SPARSE ADDITIVE MODELS

Pradeep Ravikumar

University of California, Berkeley

John Lafferty, Han Liu and Larry Wasserman

Carnegie Mellon University

February 25, 2009

We present a new class of methods for high-dimensional nonparametric regression and classification called sparse additive models (SpAM). Our methods combine ideas from sparse linear modeling and additive nonparametric regression. We derive an algorithm for fitting the models that is practical and effective even when the number of covariates is larger than the sample size. SpAM is essentially a functional version of the grouped lasso of Yuan and Lin (2006). SpAM is also closely related to the COSSO model of Lin and Zhang (2006), but decouples smoothing and sparsity, enabling the use of arbitrary nonparametric smoothers. We give an analysis of the theoretical properties of sparse additive models, and present empirical results on synthetic and real data, showing that SpAM can be effective in fitting sparse nonparametric models in high dimensional data.

1. Introduction. Substantial progress has been made recently on the problem of fitting high dimensional linear regression models of the form $Y_i = X_i^T\beta + \epsilon_i$, for $i = 1, \ldots, n$. Here Y_i is a real-valued response, X_i is a predictor and ϵ_i is a mean zero error term. Finding an estimate of β when $p > n$ that is both statistically well-behaved and computationally efficient has proved challenging; however, under the assumption that the vector β is sparse, the lasso estimator (Tibshirani (1996)) has been remarkably successful. The lasso estimator $\widehat{\beta}$ minimizes the ℓ_1-penalized sum of squares $\sum_i (Y_i - X_i^T\beta)^2 + \lambda \sum_{j=1}^p |\beta_j|$ with the ℓ_1 penalty $\|\beta\|_1$ encouraging sparse solutions, where many components $\widehat{\beta}_j$ are zero. The good empirical success of this estimator has been recently backed up by results confirming that it has strong theoretical properties; see (Bunea et al., 2007; Greenshtein and Ritov, 2004; Meinshausen and Yu, 2006; Wainwright, 2006; Zhao and Yu, 2007).

Keywords and phrases: lasso, nonparametric regression, additive model, sparsity

The nonparametric regression model $Y_i = m(X_i)+\epsilon_i$, where m is a general smooth function, relaxes the strong assumptions made by a linear model, but is much more challenging in high dimensions. Hastie and Tibshirani (1999) introduced the class of additive models of the form

$$Y_i = \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon_i. \tag{1}$$

This additive combination of univariate functions—one for each covariate X_j—is less general than joint multivariate nonparametric models, but can be more interpretable and easier to fit; in particular, an additive model can be estimated using a coordinate descent Gauss-Seidel procedure, called backfitting. Unfortunately, additive models only have good statistical and computational behavior when the number of variables p is not large relative to the sample size n, so their usefulness is limited in the high dimensional setting.

In this paper we investigate sparse additive models (SpAM), which extend the advantages of sparse linear models to the additive, nonparametric setting. The underlying model is the same as in (1), but we impose a sparsity constraint on the index set $\{j : f_j \not\equiv 0\}$ of functions f_j that are not identically zero. Lin and Zhang (2006) have proposed COSSO, an extension of lasso to this setting, for the case where the component functions f_j belong to a reproducing kernel Hilbert space (RKHS). They penalize the sum of the RKHS norms of the component functions. Yuan (2007) proposed an extension of the non-negative garrote to this setting. As with the parametric non-negative garrote, the success of this method depends on the initial estimates of component functions f_j.

In Section 3, we formulate an optimization problem in the population setting that induces sparsity. Then we derive a sample version of the solution. The SpAM estimation procedure we introduce allows the use of arbitrary nonparametric smoothing techniques, effectively resulting in a combination of the lasso and backfitting. The algorithm extends to classification problems using generalized additive models. As we explain later, SpAM can also be thought of as a functional version of the grouped lasso (Antoniadis and Fan, 2001; Yuan and Lin, 2006).

The main results of this paper include the formulation of a convex optimization problem for estimating a sparse additive model, an efficient backfitting algorithm for constructing the estimator, and theoretical results that analyze the effectiveness of the estimator in the high dimensional setting. Our theoretical results are of two different types. First, we show that, under suitable choices of the design parameters, the SpAM backfitting algorithm recovers the correct sparsity pattern asymptotically; this is a property we call *sparsistency*, as a shorthand for "sparsity pattern consistency." Second, we show that that the estimator is *persistent*, in the sense of Greenshtein and Ritov (2004), which is a form of risk consistency.

In the following section we establish notation and assumptions. In Section 3 we formulate SpAM as an optimization problem and derive a scalable backfitting algorithm. Examples showing the use of our sparse backfitting estimator on high dimensional data are included in Section 5. In Section 6.1 we formulate the sparsistency result, when orthogonal function regression is used for smoothing. In Section 6.2 we give the persistence result. Section 7 contains a discussion of the results and possible extensions. Proofs are contained in Section 8.

The statements of the Theorems in this paper were given, without proof, in Ravikumar et al. (2008). The backfitting algorithm was also presented there. Related results were obtained independently in Meier et al. (2008) and Koltchinskii and Yuan (2008).

2. Notation and Assumptions.

We assume that we are given independent data $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $X_i = (X_{i1}, \ldots, X_{ij}, \ldots, X_{ip})^T \in [0,1]^p$ and

$$Y_i = m(X_i) + \epsilon_i \tag{2}$$

with $\epsilon_i \sim N(0, \sigma^2)$ independent of X_i and

$$m(x) = \sum_{j=1}^{p} f_j(x_j). \tag{3}$$

Let μ denote the distribution of X, and let μ_j denote the marginal distribution of X_j for each $j = 1, \ldots, p$. For a function f_j on $[0,1]$ denote its $L_2(\mu_j)$ norm by

$$\|f_j\|_{\mu_j} = \sqrt{\int_0^1 f_j^2(x)\, d\mu_j(x)} = \sqrt{\mathbb{E}(f_j(X_j)^2)}. \tag{4}$$

When the variable X_j is clear from the context, we remove the dependence on μ_j in the notation $\|\cdot\|_{\mu_j}$ and simply write $\|f_j\|$.

For $j \in \{1, \ldots, p\}$, let $\mathcal{H}_j$ denote the Hilbert subspace $L_2(\mu_j)$ of measurable functions $f_j(x_j)$ of the single scalar variable x_j with zero mean, $\mathbb{E}(f_j(X_j)) = 0$. Thus, $\mathcal{H}_j$ has the inner product

$$\left\langle f_j, f_j' \right\rangle = \mathbb{E}\left(f_j(X_j) f_j'(X_j)\right) \tag{5}$$

and $\|f_j\| = \sqrt{\mathbb{E}(f_j(X_j)^2)} < \infty$. Let $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus \ldots \oplus \mathcal{H}_p$ denote the Hilbert space of functions of $(x_1, \ldots, x_p)$ that have the additive form: $m(x) = \sum_j f_j(x_j)$, with $f_j \in \mathcal{H}_j, j = 1, \ldots, p$.

Let $\{\psi_{jk}, k = 0, 1, \ldots\}$ denote a uniformly bounded, orthonormal basis with respect to $L^2[0,1]$. Unless stated otherwise, we assume that $f_j \in \mathcal{T}_j$ where

$$\mathcal{T}_j = \left\{ f_j \in \mathcal{H}_j : \ f_j(x_j) = \sum_{k=0}^{\infty} \beta_{jk}\psi_{jk}(x_j), \quad \sum_{k=0}^{\infty} \beta_{jk}^2 k^{2\nu_j} \leq C^2 \right\} \tag{6}$$

for some $0 < C < \infty$. We shall take $\nu_j = 2$ although the extension to other levels of smoothness is straightforward. It is also possible to adapt to ν_j although we do not pursue that direction here.

Let $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of a square matrix A. If $v = (v_1, \ldots, v_k)^T$ is a vector, we use the norms

$$\|v\| = \sqrt{\sum_{j=1}^{k} v_j^2}, \quad \|v\|_1 = \sum_{j=1}^{k} |v_j|, \quad \|v\|_\infty = \max_j |v_j|. \tag{7}$$

3. Sparse Backfitting. The outline of the derivation of our algorithm is as follows. We first formulate a population level optimization problem, and show that the minimizing functions can be obtained by iterating through a series of soft-thresholded univariate conditional expectations. We then plug in smoothed estimates of these univariate conditional expectations, to derive our sparse backfitting algorithm.

Population SpAM. For simplicity, assume that $\mathbb{E}(Y_i) = 0$. The standard additive model optimization problem in $L_2(\mu)$ (the population setting) is

$$\min_{f_j \in \mathcal{H}_j,\, 1 \leq j \leq p} \ \mathbb{E}\left(Y - \textstyle\sum_{j=1}^p f_j(X_j)\right)^2 \tag{8}$$

where the expectation is taken with respect to X and the noise ϵ. Now consider the following modification of this problem that introduces a scaling parameter for each function, and that imposes additional constraints:

$$\min_{\beta \in \mathbb{R}^p, g_j \in \mathcal{H}_j} \quad \mathbb{E}\left(Y - \textstyle\sum_{j=1}^p \beta_j g_j(X_j)\right)^2 \tag{9}$$

$$\text{subject to:} \quad \sum_{j=1}^{p} |\beta_j| \leq L, \tag{10}$$

$$\mathbb{E}\left(g_j^2\right) = 1, \ j = 1, \ldots, p. \tag{11}$$

noting that g_j is a function while $\beta = (\beta_1, \dots, \beta_p)^T$ is a vector. The constraint that β lies in the ℓ_1-ball $\{\beta : \|\beta\|_1 \leq L\}$ encourages sparsity of the estimated β, just as for the parametric lasso (Tibshirani, 1996). It is convenient to absorb the scaling constants β_j into the functions f_j, and re-express the minimization in the following equivalent Lagrangian form:

$$\mathcal{L}(f, \lambda) = \frac{1}{2}\mathbb{E}\left(Y - \textstyle\sum_{j=1}^{p} f_j(X_j)\right)^2 + \lambda \sum_{j=1}^{p} \sqrt{\mathbb{E}(f_j^2(X_j))}. \tag{12}$$

THEOREM 3.1. *The minimizers $f_j \in \mathcal{H}_j$ of (12) satisfy*

$$f_j = \left[1 - \frac{\lambda}{\sqrt{\mathbb{E}(P_j^2)}}\right]_+ P_j \qquad a.s. \tag{13}$$

where $[\cdot]_+$ denotes the positive part, and $P_j = \mathbb{E}[R_j \,|\, X_j]$ denotes the projection of the residual $R_j = Y - \sum_{k \neq j} f_k(X_k)$ onto $\mathcal{H}_j$.

An outline of the proof of this theorem appears in Ravikumar et al. (2008). A formal proof is given in Section 8. At the population level, the f_j's can be found by a coordinate descent procedure that fixes $(f_k : k \neq j)$ and fits f_j by equation (13), then iterates over j.

Data version of SpAM. To obtain a sample version of the population solution, we insert sample estimates into the population algorithm, as in standard backfitting (Hastie and Tibshirani, 1999). Thus, we estimate the projection $P_j = \mathbb{E}(R_j \,|\, X_j)$ by smoothing the residuals:

$$\widehat{P}_j = \mathcal{S}_j R_j \tag{14}$$

where $\mathcal{S}_j$ is a linear smoother, such as a local linear or kernel smoother. Let

$$\widehat{s}_j = \frac{1}{\sqrt{n}}\|\widehat{P}_j\| = \sqrt{\text{mean}(\widehat{P}_j^2)} \tag{15}$$

be the estimate of $\sqrt{\mathbb{E}(P_j^2)}$. Using these plug-in estimates in the coordinate descent procedure yields the SpAM backfitting algorithm given in Figure 1.

This algorithm can be seen as a functional version of the coordinate descent algorithm for solving the lasso. In particular, if we solve the lasso by iteratively minimizing with respect to a single coordinate, each iteration is given by soft thresholding; see Figure 2. Convergence properties of

SpAM Backfitting Algorithm

```
Input: Data $(X_i, Y_i)$, regularization parameter $\lambda$.
Initialize $\widehat{f}_j = 0$, for $j = 1, \ldots, p$.
Iterate until convergence:
    For each $j = 1, \ldots, p$:
    (1) Compute the residual: $R_j = Y - \sum_{k \neq j} \widehat{f}_k(X_k)$;
    (2) Estimate $P_j = \mathbb{E}[R_j \mid X_j]$ by smoothing: $\widehat{P}_j = \mathcal{S}_j R_j$;
    (3) Estimate norm: $\widehat{s}_j^2 = \frac{1}{n} \sum_{i=1}^n \widehat{P}_j^2(i)$;
    (4) Soft-threshold: $\widehat{f}_j = [1 - \lambda / \widehat{s}_j]_+ \widehat{P}_j$;
    (5) Center: $\widehat{f}_j \leftarrow \widehat{f}_j - \text{mean}(\widehat{f}_j)$.
Output: Component functions $\widehat{f}_j$ and estimator $\widehat{m}(X_i) = \sum_j \widehat{f}_j(X_{ij})$.
```

Fig 1. *The SpAM backfitting algorithm. The first two steps in the iterative algorithm are the usual backfitting procedure; the remaining steps carry out functional soft thresholding.*

variants of this simple algorithm have been recently treated by Daubechies et al. (2004, 2007). Our sparse backfitting algorithm is a direct generalization of this algorithm, and it reduces to it in case where the smoothers are local linear smoothers with large bandwidths. That is, as the bandwidth approaches infinity, the local linear smoother approaches a global linear fit, yielding the estimator $\widehat{P}_j(i) = \widehat{\beta}_j X_{ij}$. When the variables are standardized, $\widehat{s}_j = \sqrt{\frac{1}{n} \sum_{i=1}^n \widehat{\beta}_j^2 X_{ij}^2} = |\widehat{\beta}_j|$ so that the soft thresholding in step (4) of the SpAM backfitting algorithm is the same as the soft thresholding in step (3) in the coordinate descent lasso algorithm.

Basis Functions. It is useful to express the model in terms of basis functions. Recall that $B_j = (\psi_{jk} : \ k = 1, 2, \ldots)$ is an orthonormal basis for $\mathcal{T}_j$ and that $\sup_x |\psi_{jk}(x)| \leq B$ for some B. Then

$$f_j(x_j) = \sum_{k=1}^{\infty} \beta_{jk} \psi_{jk}(x_j) \tag{16}$$

where $\beta_{jk} = \int f_j(x_j) \psi_{jk}(x_j) dx_j$.

Let us also define

$$\widetilde{f}_j(x_j) = \sum_{k=1}^{d} \beta_{jk} \psi_{jk}(x_j) \tag{17}$$

Coordinate Descent Lasso

```
Input: Data $(X_i, Y_i)$, regularization parameter $\lambda$.
Initialize $\widehat{\beta}_j = 0$, for $j = 1, \dots, p$.
Iterate until convergence:
    For each $j = 1, \dots, p$:
    (1) Compute the residual: $R_j = Y - \sum_{k \neq j} \widehat{\beta}_k X_k$;
    (2) Project residual onto $X_j$: $P_j = X_j^T R_j$
    (3) Soft-threshold: $\widehat{\beta}_j = \left[1 - \lambda/|P_j|\right]_+ P_j$;
Output: Estimator $\widehat{m}(X_i) = \sum_j \widehat{\beta}_j X_{ij}$.
```

Fig 2. *The SpAM backfitting algorithm is a functional version of the coordinate descent algorithm for the lasso, which computes* $\widehat{\beta} = \arg\min \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$.

where $d = d_n$ is a truncation parameter. For the Sobolev space $\mathcal{T}_j$ of order two we have that $\left\|f_j - \widetilde{f}_j\right\|^2 = O(1/d^4)$. Let $S = \{j : \; f_j \neq 0\}$. Assuming the sparsity condition $|S| = O(1)$ it follows that $\|m - \widetilde{m}\|^2 = O(1/d^4)$ where $\widetilde{m} = \sum_j \widetilde{f}_j$. The usual choice is $d \asymp n^{1/5}$ yielding truncation bias $\|m - \widetilde{m}\|^2 = O(n^{-4/5})$.

In this setting, the smoother can be taken to be the least squares projection onto the truncated set of basis functions $\{\psi_{j1}, \dots, \psi_{jd}\}$; this is also called orthogonal series smoothing. Let Ψ_j denote the $n \times d_n$ matrix given by $\Psi_j(i, \ell) = \psi_{j,\ell}(X_{ij})$. The smoothing matrix is the projection matrix $\mathcal{S}_j = \Psi_j(\Psi_j^T \Psi_j)^{-1}\Psi_j^T$. In this case, the backfitting algorithm in Figure 1 is a coordinate descent algorithm for minimizing

$$\frac{1}{2n}\left\|Y - \textstyle\sum_{j=1}^p \Psi_j \beta_j\right\|_2^2 + \lambda \sum_{j=1}^p \sqrt{\frac{1}{n}\beta_j^T \Psi_j^T \Psi_j \beta_j}$$

which is the sample version of (12). This is the Lagrangian of a second-order cone program (SOCP), and standard convexity theory implies existence of a minimizer. In Section 6.1 we prove theoretical properties of SpAM assuming that this particular smoother is being used.

Connection with the Grouped Lasso. The SpAM model can be thought of as a functional version of the grouped lasso (Yuan and Lin, 2006) as we now explain. Consider the following linear regression model with multiple factors,

$$Y = \sum_{j=1}^{p_n} X_j \beta_j + \epsilon = X\beta + \epsilon, \tag{18}$$

where Y is an $n \times 1$ response vector, ϵ is an $n \times 1$ vector of iid mean zero noise, X_j is an $n \times d_j$ matrix corresponding to the j-th factor, and β_j is the corresponding $d_j \times 1$ coefficient vector. Assume for convenience (in this subsection only) that each X_j is orthogonal, so that $X_j^T X_j = I_{d_j}$, where I_{d_j} is the $d_j \times d_j$ identity matrix. We use $X = (X_1, \dots, X_{p_n})$ to denote the full design matrix and use $\beta = (\beta_1^T, \dots, \beta_{p_n}^T)^T$ to denote the parameter.

The *grouped lasso* estimator is defined as the solution of the following convex optimization problem:

$$\widehat{\beta}(\lambda_n) = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda_n \sum_{j=1}^{p_n} \sqrt{d_j} \|\beta_j\| \tag{19}$$

where $\sqrt{d_j}$ scales the jth term to compensate for different group sizes.

It is obvious that when $d_j = 1$ for $j = 1, \dots, p_n$, the grouped lasso becomes the standard lasso. From the KKT optimality conditions, a necessary and sufficient condition for $\widehat{\beta} = (\widehat{\beta}_1^T, \dots, \widehat{\beta}_p^T)^T$ to be the grouped lasso solution is

$$\begin{aligned} -X_j^T (Y - X\widehat{\beta}) + \frac{\lambda \sqrt{d_j} \widehat{\beta}_j}{\|\widehat{\beta}_j\|} &= \mathbf{0}, \quad \forall \widehat{\beta}_j \neq \mathbf{0}, \\ \|X_j^T (Y - X\widehat{\beta})\| &\leq \lambda \sqrt{d_j}, \quad \forall \widehat{\beta}_j = \mathbf{0}. \end{aligned} \tag{20}$$

Based on this stationary condition, an iterative blockwise coordinate descent algorithm can be derived; as shown by Yuan and Lin (2006), a solution to (20) satisfies

$$\widehat{\beta}_j = \left[1 - \frac{\lambda \sqrt{d_j}}{\|S_j\|} \right]_+ S_j \tag{21}$$

where $S_j = X_j^T (Y - X\beta_{\setminus j})$, with $\beta_{\setminus j} = (\beta_1^T, \dots, \beta_{j-1}^T, \mathbf{0}^T, \beta_{j+1}^T, \dots, \beta_{p_n}^T)$. By iteratively applying (21), the grouped lasso solution can be obtained.

As discussed in the introduction, the COSSO model of Lin and Zhang (2006) replaces the lasso constraint on $\sum_j |\beta_j|$ with a RKHS constraint. The advantage of our formulation is that it decouples smoothness ($g_j \in \mathcal{T}_j$) and sparsity ($\sum_j |\beta_j| \leq L$). This leads to a simple algorithm that can be carried out with any nonparametric smoother and scales easily to high dimensions.

4. Choosing the Regularization Parameter. We choose λ by minimizing an estimate of the risk. Let ν_j be the effective degrees of freedom for the smoother on the j^{th} variable, that is, $\nu_j = \text{trace}(\mathcal{S}_j)$ where $\mathcal{S}_j$ is the smoothing matrix for the j-th dimension. Also let $\widehat{\sigma}^2$ be an estimate of the variance. Define the total effective degrees of freedom as

$$\text{df}(\lambda) = \sum_j \nu_j I\left(\left\|\widehat{f}_j\right\| \neq 0\right). \tag{22}$$

Two estimates of risk are

$$C_p = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}\widehat{f}_j(X_j)\right)^2 + \frac{2\widehat{\sigma}^2}{n}\,\text{df}(\lambda) \tag{23}$$

and

$$\text{GCV}(\lambda) = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_j \widehat{f}_j(X_{ij}))^2}{(1-\text{df}(\lambda)/n)^2}. \tag{24}$$

The first is C_p and the second is generalized cross validation but with degrees of freedom defined by df(λ). A proof that these are valid estimates of risk is not currently available; thus, these should be regarded as heuristics.

Based on the results in Wasserman and Roeder (2007) about the lasso, it seems likely that choosing λ by risk estimation can lead to overfitting. One can further clean the estimate by testing $H_0 : f_j = 0$ for all j such that $\widehat{f}_j \neq 0$. For example, the tests in Fan and Jiang (2005) could be used.

5. Examples. To illustrate the method, we consider a few examples.

Synthetic Data. We generated $n = 100$ observations for an additive model with $p = 100$ and four relevant variables,

$$Y_i = \sum_{j=1}^{4} f_j(X_{ij}) + \epsilon_i,$$

where $\epsilon_i \sim N(0,1)$ and the relevant component functions are given by

$$f_1(x) = 2\exp(-2x) + 4\cos(x) + 2x^2 + 3x^3$$
$$f_2(x) = 2(x-1)^2 + \phi(x) + \sin(2\pi x) + \cos(2\pi x^2)$$
$$f_3(x) = \frac{\sin(2\pi x)}{2-\sin(2\pi x)} + 5\cdot\text{Beta}(x,2,3)$$
$$f_4(x) = \Phi\left(0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.2\sin^2(2\pi x) + 0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x)\right)^{-1}.$$

where ϕ and Φ are pdf and cdf for the standard Gaussian and Beta$(3,2)$ is the pdf function for a Beta distribution with parameters 3 and 2. These data therefore have 96 irrelevant dimensions. The covariates are generated as

$$X_j = (W_j + tU)/(1+t), j = 1, \ldots, 100$$

where $W_1, \ldots, W_{100}$ and U are i.i.d. sampled from Uniform$(-2.5, 2.5)$. Thus, the correlation between X_j and $X_{j'}$ is $t^2/(1+t^2)$ for $j \neq j'$. In the illustrations below, we set $t = 1$.

The results of applying SpAM with the plug-in bandwidths are summarized in Figure 3. The top-left plot in Figure 3 shows regularization paths as the parameter λ varies; each curve is a plot of $\|\widehat{f}_j(\lambda)\|$ versus

$$\frac{\sum_{k=1}^{p} \|\widehat{f}_k(\lambda)\|}{\max_\lambda \sum_{k=1}^{p} \|\widehat{f}_k(\lambda)\|} \tag{25}$$

for a particular variable X_j. The estimates are generated efficiently over a sequence of λ values by "warm starting" $\widehat{f}_j(\lambda_t)$ at the previous value $\widehat{f}_j(\lambda_{t-1})$. The top-right plot shows the C_p statistic as a function of regularization level.

Functional Sparse Coding. Olshausen and Field (1996) propose a method of obtaining sparse representations of data such as natural images; the motivation comes from trying to understand principles of neural coding. In this example we suggest a nonparametric form of sparse coding.

Let $\{y^i\}_{i=1,\ldots,N}$ be the data to be represented with respect to some learned basis, where each instance $y^i \in \mathbb{R}^n$ is an n-dimensional vector. The linear sparse coding optimization problem is

$$\min_{\beta, X} \quad \sum_{i=1}^{N} \left\{ \frac{1}{2n} \left\| y^i - X\beta^i \right\|^2 + \lambda \left\| \beta^i \right\|_1 \right\} \tag{26}$$

$$\text{such that} \quad \|X_j\| \leq 1 \tag{27}$$

Here X is an $n \times p$ matrix with columns X_j, representing the "dictionary" entries or basis vectors to be learned. It is not required that the basis vectors are orthogonal. The ℓ_1 penalty on the coefficients β^i encourages sparsity, so that each data vector y^i is represented by only a small number of dictionary elements. Sparsity allows the features to specialize, and to capture salient properties of the data.

This optimization problem is not jointly convex in β^i and X. However, for fixed X, each weight vector β^i is computed by running the lasso. For fixed β^i, the optimization is similar to ridge regression, and can be solved efficiently.

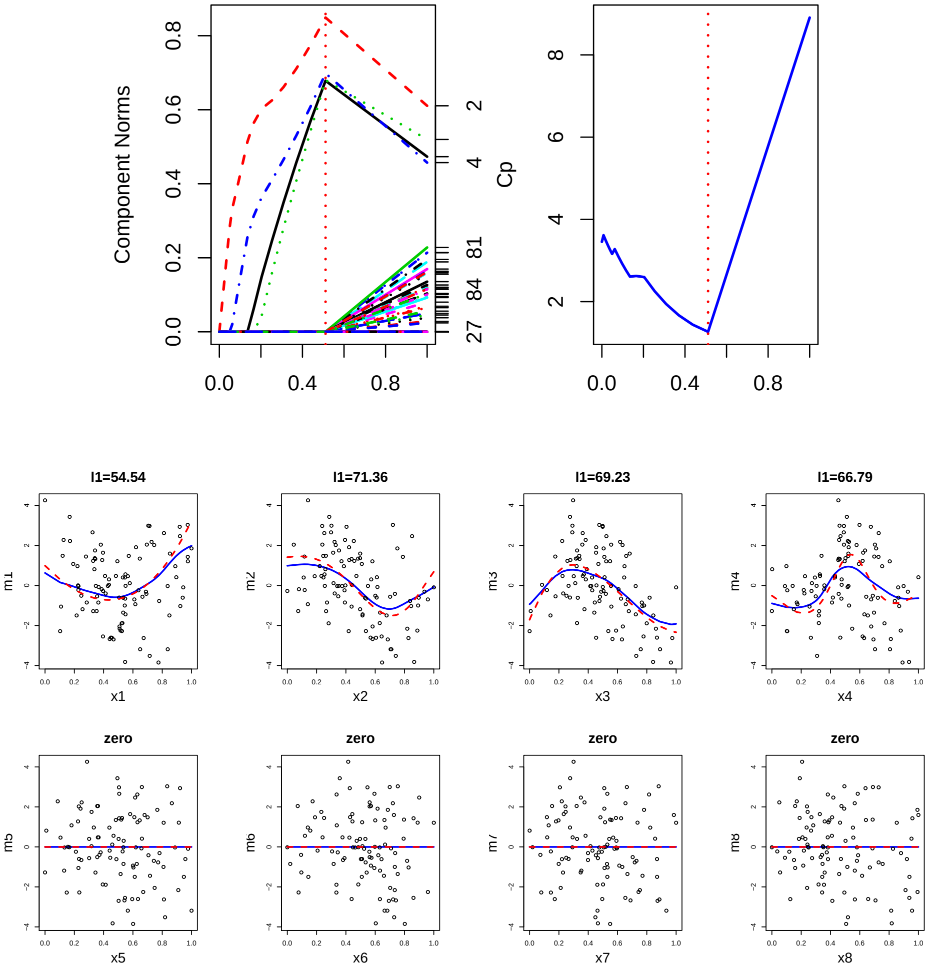

FIG 3. *(Simulated data) Upper left: The empirical ℓ_2 norm of the estimated components as plotted against the regularization parameter λ; the value on the x-axis is proportional to $\sum_j \|\widehat{f_j}\|$. Upper right: The C_p scores against the amount of regularization; the dashed vertical line corresponds to the value of λ which has the smallest C_p score. Lower two rows: Estimated (solid lines) versus true additive component functions (dashed lines) for the first four relevant dimensions, and the first four irrelevant dimensions; the remaining components are zero.*

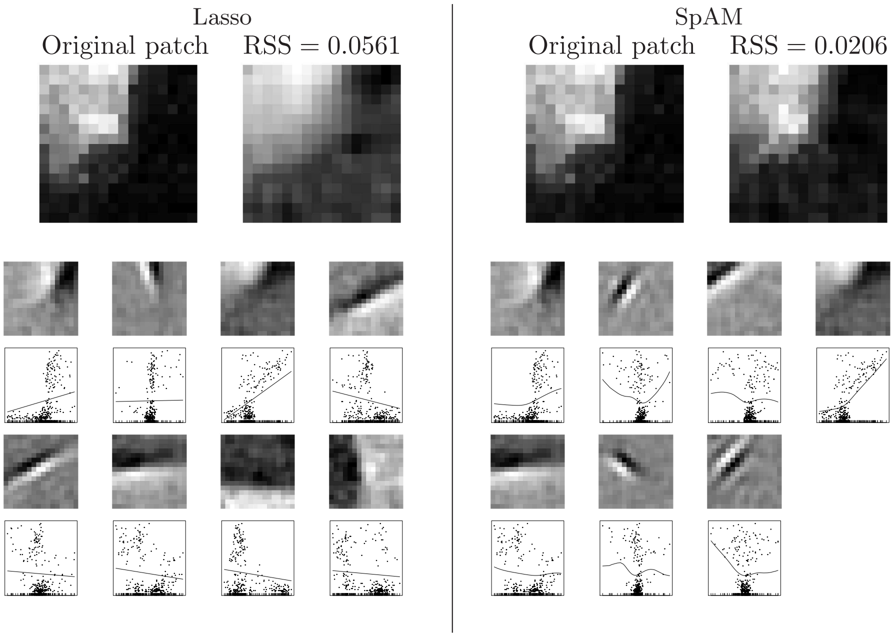

FIG 4. *Comparison of sparse reconstruction using the lasso (left) and SpAM (right).*

Thus, an iterative procedure for (approximately) solving this optimization problem is easy to derive.

In the case of sparse coding of natural images, as in Olshausen and Field (1996), the basis vectors X_j encode basic edge features at different scales and spatial orientations. In the functional version, we no longer assume a linear, parametric fit between the dictionary X and the data y. Instead, we model the relationship using an additive model. This leads to the following optimization problem for functional sparse coding:

$$\min_{f,X} \quad \sum_{i=1}^{N} \left\{ \frac{1}{2n} \left\| y^i - \textstyle\sum_{j=1}^{p} f_j^i(X_j) \right\|^2 + \lambda \sum_{j=1}^{p} \left\| f_j^i \right\| \right\} \tag{28}$$

$$\text{such that} \quad \|X_j\| \leq 1,\ j = 1, \ldots, p. \tag{29}$$

Figure 4 illustrates the reconstruction of different image patches using the sparse linear model compared with the sparse additive model. Local linear smoothing was used with a Gaussian kernel having fixed bandwidth $h = 0.05$ for all patches and all codewords. The codewords X_j are those obtained using the Olshausen-Field procedure; these become the design points in the regression estimators. Thus, a codeword for a 16×16 patch corresponds to a vector X_j of dimension 256, with each X_{ij} the gray level for a particular pixel.

6. Theoretical Properties.

6.1. *Sparsistency.* In the case of linear regression, with $f_j(X_j) = \beta_j^{*T} X_j$, several authors have shown that, under certain conditions on n, p, the number of relevant variables $s = |\mathrm{supp}(\beta^*)|$, and the design matrix X, the lasso recovers the sparsity pattern asymptotically; that is, the lasso estimator $\widehat{\beta}_n$ is *sparsistent*:

$$\mathbb{P}\left(\mathrm{supp}(\beta^*) = \mathrm{supp}(\widehat{\beta}_n)\right) \to 1. \tag{30}$$

Here, $\mathrm{supp}(\beta) = \{j : \ \beta_j \neq 0\}$. References include Wainwright (2006), Meinshausen and Bühlmann (2006), Zou (2005), Fan and Li (2001), and Zhao and Yu (2007). We show a similar result for sparse additive models under orthogonal function regression.

In terms of an orthogonal basis ψ, we can write

$$Y_i = \sum_{j=1}^{p} \sum_{k=1}^{\infty} \beta_{jk}^* \psi_{jk}(X_{ij}) + \epsilon_i. \tag{31}$$

To simplify notation, let β_j be the d_n dimensional vector $\{\beta_{jk}, k = 1, \ldots, d_n\}$ and let Ψ_j be the $n \times d_n$ matrix $\Psi_j[i,k] = \psi_{jk}(X_{ij})$. If $A \subset \{1, \ldots, p\}$, we denote by Ψ_A the $n \times d|A|$ matrix where for each $j \in A$, Ψ_j appears as a submatrix in the natural way.

We now analyze the sparse backfitting algorithm of Figure 1 assuming an orthogonal series smoother is used to estimate the conditional expectation in its Step (2). As noted earlier, an orthogonal series smoother for a predictor X_j is the least squares projection onto a truncated set of basis functions $\{\psi_{j1}, \ldots, \psi_{jd}\}$. Our optimization problem in this setting is

$$\min_{\beta} \frac{1}{2n} \left\| Y - \textstyle\sum_{j=1}^{p} \Psi_j \beta_j \right\|_2^2 + \lambda \sum_{j=1}^{p} \sqrt{\frac{1}{n} \beta_j^T \Psi_j^T \Psi_j \beta_j}. \tag{32}$$

Combined with the soft-thresholding step, the update for f_j in algorithm of Figure 1 can thus be seen to solve the following problem,

$$\min_{\beta} \frac{1}{2n} \|R_j - \Psi_j \beta_j\|_2^2 + \lambda_n \sqrt{\frac{1}{n} \beta_j^T \Psi_j^T \Psi_j \beta_j}$$

where $\|v\|_2^2$ denotes $\sum_{i=1}^n v_i^2$ and $R_j = Y - \sum_{l \neq j} \Psi_l \beta_l$ is the residual for f_j. The sparse backfitting algorithm thus solves

$$\begin{aligned}(33) \qquad \min_\beta \left\{ R_n(\beta) + \lambda_n \Omega(\beta) \right\} \quad &= \quad \min_\beta \frac{1}{2n} \left\| Y - \sum_{j=1}^p \Psi_j \beta_j \right\|_2^2 \\ &\qquad + \lambda_n \sum_{j=1}^p \left\| \frac{1}{\sqrt{n}} \Psi_j \beta_j \right\|_2\end{aligned}$$

where R_n denotes the squared error term and Ω denotes the regularization term, and each β_j is a d_n-dimensional vector. Let S denote the true set of variables $\{j : f_j \neq 0\}$, with $s = |S|$, and let S^c denote its complement. Let $\widehat{S}_n = \{j : \widehat{\beta}_j \neq 0\}$ denote the estimated set of variables from the minimizer $\widehat{\beta}_n$, with corresponding function estimates $\widehat{f}_j(x_j) = \sum_{k=1}^{d_n} \widehat{\beta}_{jk} \psi_{jk}(x_j)$. For the results in this section, we will treat the covariates as fixed. A preliminary version of the following result is stated, without proof, in Ravikumar et al. (2008).

THEOREM 6.1. *Suppose that the following conditions hold on the design matrix X in the orthogonal basis ψ:*

$$(34) \qquad \Lambda_{\max}\left(\frac{1}{n}\Psi_S^T \Psi_S\right) \;\leq\; C_{\max} \;<\; \infty$$

$$(35) \qquad \Lambda_{\min}\left(\frac{1}{n}\Psi_S^T \Psi_S\right) \;\geq\; C_{\min} \;>\; 0$$

$$(36) \qquad \max_{j \in S^c} \left\| \left(\tfrac{1}{n}\Psi_j^T \Psi_S\right) \left(\tfrac{1}{n}\Psi_S^T \Psi_S\right)^{-1} \right\| \;\leq\; \sqrt{\frac{C_{\min}}{C_{\max}}} \frac{1-\delta}{\sqrt{s}}, \quad \textit{for some } 0 < \delta \leq 1.$$

Assume that the truncation dimension d_n satisfies $d_n \to \infty$ and $d_n = o(n)$. Furthermore, suppose the following conditions, which relate the regularization parameter λ_n to the design parameters n, p, the number of relevant variables s, and the truncation size d_n:

$$(37) \qquad \frac{s}{d_n \lambda_n} \;\longrightarrow\; 0$$

$$(38) \qquad \frac{d_n \log\left(d_n(p-s)\right)}{n\lambda_n^2} \;\longrightarrow\; 0$$

$$(39) \qquad \frac{1}{\rho_n^*}\left(\sqrt{\frac{\log(s d_n)}{n}} + \frac{s^{3/2}}{d_n} + \lambda_n \sqrt{s d_n} \right) \;\longrightarrow\; 0$$

where $\rho_n^* = \min_{j\in S} \|\beta_j^*\|_\infty$. *Then the solution* $\widehat{\beta}_n$ *to* (32) *is unique and satisfies* $\widehat{S}_n = S$ *with probability approaching one.*

This result parallels the theorem of Wainwright (2006) on model selection consistency of the lasso; however, technical subtleties arise because of the truncation dimension d_n which is increasing with sample size, and the matrix $\Psi_j^T\Psi$ which appears in the regularization of β_j. As a result, the operator norm rather than the ∞-norm appears in the incoherence condition (36). Note, however, that condition (36) implies that

$$\left\|\Psi_{S^c}^T\Psi_S\left(\Psi_S^T\Psi_S\right)^{-1}\right\|_\infty = \max_{j\in S^c}\left\|\Psi_j^T\Psi_S\left(\Psi_S^T\Psi_S\right)^{-1}\right\|_\infty \tag{40}$$

$$\leq \sqrt{\frac{C_{\min}\, d_n}{C_{\max}}}\,(1-\delta) \tag{41}$$

since $\frac{1}{\sqrt{n}}\|A\|_\infty \leq \|A\| \leq \sqrt{m}\,\|A\|_\infty$ for an $m\times n$ matrix A. This relates it to the more standard incoherence conditions that have been used for sparsistency in the case of the lasso.

The following corollary, which imposes the additional condition that the number of relevant variables is bounded, follows directly. It makes explicit how to choose the design parameters d_n and λ_n, and implies a condition on the fastest rate at which the minimum norm ρ_n^* can approach zero.

Corollary 6.2. *Suppose that* $s = O(1)$, *and assume the design conditions* (34), (35) *and* (36) *hold. If the truncation dimension* d_n, *regularization parameter* λ_n, *and minimum norm* ρ_n^* *satisfy*

$$d_n \asymp n^{1/3} \tag{42}$$

$$\lambda_n \asymp \frac{\log np}{n^{1/3}} \tag{43}$$

$$\frac{1}{\rho_n^*} = o\left(\frac{n^{1/6}}{\log np}\right) \tag{44}$$

then $\mathbb{P}\left(\widehat{S}_n = S\right)\to 1$.

The following proposition clarifies the implications of condition (44), by relating the sup-norm $\|\beta_j\|_\infty$ to the function norm $\|f_j\|_2$.

Proposition 6.3. *Suppose that* $f(x) = \sum_k \beta_k\psi_k(x)$ *is in the Sobolev space of order* $\nu > 1/2$, *so that* $\sum_{i=1}^\infty \beta_i^2 i^{2\nu} \leq C^2$ *for some constant* C. *Then*

$$\|f\|_2 = \|\beta\|_2 \leq c\|\beta\|_\infty^{\frac{2\nu}{2\nu+1}} \tag{45}$$

for some constant c.

For instance, the result of Corollary 6.2 allows the norms of the coefficients β_j to decrease as $\|\beta_j\|_\infty = \log^2(np)/n^{1/6}$. In the case $\nu = 2$, this would allow the norms $\|f_j\|_2$ of the relevant functions to approach zero at the rate $\log^{8/5}(np)/n^{2/15}$.

6.2. *Persistence.* The previous assumptions are very strong. They can be weakened at the expense of getting weaker results. In particular, in the section we do not assume that the true regression function is additive. We use arguments like those in Juditsky and Nemirovski (2000) and Greenshtein and Ritov (2004) in the context of linear models. In this section we treat X as random and we use triangular array asymptotics, that is, the joint distribution for the data can change with n. Let (X, Y) denote a new pair (independent of the observed data) and define the predictive risk when predicting Y with $v(X)$ by

$$R(v) = \mathbb{E}(Y - v(X))^2. \tag{46}$$

When $v(x) = \sum_j \beta_j g_j(x_j)$ we also write the risk as $R(\beta, g)$ where $\beta = (\beta_1, \ldots, \beta_p)$ and $g = (g_1, \ldots, g_p)$. Following Greenshtein and Ritov (2004) we say that an estimator $\widehat{m}_n$ is persistent (risk consistent) relative to a class of functions $\mathcal{M}_n$, if

$$R(\widehat{m}_n) - R(m_n^*) \xrightarrow{P} 0 \tag{47}$$

where

$$m_n^* = \underset{v \in \mathcal{M}_n}{\arg\min}\, R(v) \tag{48}$$

is the predictive oracle. Greenshtein and Ritov (2004) show that the lasso is persistent for $\mathcal{M}_n = \{\ell(x) = x^T\beta : \ \|\beta\|_1 \leq L_n\}$ and $L_n = o((n/\log n)^{1/4})$. Note that m_n^* is the best linear approximation (in prediction risk) in $\mathcal{M}_n$ but the true regression function is not assumed to be linear. Here we show a similar result for SpAM.

In this section, we assume that the SpAM estimator $\widehat{m}_n$ is chosen to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_j \beta_j g_j(X_{ij}))^2 \tag{49}$$

subject to $\|\beta\|_1 \leq L_n$ and $g_j \in \mathcal{T}_j$. We make no assumptions about the design matrix. Let $\mathcal{M}_n \equiv \mathcal{M}_n(L_n)$ be defined by

(50)
$$\mathcal{M}_n = \left\{ m : \; m(x) = \sum_{j=1}^{p_n} \beta_j g_j(x_j) : \; \mathbb{E}(g_j) = 0, \; \mathbb{E}(g_j^2) = 1, \; \sum_j |\beta_j| \leq L_n \right\}$$

and let $m_n^* = \arg\min_{v \in \mathcal{M}_n} R(v)$.

THEOREM 6.4. *Suppose that $p_n \leq e^{n^\xi}$ for some $\xi < 1$. Then,*

$$R(\widehat{m}_n) - R(m_n^*) = O_P\left(\frac{L_n^2}{n^{(1-\xi)/2}} \right) \tag{51}$$

and hence , if $L_n = o(n^{(1-\xi)/4})$ then SpAM is persistent.

7. Discussion.

The results presented here show how many of the recently established theoretical properties of ℓ_1 regularization for linear models extend to sparse additive models. The sparse backfitting algorithm we have derived is attractive because it decouples smoothing and sparsity, and can be used with any nonparametric smoother. It thus inherits the nice properties of the original backfitting procedure. However, our theoretical analyses have made use of a particular form of smoothing, using a truncated orthogonal basis. An important problem is thus to extend the theory to cover more general classes of smoothing operators. Convergence properties of the SpAM backfitting algorithm should also be investigated; convergence of special cases of standard backfitting is studied by Buja et al. (1989).

An additional direction for future work is to develop procedures for automatic bandwidth selection in each dimension. We have used plug-in bandwidths and truncation dimensions d_n in our experiments and theory. It is of particular interest to develop procedures that are adaptive to different levels of smoothness in different dimensions. It would also be of interest is to consider more general penalties of the form $p_\lambda(\|f_j\|)$, as in Fan and Li (2001).

Finally, we note that while we have considered basic additive models that allow functions of individual variables, it is natural to consider interactions, as in the functional ANOVA model. One challenge is to formulate suitable incoherence conditions on the functions that enable regularization based procedures or greedy algorithms to recover the correct interaction graph. In the parametric setting, one result in this direction is Wainwright et al. (2007).

8. Proofs.

Proof of THEOREM 3.1. Consider the minimization of the Lagrangian

$$\min_{\{f_j\in\mathcal{H}_j\}} \mathcal{L}(f,\lambda) \equiv \frac{1}{2}\mathbb{E}\left(Y - \textstyle\sum_{j=1}^p f_j(X_j)\right)^2 + \lambda\sum_{j=1}^p \sqrt{\mathbb{E}(f_j(X_j)^2)} \tag{52}$$

with respect to $f_j \in \mathcal{H}_j$, holding the other components $\{f_k,\, k \neq j\}$ fixed. The stationary condition is obtained by setting the Fréchet derivative to zero. Denote by $\partial_j\mathcal{L}(f,\lambda;\eta_j)$ the directional derivative with respect to f_j in the direction $\eta_j(X_j) \in \mathcal{H}_j$ ($\mathbb{E}(\eta_j) = 0$, $\mathbb{E}(\eta_j^2) < \infty$). Then the stationary condition can be formulated as

$$\partial_j\mathcal{L}(f,\lambda;\eta_j) = \frac{1}{2}\mathbb{E}\left[(f_j - R_j + \lambda v_j)\;\eta_j\right] = 0 \tag{53}$$

where $R_j = Y - \sum_{k\neq j} f_k$ is the residual for f_j, and $v_j \in \mathcal{H}_j$ is an element of the subgradient $\partial\sqrt{\mathbb{E}(f_j^2)}$, satisfying $v_j = f_j/\sqrt{\mathbb{E}(f_j^2)}$ if $\mathbb{E}(f_j^2) \neq 0$ and $v_j \in \{u_j \in \mathcal{H}_j |\; \mathbb{E}(u_j^2) \leq 1\}$ otherwise.

Using iterated expectations, the above condition can be rewritten as

$$\mathbb{E}\left[(f_j + \lambda v_j - \mathbb{E}(R_j|X_j))\;\eta_j\right] = 0. \tag{54}$$

But since $f_j - \mathbb{E}(R_j|X_j) + \lambda v_j \in \mathcal{H}_j$, we can compute the derivative in the direction $\eta_j = f_j - \mathbb{E}(R_j|X_j) + \lambda v_j \in \mathcal{H}_j$, implying that

$$\mathbb{E}\left[\left(f_j(x_j) - \mathbb{E}(R_j|X_j = x_j) + \lambda v_j(x_j)\right)^2\right] = 0; \tag{55}$$

that is,

$$f_j + \lambda v_j = \mathbb{E}(R_j|X_j) \quad \text{a.e.} \tag{56}$$

Denote the conditional expectation $\mathbb{E}(R_j|X_j)$—also the projection of the residual R_j onto $\mathcal{H}_j$—by P_j. Now if $\mathbb{E}(f_j^2) \neq 0$, then $v_j = \frac{f_j}{\sqrt{\mathbb{E}(f_j^2)}}$, which from condition (56) implies

$$\begin{aligned}
\sqrt{\mathbb{E}(P_j^2)} &= \sqrt{\mathbb{E}\left[\left(f_j + \lambda f_j/\sqrt{\mathbb{E}(f_j^2)}\right)^2\right]} && (57)\\
&= \left(1 + \frac{\lambda}{\sqrt{\mathbb{E}(f_j^2)}}\right)\sqrt{\mathbb{E}(f_j^2)} && (58)\\
&= \sqrt{\mathbb{E}(f_j^2)} + \lambda && (59)\\
&\geq \lambda. && (60)
\end{aligned}$$

If $\mathbb{E}(f_j^2) = 0$, then $f_j = 0$ a.e., and $\sqrt{\mathbb{E}(v_j^2)} \le 1$. Equation (56) then implies that

$$\sqrt{\mathbb{E}(P_j^2)} \;\; \le \;\; \lambda. \tag{61}$$

We thus obtain the equivalence

$$\sqrt{\mathbb{E}(P_j^2)} \le \lambda \;\Leftrightarrow\; f_j = 0 \quad \text{a.e.} \tag{62}$$

Rewriting equation (56) in light of (62), we obtain

$$\begin{aligned} \left(1 + \frac{\lambda}{\sqrt{\mathbb{E}(f_j^2)}}\right) f_j = P_j & \qquad \text{if } \sqrt{\mathbb{E}(P_j^2)} > \lambda \\ f_j = 0 & \qquad \text{otherwise.} \end{aligned}$$

Using (59), we thus arrive at the soft thresholding update for f_j:

$$f_j = \left[1 - \frac{\lambda}{\sqrt{\mathbb{E}(P_j^2)}}\right]_+ P_j \tag{63}$$

where $[\cdot]_+$ denotes the positive part and $P_j = \mathbb{E}[R_j \mid X_j]$. □

Proof of THEOREM 6.1. A vector $\widehat{\beta} \in \mathbb{R}^{d_n p}$ is an optimum of the objective function in (33) if and only if there exists a subgradient $\widehat{g} \in \partial\Omega(\widehat{\beta})$, such that

$$\frac{1}{n}\Psi^\top \left(\sum_j \Psi_j \widehat{\beta}_j - Y\right) + \lambda_n \widehat{g} = 0. \tag{64}$$

The subdifferential $\partial\Omega(\beta)$ is the set of vectors $g \in \mathbb{R}^{pd_n}$ satisfying

$$\begin{aligned} g_j \;\; &= \;\; \frac{\frac{1}{n}\Psi_j^T \Psi_j \beta_j}{\sqrt{\frac{1}{n}\beta_j^T \Psi_j^T \Psi_j \beta_j}} && \text{if } \beta_j \neq 0, \\ g_j^T \left(\frac{1}{n}\Psi_j^T \Psi_j\right)^{-1} g_j \;\; &\le \;\; 1 && \text{if } \beta_j = 0. \end{aligned}$$

Our argument is based on the technique of a *primal-dual witness*, used previously in the analysis of the Lasso (Wainwright, 2006). In particular, we construct a coefficient-subgradient pair $(\widehat{\beta}, \widehat{g})$ which satisfies $\text{supp}(\widehat{\beta}) =$

$\mathrm{supp}(\beta^*)$, and in addition satisfies the optimality conditions for the objective (33) with high probability. Thus, when the procedure succeeds, the constructed coefficient vector $\widehat{\beta}$ is equal to the solution of the convex objective (33), and $\widehat{g}$ is an optimal solution to its dual. From its construction, the support of $\widehat{\beta}$ is equal to the true support $\mathrm{supp}(\beta^*)$, from which we can conclude that the solution of the objective (33) is sparsistent. The construction of the primal-dual witness proceeds as follows:

(a) Set $\widehat{\beta}_{S^c} = 0$.
(b) Set $\widehat{g}_S = \partial\Omega(\beta^*)_S$.
(c) With these settings of $\widehat{\beta}_{S^c}$ and $\widehat{g}_S$, obtain $\widehat{\beta}_S$ and $\widehat{g}_{S^c}$ from the stationary conditions in (64).

For the witness procedure to succeed, we have to show that $(\widehat{\beta}, \widehat{g})$ is optimal for the objective (33), meaning that

$$\widehat{\beta}_j \neq 0 \text{ for } j \in S. \tag{65a}$$

$$g_j^T \left(\frac{1}{n}\Psi_j^T \Psi_j\right)^{-1} g_j < 1 \text{ for } j \in S^c. \tag{65b}$$

For uniqueness of the solution, we require strict dual feasibility, meaning strict inequality in (65b). In what follows, we show these two conditions hold with high probability.

Condition (65a). Setting $\widehat{\beta}_{S^c} = 0$ and $\widehat{g}_j = \frac{\frac{1}{n}\Psi_j^T\Psi_j\beta_j^*}{\sqrt{\frac{1}{n}\beta_j^{*T}\Psi_j^T\Psi_j\beta_j^*}}$ for $j \in S$, the stationary condition for $\widehat{\beta}_S$ is given by,

$$\frac{1}{n}\Psi_S^\top\left(\Psi_S\widehat{\beta}_S - Y\right) + \lambda_n \widehat{g}_S = 0. \tag{66}$$

Let $V = Y - \Psi_S\beta_S^* - W$ denote the error due to finite truncation of the orthogonal basis, where $W = (\epsilon_1, \ldots, \epsilon_n)^T$. Then the stationary condition (66) can be simplified as,

$$\tfrac{1}{n}\Psi_S^T\Psi_S\left(\widehat{\beta}_S - \beta_S^*\right) - \tfrac{1}{n}\Psi_S^T W - \tfrac{1}{n}\Psi_S^T V + \lambda_n\widehat{g}_S = 0, \quad \text{so that,}$$

$$\widehat{\beta}_S - \beta_S^* = \left(\tfrac{1}{n}\Psi_S^T\Psi_S\right)^{-1}\left(\tfrac{1}{n}\Psi_S^T W + \tfrac{1}{n}\Psi_S^T V - \lambda_n\widehat{g}_S\right), \tag{67}$$

where we have used the assumption that $\frac{1}{n}\Psi_S^T\Psi_S$ is nonsingular. Recalling our definition of the minimum function norm $\rho_n^* = \min_{j\in S}\|\beta_j^*\|_\infty > 0$, it suffices to show that $\|\widehat{\beta}_S - \beta_S^*\|_\infty < \frac{\rho_n^*}{2}$, in order to ensure that

$$\mathrm{supp}(\beta_S^*) = \mathrm{supp}(\widehat{\beta}_S) = \left\{j \,:\, \|\widehat{\beta}_j\|_\infty \neq 0\right\},$$

so that condition (65a) would be satisfied. Using $\Sigma_{SS} = \frac{1}{n}\left(\Psi_S^\top \Psi_S\right)$ to simplify notation, we have the ℓ_∞ bound,

(68)
$$\|\widehat{\beta}_S - \beta^*_S\|_\infty \; \le \; \underbrace{\left\|\Sigma_{SS}^{-1}\left(\tfrac{1}{n}\Psi_S^\top W\right)\right\|_\infty}_{T_1} + \underbrace{\left\|\Sigma_{SS}^{-1}\left(\tfrac{1}{n}\Psi_S^\top V\right)\right\|_\infty}_{T_2} + \lambda_n \; \underbrace{\left\|\Sigma_{SS}^{-1}\widehat{g}_S\right\|_\infty}_{T_3}.$$

We now proceed to bound the quantities T_1, T_2, T_3.

Bounding T_3. Note that for $j \in S$,

$$1 = g_j^T\left(\tfrac{1}{n}\Psi_j^T\Psi_j\right)^{-1} g_j \ge \frac{1}{C_{\max}}\|g_j\|^2,$$

and thus $\|g_j\| \le \sqrt{C_{\max}}$. Noting further that,

$$\|g_S\|_\infty = \max_{j\in S}\|g_j\|_\infty \le \max_{j\in S}\|g_j\|_2 \le \sqrt{C_{\max}}, \tag{69}$$

it follows that,

$$T_3 := \left\|\Sigma_{SS}^{-1}\widehat{g}_S\right\|_\infty \le \sqrt{C_{\max}}\left\|\Sigma_{SS}^{-1}\right\|_\infty. \tag{70}$$

Bounding T_2. We proceed in two steps; we first bound $\|V\|_\infty$, and use this to bound $\left\|\frac{1}{n}\Psi_S^T V\right\|_\infty$. Note that, as we are working over the Sobolev spaces $\mathcal{S}_j$ of order two,

$$\begin{aligned}
|V_i| \;&=\; \left|\sum_{j\in S}\sum_{k=d_n+1}^{\infty}\beta^*_{jk}\Psi_{jk}(X_{ij})\right| \;\le\; B\sum_{j\in S}\sum_{k=d_n+1}^{\infty}\left|\beta^*_{jk}\right| \\
&=\; B\sum_{j\in S}\sum_{k=d_n+1}^{\infty}\frac{\left|\beta^*_{jk}\right|k^2}{k^2} \;\le\; B\sum_{j\in S}\sqrt{\sum_{k=d_n+1}^{\infty}\beta^{*2}_{jk}k^4}\sqrt{\sum_{k=d_n+1}^{\infty}\frac{1}{k^4}} \\
&\le\; sBC\sqrt{\sum_{k=d_n+1}^{\infty}\frac{1}{k^4}} \;\le\; \frac{sB'}{d_n^{3/2}},
\end{aligned}$$

for some constant $B' > 0$. It follows that,

$$\left|\frac{1}{n}\Psi_{jk}^\top V\right| \le \left|\frac{1}{n}\sum_i \Psi_{jk}(X_{ij})\right| \|V\|_\infty \le \frac{Ds}{d_n^{3/2}}, \tag{71}$$

where D denotes a generic constant. Thus,

$$T_2 := \left\|\Sigma_{SS}^{-1}\left(\tfrac{1}{n}\Psi_S^\top V\right)\right\|_\infty \le \left\|\Sigma_{SS}^{-1}\right\|_\infty \frac{Ds}{d_n^{3/2}} \tag{72}$$

Bounding T_1. Let $Z = T_1 = \Sigma_{SS}^{-1}\left(\frac{1}{n}\Psi_S^\top W\right)$. Note that $W \sim N(0, \sigma^2 I)$, so that Z is Gaussian as well, with mean zero. Consider its l-th component, $Z_l = e_l^\top Z$. Then $\mathbb{E}[Z_l] = 0$, and

$$\operatorname{Var}(Z_l) = \frac{\sigma^2}{n} e_l^\top \Sigma_{SS}^{-1} e_l \leq \frac{\sigma^2}{C_{\min} n}.$$

By Gaussian comparison results (Ledoux and Talagrand, 1991), we have then that

$$\mathbb{E}\left[\|Z\|_\infty\right] \;\leq\; 3\sqrt{\log(sd_n)\,\|\operatorname{Var}(Z)\|_\infty} \;\leq\; 3\sigma\sqrt{\frac{\log(sd_n)}{nC_{\min}}}. \tag{73}$$

Substituting the bounds for T_2, T_3 from equations (72),(70) respectively into equation (68), and using the bound for the expected value of T_1 from (73), it follows from an application of Markov's inequality that,

$$\begin{aligned}
&\mathbb{P}\left(\|\widehat{\beta}_S - \beta_S^*\|_\infty > \frac{\rho_n^*}{2}\right) \\
&\quad\leq\; \mathbb{P}\left(\|Z\|_\infty + \left\|\Sigma_{SS}^{-1}\right\|_\infty \left(Dsd_n^{-3/2} + \lambda_n\sqrt{C_{\max}}\right) > \frac{\rho_n^*}{2}\right) \\
&\quad\leq\; \frac{2}{\rho_n^*}\left\{\mathbb{E}\left[\|Z\|_\infty\right] + \left\|\Sigma_{SS}^{-1}\right\|_\infty \left(Dsd_n^{-3/2} + \lambda_n\sqrt{C_{\max}}\right)\right\} \\
&\quad\leq\; \frac{2}{\rho_n^*}\left\{3\sigma\sqrt{\frac{\log(sd_n)}{nC_{\min}}} + \left\|\Sigma_{SS}^{-1}\right\|_\infty \left(\frac{Ds}{d_n^{3/2}} + \lambda_n\sqrt{C_{\max}}\right)\right\},
\end{aligned}$$

which converges to zero under the condition that

$$\frac{1}{\rho_n^*}\left\{\sqrt{\frac{\log(sd_n)}{n}} + \left\|\left(\tfrac{1}{n}\Psi_S^T\Psi_S\right)^{-1}\right\|_\infty \left(\frac{s}{d_n^{3/2}} + \lambda_n\right)\right\} \longrightarrow 0. \tag{74}$$

Noting that

$$\left\|\left(\tfrac{1}{n}\Psi_S^T\Psi_S\right)^{-1}\right\|_\infty \;\leq\; \frac{\sqrt{sd_n}}{C_{\min}}, \tag{75}$$

it follows that condition (74) holds when

$$\frac{1}{\rho_n^*}\left(\sqrt{\frac{\log(sd_n)}{n}} + \frac{s^{3/2}}{d_n} + \lambda_n\sqrt{sd_n}\right) \;\longrightarrow\; 0. \tag{76}$$

But this is satisfied by assumption (39) in the theorem. We have thus shown that condition (65a) is satisfied with probability converging to one.

Condition (65b). We now have to consider the dual variables $\widehat{g}_{S^c}$. Recall that we have set $\widehat{\beta}_{S^c} = \beta^*_{S^c} = 0$. The stationary condition for $j \in S^c$ is thus given by

$$\frac{1}{n}\Psi_j^\top \left(\Psi_S\widehat{\beta}_S - \Psi_S\beta^*_S - W - V\right) + \lambda_n\widehat{g}_j = 0.$$

It then follows from equation (67) that

$$\begin{aligned}\widehat{g}_{S^c} &= \frac{1}{\lambda_n}\left\{\frac{1}{n}\Psi_{S^c}^\top\Psi_S\left(\beta^*_S - \widehat{\beta}_S\right) + \frac{1}{n}\Psi_{S^c}^\top(W+V)\right\} \\ &= \frac{1}{\lambda_n}\left\{\frac{1}{n}\Psi_{S^c}^\top\Psi_S\left(\frac{1}{n}\Psi_S^T\Psi_S\right)^{-1}\left(\lambda_n\widehat{g}_S - \frac{1}{n}\Psi_S^TW - \frac{1}{n}\Psi_S^TV\right)\right. \\ &\qquad \left. + \frac{1}{n}\Psi_{S^c}^\top(W+V)\right\},\end{aligned}$$

so that,

$$\widehat{g}_{S^c} = \frac{1}{\lambda_n}\left\{\Sigma_{S^cS}\Sigma_{SS}^{-1}\left(\lambda_n\widehat{g}_S - \frac{1}{n}\Psi_S^TW - \frac{1}{n}\Psi_S^TV\right) + \frac{1}{n}\Psi_{S^c}^\top(W+V)\right\}. \tag{77}$$

Condition (65b) requires that

$$g_j^T\left(\frac{1}{n}\Psi_j^T\Psi_j\right)^{-1}g_j \;<\; 1, \tag{78}$$

for all $j \in S^c$. Since

$$g_j^T\left(\frac{1}{n}\Psi_j^T\Psi_j\right)^{-1}g_j \;\le\; \frac{1}{C_{\min}}\|g_j\|^2 \tag{79}$$

it suffices to show that $\max_{j\in S^c}\|g_j\| < \sqrt{C_{\min}}$. From (77), we see that $\widehat{g}_j$ is Gaussian, with mean μ_j as

$$\mu_j = \mathbb{E}(\widehat{g}_j) = \Sigma_{jS}\Sigma_{SS}^{-1}\left(\widehat{g}_S - \frac{1}{\lambda_n}\left(\frac{1}{n}\Psi_S^TV\right)\right) - \frac{1}{\lambda_n}\left(\frac{1}{n}\Psi_j^TV\right).$$

This can be bounded as

$$\begin{aligned}\|\mu_j\| &\le \left\|\Sigma_{jS}\Sigma_{SS}^{-1}\right\|\left(\|\widehat{g}_S\| + \frac{1}{\lambda_n}\left\|\tfrac{1}{n}\Psi_S^TV\right\|\right) + \frac{1}{\lambda_n}\left\|\tfrac{1}{n}\Psi_j^TV\right\| \\ &= \left\|\Sigma_{jS}\Sigma_{SS}^{-1}\right\|\left(\sqrt{sC_{\max}} + \frac{1}{\lambda_n}\left\|\tfrac{1}{n}\Psi_S^TV\right\|\right) + \frac{1}{\lambda_n}\left\|\tfrac{1}{n}\Psi_j^TV\right\|. \qquad (80)\end{aligned}$$

Using the bound $\|\Psi_j^T V\|_\infty \leq D\, s/d_n^{3/2}$ from equation (71), we have,

$$\begin{aligned} \|\tfrac{1}{n}\Psi_j^T V\| &\leq \sqrt{d_n}\, \|\tfrac{1}{n}\Psi_j^T V\|_\infty \;\leq\; \frac{Ds}{d_n}, \quad \text{and hence,} \\ \|\tfrac{1}{n}\Psi_S^T V\| &\leq \sqrt{s}\, \|\tfrac{1}{n}\Psi_S^T V\|_\infty \;\leq\; \frac{Ds^{3/2}}{d_n}. \end{aligned}$$

Substituting in the bound (80) on the mean μ_j,

$$\|\mu_j\| \;\leq\; \left\|\Sigma_{jS}\Sigma_{SS}^{-1}\right\| \left(\sqrt{sC_{\max}} + \frac{Ds^{3/2}}{\lambda_n d_n} \right) + \frac{Ds}{\lambda_n d_n}. \tag{81}$$

Assumptions (36) and (37) of the theorem can be rewritten as,

$$\left\|\Sigma_{jS}\Sigma_{SS}^{-1}\right\| \;\leq\; \sqrt{\frac{C_{\min}}{C_{\max}}}\, \frac{1-\delta}{\sqrt{s}} \quad \text{for some } \delta > 0 \tag{82}$$

$$\frac{s}{\lambda_n d_n} \;\to\; 0. \tag{83}$$

Thus the bound on the mean becomes

$$\|\mu_j\| \;\leq\; \sqrt{C_{\min}}(1-\delta) + \frac{2Ds}{\lambda_n d_n} < \sqrt{C_{\min}},$$

for sufficiently large n. It therefore suffices in order for condition (65b) to be satisfied, to show that

$$\mathbb{P}\left(\max_{j\in S^c} \|\widehat{g}_j - \mu_j\|_\infty > \frac{\delta}{2\sqrt{d_n}} \right) \longrightarrow 0, \tag{84}$$

since this implies that

$$\begin{aligned} \|\widehat{g}_j\| &\leq \|\mu_j\| + \|\widehat{g}_j - \mu_j\| \\ &\leq \|\mu_j\| + \sqrt{d_n}\|\widehat{g}_j - \mu_j\|_\infty \\ &\leq \sqrt{C_{\min}}(1-\delta) + \frac{\delta}{2} + o(1), \end{aligned}$$

with probability approaching one. To show (84), we again appeal to Gaussian comparison results. Define

$$Z_j \;=\; \Psi_j^T \left(I - \Psi_S(\Psi_S^T\Psi_S)^{-1}\Psi_S^T \right) \frac{W}{n}, \tag{85}$$

for $j \in S^c$. Then Z_j are zero mean Gaussian random variables, and we need to show that

$$\mathbb{P}\left(\max_{j\in S^c} \frac{\|Z_j\|_\infty}{\lambda_n} \geq \frac{\delta}{2\sqrt{d_n}} \right) \longrightarrow \infty. \tag{86}$$

A calculation shows that $\mathbb{E}(Z_{jk}^2) \leq \sigma^2/n$. Therefore, we have by Markov's inequality and Gaussian comparison that

$$\begin{aligned}
\mathbb{P}\left(\max_{j\in S^c} \frac{\|Z_j\|_\infty}{\lambda_n} \geq \frac{\delta}{2\sqrt{d_n}}\right) &\leq \frac{2\sqrt{d_n}}{\delta\lambda_n}\, \mathbb{E}\left(\max_{jk} |Z_{jk}|\right) \\
&\leq \frac{2\sqrt{d_n}}{\delta\lambda_n} \left(3\sqrt{\log((p-s)d_n)}\, \max_{jk} \sqrt{\mathbb{E}\left(Z_{jk}^2\right)}\right) \\
&\leq \frac{6\sigma}{\delta\lambda_n}\sqrt{\frac{d_n \log((p-s)d_n)}{n}},
\end{aligned}$$

which converges to zero given the assumption (38) of the theorem that

$$\frac{\lambda_n^2 n}{d_n \log((p-s)d_n)} \longrightarrow \infty.$$

Thus condition (65b) is also satisfied with probability converging to one, which completes the proof. □

Proof of PROPOSITION 6.3. For any index k we have that

$$\begin{aligned}
\|f\|_2^2 &= \sum_{i=1}^{\infty} \beta_i^2 &(87)\\
&\leq \|\beta\|_\infty \sum_{i=1}^{\infty} |\beta_i| &(88)\\
&= \|\beta\|_\infty \sum_{i=1}^{k} |\beta_i| + \|\beta\|_\infty \sum_{i=k+1}^{\infty} |\beta_i| &(89)\\
&\leq k\|\beta\|_\infty^2 + \|\beta\|_\infty \sum_{i=k+1}^{\infty} \frac{i^\nu |\beta_i|}{i^\nu} &(90)\\
&\leq k\|\beta\|_\infty^2 + \|\beta\|_\infty \sqrt{\sum_{i=1}^{\infty} \beta_i^2 i^{2\nu}} \sqrt{\sum_{i=k+1}^{\infty} \frac{1}{i^{2\nu}}} &(91)\\
&\leq k\|\beta\|_\infty^2 + \|\beta\|_\infty C \sqrt{\frac{k^{1-2\nu}}{2\nu-1}}, &(92)
\end{aligned}$$

where the last inequality uses the bound

$$\sum_{i=k+1}^{\infty} i^{-2\nu} \leq \int_k^\infty x^{-2\nu}\, dx = \frac{k^{1-2\nu}}{2\nu - 1}. \tag{93}$$

Let $k^\star$ be the index that minimizes (92). Some calculus shows that $k^\star$ satisfies

$$c_1\|\beta\|_\infty^{-2/(2\nu+1)} \le k^\star \le c_2\|\beta\|_\infty^{-2/(2\nu+1)} \tag{94}$$

for some constants c_1 and c_2. Using the above in (92) then yields

$$\|f\|_2^2 \;\le\; \|\beta\|_\infty \left(c_2\|\beta\|_\infty^{(2\nu-1)/(2\nu+1)} + c_1'\|\beta\|_\infty^{(2\nu-1)/(2\nu+1)}\right) \tag{95}$$

$$= \; c\|\beta\|_\infty^{4\nu/(2\nu+1)} \tag{96}$$

for some constant c, and the result follows. □

Proof of Theorem 6.4. We begin with some notation. If $\mathcal{M}$ is a class of functions then the L_∞ bracketing number $N_{[\,]}(\epsilon, \mathcal{M})$ is defined as the smallest number of pairs $B = \{(\ell_1, u_1), \ldots, (\ell_k, u_k)\}$ such that $\|u_j - \ell_j\|_\infty \le \epsilon$, $1 \le j \le k$, and such that for every $m \in \mathcal{M}$ there exists $(\ell, u) \in B$ such that $\ell \le m \le u$. For the Sobolev space $\mathcal{T}_j$,

$$\log N_{[\,]}(\epsilon, \mathcal{T}_j) \le K\left(\frac{1}{\epsilon}\right)^{1/2} \tag{97}$$

for some $K > 0$. The bracketing integral is defined to be

$$J_{[\,]}(\delta, \mathcal{M}) = \int_0^\delta \sqrt{\log N_{[\,]}(u, \mathcal{M})}du. \tag{98}$$

From Corollary 19.35 of van der Vaart (1998),

$$\mathbb{E}\left(\sup_{g\in\mathcal{M}} |\widehat{\mu}(g) - \mu(g)|\right) \le \frac{C\, J_{[\,]}(\|F\|_\infty, \mathcal{M})}{\sqrt{n}} \tag{99}$$

for some $C > 0$, where $F(x) = \sup_{g\in\mathcal{M}} |g(x)|$, $\mu(g) = \mathbb{E}(g(X))$ and $\widehat{\mu}(g) = n^{-1}\sum_{i=1}^n g(X_i)$.

Set $Z \equiv (Z_0, \ldots, Z_p) = (Y, X_1, \ldots, X_p)$ and note that

$$R(\beta, g) = \sum_{j=0}^p \sum_{k=0}^p \beta_j\beta_k \mathbb{E}(g_j(Z_j)g_k(Z_k)) \tag{100}$$

where we define $g_0(z_0) = z_0$ and $\beta_0 = -1$. Also define

$$\widehat{R}(\beta, g) = \frac{1}{n}\sum_{i=1}^n\sum_{j=0}^p\sum_{k=0}^p \beta_j\beta_k g_j(Z_{ij})g_k(Z_{ik}). \tag{101}$$

Hence $\widehat{m}_n$ is the minimizer of $\widehat{R}(\beta, g)$ subject to the constraint $\sum_j \beta_j g_j(x_j) \in \mathcal{M}_n(L_n)$ and $g_j \in \mathcal{T}_j$. For all (β, g),

$$|\widehat{R}(\beta, g) - R(\beta, g)| \le \|\beta\|_1^2 \max_{jk} \sup_{g_j \in \mathcal{S}_j, g_k \in \mathcal{S}_k} |\widehat{\mu}_{jk}(g) - \mu_{jk}(g)| \tag{102}$$

where $\widehat{\mu}_{jk}(g) = n^{-1} \sum_{i=1}^n \sum_{jk} g_j(Z_{ij}) g_k(Z_{ik})$ and $\mu_{jk}(g) = \mathbb{E}(g_j(Z_j) g_k(Z_k))$. From (97) it follows that

$$\log N_{[\,]}(\epsilon, \mathcal{M}_n) \le 2 \log p_n + K \left(\frac{1}{\epsilon}\right)^{1/2}. \tag{103}$$

Hence, $J_{[\,]}(C, \mathcal{M}_n) = O(\sqrt{\log p_n})$ and it follows from (99) and Markov's inequality that

$$\max_{jk} \sup_{g_j \in \mathcal{S}_j, g_k \in \mathcal{S}_k} |\widehat{\mu}_{jk}(g) - \mu_{jk}(g)| = O_P\left(\sqrt{\frac{\log p_n}{n}}\right) = O_P\left(\frac{1}{n^{(1-\xi)/2}}\right). \tag{104}$$

We conclude that

$$\sup_{g \in \mathcal{M}} |\widehat{R}(g) - R(g)| = O_P\left(\frac{L_n^2}{n^{(1-\xi)/2}}\right). \tag{105}$$

Therefore,

$$\begin{aligned} R(m^*) &\le R(\widehat{m}_n) \le \widehat{R}(\widehat{m}_n) + O_P\left(\frac{L_n^2}{n^{(1-\xi)/2}}\right) \\ &\le \widehat{R}(m^*) + O_P\left(\frac{L_n^2}{n^{(1-\xi)/2}}\right) \le R(m^*) + O_P\left(\frac{L_n^2}{n^{(1-\xi)/2}}\right) \end{aligned}$$

and the conclusion follows. □

9. Acknowledgements. This research was supported in part by NSF grant CCF-0625879 and a Siebel Scholarship to PR. A preliminary version of part of this work appears in Ravikumar et al. (2008).

References.

Antoniadis, A. and Fan, J. (2001). Regularized wavelet approximations (with discussion). *Journal of American Statistical Association* **96** 939–967.

Buja, A., hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics* **17** 453–510.

Bunea, F., Tsybakov, A. and Wegkamp, M. (2007). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics* **1** 169194.

Daubechies, I., Defrise, M. and DeMol, C. (2004). An iterative thresholding algorithm for linear inverse problems. *Comm. Pure Appl. Math* **57**.

Daubechies, I., Fornasier, M. and Loris, I. (2007). Accelerated projected gradient method for linear inverse problems with sparsity constraints. Tech. rep., Princeton University. ArXiv:0706.4297.

Fan, J. and Jiang, J. (2005). Nonparametric inference for additive models. *Journal of the American Statistical Association* **100** 890–907.

Fan, J. and Li, R. Z. (2001). Variable selection via penalized likelihood. *Journal of American Statistical Association* **96** 1348–1360.

Greenshtein, E. and Ritov, Y. (2004). Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Journal of Bernoulli* **10** 971–988.

Hastie, T. and Tibshirani, R. (1999). *Generalized additive models.* Chapman & Hall Ltd.

Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric regression. *The Annals of Statistics* **28** 681–712.

Koltchinskii, V. and Yuan, M. (2008). Sparse recovery in large ensembles kernel machines. *COLT* .

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces: Isoperimetry and Processes.* Springer-Verlag.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics* **34** 2272–2297.

Meier, L., van de Geer, S. and Bühlmann, P. (2008). High-dimensional additive modelling. *arXiv* .

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34** 1436–1462.

Meinshausen, N. and Yu, B. (2006). Lasso-type recovery of sparse representations for high-dimensional data. Tech. Rep. 720, Department of Statistics, UC Berkeley.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381** 607–609.

Ravikumar, P., Liu, H., Lafferty, J. and Wasserman, L. (2008). Spam: Sparse additive models. In *Advances in Neural Information Processing Systems 20* (J. Platt, D. Koller, Y. Singer and S. Roweis, eds.). MIT Press, Cambridge, MA, 1201–1208.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, Methodological* **58** 267–288.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.

Wainwright, M. (2006). Sharp thresholds for high-dimensional and noisy recovery of sparsity. Tech. Rep. 709, Department of Statistics, UC Berkeley.

Wainwright, M. J., Ravikumar, P. and Lafferty, J. D. (2007). High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In *Advances in Neural Information Processing Systems 19* (B. Schölkopf, J. Platt and T. Hoffman, eds.). MIT Press, Cambridge, MA, 1465–1472.

Wasserman, L. and Roeder, K. (2007). Multi-stage variable selection: Screen and clean arXiv:0704.1139.

Yuan, M. (2007). Nonnegative garrote component selection in functional ANOVA models. *Proceedings of AI and Statistics, AISTATS* .

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* **68** 49–67.

Zhao, P. and Yu, B. (2007). On model selection consistency of lasso. *J. of Mach. Learn. Res.* **7** 2541–2567.

Zou, H. (2005). The adaptive lasso and its oracle properties. *Journal of the American*

Statistical Association **101** 1418–1429.

University of California, Berkeley
Berkeley, CA 94720 USA

Carnegie Mellon University
Pittsburgh, PA 15213 USA