectroscopy and other signal processing problems.

Previous work in this field has provided many fixed-design estimators with good performance over such functions (Donoho et al., 1995; Fan and Gijbels, 1995; Lepski, Mammen and Spokoiny, 1997; Donoho and Johnstone, 1998; Fan et al., 1999). With adaptive sensing, however, we can obtain further improvements: if we place more design points in regions where f is rough, our estimates $\hat{f}_n$ will become more accurate overall.

To quantify this, we will need to introduce new classes describing spatially-inhomogeneous functions, over which our algorithm will be shown to obtain improved rates of convergence. While these classes are novel, they will be shown to contain quasi-all functions from standard classes in the literature. Furthermore, our algorithm will be shown to adaptively obtain near-optimal rates over both the new and standard function classes.

Smoothness classes similar to our own have arisen in the study of adaptive

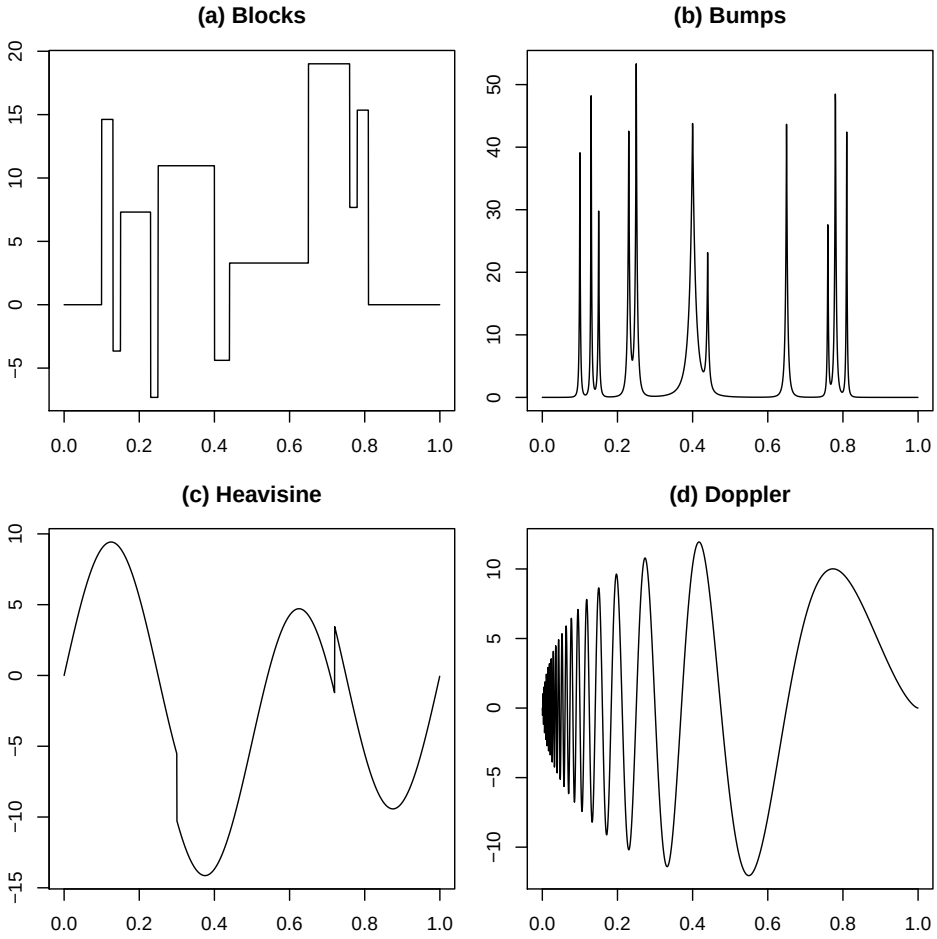

FIG 1. *Examples of spatially-inhomogeneous functions from Donoho and Johnstone (1994). Each function is scaled to have* $\mathrm{sd}(f) = 7$.

nonparametric inference (Picard and Tribouley, 2000; Giné and Nickl, 2010; Bull, 2012a), and more generally also in the study of turbulence (Frisch and Parisi, 1980; Jaffard, 2000). As in those papers, we find that for complex nonparametric problems, the standard smoothness classes may be insufficient to describe behaviour of interest; by specifying our target functions more carefully, we can achieve more powerful results.

We might also compare this phenomenon to results in sparse regression, where good rates are often dependent on specific assumptions about the design matrix or unknown parameters (Fan and Lv, 2008; van de Geer and Bühlmann, 2009; Meinshausen and Bühlmann, 2010). As there, we can use the nature of our assumptions to provide insight into the kinds of problems on which we can expect to perform well.

We will test our algorithm by estimating the functions in Figure 1 under Gaussian noise. We will see that, by sensing adaptively, we can make significant improvements to accuracy; we thus conclude that adaptive sensing can be of value in nonparametric regression whenever the unknown function may be spatially inhomogeneous.

In Section 2, we describe our adaptive-sensing algorithm. In Section 3, we describe our model of spatial inhomogeneity, and show that adaptive sensing can lead to improved performance over such functions. In Section 4, we discuss the implementation of our algorithm, and provide empirical results. Detailed proofs are available in the supplemental article (Bull, 2012b).

2. The adaptive-sensing algorithm.

We now describe our adaptive-sensing algorithm in detail. We first discuss how we estimate f under varying designs; we then move on to the choice of design itself.

2.1. *Estimation under varying designs.*

Given observations $Y_1, \dots, Y_n$ at a set of design points $\xi_n := \{x_1, \dots, x_n\}$, we will estimate the function f using the technique of wavelet thresholding, which is known to give spatially-adaptive estimates (Donoho and Johnstone, 1994). To begin, we will need to choose our wavelet basis; for $j_0 \in \mathbb{N}$, let

$$\varphi_{j,k} \text{ and } \psi_{j,k}, \quad j,k \in \mathbb{Z},\ j \geq j_0,\ 0 \leq k < 2^j,$$

be a compactly-supported wavelet basis of $L^2([0,1])$, such as the construction of Cohen, Daubechies and Vial (1993).

In the following, we will assume the wavelets $\psi_{j,k}$ have $N \in \mathbb{N}$ vanishing moments,

$$\int x^n \psi_{j,k}(x)\, dx = 0, \quad n \in \mathbb{Z},\ 0 \leq n < N,$$

and both $\varphi_{j,k}$ and $\psi_{j,k}$ are zero outside intervals $S_{j,k}$ of width $2^{-j}(2L-1)$,

$$S_{j,k} \coloneqq 2^{-j}[k-L+1, k+L) \cap [0,1).$$

For any $i \in \mathbb{N},\ i \geq j_0$, we may then write an unknown function $f \in L^2([0,1])$ in terms of its wavelet expansion,

$$f = \sum_{k=0}^{2^i-1} \alpha_{i,k}\varphi_{i,k} + \sum_{j=i}^{\infty}\sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k},$$

and estimate f in terms of the coefficients $\alpha_{j_0,k}, \beta_{j,k}$.

When the design is uniform, we can estimate these coefficients efficiently in the standard way, using the fast wavelet transform (Donoho and Johnstone, 1994). Suppose, as will always be the case in the following, that the design points x_n are distinct, so we may denote the observations Y_n as $Y(x_n)$. Given $i \in \mathbb{N},\ i \geq j_0$, suppose also that we have observed f on a grid of design points $2^{-i}k,\ k \in \mathbb{Z},\ 0 \leq k < 2^i$.

We may then estimate the scaling coefficients $\alpha_{i,k}$ of f as

$$\hat{\alpha}^i_{i,k} \coloneqq 2^{-\frac{i}{2}} Y(2^{-i}k),$$

since for i large,

$$\alpha_{i,k} = \int_{S_{i,k}} f(x)\varphi_{i,k}(x)\ dx \approx 2^{-\frac{i}{2}} f(2^{-i}k). \tag{1}$$

By an orthogonal change of basis, we can produce estimates $\hat{\alpha}^i_{j_0,k}$ and $\hat{\beta}^i_{j,k}$ of the coefficients $\alpha_{j_0,k}$ and $\beta_{j,k}$, given by the relationship

$$\sum_{k=0}^{2^{j_0}-1} \hat{\alpha}^i_{j_0,k}\varphi_{j_0,k} + \sum_{j=j_0}^{i-1}\sum_{k=0}^{2^j-1} \hat{\beta}^i_{j,k}\psi_{j,k} \coloneqq \sum_{k=0}^{2^i-1} \hat{\alpha}^i_{i,k}\varphi_{i,k}. \tag{2}$$

These estimates can be computed efficiently by applying the fast wavelet transform to the vector of values $2^{-\frac{i}{2}} Y(2^{-i}k)$.

Since we will be considering non-uniform designs, this situation will often not apply directly. Many approaches to applying wavelets to non-uniform designs have been considered in the literature, including transformations of the data, and design-adapted wavelets (see Kerkyacharian and Picard, 2004, and references therein). In the following, however, we will use a simple method, which allows us to simultaneously control the accuracy of our estimates for many different choices of design.

To proceed, we note that the value of an estimated coefficient $\hat{\alpha}^i_{j,k}$ or $\hat{\beta}^i_{j,k}$ depends only on observations $Y(x)$ at points $x \in S_{j,k} \cap 2^{-i}\mathbb{Z}$. We may therefore estimate the wavelet coefficients $\alpha_{j_0,k}$ and $\beta_{j,k}$ by

$$\hat{\alpha}_{j_0,k} \coloneqq \hat{\alpha}^{i_n(j_0,k)}_{j_0,k} \qquad \text{and} \qquad \hat{\beta}_{j,k} \coloneqq \hat{\beta}^{i_n(j,k)}_{j,k}, \tag{3}$$

where the indices $i_n(j,k)$ are chosen so that these estimates use as many observations as possible,

$$i_n(j,k) \coloneqq \max\left\{i \in \mathbb{N} : i > j,\ S_{j,k} \cap 2^{-i}\mathbb{Z} \subseteq \xi_n\right\}. \tag{4}$$

To ease notation, for now we will estimate coefficients only up to a maximum resolution level $j^{\max}_n \in \mathbb{N}$, $j^{\max}_n > j_0$, chosen so that $2^{j^{\max}_n} \sim n/\log(n)$. We will then be able to guarantee that the set in (4) is non-empty.

Using these estimates directly will lead to a consistent estimate of f, but one converging very slowly; to obtain a spatially-adaptive estimate, we must use thresholding. We fix $\kappa > 1$, and for

$$e_n(j,k) \coloneqq \sigma 2^{-\frac{1}{2}i_n(j,k)}\sqrt{2\log(n)}, \tag{5}$$

define the hard-threshold estimates

$$\hat{\beta}^T_{j,k} \coloneqq \begin{cases} \hat{\beta}_{j,k}, & |\hat{\beta}_{j,k}| \geq \kappa e_n(j,k), \\ 0, & \text{otherwise.} \end{cases}$$

We then estimate f by

$$\hat{f}_n \coloneqq \sum_{k=0}^{2^{j_0}-1} \hat{\alpha}_{j_0,k}\varphi_{j_0,k} + \sum_{j=j_0}^{j^{\max}_n-1}\sum_{k=0}^{2^j-1} \hat{\beta}^T_{j,k}\psi_{j,k}. \tag{6}$$

Given a uniform design $\xi_n = 2^{-i}\mathbb{Z} \cap [0,1)$, this is a standard hard-threshold estimate; otherwise it gives a generalisation to non-uniform designs.

2.2. *Adaptive design choices.* So far, we have only discussed how to estimate f from a fixed design. However, we can also use these estimates to choose the design points adaptively. We will choose the design points in stages, at stage m selecting points $x_{n_{m-1}+1}$ to x_{n_m} in terms of previous observations $Y_1, \ldots, Y_{n_{m-1}}$. The number of design points in each stage can be chosen freely, subject only to the conditions that n_0 is a power of two, and the ratios n_m/n_{m-1} are bounded away from 1 and ∞. We may, for example, choose

$$n_m \coloneqq \lfloor 2^{j+\tau m} \rfloor, \tag{7}$$

for some $j \in \mathbb{N}$, and $\tau > 0$.

In the initial stage, we will choose n_0 design points spaced uniformly on $[0,1]$,

$$x_i \coloneqq (i-1)/n_0, \quad 1 \le i \le n_0.$$

At further stages $m \in \mathbb{N}$, we will construct a *target density* p_m on $[0,1]$, and then select design points $x_{n_{m-1}+1}, \dots, x_{n_m}$ so that the design ξ_{n_m} approaches a draw from this density. We will choose p_m to be concentrated in regions of $[0,1]$ where we believe the function f is difficult to estimate, ensuring that later design points will adapt to the unknown shape of f.

At time n_{m-1}, for each $j \in \mathbb{N},\ j_0 \le j < j^{\max}_{n_{m-1}}$, we rank the 2^j thresholded estimates $\hat{\beta}^T_{j,k}$ in decreasing order of size. We then have

$$\left|\hat{\beta}^T_{j,r_j^{-1}(1)}\right| \ge \cdots \ge \left|\hat{\beta}^T_{j,r_j^{-1}(2^j)}\right|,$$

for a bijective ranking function r_j. We will choose the target density so that, in the support of each significant term $\beta_{j,k}\psi_{j,k}$ in the wavelet series, the density will be, up to log factors, at least $2^j/r_j(k)$. The density will thus be concentrated in regions where the wavelet coefficients are known to be large. To ensure that all coefficients are estimated accurately, we will also require the density to be bounded below by a fixed constant, given by a choice of parameter $\lambda > 0$.

Split the interval $[0,1]$ into sub-intervals

$$I_{l,m} \coloneqq 2^{-j^{\max}_{n_m}}[l, l+1), \quad l \in \mathbb{Z},\ 0 \le l < 2^{j^{\max}_{n_m}}.$$

We define the target density on $I_{l,m}$ to be

$$p_{l,m} \coloneqq A \max\Bigg(\{\lambda\} \cup \left\{ \frac{2^j}{r_j(k)(j^{\max}_{n_{m-1}})^2} : j \in \mathbb{N},\ j_0 \le j < j^{\max}_{n_{m-1}},\ I_{l,m} \subseteq S_{j,k},\ \hat{\beta}^T_{j,k} \ne 0 \right\} \Bigg),$$

where the fixed constant $A > 0$ is chosen so that the density p_m always integrates to at most one, $2^{-j^{\max}_{n_m}} \sum_l p_{l,m} \le 1$. The specific value of A is unimportant, but note that

$$2^{-j^{\max}_{n_m}} \sum_{l=0}^{2^{j^{\max}_{n_m}}-1} p_{l,m} \lesssim 1 + (j^{\max}_{n_{m-1}})^{-2} \sum_{j=0}^{j^{\max}_{n_{m-1}}} \sum_{k=1}^{2^j} k^{-1} \lesssim 1,$$

so such a choice of A exists.

We now aim to choose new design points $x_{n_{m-1}+1}, \ldots, x_{n_m}$ so that the design ξ_n approximates a draw from p_m. To simplify notation, we first include any points $x \in 2^{-j_{n_m}^{\max}}\mathbb{Z} \cap [0,1)$ not already in the design. We will assume the n_m and $j_n^{\max}$ are chosen so that this requires no more than $n_m - n_{m-1}$ design points; since $j_n^{\max}$ is defined only asymptotically, and $2^{j_{n_m}^{\max}} = o(n_m - n_{m-1})$, such a choice is always possible.

We then construct an *effective density* $q_{m,n}$, describing a nominal density generating the design ξ_n. This density will be at least $2^i/n$ on any region where the design contains the grid $2^{-i}\mathbb{Z}$; it will thus describe the density of all design points on regular grids. We define the effective density on $I_{l,m}$ at time n to be

$$q_{l,m,n} \coloneqq n^{-1} \max\left\{2^i : i \in \mathbb{N},\ 2^{-i}\mathbb{Z} \cap I_{l,m} \subseteq \xi_n\right\}.$$

Again, note this density integrates to at most one, $2^{-j_{n_m}^{\max}} \sum_l q_{l,m,n} \leq 1$.

Our remaining goal is to choose the new design points so that the effective density approaches the target density. In our proofs, we will require control over the maximum discrepancy from p_m to $q_{m,n}$,

$$\max_{l=0}^{2^{j_{n_m}^{\max}}-1} p_{l,m}/q_{l,m,n}. \tag{8}$$

To choose the next stage of design points, having selected points $x_1, \ldots, x_n$, we therefore pick an l maximising (8); note that this does not require us to calculate A. We then add points $2^{-i}\mathbb{Z} \cap I_{l,m}$ to the design, choosing the smallest index i for which at least one such point is not already present.

In doing so, we halve the largest value of $p_{l,m}/nq_{l,m,n}$, while leaving all other such values unchanged. Repeating this process, we will therefore add design points on grids $2^{-i}\mathbb{Z} \cap I_{l,m}$ so as to minimise (8). We continue until we have selected a total of n_m design points; for convenience, let $q_{l,m} \coloneqq q_{l,m,n_m}$ denote the effective density on $I_{l,m}$ once we are done.

The final algorithm is thus described by Algorithm 1; it can be implemented efficiently using a priority queue to find values of l maximising (8). We will show that this algorithm ensures the final effective density q_m is close to the target density p_m, and that estimates made under it are therefore spatially-adaptive for a wide variety of functions.

3. Theoretical results. We now provide theoretical results on the performance of our algorithm. We begin by defining the relevant function classes, then discuss our choice of functions considered; we conclude with our results on convergence rates.

Algorithm 1 Spatially-adaptive sensing

$n \leftarrow n_0$
$x_1, \dots, x_n \leftarrow n^{-1}\mathbb{Z} \cap [0,1)$
observe $Y_1, \dots, Y_n$
$m \leftarrow 1$
loop
  $x_{n+1}, \dots, x_{n'} \leftarrow 2^{-j_{n_m}^{\max}}\mathbb{Z} \cap [0,1) \setminus \xi_n$
  $n \leftarrow n'$
  while $n < n_m$ **do**
    choose l maximising $p_{l,m}/q_{l,m,n}$
    $S \leftarrow 2^{-i}\mathbb{Z} \cap I_{l,m} \setminus \xi_n$, for the smallest i such that $S \neq \emptyset$
    repeat
      $n \leftarrow n+1$
      choose $x_n \in S$
      $S \leftarrow S \setminus \{x_n\}$
    until $S = \emptyset$ or $n = n_m$
  end while
  observe $Y_{n_{m-1}+1}, \dots, Y_n$
  estimate f by $\hat{f}_n$
  $m \leftarrow m+1$
end loop

3.1. *Function classes.* We first define the function classes we will consider in the following. We will assume we have a wavelet basis $\psi_{j_0,k}, \varphi_{j,k}$ satisfying the assumptions of Section 2.1; we can then describe any function $f \in L^2([0,1])$ by its wavelet series,

$$f = \sum_{k=0}^{2^{j_0}-1} \alpha_{j_0,k}\varphi_{j_0,k} + \sum_{j=j_0}^{\infty}\sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k}.$$

The smoothness of f, and thus the ease with which it can be estimated, is determined by the size of the coefficients $\alpha_{j_0,k}, \beta_{j,k}$; f is smooth, and easily estimated, when these coefficients are small. The smoothness of a function can be described in terms of its membership of standard function classes. While there are many such classes, in what follows we will be interested primarily in the Hölder and Besov scales (Härdle et al., 1998).

For $s \in \mathbb{N}$, the Hölder classes $C^s(M)$ contain all functions which are at least s-times differentiable, with s-th derivative bounded by M; the local Hölder classes $C^s(M, I)$ instead require this condition to hold only over an interval I. These definitions can also be extended to non-integer s, and given in terms of wavelet coefficients. We note that while the wavelet definitions are in general slightly weaker than the classical ones, this will not fundamentally affect our results in what follows.

DEFINITION 1. *For $s \in (0, N)$, $M > 0$, and $I \subseteq [0,1]$, $C^s(M, I)$ is the class of functions $f \in L^2([0,1])$ satisfying*

$$\max\left(2^{j_0\left(s+\frac{1}{2}\right)} \sup_{k:S_{j_0,k}\subseteq I} |\alpha_{j_0,k}|,\ \sup_{j=j_0}^{\infty} 2^{j\left(s+\frac{1}{2}\right)} \sup_{k:S_{j,k}\subseteq I} |\beta_{j,k}|\right) \leq M.$$

For $I = [0,1]$, we denote this class $C^s(M)$.

The Besov classes $B^r_{p,\infty}(M)$ are more general. For $p = \infty$, they coincide with our definition of the Hölder classes $C^r(M)$. For $p < \infty$, they allow functions with some singularities, provided they are still, on average, r-times differentiable; smaller values of p correspond to more irregular functions.

DEFINITION 2. *For $r \in (0, N)$, $p \in [1, \infty)$, and $M > 0$, $B^r_{p,\infty}(M)$ is the class of functions $f \in L^2([0,1])$ satisfying*

$$\max\left(2^{j_0\left(r+\frac{1}{2}-\frac{1}{p}\right)} \left(\sum_k |\alpha_{j_0,k}|^p\right)^{\frac{1}{p}},\ \sup_{j=j_0}^{\infty} 2^{j\left(r+\frac{1}{2}-\frac{1}{p}\right)} \left(\sum_k |\beta_{j,k}|^p\right)^{\frac{1}{p}}\right) \leq M.$$

For $p = \infty$, we define $B^r_{\infty,\infty}(M) \coloneqq C^r(M)$.

Many other standard classes are related to these Besov classes, including the Sobolev classes $W^{r,p}(M) \subseteq B^r_{p,\infty}(M)$, the Sobolev Hilbert classes $H^r(M) \subseteq B^r_{2,\infty}(M)$, and the functions of bounded variation $BV(M) \subseteq B^1_{1,\infty}(M)$. In each case, convergence rates are unchanged by considering the containing Besov class, meaning we need consider only Besov classes in what follows.

Besov classes can also be thought of as describing functions whose wavelet expansions are *sparse*. From the above definitions, we can see that, compared to a Hölder class, functions in a Besov class can have a number of larger wavelet coefficients, provided there are not too many. In other words, functions in a Besov class can have wavelet expansions where most, but not all, coefficients are small.

Besov classes are often used to describe spatially-inhomogeneous functions; we can see why by considering Figure 2, which plots the wavelet coefficients of the functions in Figure 1. As above, the coefficients are often, but not always, small.

Our final function class is a new definition, which we will argue captures another typical feature of spatially-inhomogeneous functions, and is necessary to obtain improved rates of convergence. From Figure 2, we can see that, in regions where the functions f are rough, their wavelet coefficients

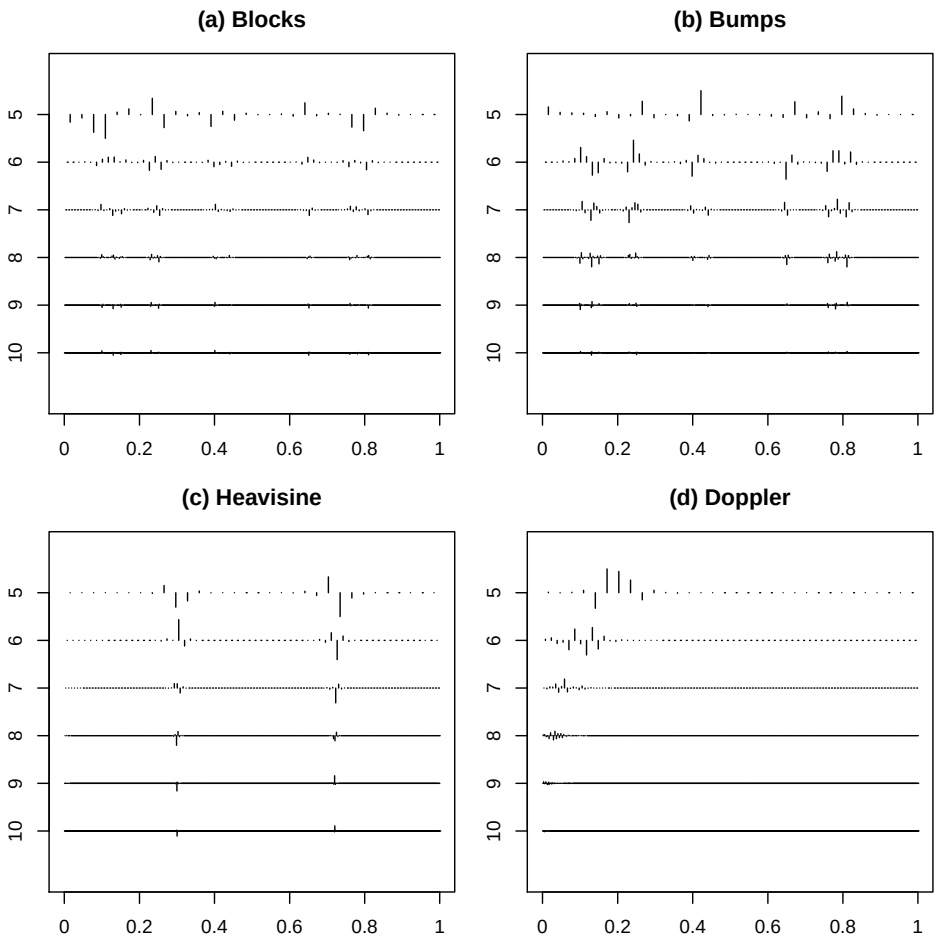

FIG 2. *CDV-8 wavelet coefficients for the functions in Figure 1. The height of each line corresponds to a wavelet coefficient* $\beta_{j,k}$; *the* x*-axis plots the location* $2^{-j}k$, *and the* y*-axis the scale* j.

are often large; in regions where they are smooth, their coefficients are small. In other words, if f is difficult to approximate in some region at high resolution, it will also be difficult to approximate there at lower resolutions.

We will call such functions *detectable*, and describe them in terms of detectable classes $D_t^s(M, I) \subseteq C^s(M, I)$. The additional parameter $t \in (0, 1)$ controls the strength of our condition; larger t corresponds to a stronger condition on functions f.

DEFINITION 3. *For $s \in (0, N)$, $t \in (0, 1)$, $M > 0$, and an interval $I \subseteq [0, 1]$, $D_t^s(M, I)$ is the class of functions $f \in C^s(M, I)$ which also satisfy*

$$\begin{aligned}(9)\quad \forall\, j \in \mathbb{N},\ j \geq \lceil \tfrac{j_0}{t} \rceil,\ k : S_{j,k} \cap I \neq \emptyset,\\ \exists\, j' \in \mathbb{N},\ \lfloor tj \rfloor \leq j' < j,\ k' : S_{j',k'} \supset S_{j,k},\\ |\beta_{j',k'}| \geq (j'/j) 2^{(j-j')\left(s+\frac{1}{2}\right)} |\beta_{j,k}|.\end{aligned}$$

The definition thus requires that each term in the wavelet series on I, at a fine scale j, lies within the support of another term, of comparable size, at a coarser scale j'. The parameter s controls how large this second term must be, and t controls how far apart the scales j and j' can be.

In Section 3.2, we will discuss why such conditions may be natural to consider for this problem. First, however, we will establish that a typical locally Hölder function will be detectable; indeed, similarly to results in Jaffard (2000) and Giné and Nickl (2010), we can show that the set of functions which are locally Hölder but not detectable is topologically negligible. We may therefore sensibly restrict to detectable functions in what follows.

PROPOSITION 4. *For $s \in (0, N)$, $M > 0$, and any interval $I \subseteq [0, 1]$, define*

$$\mathcal{D} \coloneqq C^s(M, I) \setminus \bigcup_{t \in (0,1)} D_t^s(M, I).$$

Then $\mathcal{D}$ is nowhere dense in the norm topology of $C^s(M, I)$.

3.2. *Spatially-inhomogeneous functions.* We now discuss our choice of functions to consider. We begin with some well-known results, which describe the limitations of adaptive sensing over Hölder classes. Let

$$\alpha(s) \coloneqq s/(2s + 1),$$

and define an adaptive-sensing algorithm to be a choice of design points $x_n = x_n(Y_1, \ldots, Y_{n-1})$, together with an estimator $\hat{f}_n = \hat{f}_n(Y_1, \ldots, Y_n)$.

Then, up to log factors, a spatially-adaptive estimate can attain the rate $n^{-\alpha(s)}$ over any local Hölder class $C^s(M, I)$, and this cannot be improved upon by adaptive sensing.

THEOREM 5. *Using a uniform design $x_i = (i-1)/n$, there exists an estimator $\hat{f}_n$, which satisfies*

$$\sup_{x \in J} |\hat{f}_n(x) - f(x)| = O_p(c_n),$$

uniformly over $f \in C^s(M, I) \cap C^{\frac{1}{2}}(M)$, for any $s \in [\frac{1}{2}, N)$, $M > 0$, I an interval open in $[0, 1]$, $J \subseteq I$ a closed interval, and

$$c_n = (n/\log(n))^{-\alpha(s)}.$$

THEOREM 6. *Let $s \geq \frac{1}{2}$, $M > 0$, I an interval open in $[0, 1]$, and $J \subseteq I$ a closed interval. Given an adaptive-sensing algorithm with estimator $\hat{f}_n$, if*

$$\sup_{x \in J} |\hat{f}_n(x) - f(x)| = O_p(c_n)$$

uniformly over $f \in C^s(M, I) \cap C^{\frac{1}{2}}(M)$, then

$$c_n \gtrsim n^{-\alpha(s)}.$$

To benefit from adaptive sensing, we will need to exploit two features of the functions in Figure 1. The first is that, as discussed in Section 3.1, these functions are sparse: they are rougher in some regions than others. This sparsity is necessary to benefit from adaptive sensing: it is the difference between rough and smooth which allows us to improve performance, placing more design points in rougher regions.

Sparsity is commonly measured in terms of Besov, rather than Hölder, classes. This change alone, however, is not enough to allow us to benefit from adaptive sensing. Since $B^r_{p,\infty}(M) \subseteq C^{r-1/p}(M)$, over this class we can achieve the rate $n^{-\alpha(r-1/p)}$, up to log factors, with the fixed-design method of Theorem 5. We can further show that, in this case, adaptive sensing offers little improvement.

THEOREM 7. *Let $p \in [1, \infty]$, $r \geq \frac{1}{2} + \frac{1}{p}$, $M > 0$, and I be an interval in $[0, 1]$. Given an adaptive-sensing algorithm with estimator $\hat{f}_n$, if*

$$\sup_{x \in I} |\hat{f}_n(x) - f(x)| = O_p(c_n),$$

uniformly over $f \in B^r_{p,\infty}(M)$, *then*

$$c_n \gtrsim n^{-\alpha\left(r-\frac{1}{p}\right)}.$$

To benefit from adaptive sensing, it is not enough to have regions in which the function is rough or smooth; we must also be able to detect where those regions are. This is the rationale behind our detectable classes $D^s_t(M, I)$: if a function is detectable, its roughness at high resolutions will be signalled by corresponding roughness at low resolutions, which we can observe in advance.

We will be interested in functions f which are both sparse and detectable. For $p \in [1, \infty]$, $r \in [\frac{1}{2} + \frac{1}{p}, N)$, $s \in [r - \frac{1}{p}, N)$, $t \in (0, 1)$, $M > 0$, and any interval $I \subseteq [0, 1]$, let

$$\mathcal{F} = \mathcal{F}(p, r, s, t, M, I) \coloneqq B^r_{p,\infty}(M) \cap D^s_t(M, I) \tag{10}$$

denote a class of sparse and detectable functions.

We note that this class has two smoothness parameters: r governs the average global smoothness of a function $f \in \mathcal{F}$, while s governs its local smoothness over I. Since functions in $B^r_{p,\infty}$ are everywhere at least $(r - \frac{1}{p})$-smooth, we have restricted to the interesting case $s \geq r - \frac{1}{p}$.

From Proposition 4, we know that quasi-all locally Hölder functions are detectable; we can likewise show that under a fixed design, restricting to sparse and detectable functions does not alter the minimax rate of estimation. We may thus conclude that requiring sparsity and detectability thus does not make estimation fundamentally easier.

THEOREM 8. *Using a fixed design, if an estimator* $\hat{f}_n$ *satisfies*

$$\sup_{x \in I} |\hat{f}_n(x) - f(x)| = O_p(c_n),$$

uniformly over $f \in \mathcal{F}$, *then*

$$c_n \gtrsim (n/\log(n))^{-\alpha(s)}.$$

3.3. *Benefits of adaptive sensing.* With adaptive sensing, however, we can take advantage of these conditions to obtain improved rates of convergence. We even can show that Algorithm 1 achieves this without knowledge of the class $\mathcal{F}$; we can thus adapt not only to the regions where f is rough, but also to the overall smoothness and sparsity of f.

THEOREM 9. *Algorithm 1 satisfies*

$$\sup_{x\in I}|\hat{f}_n(x) - f(x)| = O_p(c_n),$$

uniformly over $f \in \mathcal{F}$, *for* $u \coloneqq \max(r-s, 0)$,

$$r' \coloneqq s + tu, \qquad s' \coloneqq s/(1-ptu), \tag{11}$$

and

$$c_n \coloneqq \left(n/\log(n)^3\right)^{-\alpha(\min(r',s'))} \log(n)^{1(r'=s')}. \tag{12}$$

We thus obtain, up to log factors, the weaker of the two rates $n^{-\alpha(r')}$ and $n^{-\alpha(s')}$. Both of these rates are at least as good as the $n^{-\alpha(s)}$ bound faced by a fixed design; when $s < r$, and the function f may be locally rough, the rates are strictly better. In that case, we obtain the $n^{-\alpha(r')}$ rate when $s/r \geq (1-t)/(2-t)$, and the $n^{-\alpha(s')}$ rate otherwise.

The improvement is driven by two parameters: t, which governs how easy it is to detect irregularities of f, and u, which governs how much rougher f is locally than on average. When both t and u are large, the rates we obtain are significantly improved; in the most favourable case, when $u = 1$, and $t \approx 1$, this result is equivalent to gaining an extra derivative of f. We can even show that these rates are near-optimal over classes $\mathcal{F}$.

THEOREM 10. *Given an adaptive-sensing algorithm with estimator* $\hat{f}_n$, *if*

$$\sup_{x\in I}|\hat{f}_n(x) - f(x)| = O_p(c_n),$$

uniformly over $f \in \mathcal{F}$, *then*

$$c_n \gtrsim n^{-\alpha(\min(r',s'))},$$

for r', s' *given by* (11).

Furthermore, we also have that, even in the absence of sparsity or detectability, we still achieve the spatial adaptation properties of a fixed design. We may thus use our adaptive-sensing algorithm with the confidence that, even if f is spatially homogeneous, we will not pay an asymptotic penalty.

THEOREM 11. *Algorithm 1 satisfies*

$$\sup_{x\in J}|\hat{f}_n(x) - f(x)| = O_p\left(c_n\right),$$

uniformly over $f \in C^s(M, I) \cap C^{\frac{1}{2}}(M)$, *for any* $s \in [\frac{1}{2}, N)$, $M > 0$, I *an interval open in* $[0,1]$, $J \subseteq I$ *a closed interval, and*

$$c_n \coloneqq \left(n/\log(n)\right)^{-\alpha(s)}.$$

4. Implementation and experiments. We now give some implementation details of Algorithm 1, and provide empirical results. Before we test the algorithm, we must describe how we compute $\hat{f}_n$, and choose the parameters governing the algorithm's behaviour.

4.1. *Estimating functions.* For simplicity, in (6) we defined $\hat{f}_n$ in terms of wavelets only up to the resolution level $j_n^{\max}$. While asymptotically this carries no penalty, in finite time we may do better by estimating all the wavelets for which we have available data. In other words, we use the estimate

$$\hat{f}_n \coloneqq \sum_{k=0}^{2^{j_0}-1} \hat{\alpha}_{j_0,k}\varphi_{j_0,k} + \sum_{j=j_0}^{\infty}\sum_{k=0}^{2^j-1} \hat{\beta}^T_{j,k}\psi_{j,k}, \tag{13}$$

where for $j \geq j_n^{\max}$, if the set in (4) is empty, we let $i_n(j,k) \coloneqq -\infty$, forcing $\hat{\beta}^T_{j,k} = 0$. We note that since there are finitely many design points, the sum in (13) must have finitely many non-zero terms; define

$$J_n \coloneqq 1 + \sup\{j \in \mathbb{N} : \exists\ k, i_n(j,k) > -\infty\}$$

to be the resolution level at which this sum terminates.

To compute these estimates $\hat{f}_n$, we must convert the estimated coefficients $\hat{\alpha}_{j_0,k}$, $\hat{\beta}^T_{j,k}$ back into function values $\hat{f}_n(x)$. For $i \in \mathbb{N},\ i \geq J_n$, to evaluate $\hat{f}_n$ at points $x = 2^{-i}k,\ k \in \mathbb{Z},\ 0 \leq k < 2^i$, we make the approximation

$$\hat{f}_n(2^{-i}k) \approx 2^{\frac{i}{2}}\hat{\alpha}^T_{i,k}, \tag{14}$$

where the post-thresholding scaling coefficients $\hat{\alpha}^T_{i,k}$ are defined by

$$\sum_{k=0}^{2^i-1} \hat{\alpha}^T_{i,k}\varphi_{i,k} \coloneqq \hat{f}_n.$$

These can again be computed efficiently using the fast wavelet transform.

Given a uniform design, and predicting f only at the design points, this is enough to give estimates $\hat{f}_n$; if we set $\kappa = 1$, we have just described a standard hard-threshold wavelet estimate (Donoho and Johnstone, 1994). In that case, the observations and predictions are always made at the same scale, $i_n(j,k) = i$, so the errors in (1) and (14) tend to cancel out. In other cases, however, the observations and predictions may be at different scales; these errors then may build up, making $\hat{f}_n$ look like a translation of f.

To resolve the issue, we will use a slightly different definition of the estimated coefficients $\hat{\alpha}_{j_0,k}$ and $\hat{\beta}_{j,k}$, which ensures the scales of observation

and prediction are the same. Given $i \in \mathbb{N},\ i \geq J_n$, to estimate f at points $x = 2^{-i}k,\ k \in \mathbb{Z},\ 0 \leq k < 2^i$, we set $x_{n,k} := \sup\{x \in \xi_n : x \leq 2^{-i}k\}$, and let

$$\hat{\alpha}_{i,k} := 2^{-\frac{i}{2}} Y(x_{n,k}).$$

We then define the estimates $\hat{\alpha}_{j_0,k}$ and $\hat{\beta}_{j,k}$ by

$$\sum_{k=0}^{2^{j_0}-1} \hat{\alpha}_{j_0,k}\varphi_{j_0,k} + \sum_{j=j_0}^{i-1}\sum_{k=0}^{2^j-1} \hat{\beta}_{j,k}\psi_{j,k} := \sum_{k=0}^{2^i-1} \hat{\alpha}_{i,k}\varphi_{i,k},$$

using the fast wavelet transform as before.

We note that this definition is approximately the same as the one in (3); while it is harder to control theoretically, it gives improved experimental behaviour. We also note that, with a uniform design, if we wish to predict f only at the design points, this again reduces to a standard wavelet estimate.

4.2. *Choosing parameters.* To apply Algorithm 1, we must choose the parameters κ, λ and τ, and also estimate σ if it is not already known. The parameter κ governs the size of our wavelet thresholds: larger κ means we will be more conservative. While our theoretical results are proved for choices $\kappa > 1$, in our empirical tests we took $\kappa = 1$. This gives a simple choice of threshold which performs well, and allows us to compare our results with standard hard-threshold estimates.

The parameter λ controls how uniform we make our design points: for $\lambda \gg 0$ the design points will be mostly uniform, while for $\lambda \approx 0$ they will be concentrated at irregularities of f. The parameter τ likewise controls how many design points we choose at each stage: for $\tau \gg 0$ there will be a few large stages, while for $\tau \approx 0$ there will be many small ones. Empirically, we found the values $\lambda = \tau = \frac{1}{2}$ gave good trade-offs.

Finally, for uniform designs, Donoho and Johnstone (1994) suggest estimating σ by the median size of the $\hat{\beta}_{j,k}$ at fine resolution scales. Our designs may not be uniform, but they are guaranteed to provide us with estimates $\hat{\beta}_{j,k}$ up to level $j_n^{\max} - 1$. We will therefore use the similar estimate

$$\hat{\sigma}_n := \text{median}\{2^{\frac{1}{2}i}|\hat{\beta}_{j,k}| : j \geq j_n^{\max} - 1, i_n(j,k) > -\infty\}/0.6745,$$

which includes all estimated coefficients at scales at least this fine.

4.3. *Empirical results.* We now describe the results of using Algorithm 1 to estimate the functions in Figure 1, observing under $N(0,\sigma^2)$ noise. Figure 3 plots $n = 2^{11}$ noisy samples of each function, under a uniform design,

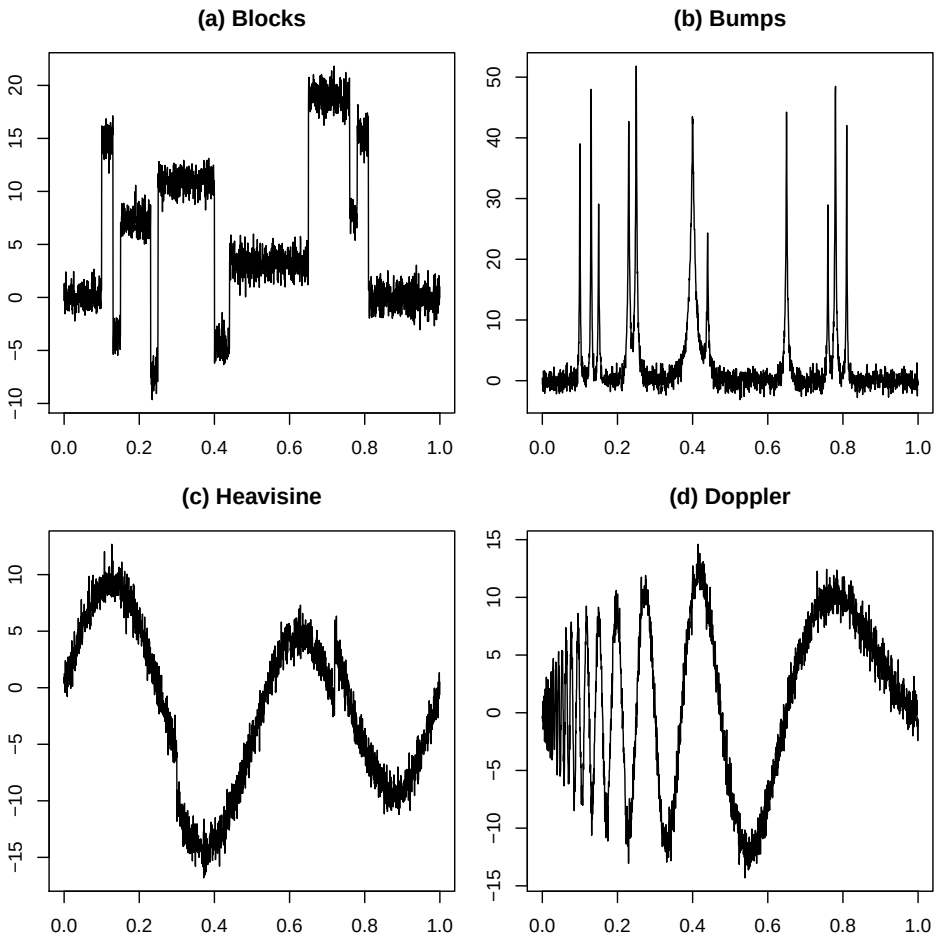

FIG 3. *Noisy samples from the functions in Figure 1.*

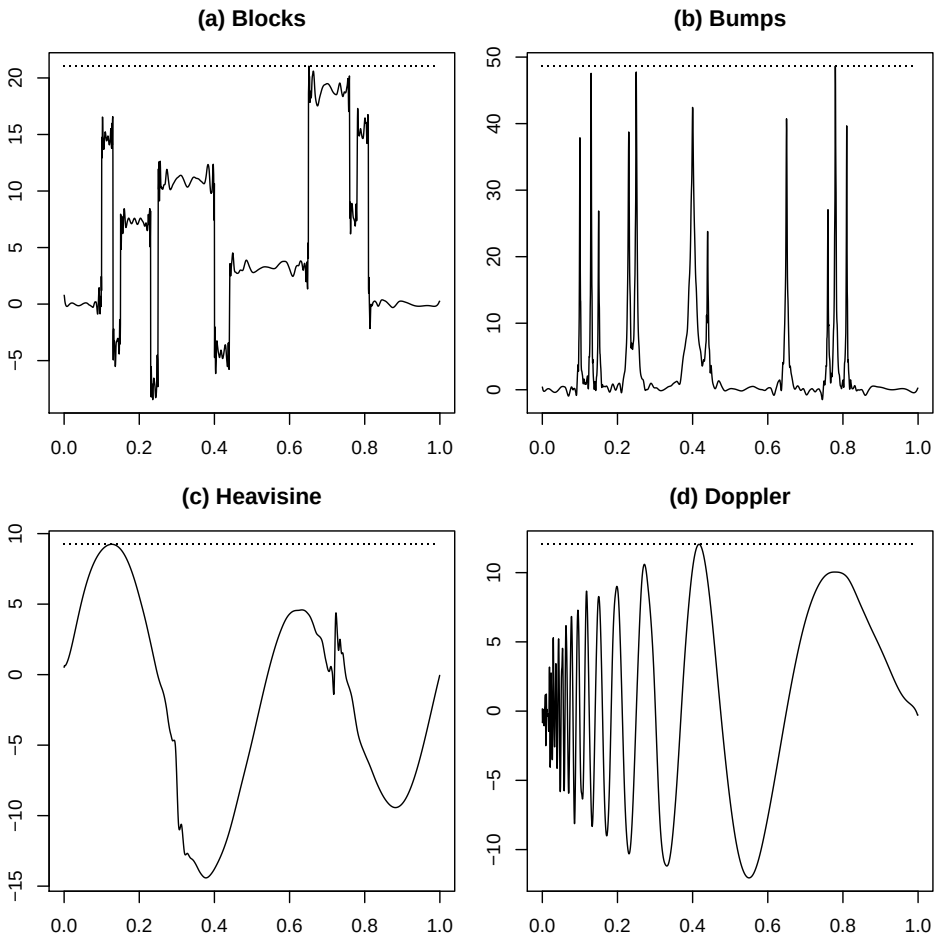

FIG 4. *Fixed-design estimates of the functions in Figure 1.*

with $\sigma = 1$, while Figure 4 plots a standard wavelet threshold estimate from these samples.

Figure 5 plots typical results of using Algorithm 1 under these conditions. The algorithm was again given access to $n = 2^{11}$ observations, with $\sigma = 1$; we set $n_0 = 2^6$, and chose the parameters $\kappa,\ \lambda,\ \tau$, and $\hat{\sigma}_n$ as in Section 4.2. We used the family of wavelet bases described by Cohen, Daubechies and Vial (1993), and implemented in Nason (2010); we took wavelets with $N = 8$ vanishing moments, set $j_0 = 5$, and $j_n^{\max} = \max(j_0 + 1, \lfloor n/\log(n) \rfloor)$.

The dots along the top of each plot are drawn proportionally to the number of design points. We can see that, for the Heavisine and Doppler functions, the adaptive design concentrated in the regions where the function is rough; as a result, the adaptive estimates are noticeably better at recovering the shape of these curves.

For the Blocks and Bumps functions, which have more complicated patterns of spatial inhomogeneity, with these measurements the adaptive design was not able to locate all the areas where these functions are rough. However, we might expect performance on all the above functions to improve as the number of design points increases; to this end, we next considered performance with up to $n = 2^{14}$ design points.

At this level of detail, it becomes harder to visually compare estimates; instead, to numerically measure the spatial adaptivity of our estimates, we evaluated procedures in terms of their maximum error over $[0, 1]$, approximated by

$$\max_{x \in 2^{-j}\mathbb{Z} \cap [0,1)} |\hat{f}_n(x) - f(x)|$$

for j large. In the following, we took $j = 17$, to avoid biasing the performance measure towards a uniform design.

Figure 6 compares the performance of the two methods on the Doppler function, with $\sigma = 1$; the values plotted are sample medians after 250 runs, together with 95% confidence intervals for the true median. We can see that for n large, the adaptive design significantly outperforms the uniform one, consistent with a difference in the asymptotic rate of estimation.

Table 1 compares performance on all the functions in Figure 1, given $n = 2^{14}$ observations, and varying levels of σ. We again report sample medians after 250 runs, together with the p-value of a two-sided Mann-Whitney-U test for difference in medians. (We note that the large errors reported for the Blocks function are due to the large discontinuities present, which are difficult to estimate uniformly over $[0, 1]$.)

We can see that for the Blocks function, the uniform design fared slightly better, as the adaptive algorithm still struggled to choose a good design.

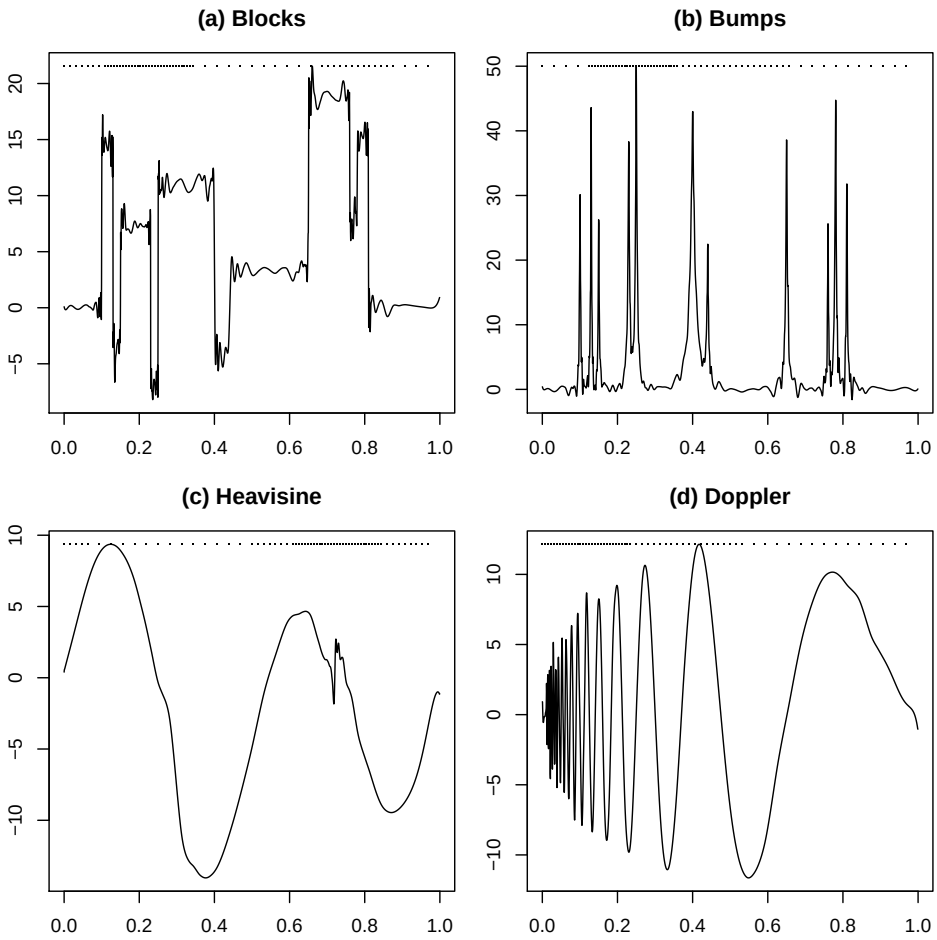

FIG 5. *Adaptive-sampling estimates of the functions in Figure 1.*

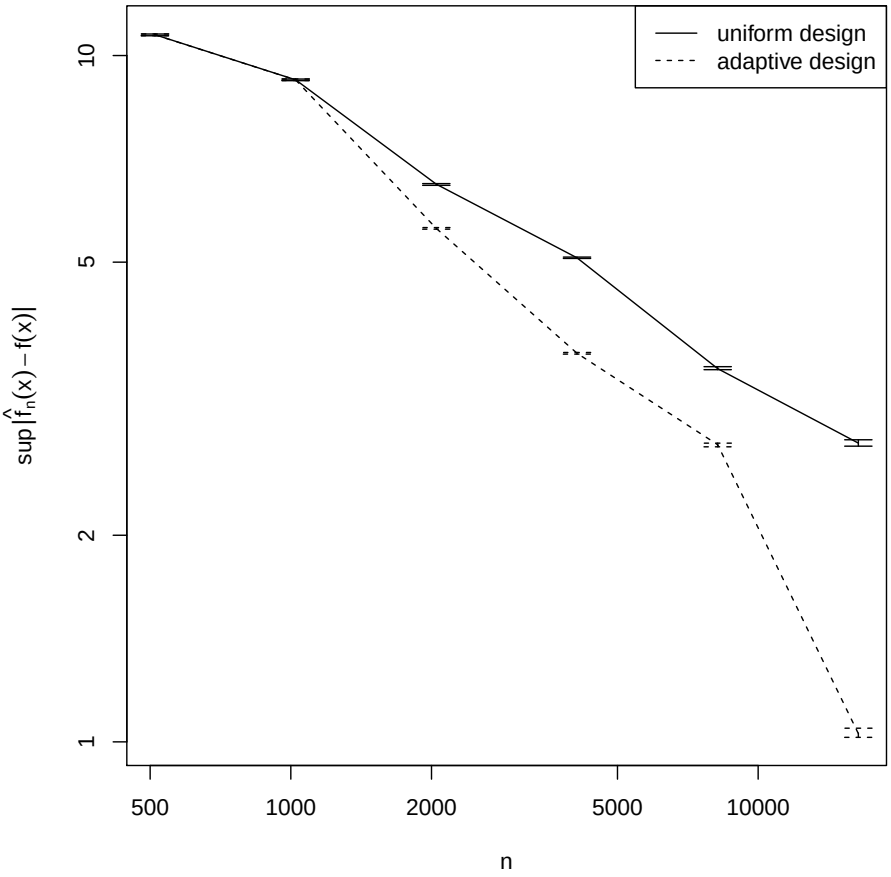

FIG 6. *Log-log plot of empirical performance on the Doppler function.*

	uniform design	adaptive design	p-value
$\sigma = 0.5$			
Blocks	**13.284**	13.3882	< 0.001
Bumps	3.553	**3.0086**	< 0.001
Heavisine	2.646	**2.5902**	< 0.001
Doppler	1.783	**0.5138**	< 0.001
$\sigma = 1$			
Blocks	**12.355**	12.799	< 0.001
Bumps	5.721	**5.487**	< 0.001
Heavisine	3.260	**3.204**	< 0.001
Doppler	2.725	**1.028**	< 0.001
$\sigma = 2$			
Blocks	11.121	10.988	0.428
Bumps	8.947	**7.964**	< 0.001
Heavisine	3.053	3.061	0.815
Doppler	3.621	**2.984**	< 0.001

TABLE 1
Empirical performance on the functions in Figure 1 for $n = 2^{14}$.

However, for the other three functions, adaptive sampling provided a significant improvement; the improvement was largest for small σ, but still significant for two of the three functions even with large σ. We thus conclude that adaptive sensing can be of value in nonparametric regression whenever the function f may be spatially inhomogeneous.

Acknowledgements. We would like to thank Richard Nickl and several anonymous referees for their valuable comments and suggestions.

SUPPLEMENTARY MATERIAL

Proofs for "Spatially-adaptive sensing in nonparametric regression."
(doi: ???http://lib.stat.cmu.edu/aos/???/???; .pdf).

References.

ARIAS-CASTRO, E., CANDES, E. J. and DAVENPORT, M. (2011). On the Fundamental Limits of Adaptive Sensing Arxiv preprint report No. arXiv:1111.4646.

BEYGELZIMER, A., DASGUPTA, S. and LANGFORD, J. (2009). Importance weighted active learning. In *Proceedings of the 26th Annual International Conference on Machine Learning* 49–56.

BOUFOUNOS, P., CEVHER, V., GILBERT, A., LI, Y. and STRAUSS, M. (2012). What's the Frequency, Kenneth?: Sublinear Fourier Sampling Off the Grid. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques* 61–72.

Bull, A. D. (2012a). Honest adaptive confidence bands and self-similar functions. *Electronic Journal of Statistics* **6** 1490–1516.

Bull, A. D. (2012b). Supplement to "Spatially-adaptive sensing in nonparametric regression.".

Castro, R. M. (2012). Adaptive Sensing Performance Lower Bounds for Sparse Signal Estimation and Testing Arxiv preprint report No. arXiv:1206.0648.

Castro, R. M. and Nowak, R. D. (2008). Minimax bounds for active learning. *Institute of Electrical and Electronics Engineers. Transactions on Information Theory* **54** 2339–2353.

Castro, R., Willett, R. and Nowak, R. (2006). Faster rates in regression via active learning. *Advances in Neural Information Processing Systems* **18** 179.

Cohen, A., Daubechies, I. and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis* **1** 54–81.

Cohn, D., Atlas, L. and Ladner, R. (1994). Improving generalization with active learning. *Machine Learning* **15** 201–221.

Cohn, D. A., Ghahramani, Z. and Jordan, M. I. (1996). Active Learning with Statistical Models. *Journal of Artificial Intelligence Research* **4** 129–145.

Davenport, M. A. and Arias-Castro, E. (2012). Compressive binary search Arxiv preprint report No. arXiv:1202.0937.

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

Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *The Annals of Statistics* **26** 879–921.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society B* **57** 301–369.

Efromovich, S. (2008). Optimal Sequential Design in a Controlled Non-parametric Regression. *Scandinavian Journal of Statistics* **35** 266–285.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society B* **57** 371–394.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society B* **70** 849–911.

Fan, J., Hall, P., Martin, M. and Patil, P. (1999). Adaptation to high spatial inhomogeneity using wavelet methods. *Statistica Sinica* **9** 85–102.

Faraway, J. (1990). Sequential design for the nonparametric regression of curves and surfaces. In *Computing Science and Statistics* **90** 104–110.

Frisch, U. and Parisi, G. (1980). On the singularity structure of fully developed turbluence. In *Annals of the New York Academy of Sciences*, **357** 84–88.

Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *The Annals of Statistics* **38** 1122–1170.

Hall, P. and Molchanov, I. (2003). Sequential methods for design-adaptive estimation of discontinuities in regression curves and surfaces. *The Annals of Statistics* **31** 921–941.

Hanneke, S. (2011). Rates of convergence in active learning. *The Annals of Statistics* **39** 333–361.

Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998). *Wavelets, Approximation, and Statistical Applications. Lecture Notes in Statistics* **129**. Springer-Verlag, New York.

Haupt, J., Castro, R. M. and Nowak, R. (2011). Distilled sensing: adaptive sampling for sparse detection and estimation. *Institute of Electrical and Electronics Engineers. Transactions on Information Theory* **57** 6222–6235.

Iwen, M. A. (2009). Group testing strategies for recovery of sparse signals in noise. In *Signals, Systems and Computers, 2009 Conference Record of the Forty-Third Asilomar Conference on* 1561–1565.

Jaffard, S. (2000). On the Frisch-Parisi conjecture. *Journal de Mathématiques Pures et Appliquées. Neuvième Série* **79** 525–552.

Kerkyacharian, G. and Picard, D. (2004). Regression in random design and warped wavelets. *Bernoulli* **10** 1053–1105.

Koltchinskii, V. (2010). Rademacher complexities and bounding the excess risk in active learning. *Journal of Machine Learning Research* **11** 2457–2485.

Lepski, O. V., Mammen, E. and Spokoiny, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics* **25** 929–947.

Malloy, M. and Nowak, R. (2011a). On the Limits of Sequential Testing in High Dimensions Arxiv preprint report No. arXiv:1105.4540.

Malloy, M. and Nowak, R. (2011b). Sequential Analysis in High Dimensional Multiple Testing and Sparse Recovery Arxiv preprint report No. arXiv:1103.5991.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society B* **72** 417–473.

Nason, G. (2010). Wavethresh: wavelet statistics and transforms. R package version 4.5.

Picard, D. and Tribouley, K. (2000). Adaptive confidence interval for pointwise curve estimation. *The Annals of Statistics* **28** 298–335.

van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics* **3** 1360–1392.

Statistical Laboratory, Cambridge CB3 0WB, UK,
E-mail: a.bull@statslab.cam.ac.uk